Understanding Associations Between Smoking Behaviour and Poorer Health: Conventional and Mendelian Randomization Approaches

By

Mohammad A. Jareebi

November 4th, 2022

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy (Public Health), School of Health and Wellbeing, College of Medical, Veterinary & Life Sciences University of Glasgow, November 2022

WORD COUNT

47,506 words

(Excluding preliminary pages, references, and supplementary materials)

71,116 words

(Whole thesis)

ACKNOWLEDGMENT

First and foremost, I want to thank my supervisors, Dr Donald Lyall, Professor Daniel Mackay, and Dr Michael Fleming, for their invaluable advice, continuous encouragement, and patience throughout my PhD studies. Their immense knowledge and vast experience have inspired me throughout my academic research and daily life. I would also like to express my gratitude to the University of Glasgow, the UK Biobank, Jazan University, and the Saudi Embassy in London. Their kind assistance and support have made my studies and the overall journey a wonderful experience. Finally, I am very grateful to my wife and family. Their confidence in me has kept my spirits and motivation up throughout this journey.

For my parents' souls

"My Lord! Have mercy on them both as they did care for me when I was little"

Author's Declaration

I declare that this thesis was written entirely by myself and that it has not previously been presented, in whole or in part, in any application for a degree. The work offered is all my own, unless otherwise stated by reference or acknowledgement.

Mohammad Jareebi
November 2022

ABSTRACT

Background: Cigarette smoking is the leading preventable risk factor of morbidity and mortality in the world. Many studies have examined the association between smoking and health outcomes. Observational, cross-sectional studies can be confounded, and hence the casualty of associations between smoking and health outcomes cannot be established. A genetic epidemiological approach such as Mendelian randomization (MR) can be informative concerning potential causal associations between smoking and health outcomes. MR leverages the availability of genetic data on smoking and health outcomes to estimate confounder-free associations. The current thesis was carried out to investigate the observational and causal associations between smoking behaviour and cardiometabolic diseases, stroke, and lipid biomarkers.

Methods: Firstly, detailed reviews of prior research were conducted, highlighting that the majority of previous research was observational in nature. The thesis utilised the relatively large sample of UK Biobank (N=~502k) to conduct observational as well as MR-based analyses. The observational approach was based on self-report for multiple smoking phenotypes (smoking status, smoking intensity, and age at smoking initiation), clinical diagnoses for cardiometabolic diseases (CMDs; coronary heart disease (CHD), hypertension (HTN), and diabetes mellitus (DM)), stroke, and lipid biomarkers (total cholesterol, low-density lipoproteins, triglycerides, and high-density lipoproteins). The genetic analysis was based on 14 single-nucleotide polymorphisms (SNPs) for smoking intensity (cigarettes smoked per day: CperD) and 15 SNPs for smoking history. The genetic analysis was conducted in the UK Biobank sample (one-sample MR) as well as publicly available 'summary statistic' genetic data (two-sample MR). The analyses were conducted using R software and the MR-Base platform.

Results: Observationally (analysis: chapter four), current smokers had a higher risk of CHD (odds ratio [OR]: 1.61, P<0.001), stroke (OR: 1.64, P<0.001), and DM (OR: 1.12, P<0.001), and lower risk for HTN (OR=0.89, P<0.001) compared to never smokers. Additionally, as individuals smoke one more cigarette per day on average (smoking intensity), the risk for all CMDs increases (CHD, stroke, and HTN: OR=1.01, DM: 1.02, all P<0.001 per average daily cigarette). Finally, as an individual initiates smoking one year later in life, the risk of all CMDs decreases except for HTN (CHD and stroke: OR = 0.96, P<0.001, DM: OR=0.99, P>0.05, and HTN: 1.01, P<0.001). For lipid

biomarkers (analysis: chapter five), current smokers showed higher levels of cholesterol (β: 0.05 mmol/L, P<0.001), LDL (β: 0.06 mmol/L, P<0.001), and TG (β: 0.09 mmol/L, P<0.001), and lower level of HDL (β=-0.14 mmol/L, P<0.001) compared to never smokers. Similarly, as individuals smoke one more cigarette per day (smoking intensity), the levels of cholesterol, LDL, and TG increase, and the level of HDL decreases (cholesterol: β=0.02 mmol/L, LDL: β=0.03 mmol/L, TG: β=0.02 mmol/L, and HDL: β=-0.04 mmol/L, all P<0.001). Lastly, as an individual starts to smoke one year later in life, the levels of all lipid biomarkers increase except for TG (cholesterol: β=0.01 mmol/L, P=0.026, LDL: β=0.001 mmol/L, P>0.05, TG: β=-0.01 mmol/L, P<0.001, HDL: β=0.04 mmol/L, P<0.001). In terms of MR-based causal estimates (analysis: chapter four), there was no evidence of any causal relationship between smoking behaviour variables with CHD, stroke, and lipid biomarkers (analysis: chapter five) in the UK Biobank sample (one-sample MR) nor in other samples or approaches (summary-level in MR-Base platform or R). The only significant causal associations were observed in two isolated MR analyses; one between smoking status (ever) and HTN in one sample MR in the UKB sample and the other was between smoking intensity (CperD) and DM in two sample MR in R.

Conclusion: The observational findings indicated that cigarette smoking increases the risk of CHD, stroke, DM, and levels of total cholesterol, LDL, and TG observationally, but this was not supported by 'causal' genetic evidence. Smoking behaviour seems to be associated with lower blood pressure (observationally and genetically) and HDL levels (observationally, not genetically). Finally, findings on HTN, cholesterol, LDL, and HDL have varied depending on the smoking variable. These ambiguous findings point toward some of smoking's association with poorer health perhaps being due to poor lifestyle generally and not smoking itself in isolation. Evidence of potentially protective findings of smoking is likely to be driven by instrumentation or attrition bias. More research is needed to meticulously determine the impact of each smoking variable on health outcomes, both observationally and genetically.

Keywords: Smoking behaviour, Cardiometabolic disease, Stroke, Lipid biomarkers, Mendelian randomization, UKB

TABLE OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
A, C, T, G	Adenine, Cytosine, Thymine, Guanine
CHD	Coronary Heart Diseases
CI	Confidence Interval
CMD	Cardiometabolic Disease
CVD	Cardiovascular diseases
CO	Carbone Monoxide
COHb	Carboxyhaemoglobin
CPD or (CperD)	Cigarette Per Day (smoking intensity)
DM	Diabetes Mellitus
DNA	Deoxyribonucleic acid
GV(s)	Genetic Variant(s)
GWAS	Genome-Wide Association Studies
HDL	High-Density Lipoproteins
HTN	Hypertension
IV(s)	Instrumental Variable(s)
IVW	Inverse-variance weighted
LD	Linkage Disequilibrium
LDL	Low-Density Lipoproteins
MAF	Minor Allele Frequency
mmHg	Millimetre of Mercury
MR	Mendelian Randomization
NHS	National Health Service
NO	Nitric Oxide
OR	Odds Ratio
P	P Value
RR	Relative Risk
SI	Smoking Initiation
SNP(s)	Single Nucleotide Polymorphism(s)
TG	Triglycerides
TC	Total cholesterol
2SLS	Two-Stage Least Squares
UKB	UK Biobank

1. Chapter One: Introduction

1.1. Overview

Smoking is the single largest avoidable cause of death worldwide [1]. It is associated with a large number of health-related conditions ranging from a simple cough to chronic heart disease and lung cancer [1]. The burden of smoking on individuals and communities has motivated healthcare professionals and governments to understand the aetiology of the effects of smoking, smokers' behaviour and how to address this public health issue [2]. With a significant amount of money and lives lost attributed to smoking and smoking-related illnesses, researchers have explored the area of smoking extensively. These efforts were dedicated to understanding smoking behaviour and its impact on health [3].

1.2. Smoking burden

Globally, more than 1.1 billion people smoke, presenting a significant threat to public health [1]. Smoking kills around 7 million people each year, and this is expected to increase to 10 million a year by 2030 [1,4]. In the United Kingdom, in 2021, 13.3% of people aged 18 and above are smokers, with females less likely to smoke than males, 11.5% and 15.1% respectively [2]. Around 7.4 million of the UK adult population smoke with almost 100,000 fatalities attributed to smoking per year [5]. The most deprived areas have the highest proportion of smokers (30% compared to the least deprived areas 15%) [6]. White British and mixed groups have the highest proportion of smokers: 28% for mixed race and 20% for white. Black British and Asian British have the lowest proportion of smokers, with 15% for black and 12% for Asians [7]. Smokers have a shorter lifespan of almost 10 years compared to non-smokers [8]. Smoking attacks almost every organ in the body including the lung, heart, brain, and kidneys. For example, smokers are at almost double the risk of having a heart attack compared to non-smokers [5]. It is the major risk factor for lung cancer, and in the US,

smoking cigarettes is responsible for between 80 and 90 per cent of lung cancer deaths. The risk of developing or dying from lung cancer is 15 – 30 times more likely in smokers than in non-smokers [9]. In the UK, 72% of lung cancer cases are attributed to exposure to tobacco smoking [6]. Smokers have a higher risk of type 2 diabetes compared to non-smokers [10] and higher systolic blood pressure than non-smokers [11]. Smoking is also associated with an elevated risk of early menopause in women [12] and an increased risk of impotence in men [13]. In addition to the hazardous effects of smoking on health, it costs the National Health Service (NHS) around 2.5 billion pounds annually in England alone [5]. Smoking leads to substantial productivity losses that cost the UK economy around 8.4 billion pounds per year. Smokers need more social care in later life with an estimated cost of 1.4 billion pounds each year. In the UK, smoking costs approximately 12.6 billion pounds a year [3]. Additionally, smokers in England alone spend approximately 14 billion pounds on tobacco each year [5]. In the following section, a detailed outline of smoking and related health conditions will be discussed.

Smoking prevalence in the United Kingdom (UK)

In 2018, the Office for National Statistics in the UK revealed that 7.2 million (14.7%) adults in the UK were active smokers [14]. The report stated that Scotland had the largest proportion of active smokers compared to other UK regions. Scotland had 16.3% of smokers, Wales had 15.9%, Northern Ireland had 15.5%, and England had 14.4%. More than 7,900 deaths have been linked to smoking-related health issues, according to the report. In addition, nearly half a million people were admitted to UK hospitals due to smoking-related illnesses such as coronary heart disease and lung cancer. Compared to the 2018 report, the prevalence of smoking in the UK has declined in the last report of the office for National Statistics in the UK in 2019 [15]. Among

adults (≥18 years), the prevalence of current smokers has fallen from 14.7% to 14.1% (from 7.2 million to 6.9 million). The office reported that Northern Ireland had the highest proportion of current smokers compared to other UK regions (Northern Ireland: 15.6%, Wales: 15.5%, Scotland: 15.4, and England: 13.9%). Additionally, the proportion of smoking among men is higher compared to women (men: 15.9%, women: 12.5%). Despite this minor decline in the number of smokers in the UK, cigarette smoking remains a major public health concern. Various epidemiological approaches were carried out to determine the effects of smoking behaviour on health. The degree of the evidence supporting this effect of smoking differs depending on the type of epidemiological investigation.

1.3. Hierarchy of evidence in epidemiological studies

One of the fundamental aims of epidemiological research is to estimate the effect of exposure on an outcome. This goal is known as the causal effect of exposure on the outcome [16]. Observational studies might provide good insight and valuable correlations, but correlation does not imply causation [17]. In observational data, inferring a correlation between an exposure and an outcome as a causal relationship depends on unstable and implausible assumptions, such as the absence of unknown confounders and reverse causation. Such assumptions have led to incorrect causal estimation and improper public health intervention, prevention, and assessment measures [16]. For example, observational studies suggested a strong inverse statistically significant association between vitamin C and the risk of coronary heart disease (CHD) even after accounting for several confounding variables [18]. However, findings from a randomised controlled trial (RCT) have discredited this relationship and shown a non-significant causal association between vitamin C and CHD [19]. Similar incompatible findings were found between observational and experimental

associations between beta-carotene and smoking-related cancer [20,21], and between vitamin E and CHD [22]. More worryingly is the beneficial effect proposed by the observational studies of hormone-replacement therapy on breast cancer and cardiovascular diseases (CVDs) which were subsequently shown to increase mortality in an RCT [23].

Uncertainty of smoking behaviour using observational approaches

Examining smoking behaviour using a person's smoking history might produce biased results. A bias is a difference between a parameter's true value and its average estimated value [24]. These results might be due to external factors such as confounders. For example, when examining the association between smoking and coronary heart disease without considering BMI, age, sex, or other variables that affect heart disease and smoking, the findings obtained can be inaccurate, biased, and erroneous. These factors might nullify or magnify the conclusions of the study [25]. Furthermore, basing the results on people's memories or self-reports may introduce biases. For example, how much a person smokes, how often, for how long, when they start smoking, when they quit, how different ex-smokers are from current smokers, etc. may all contribute to recall bias, especially in long-term smokers and elderly [26]. One problem that might arise when using self-report is social desirability bias in which some individuals will not report their reality fearing people's judgment [27]. Moreover, when individuals experience a smoking-related condition and then quit smoking and are labelled as ex-smokers, this may result in reverse causation. Reverse causation arises when the outcome preceded the exposure [25]. Such a situation will be interpreted as if ex-smokers are more likely to be associated with such conditions compared to current smokers, making the results obtained from a smoking history or self-report approach uncertain. Considering these external factors as well as other biases that might arise

because of conventional approaches, more robust approaches should be considered. One approach is using an experimental study, namely a randomised control trial (RCT).

Randomised control trials (RCTs)

RCTs are considered the gold standard in epidemiological studies, ranking at the top of hierarchical evidence [28]. Compared to observational studies, RCTs overcome the weaknesses associated with observational studies such as confounders, reverse causation and biases such as selection bias [29]. RCTs reduce bias and ensure a stringent approach to examining the cause-effect relationships between exposures and outcomes [30]. RCTs use randomisation to balance known and unknown confounders between subgroups, resulting in confounding-free estimates [16]. The strengths of RCTs also include the adoption of a prospective approach with firm inclusion and exclusion criteria, a distinct intervention, as well as well-defined endpoints [29]. Whereas RCT is the gold standard design to determine the causal status of a particular risk factor, it does, however, have some limitations.

Randomised controlled trials can be expensive, laborious, and time-consuming, especially with rare outcomes or outcomes that require a long period of follow-up [16]. Moreover, several risk factors cannot be randomly assigned for pragmatic or ethical reasons. For example, when examining the impact of red wine on CHD, it would not be possible to recruit subjects to be randomly allocated to either drink or abstain from red wine over, for instance, 20 years. Finally, the subjects in RCTs are typically not representative of the larger population of interest [31]. To overcome the uncertainty around observational approaches and the difficulties experienced when conducting RCTs, a new approach, called Mendelian randomization, that combines observational data with the robustness of an RCT design can be considered. Mendelian randomization

is a genetic-based method that uses observational data to assess the causal relationships between exposures and outcomes.

1.4. Mendelian randomization (MR) approach

Introduction

Before explaining MR, a number of genetic-related terms need to be described briefly. DNA is a long double helix (two-stranded) molecule that has unique genetic information and makes up the human genome. There is a sequence of nucleotides that line up in each strand of the DNA. These nucleotides are adenine (A), cytosine (C), guanine (G), and thymine (T). The bonding between these nucleotides across the DNA strands is mostly as follows: A bonds with T and C bonds with G. For example, if a strand of DNA has this sequence "ACGTGCTA", the complementary strand will have "TGCACGAT". A change in the DNA sequence is called a gene variation. If one single nucleotide changes within the DNA sequence, this is called a Single-Nucleotide Polymorphism (SNP). An example of a SNP is when the previous sequence (ACGTGCTA) becomes "AC**A**TGCTA", A instead of A. These variations are responsible for the differences in certain characteristics (traits) between people, for instance, hair or eye colour. Additionally, genetic variations can also explain disease susceptibility [32]. The associations between certain characteristics (traits) such as smoking status and these genetic variants such as SNPs can be examined using Genome-Wide Association Studies (GWAS). After providing enough evidence of a significant (GWAS) association between a genetic variant and a trait, MR can be performed to examine the causal association between the genetically (not observationally) proxied trait and the outcomes of interest.

Mendelian randomization (MR) in practice

Mendelian randomization (MR) is a technique that uses genetic variants in observational data to establish causal inferences about the effect of a modifiable exposure on an outcome [16]. A Genetic variant (GV) is a piece of genetic code that naturally varies between individuals (i.e., randomly allocated). The genetic variant that will be used to conduct the MR analysis in this thesis is Single-Nucleotide Polymorphisms (SNPs) which is a variation of a single nucleotide at a specific genomic position [33]. With the help of Genome-Wide Association Studies (GWAS), SNPs are examined for significant association with a particular trait such as smoking status. Once the SNPs reached the GWAS level of significance ($p<5e-8$) [34], they can be used as a proxy for that trait to test the causal relationship with the outcome of interest using Mendelian randomization, assuming they are 'non-pleiotropic' – i.e. associate with that trait specifically alone [16]. In the MR approach, these variants are used as instrumental variables for assessing the causal effect of the exposure on the outcome [16]. An instrumental variable (IV) is a measurable quantity which is associated with the exposure of interest, but not associated with any other competing variables (confounders). In addition, the IV should ideally not be associated with the outcome of interest, except through the causal pathway via the exposure [35] (Figure 1.1). SNPs are the genetic variants that will be used and analysed throughout this thesis. Further details of MR, instrumental variables, GVs, and GWAS will be discussed in the following chapters (two and three).

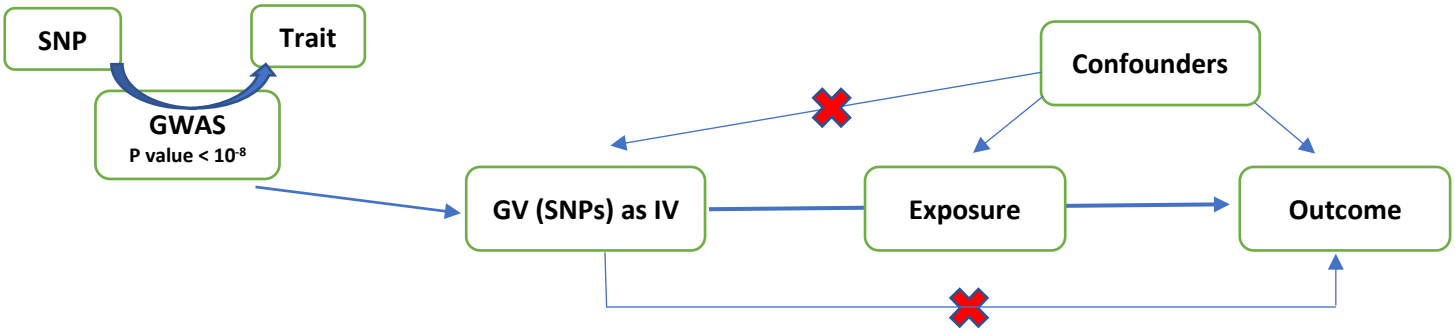


Figure 1.1. Mendelian randomization design

Types of Mendelian randomization

The MR approach can use one-sample or two-sample to achieve causal inferences [36]. One-sample (individual-level) MR uses one cohort in which both exposure and outcome data are from the same population. In contrast, two-sample MR uses two cohorts to obtain the data, exposure data from one population and outcome data from another population [36]. Further details on both approaches will be discussed in the methods section. The following section is an example of the use of MR in the literature.

Examples of Mendelian randomization applied to smoking and health outcomes

A study published in 2014 by Taylor et al. investigated the association of smoking behaviour with depression and anxiety and psychological distress using observational and MR approaches among 127,632 individuals [37]. They used self-reported data for the observational analyses and rs16969968/rs1051730 as a genetic variant (SNP) proxying smoking intensity for the MR approach. The results differed between the observational and MR approaches. In the observational analyses, current smokers had a statistically significantly higher risk for depression compared to never smokers (OR = 1.85, 95% CI: 1.65 – 2.07). These risks among current smokers were also higher and statistically significant for anxiety and psychological stress respectively (OR = 1.71, 95% CI: 1.54 – 1.90, OR = 1.69, 95% CI: 1.56 – 1.83). However, in the MR approach, rs16969968/rs1051730 was not associated with depression, anxiety, or psychological

stress (OR=1.00, 95% CI 0.95 – 1.05, OR=1.02, 95% CI 0.97 –1.07, OR=1.02, 95% CI 0.98 – 1.06, respectively). The authors concluded that there is no evidence of a causal relationship between smoking and these outcomes, suggesting the original observational association was due to confounding and/or bias. This example demonstrates the use of MR in the literature as well as how the associations based on observational approaches might be uncertain, especially in the presence of external factors.

Conclusion

MR is an analytical method that offers evidence about assumed causal associations between modifiable risk factors (exposures) and outcomes, using genetic variants as natural randomization tools. It provides an independent source of evidence which can be added to the current observational and RCT approaches. MR shows an advantage over conventional studies by avoiding confounding variables and reverse causation as well as overcoming the high cost and time factors accompanying RCTs.

1.5. Current study

The current thesis focuses on examining the associations between smoking and health outcomes observationally and genetically using the MR approach. MR uses genetic variants (SNPs), which are shown to be associated with smoking, to test the causal relationship between smoking and health outcomes. The health outcomes in this thesis will be divided into two parts. The first section will test for associations between smoking and cardiometabolic diseases (coronary heart disease (CHD), hypertension (HTN), and diabetes mellitus (DM)), in addition to stroke. CHD and stroke will also be sometimes referred to as cardiovascular diseases (CVDs). The second section will examine the associations between smoking and lipid biomarkers, which will include: total cholesterol, low-density lipoproteins (LDL), triglycerides (TG), and high-density

lipoproteins (HDL). The covariates in this thesis will include age, sex, deprivation score, body mass index (BMI), educational attainment, and ethnicity. The study will use UK Biobank (UKB) data to test these associations [38]. These variables were chosen based on the clinical literature concerning the effect of cigarette smoking on these outcomes as well as the availability of the data in the UKB (details will be provided in chapters two and three).

Both observational and MR approaches will be performed. The cross-sectional approach will be used to establish the ***observational*** associations between smoking and these outcomes, and the MR approach will be used to investigate the potential ***causal*** relationship between them. The study will also compare results from observational and MR associations as well as the magnitude of these associations. Finally, in addition to one-sample (individual-level) MR, the study will examine smoking associations using two-sample MR (using summary statistics).

The ultimate objective of this thesis is to enable casual associations between smoking and health outcomes to be estimated more reliably using Mendelian randomization approaches among UKB participants.

Contribution to the literature

The current thesis stands on three key elements: 1) a wide range of health outcomes including biomarkers, 2) a relatively large sample size, and 3) detailed covariates. It will examine the relationship between smoking behaviour and a wide range of outcomes in a large sample ($n \approx 502k$). The use of UKB data is a major strength considering the large sample size, methods of recruitment, and the availability and diversity of the traits, covariates, biomarkers, and genetic data. Moreover, previous studies examining smoking behaviour did not include all the variables, the SNPs or the covariates included in this thesis. As this study is mostly using one-sample MR, this

large number of participants is not comparable to other studies that used one-sample MR for smoking and these outcomes [39]. Additionally, the use of the MR approach in this UKB sample and the wide range of outcomes will establish a robust causal background about these relationships. Furthermore, testing the validity of the genetic variants (SNPs) proxying to smoking behaviour will provide a good understanding of these SNPs in the UKB which can be used with confidence in the future. Finally, the use of the MR approach will help to clarify the uncertainty in some relationships such as smoking with HTN and with lipid biomarkers as well as how generally different MR results are compared to observational ones. For example, smoking and lipid biomarkers have contradictory results in which some studies reported a positive association while others reported the opposite [40,41]. Such uncertainty can be alleviated using a variety of approaches, large sample sizes, different covariates as well as genetic analysis using MR. Therefore, this thesis will contribute significantly to the literature as well as to the UKB community.

Research aims and objectives

The primary objective of this study is to provide a detailed overview of the observational and causal relationships between smoking and different outcomes of interest among middle-aged to old-age adults using MR. The study will use quantitative data to assess the causal estimates of the relationship between smoking behaviour and cardiometabolic diseases (CMDs) related health outcomes (CHD, HTN, and DM) in addition to stroke and relevant biomarkers (total cholesterol, LDL, HDL, and TG) among middle-aged to older adults (40-70 years) in the UKB cohort population. The following are other objectives that will help to achieve the main goal of the thesis:

- Review the scientific evidence on the link between smoking and the outcomes of interest among middle-aged to older adults.

- Investigate the causal estimate of the relationship between smoking as an exposure vs. different outcomes (CHD, stroke, HTN, and DM) and biomarkers (total cholesterol, LDL, HDL, and TG) using the MR approach in the UKB population.
- Test (using MR) if the established relationships between smoking and the outcomes of interest are causal or not.
- Support (or not) the findings of the observational studies relating smoking to the outcomes of interest.
- Examine the validity of the SNPs associated with smoking behaviour in the UKB.
- Examine the causal associations of smoking with the outcomes of interest using one-sample and two-sample MR.

Research questions

The following are the specific research questions that will guide the process of identifying the associations between smoking behaviour and the outcomes of interest.

1- Are the instrumental variables valid to be used as a proxy for smoking behaviour in the UKB cohort population?
2- Is there a relationship between smoking behaviour and cardiometabolic disease (CMD) related health outcomes (CHD, HTN, and DM) and stroke, and if yes, is it causal?
3- Is there a relationship between smoking behaviour and lipid biomarkers (total cholesterol, LDL, HDL, and TG), and if yes, is it causal?
4- Do the findings drawn from the Mendelian randomization approach match the ones from the observational associations?

5- Do one-sample MR results (UKB cohort) match the ones from two-sample MR using other cohorts or other approaches?

Thesis roadmap

Table 1.1 shows the roadmap for this thesis demonstrating the main contents of each chapter.

Table 1.1. Thesis Roadmap

Chapter	Contents
Thesis main question	Examining the associations between smoking and CMDs as well as smoking and lipid biomarkers, observationally and genetically (MR).
Chapter 1: Introduction	Overview of smoking behaviour and exploring thesis outline, rationale, objectives, and questions.
Chapter 2: Literature review	A detailed review of smoking effects on the body followed by summary reviews of the observational and genetic (MR) associations between smoking and the outcomes of interest in the literature.
Chapter 3: Methods	Overview of the UKB and the methods of observational approach as well as the MR approach that will be used in this thesis.
Chapter 4: Smoking vs CMDs [analysis]	Detailed analysis of the relationship between smoking and CMDs; observationally and genetically.
Chapter 5: Smoking vs Lipid biomarkers [analysis]	Detailed analysis of the relationship between smoking and lipid biomarkers; observationally and genetically.
Chapter 6: Discussion and conclusion	Exploring and interpreting the results obtained from the analyses as well as answering the research questions and fulfilling the objectives proposed in this thesis and finally discussing the limitations, strengths, implications, and future research.

2. Chapter Two: Literature Review

2.1. Overview

The literature on smoking is quite substantial. Many different facets of smoking have been investigated. This chapter will provide a detailed review of smoking as a risk factor. The pathophysiological aspects of smoking, as well as the associations between smoking and CMDs, stroke and lipid biomarkers, will be covered in this review. It will also include a detailed review of Mendelian randomization and the use of smoking-associated genetic variants in the literature.

2.2. Pathophysiological aspects of cigarette smoking

When an individual smokes, the toxins from the tar in the cigarettes find their way into the bloodstream [42]. These toxins include tar and carbon monoxide (CO), nicotine and over 400 other toxins. Nicotine has stimulant and depressant effects. It deregulates cardiac autonomic function (the system that controls involuntary physiological processes such as heart rate) as follows: it increases the heart rate, narrows heart arteries (coronary), raises blood pressure, and stimulates the adrenal gland to release its hormones such as epinephrine and norepinephrine (catecholamines) [43]. Furthermore, nicotine has been linked to insulin resistance (a hormone that regulates blood sugar), higher lipid levels, and inflammation within blood vessels, all of which lead to the development of fatty substances within the arteries (atherosclerosis) [44]. Such an effect of nicotine magnifies when concord with the effect of free radicals.

Cigarette and free radicals

Smoke exists in two states: gaseous which contains CO and solid which contains tar. In both states, it has a large number of free radicals [45]. Free radicals are unstable molecules that can donate or accept an electron from other molecules [46]. These radicals can be oxidants or reductants with a harmful effect on cell function and homeostasis [47]. From cigarette smoke, 1 gram (g) of tar encompasses more than 10^{17}

long-lived free radicals (hours to months), while 1g of the gaseous portion has 10^{15} short-lived free radicals (seconds) [48]. Chronic exposure to cigarette smoke reduces the antioxidant defence mechanism that controls such a massive number of free radicals caused by smoking, leading to a substantial rise in oxidative stress [44]. Oxidative stress refers to imbalance between the synthesis and accumulation of oxygen reactive species (ROS) in cells and tissues which interfere with the detoxifying ability of the body [49]. Oxidative stress, and oxidation of proteins, DNA and lipids are related to the formation of fatty plaques within the arteries (atherogenesis) [50]. As a result, molecules such as isoprostanes (indexes of lipid peroxidation and oxidative damage) were found to be higher in smokers compared to non-smokers [49]. The abundance of free radicals affects the function of nitric oxide (NO) and alters its functions. Nitric oxide (NO) is an important molecule for blood vessel health. NO works as a vasodilator, which means it relaxes the inner muscles of blood vessels, causing them to dilate. Nitric oxide thus increases blood flow while decreasing blood pressure. The free radicals reduce the nitric oxide bioavailability, interfering with its anti-inflammatory and vasodilatory effects, in addition to its influence on endothelium permeability and myocardial function [51].

Carbon monoxide (CO) in cigarettes

Another cigarette component that alters NO is carbon monoxide. Carbone monoxide is substantially elevated in smokers resulting in the inhibition of the production of NO and replacing its position in haemoglobin bonds, hence, resulting in less oxygen delivery to the body's tissues [52]. Carbon monoxide (CO) is a product of the incomplete burning of carbon-containing materials such as tobacco [53]. The quantity of CO in cigarette smoke is 3-6% higher than normally encountered [54]. Carbon monoxide contributes to the build-up of cholesterol in the aorta and coronary arteries as well as enhancement of endothelial damage resulting in detrimental effects of

ischemic heart disease [55,56]. Hypoxia (reduction of oxygen supply to the tissues) is the key mechanism by which CO causes its effect on the heart [57]. These effects of CO poisoning will eventually lead to coronary heart disease, arrhythmias, and congestive heart failure.

In conclusion, toxins in cigarettes make the blood becomes denser with a rising risk of clot formation. Additionally, these toxins narrow the arteries, increase arterial wall thickness, and reduce the amount of oxygen-rich blood to be distributed throughout the body. This will increase the heart rate and blood pressure which eventually demands more effort from heart muscles. These pathophysiological changes are the main drivers of the health-related consequences of smoking.

2.3. Observational associations review

Several studies have investigated the associations between smoking and other common outcomes such as cardiovascular diseases, hypertension, diabetes, and biomarkers known to underlie poorer physical health. These observational studies have provided a holistic overview of the correlation between smoking and these outcomes. This review will focus on the observational relationship between smoking and the outcomes of interest. The same strategy of the literature review has been followed throughout this thesis including the observational and MR reviews. This section will be dedicated to reviewing the relationships between cardiometabolic diseases (CHD, HTN, and DM) and stroke as outcomes and smoking as a risk factor. The next section will focus on the relationship between smoking and biomarkers, namely, total cholesterol, LDL, triglycerides, and HDL. In a later review, pertinent examples of Mendelian randomization regarding smoking and other outcomes will be provided.

Literature review strategy

The literature related to the research topic was searched for on different digital platforms: PubMed (National Library of Medicine, Bethesda, Maryland), Web of Science Core Collection, NCBI (National Centre for Biotechnology Information), ScienceDirect (by Elsevier, March 1997), Cochrane CENTRAL (John Wiley & Sons, Inc., Hoboken) and Google Scholar. These platforms were chosen based on their medical and healthcare specialisation, relevance to the current thesis topic, and records coverage. The search was done using different keywords for smoking such as nicotine and tobacco. The search also involved other keywords such as smoking and cardiometabolic diseases (CMD) which included smoking and cardiovascular diseases (CVD), smoking and coronary heart disease (CHD), smoking and stroke, smoking and hypertension (HTN) and finally smoking and diabetes mellitus (DM). Additionally, it involved smoking and lipid biomarkers. Similarly, the platforms used to search the Mendelian randomization (MR) approach for smoking and other outcomes. Finally, it included searching for Genome-Wide Association Studies (GWAS) related to smoking and genetic variants (GV) such as Single Nucleotide Polymorphisms (SNPs). These terms were combined by the Boolean logic "AND" and "OR" to formulate the search strategy (smoke* OR cigarette OR tobacco) AND ("cardiovascular disease" OR "coronary heart disease" OR "stroke" OR "cerebrovascular disease" OR "hypertension" OR "high blood pressure" OR "diabetes mellitus" OR "high blood sugar" OR biomarkers OR cholesterol OR triglycerides OR LDL OR HDL). To search for the genetic data associated with smoking, smoking has been combined with ("Mendelian randomization" OR "Mendelian randomisation" OR "genome-wide association studies" OR "GWAS" OR "SNPs"). Smoking and smoking variants/synonyms was searched with each of the outcome variable in the thesis; for

example, (smoke* OR cigarette) AND (stroke OR “cerebrovascular disease”). This process was performed for smoking and all other variables. The search started in 2019 and revisited in 2020 and 2021. The references of the shortlisted articles were also evaluated to identify other relevant publications. The search was limited to human studies published in English.

The studies that were included (‘inclusion criteria’) had to have outcomes of tobacco/cigarette smoking, coronary heart disease, stroke, diabetes, hypertension, cholesterol level, LDL, HDL, and triglycerides among adults. Additionally, the studies included has to have these criteria: relevant research question, robust methods, large sample size, multiple covariates, having the same smoking variables examined in the current thesis. The inclusion criteria also involved being written in English and having undergone peer review. Studies that were not relevant to the title based on the title or the abstract were excluded. The studies had to portray smoking as a risk factor for the aforementioned outcomes. Non-human studies, small sample studies (less than 30: to ensure more robust, accurate, and generalisable findings), and other forms of smoking (non-cigarette) studies were excluded (Figure 2.1). The majority of these papers reported positive findings (only 10-15 papers have reported null/negative findings between smoking behaviour and HTN, total cholesterol, and LDL.

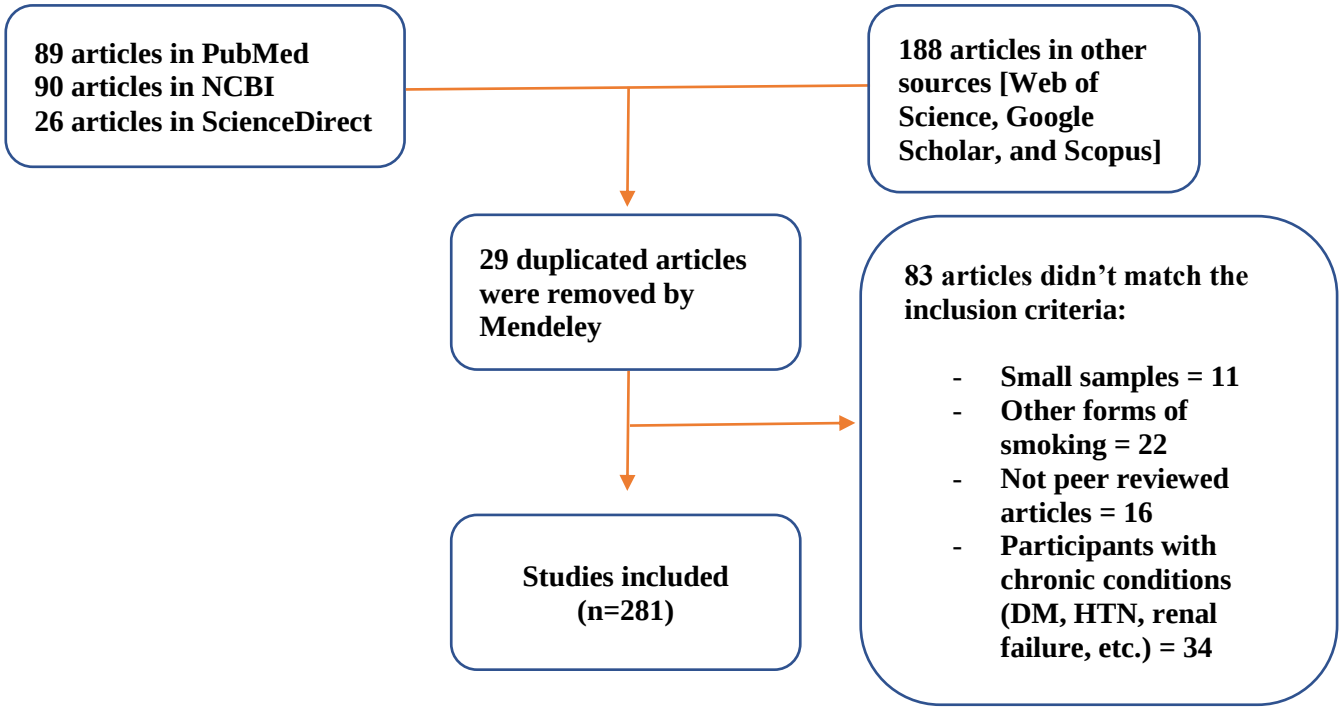


Figure 2.1. Research strategy graph (PRISMA)

Smoking and cardiometabolic diseases (CMDs)

Cardiometabolic diseases are a group of disorders that affect the cardiovascular, metabolic and renal systems [58]. This thesis will include coronary heart disease (CHD), hypertension (HTN) and diabetes mellitus (DM) as well as stroke. The following review will focus on smoking with CHD and stroke, then smoking with HTN and finally smoking with DM.

Smoking and CHD and stroke (CVDs)

Overview

Cardiovascular diseases (CVDs) are the number one cause of death worldwide. In 2016, approximately 17.9 million people died due to CVDs, this represents 31% of all global deaths. 85% of these deaths are attributed to heart attack and stroke [59]. Most CVDs can be stopped and prevented by monitoring modifiable risk factors such as smoking [60]. Smoking is a well-known risk factor for cardiovascular disease including coronary heart disease and stroke [61]. Even low-tar cigarettes and smokeless tobacco are shown

to increase the risk of cardiovascular events in comparison to non-smokers [44]. Each year cigarette smoking is accounted for around 140,000 premature fatalities from cardiovascular diseases [62]. In the UK, the Office for National Statistics (ONS) [2] reported that around half a million hospital admissions and 77,900 deaths are attributable to smoking. Cigarette smoking accounts for approximately 13% of all deaths in cardiovascular diseases [7]. Cigarette smoking affects the cardiovascular system directly via the chemicals inside the smoke, as well as the synergistic effects of other risk factors [62].

How smoking affects the cardiovascular system

Cigarette smoke contains more than 7000 chemicals with a harmful effect on cardiovascular function [63]. Cigarette smoking does not only directly affect the risk of CVDs, but it influences other sub-clinical cardiovascular risk factors, such as serum lipid levels (where low-density lipoproteins, aka LDL, are deleterious). The impact of smoking on CVDs risk is independent of its influence on the other risk factors [64]. Smoking seems to have a multiplicative interaction with other risk factors for heart disease [62]. For instance, the presence of smoking alone doubles the risk of CVDs, but the presence of another risk factor along with smoking will increase the risk by 4-fold (2 X 2), and if two more risk factors exist alongside smoking, the risk of CVDs will increase by 8-fold (2 X 2 X 2) [65].

Chemicals in cigarette smoke cause excess fluid to be trapped in the body's tissues (oedema) and inflammation in the blood vessel lining, which in turn cause most consequences of smoking on the cardiovascular system [60]. Atherosclerosis, one of the ramifications of oedema and inflammatory response caused by smoking, occurs when fat, cholesterol, and other substances in the blood form a plaque that accumulates in the arterial walls. The accumulated plaques lead to the narrowing of the arteries with

less blood flow throughout these arteries. Smoking enhances the formation of these plaques [66]. When the arteries supplying the cardiac muscles are narrowed by these plaques or blocked by clots, heart attack and sudden death can occur. Chemicals in cigarette smoke as mentioned in the previous chapter promote endothelial damage and clot formation inside cardiac veins and arteries [60]. In addition to atherosclerosis, stroke is one of the consequences of a shortage of blood flow to the brain caused by cigarette smoking. Smoking doubles the risk of death from stroke [67]. If an individual smokes 20 cigarettes per day, the risk of stroke is six times greater when compared to non-smokers [63]. The shortage of blood supply, higher cholesterol, clot formation, high blood pressure, and atherosclerosis are the mechanisms by which cigarette smoking trigger stroke [68]. The relationship between smoking and CVDs has been studied in depth and intensively in the literature. In the following section, a review of the literature on such a relationship will be provided. Significant numbers of observational studies exist in the literature examining the association between smoking and CVDs [64]. The following review will investigate the association between smoking and coronary heart disease and stroke.

Smoking and coronary heart disease (CHD)

A meta-analysis including 55 publications comprising 141 cohort studies has examined the association between cigarette smoking behaviour and cardiovascular disease [69]. The review included English-language articles published between 1946 and 2015 in Medline that described the association between smoking and coronary heart disease and stroke. More than 13861 abstracts were reviewed. They included prospective cohorts with at least 50 cardiovascular disease events. This was done to avoid including large but unreliable results seen in small studies. The term reporting bias includes a wide range of biases that revolve around what has been included/reported/published and what

has not [70]. The researchers excluded studies that have included high-risk individuals. The results of selected studies should be given separately for men and women and if the study is based on both combined, the results were adjusted for age and sex [69]. The measures of association were hazard ratio and relative risk. The study characteristics were country, time, period, sex, smoking categories, incidence, number of participants and confounding factors each study adjusted for. These factors include cholesterol, blood pressure, education, and BMI. They extracted hazard ratios (HR) and relative risks (RR) for coronary heart disease (CHD) and stroke [69]. From 24 studies, the researchers found a 48% higher risk of CHD for men who smoked one cigarette per day (RR = 1.48, 95% confidence interval 1.30 to 1,69), and 58% (RR = 1.58, 95% CI: 1.39 to 1.80) for those who consumed five cigarettes per day compared to non-smokers. The risk almost doubled for the individuals who smoked 20 cigarettes per day, RR = 2.04 (95% CI: 1.86 to 2.24). From 18 reports, the relative risks for women smokers were reported as follows, 1.57 (95% CI: 1.29 to 1.91), 1.76 (95% CI: 1.46 to 2.13) and 2.84 (95% CI: 2.21 to 3.64) for one cigarette per day, 5 cigarettes per day and 20 cigarettes per day respectively [69]. For stroke, among men who smoked one cigarette per day, the relative risk was 1.25 (95% CI: 1.13 to 1.83), for women, RR = 1.31 (95% CI: 1.13 to 1.52). the corresponding relative risks in 20 cigarettes per day were 1.64 (95% CI: 1.48 to 1.82) and 2.16 (95% CI: 1.69 to 2.75). From the above report, the risk of coronary heart disease increases with smoking and seems to be dose-dependent (the higher the individual smokes, the higher the risk of CHD: adjusted RR for men smoking 1 cigarette per day is 1.74 compared to 20 cigarettes per day RR = 2.27, this dependency stands for men and women, for CHD and stroke) [71]. Although this meta-analysis is robust including data from 141 separate cohort studies, based on around 3.07 million participants, it might have a few limitations. The smoking status reported in the studies

was not a specific quantity of smoking rather than categories which limits the regression modelling using the whole number rather than categories. Additionally, the number of cigarettes smoked is not the same each day which makes categorising the individuals not so accurate besides reporting and recall biases. Finally, the observation studies, in general, might not survive the unknown or residual confounders as well as the reverse causation [16].

Dose-response relationship between smoking and CHD

The prior results from the association between smoking and CVDs are consistent with many other observational studies [64]. The results of Law et al. (2011) review which included 19 studies are consistent with the previous review. The review was carried out on published data on environmental smoke exposure and CHD. They used Medline to identify the relevant studies in which exposure to smoking and the incidence of CHD was the key criteria [72]. Then, to determine the risk of CHD linked to a low dose of smoking, the researchers examined the dose-response relationship between smoking and CHD from five cohort studies for men recruited in the 1950s. They categorised the risk of CHD based on the number of cigarettes smoked, ranging from non-smokers who lived with smokers (low exposure) to one cigarette per day active smokers. The risk of CHD in non-smokers who lived with smokers was 30% (RR = 1.30, 95% CI: 1.22 to 1.38, P<0.001). The risk of CHD in the individual who actively smoked one cigarette per day was 39% (RR = 1.39, 95% CI: 1.18 to 1.64, P<0.001). The researchers also examine the risk of death from CHD in men who had quit smoking 20 years or more [72]. They discovered that the risk of death from CHD decreased to 6% (RR = 1.06, 95% CI: 1.02 to 1.10) when the individual stopped smoking. This reduction is considerable when compared to the risk of death from CHD in smokers in the Thun et al review (RR = 2.86, 95% CI, 2.65 to 3.08) [73]. According to this review, living with

smokers will increase the risk of CHD by 30% compared to a cigarette smoke-free environment. The review included only English language studies which might miss a huge number of studies of different languages and ethnicities. The exposure to smoking cannot be measured and it might differ from one individual to another, how much time the individual stays at home, sitting with direct contact with the smokers. Finally, there are numerous confounding factors in relation to CHD, such as hypertension, diabetes, cholesterol level, LDL level, physical activity, and diet type, that may influence the relationship between smoking and CHD.

Risk of death from CHD linked to smoking

One aspect of assessing the relationship between smoking and CHD is through the risk of death from CHD attributed to smoking, Thun et al examined such risk [73]. The study was a prospective cohort across three time periods (1959–1965, 1982-1988, and 2000-2010). The first period was from 1959 through September 1965 which included 183,060 men and 335,922 women. The second was from 1982 to December 1988 including 293,592 men and 452,893 women. The third period was from 2000 to 2010 and included the contemporary five most recent cohort studies and included 421.702 men and 535,054 women. The participants were aged 55 years or older. The majority of participants were white (90% – 97% among different studies), married (64% – 94%) and had a higher level of education (high school or less category: from 20.7% – 65.2%, some college: from 16.3% – 29.9%, and college or nursing school or more: from 13.5% – 49.4%). They defined the smoking status as follows: current, former, and never smoked. They discovered that current smokers have a higher risk of death from CHD compared to never smoked (RR = 2.86, 95% CI, 2.65 to 3.08) [73]. In further analysis, the researchers found that the risk of death from CHD among current smokers aged 55 - 74 has tripled (RR = 3.6 - 3.9, 95% CI: 2.9 to 5) [73]. The review included only

whites, married and so it may be hard to generalize such findings. As for most observational studies, it is hard to ignore the fact that confounding factors might play a major role in such a relationship between smoking and CHD or stroke in which a significant amount of risk factors are not measured or accounted for.

Smoking initiation and CHD

The Atherosclerosis Risk in Communities (ARIC) study highlighted the impact of age of smoking initiation, dosage, and time since quitting on CVDs in White and African Americans [74]. ARIC is a prospective cohort study of CHD, especially atherosclerosis, in different US communities. It included 14,200 men and women, aged 45 – 64 at baseline from 1987 – 1989. The participants were followed for 17 years in which they have been examined in three time periods, at baseline between 1987 - 1989, then revisited between 1996 – 1998, and finally by the end of December 2007. The researchers conducted an interview and clinical examination including cardiovascular risk factors and conditions. Questionnaires were utilised to measure the educational level, total yearly income, alcohol consumption, leisure time sports contribution, usage of antihypertensive and diabetic medications, and diagnosis of diabetes, coronary heart disease (CHD), or stroke [74]. The smoking status was defined as current, former, and never smokers with further analysis of current smokers based on the number of cigarettes smoked per day (CperD). The current smokers are classified into < 15 CperD, 15-24 CperD, 25-34 CperD, and ≥ 35 CperD. The age when the individual started to smoke (≤12 years, 13-15 years, 16-18 years, 19-21 years, and ≥22 years), packs per year, and smoking cessation (1-3 years, 4-9 years, ≥10 years quitting) during the follow up also collected [74]. The incidence of CHD was extracted from hospital discharge records and death certificates. The researchers found a higher hazard ratio (HR) of smoking on CVD in current smokers compared to non-smokers and previous smokers.

The adjusted HRs were 1.72 (95% CI: 1.30, 2.26), 1.67 (1.43, 1.95), 2.36 (1.88, 2.96), 2.69 (2.26, 3.19) in African-American men, white men, African-American women, and white women, respectively [74]. The researchers also explored the relationship between the number of cigarettes per day and CVDs. The overall adjusted HRs were 1.83 (1.27, 2.62), 2.68 (1.86, 3.85), 2.65 (1.79, 3.91), and 2.65 (1.79, 3.95) in < 15 CperD, 15-24 CperD, 25-34 CperD, and ≥ 35 CperD, respectively. The dose-response relationship between the cigarettes smoked per day and the risk of CVD is more prominent when examining the upper bound of the confidence interval and the P-value of the trend (P<0.0001) [74]. The study also examined the age of smoking initiation and the risk of CVDs and found the earlier the individual smoked, the higher the risk of CVD. The overall adjusted HRs (≥22 is the reference group) for age starting to smoke and the risk of CVDs were 2.52 (1.74, 3.63), 1.34 (1.01, 1.78), 1.28 (1.01, 1.62), and 1.14 (0.90, 1.45) in ≤12 years, 13-15 years, 16-18 years, 19-21 years, correspondingly. The researchers also assessed the smoking cessation and the risk of CVDs and discovered that the longer the period since quitting, the lower the odds of having CVDs but the results were statistically non-significant. The analysis of smoking cessation and CVDs risk revealed that the adjusted odds ratio (OR) were 0.87 (0.67, 1.14), 0.90 (0.96, 1.16), and 0.67 (0.45, 1.01) in 1-3 years, 4-9 years, ≥10 years quitting, respectively with P value of trend (P=0.06) [74]. As the study finding suggests, there is an increased hazard of CVDs among all current smokers compared to non-smokers, the African males, and all women. There was also a higher risk of CVDs when the individual started to smoke early and consumed higher cigarettes per day. The study also suggested the odds of having CVDs is reduced when you quit smoking, although the results were statistically non-significant. This robust cohort study was based on self-reported smoking status and not on objective measurements like cotinine which might lead to misclassification of

tobacco exposure. Additionally, in such long-term follow-up, reverse causation might be a problem in which cases could have quit smoking after the diagnosis of CVD. The previous papers were a detailed review of the relationship between smoking and CHD. In the next review and subsequent reviews, summary tables will sum up the associations between smoking and stroke as well as other CMDs (HTN and DM). These tables will include the design of the study, authors, exposure, outcome, sample size (N), findings and limitations.

Smoking and stroke review

Smoking is a major risk factor for all kinds of strokes. Current smokers have a two to four folds elevated risk of stroke compared to non-smokers [68]. In 2003, Kurth et al. published two papers on the risk of smoking on haemorrhagic stroke, one explored the risk in females and the other was in males [75,76]. Additionally, stroke risk increases as an individual smokes more, such a dose-response relationship was observed in two case-control studies; Bhat et al. and Fogelhom et al. [77]. Tables $2.1_{(a\text{-}d)}$ summarise some of the papers that examined the associations between smoking and stroke.

Tables 2.1 (a-d). Review of the associations between smoking and stroke

a) Study design and purpose	Prospective cohort (17.8 years) to study the risk of stroke among smokers in US physicians
Authors	Kurth et al [75].
Exposure	Smoking status (self-report).
Outcome	Stroke (haemorrhagic) (medical records). Covariates included were age, alcohol, and physical activity.
Sample Size (N)	22,022 (Males).
Findings	A statistically significant higher risk of stroke was found among smokers compared to never smokers: • **Previous**: RR = 2.36, 95% CI: 1.38 – 4.02. • **Current** (<20 CperD) RR = 2.06, 95% CI: 1.08 – 3.96. • **Current** (≥20 CperD) RR = 3.22, 95% CI: 1.26 – 8.18
Limitations	• The study included only the health workers which might bias the results (unrepresentative). • The study did not include other risk factors of stroke such as cholesterol level, drugs, blood diseases, and hypertension.
Conclusion	The study concluded that smokers have a higher risk of stroke compared to non-smokers.

b) Study design and purpose	Prospective cohort (9 years) to study the risk of stroke among smokers in US physicians
Authors	Kurth et al [76].
Exposure	Smoking status (self-report).
Outcome	Stroke (haemorrhagic) (medical records). Covariates: same as the previous (men) study.
Sample Size (N)	39,783 (Females).
Findings	Smokers had a statistically significant increased risk of stroke compared to non-smokers: • **Previous**: RR = 3.29, 95% CI: 1.72 – 6.29 • **Current** (<15 CperD) RR = 2.67, 95% CI: 1.04-6.90 • **Current** (≥15 CperD) RR = 4.02, 95% CI: 1.63 – 9.89
Limitations	• Only the health workers were included in the study which might bias the results. • The study did not include other risk factors of stroke such as cholesterol level, drugs (oral contraceptive pills), blood diseases, and hypertension.
Conclusion	According to the study, smokers were at an increased risk of stroke compared to non-smokers.

c) Study design and purpose	Case-control study (Dose-response relationship between smoking and the risk of ischemic stroke among 15-45 aged women).
Authors	Bhat et al [77].
Exposure	Smoking status (self-report). Covariates included: age, race, education category, HTN, DM, CHD, TC, and BMI.
Outcome	Stroke (ischemic) (medical records).
Sample Size (N)	* 466 cases (Hospital-based diagnosis). * 604 controls (Not diagnosed with stroke).
Findings	The odds of having stroke increase as the number of cigarettes smoked per day increases: • **Current (1-10 CperD):** OR = 2.2, P<0.0001. • **Current (11-20 CperD):** OR = 2.5, P<0.0001. • **Current (21-39 CperD):** OR = 4.3, P<0.0001. • **Current (>40 CperD):** OR = 9.1, P<0.0001. • The OR was 2.1 (P<0.0004) for 1-10 packs/year, 2.7 (P<0.0001) for 11-20 packs/year, and 4.8 (P<0.0001) for >21 packs/year.
Limitations	• The smoking status was based on self-report and not on objective measurements such as cotinine level. • Recall bias might be a problem in patients with stroke.
Conclusion	They concluded that there was a dose-response relationship between cigarettes smoked per day and the risk of stroke among smokers.

d) Study design and purpose	Case-control study (Dose-response relationship between smoking and the risk of haemorrhagic stroke)
Authors	Fogelholm et al [78].
Exposure	Smoking status (self-report).
Outcome	Stroke (haemorrhagic) - Computed Tomography (CT)-based diagnosis.
Sample Size (N)	158 patients confirmed stroke.
Findings	The odds of having a stroke are statistically significantly higher as the individual smokes more: • Current (1-20 CperD): OR = 3.33 (95% CI: 1.05,10.6) • Current (>40 CperD): OR = 9.78 (95% CI: 2.25,42.5)
Limitations	• The study has a small sample size, and the proportion of smokers is less than 20% of the sample. • Patients with ICH may have a memory problem which gives rise to recall bias. • The measurements of data collection are based on self-report, especially smoking status and other risk factors.
Conclusion	The study concluded that smokers have a higher risk of stroke compared to non-smokers. Additionally, this risk increases as the smoking intensity increases.

Summary

The review of the relationship between smoking and CHD as well as smoking and stroke revealed that cigarette smoking is a major risk factor for developing CHD and stroke. The earlier the individual started to smoke and the higher the dose (CperD), the higher the risk of CHD and stroke. The risk of CHD and stroke is reduced when individuals quit smoking, the earlier to quit, the lower the risk. The majority of studies were observational, and the effect of confounding variables and reverse causation is inevitable. Finally, most studies are based on self-reporting, hence reporting bias and recall bias might be a problem.

Smoking and hypertension (HTN)

Overview

Blood pressure (BP) is the pressure of blood against the arterial walls. It is measured by two properties, systolic and diastolic pressure [79]. Systolic pressure is the maximum pressure in the aorta during heart contraction, diastolic pressure is the minimum pressure in the aorta during heart relaxation [79]. In adults, the normal resting blood pressure ranges from 120 mmHg (systolic) and 80 mmHg (diastolic) [80]. Hypertension (HTN) is defined when the systolic blood pressure (SBP) is at least 140 mmHg and the diastolic blood pressure (DBP) is at least 90 mmHg [81]. Hypertension is a major public health problem accounting for more than 7.5 million deaths (12.8%) of all deaths annually [82]. In England, 1 in 4 adults suffer from hypertension and more than 5.5 million people have undiagnosed HTN costing the NHS over 2.1 billion pounds per year [83]. Hypertension is the 3rd largest risk factor of premature death after smoking, accounting for half of heart attacks and strokes. Hypertension is 30% higher in the most deprived areas [83]. Both hypertension and smoking play a major role in the development of cardiovascular diseases and other health conditions such as kidney diseases, but do they work synergistically?

The mechanism by which smoking interferes with blood pressure

The relationship between cigarette smoking and hypertension is not well understood. Some studies suggested that smoking increases blood pressure, but others do the opposite [84]. Nicotine found in cigarettes seems to the responsible for the changes in blood pressure. Acutely, it increases the systolic blood pressure by the vasoconstriction property on blood vessels, followed by the depressant effect of nicotine itself causing low blood pressure [84]. The debate is whether smoking contributes to high blood pressure, or the hypertensives tend to be smokers. It is not easy to answer such a

question, especially through conventional studies; however, great efforts have been done to explore the relationship between smoking and hypertension. The next sections will cover a review of the observational studies that examined the relationship between smoking and HTN in the form of summary tables.

Smoking and HTN review

This section presented a review of the relationship between smoking and HTN as portrayed in the literature. A summary of the main findings obtained from the observational approaches is shown in Tables 2.2$_{(a-d)}$.

Table 2.2 (a-d). Review of the associations between smoking and blood pressure

a) Study design and purpose	Systematic review To examine the findings on smoking and blood pressure in the literature across active and passive smokers
Author(s)	Leone A [84–87].
Exposure	Smoking status (self-report).
Outcome	Blood pressure (measurements and doctor diagnosis).
Findings	• Cigarette smoking in males was associated with a reduction in systolic blood pressure (SBP) by 1.3 mmHg (1.1%) in light smokers, 3.8 mmHg (3.1%) in moderate smokers, and 4.6 mmHg (3.7%) in heavy smokers compared to non-smokers [85]. • The same findings have been obtained by Gordon et al. which demonstrated the higher the cigarettes smoked, the lower the blood pressure [86]. • On the contrary, a trial conducted to evaluate the immediate effect of smoking on blood pressure found that SBP and DBP increased by about 10% and 7%, respectively [87].
Limitations	• This review is based on cross-sectional surveys and self-report smoking status. • It is difficult if not possible to ensure temporality in such types of studies. • Confounders play a huge role in observational analyses, especially in cross-sectional random surveys.
Conclusion	The article concluded that there was no correct answer to the question of the relationship between smoking and blood pressure as the studies continued to provide conflicting findings regarding this relationship.

b) Study design and purpose	Cross-sectional study To examine the relationship between smoking and uncontrolled blood pressure among hypertensive patients
Author(s)	Liu and Byrd [88].
Exposure	Smoking status (self-report).
Outcome	Blood pressure (measurements and doctor diagnosis).
Sample size	7,829 adult participants of both genders aged 18 years and above.
Findings	• Among hypertensive participants, the current smokers and former smokers have lower DBP by 1.3 mmHg (95% CI: -2.8, -0.2, P=0.02) and 0.9 mmHg (95% CI: -1.7, -0.03, P=0.04), respectively, compared to non-smokers. • The current smokers were 22% less to have uncontrolled BP (OR = 0.78, 95% CI: 0.64, 0.94, P<0.01) compared to non-smokers[88].
Limitations	• It is hard to infer causality from such a relationship or even temporality between smoking and hypertension. • Many confounding factors might have played a major role in this relationship, especially diet, lipid profile, physical activity, and medical history of diseases. • Smoking status based on self-report, and the duration of hypertension, antihypertensive drugs, and doses were not obtained which might prone the study to measurement bias.
Conclusion	Current smokers had better control of their blood pressure compared to non-smokers.

c) Study design and purpose	Cross-sectional study To examine the relationship between cigarette smoking and HTN
Author(s)	Alomari and Al-Sheyab [89].
Exposure	Self-reported questionnaire about smoking consumption. Covariates included: age, waist circumference, and BMI.
Outcome	Blood pressure (measurements: automatic oscillatory method).
Sample size	244 healthy youth of both genders aged 14-16 years.
Findings	• The results showed that the smokers were younger (P=0.001), less weight (P=0.001), and shorter (P=0.001) compared to non-smokers. • The smoking status explained 20.6% of the changes in SBP (R^2=0.206, F =46, P<0.001) and 5% of DBP (R^2=0.05, F=9.4, P<0.003). • The mean SBP and DBP in smokers were 108.8, and 55.4, respectively (P<0.001) and the mean SBP and DBP among non-smokers were 118.5, and 59.3, respectively (P<0.02). • SBP and DBP both were lower in current smokers compared to non-smokers (P<0.05)
Limitations	• The students in general have low blood pressure which was more prominent among smokers. • The study was cross-sectional with good survey information but not temporality or causation purposes. • The sample size was small, and the self-reported smoking status is unreliable, particularly in this age group. • The confounding variables and reverse causation in a cross-sectional study are almost unavoidable
Conclusion	The study concluded that smoking behaviour was having an inverse effect on blood pressure.

d) Study design and purpose	Cross-sectional study To examine the relationship between smoking and blood pressure
Author(s)	Li G et al [81].
Exposure	Self-reported smoking status. Covariates included were age, BMI, alcohol, and ethnicity.
Outcome	Blood pressure (measurements: digital device).
Sample size	1248 healthy men aged 20 -80 years.
Findings	• DBP and SBP were lower among current smokers compared to non-smokers (P<0.05). • In comparison to never-smokers, the odds of having hypertension among current smokers seem to be protective, 13% decrease in blood pressure (OR = 0.83, 95% CI: 0.61, 1.12), although the result was statistically non-significant, the odds are higher among former smokers, 48% (OR = 1.48, 95% CI: 1.01, 2.18).
Limitations	• The study is cross-sectional giving good survey information, but no causal estimation can be drawn. • The study is based on self-report smoking status as well as health conditions which give rise to inaccurate responses or different types of biases (recall, social desirability …etc.). • Many confounders might have magnified or nullified this relationship and even after statistical adjustment some residual or hidden confounders can bias the results, especially diet, physical activity, family history, and other medical conditions.
Conclusion	The study showed a consistent finding of an inverse relationship between smoking and blood pressure

On the contrary, some studies have unveiled that cigarette smoking is associated with higher blood pressure compared to non-smokers. A prospective cohort of 13,529 participants followed for 14.5 years to assess the relationship between smoking and hypertension. The researchers found that smokers have a higher risk to develop HTN compared to non-smokers (RR=1.15, 95% CI: 1.03-1.27, P=0.006) [90]. Tables 2.3$_{(a-b)}$ portray such findings.

Table 2.3 (a-b). Review of the associations between smoking and blood pressure

a) Study design and purpose	Cross-sectional study To examine the relationship between smoking and hypertension
Author(s)	McNagny SE et al [91].
Exposure	Self-reported smoking status. Covariates included were age, sex, alcohol, marital status, and education.
Outcome	Blood pressure (measurements: sphygmomanometer).
Sample size	216 hypertensive patients.
Findings	• The current smokers have higher odds of being severe uncontrolled hypertensives as well as less compliant with medications (OR = 4.17, 95% CI: 1.8, 9.5, OR = 2.33, 95% CI: 1.3, 4.1, respectively) compared to former smokers. • Both current and non-smokers were associated with uncontrolled HTN in compliant patients (OR = 14.4, 95% CI: 3.3, 63.3 and OR = 5.7, 95% CI: 1.5, 21.7, respectively) compared to former smokers.
Limitations	• The study was conducted in a very poor disadvantaged neighbourhood with less education and a high poverty rate which make it challenging to assume the generalisability of these results as well as the confounding effects of such factors on smoking and hypertension. • The diet, physical activity and weight were not recorded which might be responsible for the difference in the results especially when non-smokers have uncontrolled HTN despite compliance. • It is hard to infer causal estimation from a cross-sectional approach.
Conclusion	The finding of this study suggested that current smokers seem to have a higher risk of severe uncontrolled hypertension compared to former smokers

b) Study design and purpose	Cross-sectional study To examine the correlation between smoking and blood pressure
Author(s)	Al-Safi SA et al [92].
Exposure	Self-reported smoking status (including CperD): 1-10, 11-20, 21-30 and >31 cigarettes per day).
Outcome	Blood pressure (measurements: sphygmomanometer).
Sample size	14,310 healthy adults of both genders [Males were 7400 and females were 6910].
Findings	• Males: The SBP and DBP were significantly higher in smokers compared to non-smokers: Mean SBP: 126.24, 127.74, 129.67 and 129.11 in 1-10, 11-20, 21-30 and >31 cigarettes per day, respectively (P<0.0001). Mean DBP: 80.76, 80.97, 81.59 and 82.28 in 1-10, 11-20, 21-30 and >31 cigarettes per day, respectively (P<0.0001)). • The mean SBP and DBP of female smokers were significantly higher compared to non-smokers (P<0.001). • There was a positive dose effect of the correlation between smoking and hypertension as explained in the findings above.
Limitations	• Family history of hypertension was one of the confounding factors that played a major role in these findings (mean SBP was 120.99 (non-smokers + negative family history), 123,05 (non-smokers + positive family history, P<0.0001), 125.34 (smokers + negative family history), and 129.62 (smokers + positive family history, P<0.0001). • The results could have been more informative if logistic regression was done to measure the relationship between smoking and blood pressure with ORs to be presented as a measure of association with further adjustment for confounders such as family history, BMI and others. • The study was cross-sectional which provided good information about the prevalence of smoking and blood pressure among a large number, but it does not provide temporal relations or causation.
Conclusion	The study revealed that the SBP and DBP were significantly higher in smokers compared to non-smokers.

Summary

The relationship between smoking and hypertension seems to be ambiguous and a lot of confounding variables play a significant role in this association, such as family

history, diet, BMI, physical activity, and secondary causes of hypertension. Experimental longitudinal studies and other cheaper and more robust approaches like Mendelian randomization might be needed to prove or disprove the relationship between smoking and hypertension.

Smoking and diabetes mellitus (DM)

Overview and mechanism by which smoking affects blood sugar

Diabetes mellitus (DM) is a medical condition in which the level of blood glucose is increased. Glucose is a sugar that is produced from carbohydrate digestion and controlled by a hormone called insulin [93]. There are two types of DM, type 1 diabetes which develops when the insulin-secreting cells in the pancreas were destroyed, so the body cannot process the glucose which leads to the accumulation of this sugar in the blood and causes multiple complications [94]. The destruction of insulin-producing cells is thought to be caused by the immune system (autoimmune) with unknown aetiology behind this behaviour of the immune system [93]. Type 1 diabetes usually develops early during childhood, leaving the affected children dependent on insulin injections for life [94]. Type 2 Diabetes (DM) develops when the body cannot produce enough insulin, or the secreted insulin is not working properly. This type is more common in adults and is usually caused by modifiable environmental factors, such as obesity, physical inactivity or genetic predisposition like a family history of diabetes [95]. In 2016, the World Health Organization (WHO) reported that there are 422 million adults diagnosed with diabetes [93]. Annually, diabetes accounts for 1.5 million deaths worldwide. These deaths are from the complications of diabetes, such as CHD, stroke, kidney failure, and infections [93]. In the UK, 3.6 million people have diabetes (6% of the population), and 1.1 million have undiagnosed cases. Approximately, 700 people

are diagnosed with diabetes every day and more than 24,000 premature deaths in England and Wales are attributable to DM each year [96,97].

Smoking is an acknowledged risk factor for type 2 diabetes. Globally, smokers have a 30% - 40% higher risk for diabetes compared to non-smokers [98]. In the UK and US, it's estimated that smoking is responsible for 12% of DM which may account for 360,000 diabetic cases [99,100]. In the UK, more than 4.9 million people have diabetes. Additionally, around 13.6 million people are at risk of type 2 diabetes [101]. According to data from the UK Biobank, the prevalence of diabetes among smokers in the UK is approximately 12%. This is significantly higher than the prevalence among non-smokers, which is around 7%. It is important to note that these figures represent the overall prevalence of diabetes among smokers and non-smokers in the UK and do not consider other factors that may affect diabetes risk, such as diet, physical activity, and family history [102]. The risk of diabetes increases as the smoked cigarettes increase [103]. Smoking-related risk of diabetes is attributable to many mechanisms; first, its effect on insulin, it is believed that smoking is a risk factor for insulin resistance which is the core mechanism by which DM arises [104]. Secondly, tobacco smoking also inhibits glucose metabolism which gives rise to DM [103]. Finally, the mechanism by which smoking is believed to increase the risk of diabetes is oxidative stress, inflammatory responses to cigarette toxins, and abdominal obesity among smokers [98].

Smoking and DM review

A considerable number of studies have assessed the relationship between smoking and diabetes (DM), suggesting that cigarette smoking could independently interfere with glucose leading to impaired fasting glucose and DM, therefore smoking is believed to

be a modifiable risk factor for DM [105]. Tables 2.4$_{(a-c)}$ summarise some of the papers that examined the associations between smoking and DM.

Table 2.4 (a-c). Review of the associations between smoking and DM

a) Study design and purpose	Meta-analysis To examine the incidence of DM among smokers (between 1966 to 2007)
Author(s)	Willi et al [104].
Exposure	Self-reported smoking status.
Outcome	Diabetes (biological screening (blood or urine tests), personal or physician report of diabetes).
Sample size	25 articles with more than 1.2 million study participants.
Findings	• The pooled crude relative risk of smoking on DM in all studies was 1.89 (95% CI: 1.58 -2.27). • In a fully adjusted pooled RR, active smokers have a 44% increased risk of developing DM compared to non-smokers (RR = 1.44, 95% CI: 1.31 – 1.58). • Further analysis of active smokers based on cigarettes smoked per day showed a dose-response relationship, heavy smokers (≥ 20 CperD) were found to have a 61% increased incidence of DM compared to lighter smokers (29%) and former smokers (23%) (RR = 1.61, 95% CI: 1.43 – 1.80, RR = 1.29, 95% CI: 1.13 – 1.48, RR = 1.23, 95% CI: 1.14 – 1.33, respectively).
Limitations	• The review is based on observational studies which makes it hard to confirm causality, whether because of confounders (diet, physical activity, socioeconomic status and secondary causes of DM) or reverse causation. • They included old studies with a lack of information on the quality of participants and measures of recruitment. • The criteria to diagnose diabetes were old with a higher threshold of diagnosis which might have missed many cases of diabetes.
Conclusion	This meta-analysis of 25 studies showed a higher risk of DM among current smokers compared to never smokers with a dose-response relationship between smoking and DM.

b) Study design and purpose	Prospective cohort study (follow-up period of 23.5 years) To examine various predictors of DM
Author(s)	Lyssenko et al [106].
Exposure	Self-reported smoking status. Covariates included were age, sex, family history of diabetes, and, BMI.
Outcome	Diabetes (oral glucose tolerance test to measure blood glucose and insulin as well as fasting blood glucose).
Sample size	18,831 participants.
Findings	• In baseline unadjusted clinical factors only, current smokers were having 30% higher odds of developing type 2 diabetes compared to non-smokers (OR = 1.30, 95% CI: 1.18-1.43). • After adjustment, the risk has increased to 43% (OR = 1.43, 95% CI: 1.25 – 1.63, $P=1.4x10^{-9}$). • When genetic factors were added to the clinical factors, current smokers had a 39% risk of developing DM (OR = 1.39 (95% CI: 1.29 – 1.61, $P=6.3X10^{-8}$) compared to non-smokers.
Limitations	• In general, observational studies can estimate the risk but it is hard to infer causality. Confounding variables such as physical activity, diet, family history, socioeconomic status and secondary causes of DM might nullify or magnify the association between smoking and diabetes.
Conclusion	This powerful study has concluded that smoking is a strong predictor risk factor for DM.

c) Study design and purpose	A prospective study (follow-up period of 12 years) To examine the relationship between cigarette smoking and the incidence of DM
Author(s)	Manson et al [107].
Exposure	Self-reported smoking status. Covariates included were age, BMI, HTN, cholesterol, and physical activity.
Outcome	Diabetes (Self-reported).
Sample size	21,068 male participants.
Findings	• Compared to non-smokers, the relative risk of developing DM was: 1.7 (95% CI: 1.3 to 2.3) for current smokers of ≥ 20 CperD, 1.5 (95% CI: 1.0 to 2.2) for current smokers of < 20 CperD, and 1.1 (95% CI: 1.0 to 1.4) for past smokers. • They found a statistically significant association in smokers who smoke > 20 packs/year and non-significant results in less than 20 packs/year (1 – 19.9 pack/year: RR = 1, 95% CI: 0.8 – 1.3, 20 – 39.9 packs/year: RR = 1.3, 95% CI: 1 – 1.6, ≥ 40 packs/year: RR = 1.6, 95% CI: 1.3 – 2.1, P for trend <0.001).
Limitations	• The study was based on a self-report approach for both smoking status and diabetes which might have led to reporting and recall bias, especially regarding the cigarettes per day and packs per year. • There were no medical records on the health conditions and information was based only on participants' self-report. • The information about the family history of diabetes was not available and such variable plays a major role in predicting diabetes. • Finally, as an observational study, the causality is hard to be inferred and residual or hidden confounders might be a problem in this association
Conclusion	The study concluded that cigarette smoking is an independent modifiable risk factor of DM with a dose-effect phenomenon.

Summary

The relationship between smoking and DM seems to be robust and the risk of DM is higher among smokers compared to non-smokers. However, these studies are observational and mostly based on self-report smoking status. Observational studies usually suffer from the effect of known and/or unknown confounders. In the

relationship between smoking and diabetes, plenty of players alter the association between smoking and DM, for instance, diet, BMI, family history of diabetes as well as lipid biomarkers. Additionally, self-reported smoking or health conditions expose these studies to different biases such as recall bias, social desirability bias and selection bias. More robust approaches needed to be considered to eliminate or minimise the impact of the confounders and the biases associated with observational studies.

Smoking and lipid biomarkers

Overview and the mechanism by which smoking might affect lipid biomarkers

A biomarker is an objective (quantifiable) tool that measures normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention [108]. The biomarkers can be chemical, physical, or biological. Examples of biomarkers include body temperature, blood pressure, pulse, and serum LDL to more advanced imaging and molecular tests of tissues and blood [109]. Cigarette smoking is believed to be associated with significant changes among some biomarkers which include high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides (TG), and total cholesterol [40,110–112]. In the following review, the focus will be on the relationship between smoking behaviour and the aforementioned markers.

Cigarette smoking is believed to be associated with an increase in triglycerides (TG), LDL and cholesterol levels, and a reduction in HDL [40]. Smoking seems to affect lipids through nicotine which increases the secretion of free fatty acids and triglycerides along with lipoproteins from the liver into the bloodstream. This mechanism is enhanced by the stimulatory effect of nicotine on catecholamines (epinephrine and norepinephrine) secretion which leads to sympathetic stimulation resulting in increased lipolysis (the breakdown of fat) [113]. Cigarette smoking is also associated with an increased level of Homocysteine level which promotes the oxidative

alteration of LDL and decreases HDL [114]. A summary review of the relationship between smoking behaviour and lipid profile will be discussed in the following sections.

Smoking and lipid biomarkers review

The relationship between smoking and lipid biomarkers has been explored widely in the literature. The findings in the literature are controversial. For example, some studies found positive associations between smoking behaviour and total cholesterol, LDL, and TG, and negative association with HDL [40,110,115]. However, some studies found negative or non-significant associations between smoking and cholesterol and LDL [116,117]. Tables $2.5_{(a-e)}$ summarise some papers that examined the associations between smoking and lipid biomarkers.

Table 2.5 (a-e). Review of the associations between smoking and Lipids

a) Study design and purpose	Prospective cohort study To examine the effect of smoking on lipoprotein concentrations (LDL, HDL), total cholesterol and TG among current smokers.
Author(s)	Gossett et al [110].
Exposure	Self-reported smoking status (current vs non-smokers). Covariates included were age, sex, race, waist circumference, alcohol, physical activity, and use of lipid-lowering medications.
Outcome	Total cholesterol, LDL, TG and HDL.
Sample size	1,504 subjects of male and female participants.
Findings	• The study revealed that: HDL and HDL particles were low among current smokers (42 mg/dL, 30.3 μmol/L, respectively). • Cigarettes smoked per day (CperD) predicted higher total cholesterol (P=0.009), LDL (P=0.02) and total triglycerides (P=0.002).
Limitations	• The smoking status, medical history, and medical conditions were based on self-report which might bias these results. • Confounding factors such as diet, BMI, socioeconomic status and other causes of hyperlipidaemia might play a major role in this relationship.
Conclusion	They concluded that current smokers have a higher level of total cholesterol, LDL and TG and lower HDL levels compared to non-smokers.

b) Study design and purpose	Cross-sectional study To evaluate the effect of smoking on lipoprotein subfractions among current smokers compared to former and non-smokers.
Author(s)	Zhang et al [118].
Exposure	Self-reported smoking status (current vs former vs non-smokers).
Outcome	Total cholesterol, LDL, TG and HDL (electrophoretic technology).
Sample size	877 participants.
Findings	• The study found that the current smokers had a significant reduction in mean (± SD) HDL compared to non-smokers and former smokers (1.01 ± 0.26 vs. 1.06 ± 0.32 vs. 1.17 ± 0.36 mmol/L, P<0.001, after adjusted P=0.006, respectively). • The mean (± SD) of LDL was highest among current smokers compared to former smokers and non-smokers but was statistically non-significant (3.19 ± 0.87, 3.12 ± 0.88, 3.18 ± 0.87 mmol/L, P=0.707, adjusted P=0.554, respectively). • The mean (± SD) of TG among current smokers was higher compared to former and non-smokers but was statistically non-significant after adjustment (1.82 ± 0.81, 1.64 ± 0.68, 1.64 ± 0.79 mmol/L, P=0.002, adjusted P=0.09, respectively). • The mean (± SD) of total cholesterol was highest among non-smokers compared to current smokers and former smokers but was statistically non-significant after adjustment (4.8 ± 0.92, 4.70 ± 0.87, 4.65 ± 0.95 mmol/L, P=0.046, adjusted P=0.554, respectively).
Limitations	• The smoking status, medical conditions, and family history were self-reported which prone the study to recall and reporting bias. • The causality is hard to be inferred from such a design, in addition to confounding variables that might interfere with lipid parameters and smoking associations.
Conclusion	The study concluded that smoking is associated with a low level of HDL, and a higher level of LDL.

c) Study design and purpose	Screening To examine the effect of smoking on plasma cholesterol.
Author(s)	Muscat et al [115].
Exposure	Self-reported smoking status (current vs former vs non-smokers).
Outcome	Total cholesterol.
Sample size	51,723 US male and female participants.
Findings	• The plasma cholesterol levels were raised by 0.33 mg/dL (male, P<0.001) and 0.48 mg/dL (female, P<0.001) for each cigarette smoked compared to non-smokers and ex-smokers.
Limitations	• The study was a screening with no temporality nor causation to be drawn from this association. • The diet, socioeconomic status, education level, and physical activity are major confounders that might bias these findings.
Conclusion	They concluded that plasma cholesterol increased among current smokers with a dose-response relationship between CperD and plasma cholesterol.

d) Study design and purpose	Survey To examine the effect of smoking on lipid biomarkers.
Author(s)	Willett et al [40].
Exposure	Self-reported smoking status (current vs non-smokers). Covariates included were age, weight, height, blood glucose, resting pulse, and oral contraceptive use.
Outcome	Cholesterol, TG and HDL.
Sample size	191 female participants.
Findings	• The adjusted mean difference for TG and cholesterol is higher among current smokers compared to non-smokers (adjusted difference: 49.5 and 7.9, P<0.005, respectively). • The adjusted mean difference of HDL was lower among current smokers compared to non-smokers (- 7.3, P<0.005).
Limitations	• The study was a survey among a small sample size, which only included women. The data was based on self-report. These features might bias the results and limit the power of the study, with the possibility of increased variability and limited generalisability.
Conclusion	They concluded that smoking increases the level of TG and total cholesterol and decreases the level of HDL.

e) Study design and purpose	A systematic review of 54 published studies To examine the association between cigarette smoking in adults and serum lipid and lipoprotein concentrations.
Author(s)	Craig WY et al [119].
Exposure	Self-reported smoking status (current vs non-smokers).
Outcome	Cholesterol, LDL, TG and HDL.
Sample size	46557 participants.
Findings	• Among current smokers, the serum concentration of cholesterol, TG and LDL were 3% (P<0.001), 9.1% (P<0.001) and 1.7% (P<0.001), respectively, higher than non-smokers. • There was a dose-response relationship between light, moderate, and heavy smokers, compared to non-smokers and serum concentrations of lipids and lipoproteins. • The percentage differences increase for cholesterol, TG and LDL and decrease for HDL as the CperD increases (light, moderate and heavy smokers): Cholesterol: 1.8, 4.3, and 4.5%, respectively, TG: 10.7, 11.5, and 18%, respectively, LDL: -1.1, 1.4, and 11%, respectively, (P for trend <0.001), HDL: -4.6, -6.3, and -8.9% (P<0.001).
Limitations	• The review included studies that did not account for confounders like diet, physical activity, previous medical conditions, and others that might affect this relationship. • Some studies included in this review had a very small sample size which makes the results obtained from those studies questionable. • The results could have been more informative if beta coefficients were reported to quantify the relationship between smoking and lipid profile with further adjustment for confounders.
Conclusion	They concluded that the higher the individual smokes, the higher TG, LDL, and total cholesterol, and the lower HDL.

However, some studies found no significant relationship between cigarette smoking and lipid biomarkers. Tables 2.6$_{(a-c)}$ summarise these studies.

Table 2.6 (a-c). Review of the relationship between smoking and lipids

a) Study design and purpose	A cross-sectional study from National Health and Nutrition Examination Survey [NHANES] (1999-2012) To examine the association between cigarette smoking in individuals ≥20 years and lipid biomarkers.
Author(s)	R. Jain and A. Ducatman [117].
Exposure	Self-reported smoking status (smokers vs non-smokers) and cotinine levels. Covariates included were sex, ethnicity, alcohol, caffeine, BMI, and poverty income ratio.
Outcome	Cholesterol, LDL, TG and HDL (mg/dL).
Sample size	15276 participants.
Findings	• Smokers and non-smokers have the same levels of adjusted cholesterol and LDL levels: Cholesterol (smokers): 193.9 (95%CI: 185.6 – 202.6, P>0.05) Cholesterol (non-smokers): 193.9 (95%CI: 185.5 – 202.7, P>0.05) LDL (smokers): 113.3 (95% CI: 106.6 – 120.4, P>0.05) LDL (non-smokers): 113.6 (95% CI: 106.9 – 120.7, P>0.05) • Smokers have lower levels of HDL compared to non-smokers (48.8, 51.4, respectively, P<0.01). • Smokers have higher levels of TG compared to non-smokers (124.4, and 111.9, respectively, P<0.01).
Limitations	• The study was a survey with no temporality nor causation to be drawn from this association. • The study included a wide range of covariates however the adjusted R^2 barely reached 28% which might point toward confounded findings.
Conclusion	They found that smoking was associated with an adverse effect on lipid biomarkers. However, there was no significant difference between smokers and non-smokers concerning cholesterol and LDL levels.

b) Study design and purpose	Cross-sectional study To examine the effects of cigarette smoking on serum lipids.
Author(s)	Saengdith. P [116].
Exposure	Self-reported smoking status (non-smokers, ex-smokers, current smokers).
Outcome	Cholesterol and TG (mg/dL).
Sample size	401 priests.
Findings	• Cholesterol: No statistically significant difference among all smoking categories (P=0.22). • Triglycerides: There was a statistically significant difference between smoking groups (P=0.02).
Limitations	• The study was cross-sectional with self-reported smoking data. • There were no covariates added to the analysis which makes these finding highly susceptible to confounding effects. • There were no models built to adjust for other variables that might affect these findings. • These findings are hardly generalisable as the study included Thai priests.
Conclusion	The researcher concluded that cigarette smoking increases the level of TG but not cholesterol.

c) Study design and purpose	Cross-sectional study To examine the association between smoking habits and lipids
Author(s)	Moradinazar et al [120].
Exposure	Self-reported smoking status (non-smokers, former smokers, current smokers). Covariates included were gender, age, physical activity, and wealth index.
Outcome	Cholesterol, LDL, TG and HDL (abnormal vs normal level).
Sample size	7586 participants.
Findings	• The current smokers are compared to non-smokers if they have any abnormal lipid biomarkers. • No significant association between smoking status and abnormal cholesterol and LDL: Cholesterol (smokers compared to non-smokers): OR = 0.85 (95% CI: 0.65 – 1.13) LDL (smokers compared to non-smokers): OR = 0.50 (95% CI: 0.26 – 1.08) • Compared to non-smokers, current smokers have a higher risk for abnormal HDL and TG (OR= 2.28, 95% CI: 1.98 -2.62 and OR= 1.37, 95% CI: 1.15 -1.67, respectively).
Limitations	• The study was based on self-reported smoking status (recall bias). • Causality can hardly be inferred from such an approach considering the lack of temporality as well as the presence of confounding variables.
Conclusion	Current smokers reported more risk for abnormal HDL and TG compared to non-smokers. However, there were no significant associations observed between smokers and cholesterol and LDL variables.

Summary

Cigarette smoking seems to increase total cholesterol, LDL and TG and decrease HDL as the previous review suggested. The review also suggested a dose-response relationship between smoking and lipid biomarkers. However, some papers found no association between smoking and cholesterol or smoking and LDL. The findings obtained from this review are based on observational studies. This approach gives us an overview of the prevalence/association between variables but not causality as confounding variables and reverse causation are almost inevitable in the observational approaches. A self-report measure of obtaining smoking status and other medical

conditions might prone the observational studies to different biases which might affect the validity of the study as well as the findings obtained from such approaches. To overcome these issues, another approach, such as MR, that ensures valid measures and robust inferences would be recommended to examine smoking and other variables.

Conclusion

Smoking has a detrimental impact on health with a significant burden on public health, governments, and individuals. Billions of pounds have been invested to combat smoking, and billions have been spent to overcome complications caused by smoking and smoking-related conditions. Millions of lives are lost attributable to smoking through varieties of health conditions such as CVDs, HTN, DM, lung cancer, and others. The literature review in this section showed that smoking is associated with CVDs, HTN, DM and lipid modification. These relationships were observational which makes it hard to infer causality and sometimes even temporality. Moreover, the existence of confounding variables (hidden and residual) and reverse causation make it even harder to establish a causal relationship. Finally, the self-reported approach might give rise to biases which affect the results obtained from the observational studies that follow this approach, especially in smoking status. In such a situation, other approaches such as MR, which uses genetic proxies for variables, might be needed to examine the causal relationships between smoking and other outcomes.

2.4. Mendelian randomization review

Mendelian randomization as stated in the introduction is a technique of using genetic variants to infer causality when examining the association between modifiable risk factors and outcomes [16]. The following sections will explore the MR approach in detail. This included the rationale for using MR, the genetic variants (GVs), the concept of an instrumental variable (IV), the success of MR, and finally the limitations of MR.

Why Mendelian randomization?

The chief aim to use MR is to limit the role of confounders and reverse causation that can accompany conventional studies. Therefore, MR was historically named "Mendelian deconfounding" [121]. It gives estimates of the causal effect of the exposure on the outcome free of biases caused by confounders. Deconfounding is a fundamental principle in which the emphasis is to prove or disprove a hypothesis on a particular relationship between an exposure and an outcome proposed by conventional studies. However, MR extends this concept and not only confirms or refutes the hypothesis but also provides an estimate of the size of the unconfounded effect of this relationship with a measure of its uncertainty [121]. MR is a broader term which generates indirect, and unconfounded, inferences about the relationship between a trait and an outcome given direct information on the gene–outcome and gene–trait associations [122].

Unconfounded associations

A confounder is a variable that is a common cause of the exposure and the outcome, so it can magnify or nullify the relationship between them. The existence of unknown or residual confounders may bias the causal estimate between the exposure and the outcome [16]. In observational studies, it is difficult to isolate the effect of one variable while keeping all other risk factors equal, as the change in one risk factor will often be accompanied by changes in other factors. Although we can measure and adjust for the individual confounders, we will never be certain if all confounders were identified or precisely measured, which gives rise to residual confounders. Additionally, the adjustment might also include a true variable on the causal relationship between the exposure and outcome (a mediator) which lead to an over-adjustment attenuating the casual estimate [123]. When finding a genetic variant that satisfies the assumption of

the instrumental variable, MR can estimate an unconfounded association between the exposure and the outcome. As the genetic variants are assigned randomly at birth, MR is considered a natural RCT. Additionally, MR also overcomes the issue of reverse causation seen in observational studies [124].

Mendelian randomization and reverse causation

Reverse causation develops when an outcome is causing exposure. This might occur if the exposure increased in response to a pre-clinical condition, like the association between CRP (C-reactive protein) and CHD. The onset of the inflammatory response that raises CRP may be caused by atherosclerosis, a pre-clinical condition that occurs before the clinical manifestation of CHD. Due to reverse causality, it is possible in this scenario to erroneously link CRP to CHD. [16]. As MR is based on genetic variants, and these genes are determined before birth and cannot be changed, there is no way of reverse causation that can be responsible for the relationship between the exposure and the outcome [125].

Cost-effectiveness

MR can be valuable when the exposure is expensive or hard to quantify. For example, measuring an exposure like water-soluble vitamins for a large sample would be very costly and might not be affordable. Likewise, measuring fasting blood glucose requiring overnight fasting may be impractical. If the genetic variant is associated with the exposure of interest and is valid as an instrumental variable for the exposure, a causal inference of the exposure on the outcome can be established from an association between the GV and the outcome even in the absence of measurement of the exposure [16]. Considering conventional studies and RCT limitations, MR seems to be a practical choice to assess the causal estimate between an exposure and an outcome.

Genetic variants in Mendelian randomization (MR)

In the MR approach, genetic variants are used as an instrumental variable to proxy the risk factor of interest. In analogy to RCT, people are divided into subgroups, randomised by genetic variants [126]. Genetic variants (GVs) should be distributed randomly in the population, independent of environmental and other variables so that when categorising individuals based on GVs, the subgroups should not systematically differ in these variables. Furthermore, as the genes are determined before birth, there is no chance that a measured variable in a mature individual can be the cause of the GV [16]. Genetic variation is mostly randomly distributed across the population; hence, randomization can be leveraged to assess the causal relationship between variables analogous to the wings of an RCT. For instance, some people are born with a variation in the nicotine receptor gene making them more susceptible to smoking or having more complications compared to individuals not having this variant gene. The single nucleotide polymorphism (SNP) rs1051730 on chromosome 15 in the *neuronal nicotinic acetylcholine receptor gene (CHRNA 5)* is one example of such variation that has been linked to smoking intensity [127]. In such a situation, individuals can be assigned based on this variant and explore the outcomes based on this "natural" difference between them [127]. Genetic variants should also meet the assumptions of the instrumental variable to be used in MR (Figure 1.1).

Instrumental variable in Mendelian randomization

To manage the problem of confounders and reverse causation raised by using conventional studies, the instrumental variable was introduced. A technical description of MR is "instrumental variable analysis using genetic instruments" [128]. The instrumental variable is a technique used to estimate causal effects without the comprehensive familiarity of all confounders of the relationship between the exposure

and outcome [16]. In MR, genetic variants are used as instrumental variables to examine the causal influence of the exposure on the outcome [129]. For a genetic variant to be utilised to estimate a causal effect of the exposure on the outcome, it must satisfy these assumptions of an instrumental variable:

- The GV should be associated with the exposure (GWAS significance level).
- The GV should not be associated with any competing risk factor (confounders) of the exposure-outcome association.
- The GV does not affect the outcome, except through the exposure causal pathway (pleiotropy free).

The first assumption ensures that the genetic subgroups defined by the variant will have diverse levels of exposure, which guarantees systematic differences between the subgroups. If the association between the GV and the exposure is not strong, weak instrument bias arises. The second assumption ensures that all other variables (confounders) will be distributed equally between subgroups. The third assumption states that the only causal pathway between the GV and the outcome should be via exposure [16]. In addition to these assumptions, the biological plausibility of a genetic variant as an instrumental variable should be justified.

The validity of the IV is mandatory in the MR approach. The choice of GV as an IV should be principally justified biologically, but it can be substantiated statistically [16]. The GV with a well-understood biological function is more credible to be used in MR compared to the variants discovered outside the gene coding region. Biological plausibility is the backbone to justify the validity of GV as an IV in MR. To assess the biological plausibility of the validity of the GV as an IV, Bradford Hill criteria for

causation to MR can be applied [130] (Table 2.7). The statistical assessments for the IV assumptions will be discussed in the methods chapter.

Even though the MR approach used to involve one genetic variant (SNP), the trend has shifted toward using more than one SNP which can be exploited to build a genetic score [131]. The genetic score (allele score) is the sum of weights that each SNP contributes to explaining the variation within a trait [132]. Because of the small amount of variation that can be explained by a single SNP, the use of multiple SNPs (IVs) increases the statistical power and the precision of IV estimates. The same reason above explains why MR needs a very large sample size [132].

Table 2.7. Bradford Hill criteria for judging the biological plausibility of IV

Criteria	Description
Consistency	Existence of multiple GVs associated with the same exposure and all associated with the same outcome (more causal plausibility).
Biological gradient	Dose-response relationship between the GV and the exposure and the outcome.
Specificity	A more plausible causal relationship is if the GV is associated with a specific risk factor and specific outcome. A more specific relationship would be achieved when the GV is biologically close (proximal) to the exposure.
Biological plausibility	If the biological function of the GV on exposure is well-known, the causal relationship will be more plausible.
Strength	If the association between the GV and the outcome is low, then the association between the GV and the covariates would be low as well.
Coherence	The finding (outcome) of the experimental intervention on the exposure would be the same in the genetic context.

These assumptions can be violated and the GVs are less likely to be a valid instrumental variable, hence weakening the findings of the causal estimate for an exposure [16]. These violations are summarised in Table 2.8.

Table 2.8. Violations of IV assumptions

Violation	Description
Biological mechanisms	• **Pleiotropy**: when the GV is associated with multiple risk factors (one gene → > one trait) [133]. If a GV is associated with another risk factor for the outcome, this variant is invalid. The known biological functions of the GV can alleviate pleiotropy. • **Canalization**: the production of the same phenotype in a population regardless of its genetic or environmental variations. The gene might be inactive, but somehow the function is present (different compensatory biological mechanisms) [134].
Non-Mendelian inheritance	• **Linkage Disequilibrium (LD)**: non-random inheritance of the GVs caused by close physical proximity on the same chromosome [135]. Variants with correlated distributions are said to be in LD. If a GV is independently causing the variation in the exposure, then it can be used as an IV. The GV doesn't need to be a causal variant itself, being in (high) LD with a causal variant is enough to be a valid IV [25]. The problem that might arise from the LD is the association of the GV with a variant that might influence competing risk factors for the outcome which violates the second or the third assumption. This problem can be alleviated by testing the

	association between the measured GV and the known confounders [16]. • **Effect Modification**: occurs when the level/magnitude of the effect of an exposure on an outcome differs depending on a third variable (a modifier). The genetic associations with the outcome might appear only in men, for example, hence gender is a modifier and further analysis should be conducted (interaction term) [136].
Population effects	• **Population stratification**: a systematic difference in allele frequencies between subpopulations in a population. For example, a population is composed of a mixture of different ethnic groups. The variations in the population might be attributable to the subpopulation differences and not the effect of the GV. To overcome this dilemma, the study could be restricted to a specific ethnic group [16]. • **The ascertainment effect** occurs when the recruitment in the study is based on genetic variants, which could lead to unrepresentative findings of the association between the GV and the outcome in the original population. If the study cohort is taken from the general population, such an effect would not be a problem in practice.

In general, genetic variants have good theoretical and practical plausibility to be used as instrumental variables. The IV assumptions evaluate the causation in an observational situation without complete knowledge of all the confounders of the association between the exposure and the outcome. These assumptions can be violated risking the validity of the causal inference in Mendelian randomization.

The success of the Mendelian randomization (MR) approach in the literature

MR design was first proposed in 1986 [137] and described by Gray and Wheatly in 1991 [138]. With the availability of cheaper DNA sequencing techniques for large individuals, the MR approach has increasingly been used in the last few years [139–141] (Figure 2.2).

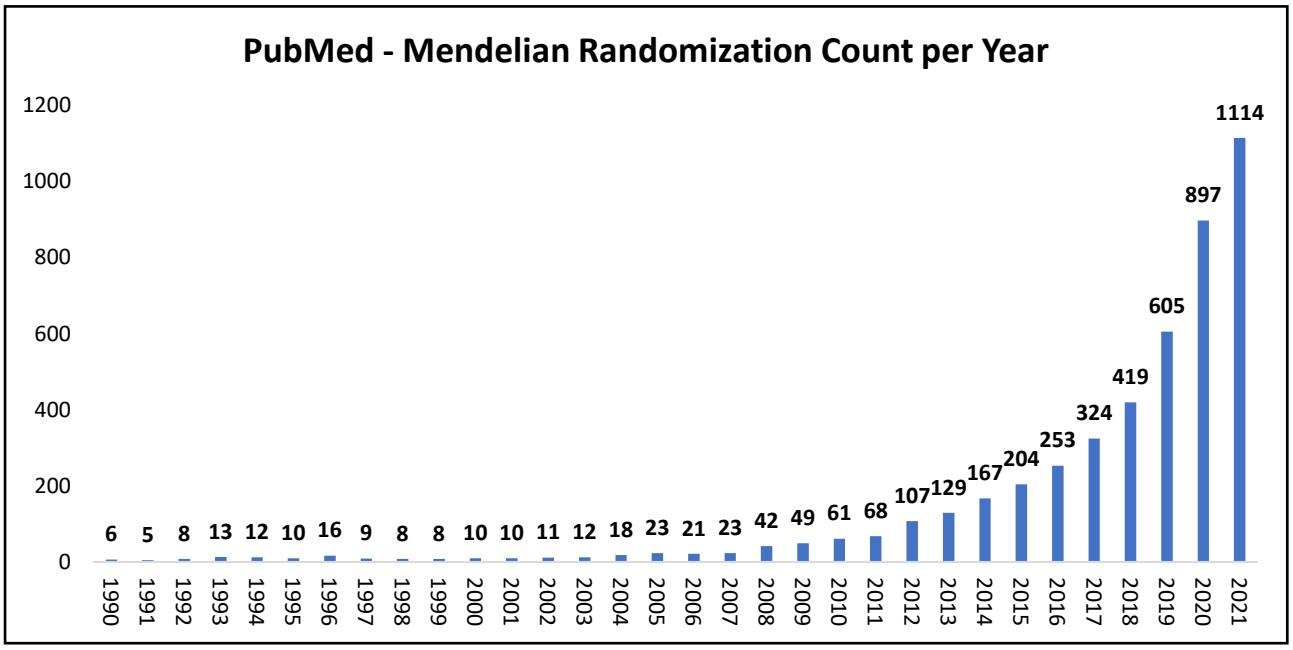


Figure 2.2: Evolution of Mendelian randomization

The use of MR has proven a success in many polygenic, multifactorial diseases such as CHD, DM, cancers, and others [142]. These diseases are to some extent genetic but also depend on modifiable risk factors such as diet, smoking, blood pressure, and others. Thousands to millions of associations between these genetic variants and

outcomes can be investigated with the increased use of genome-wide association studies (GWAS) [143]. These discoveries have added to the scientific community in which more understanding of the disease processes as well as predicting the disease risk for individuals has been achieved. However, MR has provided a breakthrough in which the estimation of causal effects of non-genetic (modifiable) risk factors can be assessed based on observational data [16]. Some examples of the use of MR in the literature are provided in Table 2.9 [16].

Table 2.9. Examples of causal relationships assessed by MR

Nature of exposure	Exposure	Outcome
Biomarkers	CRP	Insulin resistance [144].
	HDL	Myocardial Infarction (MI) [145].
Physical features	Fat mass	Academic achievement [146].
Nutritional factors	Alcohol intake	Blood pressure [147].
Pathological behaviour	Smoking	Schizophrenia [148]

A well-known example is the use of MR to test the causal relationship between HDL and myocardial infarction (MI) by Voight et al [145]. A high level of HDL was observationally associated with a lower risk of MI [149,150]. By using the MR approach, Voight et al examined if the relationship between HDL and MI is causal and then compared the size of the estimate (in terms of OR) to observational studies. The investigators conducted two analyses, they tested the relationship between HDL and MI using a single SNP and by using the genetic score comprised of 14 common SNPs. The single instrument analysis uses rs61755018 SNP in the endothelial lipase gene

(LIPG Asn396Ser) testing this SNP in 20 studies with 20913 MI cases and 95407 controls. The genetic score analysis uses 14 common SNPs exclusively associated with HDL and they tested this score vs. 12482 MI cases and 41331 controls [151]. The previous findings from the observational studies of the relationship between HDL and MI were a 13% decrease in MI risk (OR: 0.87, 95% CI: 0·84–0·91). Additionally, a one SD increase in HDL was associated with a decreased risk of MI (OR 0·62, 95% CI 0·58–0·66) [149,150]. However, the MR finding showed that the individuals with LIPG 396Ser allele have a nonsignificant association with the risk of MI (OR: 0·99, 95% CI 0·88–1·11, P=0·85). Additionally, an increase of one SD in HDL instrumented by the genetic score was not associated with the risk of MI (OR 0·93, 95% CI 0·68–1·26, P=0·63). The authors concluded that based on some genetic MR estimates, increased plasma HDL has no causal beneficial effect on reducing the risk of MI [151]. This finding almost contradicts the well-known scientific view on HDL's beneficial influence on MI. The findings of this study have matched the RCT findings on hormone replacement therapy increasing the plasma HDL but not lowering MI risk [152]. The genetic variants used in this study were biologically known functions and exclusively associated with increased plasma HDL, which makes it a reliable and valid IV to be used in MR analysis. They also tested for all potential pleiotropic effects on the cardiovascular risk factors with non-significant results have been detected.

One-sample and two-sample Mendelian randomization

Mendelian randomization can be performed using data from a single population or two populations (cohorts). Obtaining the genetic data for the exposure (SNP => exposure) and outcome (SNP => outcome) from the same population is called one-sample MR [132]. Whereas two-sample MR requires that the genetic-exposure associations are estimated from one sample and genetic-outcome data are from a second dataset.

Because there are some differences in common genotype frequencies across ethnicities, both populations should be derived from the same ancestry [153]. This thesis will mostly be using one-sample MR (and two-sample MR for comparison) utilising the UKB and publicly available data. Table 2.10 summarises each approach [124,154]. Like many other types of epidemiological studies, Mendelian randomization has its limitations especially when it comes to the fulfilment of IV assumptions and adequate power. The next section will explore these limitations.

Table 2.10. Comparison between one-sample and two-sample MR

	One-sample MR	Two-sample MR
Advantages	- Investigations are conducted on the same individuals (MR and conventional results). - Can check confounders. - Do not require harmonization of the genetic variants (one population). - Wide range of analyses with the availability of individual data (subgroup analyses).	- High power (large sample size) - Transparency (can be reproduced as the data is available publicly) - Pragmatic (easy to perform practically)
Disadvantages	- Weak instrument bias: a weak instrument explains a small amount of the variation in a specific trait [155]. As one-sample MR only uses one IV at a time, weak instruments will lead to this bias and the results will be confounded toward the observational association [24].	- The population where the samples are extracted may differ (the associations between the variants and the exposure might differ between the samples) => which affects the validity of the IV. - Limited range of analysis as dealing with the available summarised data.
Methods to estimate the causal effect	- Two-stage method (two-stage least squares/estimator)	- MR-Egger - Inverse-variance weighted (IVW)

Weak Instruments	- Ratio Method (Wald) Weak genetic variant-exposure associations will lead to bias of the exposure-outcome association in the direction of the observational association (confounded associations) => false-positive findings	Bias toward the null (no association)

Limitations of the Mendelian randomization approach

The limitations of MR revolve principally around the violation of instrumental variable assumptions. Violating any assumptions or risking the validity of a GV used as an IV, will jeopardise the integrity of the causal estimate of the relationship between the exposure and the outcome [156]. An example of such a violation is horizontal pleiotropy. Horizontal pleiotropy arises when the GV is directly associated with the outcome and not only via exposure. Furthermore, sometimes the appropriate GV to study the trait of interest might not be available, hence less reliable estimates can be inferred [125]. Finally, the lack of biological plausibility of the GV might create spurious unreliable associations [16]. The following section presented the use of genetic variants as a proxy for smoking to infer casualty using the MR approach.

Smoking and genetic variants (GVs)

With the emergence of the GWAS and Mendelian randomization, genetic variants associated with smoking behaviour have been studied and tested for understanding smoking behaviour and to assess the causal estimates between these genetic variants associated with smoking and potential outcomes [157]. The GWAS have provided a

considerable number of genetic variants associated with the trait of interest, including smoking which will be the main variable of this thesis.

GWAS is a hypothesis-free observational study of genetic variants among different individuals to identify associations between genetic regions (loci) and phenotypic traits (including diseases). GWAS measure and examine DNA sequence differences across the human genome to recognize genetic risk factors for common diseases in the population [158]. The main aim of GWAS is to identify relationships between SNPs and phenotypic traits to make predictions of disease susceptibility among individuals to formulate strategies for disease prevention and treatment [159]. GWAS also provide the GVs to be used in MR analysis to make causal inferences. Out of 3 billion human nucleotides, few GWAS-significant SNPs have been found and explored to find associations with phenotypic traits [160]. The evolution of GWAS is portrayed in Figure 2.3.

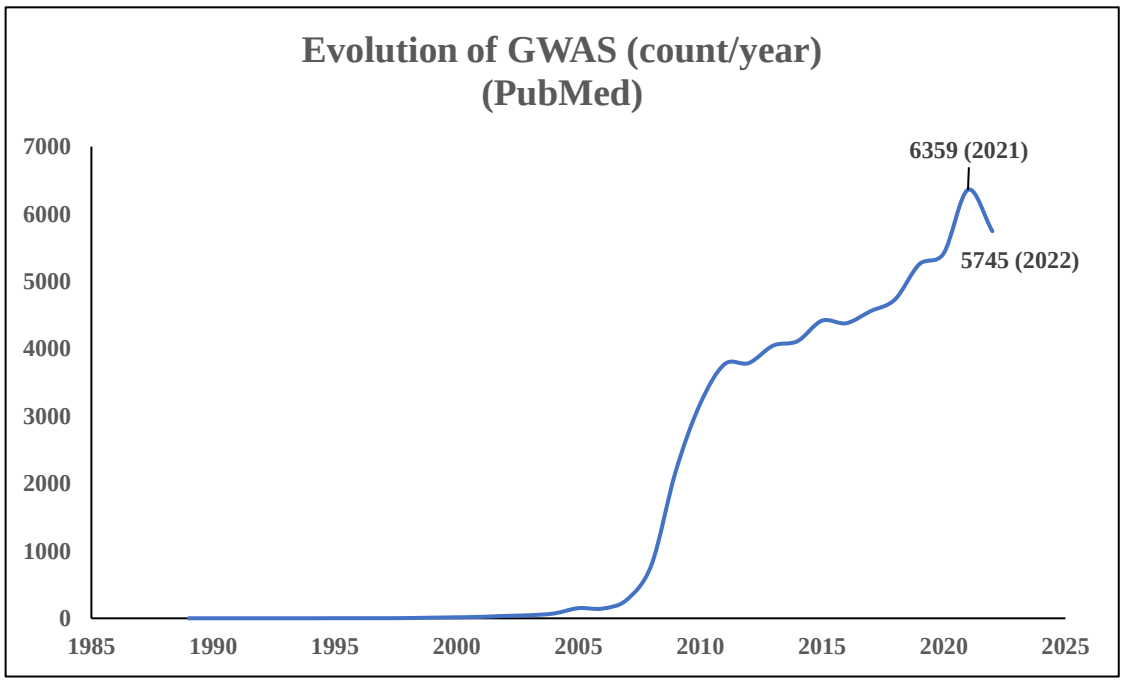


Figure 2.3. GWAS Development

GWAS have provided access to find significant loci (locations) on human genes and chromosomes. In GWAS, the threshold to encounter a SNP to be statistically significantly associated with a trait is $p<5e-8$. When testing millions of SNPs, P-value is set to be low to differentiate between true positives and false positives [161]. The genetic characteristics included in this review and all over this thesis will be of European ancestry.

One of the major risk factors of interest (phenotypic trait) that have been studied is smoking behaviour. Many SNPs are significantly associated with smoking behaviour [127,162]. These SNPs are referenced (rs) and reported along with the concerned gene and the chromosomal location. Table 2.11 summarises the smoking behaviours and some GWAS significant SNPs with their characteristics among European ancestry [127,162–164].

Table 2.11. Smoking behaviour and genetic characteristics

Genetic characteristics					
		rs (Reference SNP)	Gene(s)	Chromosome	P value
Smoking Behaviour*	CperD	rs1051730-A**	*CHRNA3 CHRNA5 and CHRNB4 (Neuronal nicotinic acetylcholine receptor)*	15q25	2.8×10^{-73} 2.4×10^{-69} 1.71×10^{-66} 9.4×10^{-19}
		rs16969968-G	*CHRNA5*	15	1.2×10^{-278} 5.57×10^{-72}
		rs55853698	*CHRNA5*	15q25	1.31×10^{-16}
		rs6495308-T	*CHRNA3*	15q25	3.3×10^{-10}
		rs1329650-G	*HECTD2-AS1*	10q25	5.7×10^{-10}
		rs1028936-A	*HECTD2-AS1*	10q25	1.3×10^{-9}
		rs3733829-G	*EGLN2, RAB4B-EGLN2*	19q13	1.0×10^{-8}
		rs4105144-C	*CYP2A (Nicotine-metabolizing enzymes)*	19q13	2.2×10^{-12}
		rs6474412-T	*CHRNA6 and CHRNB3*	8p11	1.4×10^{-8}
		rs8042374-A	*CHRNA5-A3-B4*	15	2.4×10^{-24}
	Initiation	rs6265-C	*BDNF, BDNF-AS*	11	1.8×10^{-8}
		rs462779-A	*REV3L*	6	4.52×10^{-8}
		rs12616219-C	*TMEM182*	2	2.25×10^{-16}
	Cessation	rs3025343-G	*Near DBH*	9	6.8×10^{-27}
		rs202664-T	*TOB2*	22	5.98×10^{-8}

* Across different studies
** A, G, C, T: DNA bases (nucleotides)

These statistically significant SNPs were explored widely in the literature in the last decade. Researchers have found significant relationships between SNPs and traits such as smoking [165]. The SNPs that were found to be associated with smoking explain approximately 2% of the genetic heritability of smoking behaviour [166]. The 15q25 (CHRNA3/5-CHRNB4) region holds the largest effect explaining 1% - 5% of the variation in smoking behaviour. These discoveries have given access to understanding smoking genetically and not only environmentally. These SNPs were extracted from the European ancestry population among different studies in Europe. In

this thesis, European ancestry will be the cornerstone for the analysis especially the UK population as the data will be obtained from the UKB.

Using SNPs as a proxy for smoking behaviour (e.g., smoking intensity; abbreviated to CperD)

A robust association between rs16969968 and smoking intensity has been established with biological plausibility. This SNP is in the *CHRNA5- CHRNA3-CHRNB4 nicotinic receptor subunit gene* cluster on the long arm of chromosome 15 (15q25) [164,167]. The rs16969968 is a functional variant that leads to an amino acid change (D398N) in the alpha subunit protein of the nicotine receptor [168]. The minor allele of the rs16969968 SNP was found to be associated with the increased smoking amount by one cigarette per day among smokers as well as the serum cotinine level (nicotine metabolite) [169]. The rs16969968 is in perfect linkage disequilibrium with rs1051730 ($R^2 = 1$) in the European ancestry, thus they represent the same genetic signal and can be used interchangeably (referred to as rs16969968- rs1051730) [170]. The carriers of minor allele [rs1051730 T, rs16969968 A] have been shown to have an increased risk of heavier smoking and other health-related conditions compared to wild-type [171]. As with rs16969968 SNP, each additional T allele of the rs1051730 SNP in the CHRNA3 is associated with increased CperD among smokers [169,171]. The use of rs16969968- rs1051730 as a proxy for smoking intensity is commonly used in the literature [37,170,172,173]. A review of the relationship between these SNPs and the outcomes of interest (CHD, stroke, HTN, DM, and lipid markers) will be discussed in the following sections.

Smoking behaviour and health outcomes using Mendelian randomization (MR)

Overview

The use of MR for smoking behaviour has been reported in the literature. It is a less confounded measure of smoking exposure because these genetic variants, that act as a proxy (IV) for smoking, have been determined during gamete formation and conception (i.e., before environmental exposure). Therefore, these alleles associated with smoking are paired randomly from parents to offspring and are not likely attributable to environmental factors which prone the conventional studies of smoking to confounders [174,175]. If smoking is associated with any outcome then genetic variants predicting smoking behaviour should be associated also with these outcomes attributable to smoking among current smokers but not never smokers (as their GVs are not associated with smoking intensity) [175,176].

Smoking behaviour vs CMDs and lipids (MR review)

The use of genetic variants as a proxy for smoking is widely used in the literature. Many studies have explored smoking behaviour using MR. Tables $2.12_{(a\text{-}f)}$ summarise some papers that examined smoking behaviour and thesis outcomes using the MR approach.

Table 2.12 (a-f). Review of smoking and health outcomes: MR approach

a) Study design and purpose	Mendelian randomization To examine the association of the rs1051730 T alleles with cardiovascular risk factors.
Author(s)	Åsvold BO et al [39].
Exposure	Smoking status (rs1051730-T).
Outcome	Cardiovascular risk factors (BMI, blood pressure, HDL, and glucose).
Sample size	56,625 participants.
Findings	• An additional rs1051730 allele was 0.27 mmHg (95% CI: 0.04, 0.49) lower systolic BP among the total study population (P-value <0.02) but no association was found with diastolic BP. • A 0.34% (95% CI: 0.02, 0.66) higher concentration of HDL was observed across the total study population with each additional rs1051730 T allele but not among smoking subcategories, for example, current smokers (0.37%, P=0.2). • Among current smokers, the rs1051730 T allele was associated with 1.16% (95% CI: 0.03, 2.28) lower triglyceride concentration which was attenuated after adjustment for BMI (0.03%, P=0.96). • There was no convincing association seen between rs1051730 and glucose level nor total cholesterol among current smokers.
Limitations	• This study included only a single SNP and single SNPs are weak instruments usually, a polygenic score would be more robust (The genetic score increases statistical power as well as the precision of the IV estimate). • They stated that the rs1051730 might have influenced the outcome via other routes and not only through smoking and that will violate one of the IV assumptions (positive association between rs1051730 and BMI).
Conclusion	They concluded that smoking was not a major determinant of blood pressure, serum lipid or glucose level.

b) Study design and purpose	Mendelian randomization To examine the association of the rs1051730 T alleles with cardiovascular risk factors.
Author(s)	Linneberg et al. [177].
Exposure	Smoking status (rs16969968 or rs1051730).
Outcome	Hypertension.
Sample size	141,317 participants (37,982 current smokers).
Findings	• The researchers found that the beta estimate per minor allele of rs1051730/rs16969968 with systolic blood pressure (SBP) and diastolic blood pressure (DBP) was close to null with overlapping CI (-0.20, 95% CI: -0.46, 0.06 for SBP, P=0.136, and -0.15 95% CI: -0.32, 0.02, P=0.079 for DBP) among current smokers.
Limitations	• There was no data provided on the validity of the SNPs used in the analysis. • A polygenic score would be a better instrument compared to a single SNP.
Conclusion	They concluded that there was no causal association between smoking and blood pressure.

c) Study design and purpose	Mendelian randomization (2-sample MR) To assess the causal association between smoking and stroke.
Author(s)	Larsson, Burgess, and Michaëlsson [178].
Exposure	Smoking status (372 SNPs associated with smoking initiation).
Outcome	Ischemic stroke.
Sample size	34,217 patients (404,630 control).
Findings	• SNPs explain 2.3% of the variation in smoking initiation. • A statistically significant and positive association between genetic predisposition of smoking initiation and ischemic stroke. • A unit increase in log odds of smoking initiation was associated with a 22 % increase in ischemic stroke risk (OR= 1.22, 95% CI: 1.12 – 1.34, p-value = 7.6×10^{-6}). There was no indication of horizontal pleiotropy (all P>0.24).
Limitations	• The two-sample MR provides a high-power analysis but no access to individual data. • The instrumental variables from the sample might not represent the population where the sample was obtained which might doubt the validity of the used instruments. • Further analysis will not be possible from such an approach.
Conclusion	The study concluded that there was a causal association between smoking initiation and increased risk of stroke.

d) Study design and purpose	Mendelian randomization (Summary-level) To assess the causal association between smoking and CVDs.
Author(s)	Susanna C Larsson et al. [179].
Exposure	Smoking initiation (361 SNPs associated with smoking initiation).
Outcome	CVDs
Sample size	367k European-descent individuals (UKB).
Findings	• Smoking initiation is genetically associated with 10 out of 14 CVDs (outcomes of interest will be shown below): - Heart failure: OR=1.53 (95% CI: 1.37-1.71). - Coronary heart disease: OR = 1.36 (95% CI: 1.27-1.45). - Stroke: OR = 1.30 (95% CI: 1.15-1.48).
Limitations	• The possibility of pleiotropy is not entirely ruled out. • The causal estimate might be biased as the number of UKB participants was included in both exposure and outcome datasets. • The summary-level MR provides no access to individual data if any further analysis is needed.
Conclusion	The study supported a genetic-based causal association between cigarette smoking and CVDs.

e) Study design and purpose	Mendelian randomization (Summary-level) To assess the causal association between smoking and the risk of atherosclerotic cardiovascular diseases.
Author(s)	Levin MG et al. [180].
Exposure	Lifetime smoking index and smoking initiation (126 SNVs).
Outcome	Atherosclerotic CVDs and associated risk factors.
Sample size	> 1 million individuals.
Findings	• Genetic liability for smoking was associated with increased risk for the following: CHD (OR: 1.48; 95% CI: 1.25-1.75, P<0.001) Stroke (OR: 1.40; 95% CI: 1.02-1.92, P= .04) Type 2 diabetes (OR: 1.89; 95% CI: 1.53-2.33, P<0.001) HTN (OR: 1.05; 95% CI: 1.04-1.07, p<0.001) Hyperlipidaemia (OR: 1.00; 95% CI: 1.00-1.01, P<0.001) • The study also found non-significant associations between smoking and the following: Cholesterol (β: 0.00574, 95% CI: –0.0988 to 0.1, P=0.91) LDL (β: –0.0079, 95% CI: –0.0998 to 0.08, P=0.86) TG (β: 0.069, 95% CI: –0.0373 to 0.175, P=0.20) HDL (β: –0.088, 95% CI: (–0.201 to 0.0243, P=0.12)
Limitations	• The findings seem to differ between the binary outcomes and the continuous version for the same variables (e.g., hyperlipidaemia as a binary variable was significantly associated with smoking but when the variables were examined separately as numeric variables the associations seem to be non-significant). These findings were also noted for diabetes (binary) vs fasting blood glucose and HbA1C (numeric) as well as hypertension (binary) vs systolic blood pressure (numeric).
Conclusion	The study concluded that genetic liability to smoking was associated with an increased risk of atherosclerotic CVDs and stroke.

f) Study design and purpose	Mendelian randomization (2-sample MR) To examine the causal association between smoking and DM.
Author(s)	Larsson and Yuan [181].
Exposure	Smoking status (378 SNPs).
Outcome	Type 2 Diabetes.
Sample size	Smoking initiation: a meta-analysis of GWASs which included 1,232,091 individuals. DM: 898,130 individuals (47,124 cases and 824,006 controls) from GWAS of 32 studies (DIAbetes Genetics Replication and Meta-analysis consortium).
Findings	• The study found a positive significant association between genetic-based smoking initiation and type 2 diabetes: OR= 1.28 (95% CI: 1.20-1.37, P=2.35 x 10^{-12}).
Limitations	• The two-sample MR provides a high-power analysis but no access to individual data. • The researchers reported a large overlap between the participants in the datasets of DM and smoking initiation, this might lead to bias in the estimation toward the observational association. • The instrumental variables from the sample might not represent the population where the sample was obtained which might doubt the validity of the used instruments. • Further analysis will not be possible from such an approach.
Conclusion	The study concluded that smoking initiation was causally associated with an increased risk of type 2 diabetes.

Summary

Smoking behaviour has been tested for causality in the literature using MR approaches. Some findings matched the observational studies, and others did not. One-sample MR studies in this review used a single SNP to proxy smoking status which explains less variation in smoking compared to polygenic score. Most studies examined smoking status against one or a few outcomes and either using one-sample or two-sample MR. However, the current thesis examined smoking with CMDs and stroke as well as lipid biomarkers in one-sample and two-sample MR approaches using the polygenic score in addition to a few selected single SNPs (based on the magnitude of their association

with smoking: beta coefficients and P values). The current thesis differs from previous studies by examining multiple smoking phenotypes using both observational and genetic approaches. Previous studies have often focused on examining the associations between smoking and health outcomes using either smoking history, intensity, initiation, or a combination of these variables, but rarely have all of these variables been included in the same analysis. In contrast, this thesis will consider a range of covariates and outcomes in order to more fully understand the relationships between smoking and poorer health. Additionally, this study will utilise both observational and MR approaches to establish both observational and causal associations. Moreover, this thesis will construct two genetic scores to proxy smoking behaviour, a departure from other studies which have typically used either one SNP in a one-sample MR design or two-sample MR. The combination of a wide range of outcomes, multiple smoking variables, a large sample size from the UK Biobank, and a variety of approaches make this study a novel contribution to the field and a valuable addition to the scientific literature.

2.5. Conclusion

The harmful effect of smoking on health is generally reported and the evidence showed that smoking is a risk factor for many diseases and health outcomes. Physiologically, smoking affects the body through the toxins such as nicotine, CO, and NO which cause endothelial damage and major physiological changes. These changes are believed to be the main mechanisms by which smoking causes harmful health effects. These effects include CHD, stroke, HTN, DM, and lipid biomarkers changes. The effect of smoking was largely and most commonly examined using observational approaches which are prone to bias, especially in the presence of confounders (measured or not). Inferring the causality of smoking on these outcomes is questionable when using conventional

approaches. To overcome these problems, the MR approach was introduced. In this chapter, a detailed review of observational relationships between smoking behaviour and health outcomes was provided, in addition to a detailed description of the MR approach. This is followed by a review of the use of genetic variants as a proxy for smoking behaviour in the MR approach. The next chapter discusses the methods that will be used in this thesis in addressing the key questions and aims described in the introduction. [A more detailed literature review is presented in the supplementary materials: 8.2, detailed literature review].

3. Chapter Three: Methods

3.1. Overview

This chapter describes the methods that will be applied in this thesis, which will comprise broadly a detailed review of the UKB, observational section, and MR section. Both sections include sample size calculations, the definition of the research variables, data preparation, analysis, and presentation. Additionally, characteristics of the study population and ethical considerations have been included.

3.2. The UK Biobank (UKB)

The UKB is a very large population-based prospective cohort study with thorough genetic and phenotypic data gathered on around 502k individuals across the United Kingdom [38]. The main goal of the UKB is to improve the prevention, diagnosis, and treatment of a wide range of health-related conditions by gathering an extensive and accurate assessment of exposures with wide-ranging follow-up and depiction of health-related outcomes. Additionally, it helps scientists and researchers to innovate and make contributions to the scientific community by exploiting the huge amount of freely accessible data [182]. The individuals in the UKB aged from 40 to 69 at the time of recruitment in 2006 with completed baseline data and samples in 2010. The assessment of the members was conducted in 22 assessment centres across the UK. It included a wide range of socioeconomic, ethnic, rural and urban backgrounds [38]. The assessment visit encompassed signed consent, a self-completed questionnaire, a computer-based interview, lifestyle, and health-related measures as well as a collection of biological samples (blood, urine, and saliva). These samples were stored to allow for numerous types of assay to be achieved, for example, genetic and biochemical markers [182]. All participants provided an agreement to follow up through their health-related records. By May 2018, there were more than 14,000 deaths, 79,000 participants diagnosed with cancer, and 400,000 individuals had been admitted to the hospital at

least once since recruitment [182]. The UKB has a substantial amount of data on the biochemical markers which were used to make relationships with the diseases, such as lipids for CVDs or diagnostic values such as HbA1c for DM or assessment of the biological functions such as the liver and renal function [183]. In addition to lifestyle, sociodemographic, physical, and biochemical measures, the UKB has extensive genetic data for this large number of participants.

The UKB genetic data encompasses genotypes for 488,377 individuals assayed using two similar genotyping arrays [184]. Genotyping is a technique for detecting small genetic variations that can result in significant phenotypic changes (trait). Around 49,950 participants were genotyped at 807,411 markers using Applied Biosystems UK Believe Axiom Array by Affymetrix whereas 438,427 individuals were genotyped at 820,967 markers using Applied Biosystems UKB Axiom Array [184,185]. The two arrays shared approximately 95% of the marker content. The marker content of the UKB Axiom Array was set to capture SNPs and indels (short insertions and deletions). Numerous markers were included based on acknowledged associations with, or possible roles in, phenotypic differences (95,490 markers). Moreover, the array comprised coding variants subject to estimated minor allele frequency (EMAF) ranges (111,904 markers). Finally, markers with good genome-wide coverage in European populations (total = 629,368 markers; common variants = 348,569, low-frequency variants = 280,838) were also included [184,186].

The UKB consists of approximately half a million participants with different self-reported ethnicities, however, most participants fall under the white ethnic group (94.06%) [187]. This large number of participants along with extensive sociodemographic, phenotypic, and genetic data of the European (the UK) population is the main reason to choose the UKB for this thesis. One more reason to choose the

UKB is the existence of detailed data on smoking (exposure of interest in this thesis) and the outcomes (CVDs, HTN, DM, and biomarkers). In a realistic Mendelian randomization context (moderate causal OR and non-strong correlation of GVs with exposure), thousands of cases are required to attain adequate power [16]. Therefore, with the UKB's large sample size, the power needed to detect statistically significant results when the effect is real in the population will be increased, as well as the results will be representative of the whole UK population and the European ancestry.

3.3. Observational approach

Study design

A cross-sectional analysis of the UKB cohort to examine the observational relationship between smoking and health outcomes as a reference in which MR analysis will be compared. These associations were based on a self-report questionnaire, physical measures, and haematological essays.

Sample size and power analysis

The UKB is a very large cohort with more than half a million participants. This large sample size provides greater power to detect a significant difference between study groups, offers a more accurate estimation of the population parameters, lowers the margin of error, as well as it provides a generalisable exposure-disease relationship [188]. The sample size of the current study is based on a smoking status variable in all UKB participants (n=469,598). To obtain the required sample size, G*power 3.1.9.7 software was used.

Research variables

The UKB evaluates several baseline participant characteristics. However, only those relevant to the current thesis were included and described below. Table 3.2 summarises basic information on these variables in the UKB.

Table 3.1. Basic characteristics of study variables in the UKB

Characteristics / Variable*	Data-Field	Type of variable	Sample Size	UKB Link
Smoking Status	20116	Nominal	469,598	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=20116
Smoking intensity (CperD)	3456	Numeric	35,758	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=3456
Smoking initiation (Age started smoking)	3436	Numeric	39,497	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=3436
CHD				
Stroke	6150	Binary	501,601	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=6150
HTN				
DM	2443	Binary	501,601	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=2443
Cholesterol	30690	Numeric	470,862	http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=30690
LDL	30780	Numeric	470,024	http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=30780
HDL	30760	Numeric	432,148	http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=30760
TG	30870	Numeric	470,492	http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=30870
Age	21022	Integer	502,524	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=21022
Sex	31	Binary	502,524	http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=31
Degree	6138	Binary	497,883	http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=6138
Ethnicity	21000	Binary	501,632	http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=21000
Deprivation	189	Numeric	501,901	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=189
BMI	21001	Numeric	499,518	https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=21001

***Variables**:

Green colour: independent variables

Yellow colour: dependent variables

Blue colour: covariates

- **Independent variable**

The independent variable for this thesis is smoking behaviour. The touchscreen questionnaires were used to collect smoking habits data in the UKB. The survey involved several smoking-related information including smoking status, age started smoking, smoking intensity (cigarette smoked daily; CperD), maternal smoking,

smoking duration and others [189]. The present thesis focused on smoking status, smoking intensity (CperD), and smoking initiation (SI) (n=469,598).

Smoking status is a nominal variable where participants were categorised into three groups using a three-point scale "0 = never" the person had never smoked), "1 = previous" (the person used to smoke but stopped/ ex-smoker), and "2 = current" (the person is a smoker).

Smoking intensity (CperD, n=35,758) is a numeric variable based on the question "how many cigarettes do you smoke on average each day?". The units of measurement are cigarette/day. The mean of CperD among current smokers was 14.46 cigarettes/day (± 8.44).

Smoking initiation (age started smoking: SI, n= 39,497) is a numeric variable based on the question "How old were you when you first started smoking on most days?". The units of measurement are years. The average age at which the current smokers in UKB started to smoke is 17.88 years (± 5.86).

- **Dependent variables**

The dependent variables in the study were categorised into two main categories: Cardiometabolic diseases and lipid biomarkers. The CMDs are binary variables while lipid biomarkers are numeric.

Cardiometabolic diseases (CMDs): The following diseases were included in the analysis because these are common and generally prevalent physical health conditions; CHD, stroke, HTN and DM. The participants have asked if a doctor ever told them they have one of the following [CHD, stroke, HTN, or DM]. Each one of the CMDs was coded based on a 2- points scale as "0 = NO" (doctor has not diagnosed me with disease) and "1 = yes" (doctor has diagnosed me with disease). Data on CMDs were obtained at

the assessment centres through oral interviews, touchscreen questionnaires, biological sampling, and physical measures [183].

Lipid biomarkers: The UKB has a wide range of biochemical markers in biological samples among 480,000 participants at baseline with additional 18,000 samples collected at a repeat assessment. Lipid biomarkers were measured in serum from the serum separation tube sample. Multiple immunoassays and clinical chemistry analyses were used to measure biochemistry markers [190]. The lipid measurements used in this thesis were total cholesterol, LDL, HDL, and TG. The unit of measurement is mmol/L.

- **Covariates:**

In addition to smoking behaviour, other variables that might have significant impacts on the outcomes were also analysed in this thesis. They included age, sex, educational attainment, deprivation, ethnicity, and BMI. The minimum age of the participants was 37 years, and the maximum age is 73 years (mean=56.53 ± 8.09).

Sex, education level, and ethnicity: each variable was recorded on two-point scales as follows:

Sex: 0 = Female; 1 = Male

Education level (qualifications): 0 = No College/University degree; 1= Have College/University degree.

Ethnicity: 0 = White British; 1= Other Ethnicities.

Deprivation score: The deprivation level was assessed via the Townsend deprivation index. The Townsend index is a census-based metric that measures material deprivation (unemployment, non-car ownership, non-home ownership, and overcrowded households) [191]. The higher (positive) the score, the higher the level of deprivation.

The higher deprivation score was found to be associated with adverse health effects [192].

Body mass index (BMI) represents the ratio of an individual's weight in kilograms to their height in meters. WHO has classified the BMI into; underweight, normal weight, overweight, and obese [193]. The BMI is a numeric variable obtained from the UKB participants manually at the data centres.

Data preparation, analysis, and presentation

This section summarised different tests, techniques, and tools utilised to assess, interpret, and present the study findings. The data used for this thesis were quantitative. All the analyses were performed using R (R-3.5.3). Data presentation was in the form of tables and graphs. R software was exploited to generate the tables, charts, summary statistics, and regression analyses.

To ensure unbiased and valid results from linear regression analyses, the data should meet the assumptions of parametric tests. These assumptions include normality of distribution, linearity between the variables under analysis, and homogeneity of variances [194]. Applying the log transformation (*log()* in R) ensures that the data is fulfilling the parametric assumptions. In the present thesis, the triglycerides variable was positively skewed so, it was subjected to log transformation before analysis to fulfil the normality assumption.

Descriptive statistics were used to summarise the variable cases while regression was used to establish the associations between variables. Smoking and CMD outcomes (CHD, stroke, HTN, and DM) were examined using logistic regression as these variables are binary. On the other hand, smoking and numeric lipid biomarkers (cholesterol, LDL, HDL, and TG) were examined using linear regression.

This section of the thesis intended to assess the observational association between smoking and health outcomes. Such associations are at risk of bias in the presence of confounding variables. Therefore it was crucial to adjust for the effects of these covariates on the dependent and independent variables [195]. Appropriate regression models were built to adjust for the unwanted effect of these covariates. The threshold of statistical significance was set at P = 0.05 (95% confidence level).

3.4. MR approach

Study design

A Mendelian randomization approach will be applied to make causal inferences/estimates of the relationships obtained from this observational data. To undertake an MR, genetic data (SNPs) and observational data should be available. There are two types of MR; individual-level (one-sample) and summary-level (two-sample MR) [196]. In one-sample MR, the UKB will be used to obtain smoking genetic data and observational outcome data. The SNPs that were extracted from the UKB were used as an instrumental variable for smoking status and smoking intensity (CperD) variables. After examining the assumptions of the validity of these instruments, a genetic score will be built to conduct MR analysis to test the causal relationship between smoking and health outcomes. In two-sample MR, the analysis will be performed automatically using the MR-Base platform and using R. The SNPs instrumenting smoking variables will be obtained from European populations other than the UKB cohort and the health outcomes data will be obtained from the UKB. The SNPs from the exposure data will be matched with ones in the UKB to establish the association. Finally, MR analysis will be conducted to examine the causal association between smoking behaviour and health outcomes.

Sample Size and Power

Mendelian randomization requires a large sample size because the genetic variants (SNPs) usually explain a small proportion of the variability in the risk factor. Smoking behaviour is no exception: the variability in smoking behaviour that can be explained by SNPs is small (~ 2% - 4%) [166]. To calculate the sample size for MR, the sample size for observational approaches should be divided by how much variation in the risk factor can be explained by the SNPs (R^2) [197]. In this thesis, the sample size for observational regression of the outcome required to detect a given effect size is 74 and the IV roughly explains 2% of the variation in smoking behaviour. Based on this approach, the sample size for MR analysis with significance at $P<0.05$ and 80% power is approximately 74/0.02 = 3700. Fortunately, the UKB data is far beyond this number for smoking behaviours. To ensure a valid MR approach, only Caucasian individuals were included in the analysis (European ancestry: n=25,724 for CperD, n=314K for smoking status variable). The previous estimate was based only on one SNP (IV) analysis, however, a genetic score will be used in this thesis to ensure even more statistical power as well as a better IV estimate. The details of the methods and the analysis were discussed in detail for CperD and briefly for the smoking status variable. However, the details for the smoking status variable were provided in the supplementary materials (8.4, results: smoking status MR).

Research Variables

To conduct MR, genetic variants such as SNPs are usually used to proxy the risk factor of interest, smoking status and CperD in the current thesis. In addition to the variables discussed in the observational section, SNPs rather than self-reported smoking responses were used as a proxy for smoking behaviour. SNPs associated with smoking status as well as with smoking intensity (CperD) were utilised to examine the causal association with health outcomes proposed in this thesis. These SNPs are used widely

in the literature proxying smoking behaviour [198]. Table 3.3 illustrates the SNPs characteristics in the literature including the P values, beta coefficients, the number of studies in which these effect sizes were extracted and the sample size of the original studies.

These SNPs were extracted from a recently published meta-analysis of GWAS for smoking behaviour among European ancestry [198]. Critically: these data did not include the UKB participants. The beta coefficients in Table 3.3 will be the basis for building the genetic scores for smoking variables. These scores were used to conduct one-sample MR among the UKB participants. These SNPs will be used throughout this thesis. Included SNPs based on their validity fulfilling the instrumental variable assumptions, being used in the literature, as well as availability in the UKB. Testing the validity of the IV assumptions and associated technicalities was discussed in the following section.

Table 3.2. Characteristics of SNPs proxying Smoking behaviour*

SNP-effect allele	Chromosome/Gene	Beta	P value	No. of studies	Sample size
1- rs8034191-C	*15 Intron:AGPHD1*	0.183	$4.80E^{-211}$	33	257,341
2- rs1051730-A	*15 Synonymous:CHRNA3*	0.324	$2.33E^{-202}$	33	257,341
3- rs16969968-A	*15 Nonsynonymous:CHRNA5*	0.179	$2.32E^{-200}$	33	257,341
4- rs12914385-T	*15 Intron:CHRNA3*	0.170	$2.19E^{-188}$	33	257,341
5- rs8040868-C	*15 Synonymous:CHRNA3*	0.165	$1.09E^{-184}$	33	257,341
6- rs11637630-A	*15 Intron:CHRNA3*	0.158	$1.19E^{-123}$	33	257,341
7- rs938682-A	*15 Intron:CHRNA3*	0.158	$2.17E^{-123}$	33	257,341
8- rs6474412-T	*8 Intergenic*	0.067	$3.59E^{-24}$	34	263,954
9- rs2229961-A	*15 Nonsynonymous:CHRNA5*	0.207	$1.17E^{-18}$	33	257,341
10- rs3025343-A	*9 Intergenic*	0.063	$2.62E^{-13}$	34	263,954
11- rs73229090-A	*8 Intergenic*	0.055	$2.44E^{-10}$	34	263,954
12- rs3733829-G	*19 Intron:EGLN2|RAB4B-EGLN2*	0.035	$1.77E^{-09}$	33	257,341
13- rs2273506-A	*20 Synonymous:CHRNA4*	0.061	$5.94E^{-08}$	34	263,954
14- rs215614-A	*7 Intergenic*	-0.044	$8.19E^{-15}$	34	263867
15- rs3865453-T	*19 Intergenic*	-0.102	$3.56E^{-23}$	34	244933
16- rs28399442-A	*19 Intron:CYP2A6*	-0.231	$5.1E^{-40}$	33	257,341
17- rs7260329-A	*19 Intron:CYP2B6*	-0.044	$1.18E^{-13}$	33	257,341
18- rs7599488-T	*2 Intron:BCL11A*	0.03	$1.89E^{-6}$	34	263,954
19- rs28399443-A	*19 Intron:CYP2A6*	-0.231	$2.11E^{-39}$	33	257,341
20- rs117824460-G	*19 Intergenic*	-0.230	$3.53E^{-39}$	33	257,341
21- rs4803378-A	*19 Intergenic*	-0.227	$1.99E^{-38}$	33	257,341
22- rs4243084-C	*15 Intron:CHRNA3*	0.18	3.06^{E-197}	33	257,341
23- rs1317286-G	*15 Intron:CHRNA3*	0.18	2.60^{E-203}	33	257,341
24- rs12910984-A	*15 Intron:CHRNA3*	0.16	4.34^{E-122}	33	257,341
25- rs951266-A	*15 Intron:CHRNA5*	0.18	1.79^{E-199}	33	257,341
26- rs7180002-T	*15 Intron:CHRNA5*	0.18	5.33^{E-198}	33	257,341
27- rs72740964-A	*15 Intron:CHRNA5*	0.18	9.74^{E-199}	33	257,341
28- rs17486278-C	*15 Intron:CHRNA5*	0.18	5.88^{E-201}	33	257,341
29- rs55781567-G	*15 Utr5:CHRNA5*	0.181	1.03^{E-206}	33	257,341
30- rs55853698-G	*15 Utr5:CHRNA5*	0.181	1.30^{E-206}	33	257,341

*Adopted from: https://conservancy.umn.edu/handle/11299/201564

Data Preparation

Plink (plink-1.07-dos) was used to prepare the genetic data to be later analysed in R. The quality control (QC) for the genetic data excludes heterozygosity outliers (high genetic variability), those with missingness > 10%, those whose self-reported sex did not match their genetically determined sex, those with purported sex chromosomes aneuploidy, and those whose genetic ethnic grouping is not Caucasian. Additionally, the preparation includes the following:

- GWAS significance of the association between SNPs and smoking behaviour (done using R too).
- Missingness of the SNPs for a specific individual.

- Relatedness between individuals (assuming all subjects are unrelated).
- Linkage disequilibrium (LD) between the SNPs (to examine if the SNPs are inherited together "linked" or not to ensure SNPs' independency and accurate effect on a specific trait).
- Hardy-Weinberg Equilibrium (HWE) (assuming no deviation from HWE with constant allele and genotype frequencies at P<0.000001).
- Principal Component Analysis (PCA) was also obtained to control for possible population stratification. PCA is a statistical tool used in population genetics to discover the pattern in the distribution of genetic variation across geographic locations and ethnic backgrounds [199]. Population stratification results from non-random mating between individuals, there is a systematic disparity in allele frequencies between subpopulations in a population [200].

After preparing the genetic data, R software was used to align the genetic data with observational data, generate the charts and graphs, test the IV assumptions, generate genetic (allele) scores, and perform one-sample and two-sample MR using MR packages.

Instrumental Variable

The instrumental variable (IV) is used to account for confounders between variables under the study [16]. To conduct Mendelian randomization, the instrument variable should be valid. The validity of the IV is based on three assumptions: significant GWAS level (5×10^{-8}) association of the IV with exposure, no association between the IV and the outcome except via the exposure (no pleiotropy), and finally no association between the IV and any confounders. In the current thesis, these assumptions were tested for all SNPs associated with smoking status as well as smoking intensity (CperD). The first

assumption was tested to establish GWAS's significant association between a specific SNP (IV) and smoking variables. The second assumption was tested by regressing the outcomes of interest on the IV. The third assumption was also examined using regression of the confounders on the IV. A valid instrument is significantly associated with smoking behaviour variables, not directly associated with the outcomes (except through the exposure) and has no significant association with the confounders. The SNPs that were included in generating the genetic score and subsequently used in MR analyses are summarised in Table 3.4.

Table 3.3. SNPs included in MR analysis

SNP	Beta	SD	P
rs12914385-T	0.213	0.076	9.88^{e-39}
rs1317286-G	0.162	0.078	1.22^{e-38}
rs1051730-A	0.329	0.078	1.80^{e-38}
rs16969968-A	0.314	0.078	2.47^{e-38}
rs951266-A	0.257	0.078	4.61^{e-38}
rs8034191-C	0.121	0.077	6.95^{e-38}
rs17486278-C	0.112	0.078	7.98^{e-38}
rs72740964-A	0.118	0.078	2.12^{e-37}
rs55853698-G	0.285	0.077	1.33^{e-36}
rs8040868-C	0.239	0.075	1.58^{e-35}
rs12910984-G	0.183	0.898	2.13^{e-18}
rs6474412-C	0.193	0.086	3.44^{e-15}
rs7599488-T	0.227	0.074	2.4^{e-13}
rs73229090-A	0.215	0.118	6.1^{e-13}
rs3025343-A	0.152	0.111	1.4^{e-11}
rs2229961-A	0.187	0.268	9.66^{e-10}
rs3733829-G	0.204	0.077	9.03^{e-9}
rs2273506-A	0.126	0.151	4.2^{e-9}

The genetic (allele) score was used to estimate the causal effect of smoking behaviour on outcomes in the UKB population. The genetic score will be discussed in the following section as a part of the one-sample MR approach.

One-sample Mendelian randomization

One-sample MR requires the genetic data for the exposure and the outcome to be from the same population [132]. The smoking behaviour and outcomes data for this thesis were obtained from the UKB. Rather than examining each SNP separately for the causal estimate, a genetic score was used. The genetic score increases statistical power as well as the precision of the IV estimate. The genetic score uses the sum of weights that each SNP contribute to explain the variation in the exposure. The weights of each SNP contribute to smoking behaviour (beta coefficients or log-odds for ORs) were obtained from a meta-analysis of GWASs for smoking behaviour [198]. This meta-analysis extracted the average weights across 33-34 studies. Beta coefficients show the effect per smoking-increasing allele (Table 3.4). These weights were used to construct the genetic scores for smoking behaviour.

To ensure no linkage disequilibrium between the SNPs, plink was used to examine the associations between SNPs. Linkage disequilibrium (LD) is a non-random inheritance of the SNPs caused by close physical proximity on the same chromosome [135]. Such proximity makes it difficult to distinguish the magnitude that each SNP in LD contributes to the causal estimate. After ensuring sufficiently low LD between SNPs (r^2<0.2), the genetic score was built.

The genetic score was constructed in R using the weighted method. This formula was used to create the weighted genetic scores for smoking variables [16]:

$$\text{Genetic (allelic) score} = \frac{\text{snp1}*\text{beta1} + \text{snp2}*\text{beta2} + \ldots + \text{snpn}*\text{betan}}{\text{n}}$$

Where beta is the weight that each SNP contributes to smoking behaviour (beta coefficient) and n is the total number of SNPs. The genetic score was tested for the IV assumptions, examining its association with smoking, the outcomes, and the covariates.

After ensuring the validity of the genetic score for smoking, MR was conducted using two-stage least squares (2SLS).

Two-stage least squares is a statistical approach that uses two stages to conduct MR. The first stage is regressing the risk factor on all the genetic variants in the same model and storing the fitted values of the risk factor. This is followed by the second stage in which regression is then performed with the outcome on the fitted values of the risk factor. Performing MR using 2SLS with aid of (*ivreg/systemfit*) in *ivpack/systemfit* packages in R is recommended as it takes into account the uncertainty in first-stage regression [16]. 2SLS was used throughout this thesis when conducting one-sample MR for smoking behaviour and health outcomes. The threshold of statistical significance was set at P=0.05 (95% confidence level).

Two-sample Mendelian randomization

Two-sample MR uses summary data from publicly available GWAS. The main reason for doing the two-sample MR in this thesis is to compare its findings with the ones obtained from the individual-level (one-sample) MR in the UKB. The data extracted from each population to be used in the two-sample MR are the causal estimates (beta coefficients) and the standard errors (SE). These data are often made available by large consortia. The current thesis exploited the UKB for the genetic-outcomes data and publicly available GWAS for the genetic-smoking data.

There are many methods used in the literature to conduct two-sample MR. In this thesis, MR-Egger was used to test the causal relationship and causal estimate between smoking and the health outcomes of interest. MR-Egger tests for a causal effect, an estimation of this causal effect, as well as corrects for horizontal pleiotropy [201]. MR-Egger method was conducted using *MendelianRandomization and TwoSampleMR* packages in R.

MendelianRandomization package in R provides a wide range of statistical commands for conducting the two-sample MR. After obtaining the summary data (Beta coefficients and SE) for the genetic-smoking and genetic-outcomes, *mr_egger* command in R was used to conduct a two-sample MR using the MR-Egger method. Additionally, *mr.plot* command was used to visualize the results obtained from the analysis. Sensitivity analyses for two-sample MR such as funnel plots were also included in *MendelianRandomization* package in R.

Sensitivity analysis

Sensitivity analyses are intended to check the validity of SNPs (IVs) used in MR analysis [202]. These analyses include heterogeneity, single SNP analysis, and leave-one-out analysis. *MendelianRandomization/ TwoSampleMR* packages were used in R to assess these analyses. Table 3.5 summarises the sensitivity analysis used in this thesis.

Table 3.4. Sensitivity Analysis for two-sample MR

Analysis	Details
Heterogeneity test	Examines the casual estimate variations across the SNPs. [Lower heterogeneity => better reliability of the results]
Single SNP analysis	A summary graph to examine the individual SNP effect.
Leave-one-out analysis	A graph assesses if the SNPs' effect on the outcome is consistent by leaving one SNP out each time. [To ensure no single SNP that drives all the effect]

3.5. Ethical Consideration

The data from the present thesis were obtained from the UKB after a successful application. The UKB studies have ethical approval from the NHS National Research Ethics Service. Participation in UKB studies is based on fully informed consent [183]. The process of data collection ensured a high standard of ethical considerations such as informed consent, beneficence, as well as respect for anonymity and confidentiality. The current data will serve only the purposes and objectives suggested in this thesis.

3.6. Summary

Tables 3.6$_{(a-b)}$ summarise the methods used in this thesis.

Table 3.5$_{(a-b)}$. Methods summary

a) Research Question	Required Data	Type of Analysis
Does smoking observationally associate with the health outcomes of interest?	Smoking behaviour, CMDs and stroke, lipid biomarkers, confounding variables.	Descriptive, multiple regression (R).
Do Instrumental variables valid to be used in MR analysis?	SNPs for smoking status and CperD, CMDs and stroke, lipid biomarkers, confounding variables.	Genetic preparation (Plink), multiple regression (R).
Does smoking causally associate with the health outcomes of interest (MR approach)?	SNPs for smoking status and CperD, CMD and stroke, lipid biomarkers	One-sample MR (2SLS) (R) Two-sample MR (MR-Egger) (MR-Base and R)

b) Mode of comparison	Observational	MR
Variables included	- *Independent variables*: Smoking status CperD SI - *Dependent variables*: CMD and lipid biomarkers - *Covariates*: Age, sex, ethnicity, education, deprivation, and BMI.	- *Independent variables*: SNPs for smoking status SNPs for CperD - *Dependent variables*: CMDs and lipid biomarkers
Sample size	All UKB participants *n=469,598*	White-British: Smoking status (*n=~314k*) CperD (*n=25,724*)
Outcomes	Examining observational associations between smoking and outcomes using linear and logistic regression.	Examining causal associations between smoking and outcomes using 2SLS and MR-Egger

4. Chapter Four: Observational and Mendelian randomization-based causal estimates of the association between smoking behaviour and cardiometabolic/stroke conditions

4.1. Introduction

Overview

This chapter examines the association between smoking behaviour and CMDs. The main goal of this section is to estimate if there are significant associations between smoking behaviour and CMD outcomes in the UKB observationally and genetically (using MR). The chapter starts with a brief review of the associations between smoking variables and CMDs, followed by a review of the methods used in the analysis. Next, the chapter presents the findings of the observational and genetic associations between smoking behaviour and CMDs in the results section. The analysis covers the observational associations followed by one-sample MR in the UKB sample and two-sample MR using the MR-Base platform as well as in R. Finally, the chapter investigates the finding in the discussion section followed by the overall chapter conclusion. Figure 4.1 depicts the chapter structure.

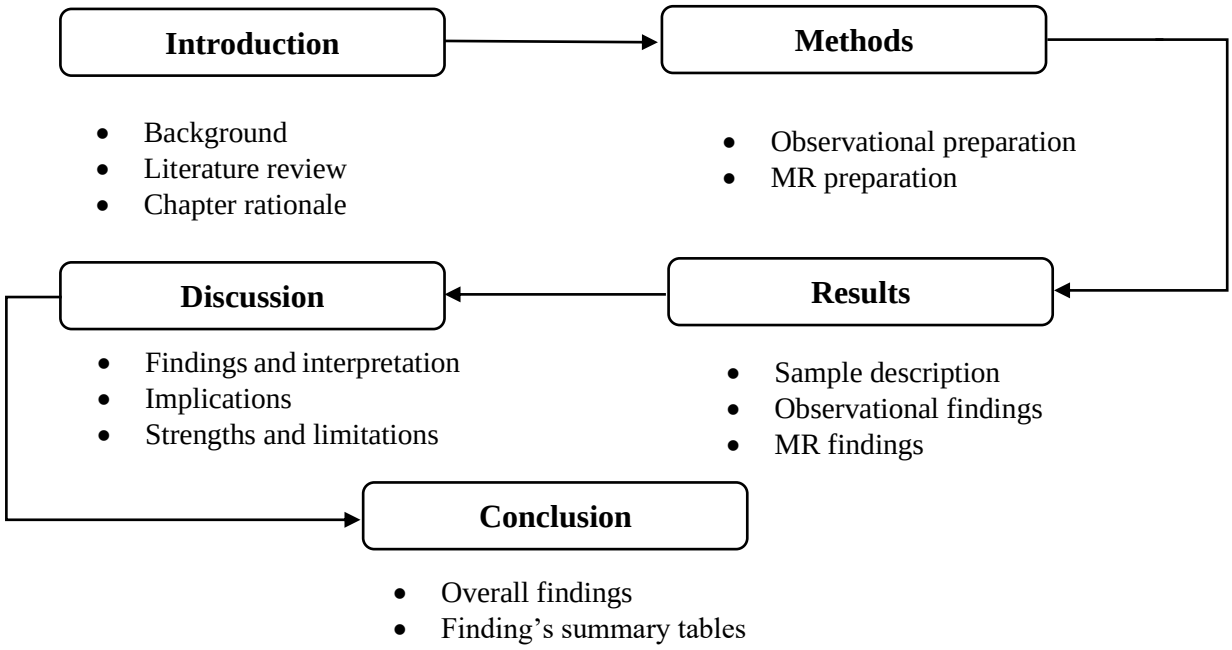


Figure 4.1. Chapter scheme I

Background

The relationships between smoking and CMDs were examined extensively in the literature. A brief review of the associations between smoking and CMDs will be provided in the next sections. A detailed literature review of the associations between smoking and CMDs (CHD, stroke, HTN, and DM) was discussed in chapter two (sections: 2.3.2 – 2.3.5).

A meta-analysis of 55 publications including 141 cohort studies revealed a positive association between cigarette smoking and the risk of CHD and stroke. The risk of CHD and stroke increases as the number of cigarettes smoked increases [69]. The results of Law et al. (2011) review which includes 19 studies are consistent with the previous findings. The risk of CHD among active smokers was 39% higher compared to non-smokers. Additionally, they found that the risk of death from CHD was reduced among smoking quitters [72]. Current smokers have a higher risk of stroke compared to non-smokers. In 2003, Kurth et al. published two prospective cohort studies (9 and 27 years) on the risk of smoking on haemorrhagic stroke. The researchers found that smokers have a higher risk for stroke compared to non-smokers. The risk was higher as individuals smoked more cigarettes per day [75,76]. Similar findings of the association between cigarette smoking and ischemic stroke were reported in Bhat et al. and Fogelhom et al case-control studies [77,78]. The association between cigarette smoking and the risk of DM is also well-established in the literature. Smoking is considered one of the modifiable risk factors for DM [105]. A meta-analysis of 25 studies (from 1966 to 2007) by Willi et al concluded that current smokers have a higher risk of DM compared to non-smokers. The risk increases as the smoking intensity increases [104]. Two prospective cohort studies by Lyssenko et al and Manson et al confirmed the above findings concluding that smoking is a risk factor for DM [106,107].

Genetically, a two-sample MR conducted by Larsson, Burgess, and Michaëlsson [178] found a causal association between smoking initiation and increased risk of ischemic stroke. Additionally, a summary-level MR was conducted to assess the causal relationship between smoking initiation and CVDs. The study found a causal association between smoking and the risk of CHD as well as stroke [179]. Finally, a summary-level MR that included more than one million participants was performed to examine the causal association between smoking and CVDs and associated risk factors [180]. The study found that genetic liability for smoking was associated with an increased risk for CHD, stroke, and DM.

The gap in the literature

The association between cigarette smoking and HTN in particular is not clear [84]. A systematic review of the association between smoking and HTN concluded that the studies reviewed provided conflicting results. The findings of a cross-sectional study by Liu and Byrd revealed that current smokers seem to have better control of their blood pressure [88]. Another cross-sectional study by Alomari and Al-Sheyab found that current smokers have lower blood pressure compared to non-smokers [89]. These findings were also confirmed by Li G et al [81].

This uncertainty was also found in a genetic examination of the relationship between smoking behaviour (rs1051730) and cardiovascular risk factors. Mendelian randomization was performed to assess such a relationship and found no causal association between smoking and HTN or DM among 56,625 participants [39]. Similarly, Linneberg et al found no causal association between smoking status (rs16969968) and HTN [177].

Chapter rationale

Prior observational associations between smoking and CMDs might be confounded as the observations were based on self-report. The known confounding variables such as age, sex, BMI, and race or unknown ones might distort the findings obtained from observational associations [25]. Reverse causation as well might be a problem for the cross-sectional data in that the variables in the analysis represent a snapshot at the time of assessment [203]. A detailed overview of the confounding and reverse causation concepts was provided in chapter one (section: 1.3).

The uncertainty concerning the observational associations makes it difficult to infer causality from such an approach. This drives the researchers to explore more robust techniques. One approach is using RCTs which are considered the gold standard in epidemiological studies [27]. However, RCTs are expensive, laborious, time-consuming and largely limited by ethical considerations [28]. These limitations were the main reasons to use the MR approach which uses the genetically available data on cigarette smoking to infer "causal" associations with the outcomes of interest. Specifically, genetic instrumentation of lifetime smoking risk is leveraged to estimate causal associations between exposures and outcomes [129].

The analysis of smoking behaviour using the MR approach was used in the literature. However, this thesis explored more variables concerning smoking behaviour as well as more outcomes compared to previous studies. This thesis also included a wide range of covariates in a large sample size such as the UKB. Moreover, different MR approaches such as one-sample and two-sample were applied to examine the causal associations of smoking behaviour using UKB data, MR-Base platform, and manual approach in R. Finally, this thesis uses a summed, multi-SNP genetic score rather than

a single SNP MR analysis to enhance the causal estimates. A detailed review of the rationale of this thesis was provided in chapter one (section: 1.3-1.5).

The next section will briefly review the methods used in the analysis and then explore the variables descriptively followed by observational associations and finally MR associations.

4.2. Methods

Overview

The methods were discussed in detail in chapter three, including UKB, covariate measurement and biomarker measurements. For the observational associations between smoking and CMDs, a cross-sectional approach (i.e., testing for associations between reported exposure vs. outcomes) of the UKB data will be used. The smoking behaviour, CMDs and covariates variables at the time of recruitment will be the basis for the analysis. A sample of 469,598 UKB participants was included in the analysis. The independent variables of interest are smoking status (never*previous*current), CperD, and SI. The dependent variables are CHD, stroke, HTN, and DM. The covariates are age, sex, ethnicity, educational attainment, deprivation score (Townsend), and BMI. R software was used for all observational and MR analyses including graphs and tables.

For the Mendelian randomization approach, a sample of 314k white individuals from the UKB will be included in the analysis for testing the causal estimate of the smoking status variable (ever vs. never) based on 15 SNPs. Additionally, the MR analysis will be conducted on a sample of 25,724 current smokers white individuals from the UKB for testing the causal estimate of differences in smoking intensity/average frequency (CperD). The independent variable will be 14 SNPs instrumenting CperD. The outcomes and covariates are the same as the observational analysis (excluding the ethnicity variable).

Observational analysis

To examine the relationship between smoking behaviour and CMDs observationally, regression analyses were performed. As the CMDs are binary outcomes, logistic regression analyses will be used to examine the associations. Frequency tables, crosstabulations as well as visualisation of the variables will be discussed in the descriptive statistics sections and the supplementary materials (8.4, results: descriptive statistics).

MR analysis

Overview

Mendelian randomization approach will be used to examine the causal associations between smoking intensity (CperD) and CMDs and briefly the smoking status variable in the whole sample. The analysis will include genetic quality control results, genetic score, instrumental variables assumptions, one-sample MR using 2SLS, and two-sample MR using MR-Egger. These results will be shown in the MR results section (4.5). Table 4.1 demonstrates the MR analysis approach. The analysis was done for both CperD and smoking status, however, the details of the analysis will be shown only for CperD. Only final MR results will be shown for the smoking status variable. The assumptions testing for the smoking status variable were provided in the supplementary materials (8.4, results, Table 8.35).

Table 4.1. MR approach scheme

Genetic preparation		
Smoking SNPs quality control		Plink output [SNPs included]
Genetic score		From included (valid) SNPs
IV assumptions	**1st**	Smoking associated with genetic score
	2nd	CHD associated with Genetic score Stroke associated with Genetic score HTN associated with Genetic score DM associated with Genetic score
	3rd	Genetic score associated with covariates
MR		
Smoking + SNPs + CMDs		2SLS [one-sample MR: UKB]
UKB: SNPs-CMDs vs MR-Base SNPs-(Smoking)		MR-Egger [two-sample MR: UKB vs MR-Base]

The next section will outline the quality assessment for the genetic data, the genetic score, the instrumental variable (IV) assumptions result for smoking intensity (CperD) and smoking status variables.

Genetic data quality control

The genetic data included in this analysis was the SNPs for smoking intensity (CperD). The smoking status variable preparation followed the same steps used for CperD, hence, only the MR results are shown in this thesis. Genetic preparation for the smoking status variable was provided in the supplementary materials (8.4, results, Table 8.35). The SNPs included were prepared using quality control (QC) measures. The quality measures used were testing these SNPs for Hardy-Weinberg Equilibrium at 1^{e-6} (HWE) and linkage disequilibrium. The LD was used to ensure independence between the SNPs to avoid redundant effects of these variants on smoking behaviour. Other quality control measures such as missingness and relatedness between individuals were also performed. These measures were done using Plink (https://zzz.bwh.harvard.edu/plink/).

The SNPs before QC were twenty-eight. One SNP was not in HWE (rs4803378). Out of all SNPs, fourteen (14) SNPs were in HWE and not in LD (r^2

< 0.20, SNPs that have ≥80% correlation were removed). These SNPs were included in the analysis for MR. Table 4.2 summarises the SNPs across the QC process.

Table 4.2. QC of the SNPs included in the analysis

SNP before QC	SNPs not in HWE	SNPs not in LD	Final SNPs
rs7599488	rs4803378	rs55853698	rs7599488
rs215614		rs17486278	rs215614
rs73229090		rs72740964	rs73229090
rs6474412		rs951266	rs6474412
rs3025343		rs16969968	rs3025343
rs8034191		rs1051730	rs8034191
rs55853698		rs1317286	rs2229961
rs17486278		rs938682	rs12910984
rs72740964		rs12914385	rs3733829
rs951266		rs11637630	rs3865453
rs2229961		rs8040868	rs28399443
rs16969968		rs4803378	rs117824460
rs12910984		rs28399442	rs7260329
rs1051730			rs2273506
rs1317286			
rs938682			
rs12914385			
rs11637630			
rs8040868			
rs3733829			
rs4803378			
rs3865453			
rs28399443			
rs28399442			
rs117824460			
rs7260329			
rs2273506			

Genetic score

The genetic score was used instead of using individual SNP. Smoking will be proxied by this score. The genetic score increases statistical power as well as the precision of the IV estimate. The genetic score was built using the weighted score of each SNP (taken from prior reports) contributing to smoking. The weights of each SNP were obtained from a recently published meta-analysis of GWAS for CperD among

European ancestry (Table 3.4: chapter three) [198]. The details of the genetic score and how it was built were discussed in the methods section.

The genetic score was built using R. The score included only the CperD-increasing allele in which the SNPs that were associated with increased cigarette smoking per day were included. The SNPs were coded as 0, 1 and 2 in which 0 is the homozygous normal (unaffected) allele, 1 is heterozygous and 2 is homozygous for the effect allele. This formula was used to calculate the weighted genetic score.

$$\text{Genetic (allelic) score} = \frac{(\text{rs6474412\_C} * 0.067 + \text{rs12910984\_G} * 0.16 + \cdots \text{snpn} * \text{betan})}{\text{n}}$$

The genetic score was based on valid SNPs. The SNPs included were significantly associated with smoking intensity (CperD), not associated with the outcomes nor the covariates (mostly, for example, some SNPs were significantly associated with HTN and DM, hence, removed and not used to build the genetic score). In other words, each SNP should meet the IV assumption before being added to the genetic score. After building the genetic score, the IV assumptions were tested to ensure the validity of this score. Finally, the genetic score was used to perform MR analyses between smoking and the outcomes. The next section will examine the validity of the genetic score.

IV assumptions results (smoking intensity: CperD)

After building the genetic score, this section focuses on examining the validity of the score to be used as an instrumental variable for smoking. The IV assumptions to be examined here are the significant association between the genetic score and CperD, the association between the genetic score and covariates and the association between the score and the outcomes. R software was used to examine these associations.

..1.2.1 Genetic score vs smoking intensity (CperD) (first IV assumption)

The genetic score was statistically significantly associated with CperD (GWAS level significance). One unit increase in the genetic score will increase the estimate of CperD by 0.223 (B=0.223, P= $1.25x10^{-10}$).

..1.2.2 Genetic score vs CMDs (second IV assumption)

The genetic score was not associated with any of the CMD variables. Table 4.3 summarises the association findings.

Table 4.3. 2nd IV assumption results (genetic score vs CMDs)

CMDs	**Genetic Score**		
	OR	*95% CI*	*p*
CHD	0.98	0.95 – 1.02	0.400
Stroke	0.97	0.88 – 1.07	0.550
HTN	1.01	0.99 – 1.03	0.299
DM	0.98	0.94 – 1.02	0.245

..1.2.3 Genetic score vs covariates (third IV assumption)

The genetic score was not associated with the covariates (except BMI). The genetic score was negatively associated with BMI (B=-0.01, P= 0.01). To account for population stratification, PCA was generated and examined against the genetic score. The ten highest principal components (PCs) were included in the analysis. Overall, the highest PC score was less than 40% factor loadings followed by the second highest PC score which accounts only for 10%. Table 4.4 summarises these findings.

Table 4.4. 3rd IV assumption results (genetic score vs covariates)

Covariates	**Genetic Score**		
	Estimates	*95% CI*	*p*
Age	0.005	-0.00 – 0.00	0.401
Degree	-0.02	-0.07 – 0.03	0.529
Sex	-0.01	-0.05 – 0.03	0.620
Townsend	-0.001	-0.01 – 0.00	0.791
BMI	-0.01	-0.01 – -0.00	**0.010**
PC1	-0.21	-0.22 – -0.21	**<0.001**
PC2	-0.01	-0.02 – 0.00	0.271
PC3	0.61	0.60 – 0.62	**<0.001**
PC4	-0.09	-0.10 – -0.07	**<0.001**
PC5	0.12	0.10 – 0.14	**<0.001**
PC6	0.46	0.44 – 0.48	**<0.001**
PC7	-0.09	-0.11 – -0.07	**<0.001**
PC8	0.26	0.24 – 0.28	**<0.001**
PC9	0.26	0.24 – 0.28	**<0.001**
PC10	0.04	0.02 – 0.06	**<0.001**

The genetic score was statistically significantly associated with the exposure (CperD), not associated with any of the CMD outcomes. Additionally, the associations between the genetic score and the covariates were non-significant with age, degree, sex, and deprivation score. However, the genetic score was significantly associated with BMI and most PCs. This genetic score is valid but not the most ideal IV to proxy smoking intensity.

IV assumptions results (smoking status: never vs ever)

This section briefly investigates the IV assumptions for smoking status (never vs ever) genetic score. One observation worth mentioning here upon examining the IV assumptions for smoking status is that the genetic score for smoking was significantly associated with HTN, education level, BMI, and most PCs. There is a violation of IV assumptions which makes this genetic score not the most ideal proxy for smoking status. The details of the analysis are in the supplementary materials (8.4, results, MR,

smoking status). After exploring the methods of the observational and genetic analysis, the results section follows.

4.3. Results

Overview

This section will provide a detailed description of the variables included in this chapter. The nominal variables will be summarised using tables and pie charts and the numeric/integer variables will be summarised using summary statistics tables. Additionally, crosstabulations and summary statistics of the descriptive associations of the variables will be performed. Finally, line plots for numeric variables across different categories as well as covariates' descriptive associations with smoking variables and CMDs were provided in the supplementary materials (8.4, results, Tables 8.4 – 8.9, and Figures 8.1 – 8.9). After sample description, observational analysis of smoking and CMDs will be performed followed by one-sample then two-sample MR in MR-Base and R.

Sample characteristics

The observational analysis includes 469,598 participants from the UKB population with genetic data after QC. This sample was based on the smoking status variable in which participants who did not respond to smoking questions were removed (n= 2,249). The general demographic, smoking and health-related characteristics are summarised in Table 4.5. The table is divided based on the type of variable either qualitative (binary/nominal) or quantitative (numeric/integer). The majority of participants were British (88.42%) and were without a university/college degree (67.7%). The sample mean age was 56.53 years (±8.09), and 54.5% of the sample were females. The population mean age was 56.53 years (±8.09), 54.5% were females while the rest (45.5%) were males. For MR analysis, only individuals who descended from European

ancestry and have genotyped SNPs for CperD (n=25274, mean age=54.81±8.05, female=51.91%, male=48.09%) as well as for smoking status variable (n=314K, mean age=56.8, female=54.02%, male=45.98%) were included. The rest of the covariates' characteristics are provided in supplementary materials (8.4, results, smoking vs covariates, CMDs vs covariates).

Table 4.5. Sample characteristics-observational (n=469,598)

Variable	**Level**	**Count (%)**
Smoking status	Current	52431 (10.56%)
	Previous	172216 (34.68%)
	Never	271951 (54.76%)
Sex	Male	226177 (45.5%)
	Female	270421 (54.5%)
Ethnicity	White British	439085 (88.42%)
	Other ethnicities	57513 (11.58%)
Degree (college/university)	No Degree	336334 (67.73%)
	Degree	160264 (32.27%)
Coronary Heart Disease (CHD)	No	474003 (95.5%)
	Yes	22589 (4.5%)
Stroke	No	490474 (98.77%)
	Yes	6124 (1.23%)
Hypertension (HTN)	No	377608 (76.04%)
	Yes	118990 (23.96%)
Diabetes Mellitus (DM)	No	470516 (94.75%)
	Yes	26082 (5.25%)
Variable	**Mean (SD)**	
Cigarette Smoked per Day (CperD)	15.5 (±8.39) cigarette/day	
Smoking Initiation (SI): (Age started smoking)	17.85 (±5.8) years	
Age	56.53 (±8.09) years	
Body Mass Index (BMI)	27.42 (±4.79)	
Deprivation Level (Townsend score)	-1.31 (±3.08)	

Power analysis

The appropriate sample size for the UK population with significance at P<0.05, power of 0.95, with a large effect size (0.35) and using eight predictors in multiple linear regression analysis was calculated to be 74 participants for each group. The minimum sample size for logistic regression was calculated to be N=1299 participants at OR: 1.493, P=0.05 (power=0.95). The UKB sample used in the current thesis was above the required threshold.

Descriptive statistics

In the UKB data, many smoking behaviour-associated variables are available. For example, smoking history, smoking intensity, age at initiation, parent smoking history, and smoking cessation. The independent variables (exposures) included in this thesis were smoking status, CperD and SI. The dependent variables (outcomes) are CMDs. This section will explore the descriptive statistics of these variables in the UKB.

Smoking status variable

The smoking status is a categorical (nominal) variable with three levels (current smokers, previous smokers and never smokers). Out of the whole sample of the UKB (n=502,536), almost 99% (496,598) of participants responded to the smoking status question of being current (10.56%), previous (34.68%) or never smoked (54.76%). According to the results above, the sample was almost divided equally between the people who ever smoked and never smoked (45.33% vs 54.76%). The observational analyses were based on the subsetted sample in which participants responded to the smoking status question (n=496,598). A visual representation of the smoking status categories is shown in Figure 4.2.

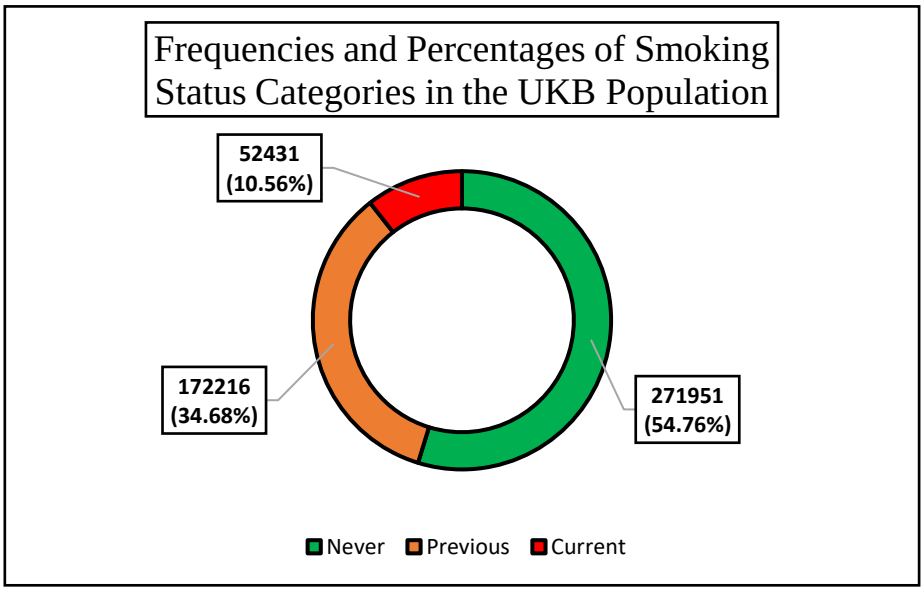


Figure 4.2. Frequencies and percentages of smoking status levels

Smoking intensity: Cigarettes smoked per day (CperD)

This variable represents how much an individual smokes per day as a current smoker. CperD (smoking intensity) is a quantitative variable. The individuals who reported their smoking intensity were 35,758. On average, the participants smoke 14.85 (SD = ±7.11, minimum: 1, maximum: 35) cigarettes per day.

Smoking initiation SI; Age started smoking)

Smoking initiation describes when current smokers started to smoke. Around 38,347 participants revealed when they started to smoke. On average, the participants started to smoke at 16.7 (±3.12 years, minimum: 10, maximum: 25). The following section describes the CMD variables.

CMDs summary statistics (dependent variables)

Cardiometabolic diseases include many health outcomes. This thesis will focus on CHD, stroke, HTN and DM because they are well-known as common, and prevalent with significant public health costs. These variables are binary. Most participants in the UKB have answered the questions regarding CMDs. The participants were categorised

based on their responses to the question of being diagnosed with these diseases or not. Out of 496,598 individuals included in the sample, 22,589 (4.5%) participants have CHD, 6124 (1.23%) participants have had a stroke, 118990 (23.96%) participants have HTN, and 26082 (5.25%) participants have DM. The descriptive analysis of the relationship between smoking behaviour vs CMDs, smoking vs covariates as well as CMDs vs covariates is provided in the supplementary materials (8.4, results, Tables 8.9 – 8.9).

Observational Analysis

This section examines the observational associations (inferentially) between smoking behaviour and CMDs. It included logistic regression analyses of the associations between smoking status, CperD and SI and CMDs. The analysis will use odds ratios (OR) as an effect size for the relationship between the binary outcomes (CMDs) and other variables. The significance level was set at 5% (95% confidence). All relationships will include unadjusted (smoking variable only vs outcome) as well as adjusted (smoking variable + covariates vs outcome). Simple logistic regression was used for unadjusted analyses while multiple logistic regression was used for adjusted association. The adjustment will help to minimise the risk of confounding variables.

Smoking status and cardiometabolic diseases (CMDs)

This section examines the associations between the smoking status variable and CMDs variables. The smoking status variable will be analysed in three categories (current, previous, and never). However, the analysis of smoking as a binary variable (ever vs never) was also performed and the results were provided in the supplementary materials (8.4, results, Tables 8.22 – 8.25). The rationale behind categorising the smoking status variable is to distinguish between people who ever smoked and people who never smoked. Adding previous smokers to the current will make the sample comparable

between people who ever smoked and people who never smoked (n=224,647, n=271,951, respectively). Each section will discuss unadjusted associations followed by adjusted results. The reference levels of the categorical variables used in the whole analysis in this thesis are summarised in Table 4.6.

Table 4.6. Reference levels of the categorical variables

Variable	Reference Level
Smoking status	Never
Sex	Female
Education	No degree
Ethnicity	Other ethnicities (non-white British)

..1.2.4 Smoking status vs coronary heart disease (CHD)

When examining smoking status against CHD (unadjusted), current smokers, as well as previous smokers, were both having a statistically significantly higher risk to have CHD compared to never smokers. Current smokers were at almost double the risk of CHD compared to never smokers (OR: 1.90, 95% CI: 1.82 – 1.99, P<0.001). Previous smokers have even more risk to develop CHD compared to never smokers (OR: 2.24, 95% CI: 2.18 – 2.31, P<0.001). After adjusting for the covariates (multiple logistic regression), current and previous smokers still have a positive and significant association with CHD. Current smokers have a 61% higher risk for CHD compared to never smokers (OR: 1.61, 95% CI: 1.54 – 1.68, P<0.001). Previous smokers have around a 50% risk of having CHD compared to never smokers (OR: 1.50, 95% CI: 1.45 – 1.54, P<0.001). After categorising smoking status into ever vs never, individuals who ever smoked have a 52% higher risk of reporting CHD compared to never-smoked individuals (OR: 1.52, 95% CI: 1.48 – 1.56, P<0.001). A summary of these findings is shown in Figure 4.3. The rest of the associations (smoking status as a binary variable) are in the supplementary materials (8.4, results, Tables 8.22 – 8.25).

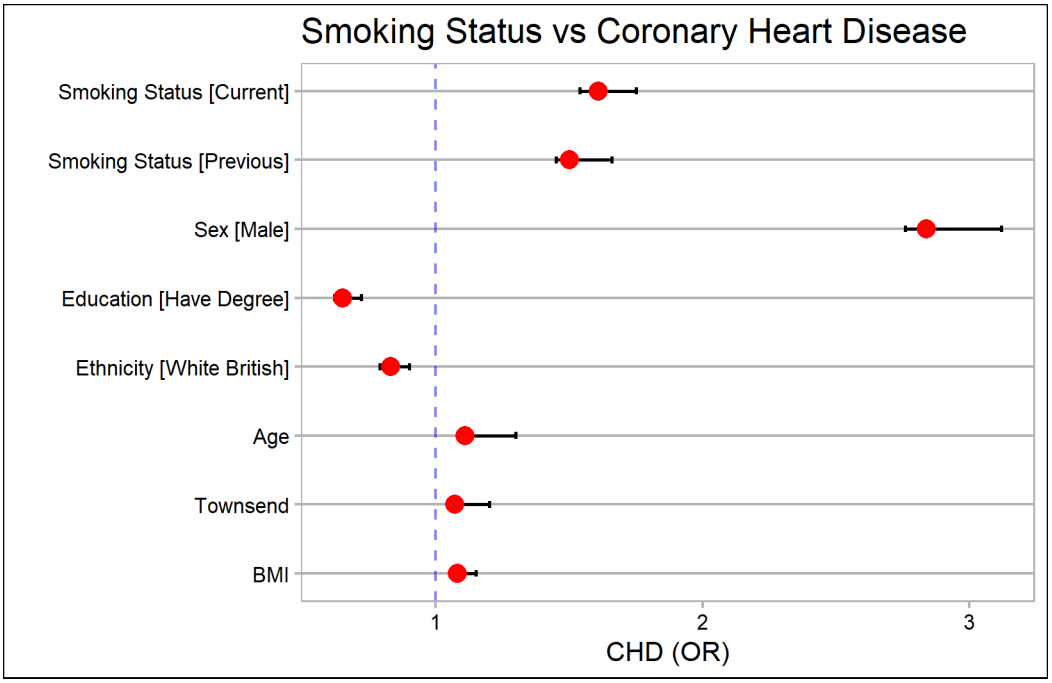


Figure 4.3. Visualisation of adjusted associations: smoking status vs CHD

..1.2.5 Smoking status vs stroke

Unadjusted association between smoking status and stroke revealed that current and previous smokers have a positive and statistically significant association with stroke compared to never smokers. Current smokers have an 82% higher risk to have a stroke compared to never smokers (OR: 1.82, 95% CI: 1.68 – 1.96, P<0.001). Previous smokers carry almost 51% higher risk of stroke compared to never smokers (OR: 1.51, 95% CI: 1.43 – 1.59, P<0.001). After including the covariates in the regression model, current and previous smokers still have a statistically significant positive association with stroke. Current smokers have a 64% higher risk to have a stroke compared to never smokers (OR: 1.64, 95% CI: 1.52 – 1.77, P<0.001). Previous smokers have around 16% risk of reporting stroke compared to never smokers (OR: 1.16, 95% CI: 1.10 – 1.23, P<0.001). After categorising smoking status into ever vs never, ever smokers have a 26% higher risk of stroke compared to never smokers (OR: 1.26, 95% CI: 1.19 – 1.33, P<0.001). A summary of these findings is shown in Figure 4.4.

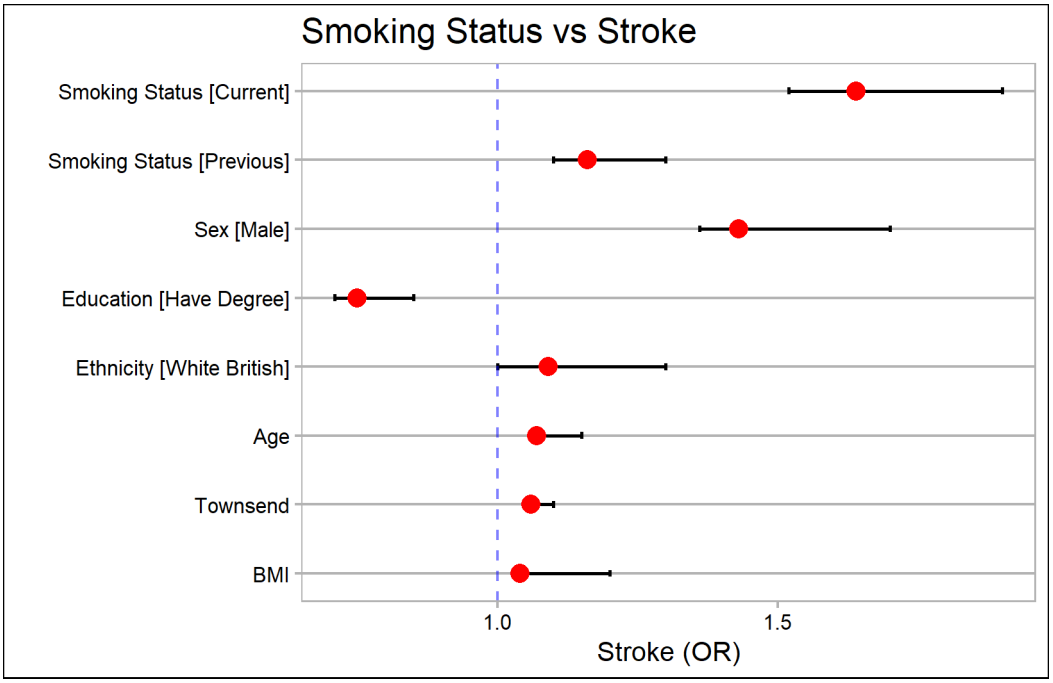


Figure 4.4. Visualisation of adjusted associations: smoking status vs stroke

..1.2.6 Smoking status vs hypertension (HTN)

Unadjusted association between smoking status and HTN revealed that current smokers have a negative and statistically significant association with HTN compared to never smokers (OR: 0.89, 95% CI: 0.87 – 0.91, P<0.001). On the contrary, previous smokers carry an almost 23% higher risk of HTN compared to never smokers (OR: 1.23, 95% CI: 1.21 – 1.24, P<0.001). After including the covariates in the regression model, the current smokers remain negatively and statistically significantly associated with HTN (OR=0.89, 95% CI: 0.87 – 0.91, P<0.001). The previous smokers became negative/borderline but statistically non-significantly associated with HTN (OR=0.99, 95% CI: 0.98 – 1.01, P=0.221). After categorising smoking status into ever vs never, ever smokers have a 3% lower risk of HTN compared to never smokers (OR: 0.97, 95% CI: 0.96 – 0.98, P<0.001). A summary of these findings is shown in Figure 4.5.

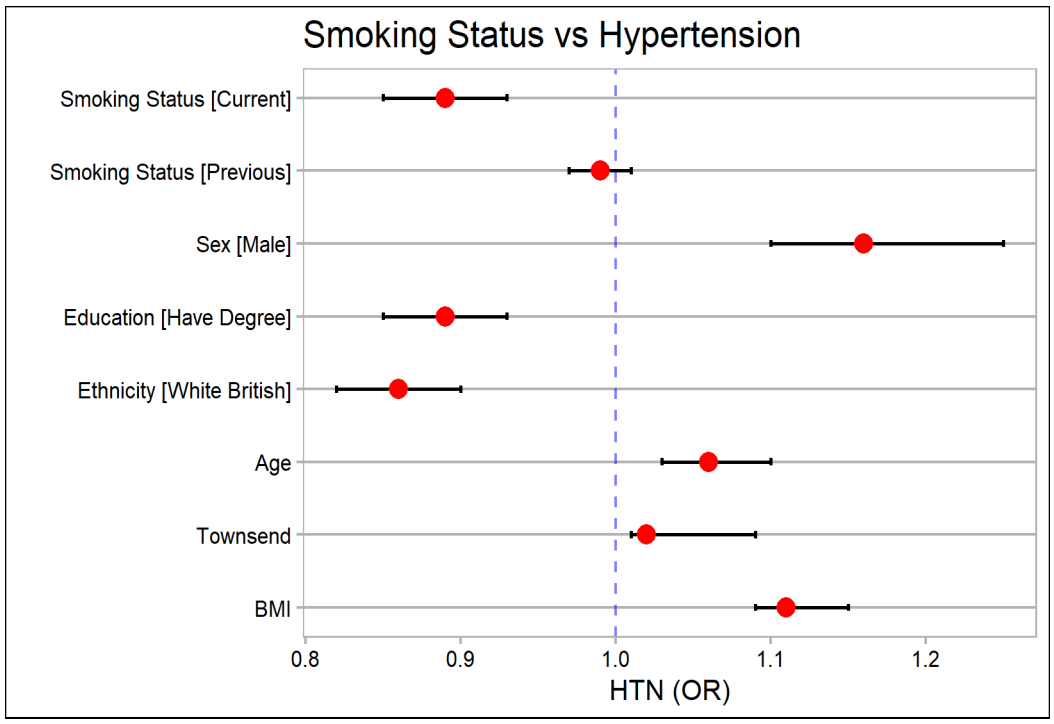


Figure 4.5. Visualisation of adjusted associations: smoking status vs HTN

..1.2.7 Smoking status vs diabetes mellitus (DM)

Unadjusted association between smoking status and DM revealed that current and previous smokers have a positive and statistically significant association with DM compared to never smokers. Current smokers have a 26% higher risk for DM compared to never smokers (OR: 1.26, 95% CI: 1.21 – 1.32, P<0.001). Previous smokers carry an almost 50% higher risk of DM compared to never smokers (OR: 1.50, 95% CI: 1.46 – 1.54, P<0.001). After including the covariates in the regression model, current and previous smokers still have a statistically significant positive association with DM. Current and previous smokers have a 12% higher risk to report DM compared to never smokers (OR: 1.12, 95% CI: 1.07 – 1.17, P<0.001, OR: 1.12, 95% CI: 1.09 – 1.16, P<0.001, respectively). Similarly, after categorising smoking status into ever vs never, ever smokers have a 12% higher risk of DM compared to never smokers (OR: 1.12, 95% CI: 1.09 – 1.15, P<0.001). A summary of these findings is shown in Figure 4.6.

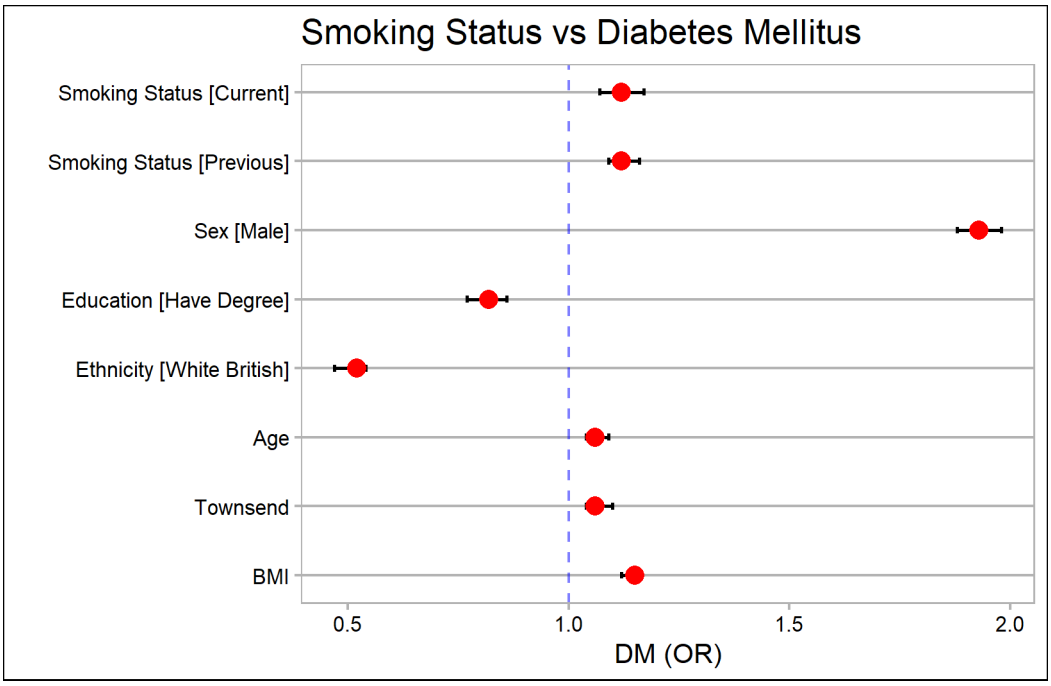


Figure 4.6. Visualisation of adjusted associations: smoking status vs DM

Smoking intensity (CperD) and cardiometabolic diseases (CMDs)

This section examines the associations between the cigarettes smoked per day (CperD) variable and CMDs variables. Each section will discuss unadjusted associations then followed by adjusted results.

..1.2.8 Smoking intensity (CperD) vs coronary heart disease (CHD)

Cigarettes smoked per day are a measure of smoking intensity. In this part, CperD will be examined against CHD. Unadjusted association between CperD and CHD revealed positive and statistically significant findings. An additional cigarette smoked per day will increase the risk of CHD by 3% (OR=1.03, 95% CI: 1.02 – 1.03, P<0.001). This effect of CperD on CHD remained after adjustment for the confounders. An additional cigarette smoked per day will increase the risk of CHD by 1% (OR=1.01, 95% CI: 1.00 – 1.02, P<0.001). A summary of these findings is shown in Figure 4.7.

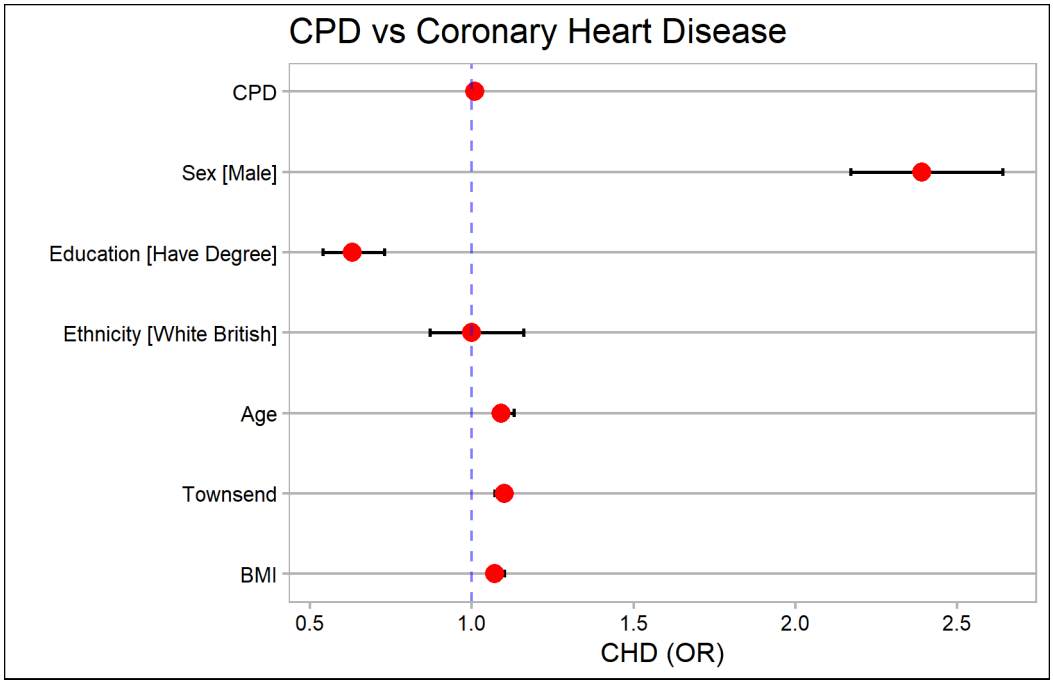


Figure 4.7. Visualisation of adjusted associations between CperD and CHD

..1.2.9 Smoking intensity (CperD) vs stroke

Unadjusted association between CperD and stroke revealed positive and statistically significant findings. An additional cigarette smoked per day will increase the risk of stroke by 2% (OR=1.02, 95% CI: 1.01 – 1.03, P<0.001). This effect of CperD on stroke persisted after the adjustment for the confounders. An additional cigarette smoked per day will increase the risk of stroke by 1% (OR=1.01, 95% CI: 1.00 – 1.02, P<0.009). A summary of these findings is shown in Figure 4.8.

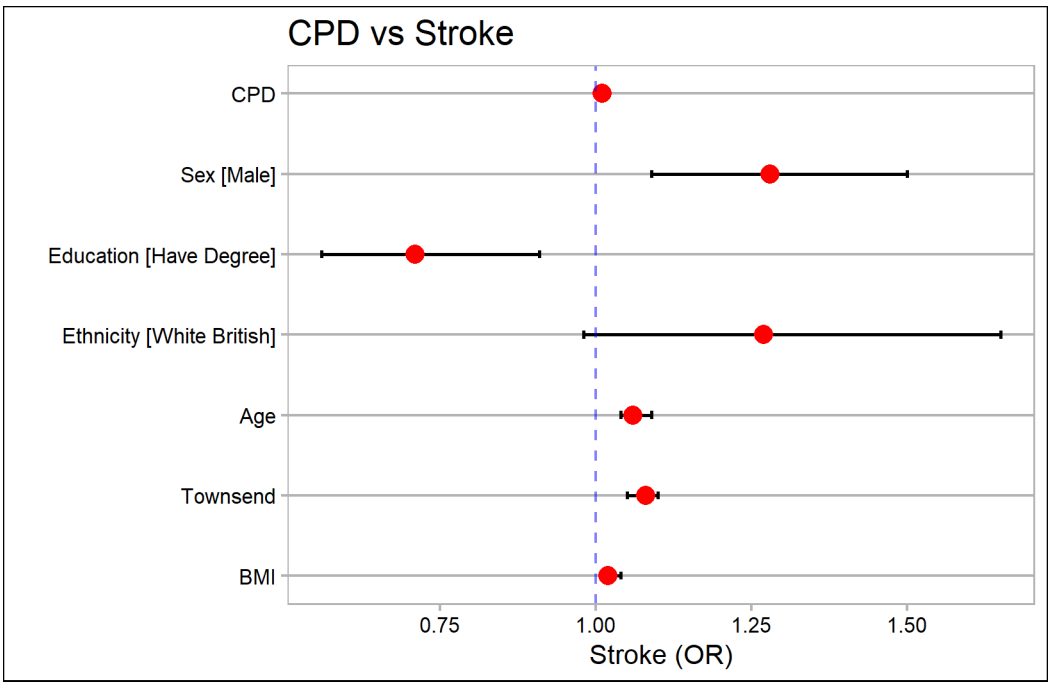


Figure 4.8. Visualisation of adjusted associations between CperD and stroke

..1.2.10 Smoking intensity (CperD) vs hypertension (HTN)

Unadjusted association between CperD and HTN revealed positive and statistically significant findings. An additional cigarette smoked per day will increase the risk of HTN by 1% (OR=1.01, 95% CI: 1.01 – 1.02, P<0.001). This effect of CperD on HTN persisted after the adjustment for the confounders. An additional cigarette smoked per day will increase the risk of HTN by 1% (OR=1.01, 95% CI: 1.00 – 1.01, P<0.001). A summary of these findings is shown in Figure 4.9.

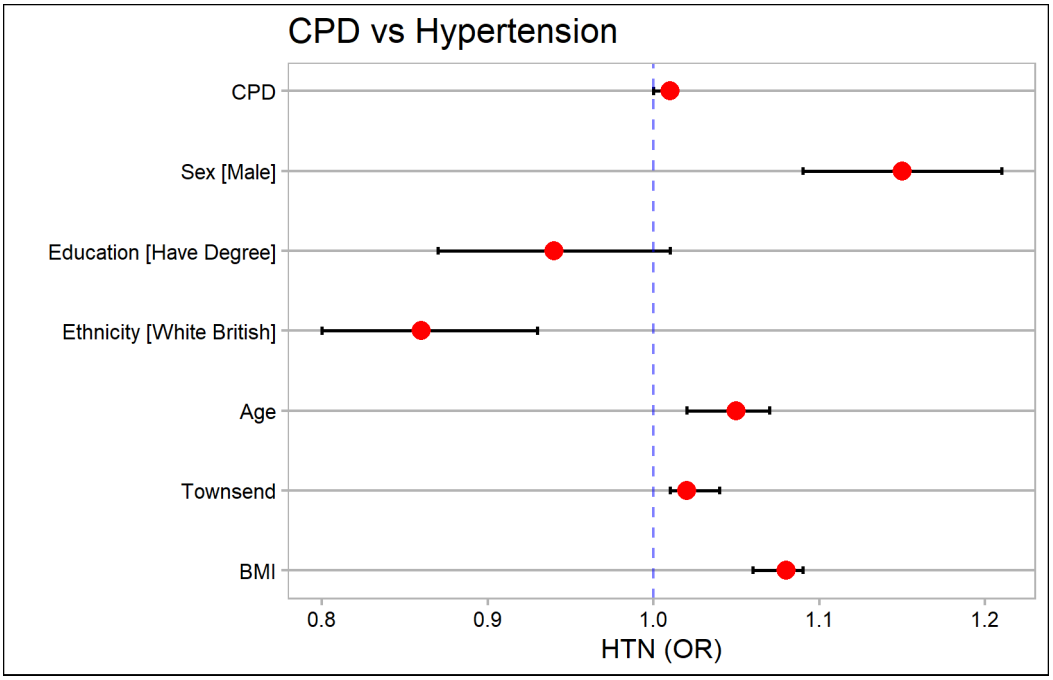


Figure 4.9: Visualisation of adjusted associations between CperD and HTN

..1.2.11 Smoking intensity (CperD) vs diabetes mellitus (DM)

Unadjusted association between CperD and DM revealed positive and statistically significant findings. An additional cigarette smoked per day will increase the risk of DM by 3% (OR=1.03, 95% CI: 1.02 – 1.03, P<0.001). This effect of CperD on DM persisted after the adjustment for the confounders. An additional cigarette smoked per day will increase the risk of DM by 2% (OR=1.01, 95% CI: 1.01 – 1.02, P<0.001). A summary of these findings is shown in Figure 4.10.

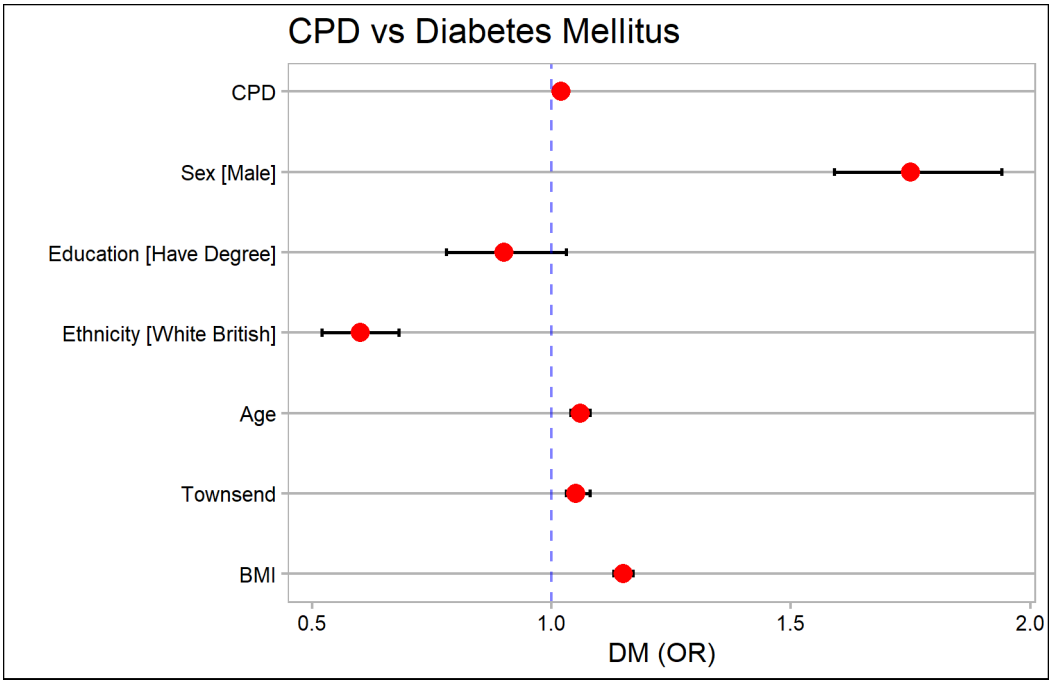


Figure 4.10. Visualisation of adjusted associations between CperD and DM

Smoking Initiation (SI) and cardiometabolic diseases (CMDs)

This section examines the associations between smoking initiation (age individuals started to smoke) (SI) variable and CMDs variables.

..1.2.12 Smoking initiation vs coronary heart disease (CHD)

Unadjusted association between SI and CHD revealed negative and statistically significant findings. The risk of CHD decreases by 5% as an individual started to smoke one year older (OR=0.95, 95% CI: 0.94 – 0.96, P<0.001). This effect of SI on CHD persisted after the adjustment for the confounders. The risk of CHD decreases by 4% when an individual started to smoke one year older (the earlier an individual started to smoke, the more the risk of CHD) (OR=0.96, 95% CI: 0.95 – 0.97, P<0.001). A summary of these findings is shown in Figure 4.11.

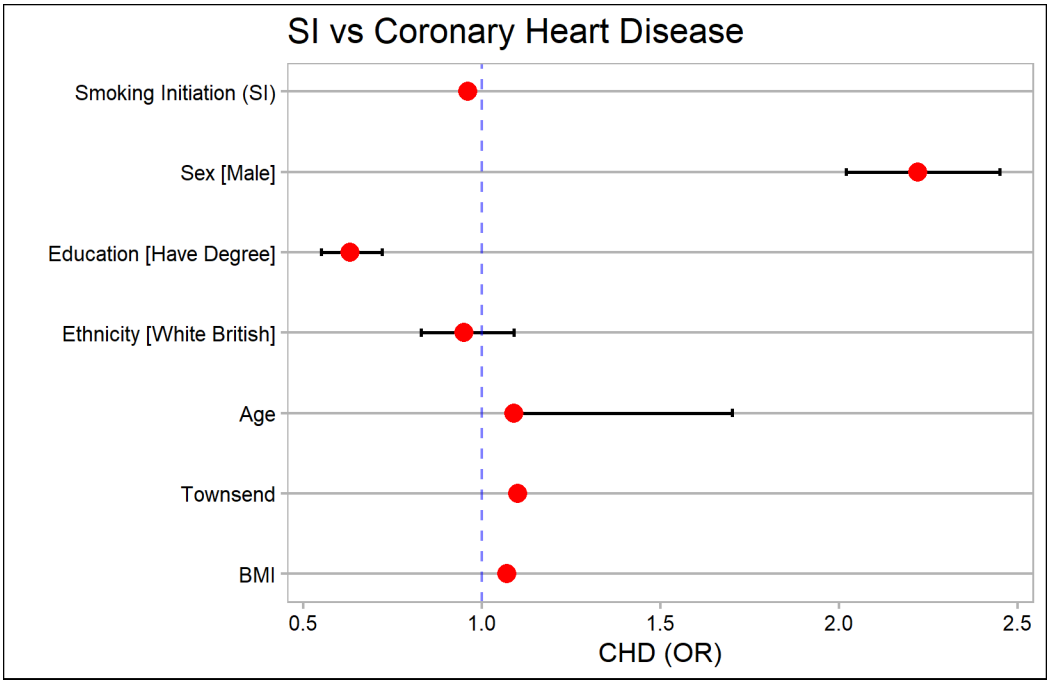


Figure 4.11. Visualisation of adjusted associations between SI and CHD

..1.2.13 Smoking initiation vs stroke

Unadjusted association between SI and stroke revealed negative and statistically significant findings. The risk of stroke decreases by 4% as an individual started to smoke one year older. (OR=0.96, 95% CI: 0.94 – 0.97, P<0.001). This effect of SI on stroke persisted after the adjustment for the confounders. The risk of stroke decreased by 4% as an individual started to smoke one year older (OR=0.96, 96% CI: 0.95 – 0.98, P<0.001). A summary of these findings is shown in Figure 4.12.

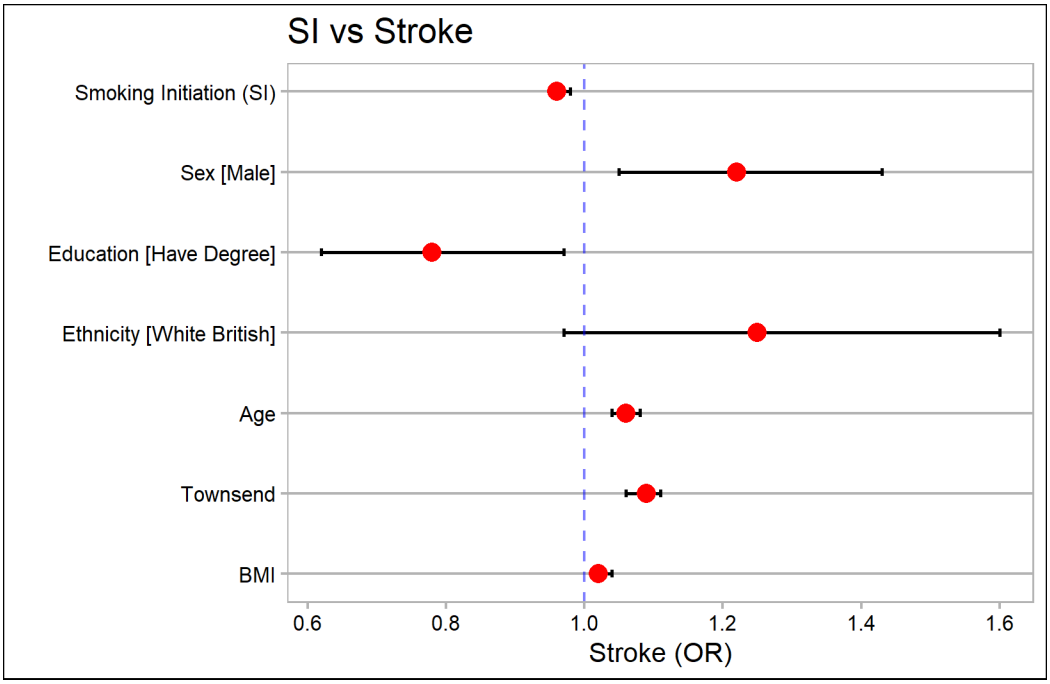


Figure 4.12. Visualisation of adjusted associations between SI and stroke

..1.2.14 Smoking initiation vs hypertension (HTN)

Unadjusted association between SI and HTN revealed positive and statistically significant findings. The risk of HTN increases by 1% as an individual started to smoke one year older. (OR=1.01, 95% CI: 1.00 – 1.01, P<0.001). This effect of SI on HTN persisted after the adjustment for the confounders. If an individual started to smoke older by one year, the risk of HTN increases by 1% (OR=1.01, 95% CI: 1.00 – 1.01, P<0.001). A summary of these findings is shown in Figure 4.13.

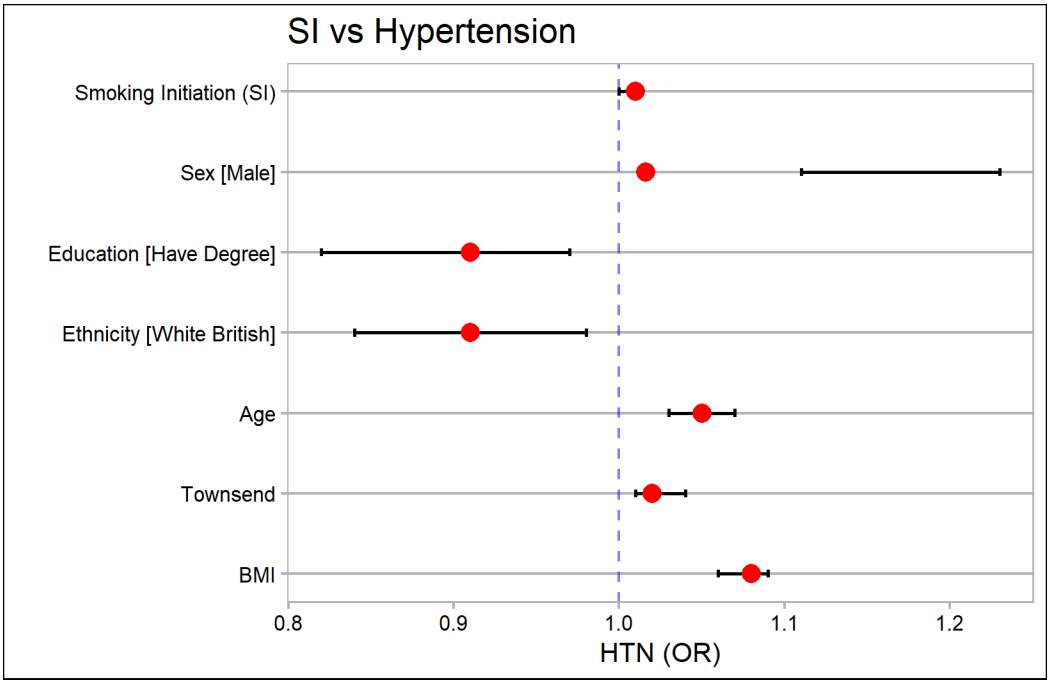


Figure 4.13. Visualisation of adjusted associations between SI and HTN

..1.2.15 Smoking initiation vs diabetes mellitus (DM)

Unadjusted association between SI and DM revealed negative and statistically non-significant findings. The risk of DM decreases by 1% as an individual started to smoke one year older. (OR=0.99, 95% CI: 0.98 – 1.00, P=0.076). This effect of SI on DM remained non-significant after the adjustment for the confounders. The older an individual started to smoke by one year, the lower the risk of DM by 1% (OR=0.99, 96% CI: 0.99 – 1.00, P=0.190). A summary of these findings is shown in Figure 4.14.

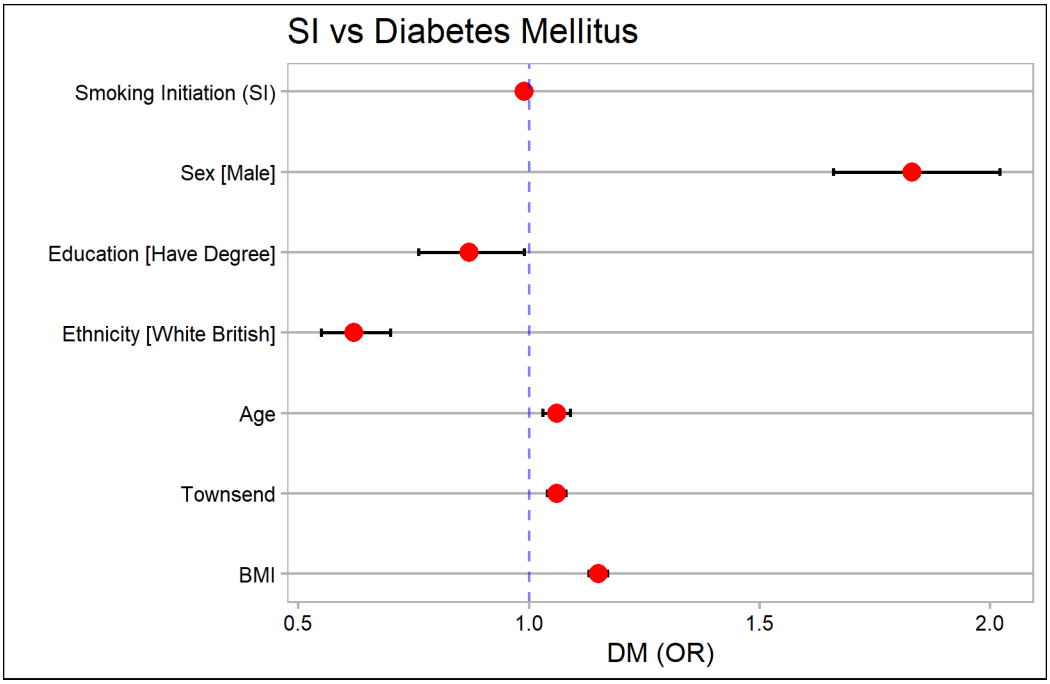


Figure 4.14. Visualisation of adjusted associations between SI and DM

Summary

The observational analysis of the associations between smoking variables and CMDs was presented in this section. Compared to never smokers, current and previous smokers were associated with increased risk of all CMDs except HTN (decreased risk only among current). Additionally, the more cigarettes an individual smokes seem to be associated with an increased risk of all CMDs. Finally, the earlier individual started to smoke the more the risk of all CMDs except HTN. Table 4.7 summarises the main findings of the observational associations between smoking variables and CMDs and stroke.

Table 4.7. Summary of the observational analysis between smoking and CMDs

Variables		**CHD**	**Stroke**	**HTN**	**DM**
Smoking Status	**Current**	Risk increased	Risk increased	Risk decreased	Risk increased
	Previous	Risk increased	Risk increased	Risk decreased (non-significant)	Risk increased
CperD (More cigarettes to smoke)		Risk increased	Risk increased	Risk increased	Risk increased
SI (To start smoking earlier)		Risk increased	Risk increased	Risk decreased	Risk increased (non-significant)

Mendelian randomization (MR) analysis

Overview

The observational analysis of the associations between smoking behaviour and CMDs revealed that smoking behaviour was associated with an increased risk of CMDs (except HTN). The CperD was associated with increased risk for all CMDs variables. However, these associations were based on a cross-sectional analysis of the observational data of the UKB participants. One of the issues that are worth noticing is the presence of the confounders which might falsify these results. These observational models only explained very little variation in the outcome variables. The pseudo-R-squared for all models was ranging from 0.4% – 13%. Additionally, the covariates used in the models were almost always statistically significantly associated with the CMDs and smoking variables suggesting some degree of confounding, including most likely in variables not assessed.

For the previous reasons, another method to examine these associations between smoking and CMDs was introduced. MR approach which uses genetic variants to proxy smoking will be used in the following sections. This approach will ensure that the association between smoking and other variables will be free of confounders and other unknown factors that might affect the observational associations, so we can infer causality from such associations.

The MR approach in this thesis will use smoking status as well as CperD variables as proxies for smoking behaviour. The sample only included individuals that descended from European ancestry (Caucasian British). The next sections will examine the genetic associations between smoking and CMDs.

One-sample Mendelian randomization

This section focuses on the causal association between smoking behaviour and CMD variables using one-sample MR. This thesis uses the UKB for both exposures and outcomes. The aim is to estimate the causal relationship between smoking behaviour and CMDs as well as the effect size of this estimate. All MR analyses were obtained using two-stage least squares (2SLS) manually and using *systemfit* package in R. The 2SLS encompasses two stages in which regression of exposure on the IV in the first stage, followed by regression of CMDs on the results obtained from the first stage.

The IV assumptions revealed a GWAS significant association between smoking status (ever) and the genetic score. The rest of the assumptions revealed no significant association with CHD, stroke, and DM. However, the genetic score was significantly associated with HTN, education level, BMI, and most PCs. Regarding CperD, there was a GWAS significant association between CperD and the genetic score. Additionally, there was no significant association between the CMDs and the genetic score. Finally, the genetic score was not independent of the covariates or the PCs. The following sections show the findings of the genetically based (MR) associations between smoking behaviour and CMDs. The odds ratio will be the causal estimate for all associations.

..1.2.16 Smoking status (n=314k) vs cardiometabolic diseases (CMDs)

The observational associations between smoking status and CMDs revealed that current smokers have a higher risk for all CMD variables. Even after categorising smoking status as a binary (ever vs never), these significant associations persisted. When using the MR approach, smokers seem to have a lower risk for CHD (OR= 0.96), stroke (OR= 0.97) and HTN (OR= 0.44) compared to non-smokers. Conversely, smokers have a slightly higher risk for DM compared to non-smokers (OR=1.01). There was no

evidence of a causal relationship between genetically estimated smoking status and CHD, stroke and DM risks compared to non-smokers (P=0.592, P=0.89 and P=0.931, respectively). However, there was evidence of a causal association between smokers and a lower risk of HTN (P=0.001). Table 4.8 summarises MR findings.

Table 4.8. Summary of MR findings: smoking status vs CMDs

Variable	**Smoking Status** (ever vs never)	
	MR Estimate	**P value**
CHD	OR=0.96	0.592
Stroke	OR=0.97	0.089
HTN	OR=0.44	**0.001**
DM	OR=1.01	0.931

..1.2.17 Smoking intensity (CperD, n=25k) vs cardiometabolic diseases (CMDs)

Observationally, as an individual smokes more cigarettes per day, the risk of CHD increases. Genetically, there was no evidence of a causal relationship between cigarettes per day vs. CHD risk. The risk of CHD was 7% lower; however, the findings were statistically non-significant (causal estimate: OR=0.93, P= 0.400). Regarding stroke, the observational association between CperD and stroke was positive and statistically significant. However, the estimate using MR analysis revealed that there was no evidence of a causal relationship between CperD and stroke risk. The risk of stroke was 13% lower for each additional cigarette smoked per day (causal estimate: OR=0.87, P= 0.550). Similarly, the observational relationship between CperD and HTN was positive and statistically significant. However, when using the MR approach, there was no evidence of a causal relationship between cigarettes per day vs. HTN risk. The risk of HTN was still higher (5%) but statistically non-significant (causal estimate: OR=1.05, p-value: 0.299). A positive and significant observational association was also found between CperD and DM. however, genetically, there was no evidence of a causal relationship between cigarettes per day vs. DM risk. The risk of DM seems to be lower

(9%) for each additional cigarette smoked per day (OR=0.91, p-value: 0.245). These findings between genetically estimated smoking intensity and CMDs were non-significant therefore, causal inference cannot be established. Table 4.9 summarises MR findings.

Table 4.9. Summary of MR findings: CperD vs CMDs

Variable	**Smoking intensity** (CperD)	
	MR Estimate	**P value**
CHD	-0.07 [OR:0.93] ↓	0.400
Stroke	-0.138 [OR:0.87] ↓	0.550
HTN	0.05 [OR=1.05] ↑	0.299
DM	-0.10 [OR:0.91] ↓	0.245

..1.2.18 Summary

This section showed a one-sample (individual-level) Mendelian randomization analysis of the relationship between genetically estimated smoking behaviour and CMDs. The genetic approach was started by selecting and examining the SNPs to be included in the analysis based on quality control of the genetic data. Additionally, the genetic score for smoking variables was built and tested for its validity as an instrument for these variables (IV assumptions). Finally, a one-sample MR was done to examine the causal relationship between smoking behaviour and CMDs. Out of twenty-eight SNPs, fourteen SNPs were not in LD and HWE. These SNPs were tested individually for IV assumptions and then exploited to build the genetic score. The genetic score was significantly (GWAS-level) associated with smoking status, not directly associated with CMDs (except HTN) and not associated with the covariates (except for education level, BMI and PCs). Additionally, the genetic score was significantly (GWAS-level) associated with CperD, not directly associated with CMDs, and not associated with the covariates (except for BMI and PCs).

The findings of MR analysis of the smoking status and CMDs revealed that ever-smokers have a lower risk for CHD, stroke and HTN and a higher risk for DM compared to never-smokers. These findings showed no evidence of a causal association between ever-smokers and CHD, stroke nor DM compared to never-smokers. However, there was evidence of a causal association between ever smokers and decreased risk of HTN (OR=0.44, P=0.001). The findings of MR analysis of the smoking intensity (CperD) revealed no evidence of causal association with CMDs. Genetic predisposition to smoking intensity (CperD) based on 14 SNPs was negatively associated with CHD, stroke and DM and positively associated with HTN. Table 4.10 summarises these findings. The final section of this chapter will explore a two-sample MR for the relationship between smoking and CMDs.

Table 4.10. Summary of MR findings for smoking behaviour vs CMDs

Variable	**Smoking status**		**Smoking intensity** (CperD)	
	MR Estimate	**P value**	**MR Estimate**	**P value**
CHD	OR=0.96 ↓	0.592	OR=0.93 ↓	0.400
Stroke	OR=0.97 ↓	0.089	OR=0.87 ↓	0.550
HTN	OR=0.44 ↓	**0.001**	OR=1.05 ↑	0.299
DM	OR=1.01 ↑	0.931	OR=0.91 ↓	0.245

Two-sample Mendelian randomization (2SMR)

This section explored the two-sample MR of the relationship between smoking behaviour and CMDs. The main goal of the two-sample MR in this thesis is to compare the results of the individual-level MR in the UKB with other samples (such as GWAS & Sequencing and Consortium of Alcohol and Nicotine use (GSCAN)). The CMDs data was obtained from the UKB in all approaches. The first MR analysis was conducted using the MR-Base platform (http://app.mrbase.org/). The CMDs data was obtained from the UKB while the smoking SNPs were acquired from different samples (GSCAN). The second analysis included smoking behaviours' SNPs (betas and SEs)

from the meta-analysis of GWAS for smoking behaviour of European ancestry, and the outcomes' SNPs (betas and SEs) were obtained from the UKB (same SNPs used in one-sample MR) [198]. The latter analysis was done in R (*TwoSampleMR* and *MendelianRandomization*).

This section began with a detailed analysis of the genetic association between smoking intensity (CperD) and CMDs; including two-sample MR in MR-Base, followed by a sensitivity analysis to evaluate MR results, then performing two-sample MR in R. The main difference between 2SMR in MR-Base and R is the latter used the same SNPs used in the one-sample MR analysis in the UKB sample. Next, a brief summary-level MR of the association between smoking status (ever-never) and CMDs using MR-Base. Finally, a brief comparison between one-sample MR and two-sample MR findings was provided.

..1.2.19 Smoking intensity vs cardiometabolic diseases (CMDs) (MR-Base)

This section focused on the results obtained from the individual-level MR in the UKB in comparison to summary-level MR results in MR-Base. MR-Base is an online platform that uses summary-level data to perform MR analysis (version 1.4.3 8a77eb). The platform has many aspects concerning summary-level MR in addition to the pertaining package in R (*TwoSampleMR*). This includes choosing instruments (exposures) from different sources and the outcomes from different GWAS studies. After choosing the exposure(s) and outcome(s), certain characteristics (methods of analysis, LD check and harmonization) should be selected. Finally, the MR analysis will be executed using the "Run MR" button (http://app.mrbase.org/) [204,205].

The CperD genetic data were obtained from the MR-Base GWAS catalogue (GWAS and Sequencing Consortium of Alcohol and Nicotine use). The sample size of

this sample was 337,334 individuals of the European Ancestry. The analysis included twenty-two (22) SNPs for CperD. These SNPs were different from the ones included in the individual-level MR. The CMDs data were from the UKB (CHD: ukb-d-19_CHD, stroke: ukb-b-8714, HTN: ukb-b-14177 and DM: ukb-a-306).

For each variable, the results of MR analysis will include the MR estimates (MR-Egger) as well as the sensitivity analysis such as heterogeneity, single SNP analysis and leave-one-out analysis. MR Egger regression is a statistical method that corrects for any horizontal pleiotropy (significant associations between individual SNPs and outcomes) [206]. In other words, MR Egger enables a valid MR estimate from an invalid instrument. The sensitivity analysis will examine the validity of the SNPs included in the two-sample MR analysis. An overview of the sensitivity analysis terms is shown in Table 3.5 (Chapter 3, section: 3.5.7).

After choosing the exposure CperD and the outcomes (CMDs), the MR analysis was performed. The analysis revealed no evidence of a causal association between CperD and all CMD variables (P>0.05 for all associations). Genetic predisposition of CperD, based on 22 SNPs, was negatively associated with CHD, HTN and DM and positively associated with stroke (based on 21 SNPs). Table 4.11 and Figure 4.15 summarise MR Egger's findings as well single SNP analysis.

Table 4.11. Two-sample MR findings of CperD and CMDs (MR-Base)

Exposure	Outcome	Method	Number of SNPs	Beta	SE	P value
CperD	**CHD**	MR Egger	22	-0.002 (OR=0.99)	0.003	0.5632
CperD	**Stroke**	MR Egger	21	0.001 (OR=1.001)	0.002	0.7634
CperD	**HTN**	MR Egger	22	-0.0003 (OR=0.99)	0.012	0.9797
CperD	**DM**	MR Egger	22	-0.0004 (OR=0.99)	0.004	0.9254

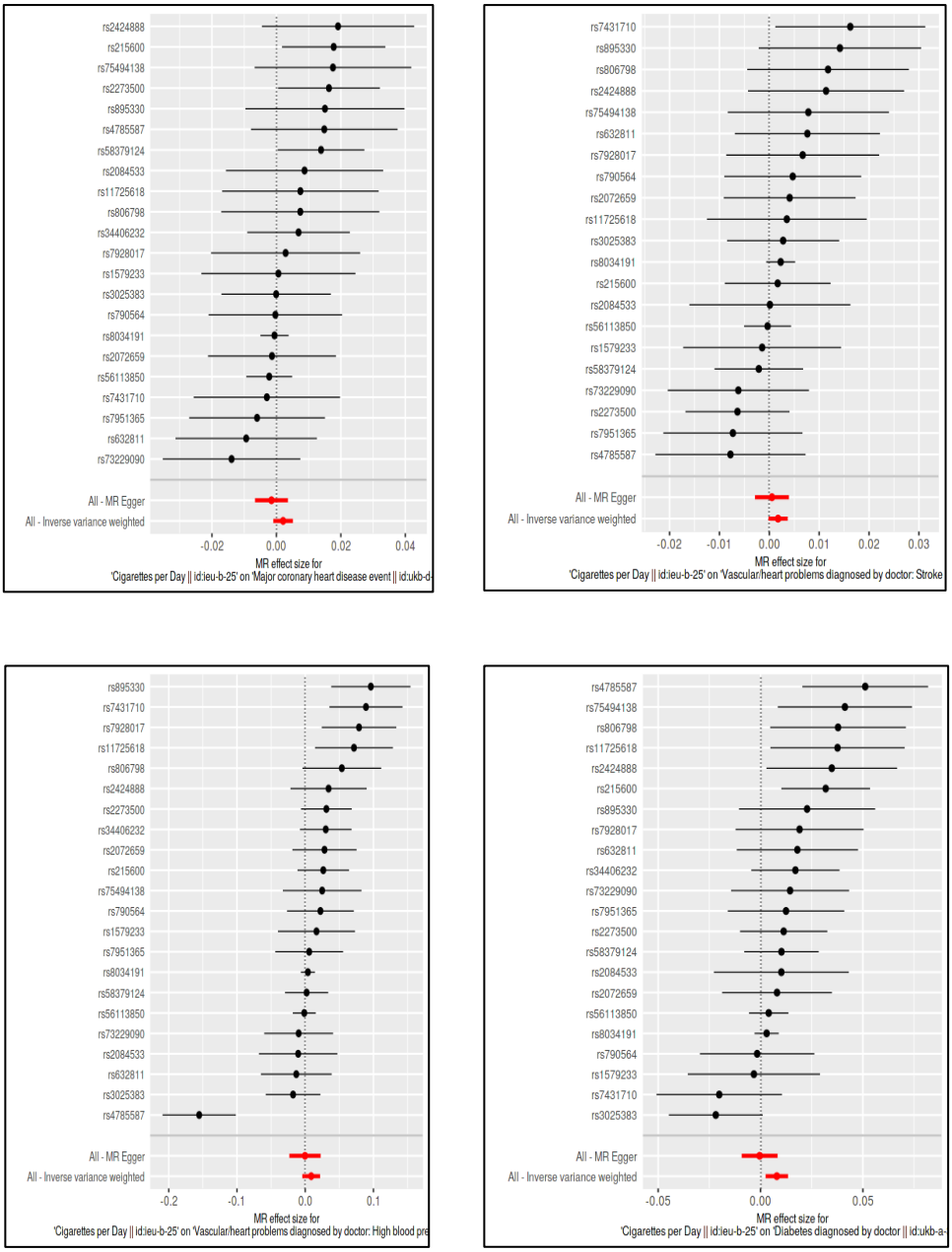


Figure 4.15. MR Egger and single SNP findings for CperD and CMDs

..1.2.20 Sensitivity analysis (MR-Base)

The MR-Base platform includes sensitivity analysis by default. The analysis included heterogeneity, single SNP analysis and leave-one-out analysis. The heterogeneity examines whether the SNPs exert their effect on the exposure and outcome concordantly or not. High heterogeneity might indicate the pleiotropic effects of the SNPs. The results of heterogeneity analysis for all CMD variables are shown in Table

4.12. Significant heterogeneity was observed for the smoking intensity (CperD) SNPs and HTN and DM risks ($P=3.34x10^{-8}$, P=0.03, respectively). In contrast, the CperD SNPs seem to be homogeneous for CHD and stroke (P=0.417, P=493, respectively).

Table 4.12: Heterogeneity findings for CperD and CMDs

Exposure	Outcome	Method	P
CperD	**CHD**	MR Egger	0.4177
CperD	**Stroke**	MR Egger	0.4937
CperD	**HTN**	MR Egger	$3.337e^{-8}$
CperD	**DM**	MR Egger	0.03093

The next sensitivity analysis to explore is a single SNP analysis. As shown in Figure 4.15, the CperD SNPs in general have a positive effect on the CMDs. The following SNPs were positively and statistically associated with the risk of CHD (rs215600, rs2273500, rs58379124), stroke (rs7431710), HTN (rs895330, rs7431710, rs7928017 and rs11725618) and DM (rs4785587, rs75494138, rs806798, rs11725618, rs2424888 and rs215600). On the contrary, only one SNP (rs4785587) was negatively and statistically significantly associated with the risk of HTN. The rest of the associations were statistically non-significant (Figure 4.15).

Finally, one of the sensitivity analyses provided by MR-Base is a leave-one-out analysis. It examines if only one SNP that drives the major effect on the outcome. By applying this analysis, each time one SNP is removed then the overall MR effect is plotted. Each point represents the MR estimate if a particular SNP was removed. The estimates of MR seem to be consistent in terms of their effect on the CMDs, however, this analysis used IVW not MR Egger regression (Figure 4.16). The next section will

explore the two-sample MR using a meta-analysis of GWAS studies for CperD using exact SNPs used in the individual-level MR in the UKB.

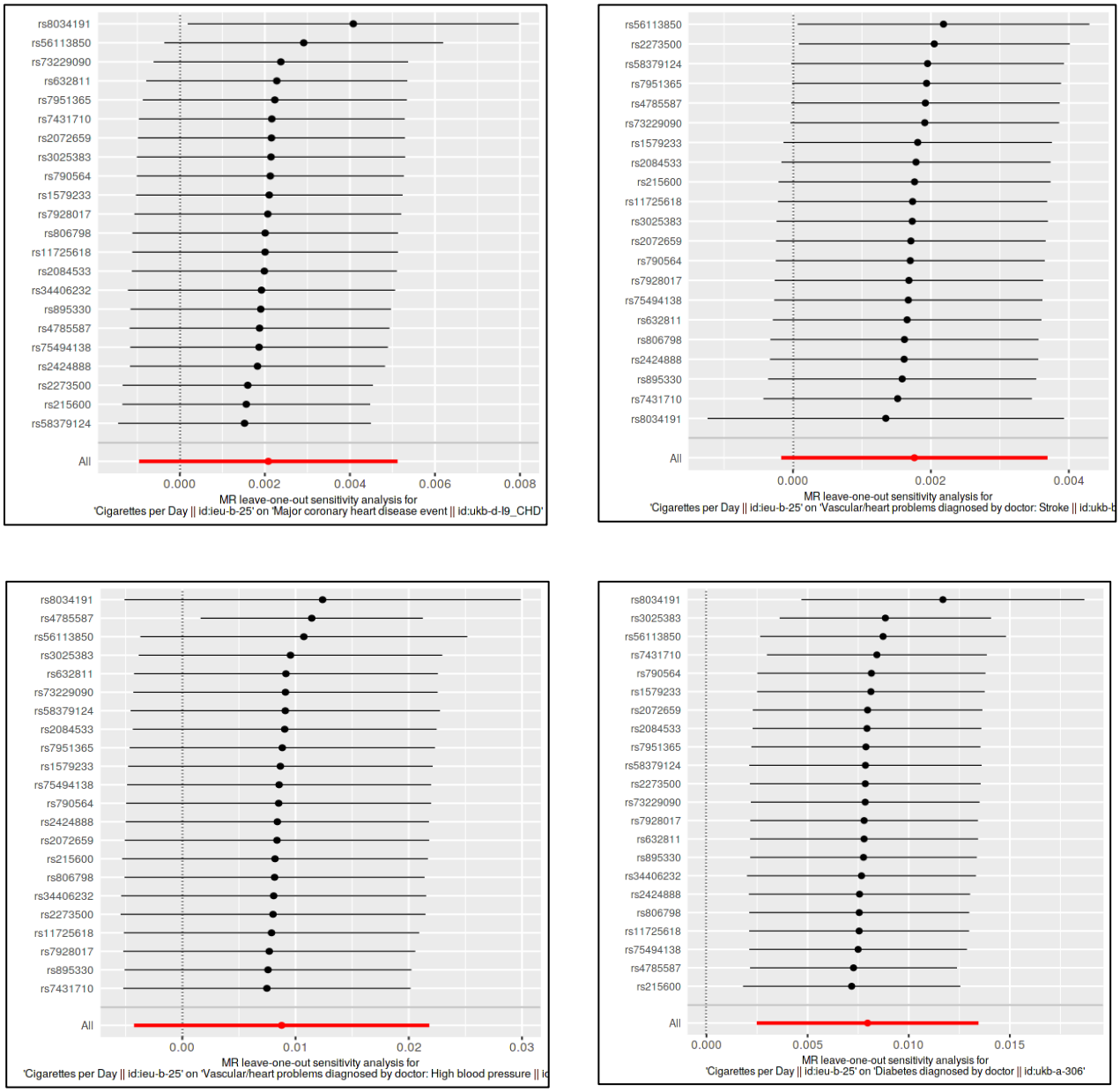


Figure 4.16. Leave-one-out analysis plots for CperD-SNPs and CMDs

..1.2.21 Two-sample MR – using R

This section examined the association between smoking intensity (CperD) SNPs and CMDs using two-sample MR. Instead of using MR-Base, this analysis was conducted in R. The CperD summary statistics (betas and SEs) were extracted from a meta-analysis of GWAS studies that included more than 1,2 million individuals (GSCAN)

[198]. The CMDs summary statistics were obtained from the UKB. The SNPs used in this analysis matched the ones used in the one-sample MR. The following SNPs were included in the analysis: rs1051730, rs7599488, rs215614, rs73229090, rs6474412, rs3025343, rs8034191, rs2229961, rs12910984, rs3733829, rs3865453, rs28399443, rs7260329, rs2273506. The details of the SNPs' summary statistics and the plots were provided in the supplementary materials (8.4, results, Tables 8.31 – 8.34, Figures 8.10).

After importing the data, the *MendelianRandomization* package was used in R to conduct a two-sample MR. An object was created that included beta estimates (B) as well as the standard errors (SE) from the regression of the SNPs on CperD and CMDs. After creating the object, MR-Egger regression was performed using *mr_egger* function. Finally, the plots for sensitivity analysis were performed using *mr_plot* function.

The analysis revealed that the genetic predisposition of CperD, based on 14 SNPs, was negatively and statistically non-significantly associated with CHD and stroke (OR=0.97, P=0.911, OR=0.90, P=0.867, respectively). The genetic predisposition of CperD was positively and statistically non-significantly associated with HTN (OR=1.08, P=0.634). Finally, the genetic predisposition of CperD was negative and statistically significantly associated with the risk of DM (OR=0.50, P=0.002). Table 4.13 summarises these findings.

Table 4.13. Two-sample MR findings of CperD and CMDs (R)

Exposure [GSCAN]	**Outcome [**UKB**]**	**Method**	**Number of SNPs**	**Beta**	**95% CI**	**P value**
CperD	**CHD**	MR Egger	14	-0.024 (OR=0.97)	-0.445, 0.397	0.911
CperD	**Stroke**	MR Egger	14	-0.102 (OR=90)	-1.299, 1.094	0.867
CperD	**HTN**	MR Egger	14	0.074 (OR=1.08)	-0.231, 0.378	0.634
CperD	**DM**	MR Egger	14	-0.693 (OR=0.50)	-1.143, -0.244	**0.002**

..1.2.22 Summary

The two-sample MR for the association between genetically estimated smoking intensity (CperD) and CMD revealed no evidence of a causal association between CperD (based on 22 SNPs) and CHD, HTN, stroke and DM. Similarly, the two-sample MR analysis using R revealed no evidence of a causal association between CperD and CMDs except DM. The genetically estimated smoking intensity was significantly associated with a lower risk of DM (OR=0.50, P=0.002). The next section will briefly examine the two-sample MR analysis of the association between the smoking status variable (ever vs never) and CMDs.

..1.2.23 Smoking status vs CMDs (two-sample MR)

This section briefly explored the two-sample MR for the smoking status variable using MR-Base. Never smokers were tested against ever smokers. The main idea is to compare the results obtained from one-sample MR to summary-level MR in MR-Base. MR-Egger was used to conceptualise the causal estimate between smoking status and CMDs.

Genetic predisposition to smoking (ever), based on 15 SNPs, was positively and non-significantly associated with CHD and stroke (OR=1.05, 1.09, P=0.644, 0.156, respectively). On the contrary, genetic predisposition to smoking among smokers

compared to non-smokers was negatively and non-significantly associated with HTN and DM (OR=0.388, 0.85, P=0.08, 0.265, respectively).

The findings obtained from one-sample MR of the UKB data were almost the opposite of the ones obtained from MR-Base for all CMD variables except HTN. Table 4.14 summarises MR Egger's findings. Sensitivity analyses were provided in the supplementary materials (8.4, results, Tables 8.36 – 8.37). The next section will compare the findings of MR results between the individual-level MR (UKB) and two-level MR.

Table 4.14. MR findings of smoking status and CMDs (UKB vs MR-Base)

Variable	**Smoking Status** (Ever) MR-Base		**Smoking Status** (Ever) UKB	
	MR Estimate	**P value**	**MR Estimate**	**P value**
CHD	OR=1.05	0.6445	OR=0.96	0.592
Stroke	OR=1.09	0.1566	OR=0.97	0.089
HTN	OR=0.38	0.0817	OR=0.44	**0.001**
DM	OR=0.85	0.265	OR=1.01	0.931

Individual-level (UKB) vs two-sample MR (MR-Base and R)

This section compares the findings obtained from individual-level MR in the UKB and those obtained from two-sample MR (MR-Base and manual analysis).

The findings of the MR analysis of the relationship between genetically estimated smoking intensity (CperD) and CMDs were statistically non-significant among all MR analyses except DM among manual analysis (OR=0.50, P=0.002). The risk of CHD and DM was lower for all MR analyses. The risk of stroke was lower among individual-level MR (UKB) and meta-analysis two-sample MR (manual) but higher in the MR-Base analysis. The risk of HTN was higher among individual-level MR (UKB) and meta-analysis two-sample MR (manual) but lower in the MR-Base analysis. Regarding the smoking status variable, there was no evidence of causal

association with CMD variables in all MR approaches except for HTN in one-sample MR (OR=0.44, P=0.001). In one-sample MR, ever-smokers have a lower risk for CHD, stroke and HTN and a higher risk for DM compared to never-smokers. While in MR-Base, ever smokers have a higher risk for CHD and stroke and a lower risk for HTN and DM.

To sum up the findings stated so far, genetically estimated smoking intensity (in smokers) is casually associated with decreased risk of DM in the two-sample MR (manual analysis). Additionally, the genetically based smoking status (ever) is causally associated with decreased risk of HTN among one-sample MR in the UKB population. Table 4.15$_{(a-b)}$ summarises these findings.

Table 4.15$_{(a-b)}$. Smoking behaviour vs CMDs (individual-level vs two-sample)

a) Genetic Predisposition of CperD vs Cardiometabolic Diseases (CMD)						
Outcomes	**MR Estimate** (One-sample UKB)	**P value**	**MR Estimate** (Two-sample: MR-Base)	**P value**	**MR Estimate** (Two-sample: R)	**P value**
CHD	↓	0.400	↓	0.5632	↓	0.911
Stroke	↓	0.550	↑	0.7634	↓	0.867
HTN	↑	0.299	↓	0.9797	↑	0.634
DM	↓	0.245	↓	0.9254	↓	**0.002**

b) Genetic Predisposition of smoking status vs Cardiometabolic Diseases (CMD)				
Outcomes	**MR Estimate** (One-sample UKB)	**P value**	**MR Estimate** (MR-Base)	**P value**
CHD	↓	0.592	↑	0.6445
Stroke	↓	0.089	↑	0.1566
HTN	↓	**0.001**	↓	0.0817
DM	↑	0.931	↓	0.265

Summary

The key findings from both observational and MR analyses are summarised in Tables 4.16 and 4.17.

Table 4.16. CperD vs CMDs: observational, one-sample and two-sample MR

Smoking intensity (CperD) vs Cardiometabolic Diseases (CMDs)								
Outcomes	**Observational Estimate**	**P value**	**MR Estimate** (One-sample)	**P value**	**MR Estimate** (Two-sample) MR-Base	**P value**	**MR Estimate** (Two-sample) R	**P value**
CHD	↑	**<0.001**	↓	0.400	↓	0.563	↓	0.911
Stroke	↑	**<0.001**	↓	0.550	↑	0.763	↓	0.867
HTN	↑	**<0.001**	↑	0.299	↓	0.979	↑	0.634
DM	↑	**<0.001**	↓	0.245	↓	0.925	↓	**0.002**

Table 4.17. Smoking status vs CMDs: observational, one and two sample MR

Smoking Status vs Cardiometabolic Diseases (CMDs)						
Outcomes	**Observational Estimate**	**P value**	**MR Estimate** (One sample) UKB	**P value**	**MR Estimate** (Two-sample) MR-Base	**P value**
CHD	↑	**<0.001**	↓	0.592	↑	0.6445
Stroke	↑	**<0.001**	↓	0.089	↑	0.1566
HTN	↓	**<0.001**	↓	**0.001**	↓	0.0817
DM	↑	**<0.001**	↑	0.931	↓	0.265

4.4. Discussion

Principal findings

The observational results obtained in this chapter support the findings of the conventional approaches in prior studies showing that smoking is a risk factor for coronary heart disease, stroke, and DM - but genetic evidence is less convincing.

This chapter explored the observational and genetic associations between smoking behaviour and CMDs. Observationally, the results found in this chapter support the positive associations between smoking behaviour (smoking status, smoking intensity (CperD) and smoking initiation (SI)) with a higher risk of CHD, stroke and DM and the conflicting results concerning HTN. For instance, current smokers have a lower risk of HTN compared to never smokers. Similarly, as an individual started to smoke early in life, the risk of HTN decreases. Conversely, as the smoking intensity increases, the risk of HTN increases. Genetically, there was limited evidence of the

causal association between smoking intensity (CperD) and CMDs except for DM. The smoking intensity was causally associated with decreased risk of DM (only in summary-level MR: OR=0.50, P=0.002). There was no causal association between smoking status (ever vs never) and CMDs (except HTN) using one-sample and summary-level MR. Smokers have a lower risk for HTN compared to non-smokers when using one-sample MR (OR=0.44, P=0.001). However, this causal association could not be established when using summary-level MR (MR-Base).

Interpretation

Observationally, smokers have a higher risk for CHD compared to non-smokers. Such impact of smoking on CHD was observed in a meta-analysis of 141 cohort studies [69] as well as a large review by Law et al [72]. Additionally, the analysis revealed that smokers who smoked more cigarettes per day (CperD) have a higher risk for CHD. These findings are consistent with Law et al review which included 19 studies [72], as well as Thun et al, found the same findings in a large prospective cohort study [73]. Finally, the findings of this thesis also revealed that as an individual started to smoke early, the risk of CHD was higher. These findings were similarly observed in a large prospective cohort study (ARIC) [74].

Similar to CHD findings, the observational analysis revealed that smokers have a higher risk for stroke compared to non-smokers. Additionally, individuals who smoked more cigarettes per day have a higher risk for stroke. Finally, As an individual starts smoking earlier in life, their risk of stroke increases. These findings were observed in two prospective cohort studies by Kurth et al among male and female physicians [75,76]. The findings were similar to two case-control studies by Bhat et al and Fogelholm et al [77,78].

The findings of the impact of smoking on DM were in correspondence to independent CHD and stroke results. The analysis revealed that smokers have a higher risk for DM compared to non-smokers. Additionally, individuals who smoked more cigarettes per day have a higher risk for DM. Finally, as an individual smokes early in life, the higher the risk of DM, however, the association was not significant. Such findings were found in a meta-analysis that included 25 articles by Willi et al [104] and in two prospective cohort studies by Lyssenko et al [106] and Manson et al [107].

The impact of smoking on HTN was not consistent. Smokers have a lower risk for HTN compared to non-smokers. A lower risk of HTN was also found as an individual started to smoke earlier. However, the smoking intensity (CperD) has a higher risk for HTN. These findings matched the conflicting results in the literature for the relationship between smoking and HTN [84,85,87,90–92]. Conflict findings of the association between smoking and HTN were observed in a systematic review by Leone A et al [77-80]. In a cross-sectional study by Liu and Byrd, smokers have better control of their blood pressure compared to non-smokers [88]. Similarly, Li G et al study revealed that smokers have lower blood pressure compared to non-smokers [81]. On the other hand, McNagny SE et al found that smokers have higher blood pressure compared to non-smokers [91]. Similar findings were also observed in a cross-sectional study by Al-Safi SA et al [92] and a prospective cohort study by Ruben et al [90].

Genetically, the results were not consistent with the observational findings nor among different MR approaches. All causal associations were statistically non-significant except for summary-level MR CperD vs. DM, and smoking status vs. HTN in the UKB sample (one-sample MR). Limited evidence of the causal associations between smoking and CMDs was found in Åsvold BO et al and Linneberg et al studies [39,207]. The negative causal association between smoking and HTN matches the

observational findings, however, opposes the positive causal association obtained by Larsson et al [179]. This causal association might be because of the violation of the instrumental variable that has a significant association with HTN. The causal association between smoking and DM found in this thesis opposed the results found in the observational studies as well as the genetic studies such as Larsson and Yuan [208].

Implications and future research

Smoking is a well-known risk factor for CMDs. Observationally, smokers have a higher risk for CHD, stroke and DM and a lower risk for HTN. Smoking intensity (CperD) has a higher risk for all CMD variables. Regarding smoking initiation, the risk of CHD, stroke and DM was higher as an individual started to smoke early in life, but the risk was lower for HTN. However, genetically, smoking is significantly associated with a lower risk of HTN and DM. Causal associations between smoking and CHD and stroke cannot be established.

These findings point toward that smoking is still a public health problem. To minimise the risk of CMDs, more strategies for smoking should be considered. The smoking intensity and age at which an individual starts to smoke play a significant role in increasing the risk of CMDs. Health education and health policy should be implemented as early as possible to reduce the risk of CMDs attributed to smoking. The inverse association between smoking and HTN/DM needs further exploration.

The causal associations between smoking and CMDs need more robust instrumental variables as well as a larger sample size. The robustness of the causal analysis can be enhanced with more SNPs as well as with a larger sample size. Additionally, the attainability of more SNPs that explained a decent amount of the variability in smoking behaviour would improve the causal inference. Finally, the

availability of SNPs that are exclusively associated with smoking would enhance the causal estimate by minimising known and unknown pleiotropic effects.

Strengths

This analysis has several strengths. First, it uses a large sample size of the UKB which provides high statistical power and precise estimation. Second, it also included many covariates e.g., deprivation to minimise the risk of confounding variables. Third, the inclusion of more than one variable to better picture smoking behaviour (smoking status, smoking intensity, and smoking initiation). Fourth, genetically, the paper uses more than one approach to examine the causal association between smoking and CMDs (one-sample MR and two-sample MR). Fifth, the genetic score was used instead of a single SNP analysis to improve the power of the estimation. Sixth, including only European ancestry in the UKB, would minimise the risk of population stratification. Finally, smoking behaviour was examined using smoking intensity and smoking status variables for a better conceptualisation of smoking against CMDs.

Limitations

The failure of reaching the causal statistical significance of the relationship between smoking and CMDs might be attributable to the small sample size of the CMD cases. Additionally, the protective effects could be due to bias where CMDs cases already died so the only people who attend assessment are very healthy. Furthermore, the number of SNPs that proxied smoking is relatively low (15 SNPs) which might jeopardise the robustness of the genetic score hence the analysis. However, the findings were not different when using different samples or summary-level data (R or MR-Base). Moreover, the precision of the causal estimate might be low due to low smoking variability that can be explained by the SNPs. To account for such limitations, multiple SNPs were utilised to build a robust genetic score proxying smoking. Finally, there was

a violation of the instrumental variable, however, the analysis uses MR-Egger to overcome such pleiotropic effect that might arise from such a violation.

4.5. Conclusion

In conclusion, this analysis found observational support for the association between increased smoking and CMDs. Smokers have a higher risk for CHD, stroke, and DM but a lower risk for HTN compared to non-smokers. The smoking intensity and smoking initiation have a higher risk for all CMD variables (except HTN with SI). The analysis also found a causal association between smoking and decreased risk of HTN and DM. No causal associations were found between smoking and CHD or stroke. A detailed exploration of these findings will be provided in the discussion chapter but fundamentally these analyses suggest that the association between smoking behaviour and poorer health may be via unmeasured variables rather than directly causal. The potentially protective effects are unlikely to be mechanistic and therefore more likely to reflect instrumentation or attrition bias e.g., where particularly unhealthy participants did not attend the assessment or have functionally declined substantially.

5. Chapter Five: Observational and Mendelian randomization-based causal estimates of the association between smoking behaviour and lipid biomarkers

5.1. Introduction

Overview

This chapter examines the association between smoking behaviour and lipid biomarkers (cholesterol, LDL, TG and HDL). The main goal of this chapter is to estimate if there are significant associations between these outcomes and smoking in the UKB observationally and genetically (using MR). The chapter starts with a brief review of the associations between smoking and lipid biomarkers, followed by a recap of the methods used in the analysis. Next, the chapter will proceed to the results section which includes the analysis of the association between smoking and lipid biomarkers. The analysis covers the observational associations followed by one-sample MR in the UKB sample and finally two-sample MR using the MR-Base platform as well as in R software. Finally, the chapter will discuss the finding in the discussion section followed by the overall chapter conclusion. Figure 5.1 describes the chapter structure.

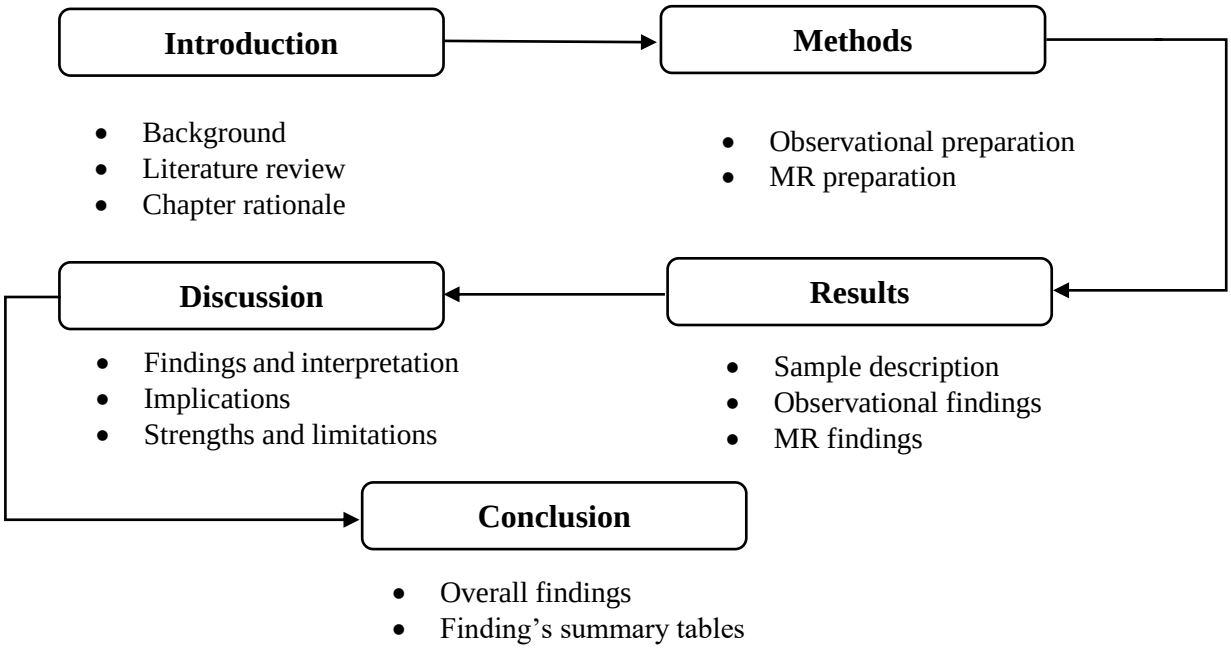


Figure 5.1. Chapter scheme II

Background

Lipid biomarkers have a significant and reliable underpinning of CMDs and physical health [209]. LDL is referred to as “bad” cholesterol while HDL is considered “good” cholesterol. High levels of cholesterol, LDL, TG, and low HDL are well-known risk factors for poorer health [210]. Lipids and lipoprotein particles play a critical role in atherosclerosis, the pathophysiology of cardiovascular diseases. They also have an impact on inflammatory processes, vascular and cardiac cell function, and the health of the heart and blood vessels [211].

Cigarette smoking is believed to be linked to higher levels of cholesterol, LDL, and TG as well as lower levels of HDL [40]. Cigarette smoke contains nicotine, which enhances the liver's release of lipoproteins, free fatty acids, and triglycerides into the bloodstream. This mechanism is strengthened by nicotine's stimulatory effects on catecholamine (epinephrine and norepinephrine) release, which results in sympathetic activation and accelerated lipolysis (the breakdown of fat) [113]. Additionally, High cholesterol level causes fatty deposits to form in blood vessels, making it difficult for adequate blood to circulate through the arteries.

The relationships between smoking and lipid biomarkers were explored widely in the literature. This chapter investigated the role of smoking on lipids because these biomarkers are well-established risk factors for CMDs and health generally and are continuous phenotypes rather than 0/1 cases vs. controls. A brief review of the associations between smoking and lipid biomarkers will be provided in the next sections. A detailed literature review of the associations between smoking and lipids was explored in chapter two (section: 2.3.6).

A systematic review of 54 articles on the effect of smoking on lipid biomarkers concluded that the higher number of cigarettes individual smokes, the higher levels of

TG, LDL, and total cholesterol, and the lower HDL [119]. Similarly, a prospective cohort study by Gossett et al found that current smokers have higher levels of cholesterol, LDL and TG and lower HDL levels compared to non-smokers [110]. These findings were also found in a cross-sectional study by Zhang et al [118], a screening by Muscat et al [115] and a survey by Willett et al [40]. Genetically, an MR study was conducted to examine the causal association between rs1051730 as a proxy for smoking behaviour and cardiovascular risk factor. The study concluded that HDL level increases with each additional rs1051730 T allele. Additionally, rs1051730 among current smokers was associated with lower TG concentration [39]. Finally, a summary-level MR was performed to examine the causal association between genetic liability for smoking and CVDs risk factors [180]. The study found a causal association between smoking and hyperlipidaemia.

The gap in the literature

These findings of the adverse effects of smoking on lipid biomarkers were not consistent. Some papers reported no association between smoking and lipids observationally as well as genetically. A cross-sectional study of 15276 participants conducted by R. Jain and A. Ducatman [117] reported no difference in the cholesterol and LDL levels among current smokers compared to non-smokers. Similar findings were also found in a cross-sectional study among 401 participants conducted by Saengdith. P [116]. Additionally, Moradinazar et al [120] included 7586 participants in a cross-sectional study that revealed that current smokers have no significant association with abnormal cholesterol nor LDL.

This uncertainty was also found genetically. A summary-level MR was conducted by Levin et al [180] to explore the causal association between smoking and CVD risk factors such as lipid biomarkers. The researchers reported no causal

association between smoking and lipid biomarkers. Furthermore, a Mendelian randomization study revealed no causal association between rs1051730 and total cholesterol among current smokers [39].

Chapter rationale

The observational approaches to examining the associations between smoking and lipid biomarkers are usually prone to uncertainty attributed to confounding effects and reverse causation. This uncertainty makes it difficult to infer a causal relationship between smoking and lipid biomarkers [25,203]. Additionally, the RCTs carry the burden of being expensive and time-consuming as well as the ethical considerations that limit such an approach [27,28]. To overcome such limitations, the MR approach was used to infer the causal association between smoking and lipid biomarkers. Specifically, genetic instrumentation of lifetime smoking risk is leveraged to estimate causal associations between smoking and lipids. This chapter will examine such associations observationally as well as genetically. A detailed review of the rationale of this thesis was provided in chapter one (section: 1.3 – 1.5) and chapter four (section: 4.1; chapter rationale). The next sections will briefly recap the methods used in the analysis.

5.2. Methods

Observational analysis

To examine the relationship between smoking behaviour and lipid biomarkers observationally, regression analyses were performed. As lipid biomarkers are continuous variables, linear regression analysis will be used to examine such a relationship. Frequency tables, crosstabulations as well as visualisation of the variables will be presented in the descriptive statistics sections and the supplementary materials (8.5, results, Tables 8.39 – 8.40, and Figure 8.11).

MR analysis

Overview

The Mendelian randomization approach will be used to examine the causal associations between smoking behaviour and lipid biomarkers. The analysis will include: the IV assumption results, one-sample (using 2SLS) and two-sample MR (using MR-Egger). The genetic quality control results, building genetic score and 1st and 3rd assumptions were performed in chapter four (section: 1.2). Table 5.1 summarises the MR analysis approach. The analysis was done for both CperD and smoking status, however, the details of the analysis will be shown only for CperD. Only final MR results will be shown for the smoking status variable.

Table 5.1. MR approach scheme

Genetic preparation		
Smoking SNPs quality control		Plink output [SNPs included]
Genetic score		From included (valid) SNPs
IV assumptions	**1st**	Smoking associated with genetic score
	2nd	Cholesterol associated with Genetic score LDL associated with Genetic score TG associated with Genetic score HDL associated with Genetic score
	3rd	Genetic score associated with covariates
MR		
Smoking + SNPs + Lipid biomarkers		2SLS [one-sample MR: UKB]
UKB: SNPs- Lipid biomarkers vs MR-Base SNPs-(Smoking**)**		MR-Egger [two-sample MR: UKB vs MR-Base]

IV assumptions results (CperD)

..1.2.24 First and third IV assumptions

The first and third IV assumptions were discussed in chapter four. The results of these associations are summarised in Table 5.2.

Table 5.2. Summary of IV assumptions (1st and 3rd)

Variables	Genetic Score		
	Estimates	*95% CI*	*p*
CperD	0.223	0.2 – 0.3	$1.25x10^{-10}$
Age	0.005	-0.00 – 0.00	0.401
Degree [No]	-0.02	-0.07 – 0.03	0.529
Sex [Male]	-0.01	-0.05 – 0.03	0.620
Townsend	-0.001	-0.01 – 0.00	0.791
BMI	-0.01	-0.01 – -0.00	**0.010**
PC1	-0.21	-0.22 – -0.21	**<0.001**
PC2	-0.01	-0.02 – 0.00	0.271
PC3	0.61	0.60 – 0.62	**<0.001**
PC4	-0.09	-0.10 – -0.07	**<0.001**
PC5	0.12	0.10 – 0.14	**<0.001**
PC6	0.46	0.44 – 0.48	**<0.001**
PC7	-0.09	-0.11 – -0.07	**<0.001**
PC8	0.26	0.24 – 0.28	**<0.001**
PC9	0.26	0.24 – 0.28	**<0.001**
PC10	0.04	0.02 – 0.06	**<0.001**

..1.2.25 Genetic score vs Lipid biomarkers (second IV assumption)

The genetic score was not associated with any of the lipid biomarkers variables. Table 5.3 shows the association findings.

Table 5.3. IV assumption 2 results (genetic score vs lipid biomarkers)

Lipid Biomarkers	Genetic Score		
	B	*95% CI*	*p*
Cholesterol	0.0003632	-0.01 – 0.01	0.942
LDL	0.001955	-0.01 – 0.01	0.613
Log (TG)	-0.003196	-0.008 – 0.001	0.156
HDL	- 0.0001	-0.003 – 0.003	0.969

The genetic score was statistically significantly associated with CperD, but not associated with any of the outcomes (lipid biomarkers). Additionally, the associations between the genetic score and the covariates were non-significant with age, degree, sex, and deprivation score. However, the genetic score was significantly associated with BMI and most PCs. This genetic score is valid but not the most ideal IV to proxy smoking intensity.

IV assumptions results (smoking status)

This section briefly examines the IV assumptions for smoking status genetic score (never vs ever). The genetic score is significantly associated with smoking status (ever). Additionally, the genetic score was not associated with any of the lipid biomarkers. Finally, the genetic score was not associated with age, sex, or deprivation score. However, the score was not independent of education attainment and most PCs.

The genetic scores for smoking variables were valid but not the most ideal proxies for smoking behaviour. The next section of the chapter will explore the results of the association between smoking and lipid biomarkers.

5.3. Results

Sample characteristics

The sample characteristics of the smoking variables and sample covariates were discussed in chapter four. Table 5.4 shows the summary statistics for the variables to be used in this chapter.

Table 5.4. Sample characteristics-observational (n=469,598)

Variable	Level	Count (%)
Smoking status	Current	52431 (10.56%)
	Previous	172216 (34.68%)
	Never	271951 (54.76%)
Sex	Male	226177 (45.5%)
	Female	270421 (54.5%)
Ethnicity	White British	439085 (88.42%)
	Other ethnicities	57513 (11.58%)
Degree (college/university)	No Degree	336334 (67.73%)
	Degree	160264 (32.27%)
Variable	**Mean (SD)**	
Cigarette Smoked per Day (CperD)	15.5 (±8.39)	
Smoking Initiation (SI): Age started smoking	17.85 (±5.8)	
Age	56.53 (±8.09)	
Body Mass Index (BMI)	27.42 (±4.79)	
Deprivation Level (Townsend score)	-1.31 (±3.08)	
Cholesterol	5.69 (±1.14)	
Low-Density Lipoproteins (LDL)	3.56 (±0.87)	
Triglycerides (TG)	***Median** = 1.48 (IQR=1.11) (Not normally distributed)	
High-Density Lipoprotein (HDL)	1.45 (±0.38)	

Descriptive statistics

The descriptive statistics for smoking variables (smoking status, smoking intensity (CperD) and smoking initiation (SI) were presented in chapter four (descriptive statistics). This section will explore the lipid biomarkers descriptively.

Lipid biomarkers summary statistics (dependent variables)

The lipid biomarkers were total cholesterol, LDL, TG, and HDL. These variables are numeric and continuous. Lipid biomarkers were measured in serum from the serum separation tube sample. Multiple immunoassays and clinical chemistry analyses were used to measure biochemistry markers [190]. The unit of measurement is mmol/L. The average cholesterol level for the UKB participants was 5.69 (±1.14) which is classified as borderline high [213]. Similarly, a borderline high level of LDL 3.56 (±0.87) was observed among the UKB participants. The median level for TG (not normally

distributed) was 1.48 (IQR=1.11), which is a desirable level. Finally, the mean HDL level for the UKB participants was 1.45 (±0.38), which is close to the best reading for HDL [213]. Table 5.5 summarises these findings.

Table 5.5. Lipid biomarkers summary statistics

Variable	Mean (SD)
Cholesterol	5.69 (±1.14)
Low-Density Lipoproteins (LDL)	3.56 (±0.87)
Triglycerides (TG)	***Median** = 1.48 (IQR=1.11) (Not normally distributed)
High-Density Lipoprotein (HDL)	1.45 (±0.38)

Observational Analysis

This section examined the observational associations between smoking behaviour and lipid biomarkers. It included linear regression analysis of the associations between smoking status, smoking intensity (CperD) and smoking initiation (SI) and lipid biomarkers. The analysis used a standardised beta coefficient (β) as the effect measure of the relationship between the numeric outcomes (cholesterol, LDL, TG and HDL) and other variables. Unstandardised beta (B) was also provided for unadjusted associations and effects visualisation. The statistical significance level was set at P = 0.05. All relationships included unadjusted (smoking variable only vs outcome) as well as adjusted (smoking variable + covariates vs outcome). Simple linear regression was used for unadjusted analyses while multiple linear regression was used for adjusted associations. The adjustment will help to reduce the risk of confounding variables.

Smoking status and lipid biomarkers

This section examined the associations between the smoking status variable and lipid variables. Smoking status was analysed as a three-categories variable (current, previous, and never) as well as a binary variable (ever vs never: provided in the supplementary materials: 8.5, results: Table 5.6$_{(a-d)}$). Each section presented unadjusted

associations followed by adjusted results. The reference levels for all categorical variables were established in chapter four (Table: 4.6).

..1.2.26 Smoking status vs total cholesterol

When examining smoking status against cholesterol (unadjusted), current smokers, as well as previous smokers, had statistically significantly lower cholesterol levels compared to never-smokers (B: -0.04, 95% CI: -0.05 – -0.03, P<0.001, B: -0.05, 95% CI: -0.06 – -0.04, P<0.001, respectively). After adjusting for the covariates (multiple linear regression), previous smokers were having significantly lower cholesterol levels compared to never-smokers (β: -0.02, 95% CI: -0.03 – -0.01, P<0.001). Conversely, current smokers became positively and significantly associated with cholesterol levels (β: 0.05, 95% CI: 0.04 – 0.06, P<0.001). Table 5.6 and Figure 5.2 summarise these findings.

Table 5.7. Linear regression analysis of smoking status vs total cholesterol

Cholesterol		**Unadjusted**			**Adjusted**			
		B	**CI**	**P**	**B**	**β**	**CI**	**P**
Smoking Status	Current	-0.04	-0.05 – -0.03	**<0.001**	0.06	0.05	0.04 – 0.06	**<0.001**
	Previous	-0.05	-0.06 – -0.04	**<0.001**	-0.02	-0.02	-0.02 – -0.01	**<0.001**
Sex [Male**]**					-0.39	-0.34	-0.40 – -0.38	**<0.001**
Education [Have Degree**]**					0.02	0.02	0.01 – 0.03	**<0.001**
Ethnicity [White British**]**					0.14	0.12	0.13 – 0.15	**<0.001**
Age					0.01	0.06	0.01 – 0.01	**<0.001**
Townsend					-0.02	-0.05	-0.02 – -0.02	**<0.001**
BMI					-0.01	-0.03	-0.01 – -0.01	**<0.001**

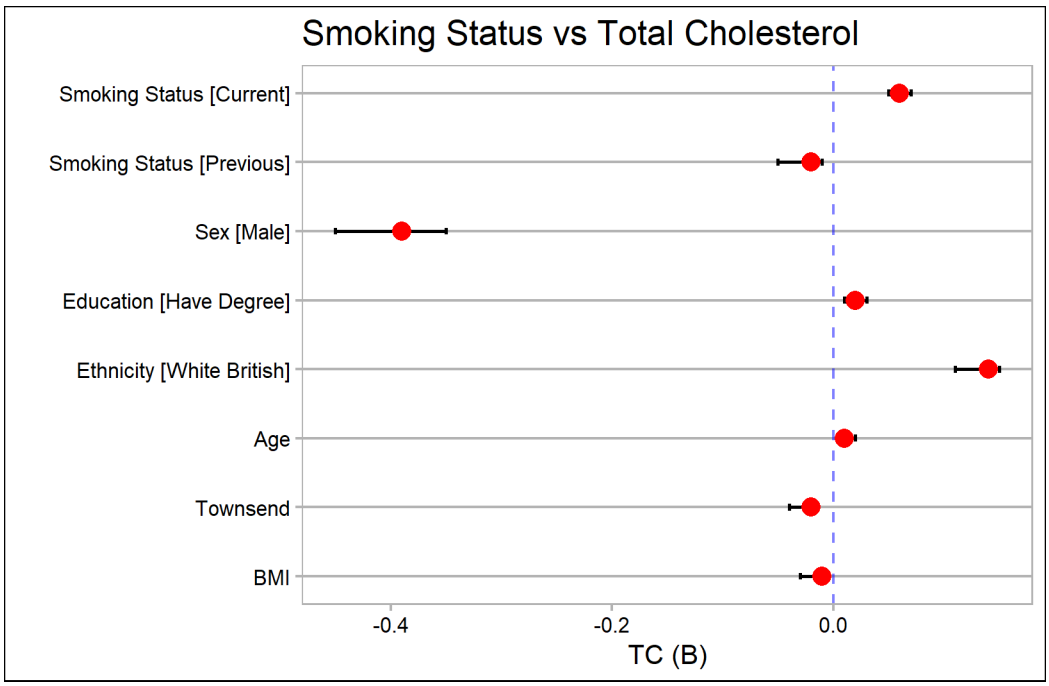


Figure 5.2. Visualisation of the adjusted associations: smoking vs cholesterol

..1.2.27 Smoking status vs low-density lipoprotein (LDL)

The unadjusted relationship between smoking status and LDL revealed that current smokers have a higher level of LDL level compared to never smokers, however, this result was statistically non-significant. Conversely, previous smokers have a statistically significant lower level of LDL compared to never-smokers. (B: 0.001, 95% CI: --0.01 – 0.01, P=0.817, B: -0.05, 95% CI: -0.05 – -0.04, P<0.001, respectively). After adjusting for the covariates, current smokers have a significantly higher level of LDL compared to never smokers. Previous smokers have a significantly lower LDL level compared to never smokers ($\beta_{current}$: 0.06, 95% CI: 0.05 – 0.07, P<0.001, $\beta_{previous}$: -0.05, 95% CI: -0.06 – -0.05, P<0.001, respectively). Table 5.7 and Figure 5.3 summarise these findings.

Table 5.8. Linear regression analysis of smoking status vs LDL

LDL		Unadjusted			Adjusted			
		B	CI	P	B	β	CI	P
Smoking Status	Current	0.001	-0.01 – 0.01	0.817	0.05	0.06	0.05 – 0.07	**<0.001**
	Previous	-0.05	-0.05 – -0.04	**<0.001**	-0.05	-0.05	-0.06 – -0.05	**<0.001**
Sex [Male**]**					-0.15	-0.17	-0.16 – -0.15	**<0.001**
Education [Have Degree**]**					0.01	0.01	0.00 – 0.01	**0.005**
Ethnicity [White British**]**					0.09	0.10	0.08 – 0.10	**<0.001**
Age					0.00	0.04	0.00 – 0.00	**<0.001**
Townsend					-0.01	-0.05	-0.01 – -0.01	**<0.001**
BMI					0.01	0.03	0.00 – 0.01	**<0.001**

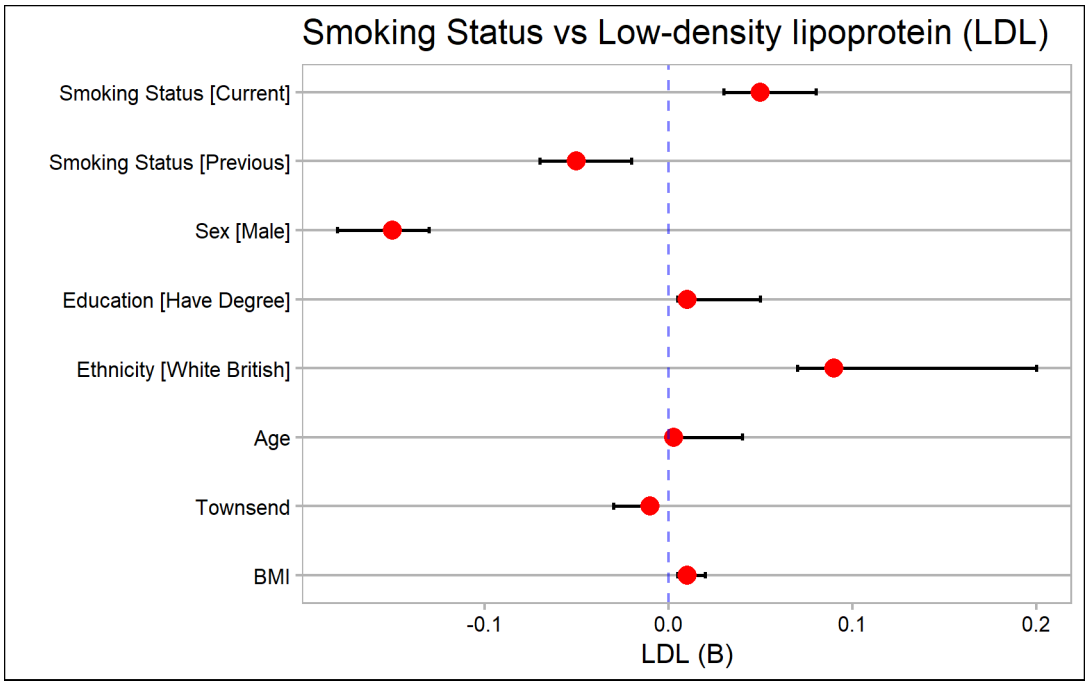


Figure 5.3. Visualisation of the adjusted associations: smoking status vs LDL

..1.2.28 Smoking status vs triglycerides (TG)

Unadjusted association between smoking and log TG showed that both current and previous smokers were significantly associated with a higher level of TG compared to never smokers (B: 0.14, 95% CI: 0.14 – 0.15, P<0.001, B: 0.07, 95% CI: 0.07 – 0.08, P<0.001, respectively). After adjusting for the covariates, both current and previous smokers hold positive and significant associations with TG levels compared to never

smokers ($\beta_{current}$: 0.09, 95% CI: 0.08 – 0.09, P<0.001, $\beta_{previous}$: 0.01, 95% CI: 0.01 – 0.01, P<0.001, respectively). Table 5.8 and Figure 5.4 show these findings.

Table 5.9. Linear regression analysis of smoking status vs TG

Log (TG)		Unadjusted			Adjusted			
		B	CI	P	B	β	CI	P
Smoking Status	Current	0.14	0.14 – 0.15	**<0.001**	0.13	0.09	0.08 – 0.09	**<0.001**
	Previous	0.07	0.07 – 0.08	**<0.001**	0.01	0.01	0.01 – 0.01	**<0.001**
Sex [Male]					0.19	0.13	0.13 – 0.13	**<0.001**
Education [Have Degree]					-0.03	-0.02	-0.02 – -0.02	**<0.001**
Ethnicity [White British]					0.04	0.03	0.02 – 0.03	**<0.001**
Age					0.01	0.03	0.03 – 0.03	**<0.001**
Townsend					-0.00	-0.00	-0.00 – -0.00	**0.048**
BMI					0.03	0.10	0.10 – 0.10	**<0.001**

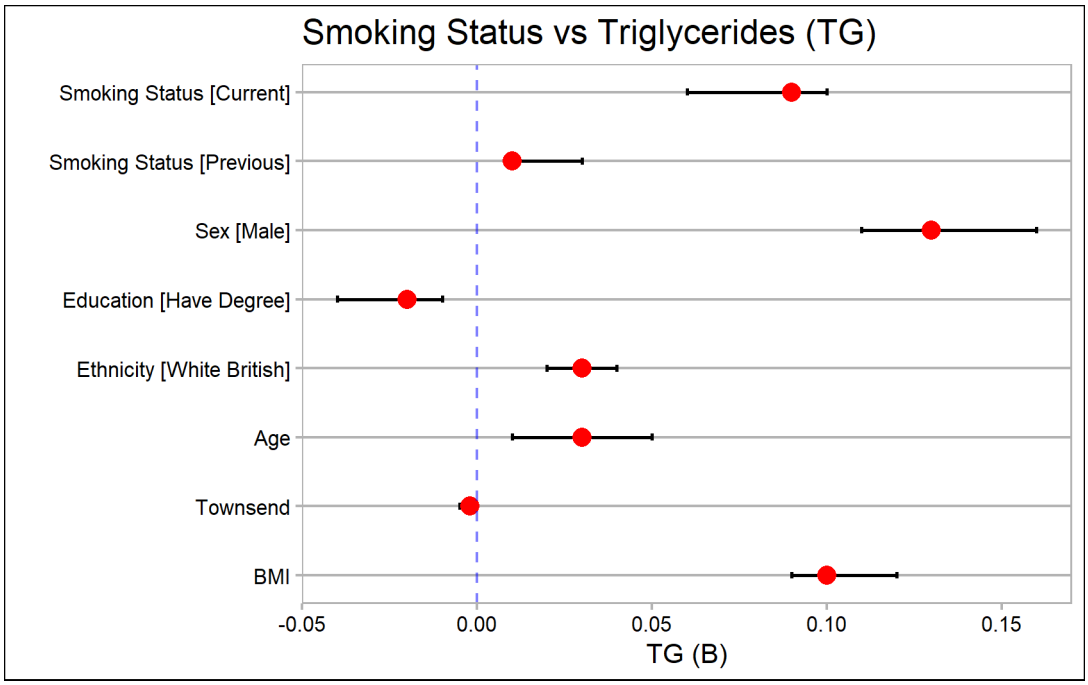


Figure 5.4. Visualisation of the adjusted associations: smoking status vs TG

..1.2.29 Smoking status vs high-density lipoprotein (HDL)

Unadjusted association between smoking and HDL showed that both current and previous smokers have significantly lower HDL levels compared to never-smokers (B: -0.10, 95% CI: -0.11 – -0.10, P<0.001, B: -0.02, 95% CI: -0.02 – -0.01, P<0.001, respectively). After adjusting for the covariates, current smokers remain negatively and

significantly associated with a lower level of HDL compared to never-smokers (β: -0.14, 95% CI: -0.15 – -0.14, P<0.001). However, previous smokers become positively associated with the HDL level compared to never smokers (β: 0.07, 95% CI: 0.07 – 0.08, P<0.001). Table 5.9 and Figure 5.5 summarise these findings.

Table 5.10. Linear regression analysis of smoking status vs HDL

HDL		Unadjusted			Adjusted			
		B	**CI**	**P**	**B**	**β**	**CI**	**P**
Smoking Status	Current	-0.10	-0.11 – -0.10	**<0.001**	-0.06	-0.14	-0.15 – -0.14	**<0.001**
	Previous	-0.02	-0.02 – -0.01	**<0.001**	0.03	0.07	0.07 – 0.08	**<0.001**
Sex [Male**]**					-0.30	-0.78	-0.78 – -0.77	**<0.001**
Education [Have Degree**]**					0.03	0.07	0.07 – 0.08	**<0.001**
Ethnicity [White British**]**					0.02	0.06	0.05 – 0.07	**<0.001**
Age					0.00	0.06	0.06 – 0.06	**<0.001**
Townsend					-0.00	-0.01	-0.02 – -0.01	**<0.001**
BMI					-0.03	-0.32	-0.32 – -0.32	**<0.001**

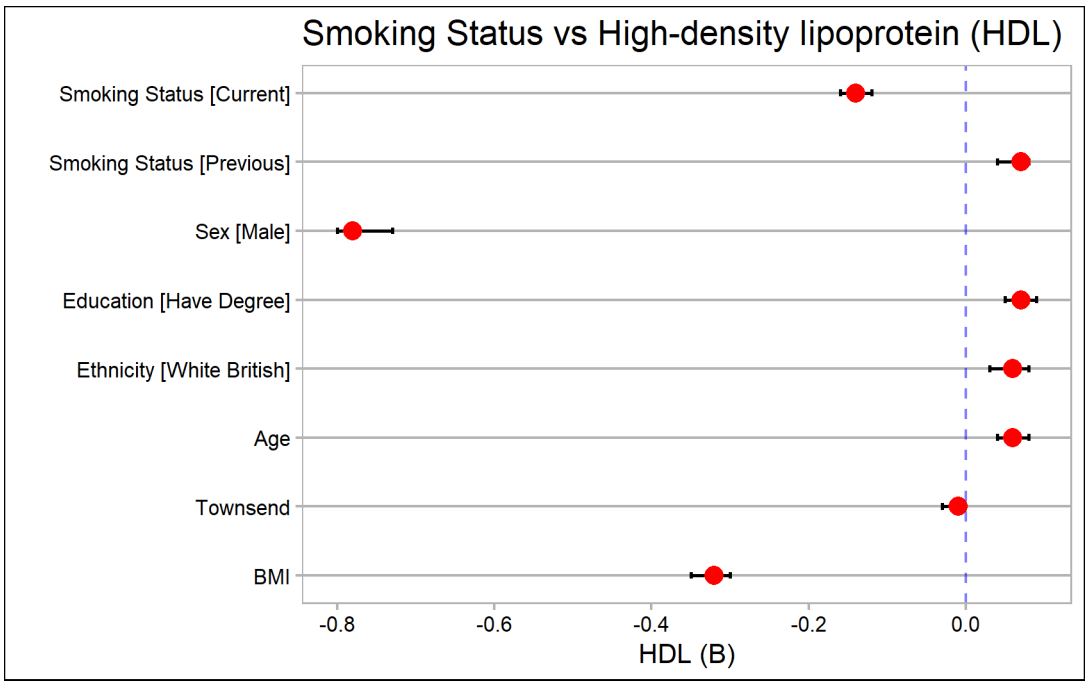


Figure 5.5. Visualisation of the adjusted associations: smoking status vs HDL

Smoking intensity (CperD) and lipid biomarkers

This section examined the associations between the smoking intensity (cigarettes smoked per day: CperD) variable and lipids variables. Each section presented the unadjusted associations followed by the adjusted results.

..1.2.30 Smoking intensity (CperD) vs total cholesterol

Unadjusted association between CperD and cholesterol revealed a negative and statistically significant association. An additional cigarette smoked per day was associated with a decreased level of cholesterol by 0.002 mmol/L (B= -0.002, 95% CI: -0.004 - -0.001, P=0.007). The effect of CperD on cholesterol remained significant after adjustment for the confounders. However, the association turned positive after the adjustment (β: 0.02, 95% CI: 0.01 – 0.03, P<0.001). These associations are summarised in Table 5.10 and visualised in Figure 5.6.

Table 5.11. Linear regression analysis of CperD vs cholesterol

Cholesterol	**Unadjusted**			**Adjusted**			
	B	**CI**	**P**	**B**	**β**	**CI**	**P**
CperD	-0.002	-0.004--0.001	**0.007**	0.003	0.02	0.01 – 0.03	**<0.001**
Sex [Male**]**				-0.37	-0.31	-0.33 – -0.29	**<0.001**
Education [Have Degree**]**				0.03	0.03	-0.00 – 0.05	0.065
Ethnicity [White British**]**				0.07	0.06	0.03 – 0.09	**<0.001**
Age				-0.002	-0.01	-0.02 – -0.00	**0.036**
Townsend				-0.02	-0.07	-0.08 – -0.06	**<0.001**
BMI				-0.0004	-0.00	-0.01 – 0.01	0.758

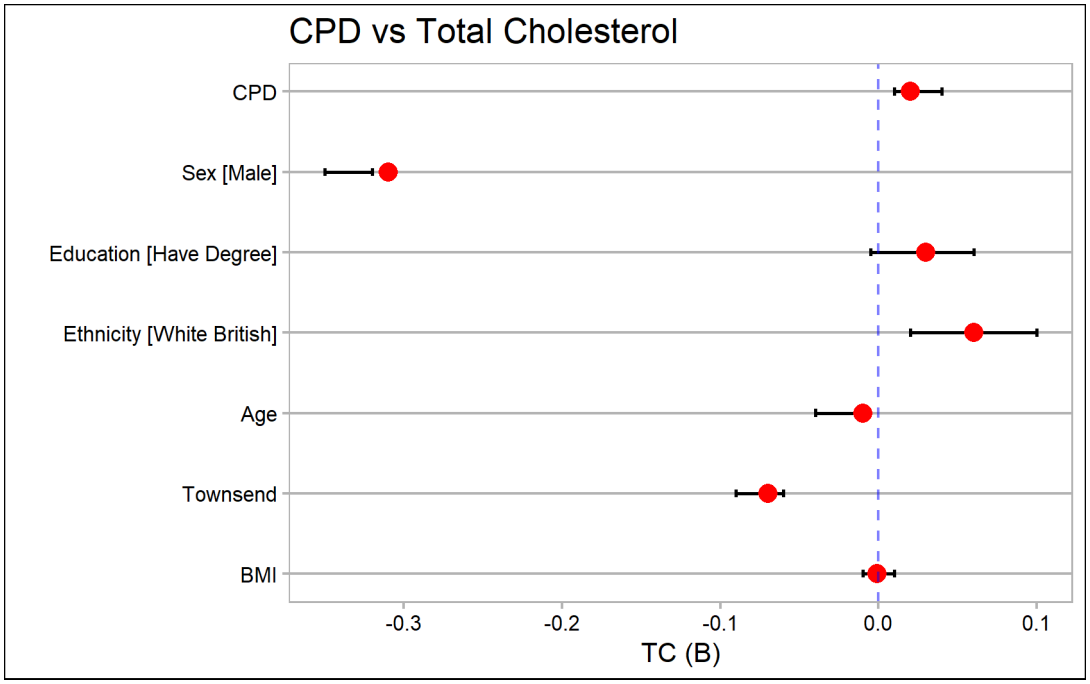


Figure 5.6. Visualisation of the adjusted associations: CperD vs cholesterol

..1.2.31 Smoking intensity (CperD) vs low-density lipoprotein (LDL)

Unadjusted association between CperD and LDL revealed a positive non-significant association. An additional cigarette smoked per day will increase the level of LDL by 0.001 mmol/L (B= 0.001, 95% CI: -0.001-0.001, P=0.280). The effect of CperD on LDL remained positive but turned statistically significant after adjustment for the confounders (β: 0.03, 95% CI: 0.02 – 0.04, P<0.001). These associations are summarised in Table 5.11 and visualised in Figure 5.7.

Table 5.12. Linear regression analysis of CperD vs LDL

LDL	**Unadjusted**			**Adjusted**			
	B	**CI**	**P**	**B**	**β**	**CI**	**P**
CperD	0.001	-0.001-0.001	0.280	0.001	0.03	0.02 – 0.04	**<0.001**
Sex [Male**]**				-0.19	-0.21	-0.23 – -0.18	**<0.001**
Education [Have Degree**]**				0.005	0.01	-0.02 – 0.03	0.709
Ethnicity [White British**]**				0.05	0.05	0.02 – 0.09	**0.001**
Age				-0.003	-0.03	-0.04 – -0.02	**<0.001**
Townsend				-0.02	-0.06	-0.07 – -0.05	**<0.001**
BMI				0.01	0.06	0.05 – 0.07	**<0.001**

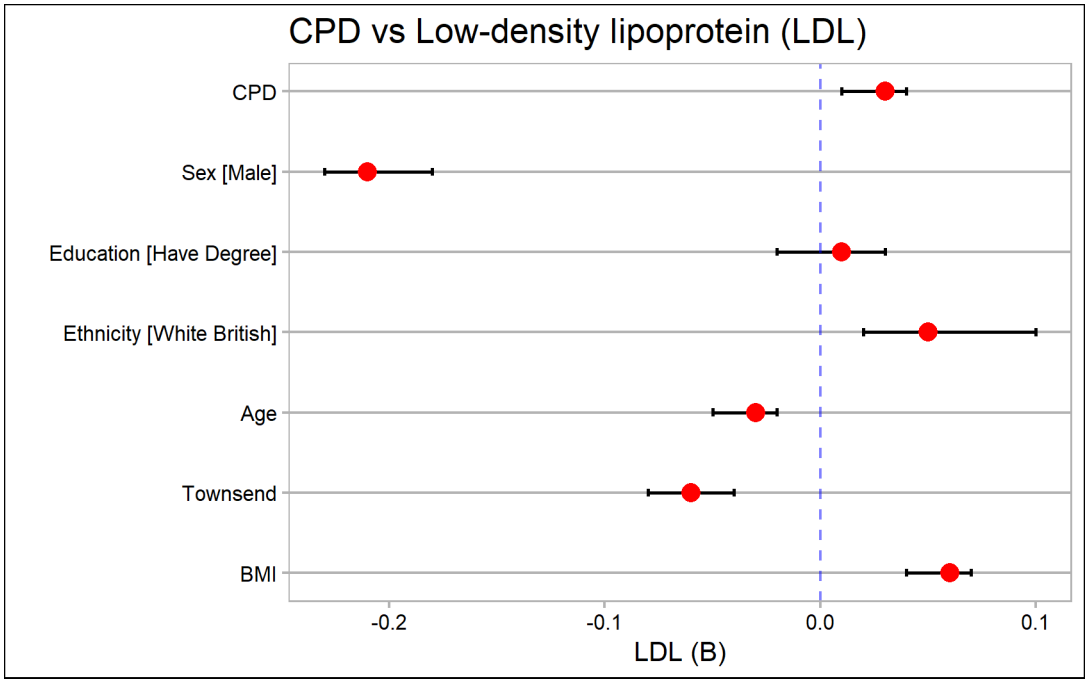


Figure 5.7. Visualisation of the adjusted associations: CperD vs LDL

..1.2.32 Smoking intensity (CperD) vs triglycerides TG

Unadjusted association between CperD and log TG revealed a positive and significant association. An additional cigarette smoked per day will increase the level of TG by 1% (0.01 mmol/L) (B= 0.01, 95% CI: 0.01 – 0.02, P<0.001). The effect of CperD on log TG remained positive and significant after adjustment for the confounders (β: 0.02, 95% CI: 0.01 – 0.02, P<0.001). These associations are summarised in Table 5.12 and visualised in Figure 5.8.

Table 5.13. Linear regression analysis of CperD vs TG

Log TG	**Unadjusted**			**Adjusted**			
	B	**CI**	**P**	**B**	**β**	**CI**	**P**
CperD	0.01	0.01 – 0.02	**<0.001**	0.01	0.02	0.01 – 0.02	**<0.001**
Sex [Male**]**				0.32	0.09	0.09 – 0.10	**<0.001**
Education [Have Degree**]**				-0.04	-0.01	-0.02 – -0.01	**0.002**
Ethnicity [White British**]**				0.08	0.03	0.02 – 0.04	**<0.001**
Age				0.002	0.01	0.01 – 0.02	**<0.001**
Townsend				-0.002	-0.00	-0.01 – 0.00	0.051
BMI				0.07	0.11	0.11 – 0.11	**<0.001**

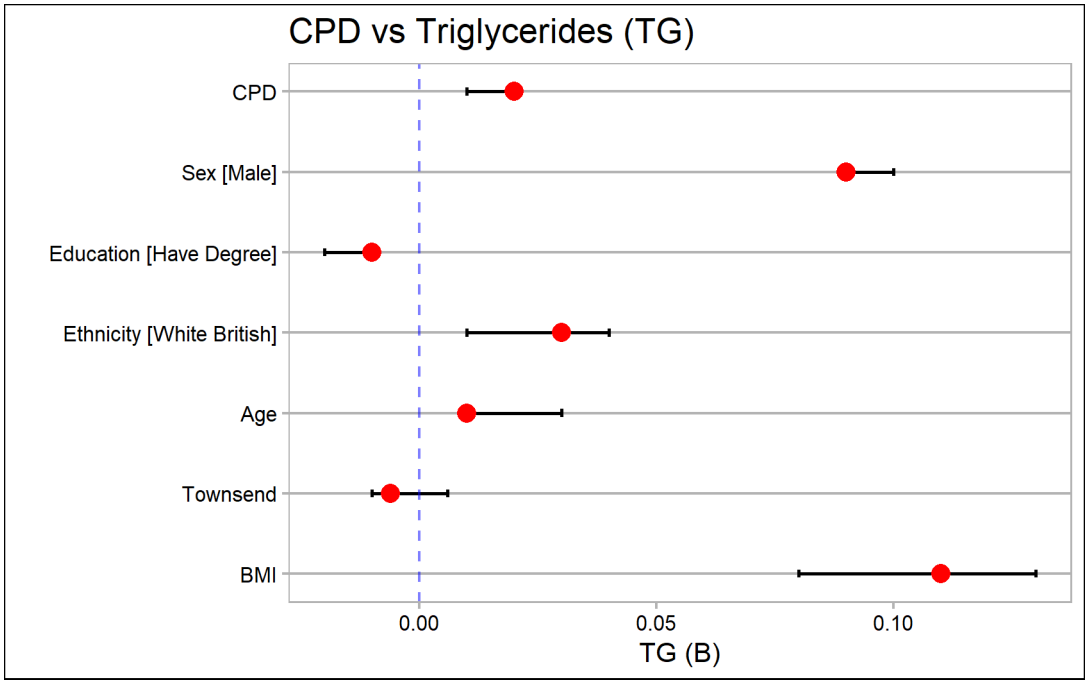


Figure 5.8. Visualisation of the adjusted associations: CperD vs TG

..1.2.33 Smoking intensity (CperD) vs high-density lipoprotein (HDL)

Unadjusted association between CperD and HDL revealed a negative and significant association. An additional cigarette smoked per day will decrease the level of HDL by 0.01 mmol/L (B= -0.01, 95% CI: -0.05 – -0.01, P<0.001). The effect of CperD on HDL remained negative and significant after adjustment for the confounders (β: -0.04, 95% CI: -0.05 – -0.02, P<0.001). These associations are summarised in Table 5.13 and visualised in Figure 5.9.

Table 5.14. Linear regression analysis of CperD vs HDL

HDL	**Unadjusted**			**Adjusted**			
	B	**CI**	**P**	**B**	**β**	**CI**	**P**
CperD	-0.01	-0.05 - -0.01	**<0.001**	-0.002	-0.04	-0.05 – -0.02	**<0.001**
Sex [Male**]**				-0.22	-0.60	-0.62 – -0.57	**<0.001**
Education [Have Degree**]**				0.04	0.11	0.08 – 0.14	**<0.001**
Ethnicity [White British**]**				-0.01	-0.02	-0.05 – 0.01	0.199
Age				0.003	0.06	0.05 – 0.07	**<0.001**
Townsend				-0.002	-0.02	-0.03 – -0.01	**<0.001**
BMI				-0.03	-0.34	-0.35 – -0.33	**<0.001**

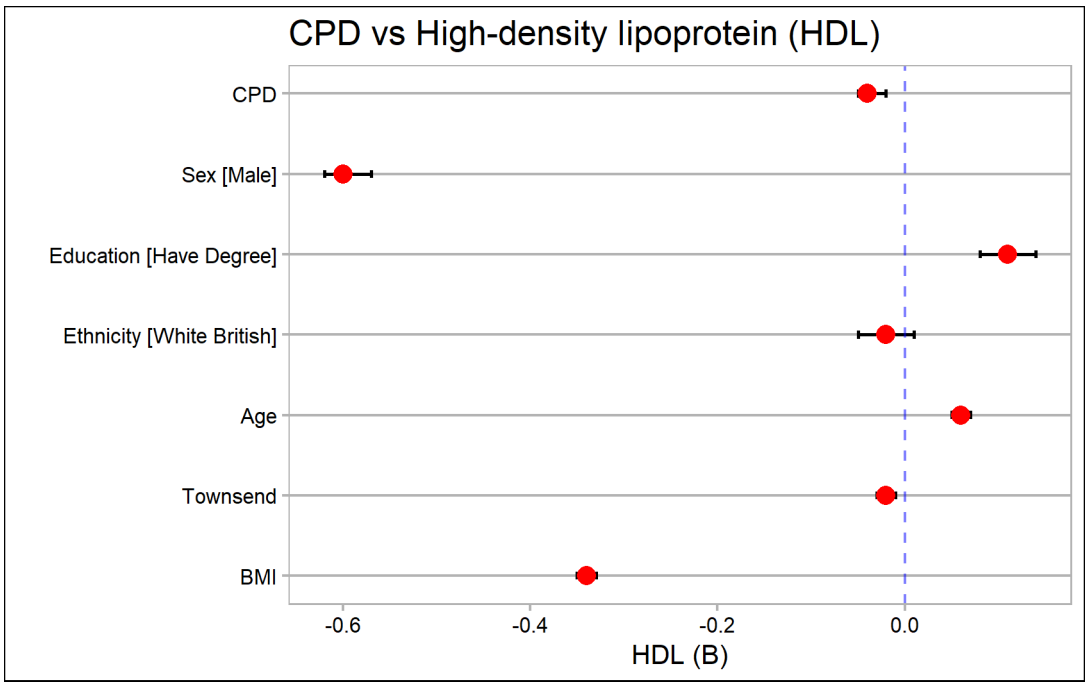


Figure 5.9. Visualisation of the adjusted associations: CperD vs HDL

Smoking initiation (SI) and lipid biomarkers

This section examined the associations between smoking initiation (age individuals started to smoke) (SI) variable and lipids variables. Each section discussed the unadjusted associations followed by adjusted results.

..1.2.34 Smoking initiation (SI) vs total cholesterol

An unadjusted association between SI and cholesterol showed a positive and significant association. The older an individual started to smoke by one year, the higher the level of cholesterol by 0.005 mmol/L (B= 0.005, 95% CI: 0.003-0.01, P<0.001). The effect of SI on cholesterol remained positive and significant after adjustment for the confounders (β: 0.01, 95% CI: 0.009 – 0.02, P=0.026). These associations are summarised in Table 5.14 and visualised in Figure 5.10.

Table 5.15. Linear regression analysis of SI vs cholesterol

Cholesterol	Unadjusted			Adjusted			
	B	CI	P	B	β	CI	P
SI	0.005	0.003-0.01	**<0.001**	0.002	0.01	0.00 – 0.02	**0.026**
Sex [Male**]**				-0.36	-0.31	-0.33 – -0.29	**<0.001**
Education [Have Degree**]**				0.03	0.03	-0.00 – 0.05	0.054
Ethnicity [White British**]**				0.08	0.07	0.04 – 0.10	**<0.001**
Age				-0.003	-0.02	-0.03 – -0.01	**<0.001**
Townsend				-0.02	-0.06	-0.07 – -0.05	**<0.001**
BMI				-0.002	-0.01	-0.02 – 0.00	0.197

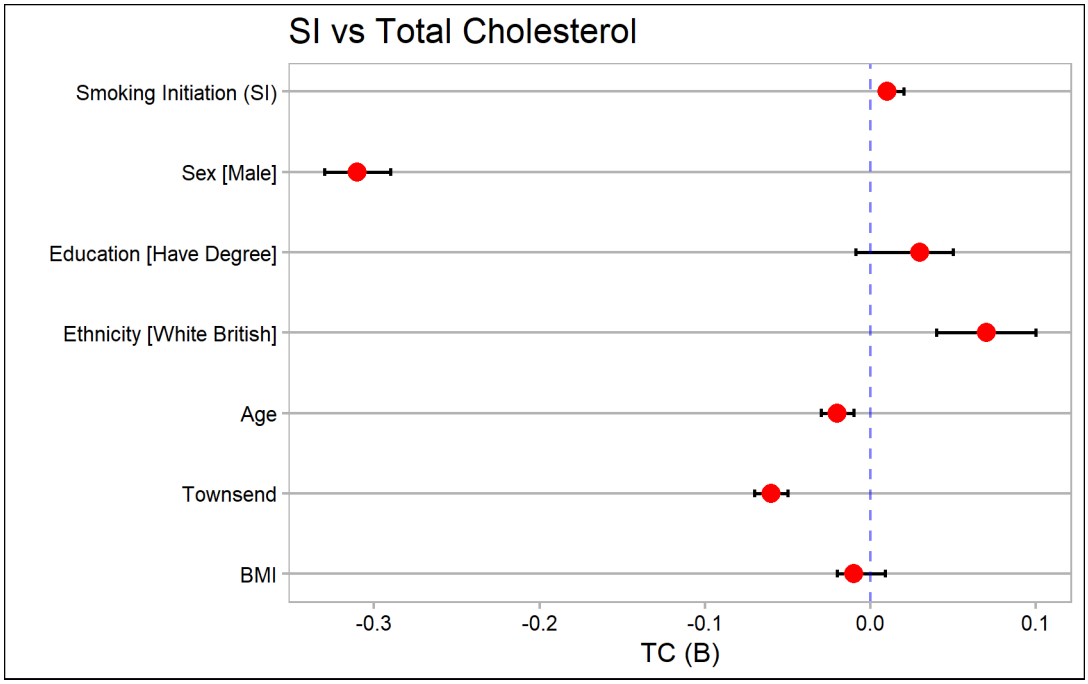


Figure 5.10. Visualisation of the adjusted associations: SI and cholesterol

..1.2.35 Smoking initiation (SI) vs low-density lipoprotein (LDL)

Unadjusted association between SI and LDL showed a positive non-significant association. The older an individual started to smoke by one year, the higher the level of LDL by 0.001 mmol/L (B= 0.001, 95% CI: -0.0003-0.003, P=0.108). The effect of SI on LDL remained positive and non-significant after adjustment for the confounders (β: 0.001, 95% CI: -0.01 – 0.01, P=0.789). These associations are summarised in Table 5.15 and visualised in Figure 5.11.

Table 5.16. Linear regression analysis of SI vs LDL

LDL	Unadjusted			Adjusted			
	B	**CI**	**P**	**B**	**β**	**CI**	**P**
SI	0.001	-0.0003-0.003	0.108	0.0002	0.001	-0.01 – 0.01	0.789
Sex [Male**]**				-0.18	-0.20	-0.22 – -0.18	**<0.001**
Education [Have Degree**]**				0.01	0.01	-0.02 – 0.03	0.609
Ethnicity [White British**]**				0.06	0.06	0.03 – 0.09	**<0.001**
Age				-0.004	-0.04	-0.05 – -0.03	**<0.001**
Townsend				-0.02	-0.06	-0.07 – -0.05	**<0.001**
BMI				0.01	0.05	0.04 – 0.06	**<0.001**

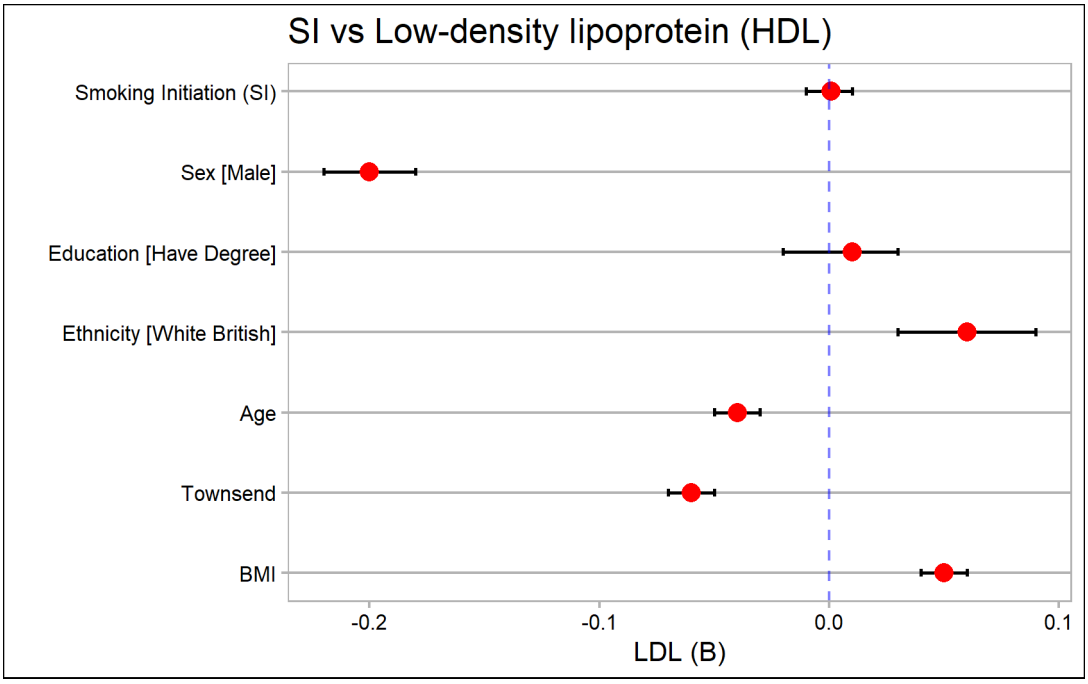


Figure 5.11. Visualisation of the adjusted associations: SI and LDL

..1.2.36 Smoking initiation (SI) vs triglycerides (TG)

Unadjusted association between SI and log TG showed a negative and significant association. The older an individual started to smoke by one year, the lower the level of log TG by 1% (0.01 mmol/L) (B= -0.01, 95% CI: -0.01-0.003, P<0.001). The effect of SI on log TG remained negative and significant after adjustment for the confounders (β: -0.01, 95% CI: -0.01 – -0.001, P<0.001). These associations are summarised in Table 5.16 and visualised in Figure 5.12.

Table 5.17. Linear regression analysis of SI vs TG

Log TG	Unadjusted			Adjusted			
	B	**CI**	**P**	**B**	**β**	**CI**	**P**
SI	-0.01	-0.01 – -0.003	**<0.001**	-0.003	-0.01	-0.01 – -0.00	**<0.001**
Sex [Male**]**				0.33	0.10	0.09 – 0.10	**<0.001**
Education [Have Degree**]**				-0.05	-0.02	-0.03 – -0.01	**<0.001**
Ethnicity [White British**]**				0.10	0.04	0.02 – 0.05	**<0.001**
Age				0.001	0.01	0.01 – 0.01	**<0.001**
Townsend				0.0002	-0.00	-0.00 – 0.00	0.644
BMI				0.07	0.11	0.11 – 0.11	**<0.001**

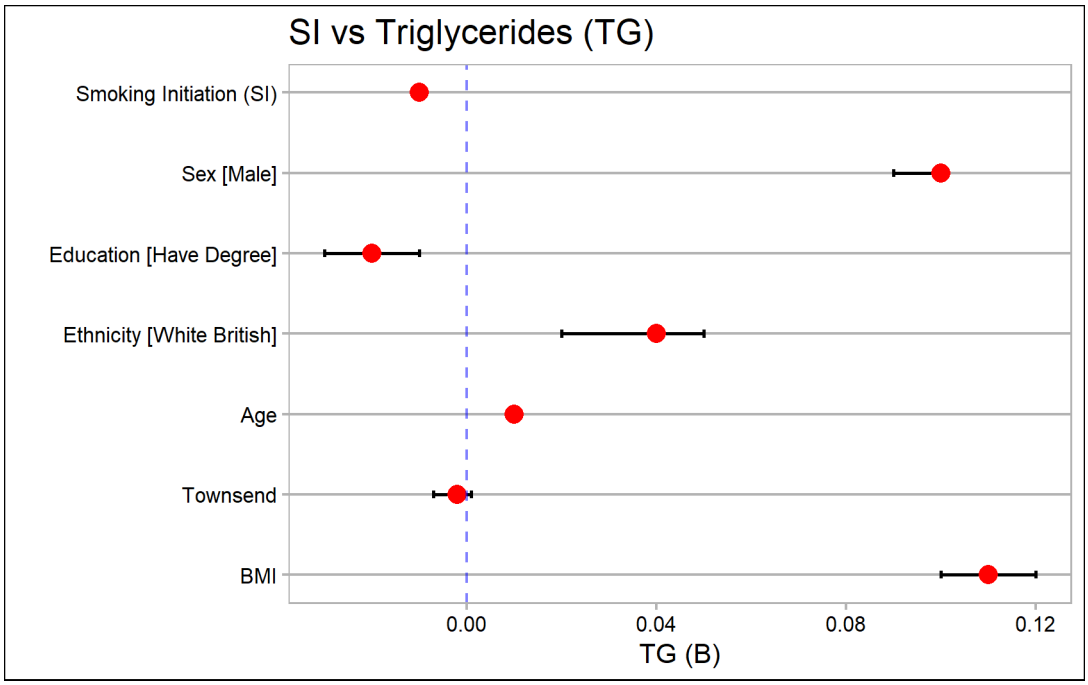


Figure 5.12. Visualisation of the adjusted associations: SI and TG

..1.2.37 Smoking initiation (SI) vs high-density lipoprotein (HDL)

Unadjusted association between SI and HDL showed a positive and significant association. The older an individual started to smoke by one year, the higher the level of HDL by 0.005 mmol/L (B= 0.005, 95% CI: 0.004-0.005, P<0.001). The effect of SI on HDL remained positive and significant after adjustment for the confounders (β: 0.04, 95% CI: -0.03 – -0.05, P<0.001). These associations are summarised in Table 5.17 and visualised in Figure 5.13.

Table 5.18. Linear regression analysis of SI vs HDL

HDL	Unadjusted			Adjusted			
	B	CI	P	B	β	CI	P
SI	0.005	0.004-0.005	<**0.001**	0.003	0.04	0.03 – 0.05	<**0.001**
Sex [Male**]**				-0.23	-0.60	-0.62 – -0.58	<**0.001**
Education [Have Degree**]**				0.04	0.11	0.09 – 0.14	<**0.001**
Ethnicity [White British**]**				-0.01	-0.02	-0.05 – 0.01	0.148
Age				0.003	0.06	0.05 – 0.07	<**0.001**
Townsend				-0.003	-0.03	-0.04 – -0.02	<**0.001**
BMI				-0.03	-0.34	-0.35 – -0.33	<**0.001**

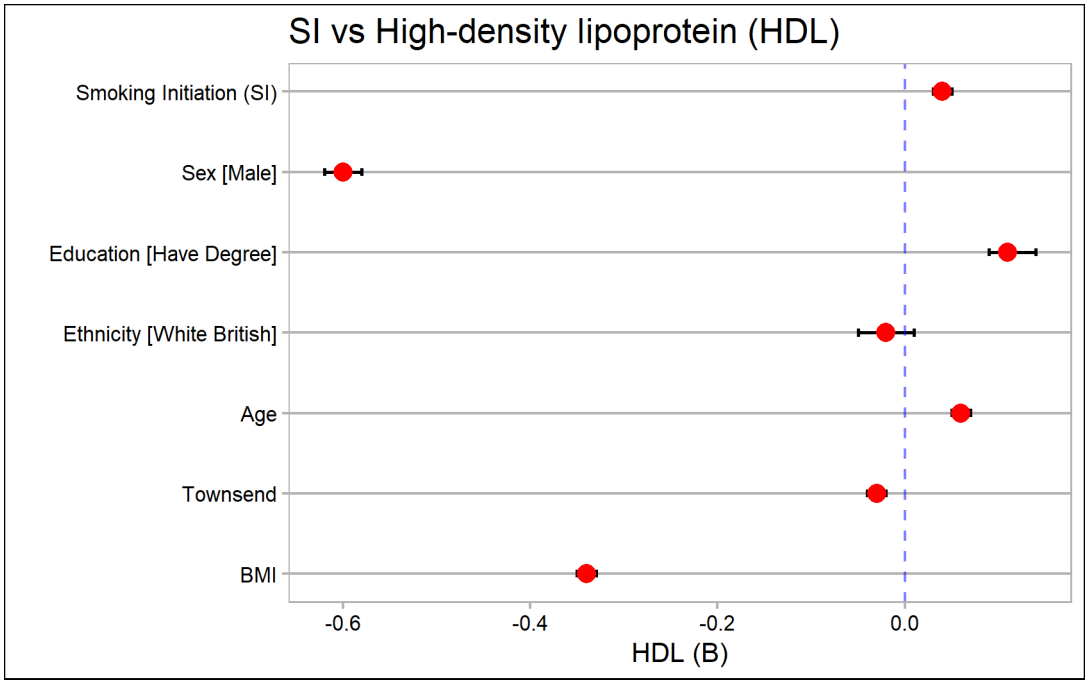


Figure 5.13. Visualisation of the adjusted associations: SI and HDL

Summary

The observational analysis of the associations between smoking behaviour and lipid biomarkers was presented in this section. Compared to never-smokers, current smokers were associated with increased levels of cholesterol, LDL and TG and decreased levels of HDL. On the contrary, the previous smokers were associated with decreased levels of cholesterol and LDL and increased levels of TG and HDL compared to never-smokers. Additionally, the more cigarettes individual smokes seem to be associated with increased levels of all lipid biomarkers except HDL. Finally, the earlier individual

started to smoke, the lower the levels of cholesterol, LDL, and HDL and higher the level of TG. Table 5.18 summarises the main findings of the observational associations between smoking variables and lipid biomarkers.

Table 5.19. The observational analysis: smoking behaviour vs lipid biomarkers

Variables		Cholesterol	LDL	TG	HDL
Smoking Status	**Current**	+	+	+	-
	Previous	-	-	+	+
CperD (more cigarettes to smoke)		+	+	+	-
SI (to start smoking older)		+	+ (non-significant)	-	+

Mendelian randomization (MR) analysis

Overview

The observational analysis of the associations between smoking behaviour and lipid biomarkers showed conflicting results. Cholesterol and LDL levels were higher among all smoking variables except previous smokers (i.e., higher in current, CperD, and SI). The TG level was higher among all smoking variables except SI. Finally, the level of HDL was lower among current smokers (vs. non-smokers), with increasing CperD and higher among previous smokers (vs. non-smokers) and SI (as an individual started to smoke later in life). As mentioned in chapter four, these associations were based on cross-sectional analysis. The presence of confounding variables is a major concern in such an approach, especially with significant associations seen with almost all covariates included in the analysis (Tables 5.6-5.17). Additionally, these observational models only explained very little variation in the lipid variables. The R-squared for all models was ranging from 2% – 22.3%. For the previous reasons, the MR approach was introduced and used to examine the associations between smoking and lipid biomarkers.

The MR analysis used smoking status (ever vs never) as well as smoking intensity (CperD) as proxies for smoking behaviour among European ancestry (Caucasian British). The next sections examined the genetic associations between smoking behaviour and lipids.

One-sample Mendelian randomization

This section focused on the casual associations between smoking behaviour and lipid biomarkers using one-sample MR. The MR approach used to examine such associations was discussed in chapter four (section: 4.3; one sample MR). The genetic scores for smoking status and CperD were valid but not the most ideal instruments to proxy smoking behaviour. The genetic scores were significantly associated with smoking variables, not associated with lipid biomarkers and not independent from the sample covariates and PCs (section: 5.2; MR analysis). The following sections explored the findings of the genetically based (MR) associations between smoking behaviour variables and lipids. Beta coefficients were used for the causal estimate for all associations.

..1.2.38 Smoking status (n=314k) vs lipid biomarkers

The observational associations between smoking status and lipid biomarkers revealed that current smokers were significantly associated with higher cholesterol, LDL and TG levels and lower HDL levels. Using the MR approach, there was no evidence of a causal relationship between smoking status (ever) vs. cholesterol (B=0.12, P=0.77), vs. LDL (B=-011, P=0.97), vs. TG (B=-0.335, P=0.10) and vs. HDL (B=23, P=0.17). The direction of the genetic associations was positive with cholesterol and HDL and negative with LDL and TG. Table 5.19 summarises these findings.

Table 5.20. MR findings for smoking status vs lipid biomarkers

Variable	**Smoking Status** (ever vs never)	
	MR Estimate	**P value**
Cholesterol	0.120	0.775
LDL	-0.0108	0.972
Log (TG)	-0.335	0.1027
HDL	0.229	0.169

..1.2.39 Smoking intensity (CperD) vs lipid biomarkers

Observationally, there was a positive and statistically significant relationship between smoking intensity (CperD) and cholesterol, LDL and TG. Conversely, the association between smoking intensity and HDL was negative. Genetically, was no evidence of a causal relationship between smoking intensity (CperD) vs. cholesterol (B=0.0016, P=0.941), vs. LDL (B=0.009, P=0.614), vs. TG (B=-0.01, P=0.969) and vs. HDL (B=-0.0002, P=0.941). The direction of the genetic associations was similar to the observational ones except for TG (CPD is positively associated with cholesterol, LDL and negatively associated with HDL). Table 5.20 summarises these findings. Visualisation of the observational vs MR findings is in the supplementary materials (8.5, results: Table 8.12). The final section of this chapter will explore a two-sample MR for the relationship between smoking behaviour and lipid biomarkers.

Table 5.21. MR findings for CperD vs lipid biomarkers

Variable	**Smoking intensity** (CperD)	
	MR Estimate	**P value**
Cholesterol	0.002 ↑	0.941
LDL	0.009 ↑	0.614
Log (TG)	-0.015 ↓	0.175
HDL	-0.0002 ↓	0.969

Two-sample Mendelian randomization (2SMR)

This section explored the summary-level MR of the relationship between smoking behaviour variables and lipid biomarkers. The main goal of the two-sample MR in this

chapter is to compare the results of the individual-level MR in the UKB with results obtained from other approaches. The lipid biomarkers data was obtained from the UKB while smoking status and smoking intensity (CperD) SNPs were acquired from different samples (Consortium of Alcohol and Nicotine use (GSCAN)). The analyses were conducted in MR-Base as well as in R. The main difference between 2SMR in MR-Base and R is the latter used the same SNPs used in the one-sample MR analysis in the UKB sample.

..1.2.40 Smoking intensity (CperD) vs lipid biomarkers – MR-Base

This section explored the associations obtained from the individual-level MR in the UKB in comparison to summary-level MR results in MR-Base. The technicalities concerning MR-Base were shown in chapter four (section: 4.3; 2SMR).

The CperD genetic data were obtained from the MR-Base GWAS catalogue (GWAS and Sequencing Consortium of Alcohol and Nicotine use, GSCAN). The sample size of this sample was 337,334 individuals of the European Ancestry. The analysis included twenty-two (22) SNPs for CperD. These SNPs were different from the ones included in the individual-level MR. The lipid biomarkers data were obtained from the UKB (Cholesterol: ukb-d-30690, LDL: ukb-d-30780, TG: ukb-d-30870 and HDL: ukb-d-30760).

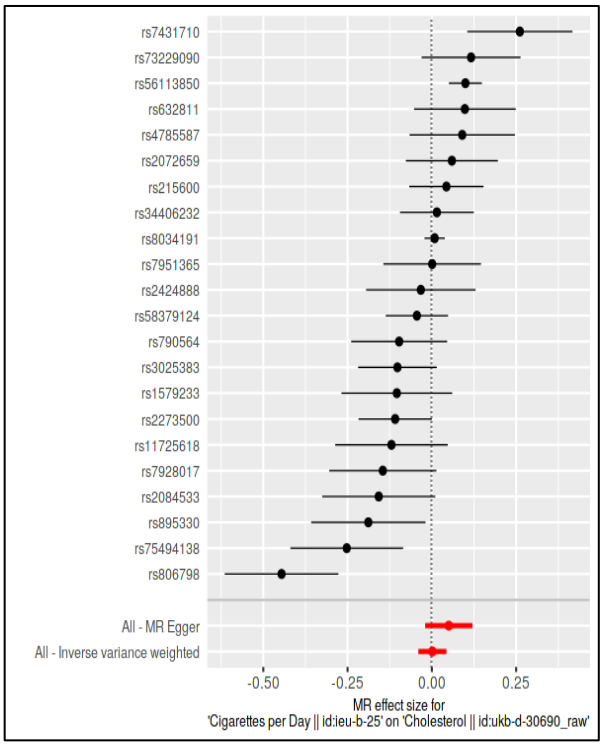
For each variable, the results of MR analysis included the MR estimates (MR-Egger) as well as the sensitivity analysis such as heterogeneity, single SNP analysis and leave-one-out analysis.

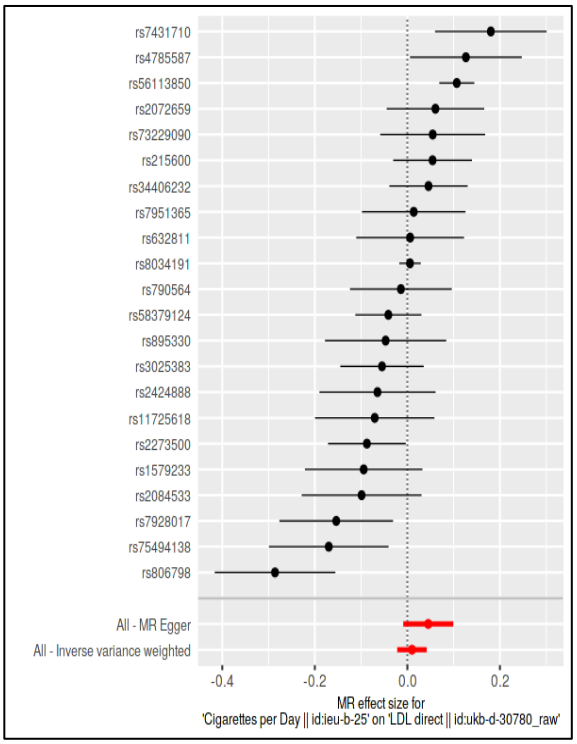
After choosing the exposure (CperD) and the outcomes (lipid variables), the MR analysis was performed. The analysis revealed no evidence of a causal association between CperD and all lipid biomarkers variables (P>0.05 for all associations). The

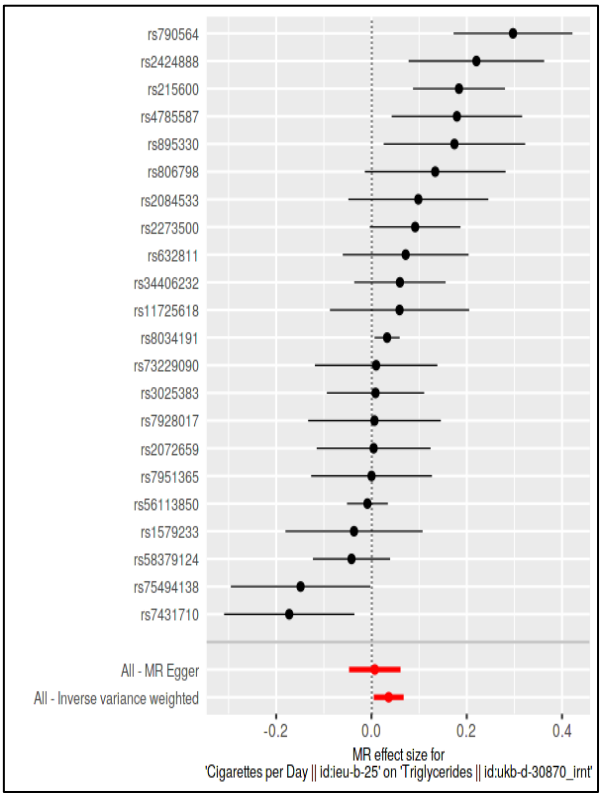
genetic predisposition of CperD, based on 22 SNPs, was positively and statistically non-significantly associated with cholesterol, LDL and TG (B = 0.05024, P=0.1755, B = 0.04514, P=0.1176 and B = 0.00697, P=0.803, respectively). On the contrary, the genetic predisposition of CperD was negatively and statistically non-significantly associated with HDL (B = - 0.005208, P=0.693). Table 5.21 and Figure 5.14 summarise MR Egger's findings as well single SNP analysis.

Table 5.22. Two-sample MR findings of CperD and lipid biomarkers (MR-Base)

Exposure (GSCAN)	Outcome (UKB)	Method	Number of SNPs	Beta	SE	P value
CperD	**Cholesterol**	MR Egger	22	0.05024	0.0358	0.1755
CperD	**LDL**	MR Egger	21	0.04514	0.0276	0.1176
CperD	**TG**	MR Egger	22	0.00697	0.02756	0.803
CperD	**HDL**	MR Egger	22	-0.005208	0.01304	0.6939







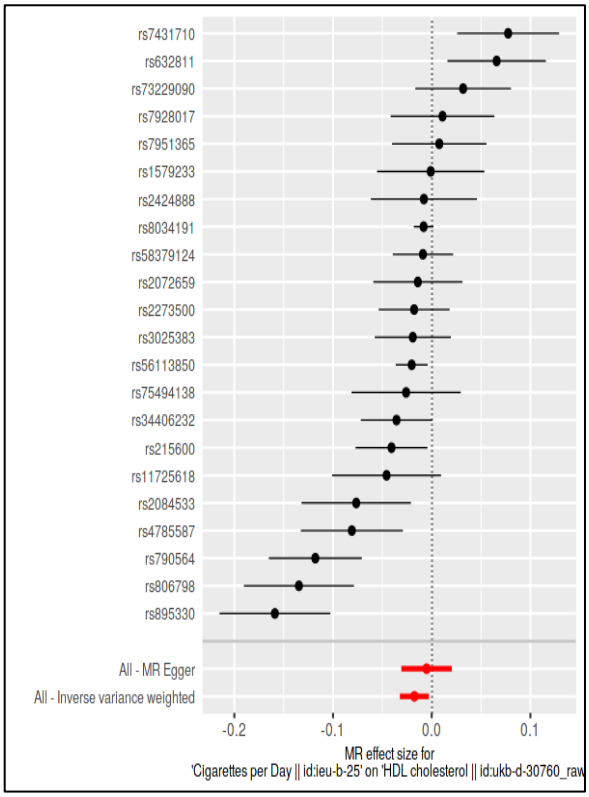


Figure 5.14: MR Egger and single SNP findings for CperD and lipid biomarkers

..1.2.41 Sensitivity analysis (MR-Base)

The results of heterogeneity analysis for all lipid biomarkers are shown in Table 5.22. Significant heterogeneity was observed for the CperD SNPs and cholesterol, LDL and TG ($P=1.547e^{-9}$, $P=1.724e^{-9}$, P=0.000002, respectively). In contrast, the CperD SNPs seem to be homogeneous for HDL (P=0.6939).

Table 5.23. Heterogeneity findings for CperD and lipid biomarkers

Exposure	Outcome	Method	P
CperD	**Cholesterol**	MR Egger	$1.547e^{-9}$
CperD	**LDL**	MR Egger	$1.724e^{-9}$
CperD	**TG**	MR Egger	0.000002
CperD	**HDL**	MR Egger	0.6939

The next sensitivity analysis to explore is a single SNP analysis. As shown in Figure 5.14 above, the CperD SNPs in general have inconsistent effects on the lipid biomarkers. The following SNPs were positively and statistically associated with cholesterol (rs7431710 and rs56113850), LDL (rs7431710, rs4785587 and rs56113850), TG (rs790564, rs2424888, rs215600, rs4785587 and rs895330) and HDL (rs7431710, rs632811). On the contrary, the following SNPs are negatively and statistically associated with cholesterol (rs3025383), LDL (rs2273500, rs7928017, rs75494138 and rs806798), TG (rs7431710 and rs75494138) and HDL (rs34406232, rs215600, rs2084533, rs4785587, rs790564, rs806798 and rs895330) (Figure 5.14).

Finally, the leave-one-out analysis revealed that the estimates of MR seem to be consistent in terms of their effect on the lipid biomarkers, however, this analysis used IVW not MR Egger regression (Figure 5.15).

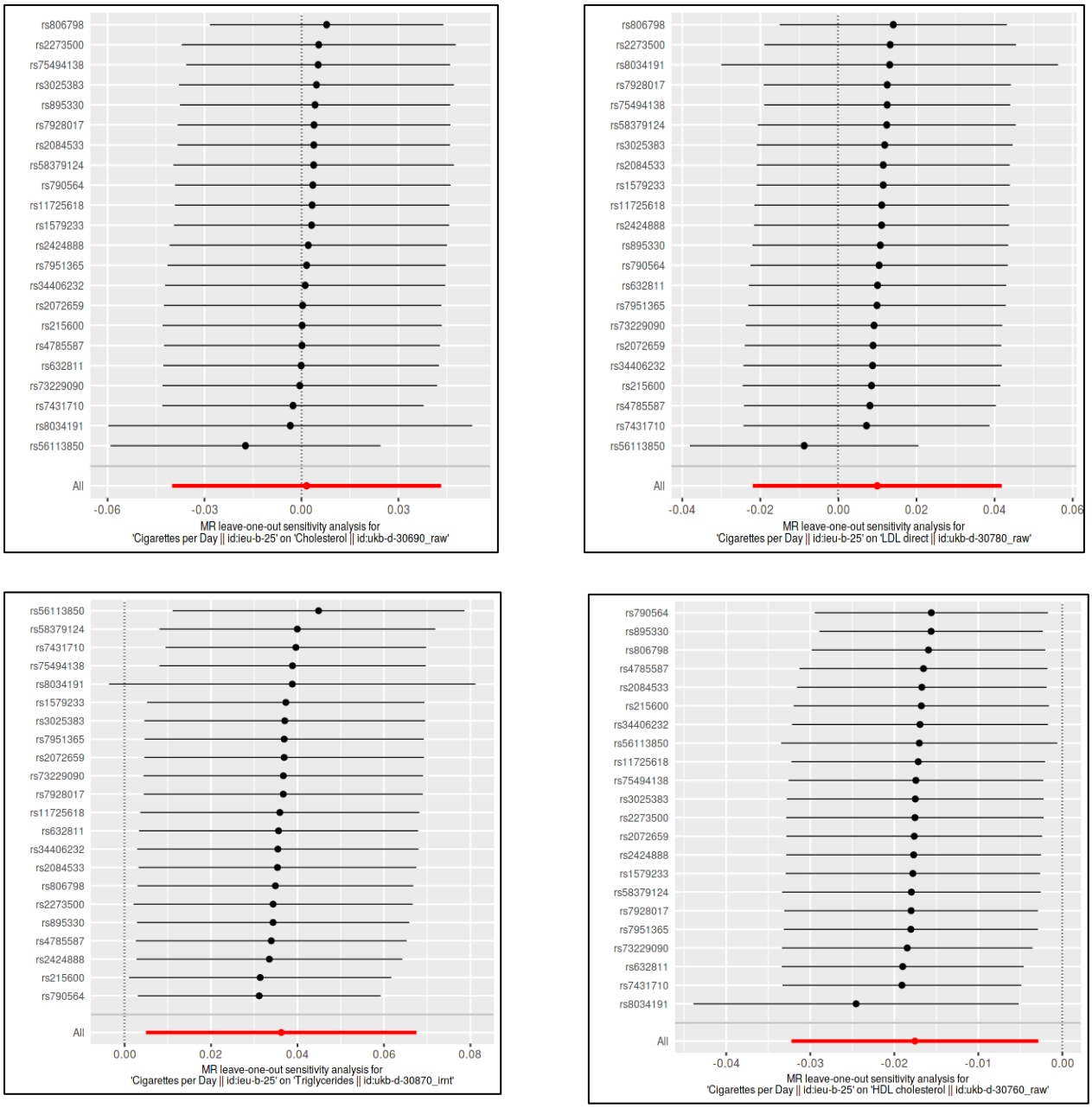


Figure 5.15. Leave-One-Out analysis plots for CperD-SNPs and lipid biomarkers

..1.2.42 Two-sample MR – in R

This section examined the association between CperD SNPs and lipid biomarkers using two-sample MR in R. The CperD summary statistics (betas and SEs) were extracted from a meta-analysis of GWAS studies that included more than 1,2 million individuals [198]. The lipid biomarkers summary statistics were obtained from the UKB. The SNPs used in this analysis matched the ones used in one-sample MR. The following SNPs

were included in the analysis: rs1051730, rs7599488, rs215614, rs73229090, rs6474412, rs3025343, rs8034191, rs2229961, rs12910984, rs3733829, rs3865453, rs28399443, rs7260329, rs2273506.

The analysis revealed that genetic predisposition of CperD, based on 14 SNPs, was positively and statistically non-significantly associated with cholesterol, LDL and TG (B=0.05, P=0.357, B=0.06, P=0.147 and B=0.02, P=0.769, respectively). Conversely, the genetic predisposition of CperD was negatively and statistically non-significantly associated with HDL (B=-0.03, P=0.117). Table 5.23 summarises these findings.

Table 5.24. Two-sample MR findings of CperD and lipid biomarkers (in R)

Exposure [GSCAN]	**Outcome [**UKB**]**	**Method**	**Number of SNPs**	**Beta**	**95% CI**	**P value**
CperD	**Cholesterol**	MR Egger	14	0.054	-0.061 - 0.169	0.357
CperD	**LDL**	MR Egger	14	0.064	-0.022 - 0.150	0.147
CperD	**TG**	MR Egger	14	0.017	-0.099 - 0.134	0.769
CperD	**HDL**	MR Egger	14	-0.031	-0.069 - 0.008	0.117

..1.2.43 Summary

The two-sample MR for the association between genetically estimated smoking intensity (CperD) and lipid biomarkers revealed no evidence of a causal association between these variables. Using MR-Base, the genetically estimated smoking intensity (CperD) (based on 22 SNPs) was positively associated with cholesterol, LDL and TG and negatively with HDL. However, these associations were statistically non-significant. Similarly, there was no evidence of a causal association between genetically estimated smoking intensity (CperD) and lipid biomarkers using R (based on 14 SNPs). The next section briefly examined the summary-level MR analysis of the association between the smoking status variable (ever vs never) and lipid biomarkers.

..1.2.44 Smoking status vs lipid biomarkers (summary-level MR, n=314k)

Genetic predisposition of smoking status, based on 15 SNPs for ever smokers, was negatively and non-significantly associated with cholesterol, LDL, and HDL (B=-0.85, -0.94, -0.32, P=0.405, 0.358, and 0.385 respectively). On the contrary, genetically estimated smoking status (ever) was positively and non-significantly associated with TG (B=1.65, P=0.331).

Individual-level (UKB) vs two-sample MR (MR-Base and R)

This section compared the findings obtained from individual-level MR in the UKB and those obtained from two-sample MR (MR-Base and R). The results of analysing the relationship between genetically estimated smoking intensity (CperD) and lipid biomarkers revealed non-significant associations across all analyses. The genetically estimated smoking intensity (CperD) has a positive association with cholesterol and LDL and a negative association with HDL across all analyses. However, the genetic predisposition of smoking intensity (CperD) was negatively associated with TG in individual-level MR and positively across two-sample MR (both approaches).

Regarding the smoking status variable, there was no evidence of a causal association with lipid biomarkers in all MR approaches. The findings obtained from one-sample MR of the UKB data were almost the opposite of the ones obtained from MR-Base for all lipid variables except HDL. Tables $5.25_{(a-b)}$ summarise these findings.

Table 5.25. Smoking behaviour vs lipid biomarkers (one-sample vs two-sample)

a) Genetic predisposition of smoking intensity (CperD) vs lipid biomarkers						
Outcomes	**MR Estimate** (One-sample UKB)	**P value**	**MR Estimate** (Two-sample: MR-Base)	**P value**	**MR Estimate** (Two-sample: R)	**P value**
Cholesterol	↑	0.941	↑	0.1755	↑	0.357
LDL	↑	0.614	↑	0.1176	↑	0.147
TG	↓	0.175	↑	0.803	↑	0.769
HDL	↓	0.969	↓	0.6939	↓	0.117

b) Genetic predisposition of smoking status vs lipid biomarkers				
Outcomes	**MR Estimate** (One-sample UKB)	**P value**	**MR Estimate** (MR-Base)	**P value**
Cholesterol	↑	0.941	↓	0.4057
LDL	↑	0.614	↓	0.3581
TG	↓	0.175	↑	0.3314
HDL	↓	0.969	↓	0.3851

Summary

Tables 5.26 and 5.27 summarise the main findings across observational and MR approaches.

Table 5.26. CperD vs lipids: observational, one-sample and two-sample MR

Smoking intensity (CperD) vs Lipid Biomarkers								
Outcomes	**Observational Estimate**	**P value**	**MR Estimate** (One-sample)	**P value**	**MR Estimate** (Two-sample) MR-Base	**P value**	**MR Estimate** (Two-sample) R	**P value**
Cholesterol	↑	**<0.001**	↑	0.941	↑	0.175	↑	0.357
LDL	↑	**<0.001**	↑	0.614	↑	0.117	↑	0.147
TG	↑	**<0.001**	↓	0.175	↑	0.803	↑	0.769
HDL	↓	**<0.001**	↓	0.969	↓	0.693	↓	0.117

Table 5.27. Smoking status vs lipids: observational, one and two-sample MR

Smoking Status vs lipids						
Outcomes	**Observational Estimate**	**P value**	**MR Estimate** (One sample) UKB	**P value**	**MR Estimate** (Summary-level) MR-Base	**P value**
Cholesterol	↑	**<0.001**	↑	0.941	↓	0.4057
LDL	↑	**<0.001**	↑	0.614	↓	0.3581
TG	↑	**<0.001**	↓	0.175	↑	0.3314
HDL	↓	**<0.001**	↓	0.969	↓	0.3851

5.4. Discussion

Principal findings

The observational findings obtained in this chapter support the results of the conventional approaches demonstrating that smoking has a conflicting effect on lipid biomarkers, but genetic evidence is less convincing.

Observationally, these results support the positive associations between smoking status (current vs. never) and smoking intensity (CperD) with a higher (i.e., worse) level of cholesterol, LDL, and TG and a lower level of HDL. Conversely, an individual who starts to smoke earlier by one year was associated with a significantly decreased level of cholesterol, HDL, and an increased level of TG. Genetically (using one-sample and two-sample MR), there was no evidence of a causal association between smoking behaviour variables (smoking intensity; CperD and smoking status) and lipid biomarkers.

Interpretation

Current smokers in this study have observationally higher levels of cholesterol, LDL, and TG and lower levels of HDL compared to never smokers. These findings were seen in a prospective cohort study conducted by Gossett et al [110]. They found that current smokers have a higher level of cholesterol, LDL, and TG and a lower level of HDL compared to never-smokers. Zhang et al [118] also found in a cross-sectional study that current smokers have a higher level of LDL and TG and a lower level of HDL compared to never-smokers. Similar findings were also found in a screening conducted by Muscat et al [115] and in a survey conducted by Willett et al [40]. Finally, a systematic review conducted by Craig WY et al found the same effect of smoking on lipid biomarkers [119].

Similar to smoking status, more cigarettes smoked per day (CperD) was associated with increased levels of cholesterol, LDL, and TG and with decreased levels of HDL. These findings were presented in the systematic review conducted by Craig WY et al. They found that as the smoking intensity increases, the level of cholesterol, LDL, and TG increases and the level of HDL decreases [119]. Similarly, Muscat et al found a positive association between smoking intensity and cholesterol level [115]. Finally, these associations were also found in a prospective cohort study conducted by Gossett et al in which smoking intensity was associated with higher levels of cholesterol, LDL, and HDL [110].

The impact of smoking on lipid biomarkers was not always consistent. For example, previous smokers have lower levels of cholesterol and LDL and higher levels of TG and HDL compared to never-smokers. In addition to smoking status, these inconsistent findings were also observed in smoking initiation. As a person begins smoking earlier in life was associated with decreased levels of cholesterol, LDL, and HDL and increased levels of TG. These findings were observed in a cross-sectional study from NHANES in which smokers and non-smokers have the same levels of cholesterol and LDL [117]. Similarly, Zhang et al also found that cholesterol level was highest among non-smokers compared to previous and current smokers [118]. Saengdith. P [41] found no association between smoking status and cholesterol level. These findings were also observed in a cross-sectional study conducted by Moradinazar et al in which no significant associations were found between smoking status and cholesterol or LDL [120].

Genetically, the findings obtained from the genetic analyses were inconsistent with the observational and between different genetic approaches. All genetic approaches (one-sample and two-sample MR) were non-significant between smoking

behaviour variables (smoking status and CperD) and lipid biomarkers. This lack of evidence of the causal association between smoking behaviour and lipid biomarkers was supported in the literature. For example, an MR study to examine the causal relationship between smoking behaviour proxied by rs1051730 and cardiovascular risk factors revealed no causal association between current smokers and HDL [39], nor rs1051730 and total cholesterol. Similarly, a summary-level MR conducted by Levin MG et al [180] found no causal association between smoking behaviour and lipid biomarkers.

Implications and future research

Observationally, smoking has a detrimental association with lipid biomarkers as smokers (including those doing so more intensely) have higher levels of cholesterol, LDL, and TG and lower levels of HDL compared to non-smokers. However, the associations between smoking behaviour and lipid biomarkers were conflicting among previous smokers and smoking initiation variables and genetically-instrumented MR associations were null. This specific effect of smoking on TG and HDL might have major consequences on cardiovascular health which needs more attention.

The genetic relationships between smoking behaviour and lipid biomarkers require more precise techniques. The availability of more robust instrumental variables that explain a large amount of smoking variability would enhance the causal inference. Additionally, implementing the genetic analysis in a larger sample size and across different populations might improve the precision and righteousness of the findings. Furthermore, Smoking is frequently linked to poorer cholesterol levels, however, there is little evidence to support this association; is possible that smoking causes negative effects. Therefore, having a poor cholesterol level is a sign of smoking, but smoking (in isolation) is not the cause - potentially.

Strengths

This chapter used a large sample size of the UKB (n=~502k) to enhance the precision of the estimates obtained from the observational and the genetic analyses. Additionally, the analyses included three variables to better picture smoking behaviour as well as a wide range of covariates to adjust for possible confounding effects. Furthermore, the chapter included observational and different approaches to the genetic analysis (one-sample and two-sample MR). Finally, the MR analyses were based on the genetic score for smoking behaviour to improve the robustness of the analyses. Further details of thesis strengths were provided in chapter four (section: 5.4; Strengths) and chapter six.

Limitations

The limitations in this chapter concern the observational as well as the genetic approaches. Observationally, the variations in the lipid biomarkers that can be explained by smoking behaviour were relatively low. The adjusted R-squared for all regression models ranged between 2% – 22.3%. This might point toward unknown confounding variables that might explain the variability in lipid biomarkers better than smoking behaviour. Genetically, Mendelian randomization is a robust technique to examine the causal association. However, the genetic approach is dependent on the validity of several steps that precede the final MR analysis. For example, GWAS and its technicalities, SNPs to be included in the analysis, the validity of these SNPs, the number of these SNPs, and the amount of variability in smoking behaviour which can be explained by these SNPs. Additionally, genetic analysis requires a very large sample size to provide a precise estimate. Finally, the number of SNPs that proxied smoking was relatively low which might produce a weak genetic score; hence, non-precise estimates are expected. However, these findings were the same with different MR approaches across different samples and different numbers of SNPs.

5.5. Conclusion

This chapter explored the relationship between smoking behaviour and lipid biomarkers observationally and genetically (using MR). This included a descriptive analysis of the variables followed by observational and MR-based causal analysis of the relationship between smoking behaviour and lipid biomarkers. Observationally, current smokers reported a higher level of cholesterol, LDL and TG and a lower level of HDL compared to never-smokers. The previous smokers have lower levels of cholesterol and LDL and higher levels of TG and HDL compared to never-smokers. Additionally, as the cigarettes smoked per day increased (CperD), the levels of cholesterol, LDL and TG increased, and the level of HDL decreased. Finally, starting smoking at a younger age was associated with decreased levels of cholesterol, LDL, HDL and increased levels of TG. These associations were statistically significant (except SI and LDL). Genetically, one-sample, as well as two-sample MR, revealed no evidence of causal associations between genetically estimated smoking status (ever) and smoking intensity (CperD) and lipid biomarkers.

6. Chapter Six: Discussion

6.1. Review of background and main findings

Gap in understanding

Cigarette smoking has previously been statistically significantly linked to an increased risk of CMDs (CHD, stroke, HTN, and DM). Additionally, smoking has been linked to alterations in lipid biomarkers, which are established risk factors for physical health conditions. These associations have been replicated in several independent studies and samples [40,66,69,103,110,214] but were based mostly on observational approaches, such as cross-sectional and (more rarely) prospective cohort studies. Such approaches are based on self-report – usually simple historical current/ever/never status - and are prone to confounding variables, reverse causation, and bias, and cannot be used to necessarily infer causality. There is therefore a gap in understanding the causal association between smoking behaviour and poorer health outcomes.

Approach

To overcome these challenges, genetic epidemiologic approaches such as Mendelian randomization were introduced to estimate causal associations between smoking behaviour and health outcomes. The current thesis differs from previous studies as follows:

1) the associations between smoking and health outcomes were examined observationally and genetically (MR) in the same sample.
2) utilising the relatively large and well-phenotyped sample of the UKB (n = approximately 502k).
3) using different MR analyses (one-sample and two-sample) to examine smoking behaviour against the multiple health outcomes of interest.

4) multiple measures of smoking behaviour: status (current; past; never), age of initiation and intensity (average use).

The primary objective of this thesis was to explore smoking behaviour and examine its impact on CMDs, stroke, and lipid biomarkers both observationally and using MR approaches.

Principle findings

The observational (i.e., cross-sectional, self-reported) results indicate that current smokers were positively and significantly associated with increased risk of CHD, stroke, DM, higher cholesterol, LDL and TG, and negatively associated with lower risk of HTN and a lower average level of HDL compared to never smokers. Previous smokers showed an increased average risk of CHD, stroke, DM, higher TG and HDL, and lower cholesterol, and LDL compared to never-smokers. Within smokers, smoking intensity (CperD) was positively and significantly associated with increased CHD, stroke, HTN, DM, cholesterol, LDL, TG, and lower HDL (i.e., consistently poorer health). Finally, smoking initiation (i.e., smoking earlier in life) was positively and significantly associated with an increased risk of CHD, stroke, and TG, and negatively associated with HTN, cholesterol, and HDL.

Genetically (one-sample and two-sample MR), there was evidence of a negative causal association between smoking status (ever) and lower HTN (one-sample MR; UKB sample), and between CperD and lower risk of DM (two-sample MR; R). The results indicate no evidence of a causal relationship between smoking behaviour (both CperD and smoking status) and CHD, stroke, or any lipid biomarkers. Table 6.1 summarises the main findings as per the research questions.

Table 6.1. A summary of the research questions and the main findings

Research Question	Summary
1) Are the instrumental variables valid to be used as a proxy for smoking behaviour in the UKB cohort population?	The instrumental variables for smoking status and CperD were valid but not the ideal tool to proxy smoking behaviour as there were some violations (2nd assumption for smoking status and 3rd assumption for both smoking status and CperD).
2) Is there a relationship between smoking behaviour and cardiometabolic disease (CMD) related health outcomes (CHD, HTN, and DM) and stroke, and if yes, is it causal?	Observationally, smoking behaviour variables were associated with increased risk of CHD, stroke, and DM and decreased risk of HTN (except for CperD which was positively associated with HTN). There was no evidence of a causal relationship between smoking behaviour and CHD or stroke. However, there was evidence of a negative and causal relationship between smoking behaviour and HTN as well as DM.
3) Is there a relationship between smoking behaviour and lipid biomarkers (total cholesterol, LDL, HDL, and TG), and if yes, is it causal?	Observationally, smoking behaviour variables were associated with higher levels of cholesterol, LDL, and TG and lower levels of HDL (except for SI which was negatively associated with cholesterol and LDL). There was no evidence of a causal relationship between smoking behaviour and lipid biomarkers.
4) Do the findings drawn from the Mendelian randomization approach match the ones from the observational associations?	Mostly no. Observational findings differ from MR findings (details below).

5) Do one-sample MR results (UKB cohort) match the ones from two-sample MR using other cohorts or other approaches?	**CMDs**: Smoking intensity (CperD) (observationally and genetically): High risk of CHD only observationally. High risk of stroke observationally and in two-sample MR (MR-Base). High risk of HTN in all approaches except two-sample MR (MR-Base). High risk of DM only observationally. Smoking status$_{(ever)}$ (observationally and genetically): High risk of CHD and stroke observationally and in two-sample MR (MR-Base). Low risk of HTN in all approaches. High risk of DM observationally and in one-sample MR (UKB sample). **Lipid biomarkers**: Smoking intensity (CperD) (observationally and genetically): High cholesterol and LDL levels among all approaches. Low HDL level among all approaches. High TG level among all approaches except one-sample MR (UKB sample). Smoking status$_{(ever)}$ (observationally and genetically): High cholesterol and LDL levels among all approaches except two-sample MR (MR-Base). High TG level among all approaches except one-sample MR (UKB sample). Low HDL level among all approaches.

6.2. Interpretation of the findings

The relationship between smoking behaviour and worse physical health outcomes is well-known and examined widely in the literature [72,117,177,215–217]. The findings obtained in the current thesis differ slightly from the findings in the literature in terms of the direction of the relationship, the significance of the findings for a different smoking variable with the outcomes, and the genetic findings (details in the following sections). All observational estimates controlled for sex, educational attainment, ethnicity, age, Townsend deprivation score, and BMI.

Smoking status (current vs never)

Cardiometabolic diseases (CMDs)

Observationally, compared with never-smokers, current smokers in the current thesis showed a 61% higher risk of coronary heart disease (CHD), 64% higher risk for stroke, and 12% higher risk of diabetes mellitus (DM) (all P<0.001). Conversely, current smokers reported an 11% lower risk of hypertension (HTN) compared to never smokers (P<0.001). Previous smokers have the same findings as the current smokers (however non-significant for HTN). Genetically, there was evidence of a causal association between smoking history and HTN where smokers have a lower risk of HTN compared to never-smokers (OR=0.44, P=0.001). There was no evidence of a causal association between smoking status and CHD, stroke, and DM (all P>0.05).

Observational findings were mostly consistent with the results found in the literature. Smokers have a higher risk of CHD, stroke, and DM compared to never smokers. Such findings were observed in many studies [69,75,77,105]. However, HTN findings were consistent with some papers but differ from others. The negative relationship (i.e. protective impact) between smoking status and HTN was found in Leone A et al [218], Liu and Byrd [88], and Alomari MA, Al-Sheyab NA [89]. On the contrary, some studies found a positive relationship (detrimental impact) between smoking and HTN [90–92]. Genetically, a lack of evidence of causal relationships between smoking and CMDs was observed in Linneberg et al and Åsvold BO et al [39,177]. However, some papers reported causal associations between smoking and CMDs [178–180]. The negative causal association between smoking status and HTN found in this thesis opposes the ones found in the literature [179,219].

Lipid biomarkers

Observationally, current smokers showed on average 0.05 mmol/L higher cholesterol, 0.06 mmol/L higher LDL, and 0.09 (9%) mmol/L higher TG levels compared to never-smokers (all $P<0.001$). Conversely, current smokers have a 0.14 mmol/L lower level of HDL compared to never smokers ($P<0.001$). Previous smokers have lower levels of cholesterol and LDL and higher levels of TG and HDL (all $P<0.05$). Genetically, there was no evidence of a causal relationship between smoking and lipid biomarkers (all $P>0.05$).

The results of observational studies were mostly in line with those found in the literature. Smokers have higher levels of cholesterol, LDL, and TG and lower levels of HDL compared to never smokers. Many studies have shown similar results [110,115,118]. However, some studies found no difference in lipid levels among smokers and never smokers [116,117,120]. Furthermore, Zhang et al found the highest level of cholesterol among never smokers compared to current smokers [118]. Genetically, the lack of evidence of the causal association between smoking status and lipid biomarkers was consistent with two MR studies by Levin MG et al and Åsvold BO et al. [39,180].

Smoking intensity (CperD)

Cardiometabolic diseases (CMDs)

Smoking intensity is positively associated with the risk of CMDs. Observationally, the findings showed that as an individual smokes more cigarettes per day the risks for CHD, stroke, HTN, and DM increase (all $P<0.05$). The findings obtained in this thesis were consistent with the findings in the literature [72,73,75,91,104]. Genetically, there was no evidence of a causal relationship between smoking intensity and CMD variables (all $P>0.05$) in all MR approaches. However, there was evidence of a causal association

between smoking intensity (CperD) and DM in two-sample MR using R. As individuals smoke more cigarettes per day, the risk of DM decreases (OR=0.50, P=0.002). Some null causal associations between smoking intensity and CMD variables were found in the literature [39,207].

Lipid biomarkers

The impact of smoking intensity (more cigarettes per day) on lipids was the same as smoking status (observationally and genetically). Observationally, as an individual smokes on average one more cigarette per day, the level of cholesterol increases by 0.02 mmol/L, LDL increases by 0.03 mmol/L, TG increases by 0.02 (2%) mmol/L, and HDL decreases by 0.04 (all P<0.001). This unfavourable impact of smoking intensity on lipid biomarkers was consistent with the findings in the literature [110,115,119]. Genetically, there was no evidence of a causal relationship between smoking intensity and lipid biomarkers. Such findings were similar to the ones found in the literature [39,180].

Smoking initiation

Cardiometabolic diseases (CMDs)

The effect of smoking initiation (i.e., earlier age at starting to smoke) on CMDs was similar to the effect of the smoking status variable. As an individual starts to smoke earlier in life the risk of CHD, stroke, and DM increases and the risk of HTN decreases. The earlier an individual smoke by one year was associated with a 4% increase in CHD risk, a 4% increase in stroke risk, a 1% increase in DM risk (non-significant) and a 1% decrease in HTN risk (all P<0.05). This unfavourable impact of smoking initiation was seen in a large prospective cohort study (ARIC) and others [74,75,77,78].

Lipid biomarkers

The impact of smoking initiation on lipid biomarkers was inconsistent with the previous findings. As an individual starts to smoke earlier, the level of cholesterol decreases by 0.01 mmol/L (P=0.026), LDL decreases by 0.26 mmol/L (P=0.789), and HDL decreases by 0.04 mmol/L (P<0.001). However, the earlier an individual starts to smoke was associated with a 0.01 (1%) mmol/L increased level of TG (P<0.001). Similar findings were found in the literature in a longitudinal prospective cohort study that examined the impact of early smoking initiation on different health outcomes [220].

Observational vs. MR findings

The findings obtained in this thesis can be classified into three categories: 1) consistent significant findings among observational and MR, 2) opposite significant findings between both approaches, and finally 3) significant findings observationally but null genetically. Table 6.2 illustrates these findings.

Table 6.2. Observational findings vs. MR findings

Consistent findings	Opposite findings	Null findings
↓ HTN in both approaches (Smoking status: ever)	↑ DM in observational (CperD) Vs. ↓ DM in MR (CperD)	Rest of associations between smoking variables and health outcomes (CMDs and lipid biomarkers)

What do these findings mean?

Consistent findings on CHD, stroke, and DM

The harmful observational association of smoking behaviour on CHD, stroke, and DM seems to be consistent among all smoking variables. These findings support the observational-based evidence found in the literature. However, smoking intensity (CperD) has a seemingly causal negative protective relationship with DM. This significant finding was only found in one approach of MR (two-sample MR in R). Furthermore, most of the MR findings concerning CperD and DM were negative

(smoking intensity increases, DM risk decreases) suggesting this finding needs replication in independent cohorts and correcting for potential selection and attrition bias.

Protective effect of smoking behaviour (negative/null findings)

The favourable impact that smoking has on HTN (observationally and genetically) seems to be different from the generally expected deleterious effect of smoking. The impact of smoking on lowering HTN risk was consistent among two smoking variables (smoking status: current vs. never) and age at smoking initiation) as well as among one-sample MR in the UKB. The genetically-estimated effect of smoking on HTN might be attributed to the violation of the IV assumption as the genetic score was significantly associated with HTN. Additionally, the observed negative/null associations between smoking behaviour and the outcomes might be attributable to attrition bias in which less healthy participants might have been less likely to attend the assessment (or died before the assessment) leaving the sample with healthier participants. Attrition bias among the UKB participants significantly affects the estimates of the associations as seen in the Lyall et al study [221]. Additionally, the participants in the UKB are known to be healthy and that might explain the null and negative findings [222]. Furthermore, the negative/null findings might suggest that it is rather that smoking proxies poor lifestyle in other ways which perhaps do cause poorer health, like poor diet, and low exercise [223,224]. Therefore, the impact of smoking behaviour needs further investigation in independent cohorts and multiple MR approaches.

Smoking behaviour variables paradox

The findings obtained in the current thesis were not always consistent. For example, observationally, the effect of smoking on lipid biomarkers differs based on the smoking variable examined. Smoking status (current vs. never) and smoking intensity variables

were associated with high levels of cholesterol, LDL, and TG and low level of HDL. Conversely, earlier age at smoking initiation was associated with lower average cholesterol, LDL, HDL and higher TG. These findings might explain the conflicting results found in the literature. The smoking variable used to test the association might be the source of conflicting findings in the published literature. This is also found in smoking intensity (CperD) and HTN (i.e., that increasing intensity is associated with a higher risk of HTN). All smoking variables were associated with decreased risk of HTN - except smoking intensity which was associated with increased HTN risk. These observations regarding smoking variables require further study including longitudinally.

Lack of genetic evidence

The lack of evidence of a causal relationship between smoking behaviour and CMDs and lipid biomarkers in the UKB sample does not rule out the causality of these associations. There may be possible reasons for such findings; 1) weak genetic score proxying smoking behaviour, 2) poor SNPs that are used to build the score in which they capture small amounts of variance in smoking behaviour as well as the violation of IV assumptions, 3) a small number of cases for the CMDs and lipid biomarkers as well as for smoking intensity (n=25k), 4) the UKB is relatively healthy and not deprived and this can lead to underestimations of effect [221]; more deprived cohorts may find different results, and 5) true findings (no direct causal association between smoking behaviour and these outcomes in isolation from other lifestyle risk factors).

Smoking behaviour is a well-known risk factor for numerous health conditions [44,72]. Smoking exerts its effect on cardiovascular and overall health via the physiological changes associated with toxins found in cigarettes. These toxins (such as tar, carbon monoxide, free radicals, and nicotine) make the blood dense and increase

the risk of clot formation which increase the blood pressure and pulse rate. This in turn leads to an extra effort on heart muscles. Additionally, these toxins make the arteries smaller, thicken the artery walls, and decrease the amount of oxygen-rich blood that can be circulated throughout the body [225]. The current findings do not suggest that smoking poses no health risks; rather, it suggests that it might not be as causally bad in isolation from other lifestyle risk factors for CMDs and lipids. Additionally, it might indicate invalid instrumentation for the MR analyses. Finally, factors such as attrition bias and healthy UKB participants might also explain such findings.

6.3. Contribution to the literature

The current thesis examined smoking behaviour using more than one measure to proxy smoking; smoking status (current/previous vs never), smoking intensity (cigarette smoked per day (CperD) and finally age at initiation. Additionally, the current thesis examined smoking behaviour and the outcomes in one relatively large sample with a standard protocol, with a wide range of covariates which could potentially confound an association.

Contribution to UKB and the scientific community

This thesis added to UKB and the scientific community a few major contributions;

1) Exploration of smoking behaviour using three variables which conceptualise smoking in more depth compared to other studies [75,115,214,226]. Inconsistency between the independent/dependent variables highlights the need for multiple phenotyping in smoking assessment.
2) Examining the association between smoking behaviour and more than eight health outcomes considering a wide range of covariates in one large sample of the UKB.

3) Investigating the causal relationship between smoking behaviour and health outcomes using multiple MR methods: one-sample and two-sample MR among UKB participants using different approaches (MR-Base and R).
4) Contrasting the observational and genetic (causal) relationships between smoking and health outcomes.
5) Proposing that the association between smoking and health may be quite confounded, or not directly causal in isolation from other lifestyle behaviours (such as poor diet or lack of daily activity) which often correlate with being a smoker [227–229].

Different smoking characteristics show different associations

One major contribution this thesis added to the literature is analysing smoking behaviour against a wide range of health outcomes. This detailed exploration produced insight into the relationship between smoking behaviour and some health outcome variables. For instance, current and early-age smokers were associated with a low risk of hypertension, but the more cigarettes an individual smokes the greater the risk of hypertension. This might explain the contradiction concerning the relationship between smoking behaviour and hypertension found in the literature [81,84,88,218]. Another example is the ambiguity of the relationship between smoking behaviour and lipid biomarkers [41,118,120]. Current smokers and smoking intensity (CperD) were associated with higher levels of cholesterol, LDL, and TG and lower level of HDL. However, as an individual starts to smoke early in life, the level of cholesterol, LDL, and HDL decrease. These findings might explain such contradiction observed in the relationship between smoking behaviour and lipid biomarkers.

Causal associations between smoking behaviour and health outcomes

The main contribution provided by the current thesis is examining the causal association between smoking behaviour and health outcomes. Causality was tested using the genetically-estimated MR approach. This thesis added to the understanding of the causal association between smoking and health outcomes as follows; 1) using two genetic scores for two smoking behaviour variables, 2) these two scores were based on multiple SNPs, 3) the validity of the SNPs and scores were tested before conducting the analyses, and 4) examining the causal associations for all health outcomes in the same sample (UKB) and other samples using two methods: one-sample and two-sample MR, utilising different approaches and platforms (R and MR-Base respectively).

This thesis found evidence of a causal relationship between smoking behaviour and low risk for HTN and DM with no evidence of a causal relationship with other health outcomes. These findings differ from others in the literature which provided evidence of a positive causal relationship between smoking behaviour and health outcomes [170,178,179,181]. The contradiction between these findings might be attributed to the nature of the samples where MR was conducted, the validity of the genetic scores used in the analyses, the SNPs used to build the genetic scores or the smoking variable that was used for the associations. Additionally, this thesis examined the causal associations between smoking behaviour and health outcomes in the same sample (UKB) which differs from other MR analyses in the literature which mostly use two-sample or summary-level MR. Conducting MR analyses in a large sample size like UKB including the same participants in two different smoking variables is the main difference between this thesis and other papers in the literature. These detailed analyses and findings provided insight into understanding the causal relationship between smoking behaviour and health outcomes, especially in the UKB sample.

6.4. Implications

The results of this thesis contribute to a better understanding of smoking behaviour and its association with health outcomes. There are three areas where this thesis can help in clinical practice and public health.

Public health intervention (smoking behaviour variables)

The observational findings support the detrimental effect of smoking behaviour on CMDs and lipid biomarkers. The current thesis explored smoking behaviour via three smoking variables, hence three public health interventions are to be considered. First, health education on the effects of smoking on health considering smoking status findings. As the finding suggested, smokers were significantly associated with higher health-related problems compared to never smokers. Second, the intervention should be directed toward lowering the number of cigarettes smoked per day (smoking intensity) in current smokers. This approach can be applied at the beginning of tackling smoking behaviour, a step toward smoking cessation [230]. Third, public health interventions, such as health education, should be implemented as early as possible to prevent smoking at a younger age [231]. The harmful impact of smoking behaviour was higher as individuals started to smoke early in life, as the current findings suggested. In addition to suggested public health implications, researchers should handle the association between smoking behaviour and health outcomes with caution. As seen in the findings, the associations might be positive or negative based on the smoking behaviour variable used for the analysis (e.g., HTN and lipid biomarkers among different smoking variables).

Public health intervention (smoking and hypertension)

The associations between smoking behaviour and hypertension revealed mostly ostensibly favourable associations with smoking. The risk of HTN was lower among

smokers compared to never smokers, and with earlier initiation life. Such a finding was also confirmed using the genetically-estimated MR approaches in the UKB. Public health intervention can be directed toward further exploration of the relationship between smoking behaviour and hypertension considering all possible smoking variables as well as the numeric measurements for blood pressure. However, these findings can be simply because the UKB participants are healthy (healthy volunteer bias/attrition bias).

Public health intervention (smoking and unhealthy lifestyle)

Based on the current findings, smoking's influence on health outcomes might not be entirely as causal as some believe; rather, smoking may serve as a proxy for other undesirable lifestyle, environmental, or unmeasured factors. For instance, smokers have an unhealthy lifestyle compared to non-smokers, they drink more alcohol and eat fewer fruits and vegetables. Additionally, smokers tend to be physically inactive compared to non-smokers [224]. Therefore, tackling smoking should be more holistic considering lifestyle overhaul and not just “stop smoking”.

Future MR analyses of smoking behaviour

To improve MR studies of smoking behaviour in the future, researchers can use larger sample sizes and more robust genetic instruments, apply sophisticated statistical methods such as MR-Egger regression and weighted median regression, and conduct studies in multiple populations to increase statistical power, precision, and generalisability of results, and help identify potential biases and sources of heterogeneity. They can also consider using MR-prospective studies, which use genetic information collected before the outcome occurs, and combining multiple MR approaches, such as instrumental variable analysis and summary data-based MR, to

provide a more comprehensive understanding of the causal effect of an exposure on an outcome and reduce the potential for reverse causality.

6.5. Strengths

The current thesis has several strengths that enhance the robustness of the analyses and the overall results.

1. Large sample size

The current thesis leveraged the large sample size of the UKB (~502k) [232] which improves the accuracy of estimates and the power to detect statistical significance derived from the observational and MR analyses [233]. Despite the healthy feature of the UKB participants [221], the risk factor associations have proven to be generalisable [188,234], hence, the conclusions from this thesis can be applied to a wider population, at least in the UK.

2. Consistent phenotyping

The current thesis included a wide range of dependent variables; four binary variables concerning cardiometabolic diseases (CHD, stroke, HTN, and DM) and four numeric (measurement) variables of lipid biomarkers (cholesterol, LDL, TG, and HDL).

3. Detailed covariates

The study considered several covariates to minimise the risk of confounding variables in the association between smoking behaviour and health-related outcomes (sex, age, deprivation score, education attainment, ethnicity, and body mass index; BMI).

4. Multiple measures of smoking behaviour

The analysis involved three variables to better depict smoking behaviour (smoking status, smoking intensity, and smoking initiation). Each smoking variable was

examined against all study outcomes. This detailed exploration of smoking behaviour provides a comprehensive reference of the smoking variables in the UKB.

5. Multiple measures of MR

The strengths of the genetic analysis in this thesis included the following: 1) the study uses genetic scores for smoking behaviour instead of using a single SNP to improve the robustness of the analysis, 2) the MR analyses included two genetic scores proxying two smoking variables (smoking status and smoking intensity) for a better conceptualisation of smoking behaviour and more precision of the findings, 3) the SNPs and the genetic scores were tested for the IV assumptions to ensure the validity of the MR estimates, 4) the genetic analyses incorporated one-sample and two-sample MR among different samples and across different platforms to ensure the validity of the causal findings, and 5) the study applied the genetic analysis for smoking status variable on around 314k participants which are considered a large sample size for individual-level MR analysis compared to the previous studies [39,177].

6.6. Limitations

The limitations of this thesis will be categorised into two parts: 1) limitations of the UKB and observational approach and 2) limitations of the genetic approach.

Observational approach (selection bias)

There are two limitations concerning the observational approach. First, the participants of the UKB are not representative of the whole UK population. They seem to be healthier compared to the general population of the UK raising the possibility of healthy volunteer bias [235]. This might explain the favourable effects of smoking behaviour on HTN and lipid biomarkers as well as the null findings of the MR analyses. Current smokers may have a lower risk of some conditions because they are healthy ones. This effect might be even more prominent when participants leave the UKB or drop the

follow-up assessments (attrition bias). Second, the variability in the outcomes which can be explained by smoking behaviour was low (adjusted R^2 for lipid biomarkers associations: 2% – 22.3%, pseudo-R^2 for CMDs: 0.4% – 13%). This might point toward unmeasured confounding variables that carry higher contributions to these outcomes. Lastly, the analyses used nominal P = 0.05 which might introduce type one error in the findings (false positive findings) [236].

MR approach (instrument validity)

The limitations of the genetic approach can be summarised as follow. First, the MR analysis is based on the validity of the genetic scores which are based on the SNPs selected for the analysis. The genetic scores in this thesis were valid but not the best instrument to proxy smoking behaviour. However, the genetic scores, as well as the SNPs, were tested for validity before the analysis. Second, the variation in smoking which can be explained by each SNP was low. However, the genetic scores were used instead of using individual SNPs to enhance the power of the analysis. Third, genetic analysis requires a very large sample size for robust and precise estimation. The MR analysis of the smoking status variable included ~314k participants, however, only 25k participants were included in the smoking intensity analysis (the smoking intensity variable in the UKB is only among current smokers so it has only ~25k participants). To overcome such limitations, the analysis included one-sample and two-sample MR as well as using different platforms across different samples. Fourth, the number of SNPs included in building the genetic score was relatively low (15 SNPs) which might produce a weak genetic score and subsequently non-precise estimates. However, the study used two genetic scores for two different smoking variables and tested across different approaches which generally produced almost the same findings. Fifth, the genetic scores (as well as some individual SNPs) have a significant association with the

outcomes (such as HTN and DM) as well as with the covariates (such as BMI). For example, the genetic score for smoking intensity (CperD) was significantly associated with a lower level of BMI. Additionally, the genetic score for smoking status was significantly associated with a lower risk of HTN and lower BMI level. These pleiotropic effects might drive the negative associations between smoking behaviour and these outcomes. However, the analysis used MR-Egger to overcome any pleiotropic effects that may have resulted from the IV assumptions violation. Finally, in isolation from other lifestyle factors that frequently correlate with smoking, the association between smoking and outcomes may be quite confounded, or not directly causal.

Despite these limitations, this thesis used a variety of methods to achieve reliable results and to provide the best possible answers to the research questions.

6.7. Future research

The current thesis investigated smoking behaviour observationally and genetically. Despite the significant effort to impartially examine the associations between smoking behaviour and health outcomes, there is always room for advancement in terms of variables, analysis, and study cohort.

Objective measurements for the study variables

The variables for smoking behaviour and health outcomes can be measured objectively whenever possible. For example, there are several variables in the UKB related to smoking behaviour [38]. Instead of using smoking variables separately, it could be better to build a score that included smoking status, smoking intensity, smoking duration, etc. Furthermore, smoking metabolites such as cotinine can be used as a numeric measurement for smoking behaviour instead of self-report [237]. Another example is the method of diagnosing hypertension and diabetes mellitus. Using numeric measurements is an objective way to classify individuals as having high blood pressure

and blood sugar. These objective and quantifiable measurements might produce more accurate and precise estimates concerning the relationship between smoking behaviour and health outcomes [238].

Subgroup analyses (stratification)

One of the improvements to be considered in the future is examining the associations among subgroups. Stratification might unveil hidden findings among certain groups. For instance, the UKB sample can be subsetted based on sex, and the association between smoking behaviour and health outcomes can be examined separately among the male sample and the female sample. A further approach is examining smoking behaviour after normalising all other variables in the study. For example, when investigating the smoking status variable against coronary heart disease, only individuals with normal readings will be included in the analysis. So, the analysis will include individuals who have normal blood pressure, normal body weight, normal blood sugar, normal cholesterol level, etc. This might minimise the risk of unwanted effects of other variables on the relationship between smoking behaviour and health outcomes.

Alternative genetic instruments

In the genetic approach, a significant area for improvement is the instrumental variable that proxies smoking behaviour. The availability of more SNPs that can explain significant variations in smoking behaviour will provide a robust and valid genetic score, resulting in precise and valid findings. Additionally, to better understand the causal impact of each smoking variable on health outcomes, each smoking behaviour variable should be instrumented separately. Furthermore, the current thesis applied genetic analysis to the UKB sample and European ancestry. The genetic association between smoking behaviour and health outcomes can be studied in different cohorts

(e.g., Generation Scotland) and other ancestries. This will provide a better understanding of the causal relationship between smoking and health outcomes, as well as the genetic approach used to test such a relationship. Finally, Mendelian randomization is a relatively new approach [239] with anticipated advancements in the future.

6.8. Conclusion

This study looked at smoking behaviour in terms of both observational and causal associations with various health outcomes primarily among the UKB participants.

The thesis used multiple smoking variables (smoking history, intensity, and age at initiation) to examine the observational relationship with coronary heart disease (CHD), stroke, hypertension (HTN), diabetes mellitus (DM), total cholesterol, low-density lipoprotein (LDL), triglycerides (TG), and high-density lipoprotein (HDL). The thesis investigated the causal association between smoking variables (smoking status and smoking intensity) and the aforementioned health outcomes using genetically-estimated Mendelian randomization (MR) causal estimates (one-sample and two-sample approaches) using R and MR-Base respectively).

The study discovered observational-based evidence of the harmful effect of smoking behaviour on CHD, stroke, DM, cholesterol, LDL, TG, and HDL. However, there was no harmful effect of smoking initiation on cholesterol and LDL. There was limited supportive evidence of the causal relationship between smoking variables and these outcomes. Contrary to other health outcomes, smoking was negatively (protective effect) associated with HTN, observationally and 'causally'.

Such findings supported the findings in the literature on the harmful effect of smoking behaviour on health, however, the current thesis extended the analysis to

include more than one variable representing smoking behaviour, both observationally and causally. The inclusion of more than one smoking variable provided more insight into the effect of each variable of smoking behaviour on health outcomes. Some health outcomes were harmfully impacted by all smoking variables (such as CHD, stroke, DM, TG, and HDL), while others (such as HTN, cholesterol, and LDL) had inconsistent findings among different smoking variables.

Based on current findings, future research should consider a wider variety of smoking variables to better understand the relationship between smoking behaviours and health. Additionally, the protective/null effect of smoking behaviour on health outcomes requires further investigation (considering the biases carefully). Furthermore, these null/protective findings suggest that some of smoking's association might be attributable to other aspects of environment/lifestyle and not smoking in isolation. Moreover, the genetic approach has more evidence to present about the causal impact of smoking behaviour through more reliable and valid SNPs that explain a significant amount of the variability in smoking behaviour, hence more robust MR analysis and more precise findings. Finally, public health initiatives to stop smoking could aim to reduce daily cigarette use, launch anti-smoking campaigns earlier in life, and consider people's overall health and lifestyles.

7. References

[1] World Health Organisation, tobacco key facts. [Internet] 2019. Available from: https://www.who.int/news-room/fact-sheets/detail/tobacco. Tobacco World Health Organisation 2018 2018:1–6.

[2] Office for National Statistics (ONS). Smoking prevalence in the UK and the impact of data collection changes - Office for National Statistics. ONS 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/drugusealcoholandsmoking/bulletins/smokingprevalenceintheukandtheimpactofdatacollectionchanges/2020 (accessed December 20, 2022).

[3] Health matters: stopping smoking - what works? [Internet]. GOV.UK. 2019. Available from: Https://www.gov.uk/government/publications/health-matters-stopping-smoking-what-works/health-matters-stopping-smoking-what-works. Gov.uk 1. 2019:1–17.

[4] Wagner EH, Groves T. Care for chronic diseases. BMJ 2002;325:913–4.

[5] Fact Sheets - Action on Smoking and Health [Internet]. Action on Smoking and Health. 2019 . Available from: Http://ash.org.uk/category/information-and-resources/fact-sheets/. Facts at a glance-key smoking statistics 2018:1–4.

[6] Lung cancer statistics [Internet]. Cancer Research UK. 2019 [cited 25 May 2019]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer. Lung Ca 2019:1–8.

[7] Lifestyle Statistics Team, Health and Social Care Information Centre. Statistics of Smoking. Nature 2018;181:1181–1181. doi:10.1038/1811181a0.

[8] Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ 2004;328:1519.

doi:10.1136/bmj.38142.554479.AE.

[9] What Are the Risk Factors for Lung Cancer? | CDC [Internet]. Cdc.gov. 2019 [cited 11 August 2019]. Available from: https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm. What Are the Risk Factors for Lung Cancer ? 2018;1:6348.

[10] Smoking and diabetes: How smoking causes type 2 diabetes. Available from: https://www.cdc.gov/tobacco/data_statistics/sgr/50th-anniversary/pdfs/fs_smoking_diabetes_508.pdf. Smoking and diabetes: How smoking causes type 2 diabetes n.d. www.smokefree.gov.

[11] Primatesta P, Falaschetti E, Gupta S, Marmot MG, Poulter NR. Association Between Smoking and Blood Pressure Evidence From the Health Survey for England Scientific Contributions. 2001.

[12] Sun L, Tan L, Yang F, Luo Y, Li X, Deng HW, et al. Meta-analysis suggests that smoking is associated with an increased risk of early natural menopause. Menopause 2012;19:126–32. doi:10.1097/gme.0b013e318224f9ac.

[13] Tengs TO, Osgood ND. The link between smoking and impotence: Two decades of evidence. Prev Med (Baltim) 2001;32:447–52. doi:10.1006/pmed.2001.0830.

[14] Adult smoking habits in the UK - Office for National Statistics [Internet]. Ons.gov.uk. 2019. Adult smoking habits in the UK: 2017. Am J Public Health 2017;76:1337–8. doi:10.2105/AJPH.76.11.1337.

[15] Windsor-Shellard B, Horton M, Scanlon S, Manders B. Office for National Statistics - Adult smoking habits in the UK: 2019. Stat Bull 2019:1–15.

[16] Burgess S, Thompson SG. Mendelian randomization: Methods for using genetic variants in causal estimation. Mendelian Randomization Methods

Using Genet. Var. Causal Estim., 2015, p. 1–207. doi:10.1201/b18084.

[17] Genest C. Correlation and Dependence. J Am Stat Assoc 2009;97:653–4. doi:10.1198/jasa.2002.s472.

[18] Khaw KT, Bingham S, Welch A, Luben R, Wareham N, Oakes S, et al. Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. European Prospective Investigation into Cancer and Nutrition. Lancet 2001;357:657–63.

[19] Collins R, Armitage J, Parish S, Sleight P, Peto R. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20 536 high-risk individuals: A randomised placebo-controlled trial. Lancet 2002;360:23–33. doi:10.1016/S0140-6736(02)09328-5.

[20] Peto R, Doll R, Buckley JD, Sporn MB. Can dietary beta-carotene materially reduce human cancer rates? Nature 1981;290:201–8. doi:10.1038/290201a0.

[21] Hennekens CH, Buring JE, Manson JE, Stampfer M, Rosner B, Cook NR, et al. Lack of Effect of Long-Term Supplementation with Beta Carotene on the Incidence of Malignant Neoplasms and Cardiovascular Disease. N Engl J Med 2002;334:1145–9. doi:10.1056/nejm199605023341801.

[22] Hooper L. Dietary fat intake and prevention of cardiovascular disease: systematic review. Bmj 2001;322:757–63. doi:10.1136/bmj.322.7289.757.

[23] Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. Jama 2002;288:321–33.

[24] Burgess S, Thompson SG. Avoiding bias from weak instruments in mendelian randomization studies. Int J Epidemiol 2011;40:755–64.

doi:10.1093/ije/dyr036.

[25] Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006;60:578–86. doi:10.1136/jech.2004.029496.

[26] Brigham J, Lessov-Schlaggar CN, Javitz HS, Krasnow RE, Tildesley E, Andrews J, et al. Validity of Recall of Tobacco Use in Two Prospective Cohorts. Am J Epidemiol 2010;172:828. doi:10.1093/AJE/KWQ179.

[27] Vogt W. Social Desirability Bias. Dict Stat Methodol 2015:10–1. doi:10.4135/9781412983907.n1826.

[28] Suresh K. An overview of randomization techniques: An unbiased assessment of outcome in clinical research. J Hum Reprod Sci 2011;4:8. doi:10.4103/0974-1208.82352.

[29] Sørensen HT, Lash TL, Rothman KJ. Beyond randomized controlled trials: A critical comparison of trials with nonrandomized studies. Hepatology 2006;44:1075–82. doi:10.1002/hep.21404.

[30] Hariton E, Locascio JJ. Randomised controlled trials-the gold standard for effectiveness research HHS Public Access. Bjog 2018;125:1716. doi:10.1111/1471-0528.15199.

[31] Gelman A. Benefits and limitations of randomized controlled trials. Soc Sci Med 2018:1–3. doi:10.1016/j.socscimed.2018.04.034.

[32] Spiegel J, Adhikari S, Balasubramanian S. The Structure and Function of DNA G-Quadruplexes. Trends Chem 2020;2:123–36. doi:10.1016/j.trechm.2019.07.002.

[33] Stram DO. Design, Analysis, and Interpretation of Genome-Wide Association Scans. 2014. doi:10.1007/978-1-4614-9443-0.

[34] Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS

Discovery. Am J Hum Genet 2012;90:7–24. doi:10.1016/j.ajhg.2011.11.029.

[35] Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res 2017;26:2333–55. doi:10.1177/0962280215597579.

[36] Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. Int J Epidemiol 2016. doi:10.1093/ije/dyw127.

[37] Taylor AE, Fluharty ME, Bjørngaard JH, Gabrielsen ME, Skorpen F, Marioni RE, et al. Investigating the possible causal association of smoking with depression and anxiety using Mendelian randomisation meta-analysis: The CARTA consortium. BMJ Open 2014;4. doi:10.1136/bmjopen-2014-006141.

[38] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med 2015. doi:10.1371/journal.pmed.1001779.

[39] Åsvold BO, Bjørngaard JH, Carslake D, Gabrielsen ME, Skorpen F, Davey Smith G, et al. Causal associations of tobacco smoking with cardiovascular risk factors: a Mendelian randomization analysis of the HUNT Study in Norway. Int J Epidemiol 2014;43:1458–70. doi:10.1093/ije/dyu113.

[40] Willett W, Hennekens CH, Castelli W, Rosner B, Evans D, Taylor J, et al. Effects of cigarette smoking on fasting triglyceride, total cholesterol, and HDL-cholesterol in women. Am Heart J 1983;105:417–21. doi:10.1016/0002-8703(83)90358-7.

[41] Saengdith P. Effects of Cigarette Smoking on Serum Lipids among Priests in Bangkok. J Med Assoc Thai 2008;91:41–5.

[42] Effects of smoking on the body | Smoke free [Internet]. Nhs.uk. 2019 [cited 26

May 2019]. Available from: https://www.nhs.uk/smokefree/why-quit/smoking-health-problems. How smoking affects your body 2017:6–9.

[43] Cournot A, Berlin I, Renout P, Duchier J, Safar M. Peripheral neurodynamic effects of smoking in habitual smokers. A methodological study. Eur J Pharmacol 1990;183:1649–50. doi:10.1016/0014-2999(90)91940-D.

[44] Ambrose JA, Barua RS. The pathophysiology of cigarette smoking and cardiovascular disease. J Am Coll Cardiol 2004;43:1731–7. doi:10.1016/j.jacc.2003.12.047.

[45] Harvard School of Public Health. The Greek Tobacco Epidemic. Center for Global Tobacco Control. Boston D 2011; available at: www. smokefreegreece. or. Center for Global Tobacco Control Mission 2019:3–5.

[46] Cheeseman KH, Slater TF. An introduction to free radical biochemistry. Br Med Bull 1993;49:481–93. doi:10.1093/oxfordjournals.bmb.a072625.

[47] Lobo V, Patil A, Phatak A, Chandra N. Free radicals, antioxidants and functional foods: Impact on human health. Pharmacogn Rev 2010;4:118–26. doi:10.4103/0973-7847.70902.

[48] PRYOR WA, STONE K. Oxidants in Cigarette Smoke Radicals, Hydrogen Peroxide, Peroxynitrate, and Peroxynitrite. Ann N Y Acad Sci 1993;686:12–27. doi:10.1111/j.1749-6632.1993.tb39148.x.

[49] Institute of Medicine (US) Committee on Secondhand Smoke Exposure and Acute Coronary Events., Source, 2010 W (DC): NAP (US); Secondhand Smoke Exposure and Cardiovascular Effects: Making Sense of the Evidence. Inst Med 2009:4.

[50] Bullen C. Impact of tobacco smoking and smoking cessation on cardiovascular risk and disease. Expert Rev Cardiovasc Ther 2008;6:883–95.

doi:10.1586/14779072.6.6.883.

[51] Gusarov I, Shatalin K, Starodubtseva M, Nudler E. Endogenous Nitric Oxide Protects Bacteria Against a Wide Spectrum of Antibiotics. Science (80-) 2009;325:1380–4. doi:10.1126/science.1175439.

[52] Coceani F. Carbon Monoxide in Vasoregulation. Circ Res 2000;86:1184–6. doi:10.1161/01.RES.86.12.1184.

[53] Rietbrock N, Kunkel S, Wörner W, Eyer P. Oxygen-dissociation kinetics in the blood of smokers and non-smokers: interaction between oxygen and carbon monoxide at the hemoglobin molecule. Naunyn Schmiedebergs Arch Pharmacol 1992;345:123–8. doi:10.1007/BF00175479.

[54] Turino GM. Effect of carbon monoxide on the cardiorespiratory system. Carbon monoxide toxicity: physiology and biochemistry. Circulation 1981;63:253A-259A.

[55] Astrup P, Kjeldsen K, Wanstrup J. Effects of Carbon Monoxide Exposure on the Arterial Walls. Ann N Y Acad Sci 1970;174:294–300. doi:10.1111/j.1749-6632.1970.tb49796.x.

[56] Zevin S, Saunders S, Gourlay SG, Jacob P, Benowitz NL. Cardiovascular effects of carbon monoxide and cigarette smoking. J Am Coll Cardiol 2001;38:1633–8. doi:10.1016/S0735-1097(01)01616-3.

[57] Lindell K. Weaver MD. Carbon Monoxide Poisoning - Carbon Monoxide Kills. N Engl J Med 2019;94:270–2.

[58] Article R, Miranda JJ, Corvalan C, Hyder AA, Lazo-porras M, Oni T. Understanding the rise of cardiometabolic diseases in low- a middle-income countries 2020:1–23.

[59] WHO. Cardio-Vascular Diseases. Lancet 2003;199:594–5. doi:10.1016/s0140-

6736(01)32941-0.

[60] George P, Mamali A, Papafloratos S, Zerva E. Effects of Smoking on Cardiovascular Function: The Role of Nicotine and Carbon Monoxide Institution of Athens (TEI-A), Greece 2. Physical Therapy Department, Technological Educational Institution of Athens (TEI-A), Greece 3. Physical Therapy Department, T. Heal Sci J 2014;8:274–90.

[61] Risk Factors: Smoking BHFA from: https://www. bhf. org. uk/informationsupport/risk-factors/smokin. [Accessed 14 A 2018]. Smoking British Heart Foundation 2016:2–3. http://www.cancer.ca/Canada-wide/Prevention/Smoking and tobacco/Why should I quit.aspx?sc_lang=en.

[62] U.S. Department of Health and Human Services. The Health Consequences of Smoking- 50 Years of Progress: A Report of the Surgeon General, Executive Summary. A Rep Surg Gen 2014:1–2.

[63] Stroke Associatioan. Smoking and the risk of Stroke. StrokeOrgUk 2017;144:540–4.

[64] Burns DM. Epidemiology of smoking-induced cardiovascular disease. Prog Cardiovasc Dis 2003;46:11–29. doi:10.1016/S0033-0620(03)00079-3.

[65] U.S. Department of Health and Human Services. How Tobacco Smoke Causes Disease. 2010.

[66] Yarnell JW. Smoking and cardiovascular disease. QJM 1996;89:493–8. doi:10.1186/1617-9625-3-29.

[67] Jacobs DR, Adachi H, Mulder I, Kromhout D, Menotti A, Nissinen A, et al. Cigarette smoking and mortality risk: twenty-five-year follow-up of the Seven Countries Study. Arch Intern Med 1999;159:733–40. doi:10.1001/ARCHINTE.159.7.733.

[68] Shah RS, Cole JW. Smoking and stroke: the more you smoke the more you stroke. Expert Rev Cardiovasc Ther 2010;8:917–32. doi:10.1586/erc.10.56.

[69] Hackshaw A, Morris JK, Boniface S, Tang J-L, Milenković D. Low cigarette consumption and risk of coronary heart disease and stroke: meta-analysis of 141 cohort studies in 55 study reports. BMJ 2018;360:j5855. doi:10.1136/bmj.j5855.

[70] Mcgauran N, Wieseler B, Kreis J, Schüler Y-B, Kölsch H, Kaiser T. Open Access REVIEW BioMed Central Reporting bias in medical research-a narrative review. vol. 11. 2010.

[71] Stallones RA. The association between tobacco smoking and coronary heart disease. Int J Epidemiol 2015;44:735–43. doi:10.1093/ije/dyv124.

[72] Law MR, Morris JK, Wald NJ. Environmental tobacco smoke exposure and ischaemic heart disease: an evaluation of the evidence. BMJ 2011;315:973–80. doi:10.1136/bmj.315.7114.973.

[73] Thun MJ, Carter BD, Feskanich D, Freedman ND, Prentice R, Lopez AD, et al. 50-Year Trends in Smoking-Related Mortality in the United States. N Engl J Med 2013;368:351–64. doi:10.1056/NEJMsa1211127.

[74] Huxley RR, Yatsuya H, Lutsey PL, Woodward M, Alonso A, Folsom AR. Impact of Age at Smoking Initiation, Dosage, and Time Since Quitting on Cardiovascular Disease in African Americans and Whites: The Atherosclerosis Risk in Communities Study. Am J Epidemiol 2012;175:816–26. doi:10.1093/aje/kwr391.

[75] Kurth T, Kase CS, Berger K, Schaeffner ES, Buring JE, Gaziano JM. Smoking and the risk of hemorrhagic stroke in men. Stroke 2003;34:1151–5. doi:10.1161/01.STR.0000065200.93070.32.

[76] Kurth T, Kase CS, Berger K, Gaziano JM, Cook NR, Buring JE. Smoking and Risk of Hemorrhagic Stroke in Women. Stroke 2003;34:2792–5. doi:10.1161/01.STR.0000100165.36466.95.

[77] Bhat VM, Cole JW, Sorkin JD, Wozniak MA, Malarcher AM, Giles WH, et al. Dose-Response Relationship Between Cigarette Smoking and Risk of Ischemic Stroke in Young Women. Stroke 2008;39:2439–43. doi:10.1161/STROKEAHA.107.510073.

[78] Fogelholm R, Murros K. Cigarette smoking and risk of primary intracerebral haemorrhage. Acta Neurol Scand 2009. doi:10.1111/j.1600-0404.1993.tb04119.x.

[79] Homan TD, Cichowski E. Physiology , Pulse Pressure. 2019.

[80] Zhou B, Bentham J, Di Cesare M, Bixby H, Danaei G, Cowan MJ, et al. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. Lancet 2017;389:37–55. doi:10.1016/S0140-6736(16)31919-5.

[81] Li G, Wang H, Wang K, Wang W, Dong F, Qian Y, et al. The association between smoking and blood pressure in men: a cross-sectional study. BMC Public Health 2017;17:797. doi:10.1186/s12889-017-4802-x.

[82] WHO. Global Health Observatory (GHO), Raised Blood Pressure. Situations Trends Availabe from Http//Www Who Int/Gho/Ncd/Risk_factors/Blood_pressure_prevalence_text/En 2014:39–41.

[83] National Institute for Health and Care Excellence (NICE). Health matters: combating high blood pressure - GOV.UK 2017:1–15.

[84] Leone A. Does Smoking Act as a Friend or Enemy of Blood Pressure? Let Release Pandora’s Box. Cardiol Res Pract 2011;2011:1–7.

doi:10.4061/2011/264894.

[85] Hughes K, Leong WP, Sothy SP, Lun KC, Yeo PPB. Relationships between cigarette smoking, blood pressure and serum lipids in the singapore general population. Int J Epidemiol 1993;22:637–43. doi:10.1093/ije/22.4.637.

[86] Gordon T, Kannel WB. Multiple risk functions for predicting coronary heart disease: The concept, accuracy, and application. Am Heart J 1982;103:1031–9. doi:10.1016/0002-8703(82)90567-1.

[87] Trap-Jensen J. Effects of smoking on the heart and peripheral circulation. Am Heart J 1988;115:263–7. doi:10.1016/0002-8703(88)90647-3.

[88] Liu X, Byrd JB. Cigarette Smoking and Subtypes of Uncontrolled Blood Pressure Among Diagnosed Hypertensive Patients: Paradoxical Associations and Implications. Am J Hypertens 2017;30:602–9. doi:10.1093/ajh/hpx014.

[89] Alomari MA, Al-Sheyab NA. Cigarette smoking lowers blood pressure in adolescents: The Irbid-TRY. Inhal Toxicol 2016;28:140–4. doi:10.3109/08958378.2016.1145769.

[90] Halperin RO, Michael Gaziano J, Sesso HD. Smoking and the risk of incident hypertension in middle-aged and older men. Am J Hypertens 2008;21:148–52. doi:10.1038/AJH.2007.36/2/M_AJH.148.T3.JPEG.

[91] McNagny SE, Ahluwalia JS, Scott Clark W, Resnicow KA. Cigarette Smoking and Severe Uncontrolled Hypertension in Inner-city African Americans. vol. 103. by Excerpta Medica, Inc; 1997.

[92] Al-Safi SA. Does smoking affect blood pressure and heart rate? Eur J Cardiovasc Nurs 2005;4:286–9. doi:10.1016/j.ejcnurse.2005.03.004.

[93] Diabetes [Internet]. Who.int. 2019 [cited 30 May 2019]. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes. Diabetes 30

2019:2019.

[94] Diabetes: the basics [Internet]. Diabetes UK. 2019 [cited 30 May 2019]. Available from: https://www.diabetes.org.uk/diabetes-the-basics. Diabetes : the basics 2019:3–5.

[95] Bao W, Michels KB, Tobias DK, Li S, Chavarro JE, Gaskins AJ, et al. Parental smoking during pregnancy and the risk of gestational diabetes in the daughter. Int J Epidemiol 2016;45:160–9. doi:10.1093/ije/dyv334.

[96] Diabetes UK. Diabetes Prevalence 2017 (November 2017). Diabetes UK 2017;2016:2016–7.

[97] Diabetes-resources-production.s3-eu-west-1.amazonaws.com. 2019 [cited 31 May 2019]. Available from: https://diabetes-resources-production.s3-eu-west-1.amazonaws.com/diabetes-storage/migration/pdf/DiabetesUK_Facts_Stats_Oct16.pdf. Diabetes UK. n.d.

[98] CDC. Smoking and diabetes: How smoking causes type 2 diabetes n.d. www.smokefree.gov (accessed May 29, 2019).

[99] Diabetes and Smoking [Internet]. Diabetes.co.uk. 2019 [cited 31 May 2019]. Available from: https://www.diabetes.co.uk/diabetes-and-smoking.html. Smoking and Diabetes UK 2019:1–6.

[100] Chang SA. Smoking and Type 2 Diabetes Mellitus. Diabetes Metab J 2012;36:399. doi:10.4093/dmj.2012.36.6.399.

[101] The British Diabetic Association. Diabetes statistics | Professionals | Diabetes UK. Prev Type 2 Diabetes 2020. https://www.diabetes.org.uk/professionals/position-statements-reports/statistics (accessed December 20, 2022).

[102] Lehrer S, Rheinstein PH. Diabetes, cigarette smoking and transcription factor

7-like 2 (Tcf7L2) in the UK Biobank cohort. Bull Acad Natl Med 2021;205:1146. doi:10.1016/J.BANM.2021.09.001.

[103] Fagard RH, Nilsson PM. Smoking and diabetes—The double health hazard! Prim Care Diabetes 2009;3:205–9. doi:10.1016/j.pcd.2009.09.003.

[104] Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. CLINICIAN ' S CORNER Active Smoking and the Risk of Type 2 Diabetes. Diabetes 2007;298:2654–64.

[105] Eliasson B. Cigarette smoking and diabetes. Prog Cardiovasc Dis 2003;45:405–13. doi:10.1053/pcad.2003.00103.

[106] Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, et al. Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes A bs tr ac t. vol. 359. 2008.

[107] Manson JAE, Ajani UA, Liu S, Nathan DM, Hennekens CH. A prospective study of cigarette smoking and the incidence of diabetes mellitus among US male physicians. Am J Med 2000;109:538–42. doi:10.1016/S0002-9343(00)00568-4.

[108] Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69:89–95. doi:10.1067/mcp.2001.113989.

[109] Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS 2010;5:463–6. doi:10.1097/COH.0b013e32833ed177.

[110] Gossett LK, Johnson HM, Piper ME, Fiore MC, Baker TB, Stein JH. Smoking intensity and lipoprotein abnormalities in active smokers. J Clin Lipidol 2009;3:372–8. doi:10.1016/j.jacl.2009.10.008.

[111] Rao Ch. S. The Effect of Chronic Tobacco Smoking and Chewing on the Lipid Profile. J Clin DIAGNOSTIC Res 2013;1:4–7. doi:10.7860/JCDR/2012/5086.2663.

[112] Tonstad S, Cowan JL. C-reactive protein as a predictor of disease in smokers and former smokers. Int J Clin Pract 2009;63:1634–41. doi:10.1111/j.1742-1241.2009.02179.x.

[113] SIMONS LA, SIMONS J, JONES AS. the Interactions of Body Weight, Age, Cigarette Smoking and Hormone Usage With Blood Pressure and Plasma Lipids in an Australian Community. Aust N Z J Med 1984;14:215–21. doi:10.1111/j.1445-5994.1984.tb03753.x.

[114] Pagán K, Hou J, Goldenberg RL, Cliver SP, Tamura T. Effect of smoking on serum concentrations of total homocysteine and B vitamins in mid-pregnancy. Clin Chim Acta 2001;306:103–9. doi:10.1016/S0009-8981(01)00402-8.

[115] Muscat JE, Harris RE, Haley NJ, Wynder EL. Cigarette smoking and plasma cholesterol. Am Heart J 1991;121:141–7. doi:10.1016/0002-8703(91)90967-M.

[116] Saengdith P. Effects of cigarette smoking on serum lipids among priests in Bangkok. J Med Assoc Thai 2008;91 Suppl 1:41–4.

[117] Jain RB, Ducatman A. Associations between smoking and lipid/lipoprotein concentrations among US adults aged ≥20 years. J Circ Biomarkers 2018;7:184945441877931. doi:10.1177/1849454418779310.

[118] Zhao X, Zhang HW, Zhang Y, Li S, Xu RX, Sun J, et al. Impact of Smoking Status on Lipoprotein Subfractions: Data from an Untreated Chinese Cohort. Biomed Environ Sci 2017;30:235–43. doi:10.3967/bes2017.033.

[119] Craig WY, Palomaki GE, Haddow JE. Cigarette smoking and serum lipid and lipoprotein concentrations: an analysis of published data. BMJ 1989;298:784–

8.

[120] Moradinazar M, Pasdar Y, Najafi F, Shahsavari S, Shakiba E, Hamzeh B, et al. Association between dyslipidemia and blood lipids concentration with smoking habits in the Kurdish population of Iran. BMC Public Health 2020;20:1–10. doi:10.1186/S12889-020-08809-Z/TABLES/2.

[121] Tobin MD, Minelli C, Burton PR, Thompson JR. Commentary: Development of Mendelian randomization: From hypothesis test to "Mendelian deconfounding." Int J Epidemiol 2004;33:26–9. doi:10.1093/ije/dyh016.

[122] Smith GD, Ebrahim S. "Mendelian randomization": Can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003. doi:10.1093/ije/dyg070.

[123] Christenfeld NJS, Sloan RP, Carroll D, Greenland S. Risk factors, confounding, and the illusion of statistical control. Psychosom Med 2004;66:868–75. doi:10.1097/01.psy.0000140008.70959.41.

[124] Davies NM, Holmes M V., Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 2018;362:k601. doi:10.1136/bmj.k601.

[125] Smith GD, Ebrahim S. Mendelian randomization: Prospects, potentials, and limitations. Int J Epidemiol 2004. doi:10.1093/ije/dyh132.

[126] Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on mendelian randomization: A comparison with randomized controlled trials. Am J Epidemiol 2006;163:397–403. doi:10.1093/aje/kwj062.

[127] Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. Sequence variants at CHRNB3–CHRNA6 and CYP2A6 affect smoking

behavior. Nat Genet 2010;42:448–53. doi:10.1038/ng.573.

[128] Wehby GL, Ohsfeldt RL, Murray JC. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. Stat Med 2008;27:2745–9. doi:10.1002/sim.3255.

[129] Thomas DC, Conti D V. Commentary: The concept of "Mendelian randomization." Int J Epidemiol 2004. doi:10.1093/ije/dyh048.

[130] Hill AB. The Environment and Disease: Association or Causation? J R Soc Med 1965. doi:10.1177/003591576505800503.

[131] Lousdal ML. An introduction to instrumental variable assumptions, validation and estimation. Emerg Themes Epidemiol 2018;15:1–7. doi:10.1186/s12982-018-0069-7.

[132] Brion M-JA, Benyamin B, Visscher PM, Smith GD. Beyond the Single SNP: Emerging Developments in Mendelian Randomization in the "Omics" Era. Curr Epidemiol Reports 2014;1:228–36. doi:10.1007/s40471-014-0024-2.

[133] Paaby AB, Rockman M V. The many faces of pleiotropy. Trends Genet 2013. doi:10.1016/j.tig.2012.10.010.

[134] WADDINGTON CH. CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS. Nature 1942. doi:10.1038/150563a0.

[135] Slatkin M. and Mapping the Medical Future. Nat Rev Genet 2016;9:477–85. doi:10.1038/nrg2361.Linkage.

[136] Corraini P, Olsen M, Pedersen L, Dekkers OM, Vandenbroucke JP. Effect modification, interaction and mediation: An overview of theoretical insights for clinical investigators. Clin Epidemiol 2017;9:331–8. doi:10.2147/CLEP.S129728.

[137] Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Int J Epidemiol 1986;33:9–9. doi:10.1093/ije/dyh312.

[138] Gray R, Wheatley K. How to avoid bias when comparing bone marrow transplantation with chemotherapy. Bone Marrow Transplant 1991.

[139] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008;26:1135–45. doi:10.1038/nbt1486.

[140] McPherson JD, Marra M, Hillier LD, Waterston RH, Chinwalla A, Wallis J, et al. A physical map of the human genome. Nature 2001. doi:10.1038/35057157.

[141] Roberts L, Davenport RJ, Pennisi E ME. A Brief History of the Human Genome Project 2001;291:11233436.

[142] Larsson SC, Burgess S. Appraising the causal role of smoking in multiple diseases: A systematic review and meta-analysis of Mendelian randomization studies. EBioMedicine 2022;82:104154. doi:10.1016/J.EBIOM.2022.104154.

[143] Burgess S, Daniel RM, Butterworth AS, Thompson SG. Network Mendelian randomization: Using genetic variants as instrumental variables to investigate mediation in causal pathways. Int J Epidemiol 2015;44. doi:10.1093/ije/dyu176.

[144] Timpson NJ, Lawlor DA, Harbord RM, Gaunt TR, Day IN, Palmer LJ, et al. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. Lancet (London, England) 2005;366:1954–9. doi:10.1016/S0140-6736(05)67786-0.

[145] Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 2012;380:572. doi:10.1016/S0140-6736(12)60312-2.

[146] von Hinke S, Davey Smith G, Lawlor DA, Propper C, Windmeijer F. Genetic markers as instrumental variables. J Health Econ 2016;45. doi:10.1016/j.jhealeco.2015.10.007.

[147] Chen L, Smith GD, Harbord RM, Lewis SJ. Alcohol Intake and Blood Pressure: A Systematic Review Implementing a Mendelian Randomization Approach. PLoS Med 2008;5:0461–71. doi:10.1371/JOURNAL.PMED.0050052.

[148] Wootton RE, Richmond RC, Stuijfzand BG, Lawn RB, Sallis HM, Taylor GMJ, et al. Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. Psychol Med 2020;50:2435. doi:10.1017/S0033291719002678.

[149] Walton ME. Major lipids, apolipoproteins, and risk of vascular disease. JAMA. 2009;302:1993–2000. Cell 2009;44:1–16. doi:10.1001/jama.2009.1619.Major.

[150] MacMahon S, Duffy S, Rodgers A, Tominaga S, Chambless L, De Backer G, et al. Blood cholesterol and vascular mortality by age, sex, and blood pressure: A meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. Lancet 2007;370:1829–39. doi:10.1016/S0140-6736(07)61778-4.

[151] Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. Lancet 2012;380:572–80. doi:10.1016/S0140-6736(12)60312-2.

[152] Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. J Am Med Assoc

1998;280:605–13. doi:10.1001/jama.280.7.605.

[153] Eisenberg DTA, Kuzawa CW, Hayes MG. Worldwide allele frequencies of the human apolipoprotein E gene: Climate, local adaptations, and evolutionary history. Am J Phys Anthropol 2010;143:100–11. doi:10.1002/ajpa.21298.

[154] Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, et al. Guidelines for performing Mendelian randomization investigations. Wellcome Open Res 2019;4:186. doi:10.12688/wellcomeopenres.15555.1.

[155] Burgess S, Thompson SG. Bias in causal estimates from Mendelian randomization studies with weak instruments 2010.

[156] Smith GD. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol 2004;33:30–42. doi:10.1093/ije/dyh132.

[157] Siedlinski M, Cho MH, Bakke P, Gulsvik A, Lomas DA, Anderson W, et al. Genome-wide association study of smoking behaviours in patients with COPD. Thorax 2011. doi:10.1136/thoraxjnl-2011-200154.

[158] Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol 2012. doi:10.1371/journal.pcbi.1002822.

[159] Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res 2018;27:e1608. doi:10.1002/mpr.1608.

[160] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature 2015;526:68–74. doi:10.1038/nature15393.

[161] Belmont JW, Boudreau A, Leal SM, Hardenbol P, Pasternak S, Wheeler DA, et al. A haplotype map of the human genome. Nature 2005;437:1299–320.

doi:10.1038/nature04226.

[162] Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, Ardissino D, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 2010;42:441–7. doi:10.1038/ng.571.

[163] Caporaso N, Gu F, Chatterjee N, Sheng-Chih J, Yu K, Yeager M, et al. Genome-Wide and Candidate Gene Association Study of Cigarette Smoking Behaviors. PLoS One 2009;4:e4653. doi:10.1371/journal.pone.0004653.

[164] Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 2010;42:436–40. doi:10.1038/ng.572.

[165] Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 2019;51:237–44. doi:10.1038/s41588-018-0307-5.

[166] Erzurumluoglu AM, Liu M, Jackson VE, Barnes DR, Datta G, Melbourne CA, et al. Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. Mol Psychiatry 2019. doi:10.1038/s41380-018-0313-0.

[167] Bierut LJ, Madden PAF, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, et al. Association study for nicotine dependence. Hum Mol Genet 2007;16:24–35. doi:10.1093/hmg/ddl441.

[168] Fowler CD, Lu Q, Johnson PM, Marks MJ, Kenny PJ. Habenular α5* nicotinic receptor signaling controls nicotine intake HHS Public Access. Nature 2011;471:597–601. doi:10.1038/nature09797.

[169] Munafò MR, Timofeeva MN, Morris RW, Prieto-Merino D, Sattar N, Brennan

P, et al. Association between genetic variants on chromosome 15q25 locus and objective measures of Tobacco exposure. J Natl Cancer Inst 2012;104:740–8. doi:10.1093/jnci/djs191.

[170] Linneberg A, Jacobsen RK, Skaaby T, Taylor AE, Fluharty ME, Jeppesen JL, et al. Effect of Smoking on Blood Pressure and Resting Heart Rate. Circ Cardiovasc Genet 2015;8:832–41. doi:10.1161/CIRCGENETICS.115.001225.

[171] Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638–42. doi:10.1038/nature06846.

[172] Lee YH. Assessing the causal association between smoking behavior and risk of gout using a Mendelian randomization study. Clin Rheumatol 2018;37:3099–105. doi:10.1007/s10067-018-4210-3.

[173] Johnsen MB, Winsvold BS, Børte S, Vie G, Pedersen LM, Storheim K, et al. The causal role of smoking on the risk of headache. A Mendelian randomization analysis in the HUNT study. Eur J Neurol 2018;25:1148-e102. doi:10.1111/ene.13675.

[174] Smith GD, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies n.d. doi:10.1093/hmg/ddu328.

[175] Davey Smith G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. Genes Nutr 2011;6:27–43. doi:10.1007/s12263-010-0181-y.

[176] Freathy RM, Kazeem GR, Morris RW, Johnson PCD, Paternoster L, Ebrahim S, et al. Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence body mass index. Int J Epidemiol 2011;40:1617–28. doi:10.1093/ije/dyr077.

[177] Linneberg A, Jacobsen RK, Skaaby T, Taylor AE, Fluharty ME, Jeppesen JL, et al. Effect of Smoking on Blood Pressure and Resting Heart Rate: A Mendelian Randomisation Meta-Analysis in the CARTA Consortium. Circ Cardiovasc Genet 2016;8:832–41. doi:10.1161/CIRCGENETICS.115.001225.Effect.

[178] Larsson SC, Burgess S, Michaëlsson K. Smoking and stroke: A mendelian randomization study. Ann Neurol 2019;86:468–71. doi:10.1002/ana.25534.

[179] Larsson SC, Mason AM, Bäck M, Klarin D, Damrauer SM, Michaëlsson K, et al. Genetic predisposition to smoking in relation to 14 cardiovascular diseases. Eur Heart J 2020;41:3304–10. doi:10.1093/EURHEARTJ/EHAA193.

[180] Levin MG, Klarin D, Assimes TL, Freiberg MS, Ingelsson E, Lynch J, et al. Genetics of Smoking and Risk of Atherosclerotic Cardiovascular Diseases. JAMA Netw Open 2021;4:e2034461. doi:10.1001/jamanetworkopen.2020.34461.

[181] Yuan S, Larsson SC. A causal relationship between cigarette smoking and type 2 diabetes mellitus: A Mendelian randomization study. Sci Rep 2019;9:1–4. doi:10.1038/s41598-019-56014-9.

[182] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018. doi:10.1038/s41586-018-0579-z.

[183] Hewitt J, Walters M, Padmanabhan S, Dawson J. Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease. BMJ Open 2016;6:e009161. doi:10.1136/bmjopen-2015.

[184] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature

2018;562:203–9. doi:10.1038/s41586-018-0579-z.

[185] Wain L V, Shrine N, Jackson VE, Ntalla I, Soler Artigas M, Allen R, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir 2015;3:769–81. doi:10.1016/S2213-2600(15)00283-0.

[186] The UK Biobank. UK Biobank Axiom Array Content Summary http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf (2014). UK Biobank Axiom Array 2014:1–7.

[187] UK Biobank. Genotyping and Quality Control of UK Biobank, a Large-Scale, Extensively Phenotyped Prospective Resource: Information for Researchers. Interim Data Release. 2015:1–27.

[188] Batty GD, Gale C, Kivimaki M, Deary I, Bell S. Generalisability of results from UK Biobank: comparison with a pooling of 18 cohort studies. MedRxiv 2019:19004705. doi:10.1101/19004705.

[189] Biobank U. About UK Biobank | UK Biobank [Internet]. Ukbiobank.ac.uk. 2020. Available from: http://www.ukbiobank.ac.uk/about-biobank-uk/ 2020:1–7.

[190] UK Biobank. UK Biobank biochemistry assay quality procedures 2019.

[191] Yousaf S, Bonsall A. UK Data Service Impact Ambassadors Workshop With The Department for Education Jointly organised by The UK Data Service and The Department for Education (DfE) UK Townsend Deprivation Scores from 2011 census data. 2017.

[192] Health Scotland N. The Scottish Burden of Disease Study (2016). 2018.

[193] World Health Organization - Global Database on Body Mass Index.

Apps.who.int. 2020. Available from: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html. WHO :BMI classification. Who 2004:2–3.

[194] Gingerich PD. Arithmetic or geometric normality of biological variation: An empirical test of theory. J Theor Biol 2000;204:201–21. doi:10.1006/jtbi.2000.2008.

[195] Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. vol. 5. 2012.

[196] Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. Int J Epidemiol 2016;45:908–15. doi:10.1093/ije/dyw127.

[197] Wooldridge JM. Introductory Econometrics. 2012.

[198] Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 2019;51:237–44. doi:10.1038/s41588-018-0307-5.

[199] McVean G. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 2009;5:e1000686. doi:10.1371/journal.pgen.1000686.

[200] Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003;361:598–604. doi:10.1016/S0140-6736(03)12520-2.

[201] Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol 2017. doi:10.1007/s10654-017-0255-x.

[202] Walker VM, Davies NM, Hemani G, Zheng J, Haycock PC, Gaunt TR, et al. Open Peer Review Using the MR-Base platform to investigate risk factors and

drug targets for thousands of phenotypes [version 2; peer review: 3 approved] 2019. doi:10.12688/wellcomeopenres.15334.1.

[203] Kivimäki M, Luukkonen R, Batty GD, Ferrie JE, Pentti J, Nyberg ST, et al. Body mass index and risk of dementia: Analysis of individual-level data from 1.3 million individuals. Alzheimer's Dement 2018;14:601. doi:10.1016/J.JALZ.2017.09.016.

[204] Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure n.d. doi:10.1101/2020.08.10.244293.

[205] Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-base platform supports systematic causal inference across the human phenome. Elife 2018;7:1–29. doi:10.7554/eLife.34408.

[206] Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression n.d. doi:10.1093/ije/dyv080.

[207] Taylor AE, Fluharty ME, Bjørngaard JH, Gabrielsen ME, Skorpen F, Marioni RE, et al. Investigating the possible causal association of smoking with depression and anxiety using Mendelian randomisation meta-analysis: The CARTA consortium. BMJ Open 2014;4. doi:10.1136/bmjopen-2014-006141.

[208] Yuan S, Larsson SC. A causal relationship between cigarette smoking and type 2 diabetes mellitus: A Mendelian randomization study. Sci Rep 2019;9:19342. doi:10.1038/s41598-019-56014-9.

[209] Bhargava S, de la Puente-Secades S, Schurgers L, Jankowski J. Lipids and lipoproteins in cardiovascular diseases: a classification. Trends Endocrinol Metab 2022;33:409–23. doi:10.1016/J.TEM.2022.02.001.

[210] Welsh C, Celis-Morales CA, Brown R, MacKay DF, Lewsey J, Mark PB, et al. Comparison of Conventional Lipoprotein Tests and Apolipoproteins in the Prediction of Cardiovascular Disease. Circulation 2019;140:542–52. doi:10.1161/CIRCULATIONAHA.119.041149.

[211] Soppert J, Lehrke M, Marx N, Jankowski J, Noels H. Lipoproteins and lipids in cardiovascular disease: from mechanistic insights to therapeutic targeting. Adv Drug Deliv Rev 2020;159:4–33. doi:10.1016/J.ADDR.2020.07.019.

[212] Holmes M V., Asselbergs FW, Palmer TM, Drenos F, Lanktree MB, Nelson CP, et al. Mendelian randomization of blood lipids for coronary heart disease. Eur Heart J 2015. doi:10.1093/eurheartj/eht571.

[213] Rosenson RS, Cannon CP. Patient education: High cholesterol and lipid treatment options (Beyond the Basics) 2021:1–9.

[214] Ostergaard SD, Mukherjee S, Sharp SJ, Proitsi P, Day F, Boehme KL, et al. Associations between potentially modifiable risk factors and Alzheimer disease: a Mendelian randomization study. Eur Neuropsychopharmacol 2017;27:S166-S167. doi:10.1016/j.euroneuro.2015.09.010.

[215] Østergaard SD, Mukherjee S, Sharp SJ, Proitsi P, Lotta LA, Day F, et al. Associations between Potentially Modifiable Risk Factors and Alzheimer Disease: A Mendelian Randomization Study. PLOS Med 2015;12:e1001841. doi:10.1371/journal.pmed.1001841.

[216] Shah RS, Cole JW. Smoking and stroke: the more you smoke the more you stroke. Expert Rev Cardiovasc Ther 2010;8:917–32. doi:10.1586/erc.10.56.

[217] DEBRA HAIRE-JOSHU, PHD RUSSELL E. GLASGOW, PHD TIFFANY L. TIBBS M. Smoking and diabetes Review 2016;22.

[218] Leone A. Smoking and Hypertension. J Cardiol Curr Res Smok Hypertens

n.d.;2. doi:10.15406/jccr.2015.02.00057.

[219] Yuan S, Larsson SC. A causal relationship between cigarette smoking and type 2 diabetes mellitus: A Mendelian randomization study. Sci Rep 2019;9:19342. doi:10.1038/s41598-019-56014-9.

[220] Aleksandrov AA, Rozanov VB, Kotova MB, Ivanova EI, Drapkina OM. Early smoking initiation and changes in body weight, blood pressure and lipid profile in males: results of a 26-year prospective study. Cardiovasc Ther Prev 2020;19:2610. doi:10.15829/1728-8800-2020-2610.

[221] Lyall DM, Quinn T, Lyall LM, Ward J, Anderson JJ, Smith DJ, et al. Quantifying bias in psychological and physical health in the UK Biobank imaging sub-sample n.d. doi:10.1093/braincomms/fcac119.

[222] Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. Lancet (London, England) 2019;393:1297. doi:10.1016/S0140-6736(18)33067-8.

[223] Gow AJ, Bastin ME, Maniega SM, Hernández MCV, Morris Z, Murray C, et al. Neuroprotective lifestyles and the aging brain. Neurology 2012;79:1802–8. doi:10.1212/WNL.0B013E3182703FD2.

[224] Lohse T, Rohrmann S, Bopp M, Faeh D. Heavy Smoking Is More Strongly Associated with General Unhealthy Lifestyle than Obesity and Underweight. PLoS One 2016;11:e0148563. doi:10.1371/JOURNAL.PONE.0148563.

[225] Powell JT. Vascular damage from smoking: disease mechanisms at the arterial wall. vol. 3. 1998.

[226] Xie X-T, Liu Q, Wu J, Wakui M. Impact of cigarette smoking in type 2 diabetes development. Acta Pharmacol Sin 2009;30:784–7. doi:10.1038/aps.2009.49.

[227] Heydari G, Heidari F, Yousefifard M, Hosseini M. Smoking and Diet in Healthy Adults: A Cross-Sectional Study in Tehran, Iran, 2010. Iran J Publ Heal 2014;43:485–91.

[228] Conway TL, Cronan TA. Smoking, exercise, and physical fitness. Prev Med (Baltim) 1992;21:723–34. doi:10.1016/0091-7435(92)90079-W.

[229] Prattala RS, Laaksonen MT, Rahkonen O. Smoking and unhealthy food habits How stable is the association? Eur J Public Health 1998;8:28–33.

[230] Begh R, Lindson-Hawley N, Aveyard P. Does reduced smoking if you can't stop make any difference? 2015. doi:10.1186/s12916-015-0505-2.

[231] Ali FRM, Agaku IT, Sharapova SR, Reimels EA, Homa DM. Onset of Regular Smoking Before Age 21 and Subsequent Nicotine Dependence and Cessation Behavior Among US Adult Smokers. Prev Chronic Dis 2020;17. doi:10.5888/PCD17.190176.

[232] Collins R. What makes UK Biobank special? Lancet 2012;379:1173–4. doi:10.1016/S0140-6736(12)60404-8.

[233] Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20:453–8. doi:10.1136/emj.20.5.453.

[234] David Batty G, Kivimäki M, Deary IJ, Bell S, Batty D. Generalisability of Results from UK Biobank: Comparison With a Pooling of 18 Cohort Studies n.d. doi:10.1101/19004705.

[235] Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. Am J Epidemiol 2017. doi:10.1093/aje/kwx246.

[236] Ilakovac V. Statistical hypothesis testing and some pitfalls. Biochem Medica

2009;19:10–6. doi:10.11613/BM.2009.002/FULLARTICLE.

[237] Williams J, Rakovac I, Loyola E, Sturua L, Maglakelidze N, Gamkrelidze A, et al. A comparison of self-reported to cotinine-detected smoking status among adults in Georgia. Eur J Public Health 2020;30:1007–12. doi:10.1093/EURPUB/CKAA093.

[238] Lee JY, Hong JH, Lee S, An S, Shin A, Park SK. Binary cutpoint and the combined effect of systolic and diastolic blood pressure on cardiovascular disease mortality: A community-based cohort study. PLoS One 2022;17:e0270510. doi:10.1371/JOURNAL.PONE.0270510.

[239] Katan MB. APOUPOPROTEIN E ISOFORMS, SERUM CHOLESTEROL, AND CANCER. Lancet 1986;327:507–8. doi:10.1016/S0140-6736(86)92972-7.

[240] Juvela S, Hillbom M, Numminen H, Koskinen P. Cigarette smoking and alcohol consumption as risk factors for aneurysmal subarachnoid hemorrhage. Stroke 1993;24:639–46. doi:10.1161/01.STR.24.5.639.

[241] Freitas SRS, Alvim RO. Smoking and Blood Pressure Phenotypes: New Perspective for an Old Problem. 554 Am J Hypertens 2017;30. doi:10.1093/ajh/hpx039.

[242] Perkins KA, Epstein LH, Marks BL, Stiller RL, Jacob RG. The Effect of Nicotine on Energy Expenditure during Light Physical Activity. N Engl J Med 2010;320:898–903. doi:10.1056/nejm198904063201404.

[243] Blood Pressure Levels for Boys by Age and Height Percentile. Nhlbi.nih.gov. 2019. Available from: https: www.nhlbi.nih.gov/files/docs/guidelines/child_tbl.pdf. Blood Pressure Levels for Boys by Age and Height Percentile. J Clin Gastroenterol 2001;33:289–94.

[244] Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 Guideline for High Blood Pressure in Adults - American College of Cardiology. J Am Coll Cardiol 2018;71:e127–248. doi:10.1016/j.jacc.2017.11.006.

[245] Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Stefansson K, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes HHS Public Access Author manuscript. Nat Genet 2018;50:524–37. doi:10.1038/s41588-018-0058-3.

[246] Mahajan A, Sarnowski C, Lecoeur C, Schurmann C, Genotyping APM, Mahajan PA. Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps Individual study design and principal investigators Europe PMC Funders Group 2018. doi:10.1038/s41588-018-0241-6.

8. Supplementary Materials

Chapter: Two (8.2)

Detailed Literature Review

Smoking and Stroke

Smoking is a major risk factor for all kinds of strokes. Current smokers have a two to four folds elevated risk of stroke compared to non-smokers [68]. In 2003, Kurth et al. published two papers on the risk of smoking on haemorrhagic stroke, one explored the risk in females and the other was in males [75,76]. In both studies, the researchers found that the risk of stroke increased compared to non-smokers. Males study was a prospective cohort with 22,022 US physician participants followed for 17.8 years [75]. They used self-report smoking status and the use of medical records for stroke. Stroke was defined as total haemorrhagic stroke, intracranial haemorrhage (ICH) and subarachnoid haemorrhage (SAH). Smoking status was categorised into 4 groups: never, current < 20 cigarettes per day (CperD) and currents with > 20 CperD. The females' paper was also a prospective cohort among 39,783 US participants followed over 9 years. They used the same criteria as the men's study regarding the smoking categories and the use of self-report for smoking status and records for stroke. The only exception was current smokers for which the cut point for CperD was 15. They discovered a higher risk for total haemorrhagic stroke, intracranial haemorrhage (ICH) and subarachnoid haemorrhage (SAH) in females smoking 15 or more cigarettes per day compared to non-smokers (RR = 3.29, 95% CI: 1.72 – 6.29, RR = 2.67, 95% CI: 1.04-6.90, RR = 4.02, 95% CI: 1.63 – 9.89, respectively) and in males smoking 20 or more cigarettes per day (RR = 2.36, 95% CI: 1.38 – 4.02, RR = 2.06, 95% CI: 1.08 – 3.96, RR = 3.22, 95% CI: 1.26 – 8.18, respectively) compared to non-smokers [75,76]. The increased risk of SAH seems to be linked to the elevated incidence of aneurysms seen among smokers. Heavy smoker (>20 CperD) men and current smoking females

have a relative risk of aneurysmal haemorrhage of 7.3 (95% CI: 3.8-14.3) and 2.1 (95% CI: 1.2-3.6, respectively, compared to non-smokers [240]. Such studies demonstrate a huge impact of smoking on stroke, especially haemorrhagic ones. The Kurth et al. studies sought only the health workers which might bias the results. The study did not include other risk factors of stroke which might play a major role in such a relationship, for example, cholesterol level, drugs, blood diseases, and hypertension.

The stroke risk increases as individual smokes more, such a dose-response relationship was observed in Bhat et al study [77]. This paper is a population-based case-control study of the risk factors of stroke in women aged 15 to 45 years. The data are from the Stroke Prevention in Young Women. It examines the relationship between cigarette smoking and ischemic stroke. The study comprises 466 cases and 604 controls. The researchers used 2 time periods, from 1992 – 1996 and 2001 – 2003. They included women, aged 15 – 49, with ischemic stroke identified by hospital diagnosis upon discharge. The controls were women with no history of stroke [77]. Smoking status was categorised as follows: never smokers, former smokers, and current smokers. Current smokers were further stratified into 4 categories: 1-10 CperD, 11-20 CperD, 21-39 CperD and >40 CperD. The researchers found that the adjusted OR for current smokers having a stroke was 2.6 (P<0,0001). The OR of having stroke increases as the number of cigarettes smoked per day increases by 2.2, 2.5, 4.3, and 9.1 in 1-10 CperD, 11-20 CperD, 21-39 CperD and >40 CperD, respectively [77]. The researchers also analysed the risk of stroke concerning the number of packs smoked per year. The OR was 2.1 (P<0.0004) for 1-10 packs/year, 2.7 (P<0.0001) for 11-20 packs/year, and 4.8 (P<0.0001) for >21 packs/year. These findings suggest a significant dose-response relationship between cigarette smoking and ischemic stroke [77]. Despite these strong findings, the study has some limitations. For instance, the smoking status was based on

self-report and not on objective measurements such as cotinine level. Recall bias might be a problem in patients with stroke. Additionally, all the variables included were based on self-reports including the medical assessment which might give rise to social desirability bias and reporting bias.

Fogelholm et al. revealed consistent findings with Bhat et al. [78]. Fogelholm et al. collected a total of 158 patients confirmed to have intracerebral haemorrhage (ICH) in the period between 1985 to 1989 in Finland. The diagnosis was confirmed by computed tomography (CT) and the smoking data based on self-report. 20% of the patients were labelled as current smokers. In this case-control study, the risk of stroke in the individuals who smoke 1 – 20 CperD was OR of 3.33 (95% CI: 1.05,10.6) compared to individuals who some > 21 CperD whom OR was 9.78 (95% CI: 2.25,42.5) [78]. The study has a small sample size, and the proportion of smokers is less than 20% of the sample. Additionally, patients with ICH may have memory problems which give rise to recall bias. Finally, the measurements of data collection are based on self-report, especially regarding smoking status and other risk factors.

Smoking and Hypertension

An article published in Cardiology and Research Practice has discussed the relationship between smoking and blood pressure [84]. The reviewer examined the findings on smoking and blood pressure in the literature across active and passive smokers. Some papers found that cigarette smoking in males was associated with a reduction in systolic blood pressure (SBP) by 1.3 mmHg (1.1%) in light smokers, 3.8 mmHg (3.1%) in moderate smokers, and 4.6 mmHg (3.7%) in heavy smokers compared to non-smokers [85]. The same findings have been obtained from Gordon et al. who demonstrated the higher the cigarettes smoked, the lower the blood pressure [86]. On the contrary, a trial conducted to evaluate the immediate effect of smoking on blood pressure found that

SBP and DBP increased by about 10% and 7%, respectively [87]. These changes in blood pressure are attributed to the toxic effects of nicotine and carbon monoxide and the structural damage to the endothelium of the arterial walls [84]. The article concluded that there is no correct answer to the question of the relationship between smoking and blood pressure as the studies continue to provide conflicting findings regarding this relationship.

This dilemma is also found in Silvia and Rafael's review [241]. Different epidemiological studies have reported that blood pressure in smokers is higher [92], lower [88] or the same [11] as the non-smokers. In the National Health and Nutrition Examination Survey (NHANES) program, Liu and Byrd reported that blood pressure is reduced among current smokers [88]. The cross-sectional study examined the relationship between smoking and uncontrolled blood pressure among hypertensive patients. The study has 7,829 adult participants of both genders aged 18 years and above. The participants included were 18 years or older, with a diagnosis of hypertension. The blood pressure is based on four readings by a trained physician using an arm sphygmomanometer. The diagnosis of hypertension is based on the reading of 140/90 mmHg or a previously confirmed diagnosis of blood pressure. The smoking status is defined as current, previous and non-smokers based on interviews. Among hypertensive participants, the current smokers and former smokers have lower DBP by 1.3 mmHg (95% CI: -2.8, -0.2, P=0.02) and 0.9 mmHg (95% CI: -1.7, -0.03, P=0.04), respectively, compared to non-smokers. The current smokers were 22% less to have uncontrolled BP (OR = 0.78, 95% CI: 0.64, 0.94, P<0.01) compared to non-smokers[88]. In this study, it seems that current smokers seem to have better control of their blood pressure compared to non-smokers, this reduction is theorised to be attributable to increase awareness of the current smokers regarding their health [242].

As a cross-sectional design study, it is hard to infer causality from such a relationship or even temporality between smoking and hypertension. Additionally, many confounding factors might play a major role in this relationship, especially diet, lipid profile, physical activity and medical history of diseases. Finally, smoking status based on self-report, and the duration of hypertension, antihypertensive drugs, and doses were not obtained which might prone the study to measurement bias.

Alomari and Al-Sheyab found consistent findings on the relationship between cigarette smoking and HTN, in which smoking is having an inverse effect on blood pressure [89]. This descriptive cross-sectional study examined the health adverse effect of smoking amongst adolescents (14 – 16 years). The data was gathered in 2015 from the Irbid tobacco Risk in Youth, Jordan. 244 healthy male participants were included with no history of medical conditions or using long-term medications. A self-reported questionnaire about smoking consumption was obtained for eligible participants and categorised as smokers and non-smokers (smokers: 90 vs non-smokers: 134). The blood pressure readings were measured using an automatic oscillatory method[89] and the average normal SBP/DBP for such age is 113/64 mmHg [243]. The results showed that the smokers were younger (P=0.001), less weight (P=0.001), and shorter (P=0.001) compared to non-smokers. The smoking status explained 20.6% of the changes in SBP (R2=0.206, F =46, P<0.001) and 5% of DBP (R2=0.05, F=9.4, P<0.003). ANCOVA was used to control for age, waist circumference, and BMI. The mean SBP and DBP in smokers were 108.8, 55.4, respectively (P<0.001) and the mean SBP and DBP among non-smokers were 118.5, 59.3, respectively (P<0.02). SBP and DBP both were lower in current smokers compared to non-smokers (P<0.05) [89]. The study was cross-sectional with good survey information but not temporality nor causation purposes. The students in general have a low blood pressure which was more prominent among

smokers. The study targeted children and hardly to have impact on adulthood smoking and HTN relationship. The sample size was small, and the self-reported smoking status is unreliable particularly from children. Finally, the confounding variables and reverse causation in a cross-sectional study are almost unavoidable.

A recently published cross-sectional study based on the China National Health Survey (CNHS) also suggested the same results of reduction of blood pressure with smoking. The researchers included 1248 healthy men aged 20 -80 years with self-reported smoking status and health conditions. The blood pressure was tested 3 times using a digital device by trained medical staff and defined as hypertensive (>140/90) or self-reported diagnosis [81]. The smoking status was classified as never, former and current smokers. Current smokers are further stratified into light smokers (0.025-5 packs/year), medium smokers (5-14 packs/year), heavy smokers (14-26 packs/year) and extreme smokers (> 26 packs/year). The study revealed that the ANCOVA of adjusted DBP and SBP were lower among current smokers compared to non-smokers ($P<0.05$). In comparison to never-smokers, the odds of having hypertension among current smokers seem to be protective, 13% decrease in blood pressure (OR = 0.83, 95% CI: 0.61, 1.12), although the result was statistically non-significant, the odds are higher among former smokers, 48% (OR = 1.48, 95% CI: 1.01, 2.18)[81]. There was no dose-dependent relationship between the number of packs/year and blood pressure. The study showed a consistent finding of an inverse relationship between smoking and blood pressure. The paper had some limitations, first, the study is cross-sectional giving good survey information, but no causal estimation can be drawn. Additionally, many confounders might magnify or nullify this relationship and even after statistical adjustment some residual or hidden confounders can bias the results, especially diet, physical activity, family history, and other medical conditions. Finally, the reading of

blood pressure might be misleading and such a diagnosis needs specific criteria like taking an average of more than 2 readings, on more than two occasions and the diagnosis should be confirmed only after the exclusion of secondary causes of hypertension, such as kidney diseases, systemic diseases like thyroid or adrenal glands or simply obesity [244]. Although many studies have reported the inverse or no effect of smoking on blood pressure, few studies have contradicted these findings.

A cross-sectional study was conducted at Grady Memorial Hospital in Atlanta, the US from March 1994 to August 1994 to examine the relationship between smoking and hypertension control among African Americans. The researchers included 216 individuals meeting the criteria of having hypertension based on the presence of antihypertensive medications in the individual's pharmacy chart, awareness of having hypertension as well as having ever taken antihypertensive medications [91]. They excluded non-English speakers or individuals having mental problems. After measuring the blood pressure using a sphygmomanometer, the patients were categorised as follows: controlled BP with SBP ≤140 mmHg and DBP ≤ 90 mmHg, severe HTN with SBP ≥ 180 and DBP ≥ 110 mmHg [91]. Patients with stage 1 or 2 HTN with SBP=141 to 179 mmHg or DBP=91 to 109 mmHg were excluded. The researchers also categorised the patients based on compliance with antihypertensive medications. After inclusion, demographic and health-related information was collected. The smoking status was defined as never, former, and current. The finding of this study suggested that current smokers have higher odds of being severe uncontrolled hypertensives as well as less compliant with medications (OR = 4.17, 95% CI: 1.8, 9.5, OR = 2.33, 95% CI: 1.3, 4.1, respectively) compared to former smokers. Additionally, both current and non-smokers were associated with uncontrolled HTN in compliant patients (OR = 14.4, 95% CI: 3.3, 63.3 and OR = 5.7, 95% CI: 1.5, 21.7, respectively) compared to former

smokers [91]. Surprisingly, in non-compliant patients, smoking status was not linked to uncontrolled hypertension. The finding of this study suggested that current smokers seem to have a higher risk of severe uncontrolled hypertension compared to former smokers. The study was conducted in a very poor disadvantaged neighbourhood with less education and a high poverty rate which makes the generalisability challenging as well as the confounding of such factors on smoking and hypertension. The diet and physical activity and weight were not recorded which might be responsible for the difference in the results especially when non-smokers have uncontrolled HTN despite compliance. Finally, in a cross-sectional approach, temporality, hence causality is hard to achieve.

One study had the same finding as the previous one of having higher blood pressure in smokers compared to non-smokers. Al-Safi and his students conducted a cross-sectional study to assess the correlation between smoking and blood pressure [92]. The study was conducted in 2004 on 14,310 healthy adults of both genders in Jordan. Males were 7400 and females were 6910. The smokers were 26.8 % and the non-smokers were 73.2%. The blood pressure was measured three times at 10 -15-minute intervals. Self-report demographic exploration was also collected, including, age, occupation, and education level. Previously diagnosed hypertension and CVDs were excluded. Blood pressure was measured three times at 10-15 minutes intervals using a sphygmomanometer. The smoking status is defined as smokers (with 1-10, 11-20, 21-30 and >31 cigarettes per day) and non-smokers. The study revealed that the SBP and DBP were significantly higher in male smokers compared to non-smokers (mean SBP: 126.24, 127.74, 129.67 and 129.11 in 1-10, 11-20, 21-30 and >31 cigarettes per day, respectively ($P<0.0001$), and mean DBP: 80.76, 80.97, 81.59 and 82.28 in 1-10, 11-20, 21-30 and >31 cigarettes per day, respectively ($P<0.0001$)). The mean SBP

and DBP of female smokers were significantly higher compared to non-smokers (P<0.001). There was a positive dose-effect of the correlation between smoking and hypertension as explained in the findings above. Despite these positive findings about smoking and blood pressure, the real factor in blood pressure was not smoking but family history [92]. When the comparison between smokers and non-smokers was conducted based on family history, the results were directed toward the family history in which the mean SBP was 120.99 (non-smokers + negative family history), 123,05 (non-smokers + positive family history, P<0.0001), 125.34 (smokers + negative family history), and 129.62 (smokers + positive family history, P<0.0001). Both findings did not meet the criteria of hypertension definition (SBP>140 mmHg and DBP>90 mmHg). The results could have been more informative if logistic regression was done to measure the relationship between smoking and blood pressure with ORs to be presented as a measure of association with further adjustment for confounders such as family history, BMI and others. The study was cross-sectional which provided good information about the prevalence of smoking and blood pressure among a large number, but it does not provide temporal relations or causation.

Smoking and Diabetes Mellitus

A significant number of studies have assessed the relationship between smoking and type 2 diabetes (DM), suggesting that cigarette smoking could independently interfere with glucose leading to impaired fasting glucose and DM, therefore smoking is believed to be a modifiable risk factor for DM [105].

A meta-analysis conducted by Willi et al. in 2007 examined the incidence of DM among smokers [104]. The researchers have conducted an extensive literature search targeting papers assessing this relationship between 1966 to 2007. They used these themes to define their search, glucose metabolism irregularity, smoking, and

prospective design studies [104]. The eligible articles should be prospective cohorts, with an adult population (≥16 years), with active smokers compared strictly to never smokers, and DM, impaired glucose tolerance or impaired fasting glucose as an outcome. They excluded studies with diabetic participants at the beginning of the study or studies with unfitting comparison groups [104]. They included 25 articles with more than 1.2 million study participants. The criteria to diagnose diabetes based on fasting blood glucose were according to WHO criteria in 1985 (≥140 mg/dL), WHO criteria in 1999 or American Diabetes Association criteria in 1997 (≥126 mg/dL) or other criteria in which the fasting blood glucose levels were ≥120 mg/dL or ≥110 mg/dL. The diagnosis of diabetes was based on biological screening (blood or urine tests), and personal or physician reports of diabetes [104]. With more than 45844 incident cases of DM in follow-up periods ranging from 5 to 30 years among 25 included studies, all except one, have found an association between active smokers and the risk of type 2 diabetes. The pooled crude relative risk of smoking on DM in all studies was 1.89 (95% CI: 1.58 -2.27). The adjusted RR ranges from 0.82 to 3.74. In a fully adjusted pooled RR, active smokers have a 44% increased risk of developing DM compared to non-smokers (RR = 1.44, 95% CI: 1.31 – 1.58)[104]. Further analysis of active smokers based on cigarettes smoked per day showed a dose-response relationship, heavy smokers (≥ 20 CperD) were found to have a 61% increased incidence of DM compared to lighter smokers (29%) and former smokers (23%) (RR = 1.61, 95% CI: 1.43 – 1.80, RR = 1.29, 95% CI: 1.13 – 1.48, RR = 1.23, 95% CI: 1.14 – 1.33, respectively) [104]. This vigorous meta-analysis of 25 studies showed a higher risk of DM among current smokers compared to never smokers with a dose-response relationship between smoking and DM [104]. The review is based on observational studies which makes it hard to confirm causality, whether because of confounders (diet, physical activity,

socioeconomic status, and secondary causes of DM) or reverse causation. They included old studies with a lack of information on the quality of participants and measures of recruitment. Finally, the criteria to diagnose diabetes were old with a higher threshold of diagnosis which might have missed many cases of diabetes.

Consistent findings of such a relationship between smoking and DM were also observed in a large prospective cohort study in Sweden [106]. The purpose of the study is to examine various predictors of DM with and without the inclusion of genetic factors. Smoking is one of these predictors and the one that will be concentrated on in this review [106]. The study encompasses two large prospective cohorts with a follow-up period of 23.5 years. The total number of participants in both cohorts was 18,831, among them DM developed in 2201 (11.7%). They used an oral glucose tolerance test to measure blood glucose and insulin. Fasting blood glucose was also measured. Plasma glucose was measured using hexokinase and glucose oxidase methods, while plasma insulin was measured by local radioimmunoassay and enzyme-linked immunosorbent assay [106]. The risk of DM was calculated at baseline clinical factors only and then clinical factors plus genetic factors. In baseline unadjusted clinical factors only, current smokers were having 30% higher odds of developing type 2 diabetes compared to non-smokers (OR = 1.30, 95% CI: 1.18-1.43). After adjustment, the risk has increased to 43% (OR = 1.43, 95% CI: 1.25 – 1.63, $P=1.4x10^{-9}$) [106]. When genetic factors were added to the clinical factors, current smokers had a 39% risk of developing DM (OR = 1.39 (95% CI: 1.29 – 1.61, $P=6.3X10^{-8}$) compared to non-smokers. This powerful study has concluded that smoking is a strong predictor risk factor for DM [106]. In general, observational studies can estimate the risk but it is hard to infer causality. Confounding variables such as physical activity, diet, family history, socioeconomic status and

secondary causes of DM might nullify or magnify the association between smoking and diabetes.

Another prospective study that examined the relationship between cigarette smoking and the incidence of DM has revealed the same findings as in the prior papers [107]. The participants were recruited from The Physicians Health Study. The included subjects are healthy US male physicians aged from 40 to 84 years and were followed for about 12 years. More than 21,068 eligible subjects were included while subjects with DM, CHD, stroke or cancer were excluded [107]. The information about health conditions, smoking status, and sociodemographic characteristics was collected by mailed questionnaires. Smoking status was categorised as never, past only and current. Current smokers further explored using how many cigarettes per day (≥ 20 CperD and < 20 CperD). The follow-up information was collected through mail questionnaires as well, and it has done two times per year in the first year, then thereafter once every year. The incidence of DM reported was 770 cases. Compared to non-smokers, the multivariate adjustment for BMI, physical activity, hypertension, and other risk factors, the relative risk of developing DM was 1.7 (95% CI: 1.3 to 2.3) for current smokers of ≥ 20 CperD, 1.5 (95% CI: 1.0 to 2.2) for current smokers of < 20 CperD, and 1.1 (95% CI: 1.0 to 1.4) for past smokers[107]. The researchers also assessed the association between packs/year of cigarette smoking and risk of DM and found a statistically significant association in smokers who smoke > 20 packs/year and non-significant results in less than 20 packs/year (1 – 19.9 pack/year: RR = 1, 95% CI: 0.8 – 1.3, 20 – 39.9 packs/year: RR = 1.3, 95% CI: 1 – 1.6, ≥ 40 packs/year: RR = 1.6, 95% CI: 1.3 – 2.1, P for trend <0.001) [107]. The study concluded that cigarette smoking is an independent modifiable risk factor of DM with a dose-effect phenomenon. The study was based on self-reports on both smoking status and diabetes which might have led to

reporting and recall bias, especially regarding the cigarettes per day and packs per year. There were no medical records on the health conditions and information was based only on participants' self-report. The information about the family history of diabetes was not available and such variable plays a major role in predicting diabetes. Finally, as an observational study, the causality is hard to be inferred and residual or hidden confounders might be a problem in this association.

Smoking and Lipid Biomarkers

A biomarker is an objective (quantifiable) tool that measures normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention [108]. The biomarkers can be chemical, physical, or biological. Examples of biomarkers include body temperature, blood pressure, pulse, and serum LDL to more advanced imaging and molecular tests of tissues and blood [109]. Cigarette smoking is believed to be associated with significant changes among some biomarkers which include High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), triglycerides (TG), and total cholesterol [40,110–112]. In the following review, the focus will be on the relationship between smoking and the aforementioned markers.

Cigarette smoking is associated with an increase in triglycerides (TG), LDL and cholesterol levels, and a reduction in HDL. Smoking seems to affect lipids through nicotine which increases the secretion of free fatty acids and triglycerides along with lipoproteins from the liver into the bloodstream. This mechanism is enhanced by the stimulatory effect of nicotine on catecholamines (epinephrine and norepinephrine) secretion which lead to sympathetic stimulation resulting in increased lipolysis (the breakdown of fat) [113]. Cigarette smoking is also associated with an increased level of Homocysteine level which promotes the oxidative alteration of LDL and decreases

HDL [114]. A review of the relationship between smoking and lipid profile will be discussed in the following sections.

Gossett et al. have explored the effect of smoking on lipoprotein concentrations (LDL, HDL), total cholesterol and TG among current smokers in a prospective cohort study from 2005 – 2007 [110]. 1,504 subjects of male and female participants were obtained from the longitudinal, randomized, double-blind, placebo-controlled trial. The participants should be >18 years old and currently smoking. On average, they smoked 21.4 cigarettes per day. The study revealed that the HDL (42 mg/dL) and HDL particles (30.3 μmol/L) were low among current smokers. Cigarettes smoked per day (CperD) predicted higher total cholesterol (P=0.009), LDL (P=0.02) and total triglycerides (P=0.002) [110]. They concluded that current smokers have a higher level of total cholesterol, LDL and TG and lower HDL levels compared to non-smokers. The smoking status, medical history, and medical conditions were based on self-report which might bias these results. In addition, confounding factors such as diet, socioeconomic status and other causes of hyperlipidaemia might play a major role in this relationship.

Zhang et al. study had close findings to the previous study in which they found consistent results of the effect of smoking on lipid profile [118]. The researchers evaluated the effect of smoking on lipoprotein subfractions among current smokers compared to former and non-smokers. This cross-sectional study recruited 877 eligible Chinese participants, aged > 18 with angina-like chest pain, and excluded patients less than 18 years, using statins or any lipid-lowering medications, or having any end-stage medical condition, severe infection, thyroid disorder or confirmed pregnancy [118]. The smoking history is based on self-report, current, former and non-smokers. The lipid parameters (LDL, HDL, TG, and total cholesterol) were collected and examined using

electrophoretic technology. The study found that the current smokers had a significant reduction in mean (± SD) HDL, adjusted for gender, age, BMI, alcohol consumption and family history of CVD, HTN and DM, compared to non-smokers and former smokers (1.01 ± 0.26 vs. 1.06 ± 0.32 vs. 1.17 ± 0.36 mmol/L, P<0.001, after adjusted P=0.006, respectively). The mean (± SD) of LDL was highest among current smokers compared to former smokers and non-smokers but was statistically non-significant (3.19 ± 0.87, 3.12 ± 0.88, 3.18 ± 0.87 mmol/L, P=0.707, adjusted P=0.554, respectively)[118]. The mean (± SD) of TG among current smokers was higher compared to former and non-smokers but was statistically non-significant after adjustment (1.82 ± 0.81, 1.64 ± 0.68, 1.64 ± 0.79 mmol/L, P=0.002, adjusted P=0.09, respectively). The mean (± SD) of total cholesterol was highest among non-smokers compared to current smokers and former smokers but was statistically non-significant after adjustment (4.8 ± 0.92, 4.70 ± 0.87, 4.65 ± 0.95 mmol/L, P=0.046, adjusted P=0.554, respectively) [118]. The study concluded that smoking is associated with a low level of HDL, and a higher level of LDL. In this cross-sectional study, the smoking status, medical conditions, and family history were self-reported which prone the study to recall and reporting bias. The causality is hard to be inferred from such a design, in addition to confounding variables that might interfere with lipid parameters.

In a large screening of plasma cholesterol among 51,723 US participants in 1988, Muscat et al. found that plasma cholesterol increased among current smokers with a dose-response relationship between CperD and plasma cholesterol [115]. Among men and women aged 18 to 60 years, the plasma cholesterol levels raised by 0.33 mg/dL (P<0.001) and 0.48 mg/dL (P<0.001) for each cigarette smoked compared to non-smokers and ex-smokers. There was no observed association between smoking and plasma cholesterol over age 60. The study was a screening with no temporality nor

causation to be drawn from this association. The diet, socioeconomic status, education level, and physical activity are major confounders that might bias these findings.

These findings of the association between smoking and increased level of TG and total cholesterol and decreased the level of HDL were also observed in many studies such as Willett et al [40]. This small survey which included 191 women found the adjusted mean difference of TG and cholesterol is higher among current smokers compared to non-smokers (adjusted difference: 49.5 and 7.9, P<0.005, respectively) [40]. The adjusted mean difference of HDL was lower among current smokers compared to non-smokers (- 7.3, P<0.005). The study was a survey among a small sample, in which only women have been screened and self-reported data were obtained which could bias the results and limit the power of the study, with increased variability and limited generalizability.

The final paper to consider here is the review of 54 published studies conducted by Craig et al. which examined the association between cigarette smoking in adults and serum lipid and lipoprotein concentrations [119]. They found that current smokers had higher cholesterol, TG, LDL, and lower HDL compared to non-smokers. Among current smokers, the serum concentration of cholesterol, TG and LDL were 3% (P<0.001), 9.1% (P<0.001) and 1.7% (P<0.001), respectively, higher than non-smokers. There was a dose-response relationship between light, moderate, and heavy smokers, compared to non-smokers and serum concentrations of lipids and lipoproteins. The percentage difference from non-smokers increased as the CperD (light, moderate and heavy) increased (cholesterol: 1.8, 4.3, and 4.5%, respectively, TG: 10.7, 11.5, and 18%, LDL: -1.1, 1.4, (P for trend <0.001) and 11%, HDL: -4.6, -6.3, and -8.9% (P<0.001) [119]. This robust review showed a significant association between smoking and lipid serum concentrations and lipoproteins, with a significant dose-response

relationship, in which the higher the individual smoked, the higher TG, LDL, and total cholesterol, and the lower HDL. The review did not account for confounders like diet, physical activity, previous medical conditions, and others that might affect this relationship. Additionally, the results could have been more informative if beta coefficients were reported to quantify the relationship between smoking and lipid profile with further adjustment for confounders.

MR Review

A Mendelian randomization analysis was conducted in Norway and examined the association of rs1051730 T alleles with cardiovascular risk factors [39]. The researchers included 56,625 participants aged 20 years or above whom were interviewed about smoking habits and health conditions as well as underwent clinical and laboratory examinations evaluating CVD risk factors (BMI, blood pressure, HDL and glucose) [39]. The rs1051730 polymorphism was successfully genotyped at HUNT Biobank for 56,664 and 56,625 (99.9%) reported the smoking status, thus included in the study. The researchers found that rs1051730 T alleles carriers have higher CperD as well as more likely to be current smokers and slightly younger compared to never and former smokers. They revealed that an additional rs1051730 allele was associated with a 0.27 mmHg (95% CI: 0.04, 0.49) lower systolic BP among the total study population (P-value <0.02) but no association was found with diastolic BP [39]. Moreover, a 0.34% (95% CI: 0.02, 0.66) higher concentration of HDL was observed across the total study population with each additional rs1051730 T allele but not among smoking subcategories, for example, current smokers (0.37%, P=0.2). Furthermore, among current smokers, the rs1051730 T allele was associated with 1.16% (95% CI: 0.03, 2.28) lower triglyceride concentration which was attenuated after adjustment for BMI (0.03%, P=96). Finally, there was no convincing association was seen between

rs1051730 and glucose level or total cholesterol among current smokers. Further analysis of current smokers based on CperD, the researchers found that the association between rs1051730 T alleles and cardiovascular risk factors was similar among light and heavy smokers [39]. They concluded that smoking is not a major determinant of blood pressure, serum lipid or glucose level. Although the large sample size, the non-fasting sampling might bias the results of the biomarkers. They stated that the rs1051730 might have influenced the outcome via other routes and not only through smoking and that will violate one of the IV assumptions (positive association between rs1051730 and BMI).

Similar findings regarding HTN and smoking were seen in a meta-analysis conducted by Lnneberg et al [177]. Data were collected on 141,317 participants from self-reported European ancestry aged > 16 years with more than 37,982 current smokers. The researchers categorised smokers as never, former, and current smokers, and smoking intensity among current smokers was analysed based on cigarettes smoked per day (CperD). The HTN diagnosis was based on the SBP>140 and DBP>90 or taking antihypertensive medications. They used rs16969968 or rs1051730 as a proxy for smoking intensity [177]. An additive genetic model was assumed, in which the difference in the risk of outcome per each additional copy of the risk allele characterizes the risk estimates. The minor allele frequency (MAF) for both SNPs ranged between 0.29 and 0.36. The researchers found that the beta estimate per minor allele of rs1051730/rs16969968 with systolic blood pressure (SBP) and diastolic blood pressure (DBP) was close to null with overlapping CI (-0.20, 95% CI: -0.46, 0.06 for SBP, P=0.136, and -0.15 95% CI: -0.32, 0.02, P=0.079 for DBP) among current smokers [177]. They concluded that there is no causal association between smoking and blood pressure. The HTN was based on the self-report and not established clinical diagnosis

which might prone this analysis to misclassification bias. There was no data provided on the validity of the SNPs used in the analysis.

A Mendelian randomization study was conducted by Larsson, Burgess, and Michaëlsson to assess the causal association between smoking and stroke [178]. The study uses summary statistics data from the MEGASTROKE consortium for 438,847 individuals of European ancestry [245]. It included 34,217 patients with ischemic stroke (404,630 control). The study also included an analysis of intracerebral haemorrhage, but the review will include only ischemic stroke. The smoking behaviour of interest was smoking initiation. A genome-wide association meta-analysis of 1,232,091 individuals identified 372 SNPs associated with smoking initiation and included them in the analysis of ischemic stroke [165]. These SNPs explain 2.3% of the variation in smoking initiation. The statistical analyses were performed using *mrrobust* and *MendelianRandomization* packages. The researchers found a statistically significant positive association between the genetic predisposition of smoking initiation and ischemic stroke. A unit increase in log odds of smoking initiation was associated with odds ratios of 1.22 (95% CI: 1.12 – 1.34, p-value = 7.6×10^{-6}). There was no indication of horizontal pleiotropy (all P>0.24). The study found no significant association between smoking intensity (CperD) and ischemic stroke. The study concluded that there is a causal association between smoking initiation and increased risk of stroke. The study is based on a large sample size for stroke which gives a high power to detect any weak associations. It included a robust genetic instrument using 372 SNPs for smoking initiation. The two-sample MR provides a high-power analysis but no access to individual data. The instrumental variables from the sample might not represent the population where the sample was obtained which might doubt the validity

of the used instruments. Finally, further analysis will not be possible from such an approach.

A two-sample MR was conducted recently by Larsson and Yuan to examine the causal association between smoking and type 2 diabetes [181]. The study uses publicly available summary-level data (beta coefficients and standard errors). The smoking behaviour of interest in this study was smoking initiation. The summary-level data on diabetes were obtained from GWAS of 32 studies ((DIAbetes Genetics Replication and Meta-analysis consortium) [246]. These GWAS included 898,130 individuals (47,124 cases and 824,006 controls) of European ancestry. The instrumental variable data for smoking initiation is based on a published meta-analysis of GWASs which included 1,232,091 individuals of European ancestry [198]. Up to 378 SNPs were associated with smoking initiation at the GWAS significance level. All SNPs (except one) were available in the type 2 diabetes dataset. The study found a positive significant association between genetic-based smoking initiation and type 2 diabetes. The odds ratios of DM were 1.28 (95% CI: 1.20-1.37, $P=2.35 \times 10^{-12}$). The study concluded that smoking initiation is causally associated with an increased risk of type 2 diabetes. The researchers reported a large overlap between the participants in the datasets of DM and smoking initiation, this might lead to bias in the estimation of the observational association.

The use of Mendelian randomization for smoking behaviour has been pronounced in the literature. It is an unconfounded measure of smoking exposure because these genetic variants, that act as a proxy (IV) for smoking, have been determined during gamete formation and conception. Therefore, these alleles associated with smoking are paired randomly from parents to offspring and are not likely attributable to environmental factors which prone the conventional studies of

smoking to confounders [174,175]. If smoking is associated with any outcome then genetic variants predicting smoking behaviour should be associated also with these outcomes attributable to smoking among current smokers but not never smokers (as their GVs are not associated with smoking intensity) [175,176].

In conclusion, smoking behaviour has been tested for causality in the literature using Mendelian randomization, however, there is no study has examined all outcomes that this study proposed, especially in the UKB. The wide range of outcomes proposed in this thesis about smoking behaviour, especially in a very large number of participants in the UKB, make it unique and worth reading in the scientific community.

Chapter: Four (8.4)

Methods

Logistic regression assumptions

The assumptions for logistic regression were met before proceeding to the analysis. These assumptions are:

1] Cases are randomly sampled

2] Binary DV

3] No outliers [bivariate and multivariate]

4] Associations between continuous predictors and logit DV are linear

5] No multicollinearity

The first assumption was met based on the nature of the UKB sampling assuming random sampling techniques. The second assumption was met as the CMD variables are binary. The rest of the assumptions were tested in R using *lessR package*. The outliers were tested using Cooks distance which revealed no potential outliers. Additionally, the linear association between the quantitative variables (Age, BMI and deprivation score) and logit CMD variables were tested using the Box-Tidwell test as well as visually using scatterplot. The linearity assumption was also met visually and using a correlation coefficient, however not all variables were linearly associated with the logit transformation of the outcomes when using Box-Tidwell ($P>0.05$). Finally, there was no multicollinearity (high correlation) between the predictors. The analysis was done using variance inflation fact (VIF) in *car package* in R. The VIF score for all variables was around 1, suggesting a low correlation among the predictors. Examination of regression assumptions was performed for all CMD variables.

Sample characteristics

..1.2.45 Observational sample

The average BMI of the participants was 27.42 (SD = ±4.79) which lies in an overweight zone, according to WHO BMI categorisation. WHO has classified the BMI as underweight (<18.5), normal/healthy weight (18.5 – 24.9), overweight (25 – 29.9) and obese (30 and above) [193]. When categorising the BMI variable in this sample, the participants were 0.5% (underweight), 31.6% (healthy weight), 42.9% (overweight) and 25% (obese). This makes almost 68% of the participants either overweight or obese. According to Townsend's score, the participants were not materially deprived (mean = -1.31 ±3.08, minimum = - 6.26, maximum = 11), and higher scores = more deprivation).

..1.2.46 MR sample

For MR data, individuals who descended from European ancestry and have the genotyped SNPs for CperD and smoking status were included. Because the MR approach is based on genetic analysis, the sample characteristics such as age, sex and other covariates will not be discussed in detail. A summary of the MR sample characteristics is shown in Table 8.1.

Table 8.1. Sample characteristics-MR (n=25274)

Variable	Level	Count (%)
Sex	Male	12155 (48.1%)
	Female	13119 (51.9%)
Degree (college/university)	No Degree	21059 (83.3%)
	Degree	4215 (16.7%)
Coronary Heart Disease (CHD)	No	23865 (94.4%)
	Yes	1409 (5.6%)
Stroke	No	25108 (99.3%)
	Yes	166 (0.7%)
Hypertension (HTN)	No	20028 (79.2%)
	Yes	5246 (20.8%)
Diabetes Mellitus (DM)	No	24018 (95%)
	Yes	1256 (5%)
Variable	**Mean (SD)**	
Cigarette Smoked per Day (CperD)	15.71 (±8.35)	
Age	54.81 (±8.05)	
Body Mass Index (BMI)	26.83 (±4.85)	
Deprivation Level (Townsend score)	0.18 (±3.48)	

Descriptive statistics

CPD and SI

Table 8.2: Summary statistics for CperD and SI

Variable	Minimum	Maximum	Mean	SD	Median	Q1	Q3
CperD	1	35	14.85	7.11	15	10	20
SI	10	25	16.7	3.12	16	15	18

CMDs

Table 8.3: Frequencies and percentages of CMDs

Variable	Levels	Count (%)
Coronary Heart Disease (CHD)	No	474003 (95.5%)
	Yes	22589 (4.5%)
Stroke	No	490474 (98.77%)
	Yes	6124 (1.23%)
Hypertension (HTN)	No	377608 (76.04%)
	Yes	118990 (23.96%)
Diabetes Mellitus (DM)	No	470516 (94.75%)
	Yes	26082 (5.25%)

Smoking behaviour and CMDs

This section will explore the associations between smoking and CMDs descriptively. This will include smoking status, CperD and SI variables vs CHD, stroke, HTN and DM.

..1.2.47 Smoking variables vs CHD and stroke

Among all individuals having CHD, 13.1% were current smokers, 55.2% were previous smokers and 36.7% were never smokers. The previous smokers have the highest number of cases of CHD 11340 (6.6%) compared to the current 2961 (5.6%) and never smokers 8288 (3%). When categorising smoking status into ever vs never, approximately 68.3% of all CHD cases are either current or have smoked in the past. For CperD, individuals with CHD have, on average, a higher number of cigarettes smoked per day compared to individuals who do not have CHD (17.57 ± 9.21, 15.4 ± 8.32, respectively). It seems that individuals with CHD started to smoke earlier in life compared to CHD-free individuals (16.64 ± 5.13, 17.93 ± 5.83, respectively).

Of individuals who reported stroke, 15.1% were current, 41.3% were previous and 43.5% were never smokers. Stroke is more prevalent among current smokers (1.8%) compared to previous (1.5%) and never-smokers (1%). When categorising

smoking status into ever vs never, approximately 56.4 % of all stroke cases are either current or have smoked in the past. For CperD, stroke patients have on average higher cigarettes smoked per day compared to individuals not diagnosed with stroke (17.22 ± 9, 15.5 ± 8.38, respectively). The stroke patients started to smoke earlier in life compared to individuals who had no stroke (16.76 ± 5.09, 17.83 ± 5.81, respectively). Table 8.4 and Figure 8.1 summarise and visualise these associations.

Table 8.4: Descriptive analysis of smoking variables vs CHD and stroke

Variable		**CHD**		**Stroke**	
Smoking status	Current	**2961** [5.6%] of all current [13.1%] of all CHD cases		**926** [1.8%] of all current [15.1%%] of all stroke cases	
	Previous	**11340** [6.6%] of all previous [50.2%] of all CHD cases		**2532** [1.5%] of all previous [41.3%] of all stroke cases	
	Never	**8288** [3%] of all never [36.7%] of all CHD cases		**2666** [1%] of all never [43.5%] of all stroke cases	
		Mean (SD)		**Mean (SD)**	
		Have CHD	No CHD	Have Stroke	No Stroke
CperD		17.57 (±9.21)	15.4 (±8.32)	17.22 (±9)	15.5 (±8.38)
SI		16.64 (±5.13)	17.93 (±5.83)	16.76 (±5.09)	17.83 (±5.81)

Figure 8.1: Prevalence of CHD and stroke across smoking categories

..1.2.48 Smoking variables vs HTN

The prevalence of hypertension is high in this sample (~24%). Out of those individuals, 9.2% were current smokers, 38.6% were previous smokers and 52.3% were never smokers. Previous smokers have the highest prevalence of HTN (26.7%) compared to current (20.8%) and never-smokers (22.9%). When categorising smoking status into ever vs never, approximately 47.8% of all HTN cases are either current or have smoked in the past. For CperD, individuals with HTN have, on average, a higher number of cigarettes smoked per day compared to individuals who do not have HTN (16.27 ± 8.69, 15.34 ± 8.3, respectively). It seems that individuals with no HTN started to smoke

relatively earlier in life compared to individuals with HTN (17.79 ± 5.64, 18.09 ± 6.35, respectively). Table 8.5 and Figure 8.2 summarise and visualise these associations.

Table 8.5: Descriptive analysis of smoking variables vs HTN

Variable		HTN
Smoking status	Current	**10898** [20.8%] of all current [9.2%] of all HTN cases
	Previous	**45912** [26.7%] of all previous [38.6%] of all HTN cases
	Never	**62180** [22.9%] of all never [52.3%] of all HTN cases
		Mean (SD)
		Have HTN / No HTN
CperD		16.27 (±8.69) / 15.34 (±8.3)
SI		18.09 (±6.35) / 17.79 (±5.64)

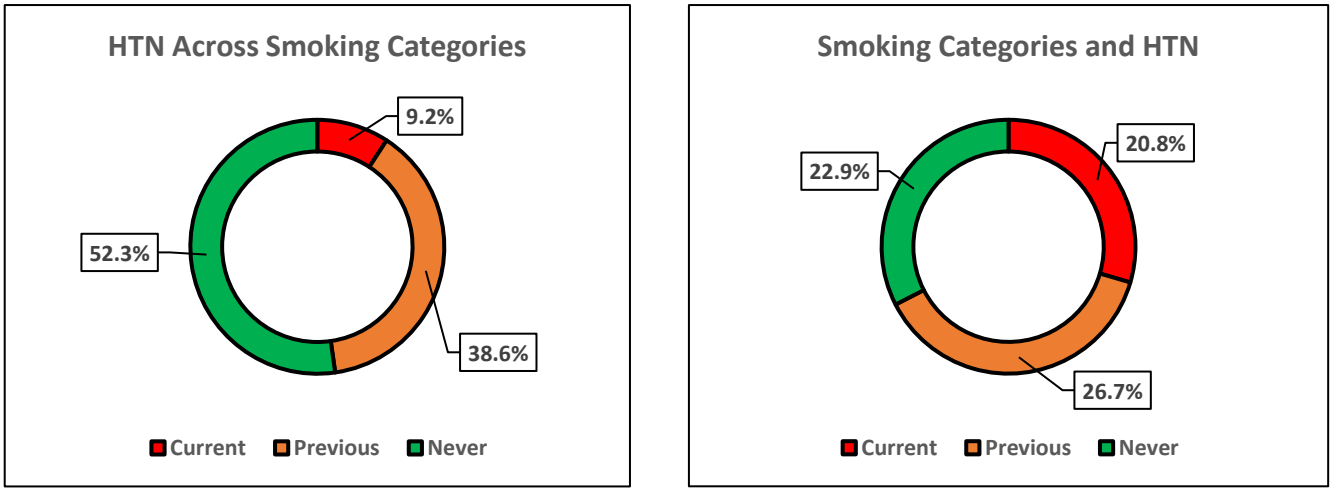


Figure 8.2: Prevalence of HTN across smoking categories

..1.2.49 Smoking variables vs DM

Among all individuals having DM, 11.1% were current smokers, 42.8% were previous smokers and 46.1% were never smokers. Previous smokers have the highest prevalence of DM (6.5%) compared to current (5.5%) and never-smokers (4.4%). When categorising smoking status into ever vs never, approximately 54% of all DM cases are either current or have smoked in the past. For CperD, individuals with DM have, on average, a higher number of cigarettes smoked per day compared to individuals who do not have DM (17.7 ± 9.56, 15.4 ± 8.3, respectively). It seems that individuals with DM started to smoke earlier in life compared to DM-free individuals (17.63 ± 6.21, 17.86 ± 5.77, respectively). Table 8.6 and Figure 8.3 summarise and visualise these associations.

Table 8.6: Descriptive analysis of smoking variables vs DM

Variable		DM
Smoking status	Current	**2895** [5.5%] of all current [11.1%] of all DM cases
	Previous	**11170** [6.5%] of all previous [42.8%] of all DM cases
	Never	**12017** [4.4%] of all never [46.1%] of all DM cases
		Mean (SD)
		Have DM / No DM
CperD		17.7 (±9.56) / 15.4 (±8.3)
SI		17.63 (±6.21) / 17.86 (±5.77)

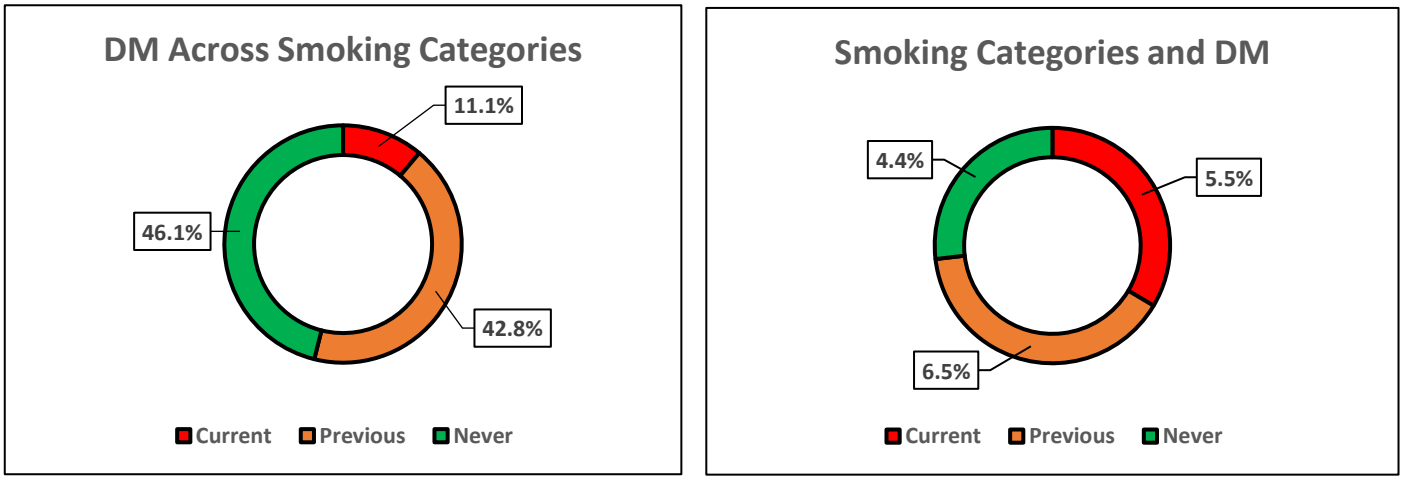


Figure 8.3: Prevalence of DM across smoking categories

Summary

This section discussed the descriptive associations of smoking status, CperD, and SI variables with CMDs and covariates. It also included a descriptive analysis of the associations between CMDs and the covariates. The descriptive analysis gives an overview of the variables in this sample. The next section will discuss the inferential analysis of these variables observationally.

Smoking behaviour and covariates

The covariates used in this thesis were sex, age, degree, ethnicity, deprivation level (Townsend score) and BMI. The covariates are divided based on their type; either qualitative or quantitative. The qualitative variables (nominal and binary) are sex, degree, and ethnicity. The quantitative variables include age, BMI, and Townsend score. This section focuses on the association between smoking behaviour and covariates descriptively.

..1.2.50 Smoking vs qualitative variables

This section focuses on the association between smoking variables and qualitative covariates. The sociodemographic qualitative variables were sex, degree and ethnicity. The prevalence of smoking among male participants was higher than among females. Approximately 12.5% of males were current smokers compared to 8.9% of female current smokers. This pattern was the same with previous smokers in which the prevalence of previous male smokers was higher compared to females (38.5%, and 31.5%, respectively). The prevalence of smoking was higher among individuals with no degree compared to those who had a degree (current smokers: 12%,7.4%, respectively, previous smokers: 36.1%, 31.7%, respectively). Regarding ethnicity, smoking was more prevalent among white British. Almost 45.5% of white British were either current (10.2%) or had previously smoked (35.3%). Smoking was also prevalent

among other ethnicities (42.8%) of which 12.9% were current and 29.9% were previous smokers.

The average smoking intensity (CperD) varied across the categories of the qualitative variables. Males on average have higher cigarettes smoked per day compared to females (17.06 ± 4.81, 14.07 ± 7.36, respectively). The individuals who had no degree smoke on average more cigarettes per day compared to individuals holding a high degree (15.94 ± 8.37, 13.68 ± 8.22, respectively). Regarding ethnicity, white British smoke on average higher CperD compared to other ethnicities (15.77 ± 8.38, 14.02 ± 8.29, respectively).

The age individuals started to smoke (smoking initiation) followed the same pattern as CperD concerning these variables. On average, male participants started to smoke relatively earlier than females (17.46 ± 5.67, 18.28 ± 5.9, respectively). The individuals who had no degree started to smoke earlier compared to individuals holding a high degree (17.5 ± 5.58, 19.34 ± 6.44, respectively). Regarding ethnicity, white British began to smoke relatively earlier than other ethnicities (17.72 ± 5.73, 18.68 ± 6.14, respectively). Detailed associations are shown in Table 8.7.

Table 8.7: Descriptive analysis of smoking variables vs covariates (qualitative variables)

Variables		**Sex**		**Degree**		**Ethnicity**	
		Female	Male	Yes	No	White British	Other ethnicities
Smoking status	Current	8.9%	12.5%	7.4%	12%	10.2%	12.9%
	Previous	31.5%	38.5%	31.7%	36.1%	35.3%	29.9%
	Never	59.6%	49%	60.8%	51.9%	54.5%	57.1%
		Mean (SD)		**Mean (SD)**		**Mean (SD)**	
CperD		14.07 (±7.36)	17.06 (±4.81)	13.68 (±8.22)	15.94 (±8.37)	15.77 (±8.38)	14.02 (±8.29)
SI		18.28 (±5.9)	17.46 (±5.67)	19.34 (±6.44)	17.5 (±5.58)	17.72 (±5.73)	18.68 (±6.14)

..1.2.51 Smoking vs quantitative variables

This section focuses on the association between smoking variables and quantitative covariates. The quantitative variables were age, Townsend score and body mass index (BMI). The mean age of all participants in the sample was 56.53 ± 8.09. On average, current smokers in the sample are younger compared to previous and never (45.67 ± 8.14, 58.18 ± 7.69, 55.84 ± 8.15, respectively). The current smokers seem to have the lowest BMI compared to never and previous (27.05 ± 4.81, 27.17 ± 4.79, 27.92 ± 4.75, respectively). All smoking categories lie above the healthy weight. Regarding deprivation level, current smokers are materially more deprived than both previous and never smokers (0.14, -1.29 ± 3.04, -1.61 ± 2.93, respectively). The Townsend score was positive among current smokers (positive score = more deprived).

The CperD and SI are quantitative variables. The correlation coefficient (r) was used to describe the relationship between CperD/SI and other quantitative covariates. The relationship between CperD/SI and these variables was generally weak. For CperD, the correlation with age and BMI was positive and very weak (almost zero) (r = 0.02, r = 0.08, respectively). Townsend score has a positive and weak (13%) correlation with CperD (r = 0.13). Smoking initiation had very weak associations with all quantitative covariates in which age was positive and BMI/Townsend was negative (age: r = 0.05, BMI: r = -0.001, Townsend: r = -0.06). Detailed associations are shown in Table 8.8.

Table 8.8: Descriptive analysis of smoking variables vs covariates (quantitative variables)

Variables		Age	BMI	Townsend
		Mean (SD)	Mean (SD)	Mean (SD)
Smoking Status	Current	45.67 (±8.14)	27.05 (±4.81)	0.14 (±3.52)
	Previous	58.18 (±7.69)	27.92 (±4.75)	-1.29 (±3.04)
	Never	55.84 (±8.15)	27.17 (±4.79)	-1.61 (±2.93)
		Correlation	**Correlation**	**Correlation**
CperD		r = 0.02	r = 0.08	r = 0.13
SI		r = 0.05	r = -0.001	r = -0.06

CMDs vs covariates

This section explores the descriptive associations between individuals with CMDs and the covariates. Percentages and means (± SD) were used to describe such associations. Coronary heart disease (CHD) was more prevalent among males (almost three times) compared to females (7.1%, and 2.4%, respectively). Individuals with no degree have a higher prevalence of CHD compared to individuals with a high degree (5.4%, and 2.7%, respectively). The prevalence of CHD was almost the same among white British compared to other ethnicities (4.5%, and 4.6%, respectively). The average BMI among individuals with CHD seems to be in overweigh category (29.47 ± 5.05). The deprivation level among individuals with CHD was above the sample mean (-0.58 ± 3.41).

The stroke followed the same pattern of prevalence among these variables. Stroke was more prevalent among males compared to females (1.5%, and 1%, respectively). Individuals with no degree have a higher prevalence of stroke compared to individuals with a high degree (1.4%, and 0.8%, respectively). The prevalence of stroke was higher among white British compared to other ethnicities (1.3%, and 1.1%, respectively). The average BMI among individuals with stroke seems to be in the

overweight category (28.7 ± 5.11). The deprivation level among individuals with stroke was above the sample mean (-0.66 ± 3.38).

The prevalence of HTN was higher among males compared to females (25.9%, and 22.3%, respectively). Individuals with no degree have a higher prevalence of HTN compared to individuals with a high degree (25.9%, and 20%, respectively). The prevalence of HTN is almost the same among British and other ethnicities (23.9%, and 24.3%, respectively). The average BMI among individuals with HTN seems to fall among the overweight and obese categories (29.29 ± 5.22). The deprivation level among individuals with HTN was relatively near the sample mean (-1.17 ± 3.16).

DM was more prevalent (almost double) among males compared to females (7%, and 3.8%, respectively). Individuals with no degree have a higher prevalence of DM compared to individuals holding a high degree (5.9%, and 3.8%, respectively). The non-white British have a higher prevalence (almost double) of DM compared to white British (8.2%, and 4.8%, respectively). The average BMI among individuals with DM seems to be in the obese category (31.34 ± 5.92). The deprivation level among individuals with DM was above the sample mean (-0.40 ± 3.42). Table 8.9 shows the findings obtained from these descriptive associations.

Table 8.9: Descriptive analysis of CMDs vs covariates

Variables	Sex		Degree		Ethnicity	
	Female	Male	Yes	No	White British	Other ethnicities
CHD	2.4%	7.1%	2.7%	5.4%	4.5%	4.6%
Stroke	1%	1.5%	0.8%	1.4%	1.3%	1.1%
HTN	22.3%	25.9%	20%	25.9%	23.9%	24.3%
DM	3.8%	7%	3.8%	5.9%	4.9%	8.2%
	Age		**BMI**		**Townsend**	
	Mean (SD)		Mean (SD)		Mean (SD)	
CHD	61.92 (±6.04)		29.47 (±5.05)		-0.58 (±3.41)	
Stroke	60.6 (±6.88)		28.7 (±5.11)		-0.66 (±3.38)	
HTN	59.13 (±7.16)		29.29 (±5.22)		-1.17 (±3.16)	
DM	59.55 (±7.21)		31.34 (±5.92)		-0.40 (±3.42)	

Plots

..1.2.52 Smoking status across covariates

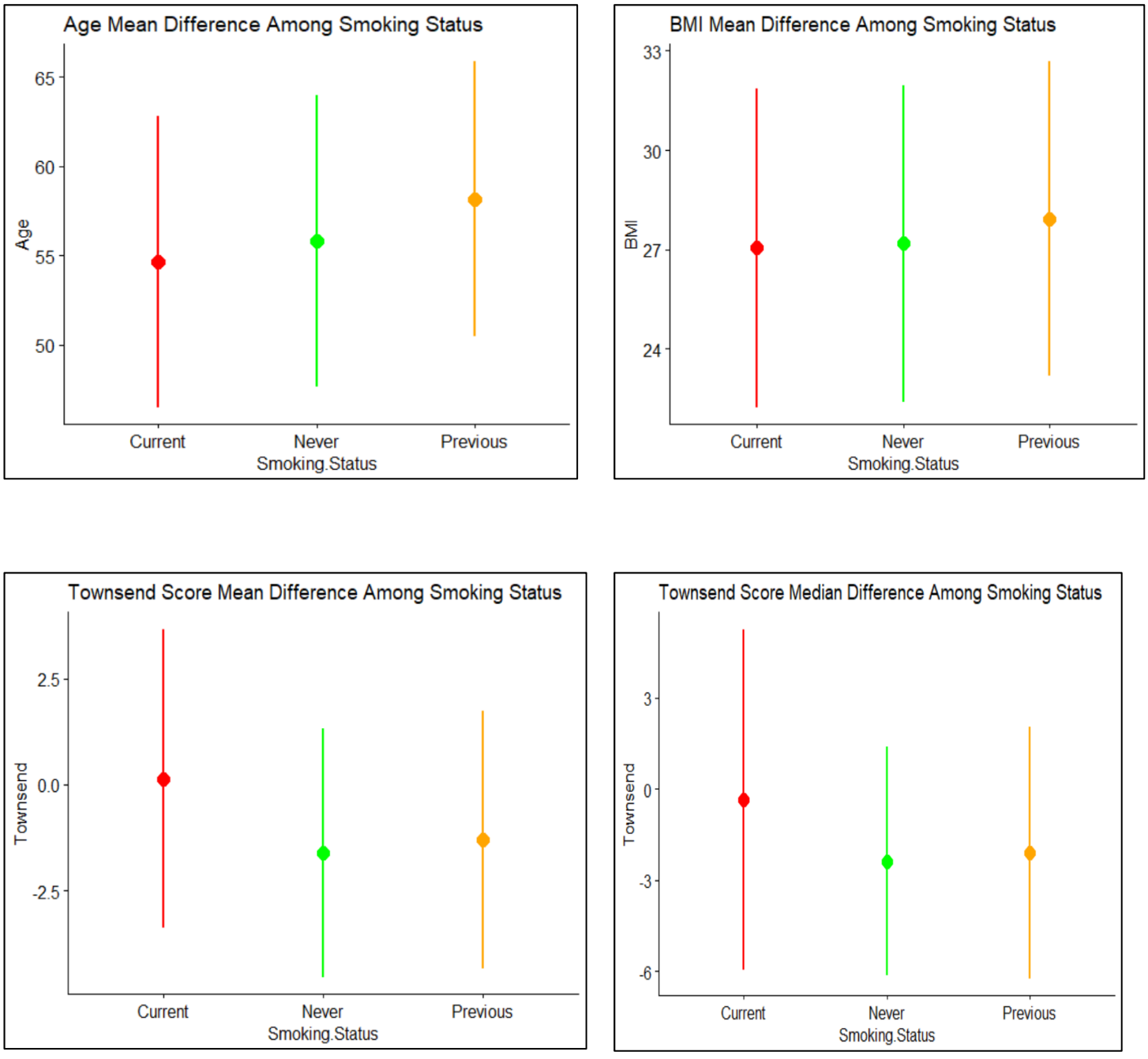


Figure 8.4: Smoking categories vs covariates

..1.2.53 CperD across CMDs

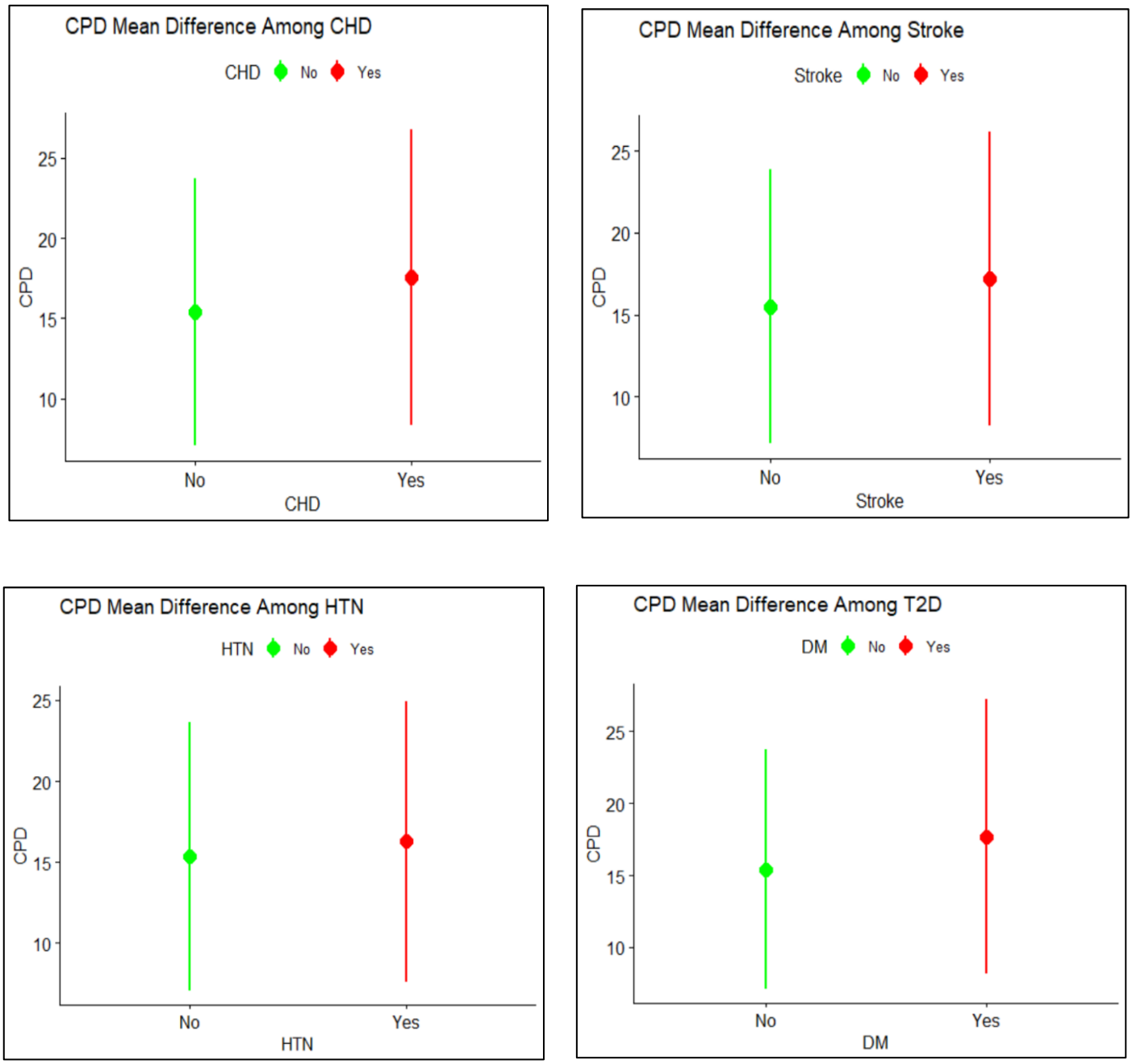


Figure 8.5: CPD across CMDs

..1.2.54 *CperD across covariates*

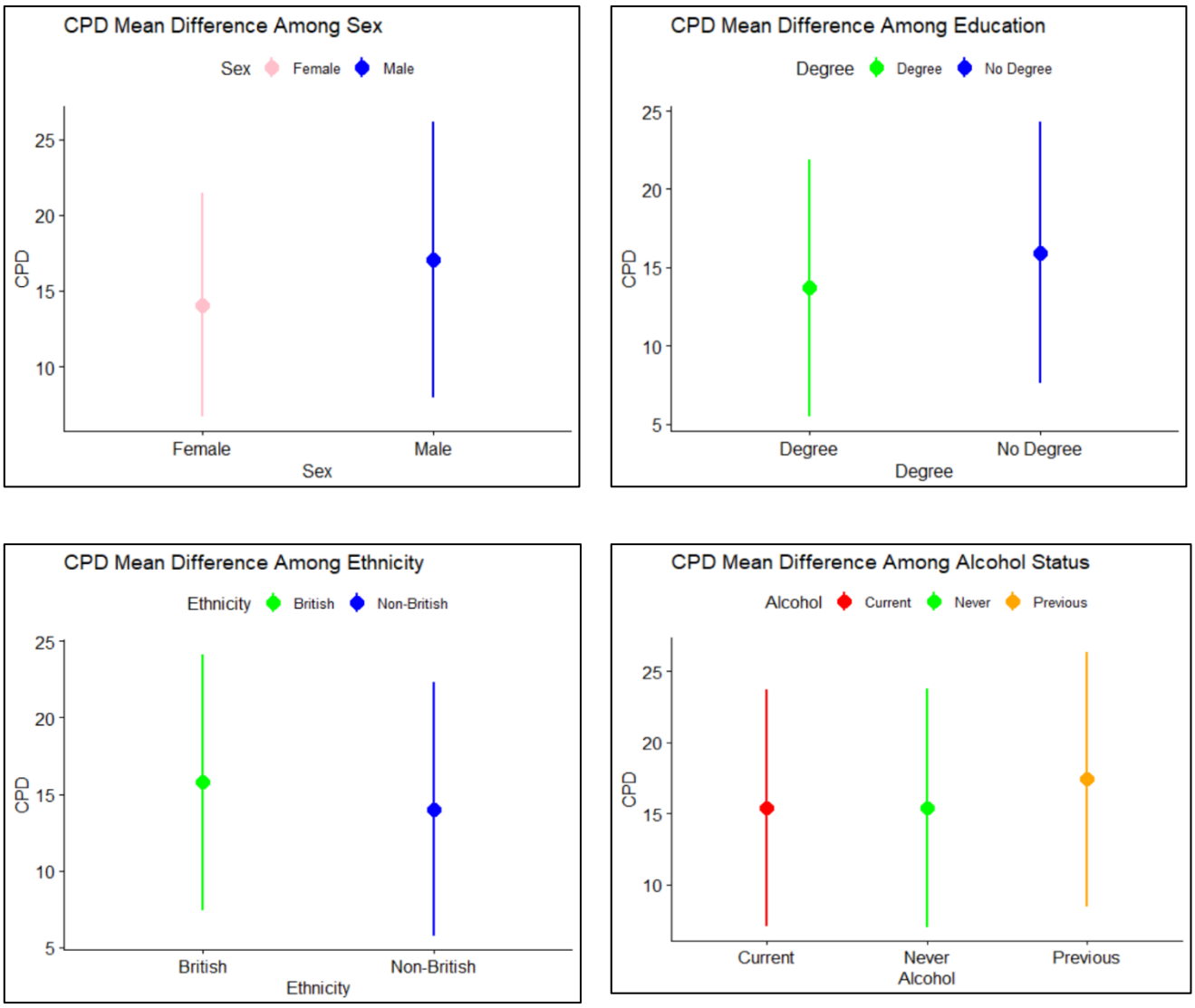


Figure 8.6: CPD vs covariates

..1.2.55 SI across CMDs

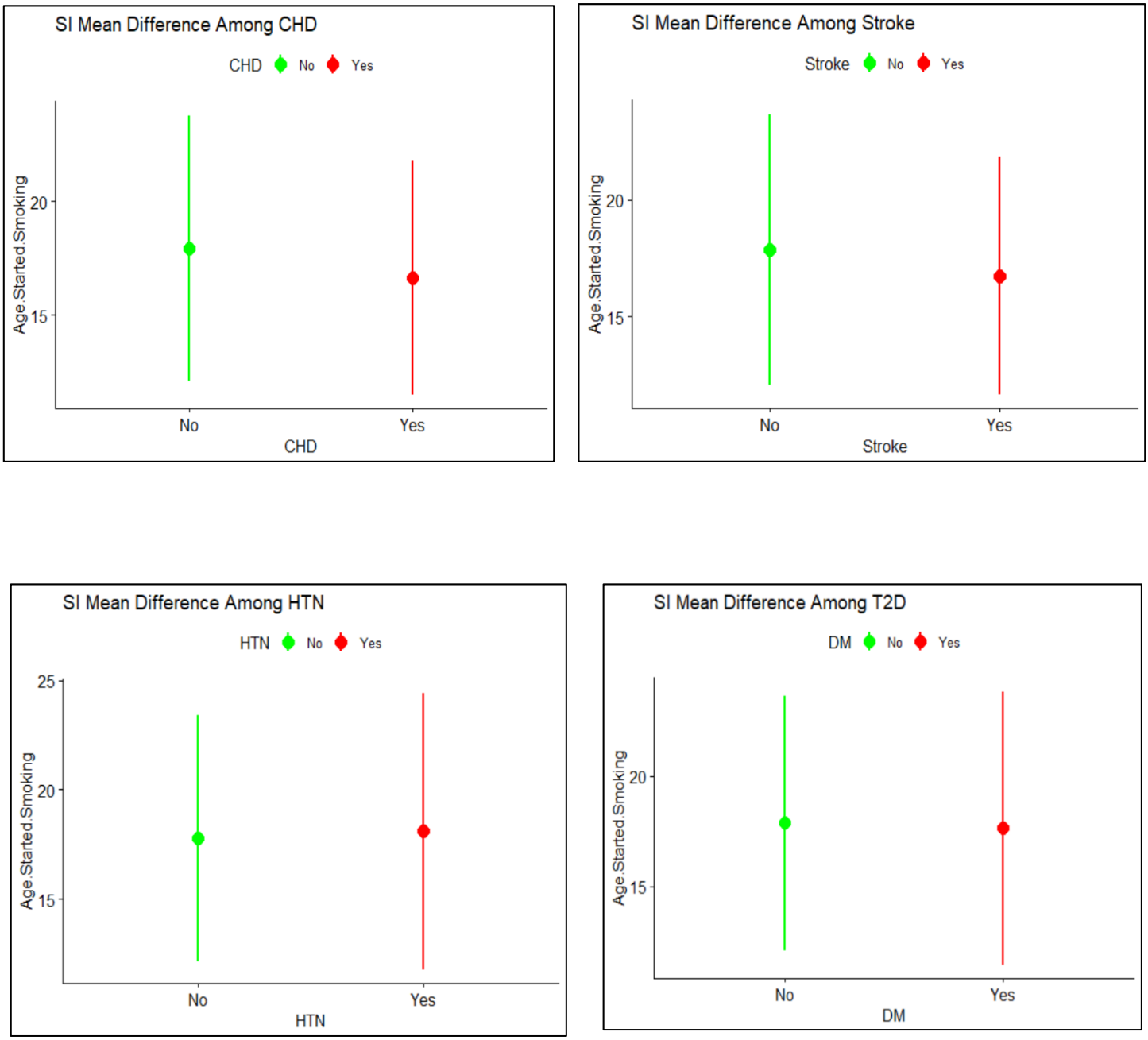


Figure 8.7: SI vs CMDs

..1.2.56 SI across covariates

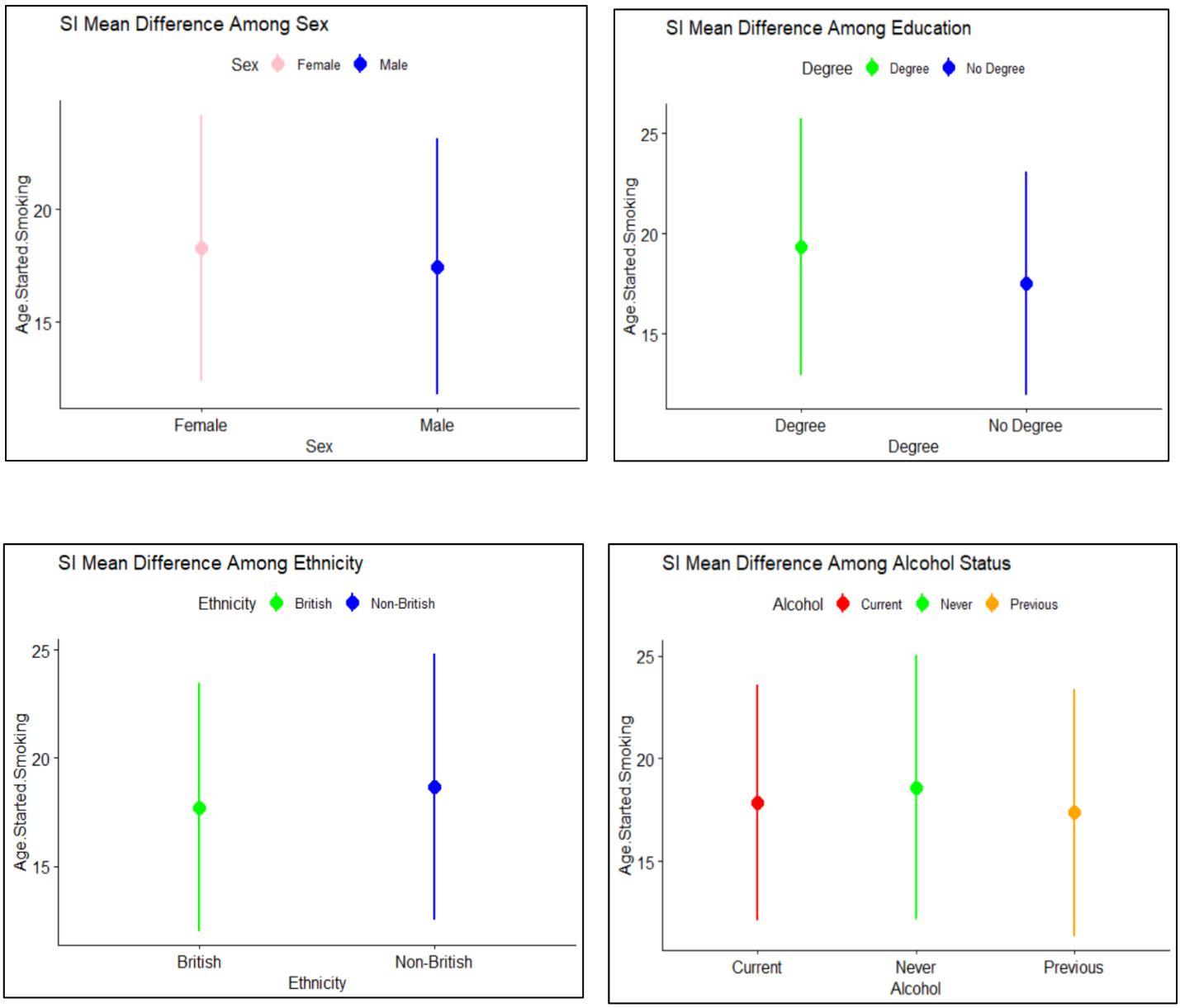
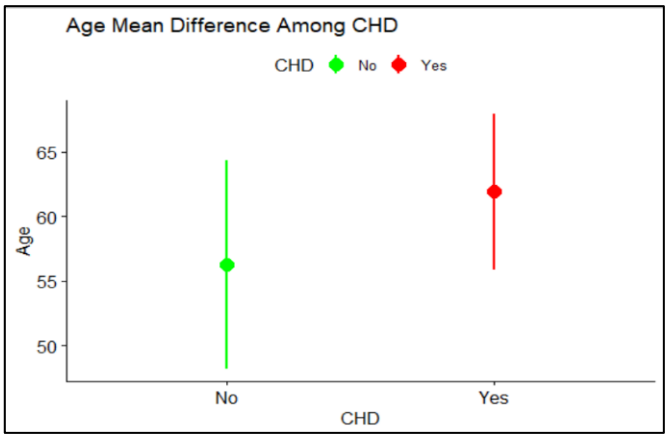
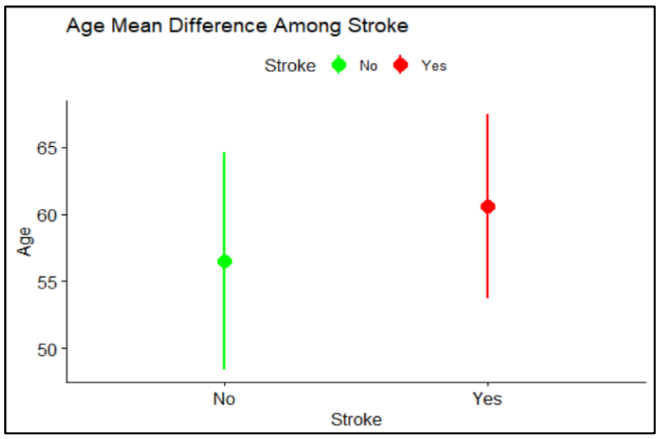
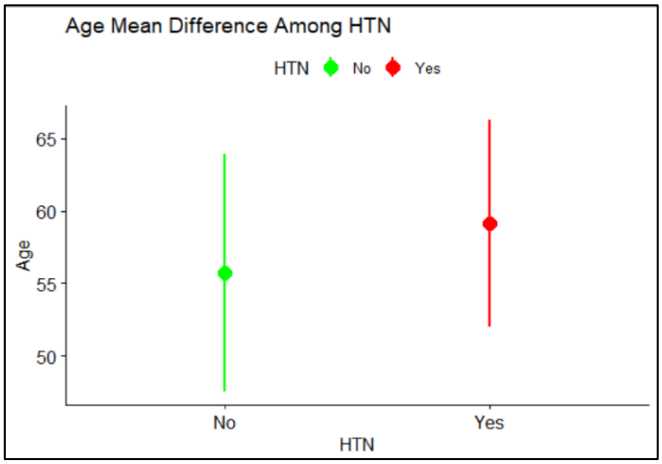


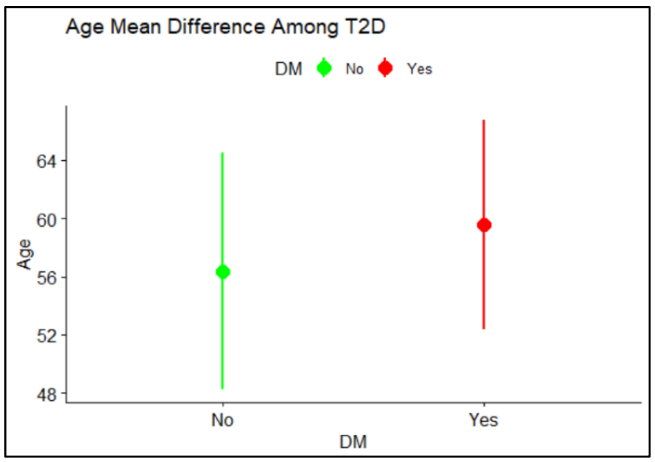
Figure 8.8: SI vs covariates

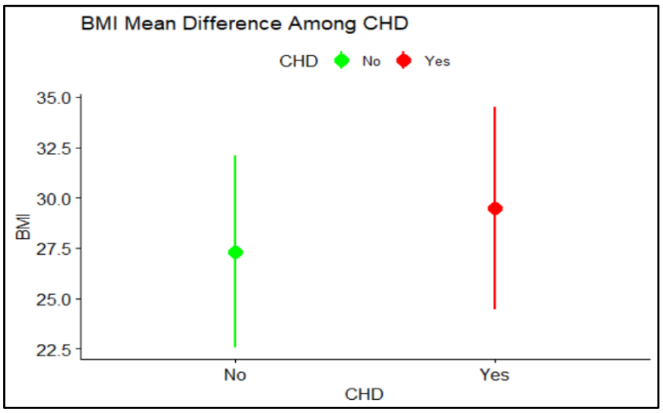
..1.2.57 CMDs across covariates

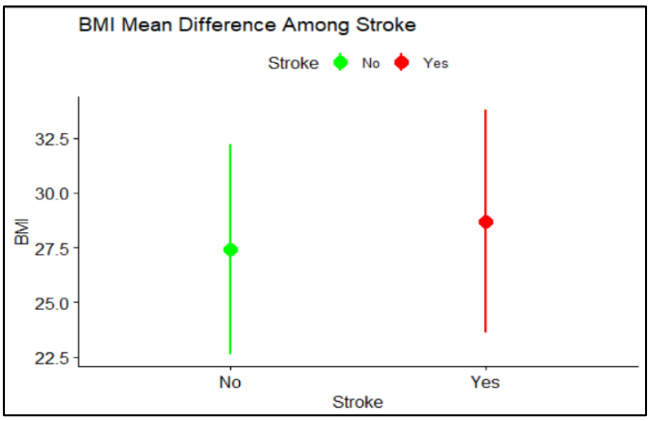












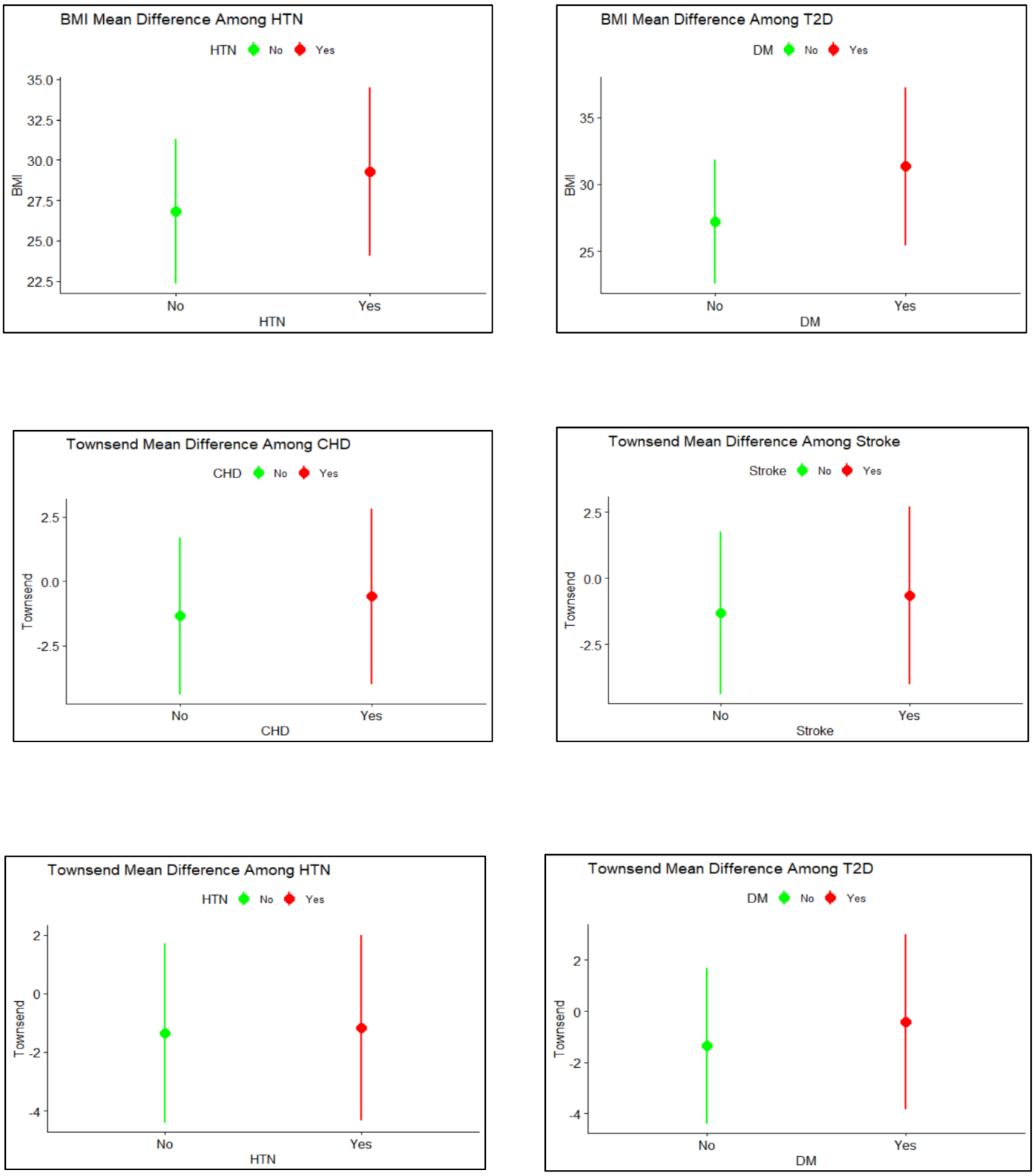


Figure 8.9: CMDs vs covariates

Observational Analysis

Smoking variables vs CMDs [tables]

..1.2.58 Smoking status vs CHD

Table 8.10: Logistic regression analysis of smoking status (current, previous, never) vs CHD

Coronary Heart Disease (CHD)		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Current	1.90	1.82 – 1.99	**<0.001**	1.61	1.54 – 1.68	**<0.001**
	Previous	2.24	2.18 – 2.31	**<0.001**	1.50	1.45 – 1.54	**<0.001**
Sex [Male**]**					2.84	2.76 – 2.93	**<0.001**
Education [Have Degree**]**					0.65	0.62 – 0.67	**<0.001**
Ethnicity [White British**]**					0.83	0.79 – 0.87	**<0.001**
Age					1.11	1.11 – 1.11	**<0.001**
Townsend					1.07	1.07 – 1.08	**<0.001**
BMI					1.08	1.07 – 1.08	**<0.001**

..1.2.59 Smoking status vs stroke

Table 8.11: Logistic regression analysis of smoking status (current, previous, never) vs stroke

Stroke		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Current	1.82	1.68 – 1.96	**<0.001**	1.64	1.52 – 1.77	**<0.001**
	Previous	1.51	1.43 – 1.59	**<0.001**	1.16	1.10 – 1.23	**<0.001**
Sex [Male**]**					1.43	1.36 – 1.50	**<0.001**
Education [Have Degree**]**					0.75	0.71 – 0.80	**<0.001**
Ethnicity [White British**]**					1.09	1.00 – 1.19	**0.050**
Age					1.07	1.07 – 1.08	**<0.001**
Townsend					1.06	1.06 – 1.07	**<0.001**
BMI					1.04	1.04 – 1.05	**<0.001**

..1.2.60 Smoking status vs HTN

Table 8.12: Logistic regression analysis of smoking status (current, previous, never) vs HTN

Hypertension (HTN)		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Current	0.89	0.87 – 0.91	<**0.001**	0.89	0.87 – 0.91	<**0.001**
	Previous	1.23	1.21 – 1.24	<**0.001**	0.99	0.98 – 1.01	0.221
Sex [Male**]**					1.16	1.14 – 1.17	<**0.001**
Education [Have Degree**]**					0.89	0.87 – 0.90	<**0.001**
Ethnicity [White British**]**					0.86	0.84 – 0.88	<**0.001**
Age					1.06	1.06 – 1.06	<**0.001**
Townsend					1.02	1.01 – 1.02	<**0.001**
BMI					1.11	1.10 – 1.11	<**0.001**

..1.2.61 Smoking status vs DM

Table 8.13: Logistic regression analysis of smoking status (current, previous, never) vs DM

Diabetes Mellitus (DM)		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Current	1.26	1.21 – 1.32	<**0.001**	1.12	1.07 – 1.17	<**0.001**
	Previous	1.50	1.46 – 1.54	<**0.001**	1.12	1.09 – 1.16	<**0.001**
Sex [Male**]**					1.93	1.88 – 1.98	<**0.001**
Education [Have Degree**]**					0.82	0.80 – 0.85	<**0.001**
Ethnicity [White British**]**					0.52	0.50 – 0.54	<**0.001**
Age					1.06	1.06 – 1.06	<**0.001**
Townsend					1.06	1.06 – 1.07	<**0.001**
BMI					1.15	1.14 – 1.15	<**0.001**

..1.2.63 Smoking intensity (CperD) vs coronary heart disease (CHD)

Table 8.14: Logistic regression analysis of CperD vs CHD

Coronary Heart Disease (CHD)	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
CperD	1.03	1.02 – 1.03	**<0.001**	1.01	1.00 – 1.02	**<0.001**
Sex [Male**]**				2.39	2.17 – 2.64	**<0.001**
Education [Have Degree**]**				0.63	0.54 – 0.73	**<0.001**
Ethnicity [White British**]**				1.00	0.87 – 1.16	0.959
Age				1.09	1.09 – 1.10	**<0.001**
Townsend				1.10	1.09 – 1.12	**<0.001**
BMI				1.07	1.06 – 1.08	**<0.001**

..1.2.64 Smoking intensity (CperD) vs stroke

Table 8.15: Logistic regression analysis of CperD vs stroke

Stroke	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
CperD	1.02	1.01 – 1.03	**<0.001**	1.01	1.00 – 1.02	**0.009**
Sex [Male**]**				1.28	1.09 – 1.50	**0.002**
Education [Have Degree**]**				0.71	0.56 – 0.91	**0.006**
Ethnicity [White British**]**				1.27	0.98 – 1.65	0.065
Age				1.06	1.05 – 1.07	**<0.001**
Townsend				1.08	1.05 – 1.10	**<0.001**
BMI				1.02	1.01 – 1.04	**0.005**

..1.2.65 Smoking intensity (CperD) vs hypertension (HTN)

Table 8.16: Logistic regression analysis of CperD vs HTN

Hypertension (HTN)	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
CperD	1.01	1.01 – 1.02	**<0.001**	1.01	1.00 – 1.01	**<0.001**
Sex [Male**]**				1.15	1.09 – 1.21	**<0.001**
Education [Have Degree**]**				0.94	0.87 – 1.01	0.074
Ethnicity [White British**]**				0.86	0.80 – 0.93	**<0.001**
Age				1.05	1.04 – 1.05	**<0.001**
Townsend				1.02	1.01 – 1.03	**<0.001**
BMI				1.08	1.07 – 1.08	**<0.001**

..1.2.66 Smoking intensity (CperD) vs diabetes mellitus (DM)

Table 8.17: Logistic regression analysis of CperD vs DM

Diabetes Mellitus (DM)	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
CperD	1.03	1.02 – 1.03	**<0.001**	1.02	1.01 – 1.02	**<0.001**
Sex [Male**]**				1.75	1.59 – 1.94	**<0.001**
Education [Have Degree**]**				0.90	0.78 – 1.03	0.121
Ethnicity [White British**]**				0.60	0.52 – 0.68	**<0.001**
Age				1.06	1.06 – 1.07	**<0.001**
Townsend				1.05	1.04 – 1.07	**<0.001**
BMI				1.15	1.14 – 1.16	**<0.001**

..1.2.67 Smoking initiation (SI) vs coronary heart disease (CHD)

Table 8.18: Logistic regression analysis of SI vs CHD

Coronary Heart Disease (CHD)	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
SI	0.95	0.94 – 0.96	**<0.001**	0.96	0.95 – 0.97	**<0.001**
Sex [Male**]**				2.22	2.02 – 2.45	**<0.001**
Education [Have Degree**]**				0.63	0.55 – 0.72	**<0.001**
Ethnicity [White British**]**				0.95	0.83 – 1.09	0.482
Age				1.09	1.08 – 1.10	**<0.001**
Townsend				1.10	1.09 – 1.11	**<0.001**
BMI				1.07	1.06 – 1.08	**<0.001**

..1.2.68 Smoking initiation (SI) vs stroke

Table 8.19: Logistic regression analysis of SI vs stroke

Stroke	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
SI	0.96	0.94 – 0.97	**<0.001**	0.96	0.95 – 0.98	**<0.001**
Sex [Male**]**				1.22	1.05 – 1.43	**0.009**
Education [Have Degree**]**				0.78	0.62 – 0.97	**0.027**
Ethnicity [White British**]**				1.25	0.97 – 1.60	0.083
Age				1.06	1.05 – 1.07	**<0.001**
Townsend				1.09	1.06 – 1.11	**<0.001**
BMI				1.02	1.01 – 1.04	**0.001**

..1.2.69 Smoking initiation (SI) vs hypertension (HTN)

Table 8.20: Logistic regression analysis of SI vs HTN

Hypertension (HTN)	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
SI	1.01	1.00 – 1.01	**<0.001**	1.01	1.00 – 1.01	**<0.001**
Sex [Male**]**				1.16	1.11 – 1.23	**<0.001**
Education [Have Degree**]**				0.91	0.85 – 0.97	**0.006**
Ethnicity [White British**]**				0.91	0.84 – 0.98	**0.014**
Age				1.05	1.04 – 1.05	**<0.001**
Townsend				1.02	1.01 – 1.03	**<0.001**
BMI				1.08	1.07 – 1.08	**<0.001**

..1.2.70 Smoking initiation (SI) vs diabetes mellitus (DM)

Table 8.21: Logistic regression analysis of SI vs DM

Diabetes Mellitus (DM)	Unadjusted			Adjusted		
	OR	**CI**	**P**	**OR**	**CI**	**P**
SI	0.99	0.98 – 1.00	0.076	0.99	0.99 – 1.00	0.190
Sex [Male**]**				1.83	1.66 – 2.02	**<0.001**
Education [Have Degree**]**				0.87	0.76 – 0.99	**0.031**
Ethnicity [White British**]**				0.62	0.55 – 0.70	**<0.001**
Age				1.06	1.06 – 1.07	**<0.001**
Townsend				1.06	1.05 – 1.07	**<0.001**
BMI				1.15	1.14 – 1.16	**<0.001**

Smoking behaviour (binary)

..1.2.71 Ever vs never smoking (smoking as a binary variable) vs CHD

After examining the associations using smoking status as a binary (ever vs never), the findings of all variables were almost identical to smoking with three categories (current, previous, and never). The individuals who ever smoked have a 52% higher risk of developing CHD compared to never smoked individuals (OR: 1.52, 95% CI: 1.48 – 1.56, P<0.001). The rest of the associations are in Table 8.22.

Table 8.22: Logistic regression analysis of smoking status (ever vs never) vs CHD

Coronary Heart Disease (CHD)		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Ever	2.16	2.10 – 2.22	**<0.001**	1.52	1.48 – 1.56	**<0.001**
Sex [Male**]**					2.84	2.76 – 2.93	**<0.001**
Education [Have Degree**]**					0.64	0.62 – 0.67	**<0.001**
Ethnicity [White British**]**					0.83	0.79 – 0.87	**<0.001**
Age					1.11	1.11 – 1.11	**<0.001**
Townsend					1.07	1.07 – 1.08	**<0.001**
BMI					1.08	1.07 – 1.08	**<0.001**

..1.2.72 Ever vs never smoking (smoking as a binary variable) vs stroke

When categorising smoking status into ever vs never, ever smokers have a 58% higher risk to develop stroke compared to never (OR=1.58, 95% CI: 1.50 – 1.66, P<0.001).

Table 8.23: Logistic regression analysis of smoking status (ever vs never) vs stroke

Stroke		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Ever	1.58	1.50 – 1.66	**<0.001**	1.26	1.19 – 1.33	**<0.001**
Sex [Male**]**					1.43	1.36 – 1.51	**<0.001**
Education [Have Degree**]**					0.74	0.70 – 0.79	**<0.001**
Ethnicity [White British**]**					1.09	1.00 – 1.19	0.055
Age					1.07	1.07 – 1.08	**<0.001**
Townsend					1.07	1.06 – 1.08	**<0.001**
BMI					1.04	1.03 – 1.05	**<0.001**

..1.2.73 Ever vs never smoking (smoking as a binary variable) vs HTN

When categorising smoking status into ever vs never, ever smokers have a 14% higher risk to develop HTN compared to never (OR=1.14, 95% CI: 1.13 – 1.16, P<0.001).

Table 8.24: Logistic regression analysis of smoking status (ever vs never) vs HTN

Hypertension (HTN)		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Ever	1.14	1.13 – 1.16	**<0.001**	0.97	0.96 – 0.98	**<0.001**
Sex [Male**]**					1.16	1.14 – 1.17	**<0.001**
Education [Have Degree**]**					0.89	0.88 – 0.90	**<0.001**
Ethnicity [White British**]**					0.86	0.84 – 0.88	**<0.001**
Age					1.06	1.06 – 1.06	**<0.001**
Townsend					1.01	1.01 – 1.02	**<0.001**
BMI					1.11	1.10 – 1.11	**<0.001**

..1.2.74 Ever vs never smoking (smoking as a binary variable) vs DM

When categorising smoking status into ever vs never, ever smokers have a 44% higher risk to develop DM compared to never (OR=1.44, 95% CI: 1.41 – 1.48, P<0.001).

Table 8.25: Logistic regression analysis of smoking status (ever vs never) vs DM

Diabetes Mellitus (DM)		Unadjusted			Adjusted		
		OR	**CI**	**P**	**OR**	**CI**	**P**
Smoking Status	Ever	1.44	1.41 – 1.48	**<0.001**	1.12	1.09 – 1.15	**<0.001**
Sex [Male**]**					1.93	1.88 – 1.98	**<0.001**
Education [Have Degree**]**					0.82	0.80 – 0.85	**<0.001**
Ethnicity [White British**]**					0.52	0.50 – 0.54	**<0.001**
Age					1.06	1.06 – 1.06	**<0.001**
Townsend					1.06	1.06 – 1.07	**<0.001**
BMI					1.15	1.14 – 1.15	**<0.001**

MR

Smoking status sample characteristics:

Table 8.26. Sample characteristics (MR-sample)

Variable	Level	Count (%)
Smoking status	Never	172,504 (55%)
	Ever	141,622 (45%)
Sex	Male	144,427 (46%)
	Female	169,699 (54%)
Degree (college/university)	No Degree	213,948 (68%)
	Degree	100,178 (32%)
Coronary Heart Disease (CHD)	No	300685 (95.72%)
	Yes	13441 (4.28%)
Stroke	No	313307 (99.74%)
	Yes	819 (0.26%)
Hypertension (HTN)	No	238222 (75.84%)
	Yes	75904 (24.16%)
Diabetes Mellitus (DM)	No	299470 (95.33%)
	Yes	14656 (4.67%)
Variable	**Mean (SD)**	
Age	56.80 (8.00)	
Body Mass Index (BMI)	27.36 (4.73)	
Deprivation Level (Townsend score)	-1.60 (2.91)	
Cholesterol	5.73 (1.14)	
LDL	3.58 (0.87)	
TG	1.75 (1.02)	
HDL	1.45 (0.38)	

IV Assumptions for all SNPs

Individual SNP assumptions
SNP vs CperD

Table 8.27. Individual SNP (1st assumption)

SNP	BETA	P
rs1317286	0.959898	2.90E-37
rs1051730	0.957472	4.88E-37
rs12914385	0.929878	3.38E-37
rs16969968	0.955912	6.00E-37
rs8034191	0.952962	6.88E-37
rs951266	0.954618	9.14E-37
rs17486278	0.951751	1.48E-36
rs72740964	0.945604	5.85E-36
rs55853698	0.929098	2.96E-35
rs8040868	0.884425	4.25E-34
rs11637630	0.756125	1.73E-18
rs938682	0.755137	1.90E-18
rs12910984	0.754276	2.13E-18
rs6474412	0.485786	1.50E-08
rs2229961	1.312994	3.20E-07

Individual SNP assumptions
SNP vs Outcomes

Table 8.28. Individual SNP (2nd assumption)

SNPs	Variables							
	CHD		Stroke		HTN		DM	
	OR	P	OR	P	OR	P	OR	P
rs1317286	1.002	0.958	0.85	0.181	0.98	0.479	0.90	0.325
rs1051730	0.99	0.911	0.86	0.197	0.98	0.468	0.90	0.413
rs12914385	0.98	0.535	0.87	0.219	0.98	0.402	0.92	0.912
rs16969968	0.99	0.879	0.86	0.202	0.99	0.554	0.89	0.523
rs8034191	1.002	0.954	0.89	0.346	0.99	0.639	0.90	0.276
rs951266	0.99	0.755	0.86	0.207	0.99	0.636	0.90	0.781
rs17486278	0.99	0.727	0.86	0.203	0.99	0.551	0.89	0.813
rs72740964	0.99	0.870	0.86	0.214	0.98	0.480	0.89	0.491
rs55853698	1.0002	0.996	0.89	0.324	0.98	0.510	0.90	0.516
rs8040868	0.97	0.495	0.94	0.553	0.99	0.532	0.93	0.07
rs11637630	0.99	0.869	1.01	0.924	1.07	0.324	1.03	0.518
rs938682	0.99	0.863	1.01	0.926	1.07	0.454	1.03	0.521
rs12910984	0.99	0.856	1.01	0.928	1.07	0.542	1.03	0.526
rs6474412	1.01	0.882	1.03	0.842	0.95	0.156	1.01	0.847
rs2229961	0.90	0.468	1.40	0.323	1.10	0.219	0.78	0.133

Individual SNP assumptions
SNP vs Covariates

Table 8.29. Individual SNP (3rd assumption)

SNPs	Age	Degree	Sex	Townsend	BMI
	P value	P value	P value	P value	P value
rs1317286	**0.04 (↓)**	0.99	0.89	0.77	**<0.001**
rs1051730	0.07	0.78	0.92	0.99	**<0.001**
rs12914385	0.12	0.58	0.97	0.94	**<0.001**
rs16969968	**0.04 (↓)**	0.723	0.843	0.951	**<0.001**
rs8034191	0.141	0.95	0.479	0.776	**<0.001**
rs951266	0.054	0.883	0.778	0.929	**<0.001**
rs17486278	0.06	0.787	0.768	0.951	**<0.001**
rs72740964	0.07	0.757	0.734	0.971	**<0.001**
rs55853698	0.06	0.927	0.976	0.706	**<0.001**
rs8040868	0.121	0.887	0.87	0.889	**<0.001**
rs11637630	0.16	0.57	0.708	0.991	**<0.001**
rs938682	0.164	0.58	0.731	0.996	**<0.001**
rs12910984	0.159	0.57	0.804	0.976	**<0.001**
rs6474412	0.324	0.927	0.583	0.786	0.994
rs2229961	0.26	0.496	0.244	0.657	0.457

..1.2.75 Individual SNPs analysis

This section will focus on performing MR for individual SNPs against CMD variables. All SNPs that have a GWAS-level significance with CperD were included in the MR analysis. The following CperD SNPs were significantly associated with decreased risk of DM (DM): rs1317286_G, rs12914385_T, rs16969968_A, rs8034191_C, rs951266_A, rs17486278_C, rs72740964_A and rs55853698_G. Additionally, the risk of HTN was lower among the following CperD-increasing SNPs: rs11637630_G, rs938682_G and rs12910984_G. Finally, the risk of HTN was 11% higher among rs6474412_C. The rest of the findings are summarised in Table 8.30.

Table 8.30: MR analysis of CperD and CMDs for individual SNPs

Variables/SNPs	CHD		Stroke		HTN		DM	
	MR Estimate	P	MR Estimate	P	MR Estimate	P	MR Estimate	P
rs1051730_A	1.23	0.252	0.55	0.313	0.92	0.406	0.91	0.634
rs1317286_G	0.99	0.958	0.85	0.181	0.98	0.479	0.90	**0.016**
rs12914385_T	0.98	0.535	0.87	0.219	0.98	0.402	0.92	**0.040**
rs16969968_A	0.99	0.879	0.86	0.202	0.99	0.554	0.89	**0.011**
rs8034191_C	1.002	0.954	0.89	0.346	0.99	0.639	0.90	**0.015**
rs951266_A	0.99	0.755	0.86	0.207	0.99	0.636	0.90	**0.013**
rs17486278_C	0.99	0.727	0.86	0.203	0.99	0.551	0.89	**0.010**
rs72740964_A	0.99	0.870	0.86	0.214	0.98	0.480	0.89	**0.011**
rs55853698_G	1.0002	0.996	0.89	0.324	0.98	0.510	0.90	**0.015**
rs8040868_C	0.97	0.495	0.93	0.553	0.99	0.532	0.92	0.07
rs11637630_G	1.01	0.869	0.98	0.924	0.92	**0.009**	0.96	0.516
rs938682_G	1.01	0.863	0.98	0.926	0.92	**0.009**	0.96	0.521
rs12910984_G	1.01	0.856	0.98	0.928	0.91	**0.007**	0.96	0.526
rs6474412_C	0.99	0.882	0.95	0.842	1.11	**0.049**	0.98	0.847
rs2229961_A	0.91	0.468	1.33	0.323	1.08	0.219	0.81	0.133

Two-Sample MR

SNPs summary statistics

CperD and CHD summary statistics

Table 8.31. Beta and SD used in summary-level MR (CperD vs CHD)

SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	-0.0045	0.041
rs7599488	0.026414	0.005544	0.02535	0.039
rs215614	-0.04448	0.005728	-0.0201	0.04004
rs73229090	0.055489	0.008763	-0.10946	0.0644
rs6474412	0.067113	0.006613	0.0069	0.04662
rs3025343	0.063352	0.008661	0.07492	0.0565
rs8034191	0.182567	0.005889	0.00237	0.04091
rs2229961	0.207114	0.023481	-0.10646	0.1465
rs12910984	0.1571	0.006687	-0.00851	0.04705
rs3733829	0.034965	0.005811	-0.0634	0.0408
rs3865453	-0.10241	0.010329	-0.0799	0.07691
rs28399443	-0.23131	0.017486	0.02309	0.1241
rs7260329	-0.04462	0.006017	0.00903	0.04216
rs2273506	0.06063	0.011182	0.08769	0.0765

Abbreviations: bx: beta for CperD, bxse: standard deviation for CperD, by: beta for outcome, byse: standard deviation for outcome

CperD and stroke summary statistics

Table 8.32. Beta and SD used in summary-level MR (CperD vs stroke)

SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	-0.154	0.119
rs7599488	0.026414	0.005544	-0.03795	0.11132
rs215614	-0.04448	0.005728	0.104	0.1123
rs73229090	0.055489	0.008763	-0.209	0.1913
rs6474412	0.067113	0.006613	0.02627	0.13169
rs3025343	0.063352	0.008661	0.03569	0.16249
rs8034191	0.182567	0.005889	-0.1116	0.1184
rs2229961	0.207114	0.023481	0.33593	0.33983
rs12910984	0.1571	0.006687	0.012	0.1328
rs3733829	0.034965	0.005811	-0.1226	0.1174
rs3865453	-0.10241	0.010329	-0.50718	0.26465
rs28399443	-0.23131	0.017486	0.11147	0.33758
rs7260329	-0.04462	0.006017	-0.061	0.12141
rs2273506	0.06063	0.011182	-0.253	0.2506

CperD and HTN summary statistics

Table 8.33. Beta and SD used in summary-level MR (CperD vs HTN)

SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	-0.0168	0.0232
rs7599488	0.026414	0.005544	-0.0217	0.0221
rs215614	-0.04448	0.005728	-0.03293	0.02266
rs73229090	0.055489	0.008763	-0.00457	0.03507
rs6474412	0.067113	0.006613	-0.05257	0.02666
rs3025343	0.063352	0.008661	0.03437	0.07023
rs8034191	0.182567	0.005889	-0.01086	0.0231
rs2229961	0.207114	0.023481	0.09478	0.07719
rs12910984	0.1571	0.006687	0.0708	0.02627
rs3733829	0.034965	0.005811	0.0107	0.0229
rs3865453	-0.10241	0.010329	-0.0505	0.0427
rs28399443	-0.23131	0.017486	0.05437	0.0707
rs7260329	-0.04462	0.006017	0.0148	0.0238
rs2273506	0.06063	0.011182	-0.03756	0.0451

CperD and DM summary statistics

Table 8.34. Beta and SD used in summary-level MR (CperD vs DM)

SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	-0.1097	0.044
rs7599488	0.026414	0.005544	0.0025	0.0412
rs215614	-0.04448	0.005728	-0.01545	0.04225
rs73229090	0.055489	0.008763	0.06242	0.0639
rs6474412	0.067113	0.006613	0.00948	0.0491
rs3025343	0.063352	0.008661	0.016	0.1314
rs8034191	0.182567	0.005889	-0.10699	0.0439
rs2229961	0.207114	0.023481	-0.2478	0.1647
rs12910984	0.1571	0.006687	0.0312	0.049
rs3733829	0.034965	0.005811	0.0646	0.04246
rs3865453	-0.10241	0.010329	0.06836	0.0767
rs28399443	-0.23131	0.017486	0.005655	0.133627
rs7260329	-0.04462	0.006017	0.01654	0.04446
rs2273506	0.06063	0.011182	-0.04043	0.08485

Smoking status MR (314k)

SNPs included in the analysis:

rs3001723_A, rs2947411_A, rs528301_G, rs13026471_T, rs13022438_G, rs6445538_C, rs9320995_G, rs3857914_C, rs12763665_A, rs1447481_T, rs10891504_G, rs2292239_T

IV assumptions results:

Table 8.35. IV assumptions (smoking status)

CMDs	**Genetic Score**	
	OR	*p*
CHD	0.997	0.588
Stroke	0.97	0.071
HTN	0.99	**<0.001**
DM	1.0003	0.931
Variables	**Genetic Score**	
	Estimates	*p*
Smoking status	0.0080 (OR=1.01)	**$1.35e^{-07}$**
Age	-0.0004	0.449
Degree [No]	-0.022	**0.0125**
Sex [Male]	-0.003	0.694
Townsend	0.001	0.433
BMI	-0.005	**$44e^{-08}$**
PC1	-0.015	**$6.9e^{-14}$**
PC2	0.1458	**$2e^{-16}$**
PC3	0.0424	**$2e^{-16}$**
PC4	0.915	**$2e^{-16}$**
PC5	-0.087	**$2e^{-16}$**
PC6	1.535	**$2e^{-16}$**
PC7	-0.167	**$2e^{-16}$**
PC8	-0.424	**$2e^{-16}$**
PC9	-1.011	**$2e^{-16}$**
PC10	0.458	**$2e^{-16}$**

Details of MR analysis

Ever vs never MR analysis

Table 8.36. Summary-level MR (Smoking status)

ID [Exposure]	Outcome	Method	Number of SNPs	Beta	SE	P value
Smoking (Ever)	CHD	MR Egger	15	0.05187	0.1098	0.6445
Smoking (Ever)	Stroke	MR Egger	15	0.09315	0.06196	0.1566
Smoking (Ever)	HTN	MR Egger	15	-0.9451	0.5009	0.08171
Smoking (Ever)	DM	MR Egger	15	-0.1676	0.1439	0.265

Sensitivity analysis

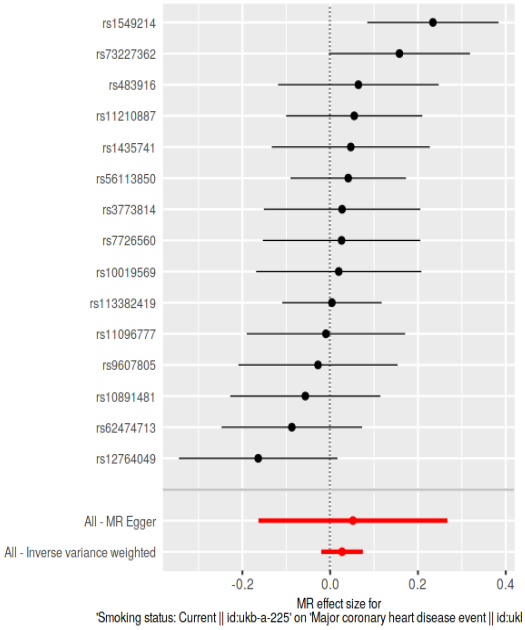
Heterogeneity

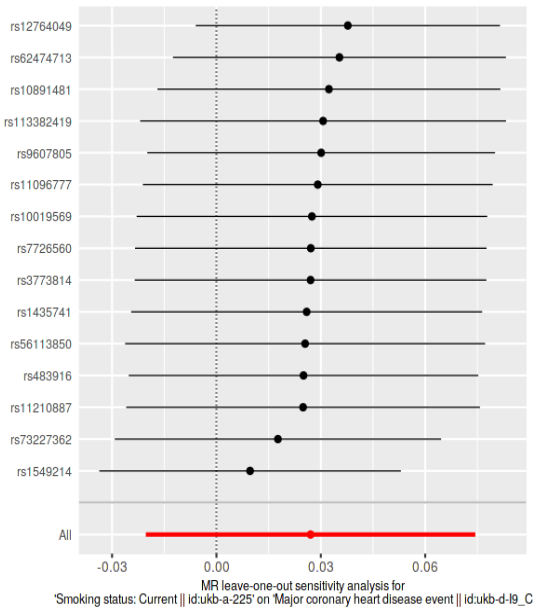
Table 8.37. Summary-level MR (Heterogeneity analysis)

ID [Exposure]	Outcome	Method	P
Smoking (Ever)	**CHD**	MR Egger	0.1557
Smoking (Ever)	**Stroke**	MR Egger	0.9444
Smoking (Ever)	**HTN**	MR Egger	$3.036e^{-9}$
Smoking (Ever)	**DM**	MR Egger	0.2002

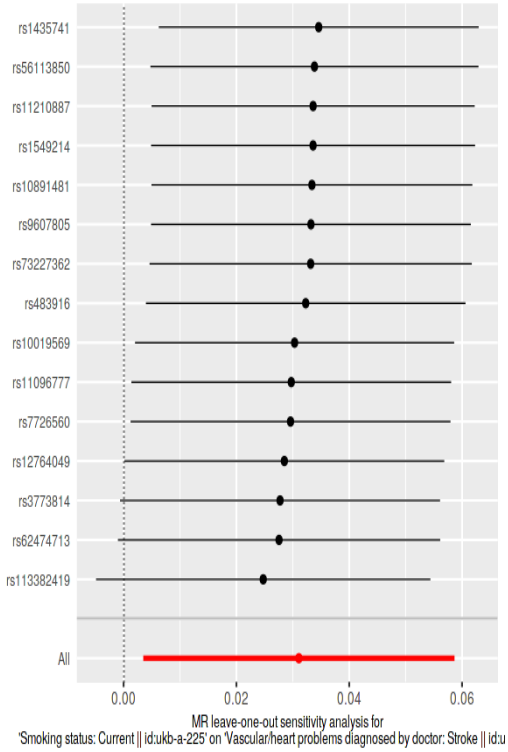
Figures 8.10. MR Egger and single SNP findings for smoking status and CMDs

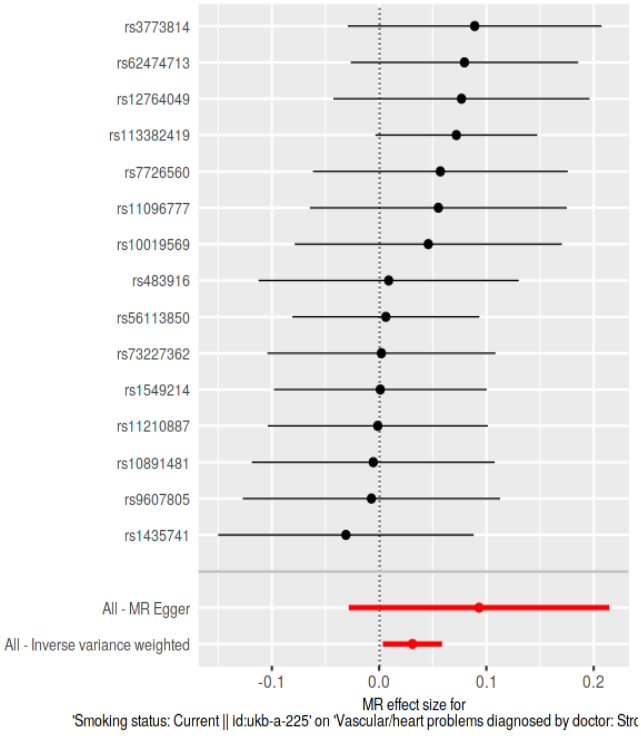
CHD:



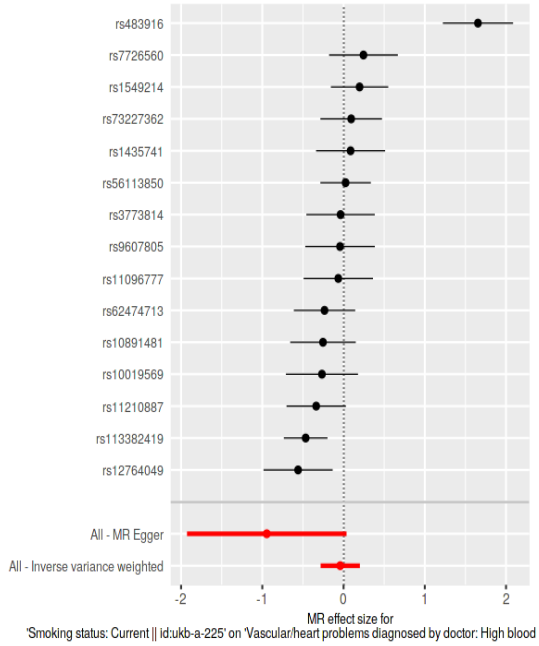


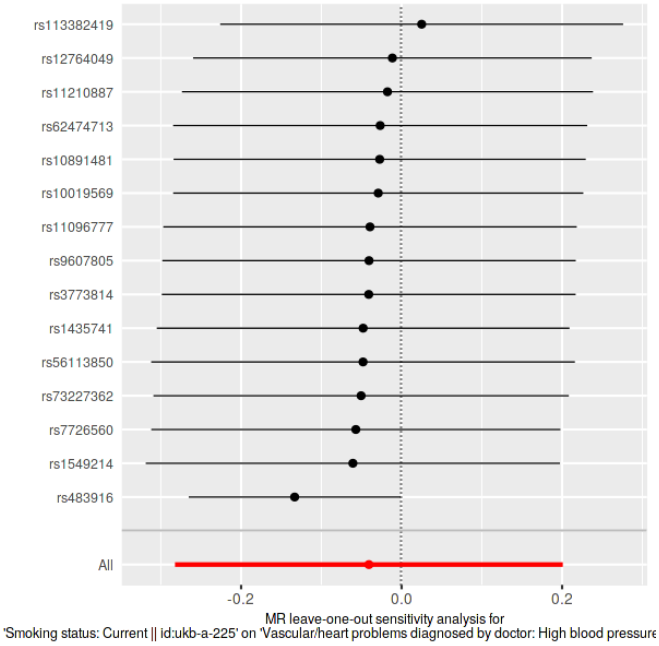
Stroke:



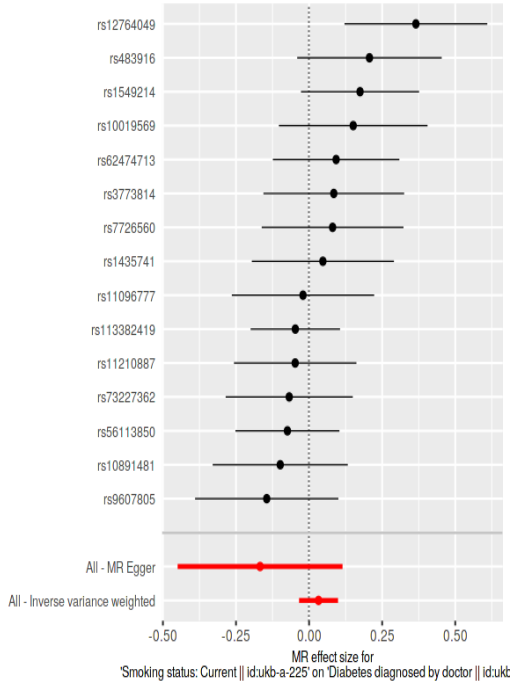


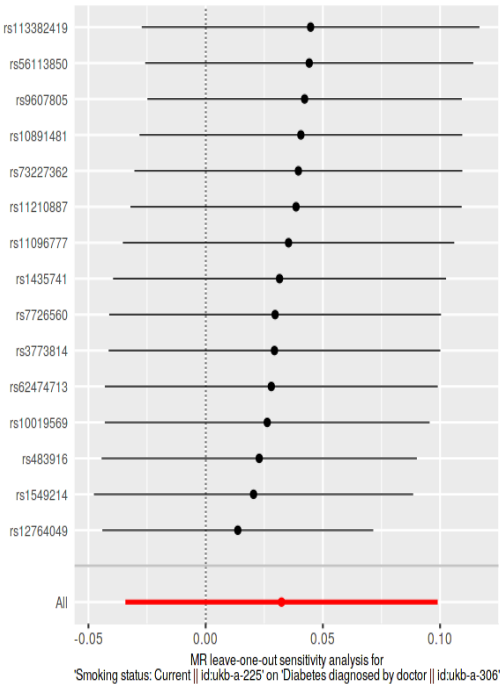
HTN:


rs483916
rs7726560
rs1549214
rs73227362
rs1435741
rs56113850
rs3773814
rs9607805
rs11096777
rs62474713
rs10891481
rs10019569
rs11210887
rs113382419
rs12764049
All - MR Egger
All - Inverse variance weighted
-2
-1
0
1
2
MR effect size for
'Smoking status: Current || id:ukb-a-225' on 'Vascular/heart problems diagnosed by doctor: High blood


rs113382419
rs12764049
rs11210887
rs62474713
rs10891481
rs10019569
rs11096777
rs9607805
rs3773814
rs1435741
rs56113850
rs73227362
rs7726560
rs1549214
rs483916
All
-0.2
0.0
0.2
MR leave-one-out sensitivity analysis for
'Smoking status: Current || id:ukb-a-225' on 'Vascular/heart problems diagnosed by doctor: High blood pressure

DM:


rs12764049
rs483916
rs1549214
rs10019569
rs62474713
rs3773814
rs7726560
rs1435741
rs11096777
rs113382419
rs11210887
rs73227362
rs56113850
rs10891481
rs9607805
All - MR Egger
All - Inverse variance weighted
-0.50
-0.25
0.00
0.25
0.50
MR effect size for
'Smoking status: Current || id:ukb-a-225' on 'Diabetes diagnosed by doctor || id:ukb


rs113382419
rs56113850
rs9607805
rs10891481
rs73227362
rs11210887
rs11096777
rs1435741
rs7726560
rs3773814
rs62474713
rs10019569
rs483916
rs1549214
rs12764049
All
-0.05
0.00
0.05
0.10
MR leave-one-out sensitivity analysis for
'Smoking status: Current || id:ukb-a-225' on 'Diabetes diagnosed by doctor || id:ukb-a-306'

Chapter: Five (8.5)

Methods

Linear regression assumptions

The assumptions for linear regression were met before proceeding to the analysis. These assumptions are:

1] Normally distributed variables

2] Numeric DV

3] Linear relationship between the predictors and outcomes

4] No multicollinearity

5] Homoscedasticity [variability of the lipid biomarkers are constant for the whole data points]

The assumptions for linear regression analysis were tested collectively using g*vlma package* in R. The package provides a decision on whether the assumptions are satisfied or not. The decision was acceptable for most of the assumptions. Testing the assumptions were also performed manually in R. The first assumption was met using a histogram. The second assumption was met as the lipid biomarkers variables are numeric. Additionally, there was no multicollinearity (high correlation) between the predictors. The analysis was done using variance inflation fact (VIF) in the *car* package in R. The VIF score for all variables was around 1 and 2, suggesting a low correlation among the predictors. Finally, the linear relationship between the lipid biomarkers and the predictors was met however it was ranging from weak to very weak. Examining the assumptions was done for lipid biomarkers.

Results

Sample characteristics

MR sample

For MR data, only individuals who descended from European ancestry and have the genotyped SNPs for (CperD: n=25274, mean age=54.81±8.05, female=51.91%, male=48.09%) and smoking status (n=314k, mean age=56.80±7.99, female=54.02%, male=45.97%) were included. Because the MR approach is based on genetic analysis, the sample characteristics such as age, sex and other covariates will not be discussed in detail. A summary of the MR sample characteristics is shown in Table 8.38.

Table 8.38. **Sample characteristics-MR (n=25274)**

Variable	Level	Count (%)
Sex	Male	12155 (48.1%)
	Female	13119 (51.9%)
Degree (college/university)	No Degree	21059 (83.3%)
	Degree	4215 (16.7%)
Variable	**Mean (SD)**	
Cigarette Smoked per Day (CperD)	15.71 (±8.35)	
Age	54.81 (±8.05)	
Body Mass Index (BMI)	26.83 (±4.85)	
Deprivation Level (Townsend score)	0.18 (±3.48)	
Cholesterol	5.74 (±1.17)	
Low-Density Lipoproteins (LDL)	3.63 (±0.91)	
Triglycerides (TG)	*__Median__ = 1.67 (IQR=1.26) (not normally distributed)	
High-Density Lipoprotein (HDL)	1.36 (±0.37)	

Descriptive statistics

Smoking behaviour and lipid biomarkers

This section will explore the associations between smoking and lipid biomarkers descriptively.

Smoking status vs lipids

The level of cholesterol among all smoking categories is on average revolving around 5.7 mmol/L. Never smokers have the highest level of cholesterol (5.74mmol/L) compared to previous and current smokers. For LDL, previous smokers have relatively lower levels compared to current and never smokers (3.53 mmol/L). The difference between the smoking categories was apparent in TG. The level of TG was highest among current smokers and lowest among never-smokers (1.66, 1.34, respectively). On the contrary, the level of HDL was highest among never-smokers and lowest among current smokers (1.47 and 1.36, respectively).

CperD vs lipids

To depict the relationship between smoking intensity (CperD) and lipid biomarkers, a correlation coefficient (r) was used. Overall, all lipid biomarkers have either a weak or very weak correlation with CperD. Cholesterol, LDL and HDL have a negative correlation with CperD while TG has a positive correlation with CperD.

SI vs lipids

Similarly, smoking initiation (SI) has either a weak or very weak correlation with lipid biomarkers. Cholesterol, LDL and HDL have a negative correlation with SI, while TG has a positive correlation.

Table 8.39. **Descriptive analysis of smoking variables vs lipid biomarkers**

Variable		Cholesterol (Mean ± SD)	LDL (Mean ± SD)	TG (Median ± IQR)	HDL (Mean ± SD)
Smoking status	Current	5.69 ± 1.16	3.59 ± 0.9	1.66 ± 2.11	1.36 ± 0.37
	Previous	5.68 ± 1.17	3.53 ± 0.88	1.55 ± 1.87	1.45 ± 0.39
	Never	5.74 ± 1.12	3.59 ± 0.86	1.43 ± 2	1.47 ± 0.38
CperD		r = -0.02	r = -0.002	r = 0.1	r = -0.12
SI		r = 0.03	r = 0.01	r = -0.03	r = 0.08

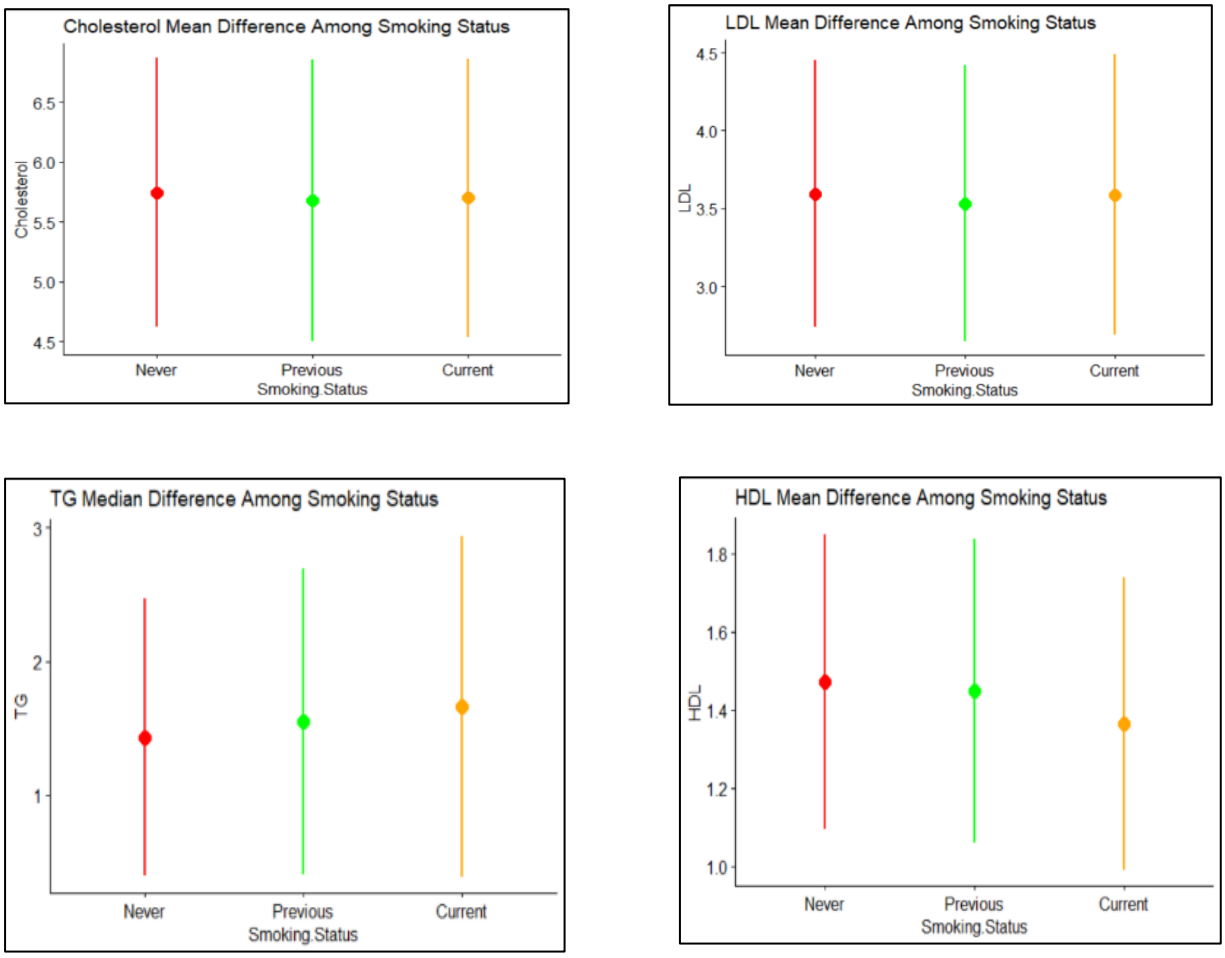


Figure 8.11: Lipid biomarkers across different smoking status categories

Lipid biomarkers vs covariates

This section explores the descriptive associations between lipid biomarkers and covariates. Means (± SD), median (± IQR) and correlation coefficient (r) was used to describe such associations.

This section describes the lipid biomarkers levels across qualitative variables. The average cholesterol and LDL levels were slightly higher among females and white British (Cholesterol: 5.9±1.1, 5.72±1.1, respectively, LDL: 3.6±0.9, 3.6±0.9, respectively). The triglycerides (TG) levels were higher among males, participants with no high degree and white British (1.7±2.2, 1.5±1.9, 1.5±1.9, respectively). Finally, HDL levels were higher in females, participants holding a high degree and white British (1.6±0.4, 1.5± 0.4, 1.5±0.4, respectively).

This section explores the correlation between lipids and quantitative covariates. The correlation between lipid variables and the covariates was generally very weak and weak. All lipid biomarkers were very weak and positively correlated with age. BMI and deprivation scores have a weak and negative correlation with cholesterol and HDL and are positively correlated with LDL and TG. Table 8.40 summarised the findings obtained from these descriptive associations.

Summary

This section discussed the descriptive associations of smoking status, CperD and SI variables with lipid biomarkers. It also included a descriptive analysis of the associations between lipid biomarkers and the covariates. The descriptive analysis gives an overview of the variables in this sample. The next section will discuss the inferential analysis of these variables observationally.

Table 8.40: Descriptive analysis of lipids vs covariates

Variables	Sex (Mean ± SD)		Degree (Mean ± SD)		Ethnicity (Mean ± SD)	
	Male	**Female**	**Yes**	**No**	**White British**	**Other ethnicities**
Cholesterol	5.5±1.1	5.9±1.1	5.7±1.1	5.7±1.2	5.72±1.1	5.5±1.1
LDL	3.5±0.9	3.6±0.9	3.6±0.9	3.6±0.9	3.6±0.9	3.5±0.9
TG (Median ± IQR)	1.7±2.2	1.3±1.7	1.4±1.8	1.5±1.9	1.5±1.9	1.4±1.7
HDL	1.3±0.3	1.6±0.4	1.5± 0.4	1.4±0.4	1.5±0.4	1.4±0.4

	Age	**BMI**	**Townsend**
Cholesterol	r = 0.06	r = -0.01	r = -0.06
LDL	r = 0.04	r = 0.02	r = 0.05
TG	r = 0.07	r = 0.29	r = 0.03
HDL	r = 0.04	r = -035	r = -0.06

Observational analysis

Ever vs never smoking (smoking as a binary variable)

After examining the associations using smoking status as a binary (ever vs never), the individuals who ever smoked have a lower level of cholesterol compared to never-smoked individuals (B: -0.05, 95% CI: -0.05 – -0.04, P<0.001). However, after adjustment, the association between smoking and cholesterol becomes non-significant (B: -0.001, 95% CI: -0.01 – 0.002, P=0.428). The rest of the associations and all lipid biomarkers are summarised in table 8.41***(a-d)***.

Table 8.41(a-d): Linear regression analysis of smoking status (ever vs never) vs lipid biomarkers

a) Cholesterol		Unadjusted			Adjusted			
		B	**CI**	**P**	**B**	**β**	**CI**	**P**
Smoking Status	Ever	-0.05	-0.05 – -0.04	**<0.001**	-0.00	-0.00	-0.01 – 0.00	0.428
Sex [Male**]**					-0.39	-0.34	-0.39 – -0.38	**<0.001**
Education [Have Degree**]**					0.02	0.02	0.01 – 0.03	**<0.001**
Ethnicity [White British**]**					0.14	0.12	0.13 – 0.15	**<0.001**
Age					0.01	0.05	0.01 – 0.01	**<0.001**
Townsend					-0.02	-0.04	-0.02 – -0.02	**<0.001**
BMI					-0.01	-0.03	-0.01 – -0.01	**<0.001**

b) LDL		Unadjusted			Adjusted			
		B	**CI**	**P**	**B**	**β**	**CI**	**P**
Smoking Status	Ever	-0.04	-0.04 – -0.03	<**0.001**	-0.02	-0.03	-0.03 – -0.02	<**0.001**
Sex [Male**]**					-0.15	-0.17	-0.15 – -0.14	<**0.001**
Education [Have Degree**]**					0.002	0.01	-0.004 – 0.01	0.113
Ethnicity [White British**]**					0.09	0.10	0.08 – 0.10	<**0.001**
Age					0.009	0.03	0.001 – 0.05	<**0.001**
Townsend					-0.01	-0.04	-0.01 – -0.01	<**0.001**
BMI					0.01	0.03	0.00 – 0.01	<**0.001**

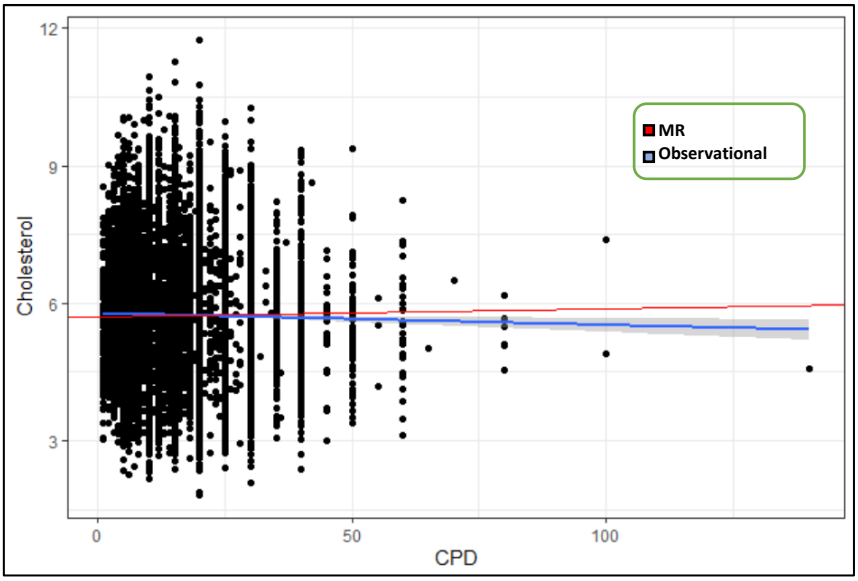
c) TG		Unadjusted			Adjusted			
		B	**CI**	**P**	**B**	**β**	**CI**	**P**
Smoking Status	Ever	0.17	0.16 – 0.17	<**0.001**	0.08	0.08	0.08 – 0.09	<**0.001**
Sex [Male**]**					0.37	0.36	0.37 – 0.38	<**0.001**
Education [Have Degree**]**					-0.07	-0.06	-0.07 – -0.06	<**0.001**
Ethnicity [White British**]**					0.06	0.06	0.06 – 0.07	<**0.001**
Age					0.01	0.04	0.00 – 0.01	<**0.001**
Townsend					0.00	0.00	0.00 – 0.00	**0.003**
BMI					0.06	0.26	0.06 – 0.06	<**0.001**

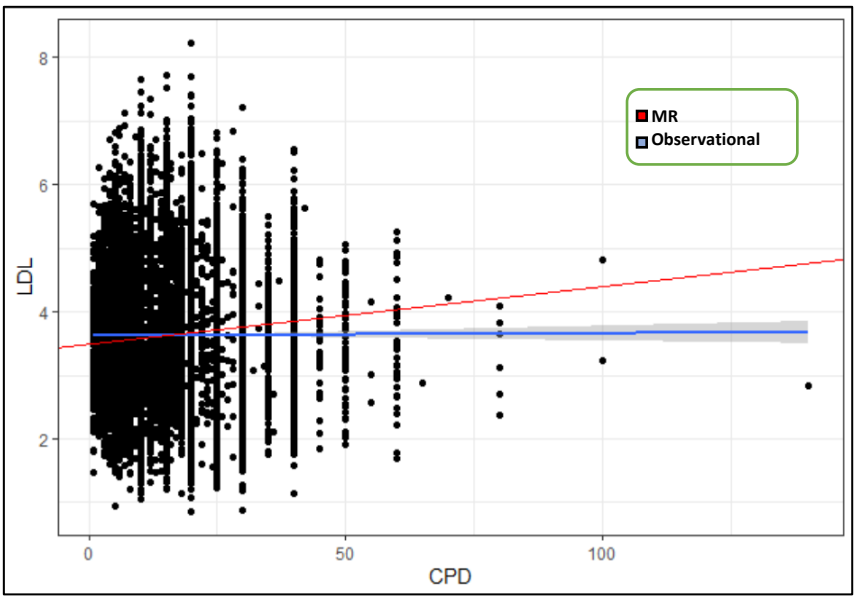
d) HDL		Unadjusted			Adjusted			
		B	**CI**	**P**	**B**	**β**	**CI**	**P**
Smoking Status	Ever	-0.04	-0.04 – -0.03	<**0.001**	0.01	0.02	0.01 – 0.01	<**0.001**
Sex [Male**]**					-0.30	-0.78	-0.30 – -0.30	<**0.001**
Education [Have Degree**]**					0.03	0.08	0.03 – 0.03	<**0.001**
Ethnicity [White British**]**					0.02	0.06	0.02 – 0.03	<**0.001**
Age					0.00	0.07	0.00 – 0.00	<**0.001**
Townsend					-0.00	-0.02	-0.00 – -0.00	<**0.001**
BMI					-0.03	-0.31	-0.03 – -0.02	<**0.001**

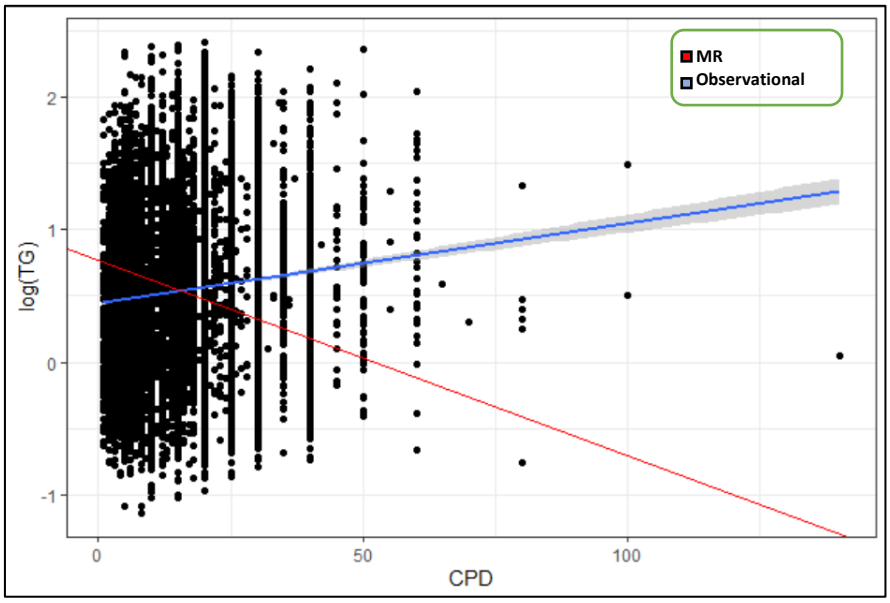
MR

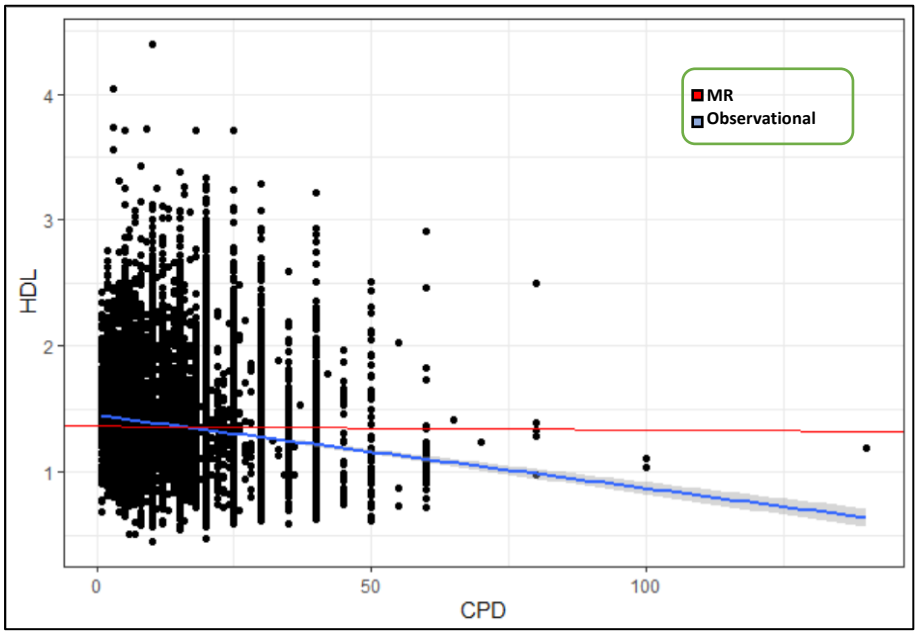
Plots

Figure 8.12: CPD vs lipid biomarkers (observational vs MR)






MR
Observational
log(TG)
2
1
0
-1
0
50
100
CPD


MR
Observational
HDL
4
3
2
1
0
50
100
CPD

Individual SNPs analysis

This section will focus on performing MR for individual SNPs against lipids variables. All SNPs that have a GWAS-level significance with CperD were included in the MR analysis. There were no significant associations between CperD and lipid biomarkers among individual SNPs (Table 8.42).

Table 8.42: MR analysis of CperD and lipid biomarkers for individual SNPs

Variables/SNPs	Cholesterol		LDL		TG		HDL	
	MR Estimate	P	MR Estimate	P	MR Estimate	P	MR Estimate	P
rs1051730_A	0.00427	0.707	0.008627	0.328	0.0016	0.750	-0.0048	0.181
rs1317286_G	0.00358	0.7524	0.00803	0.361	0.0015	0.762	-0.0045	0.208
rs12914385_T	0.0116	0.302	0.0145	0.097	0.00314	0.5319	-0.0044	0.223
rs16969968_A	0.0049	0.665	0.0090	0.307	0.00146	0.774	-0.0046	0.206
rs8034191_C	0.0111	0.332	0.0123	0.164	0.0039	0.440	-0.0027	0.458
rs951266_A	0.0051	0.652	0.0091	0.302	0.0018	0.712	-0.00508	0.164
rs17486278_C	0.00475	0.677	0.0090	0.306	0.0017	0.734	-0.0049	0.175
rs72740964_A	0.0041	0.718	0.0084	0.346	0.0016	0.747	-0.00483	0.189
rs55853698_G	0.0063	0.585	0.00964	0.286	0.0029	0.576	-0.0044	0.231
rs8040868_C	0.0125	0.283	0.01538	0.090	0.00010	0.984	-0.0029	0.434
rs11637630_G	0.0184	0.250	0.0168	0.167	0.01002	0.1567	-0.0033	0.530
rs938682_G	0.01805	0.259	0.0164	0.177	0.0101	0.1498	-0.0032	0.539
rs12910984_G	0.0179	0.262	0.0165	0.175	0.0101	0.151	-0.0034	0.506
rs6474412_C	0.0192	0.438	0.0200	0.302	-0.0019	0.8607	-0.0056	0.484
rs2229961_A	0.0079	0.796	0.0171	0.480	0.0061	0.655	-0.0127	0.212

Smoking status vs Outcomes

Table 8.43: MR analysis of the smoking status and lipid biomarkers for individual SNPs

SNPs	Variables							
	Cholesterol		LDL		TG		HDL	
	B	P	B	P	B	P	B	P
rs1317286	0.00356	0.752	0.008005	0.361	0.001542	0.763	-0.004718	0.21
rs1051730	0.004235	0.707	0.008577	0.328	0.001621	0.751	-0.004991	0.185
rs12914385	0.01133	0.3	0.014186	0.0947	0.003081	0.533	-0.004409	0.227
rs16969968	0.004889	0.665	0.008967	0.306	0.001457	0.775	-0.004722	0.21
rs8034191	0.01096	0.33	0.012232	0.162	0.003909	0.443	-0.002770	0.461
rs951266	0.005091	0.652	0.009068	0.301	0.001882	0.713	-0.005206	0.167
rs17486278	0.004709	0.677	0.008991	0.305	0.001729	0.735	-0.005069	0.179
rs72740964	0.004081	0.718	0.008298	0.345	0.001646	0.748	-0.004923	0.193
rs55853698	0.006137	0.585	0.009337	0.285	0.002828	0.578	-0.004456	0.235
rs8040868	0.01174	0.28	0.014445	0.0872	.000009	0.985	-0.002825	0.437
rs11637630	-0.01501	0.245	-0.01398	0.163	-0.0082	0.158	0.002683	0.534
rs938682	-0.01470	0.255	-0.01366	0.173	-0.00839	**0.**151	0.002628	0.542
rs12910984	-0.01461	0.258	-0.01374	0.171	-0.0083	0.153	0.002843	0.51
rs6474412	-0.01005	0.433	-0.0104	0.295	0.00102	0.86	0.00297	0.489
rs2229961	0.0099	0.796	0.02132	0.475	0.0077	0.657	-0.01597	0.211

Two-Sample MR

SNPs summary statistics

CperD and cholesterol summary statistics

Table 8.44. Beta and SD are used in summary-level MR (CperD vs cholesterol)

SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	0.0042	0.011
rs7599488	0.026414	0.005544	-0.0131	0.0107
rs215614	-0.04448	0.005728	0.0153	0.01098
rs73229090	0.055489	0.008763	0.0051	0.017
rs6474412	0.067113	0.006613	-0.01005	0.0129
rs3025343	0.063352	0.008661	-0.0128	0.0348
rs8034191	0.182567	0.005889	0.0109	0.011
rs2229961	0.207114	0.023481	0.0099	0.038
rs12910984	0.1571	0.006687	-0.0146	0.013
rs3733829	0.034965	0.005811	0.0103	0.011
rs3865453	-0.10241	0.010329	0.00179	0.020554
rs28399443	-0.23131	0.017486	-0.0085	0.035
rs7260329	-0.04462	0.006017	-0.0023	0.0116
rs2273506	0.06063	0.011182	0.0063	0.0217

CperD and LDL summary statistics

Table 8.45. Beta and SD are used in summary-level MR (CperD vs LDL)

SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	0.008	0.008
rs7599488	0.026414	0.005544	-0.0089	0.0083
rs215614	-0.04448	0.005728	0.011	0.0085
rs73229090	0.055489	0.008763	-0.003	0.0132
rs6474412	0.067113	0.006613	-0.01044	0.009
rs3025343	0.063352	0.008661	-0.015	0.0269
rs8034191	0.182567	0.005889	0.012	0.008
rs2229961	0.207114	0.023481	0.0213	0.0298
rs12910984	0.1571	0.006687	-0.013	0.01
rs3733829	0.034965	0.005811	0.0126	0.00865
rs3865453	-0.10241	0.010329	-0.0071	0.0159
rs28399443	-0.23131	0.017486	-0.0144	0.027
rs7260329	-0.04462	0.006017	-0.0022	0.009
rs2273506	0.06063	0.011182	0.013	0.0168

CperD and TG summary statistics

Table 8.46. Beta and SD are used in summary-level MR (CperD vs TG)

SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	-0.0009	0.0113
rs7599488	0.026414	0.005544	-0.0152	0.0108
rs215614	-0.04448	0.005728	0.0158	0.011
rs73229090	0.055489	0.008763	-0.02179	0.01714
rs6474412	0.067113	0.006613	-0.0034	0.01291
rs3025343	0.063352	0.008661	0.06556	0.0348
rs8034191	0.182567	0.005889	0.0033	0.0113
rs2229961	0.207114	0.023481	-0.0165	0.0387
rs12910984	0.1571	0.006687	-0.01886	0.013
rs3733829	0.034965	0.005811	-0.0106	0.0112
rs3865453	-0.10241	0.010329	0.00692	0.0207
rs28399443	-0.23131	0.017486	0.0611	0.035
rs7260329	-0.04462	0.006017	-0.004	0.0116
rs2273506	0.06063	0.011182	-0.00025	0.02189

CperD and HDL summary statistics

Table 8.47. Beta and SD are used in summary-level MR (CperD vs HDL)

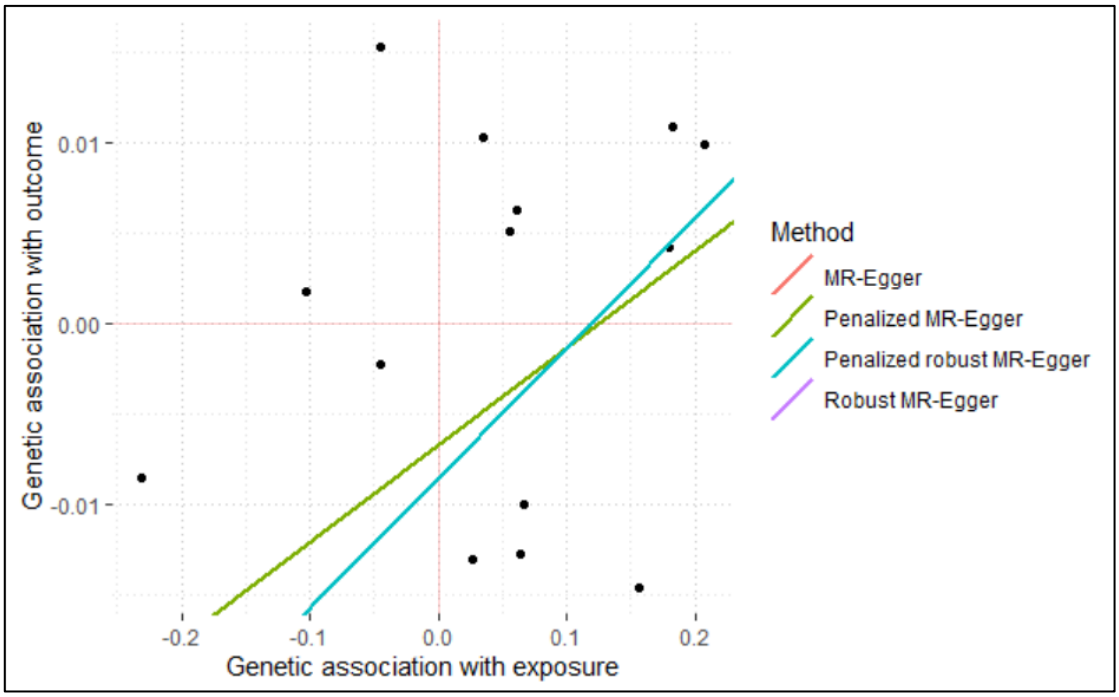
SNP	bx	bxse	by	byse
rs1051730	0.179517	0.005915	-0.00499	0.0037
rs7599488	0.026414	0.005544	0.0014	0.00358
rs215614	-0.04448	0.005728	-0.00159	0.0036
rs73229090	0.055489	0.008763	0.00781	0.0056
rs6474412	0.067113	0.006613	0.0029	0.00428
rs3025343	0.063352	0.008661	0.0099	0.0114
rs8034191	0.182567	0.005889	-0.00277	0.0037
rs2229961	0.207114	0.023481	-0.0159	0.01276
rs12910984	0.1571	0.006687	0.0028	0.0043
rs3733829	0.034965	0.005811	-0.00163	0.0037
rs3865453	-0.10241	0.010329	0.0089	0.00685
rs28399443	-0.23131	0.017486	0.0113	0.0116
rs7260329	-0.04462	0.006017	0.00097	0.00388
rs2273506	0.06063	0.011182	-0.0069	0.0072

2SMR in R

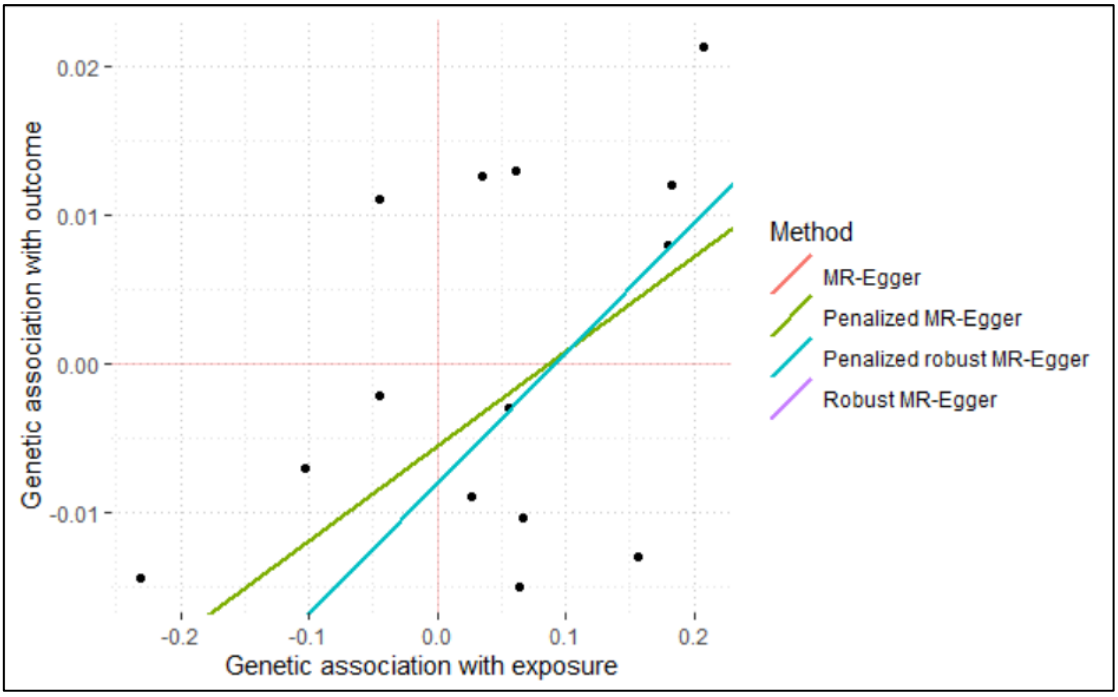
Plot

Figures 8.13: CPD vs lipid biomarkers (summary-level MR)

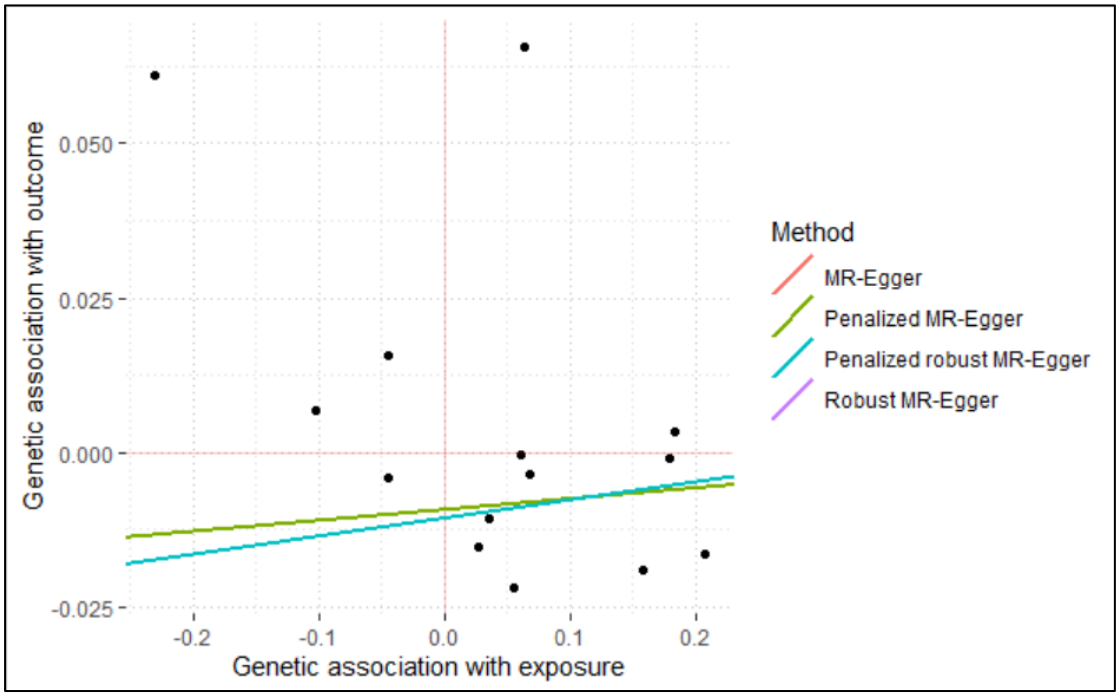
CperD vs Cholesterol



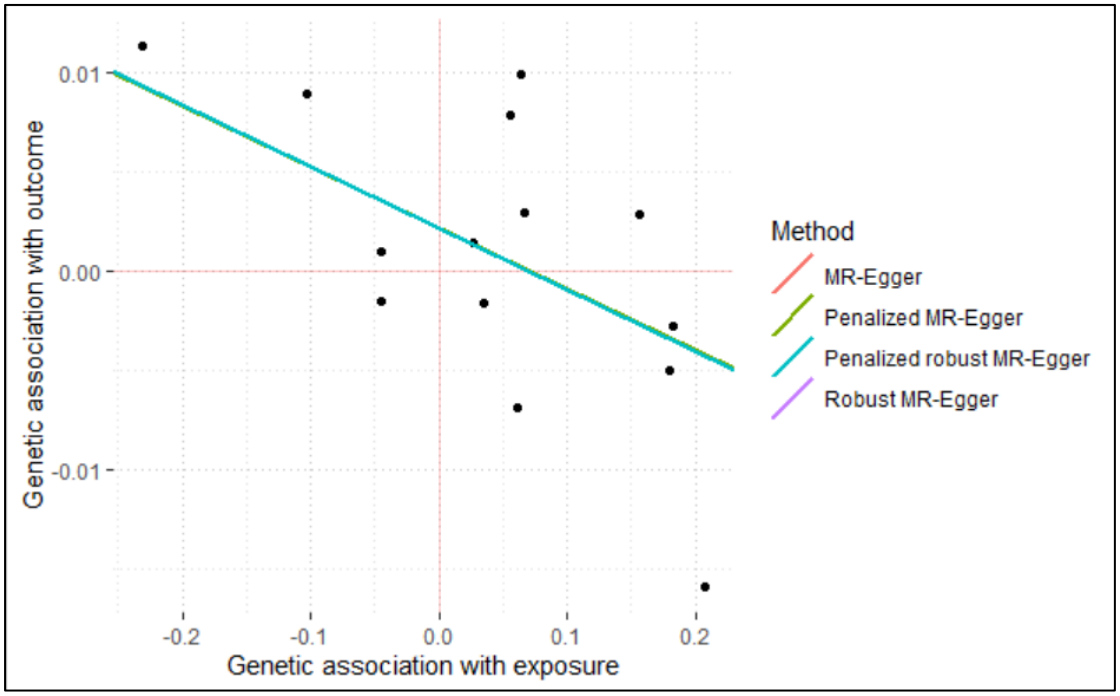
CperD vs LDL



CperD vs TG



CperD vs HDL


0.01
0.00
-0.01
Genetic association with outcome
-0.2
-0.1
0.0
0.1
0.2
Genetic association with exposure
Method
MR-Egger
Penalized MR-Egger
Penalized robust MR-Egger
Robust MR-Egger

Smoking status MR (~314k)

IV assumptions results:

Table 8.48. IV assumptions (smoking status)

CMDs	Genetic Score	
	Estimates	*p*
Cholesterol	0.0002	0.775
LDL	-2.333e^{-05}	0.972
TG	-0.0007	0.078
HDL	0.0004	0.146
Variables	**Genetic Score**	
	Estimates	*p*
Smoking status	0.0080 (OR=1.01)	**1.35e^{-09}**
Age	-0.0004	0.449
Degree [No]	-0.022	**0.0125**
Sex [Male]	-0.003	0.694
Townsend	0.001	0.433
BMI	-0.005	**44e^{-08}**
PC1	-0.015	**6.9e^{-14}**
PC2	0.1458	**2e^{-16}**
PC3	0.0424	**2e^{-16}**
PC4	0.915	**2e^{-16}**
PC5	-0.087	**2e^{-16}**
PC6	1.535	**2e^{-16}**
PC7	-0.167	**2e^{-16}**
PC8	-0.424	**2e^{-16}**
PC9	-1.011	**2e^{-16}**
PC10	0.458	**2e^{-16}**

Details of MR analysis

Ever vs never MR analysis

Table 8.49. Summary-level MR (Smoking status)

ID [Exposure]	Outcome	Method	Number of SNPs	Beta	SE	P value
Smoking (Ever) [ukb-a-225]	Cholesterol [ukb-d-30690]	MR Egger	15	-0.85	0.989	0.4057
Smoking (Ever) [ukb-a-225]	LDL [ukb-d-30780_raw]	MR Egger	15	-0.9402	0.9868	0.3581
Smoking (Ever) [ukb-a-225]	TG [ukb-d-30870_raw]	MR Egger	15	1.652	1.637	0.3314
Smoking (Ever) [ukb-a-225]	HDL [ukb-d-30760_raw]	MR Egger	15	-0.3172	0.3529	0.3851

Sensitivity analysis

Heterogeneity

Table 8.50. Summary-level MR (Heterogeneity analysis)

ID [Exposure]	Outcome	Method	P
Smoking (Ever) [ukb-a-225]	Cholesterol [ukb-d-30690]	MR Egger	0.000117
Smoking (Ever) [ukb-a-225]	LDL [ukb-d-30780_raw]	MR Egger	0.000001556
Smoking (Ever) [ukb-a-225]	TG [ukb-d-30870_raw]	MR Egger	$3.318e^{-16}$
Smoking (Ever) [ukb-a-225]	HDL [ukb-d-30760_raw]	MR Egger	0.0005869