Author:
Carter, Alice R

Title:
Strengthening causal inference in educational inequalities in cardiovascular disease

Strengthening causal inference in educational inequalities in cardiovascular disease

Alice R Carter

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of PhD in the Faculty Health Sciences

MRC Integrative Epidemiology Unit, Population Health Sciences,

University of Bristol, United Kingdom

September 2020

Word count: 56 764

Abstract

Despite reducing rates of cardiovascular disease in high income countries, individuals who are the most socioeconomically deprived remain at the highest risk of disease. The mechanisms by which the inequalities arise are still unknown. In this thesis I use causal inference methods, including Mendelian randomisation (MR), mediation analysis and polygenic scores, to understand the aetiology of educational inequalities in cardiovascular disease, using UK Biobank.

Establishing causality in epidemiology can be challenging, due to unmeasured (or mis-measured) confounding, measurement error and reverse causality. One method to overcome these sources of bias is MR. In this thesis I demonstrate using simulations and applied examples how MR can be applied to mediation analysis, identifying sources of bias and methodological limitations.

Using MR mediation methods and non-genetic (phenotypic) mediation methods I demonstrate that body mass index, systolic blood pressure and lifetime smoking behaviour mediate up to 40% of the association between education and cardiovascular disease. Intervening on these intermediate risk factors would likely reduce cases of cardiovascular disease attributable to low educational attainment.

I then investigate inequalities in prescribing of statins as a primary cardiovascular preventative medication. I identified clear inequalities, where for a given level of underlying cardiovascular risk (assessed via QRISK3 score) individuals with lower educational attainment were less likely to receive statins.

Finally, explore the role of education as an effect modifier of genetic suscepbility to cardiovascular disease.I demonstrate that on the additive scale, higher education protects against genetic susceptibility to body mass index and smoking but accentuates genetic susceptibility to low-density lipoprotein cholesterol and systolic blood pressure. On the multiplicative scale, higher education accentuates genetic susceptibility to atrial fibrillation and coronary heart disease.

This thesis demonstrates that body mass index, systolic blood pressure, smoking and statin use all likely contribute to educational inequalities in cardiovascular disease, whilst contributing to the development of methods to improve causal inference in social epidemiology.

Acknowledgements

First and foremost, my biggest thanks and gratitude goes to my supervisors, Laura Howe, Neil Davies, Amy Taylor and George Davey Smith. The support, encouragement, teaching, generosity, friendship, kindness, patience and passion they have shown me knows no bounds. They have encouraged me to follow my research interests, to break out of my comfort zone and to put myself and my work out there (as well as having a good cry when it all gets too much – extra thanks to the tissue supply in Laura's office). I feel exceptionally lucky that I have had the most enjoyable four years during my PhD, and I know this has been down to having the best group of supervisors I could have wished for, so thank you.

Secondly, I would like to thank the Medical Research Council Integrative Epidemiology Unit, not only for funding my PhD, but for providing such a welcoming, friendly and accommodating environment to complete my PhD. I have been lucky enough to work with a number of wonderful colleagues from whom I have learnt so much. I would especially like to thank all of the PALS (Postdocs and PhDs of Abi and Laura) from over the years, who have taught me so much. I would like to thank Marcus Munafò and Nic Timpson for their feedback during my annual reviews. This was invaluable and helped shape so much of this work. Special thanks also go to all of the A-team for helping arrange my supervision meetings, fixing my messed-up travel bookings and helping to organise my MR mediation workshop.

I have been lucky enough to publish frequently throughout my PhD and all co-authors have provided invaluable support, so thank you to all of you. In particular, I would like to thank Deborah Lawlor, Carolina Borges, Eleanor Sanderson, Sean Harrison, Teri-Louise North and Dipender Gill. These co-authors have gone above and beyond in their support and assistance in our publications and I look forward to continuing with these collaborations. Thank you also to the peer reviewers of these publications, which helped develop both my research and me as a researcher.

This work would not have been possible without the UK Biobank resource. Thank you to all participants who have so generously given up their time and data to make this possible.

I wouldn't have made it through these years without a wonderful set of friends, so thank you to the Epi Queens, Anastasia, Elise and Isha. I'm so glad we've been on this journey together and can't wait to see what epidemiological breakthroughs we all achieve in the years to come. To Sophie, for the friendship and numerous PhD fuelled rants over the years. Here's to many more years of cocktails and Christmas films with a little less complaining. To Mary and

George, for the lovely PhD free activities. And to Katie, you might not be here to celebrate, but I know the years of friendship shaped so much of who I am.

Finally, thank you to my family. This PhD would not have been possible without so many sacrifices from so many of my people. To Alex (and Nellie) for the being the best PhD distractions. To Nan and Grandad for always supporting me, even if you weren't quite sure why I still wanted to be a student in my late twenties. To Grace and Matt, and more recently Darcey. If it wasn't for your generosity, I would never have made it through my MSc, let alone PhD. And last but not least, thank you to my Mum for being my biggest supporter and role model over the years. You taught me to never, ever, ever, give up. This achievement is as much yours as it is mine.

Author declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: DATE: 8th September 2020

Abbreviations used in this thesis

Abbreviation	Term
CVD	Cardiovascular disease
BMI	Body mass index
C	Measured confounder
CDE	Controlled direct effect
CHD	Coronary heart disease
CI	Confidence interval
DIAGRAM	DIAbetes Genetic Replication and Meta-analysis Consortium
DNA	Deoxyribonucleic acid
GIANT	Genetic Investigation of ANthropometric traits
GBD	Global Burden of Disease
GSCAN	GWAS and Sequencing Consortium of Alcohol and Nicotine use
GWAS	Genome-wide association study
HDL	High density lipoprotein cholesterol
HES	Hospital episode statistics
HSE-SHS	Health Surveys for England and Scottish Health Surveys
ICD	International classification of disease
ISCED	International Standard Classification for Education
IV	Instrumental variable
IVW	Inverse variance weighted
LD	Linkage disequilibrium
LDL-C	Low-density lipoprotein cholesterol
M	Mediator
MI	Myocardial infarction
MR	Mendelian randomisation
MRC	Medical Research Council
MVMR	Multivariable Mendelian randomisation

NDE	Natural direct effect
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIE	Natural indirect effect
OR	Odds ratio
PC	Principal Components
PGS	Polygenic score
PHE	Public Health England
RCT	Randomised controlled trial
RoSLA	Raising of School Leaving Age
SBP	Systolic blood pressure
SD	Standard deviation
SEP	Socioeconomic position
SMR	Scottish morbidity records
SNP	Single nucleotide polymorphism
SSGAC	Social Science Genetic Association Consortium
TDI	Townsend deprivation index
TSMR	Two-step Mendelian randomisation
U	Unmeasured confounder
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
X	Exposure
Y	Outcome

Table of Contents

List of figures

List of tables

List of appendix tables

Chapter 1. Introduction

This introductory chapter outlines the aims and objectives of my thesis and provides a brief description of the main topics and work covered.

1.1 Socioeconomic inequalities in cardiovascular disease

Cardiovascular disease (CVD) is the leading cause of mortality worldwide. The 2017 Global Burden of Disease study estimated that CVD accounted for one third of all deaths globally (1). Insights into aetiological mechanisms have improved prevention through the modification of risk factors and age-standardised rates for prevalent cases of CVD in high-income countries are declining; however, CVD mortality has plateaued in these same countries (1, 2).

Socioeconomic position (SEP) has long been associated with increased morbidity and mortality from CVD in high income countries. The Whitehall study of British civil servants, which began in 1967, provided much of the early data on this; although socioeconomic differences in CVD were observed for many years prior to this (3, 4). Despite overall reductions in morbidity and mortality since the Whitehall studies, these social class differences have persisted in the population (5, 6). Additionally, there is evidence to show that the effect of SEP accumulates across the life course (7). The causal effect of education on CVD has recently been demonstrated using non-genetic instrumental variable (IV) methods and genetic IV methods (Mendelian randomisation [MR]) (8, 9).

SEP can be measured in a number of ways both at the population level, such as a postcode deprivation index or at the individual level typically from occupation, income or educational attainment (10). In this thesis I proxy SEP by measuring educational attainment, where self-reported education is mapped to the International Standard Classification for Education (ISCED) years of schooling measure United Nations Educational, Scientific and Cultural Organization (UNESCO (11).

Education may prevent CVD, in part, through its effects on modifiable risk factors for CVD, including body mass index, smoking and systolic blood pressure (12-14). Intervening on education is difficult to achieve without social and political reform. Targeting intermediate risk factors for CVD could therefore help to reduce educational inequalities in CVD risk. In this thesis I hope to identify where interventions may be possible to reduce inequalities in CVD.

1.2 Data used in this thesis

The main data used throughout this thesis is the UK Biobank study (Chapter 3-Chapter 6). The UK Biobank is a population-based cohort study which recruited 503,317 UK adults between 2006 and 2010. Participants attended baseline assessment centres involving questionnaires, interviews, anthropometric, physical and genetic measurements (15, 16) and have been periodically followed up following baseline both in clinics and online questionnaires. Additionally, the UK Biobank has data linked with hospital episode statistics (in England and Wales), Sottish morbidity records and death registers, meaning clinical outcomes are routinely updated. Full details of the UK Biobank and the data and measurements used are described in detail in each results chapter (Chapters 3-6).

1.3 Statistical methods used in this thesis

UK Biobank is an incredibly rich data source, with large amounts of both phenotypic and genetic data available. A number of methods are used in this PhD which allow for the incorporation of genetic data with phenotypic data for robust causal inference. In Chapter 3 and Chapter 4, the main method used is MR; an IV approach using genetic variants to instrument modifiable phenotypic exposures. Given the random allocation of genetic variants at meiosis, genetic variants provide suitable instruments for many exposures of interest in epidemiology; including educational attainment (17).

Where these assumptions hold, MR can be used to obtain estimates of causal effects that are robust to non-differential measurement error and confounding of the exposure-outcome relationship (18). Methods have been developed, including two-step (network) MR and multivariable MR (19-22), which allow us to estimate the mediating effects of risk factors and begin to disentangle the causal pathways. These methods, and the assumptions of them, are described in detail in Chapter 3 and applications of these methods are described in Chapter 3 and Chapter 4.

In Chapter 5, I derive a cardiovascular risk score, QRISK3 (23), used in general practices in England to determine 10-year cardiovascular risk and statin prescriptions as primary cardiovascular prevention (24-26). Using self-reported educational attainment, I consider how interactions may arise between education and QRISK3 score to lead to inequalities in access to statins.

In Chapter 6 I consider whether and how educational inequalities may exist in genetic risk for cardiovascular disease. Using a number of polygenic scores (PGS) for cardiovascular risk

factors and cardiovascular outcomes, I investigate effect modification by educational attainment.

1.4 Thesis outline

This PhD thesis begins with an introductory chapter, briefly introducing the rationale for studying socioeconomic inequalities in CVD, as well as the data and statistical methods used. Additionally, the primary aim and objectives are outlined. Chapter 2 is a literature review of what we currently understand about the association between education and CVD. It explores in detail the risk factors currently understood to be implicated in this association and the methods that have typically been used to explore the topic. It ends with a review of causal inference methods that have recently been developed to strengthen our causal understanding about the effect of education on CVD. A detailed description of UK Biobank is presented in each results chapter (Chapter 3-Chapter 6) including how variables were selected and assessed for use in this thesis as well as the statistical methods used in each chapter. Chapter 3, is the first of my analysis chapters. This chapter explores the use of mediation analysis in MR. Using the motivating example of the effect of education on systolic blood pressure, with mediation by body mass index, I present results comparing different analytical methods for mediation analysis, including in an MR framework. I explore a range of research scenarios and show which methods introduce bias, given the variables included in the analysis, and which methods are robust to bias in the examples. These methods are then used in the subsequent analysis chapter. In Chapter 4, I present and discuss results showing the mediating role of body mass index (BMI), systolic blood pressure and smoking in the association between education and CVD. In Chapter 5, I show how there are educational inequalities in access to statin treatment in the primary care setting, investigating interactions between 10-year risk of cardiovascular disease and educational attainment on statins. In Chapter 6, I then consider whether education acts as an effect modifier of genetic risk scores for cardiovascular risk factors, which might contribute to the accumulation of excess cardiovascular risk in people with low educational attainment. There is also a brief discussion of the results in all of these chapters. Finally, in Chapter 7, I discuss the findings of each analysis chapter in the context of the wider research aim and their implications. I discuss the overall strengths and limitations of my research, along with opportunities for future research. I discuss how the methods used here allow us to make causal inference about the objectives studied and how this could be used to reduce socioeconomic inequalities in CVD and improve prevention mechanisms.

1.5 Thesis aims and objectives

The overarching aim of my thesis is to understand what factors are driving socioeconomic inequalities in CVD, triangulating across causal inference methods.

This aim will be achieved by addressing the following objectives:

1) Compare and contrast methods for mediation analysis, with and without the use of genetic IVs, applied to the motivating example of the roles of BMI and low-density lipoprotein cholesterol in mediating the association between education and cardiovascular outcomes
2) Investigate the causal effects of education on cardiovascular disease subtypes and the role of BMI, systolic blood pressure and smoking in mediating the association
3) Identify whether there is an interaction between education and a clinical risk score for cardiovascular disease with respect to statin prescribing
4) Investigate whether education modifies genetic susceptibility to cardiovascular disease and cardiovascular risk factors

Chapter 2. Literature review

In this chapter I will introduce the epidemiology of cardiovascular disease (CVD), including how mortality from CVD is declining and current known risk factors for disease. I will review socioeconomic position (SEP) as a risk factor for cardiovascular disease, specifically considering education as a measure of SEP. Following this, I will explore what factors may help explain the effects of education, and SEP more widely, on cardiovascular outcomes, including potential mediators or explanatory mechanisms of the association. I will introduce the methods used in this thesis, beginning with epidemiological methods and principles, with a focus on causal inference using mediation analysis and genetic epidemiology; two of the main approaches used in this thesis. Finally, I will explore how predictive risk scores are used in clinical practice and opportunities for integrating genetic risk into these scores.

2.1 Epidemiology of cardiovascular disease

Globally, CVD remains the leading cause of death, accounting for over 17.5 million deaths annually (27). In the United Kingdom (UK), there is clear evidence from a number of studies that mortality from CVD and subtypes of CVD, such as coronary heart disease (CHD), stroke and myocardial infarction (MI) are decreasing (28-31). Bhatnagar and colleagues estimate that age-standardised absolute CVD mortality has declined by 70% from 1979 to 2013 (28). However, evidence that the incidence and prevalence of CVD is decreasing is less clear, where some studies estimate the prevalence is decreasing (32, 33) and others estimate the prevalence of CVD is stable (28, 34). For subtype specific cardiovascular mortality, the Global Burden of Disease (GBD) study estimates that CHD mortality has declined by 60% between 1990 and 2013, and for the same time period mortality from stroke has declined by 46% (29). Similar to all-cause CVD, the evidence for subtype specific incidence and prevalence decreasing is mixed. For example, Lampe *et al* estimate that between 1978 and 1996 angina symptoms decreased by 1.8%, but the prevalence of CHD diagnosis remained unchanged (34). One potential contributor to the reduction in cardiovascular mortality is the widespread prescribing of medications for the primary, and often secondary, prevention of cardiovascular disease, such as statins and antihypertensive medications (35-38).

Cardiovascular disease is a complex, multi-factorial disease (39). A number of modifiable behavioural, biological and environmental (including societal) risk factors have been identified for CVD. Behavioural risk factors include among others, alcohol consumption (40, 41), smoking (42) (43-45) and physical inactivity (46-48). Biological risk factors include increased cholesterol levels, in particular low-density lipoprotein-cholesterol (LDL-C) (49),

triglycerides (50), lipoprotein(a) (51) and elevated blood pressure (52, 53). Both individual SEP (such as income and educational attainment) and neighbourhood level SEP are risk factors for CVD (54). More recently environmental exposures such as air pollution and chemical exposure have emerged as risk factors (55). Indeed a number of risk factors will be multifactorial themselves, such as body mass index (BMI) (56-59), which can be increased by among other factors, diet and activity levels, the obesogenic environment (60) or genetics (61). Additionally, non-modifiable factors such as age and sex are risk factors for disease. Although there are distinct cardiovascular subtypes with different clinical pathologies, these typically share many of the same risk factors.

Although CVD events typically occur later in adult life, the aetiology of CVD emerges early in the life course, with precursors of disease or associations between known CVD risk factors and intermediate processes evident from infancy onwards (62-68). Additionally, risk factors for CVD, such as elevated BMI, blood pressure, or adverse lipid profiles are often present from early in life and track throughout the life course (69, 70). Therefore, early interventions to reduce harmful levels of these risk factors are crucial to reduce the burden of disease later in life.

2.2 Socioeconomic position and socioeconomic inequalities

Socioeconomic position is used to describe one of, or a combination of, resource-based measures (such as income and wealth) and prestige-based measures (evaluated by the consumption of good and services as linked to income and education) that influence populations and society (71, 72). In line with the recommendations made by Krieger, I do not refer to this as socioeconomic status or social class, which implies status as determined by societal norms, rather than material resources, such as income and wealth (72). Socioeconomic position can be measured both at the individual level, such as educational attainment, or the population level, such as index of neighbourhood deprivation; where different measures of SEP can have different effects on later life health (73).

Social inequalities have been defined by Krieger to state that these are "health disparities, within and between countries, that are judged to be unfair, unjust, avoidable, and unnecessary (meaning: are neither inevitable nor irremediable) and that systematically burden populations rendered vulnerable by underlying social structures and political, economic, and legal institutions"(72). Importantly, this definition states that these inequalities are modifiable and that they are specifically "unjust". This is an important distinction from health inequalities

which can simply mean any difference in health between groups, without specifically referencing SEP (72, 74).

2.3 Role of socioeconomic position in cardiovascular disease

The concept of socioeconomic inequalities in health is not new, where mortality differences across neighbourhoods were reported as early as the 1820s (75). Although mortality from CVD is decreasing in high income countries, the most socioeconomically deprived individuals remain at the greatest risk (3, 76). The wider determinants of health (including living and working conditions, health care services, housing) are suggested to be the most important drivers of health (77). Indeed, low SEP is one of the strongest indicators of morbidity and mortality (78-80). A number of indicators of SEP have consistently been implicated as risk factors for CVD, or cardiovascular risk factors, in high income countries. These include occupation and employment status (4, 81), education (8, 9, 82-84), income (85-87) and neighbourhood SEP (54, 88, 89). Although inequalities in CVD, and morbidity and mortality more widely, are evident in low- and middle-income countries (90-92), this PhD focuses on the United Kingdom (UK), therefore, this review of the literature will focus on the context of socioeconomic inequalities in high-income countries.

Much of the evidence base identifying these inequalities came from the occupational cohort study, the Whitehall I study of civil servants set up in 1967 (93). A number of key findings came from this study, including identifying the social gradient between occupational social class and CHD (94), and occupational social class grade and all-cause and cause-specific mortality (95, 96). These studies demonstrated that men in the lowest employment grade working as messengers had 3.6 times the mortality from CHD compared with those in the highest employment grade, working as administrators. This was similar considering all-cause mortality and mortality due to other causes.

At a similar time to the Whitehall I study, the Black report identified the worsening of inequalities in health, following the advent of the National Health Service in 1948 (97, 98). This report considered outcomes including mortality, but also wider factors such as access to health services (97). Prior to the Whitehall studies, social factors were often thought of as a potential confounder, usually adjusted for, rather than considered an exposure in their own right (99).

In recent years, the Marmot reviews (Fair Society, Health Lives and Health equity in England) have sought to characterise the extent to which health inequalities exist in England, and what evidence-based strategies exist for reducing these inequalities (100, 101). The first Marmot

review published in 2010 aimed to identify evidence of health inequalities in England and how evidence could be translated into practice to reduce inequalities. This review found that if mortality rates were equal between the least deprived individuals and most disadvantaged individuals, between 1.3 and 2.5 million extra years of life could have been lived. It was proposed that to reduce health inequalities action is required to address all of the social determinants of health, through a method termed proportionate universalism ("with a scale and intensity that is proportionate to the level of disadvantage") so as to benefit the whole of society equally (100). In 2020, the Marmot review was reviewed, with the goal of assessing how population health and health inequalities have changed during the decade. It was reported that rather than any marked improvements in health, health has deteriorated and inequalities widened in England (101). Although life expectancy has slowed down for all groups, those who live in the most deprived areas of the country have seen the greatest reductions. It was reported that the life expectancy for males born in the most deprived areas in England during 2016-2018 was 73.9 years, compared with 83.4 years for males born in the least deprived areas (difference in mortality of 9.5 years). The life expectancy for females born in the most deprived area was 78.6 years compared with 86.3 years in the least deprived area (difference in in mortality of 7.7 years) (101). These mortality differences are widening, where the equivalent difference in mortality in males in 2010-2012 was 9.1 years and the equivalent difference in females was 6.9 years (102).

Public Health England (PHE) posit that much of these differences in life expectancy are due to higher mortality from lung cancer, chronic lower respiratory diseases and CVD in more deprived areas (103). This report estimated that individuals living in the most deprived areas are four-times more likely to die prematurely (below the age of 75) from CVD compared with individuals in the least deprived areas (103, 104).

A number of studies have sought to estimate the contribution of low educational attainment to all-cause mortality and cause-specific mortality. In a 2005 analysis, Huisman and colleagues estimated in an analysis of Western European countries that the absolute rate difference for total mortality between the lowest educated and highest educated participants was 796 deaths per 100 000 person years in males and 442 deaths per 100 000 person years in females. Of these total mortality differences, it was estimated that CVD accounted for 39% of the total mortality difference in males and 60% in females. In England and Wales, the rate difference in total mortality was 1052 deaths per 100 000 person years in males and 435 deaths per 100 000 person years in females (84).

However, the methods by which socioeconomic factors cause disease are not well understood. In the years following The Black report and the Whitehall I study findings, epidemiologists have sought to characterise how these inequalities emerge and persist through the life course (105-107), what explanatory (intermediate) factors might help explain the associations (12-14), and whether the effects of SEP on cardiovascular outcomes are causal (8, 9, 108-110).

2.3.1 Mediators of the association between education and cardiovascular disease

An intermediate variable, or mediator, is one that can either wholly, or partly explain the association between an exposure and an outcome (111). These downstream factors offer an opportunity to intervene after an exposure has occurred. Where an exposure is difficult to intervene on, such as educational attainment, identifying these mediators offers an opportunity to mitigate the impact of the exposure on later outcomes. A number of modifiable risk factors, such as BMI, diet, exercise, smoking and risky drinking have been identified as mediators of low SEP and CVD (12-14, 112-117).

Lower levels of education have been shown to lead to an increase in BMI, using traditional epidemiological analyses and instrumental variable analyses (118-121). Increased education has also been shown to improve diet (122, 123) and increase physical activity (124), likely to contribute to this reduction in BMI. Similarly, increased SEP has been shown to decrease systolic blood pressure (125-128).

Lower levels of education have also been shown to increase both smoking uptake and decrease smoking cessation in those who initiate smoking (129, 130). Smoking is one of the leading causes of CVD (131). Smoking rates are declining in high income countries , including the UK (132); between 2011 and 2018, smoking prevalence has decreased from 20% to 14% (133). However, education inequalities persist between smokers and non-smokers. According to the Office of National Statistics 2018 report on adult smoking habits in the UK, 29.8% of individuals with no formal qualification were smokers, compared with 7.5% of individuals with a degree (133). In turn, BMI, smoking and systolic blood pressure have all been shown to increase the risk of CVD (45, 59, 134-136).

Méjean and colleagues identified that smoking explained 26% of the variation in CHD according to strata of educational attainment, whilst alcohol consumption explained 23% of the variance, physical activity explained 9% of the variation and dietary factors explained 48% of the variation in CHD across strata of educational attainment (117).

In phenotypic mediation analyses, Kershaw and colleagues identified that smoking behaviour explained almost 27% of the effect of education on coronary heart disease, 10% of the effect was explained by BMI and 5% of the effect explained by hypertension (114). However, these estimates may be biased by confounding, reverse causality or measurement error (see 2.7.1). Therefore, in this thesis, I use Mendelian Randomisation (MR) (see 2.7.3) to estimate the causal effect of the mediating role of BMI, systolic blood pressure and smoking in Chapter 4.

Additionally, putative biological mediators, such as low-density lipoprotein cholesterol (LDL-C) (115) and hypertension (14) have been identified as potential mediators, or as downstream effects of education (137). These intermediate variables are already targeted by clinical interventions where statins and antihypertensives are prescribed respectively.

Later life measures of SEP have been implicated as mediators of early life SEP and CVD risk, for example occupation, housing, financial stress are downstream of early life SEP and themselves, independent risk factors for CVD (112, 116, 138). In mediation analyses, Hossin and colleagues found up to 39% of the effect of childhood SEP on CVD mortality could be explained by own occupation; this increased to 59% of the effect when behavioural intermediates (such as BMI and smoking) were included. Mental health and emotional states have also been implicated as mediators (113).

2.3.2 Association between socioeconomic position and preventative medication for cardiovascular disease

Biological intermediate risk factors, such as elevated cholesterol levels and elevated blood pressure (hypertension) are already targets of cardiovascular preventative medication, where statins and antihypertensives are prescribed respectively. Inequalities in access to, or prescribing of, these preventative medications may also contribute to inequalities in CVD. Many of the factors described in section 2.3.1 are individual level factors which may explain inequalities in CVD. However, access to medication begins to allude to a wider societal contributor to inequalities in CVD.

Statins are a group of cholesterol-lowering drugs, widely prescribed for both primary (before an adverse event) or secondary (following an adverse event) prevention of CVD (139). They are one of the most commonly prescribed drugs in the UK (140). Current guidance in England states that individuals should be prescribed statins if they have a 10% of greater risk of experiencing an adverse cardiovascular event in 10 years (24-26). Typically, this is examined using a QRISK score in general practice (currently the QRISK3 version). This score incorporates a number of cardiovascular risk factors, including (among others) age, sex,

ethnicity, systolic blood pressure, area level deprivation, BMI, smoking and family history of CVD (23). In the case of some adverse cardiac events, statins will be prescribed as secondary prevention, without estimating 10-year risk of disease.

To date, there is limited evidence for the role of medication prescribing in contributing to inequalities. Although, it was described as early as The Black Report (1980), that access to healthcare was not equitable (98).

Considering access to statins, the literature is mixed in the direction to which inequalities in exist. Some studies suggest that there are no socioeconomic differences in prescribing (141), others suggest that those with lower socioeconomic position are more likely to be prescribed statins (36, 142-144), whereas some studies suggest individuals of lower socioeconomic position are less likely to be prescribed statins (87, 145-147). These inequalities have predominantly been explored in the secondary prevention setting, where they are prescribed following an adverse cardiac event to prevent further events. This is in contrast to the primary prevention setting, where prescribing aims to prevent an adverse cardiac event happening initially.

One potential reason for this mixed evidence is that individuals of a lower SEP are more likely to have a greater underlying clinical need for medication. For example, as discussed in section2.3.1, lower SEP leads to, among other factors, higher BMI, increased smoking prevalence and elevated cholesterol levels. These factors are all considered in clinical decision making, for example, these three factors all contribute to the QRISK3 model of CVD risk used to inform statin prescribing (23).

In Chapter 5 of this thesis, I explore the potential role of access to statins, as a primary prevention mechanism, contributing to socioeconomic inequalities of CVD, after controlling for measures of clinical need.

2.4 Defining socioeconomic position in this thesis

In this thesis, I focus on educational attainment to measure SEP for a number of reasons. Firstly, educational attainment is largely determined early in life. When considering mediators of SEP and CVD, it is important to consider the temporal relationship between all of the exposures, the mediators and the outcomes (111). Therefore, using an early measure of SEP, such as education, compared with a later life measure such as occupation or income, means that temporality between exposures and outcomes can be better accounted for. For example, a cardiovascular event in adulthood is unlikely to affect early life education, however it could

affect income in adulthood. Additionally, education is a strong predictor, and highly correlated with, future employment and income, both later life measures of SEP (71). Therefore, considering education as an exposure, in an adult population, is also likely to be capturing part of the effect of later life SEP. Secondly, this thesis uses MR and the integration of genetic data with epidemiological analyses. In recent years, a number of genome wide association studies (GWAS) have been carried out for education, and up to 11% of the variance in education can now be detected via these genetic variants (17, 148, 149). Thirdly, education can be easily and widely measured. Typically, an individual knows when they left school, either by being asked the age in which they left or what their highest qualification is. Education is also relatively stable during an individual's life course. Conversely, a measure such as occupation or income for example, often changes across a life course and can be difficult to measure when participants may have retired prior to study entry (71, 73).

As UK Biobank data are used in this thesis, and participants were aged 40-69 at recruitment, SEP measures such as income, employment or occupation could introduce bias to analyses (16). Given their age, it is common in UK Biobank for participants to have retired from work. Although both current and historical employment is recorded in UK Biobank, it is not straight forward how to determine their SEP from employment in this context. Similarly, participants were asked to report current average annual, monthly and weekly household income. This measure is unlikely to truly capture SEP in individuals who have retired and does not account for individual life-time income.

2.5 Intermediate variables considered in this thesis

In this thesis I focus on four main intermediate pathways between education and CVD; BMI, systolic blood pressure, smoking and medication prescribing (see section 2.3.2). These are all modifiable major public health targets, either through lifestyle interventions (150), clinical interventions (151) or, major public health campaigns (152). The total years of education completed are largely determined in early life, prior to adulthood. Although CVD is largely considered a disease of ageing (153), socioeconomic patterns of BMI, systolic blood pressure and smoking all emerge during early life and adolescence (129, 154). The incubation period between major cardiovascular risk factors and CVD is long (155), providing ample opportunity to intervene following the establishment of educational attainment and prior to developing disease.

There is likely to be significant overlap in how these mediators work together to increase the risk of CVD. For example, there is evidence of bi-directional associations between BMI and

smoking (156). The INTERSALT study also found that BMI and smoking behaviours were mediators of the association between education and systolic blood pressure (127). Additionally, many other measures of SEP and potential mediators are likely to overlap with the variables considered in this thesis, and therefore will be captured to some extent by these factors. For example, a higher income may result in a more cardioprotective diet (i.e. more fruits, vegetables and wholegrains, and less processed foods) (157) and thus lower BMI and SBP. By looking at the modifiable intermediate factors chosen here, in a causal inference framework, I hope to identify intermediates that could be used as interventional targets to reduce the burden of CVD attributable to educational inequalities.

2.6 Genetic determinants of cardiovascular disease

Cardiovascular disease is a complex multifactorial condition, encompassing a wide range of conditions, where environmental and genetic factors both contribute to the aetiology of disease (158). As previously outlined, a number of behavioural, lifestyle, societal and environmental risk factors exist for CVD. However, there is also a strong genetic component of disease. It has been estimated that the heritability of CHD lies between 40% to 60% (159), whilst for atrial fibrillation heritability is estimated at 22% (160) and heritability of ischaemic stroke lies between 34% to 42% (161). Despite the presence of distinct cardiovascular subtypes, much like shared environmental risk factors for disease subtypes (162-164), there is evidence of shared genetic contributions for subtypes of CVD (165, 166).

2.6.1 Genetic Epidemiology

Genetic epidemiology considers the contribution of genetics in disease aetiology, including understanding heritable aspects of disease and individual susceptibility (18). One of the primary aims of genetic epidemiology is to identify, isolate and understand, the genetic component of disease risk from complex, multifactorial disease states (167). The advent of genetic epidemiology was in part driven by the human genome project and wide scale genome sequencing and genotyping. More recently, the advent of large-scale human biobanks (often as part of cohort studies) have enabled genetics to be widely incorporated to epidemiology (15, 168, 169). Genetics can be used in epidemiology to answer a host of research questions, including understanding disease aetiology (how much a genetic variant can explain disease risk, and the non-genetic risk factors that affect risk of disease) and in health services research (such as the impact of using genetic tests in health services) (18, 170).

Genetic epidemiology methods provide opportunities for improving causal inference in aetiological epidemiology. These techniques can provide insights into biological mechanisms

for disease pathogenesis (171) and help prioritize targets for intervention (172). One common method in genetic epidemiology is that of Mendelian randomisation; the use of genetic variants as an instrumental variable for a phenotype (see section 2.7.3) (18).

2.6.2 Polygenic prediction of disease

Since the rapid increase in genetic data, and the explosion of GWAS summary statistics, polygenic scores (PGS) (or genetic risk scores, polygenic risk scores or weighted allele scores), have been increasingly used in epidemiology to understand how genetics contribute to the aetiology of polygenic (i.e. multiple genetic causes) phenotypes (173, 174). A PGS incorporates information from across the genome to understand the genetic component of a phenotype, where there may not be one single gene responsible for the acquisition of the trait. For example, height or BMI are examples of polygenic traits, where a number of genetic variants contribute to the trait (61), as well as interactions with the environment (175-177). Conversely, a monogenetic trait is determined by a mutation in a single genetic variant (or few genes), such as cystic fibrosis, which is caused by a mutation in the *CTFR* gene (178). Indeed, most common diseases are polygenic (170).

Typically, in polygenic traits, each genetic variant explains very little of the phenotype, but the cumulative risk across many genetic variants can begin to explain a substantial fraction of the variation in a phenotype (173, 179). As the samples size of GWAS increase, the power and predictive accuracy of PGS have been improving (180). This leads to an increase in the number of genetic variants identified to be associated with the phenotype, and an increase in the total variance in the phenotype explained by known genetic variants. For example, in the 2015 Locke *et al* GWAS of BMI, the 97 independent genetic variants identified for BMI at the genome-wide significance level explained about 2.7% of the variation in BMI (181). The updated 2018 Yengo *et al* GWAS of BMI, identified 941 near-independent genetic variants related to BMI at the genome-wide significance level explaining around 6% of the variation (61). The latter GWAS had a sample size of around 700 000 individuals, compared with almost 340,000 in the 2015 GWAS.

A PGS is usually derived by weighting the sum of the genetic variants with their relative effect sizes (182, 183). This upweights the genetic variants with the greatest effect on the phenotype, improving the explanatory power of the score, although unweighted scores can also be generated. When a PGS is derived for the purpose of disease prediction, genetic variants included may be broad, including variants in linkage disequilibrium (i.e. highly correlated with other genetic variants), or with small amounts of explanatory power. Where the PGS is to

be used as an instrument in MR, for example, genetic variants may be limited to those that, are not in linkage disequilibrium to other variants, or those that meet a stringent GWAS significance threshold of $P<5\times10^{-8}$.

Importantly, although not their sole use, PGS are used in disease prediction modelling; meaning that across the population a PGS can be used to estimate the probability of developing a phenotype, but they are not deterministic (184). Behaviour modification, or treatments, can therefore be used to modify the risk of disease. A strong association between a PGS and a trait does not necessarily mean there is a causal effect of the PGS on the trait; associations could be induced due to population structure or assortative mating, or dynastic effects (185, 186). The causal aetiology of the PGS and subsequently of the trait, can be interrogated via different study designs.

2.6.3 Applications of polygenic scores in disease prediction

Genetic information is already used in clinical practice to aid decision making. For example, in the cases of familial breast cancer, genetic testing can be offered to identify whether high-risk genetic variants, *BRCA1*, *BRCA2* or *TP53* are present (187). The identification of these genes will then feed into clinical decision making. Similarly, in the case of familial hypercholesterolaemia genetic testing may be offered to identify the presence of disease causing mutations (188). As PGS become more powerful, explaining a greater amount of the variation in a trait, there is a growing body of research that could potentially be incorporated into clinical practice.

Khera and colleagues generated PGS for a host of disease outcomes and compared the predictive ability of them with known monogenic disease-causing mutations (170). It is estimated that the mutation for familial hypercholesterolaemia causes a three-fold increase in the risk of CHD (189). In this sample, the authors identified that using the PGS for CHD 8% of the population were at the equivalent 3-fold risk of disease. Comparatively, the mutation for familial hypercholesterolaemia is found in approximately 0.4% of the population (189). Khera *et al* concluded that "it is time to contemplate the inclusion of polygenic risk prediction in clinical care". In addition to CHD, these conclusions could be made for atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer.

Similarly, Inouye and colleagues concluded that a PGS for CHD had a greater predictive power for incident CHD than individual conventional CVD risk factors, smoking, diabetes, hypertension, BMI, cholesterol or family CVD history (190). This finding was replicated for ischaemic stroke (191). Additionally, clinical trials have demonstrated that providing

information on genetic risk to participants led to greater reductions in low density lipoprotein cholesterol than information based on conventional risk factors alone (192).

As genetic testing becomes more widespread and financially feasible, more consideration is being given to how they could be used in clinical practice (193). However, this isn't without ethical considerations; especially in societies with no national health service (194). However, a number of recent studies have compared the additional predictive power of polygenic risk over and above phenotypic risk. Typically, these studies have found little improvement in predictive power when including genetic risk (195-197).

2.6.4 Gene*environment interactions in cardiovascular disease

As discussed throughout this chapter, CVD has many risk factors, both environmental and genetic. It has been widely noted that consideration needs to be given as to how genetics and the environment work together to contribute to disease risk, rather than considering either in isolation (198, 199). Whilst socioeconomic inequalities in health have been widely studied, the extent to which the interplay with genetic factors contributes to these inequalities been sparsely studied (200). An interaction analysis assesses the joint effect of two risk factors (the environment and genetics), where a joint effect greater than the sum of the individual effects indicates positive interaction (201).

Understanding gene*environment interactions are important for a number of reasons, including estimating the population attributable risk for genetic and environmental risk factors and their shared effects, helping to identify biologically plausible mechanisms for disease aetiology, identify potential therapeutic targets and tailor preventative treatments to those most at risk or identify who would benefit the most from treatment (198, 202).

Often, these interactions have typically been assessed in a candidate gene approach. For example, Hamrefors and colleagues carried out analyses to identify whether the rs4977574 allele, an allele in the chromosome 9p21 region which is suggested to increase susceptibility to CHD, interacted with smoking, education and physical activity to increase cardiovascular risk (203). Here, it was found smoking, but not education or physical activity, interact with the risk allele to increase cardiovascular risk. However, these approaches have been criticised for failing to replicate, likely due to statistical power, publication bias and low prior probabilities of hypotheses being true (204).

Therefore, more recently, gene*environment interactions have been assessed using polygenic gene prediction. Additionally, for traits such as CHD which are polygenic in nature, it may be

more informative to consider the whole of the polygenic risk susceptibility, rather than focussing on interactions with single genetic variants each with small effects on disease risk.

In Chapter 6 I explore whether polygenic susceptibility to cardiovascular risk is modified by strata of educational attainment.

2.7 Causal inference in epidemiology

Epidemiology is defined as "the study of the occurrence and distribution of health-related events, states and processes in specified populations and the application of this knowledge to control health problems" (205). Being able to make causal inference is central to epidemiology; practitioners, policy makers, clinicians and scientists want to know whether intervening on a risk factor will lead to a reduction in disease, or other outcomes. Causality has been a key aim of decades of research following Bradford and Hill's research on the causal criteria in the 1960s (206). This work suggested for a risk factor to be causal the strength of the evidence should be evaluated by 9 factors. These are i) strength of the association ii) consistency of the association iii) specificity of the association iv) temporality of the relationship v) biological gradient vi) biologically plausible vii) coherence of the effect interpretation viii) experimental evidence to support the observed effects and ix) useful analogy. Since this seminal work, the field of causal inference has worked to improve methods and knowledge to enable the study to make claims of causality (207, 208).

In this thesis, I use a number of causal inference approaches to understand educational inequalities in cardiovascular disease. In Chapter 3 I demonstrate how mediation analyses (section 2.7.2) can be carried out using MR (section 2.7.3). In Chapter 4 I use MR methods for mediation analysis to identify the role of BMI, systolic blood pressure and smoking in mediating the association between education and CVD, triangulating across data sources and methods. In Chapter 5 I triangulate across different data sources with different sources of bias to identify educational inequalities in statin treatment to prevent cardiovascular disease. In Chapter 6 I use different PGS' to identify whether there is evidence of education as an effect modifier of polygenic susceptibility to CVD.

2.7.1 Sources of bias in epidemiological research

Epidemiology aims to determine the causal effect of an exposure (cause) on an outcome, however, spurious associations can be observed for many reasons. These spurious associations are often inherent in the data, although they can be induced through analytical or design choices.

One of the main sources of bias comes from confounding, where both the exposure and outcome share a common cause, which can wholly, or more likely partly, explain some of the association between the exposure and confounder. Where all confounders are perfectly measured, and adjusted for, effect estimates will be unbiased due to confounding. Although confounding can be addressed in epidemiological analyses , such as multivariable adjusted or propensity score designs, for most study designs it is vital that information on a sufficient set of confounders are available (209). In a causal diagram approach, confounders are said to be sufficiently controlled for when all backdoor paths are blocked between the exposure and outcome and no spurious associations can exist (210). When multiple confounders are correlated, not all confounders are required to be adjusted for in the minimally sufficient model to block all backdoor pathways (see Figure 2.1) (211). If a minimally sufficient set of confounders is not available, or confounders are measured with error, residual confounding will bias estimates. Confounding can be addressed during the study design stage, for example by restriction or matching. Here, the study population are selected on key variables which may be confounders, for example selecting participants all of a similar age range or of the same sex (209, 210). Alternatively, analyses can be carried out stratified by certain key confounders, such as carrying out sex-stratified analyses (210, 212). However, where data are sparse, this can lead to inefficient analyses, or introduce additional biases (213).

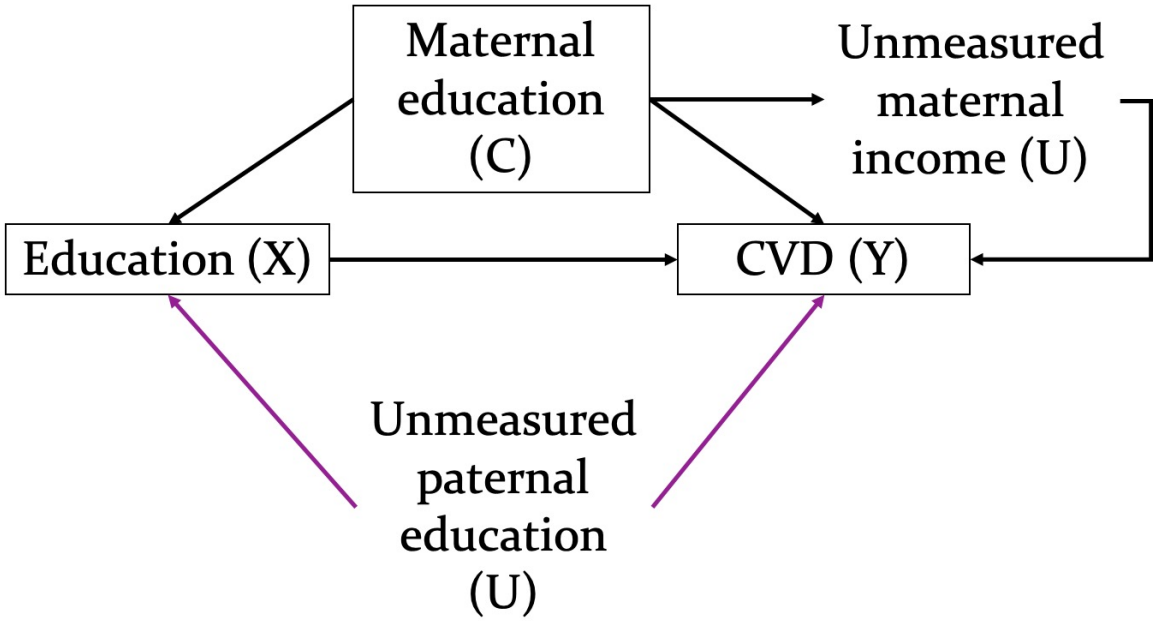

Figure 2.1: : Directed acyclic graph demonstrating the hypothetical association between education (X) and cardiovascular disease (CVD) (Y) controlling for the measured confounder (C), maternal education. Maternal income and paternal education are unmeasured confounders (U). Unmeasured paternal education biases the effect estimate (demonstrated in purple). Unmeasured maternal income is sufficiently controlled by maternal education and therefore does not lead to bias.

Estimates from multivariable adjusted regression can be biased by information bias or measurement error; either because information is not reported accurately, study equipment may not be calibrated accurately, the information being collected doesn't actually reflect the true causal variable of interest, or due to random chance (214). These biases can occur in an exposure, outcome, confounder or mediator (where included) (see Figure 2.2). Information bias and measurement error are often used similarly to describe errors in data. Measurement error occurs when the quality of measurement is poor (215). Measurement error can either be systematic, e.g. a mis-calibrated blood pressure monitor adds 5 mmHg to all readings, or random, e.g. a mis-calibrated blood pressure monitor can add or subtract up to 5 mmHg from some readings. Recall bias is an example of a systematic bias, when participants do not accurately remember past experiences, or omit specific details (205). Information bias, or misclassification bias, occurs when information is measured or recorded inaccurately (215). Information bias can either be non-differential or differential. Non-differential bias does not relate to the outcome, meaning the chance of being misclassified, is equal across all study groups and outcomes (205). For example, in a study exploring the association between hypertension and incident CHD, all participants, regardless of outcome could be incorrectly classified as being hypertensive. Conversely, where misclassification is differential, the bias varies according to the outcome of interest (216, 217). In the same hypothetical example, this may mean only those diagnosed with incident CHD can incorrectly be classified as hypertensive. Where error is non-differential, or random, in the exposure, effect estimates are typically attenuated towards the null, also known as regression dilution bias (217, 218). If error is non-differential or systematic in the exposure or outcome, bias can be present in either direction (e.g. over- or underestimate the true association). Where a confounder is measured with error, this will result in residual confounding; it is not possible to predict which direction the effect estimate would be biased by in this case (219).

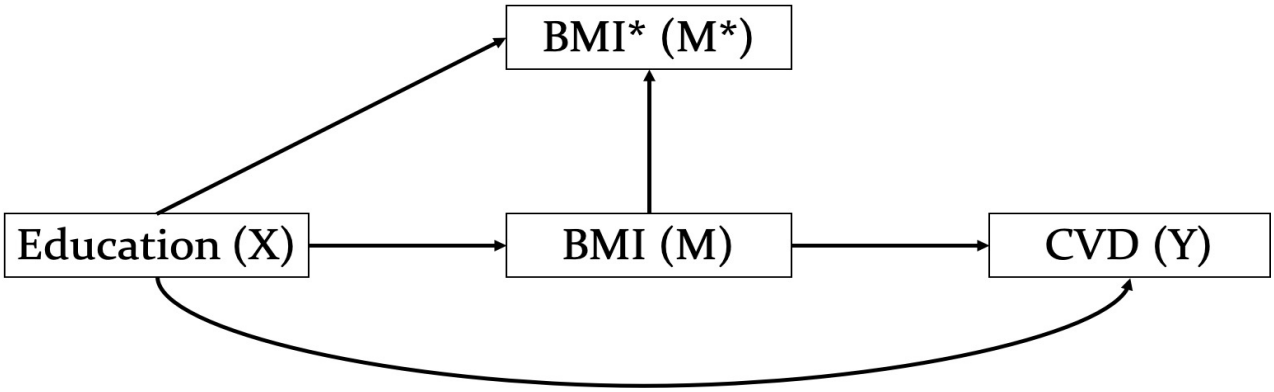

Figure 2.2: Directed acyclic graph demonstrating measurement error, where the effect of education (X) on cardiovascular disease (CVD) (Y) is mediated by the true value of body mass index (BMI). Where BMI is measured (observed) with error (BMI) this direct effect association is no longer observed.*

Reverse causality can introduce bias, when the temporality of the exposure and outcome is mis-specified and the outcome itself affects the exposure (see Figure 2.3) (220). This type of bias can frequently occur in case-control studies which often collect data on the outcome prior to the exposure, or cross-sectional studies where the exposure and outcome are measured at the same time. However, in an incorrectly specified model, this can occur in other study types. One method to minimise this bias is to collect data prospectively or maintain temporality between an exposure and outcome e.g. recall of early life exposure and later life outcome measured at the same time (cross-sectional).

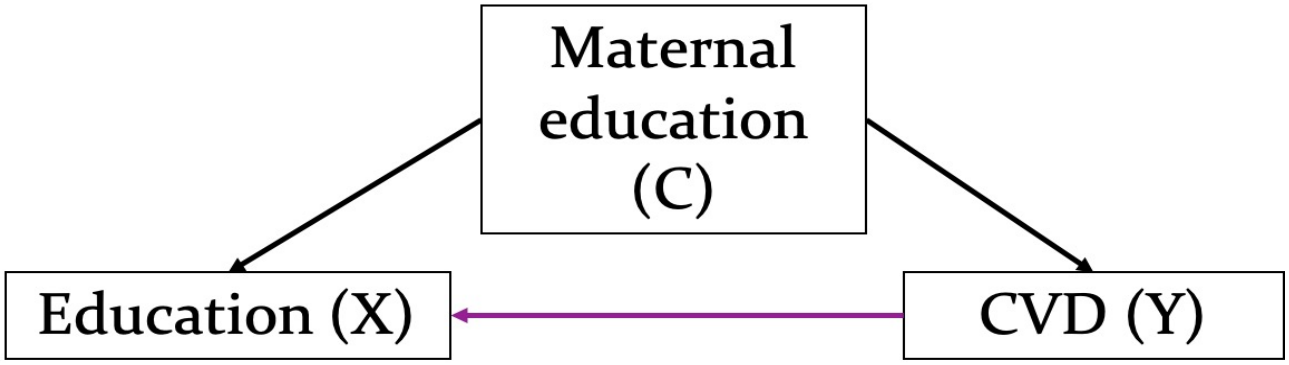

Figure 2.3: Directed acyclic graph depicting reverse causality, where the model is mis-specified and the outcome (Y), cardiovascular disease (CVD) causes the exposure (X), education

Further bias can be introduced by selection into the study, where individuals in the study population are not reflective of the wider population of interest. If selection involves conditioning on a factor that is a cause or effect of both the exposure and the outcome collider bias can be introduced (221). This bias is demonstrated in Figure 2.4. In this hypothetical

example, to be recruited into the study, a participant must be alive at recruitment, where all participants eligible have to be at least 60 years old. Both the hypothetical exposure, Lower education, and the hypothetical outcome, CVD are associated with higher mortality at younger ages. Therefore, selection into the study is conditional on being alive at age 60, which could induce a spurious association between education and CVD.

Figure 2.4: Directed acyclic graph depicting selection bias. In this hypothetical example, study entry is conditional on being alive at aged 60 or above, which is caused by both education (X), the exposure, and cardiovascular disease (CVD (Y)), the outcome. Solid boxes around a variable demonstrate conditioning on the variable, dashed lines indicate an induced spurious association

Well conducted randomised controlled trials (RCTs) provide the strongest evidence of causality, and often described as the 'gold standard' (222). Here, participants are randomised to either the exposure of interest, or a control condition and followed up to determine their outcomes. This study design examines the cause and effect relationship between two factors, controlling for temporality and allocating actual exposure (223). However, RCTs are very costly, not always ethical or practical, or timely. Particularly in the case of social epidemiology, it would be highly impractical, potentially unethical, to randomise access to education, and take a lifetime to discover the effects on health outcomes later in life. Additionally, the generalisability of RCTs is often criticised. Often the external validity of a trial (i.e. how well the trial results relate to the population of interest for the intervention) is given less consideration than the internal validity of the RCT (designing and carrying out the study with minial opportunites for bias) (224).

2.7.2 Mediation analysis

Mediation analysis has been widely used by a number of disciplines to identify intermediate variables between an exposure and an outcome, including behavioural or biological variables, which provide opportunities for interventions, or gain a wider understanding of how an exposure may be causing a disease outcome.

Sewall Wright first proposed early methods for path tracing, an early form of mediation analysis (225), which were extended to allow for decomposition of total effects in to direct and

indirect effects (226). These early methods provided the first statistical decompositions of mediated effects. Baron and Kenny formalised mediation analysis in the 1980s by proposing four steps that were required to establish mediation in a hypothesised model (227) (Figure 2.5).

1. Regress the dependent variable (Y) on the independent variable (X) and show that X is associated with Y to establish a total effect which can be mediated (path c).
2. Regress the mediator (M) on the independent variable and show that X affects M (path a).
3. Regress the dependent variable (Y) on the mediator (M) controlling for X to show that the mediator affects the dependent variable, independent of X (path b).
4. Establish that M completely mediates the X-Y, where there is no effect of X on Y once M is controlled for (path c'). If this condition is not met but the regression coefficient is smaller in step 3 than step 1, there is evidence for partial mediation.

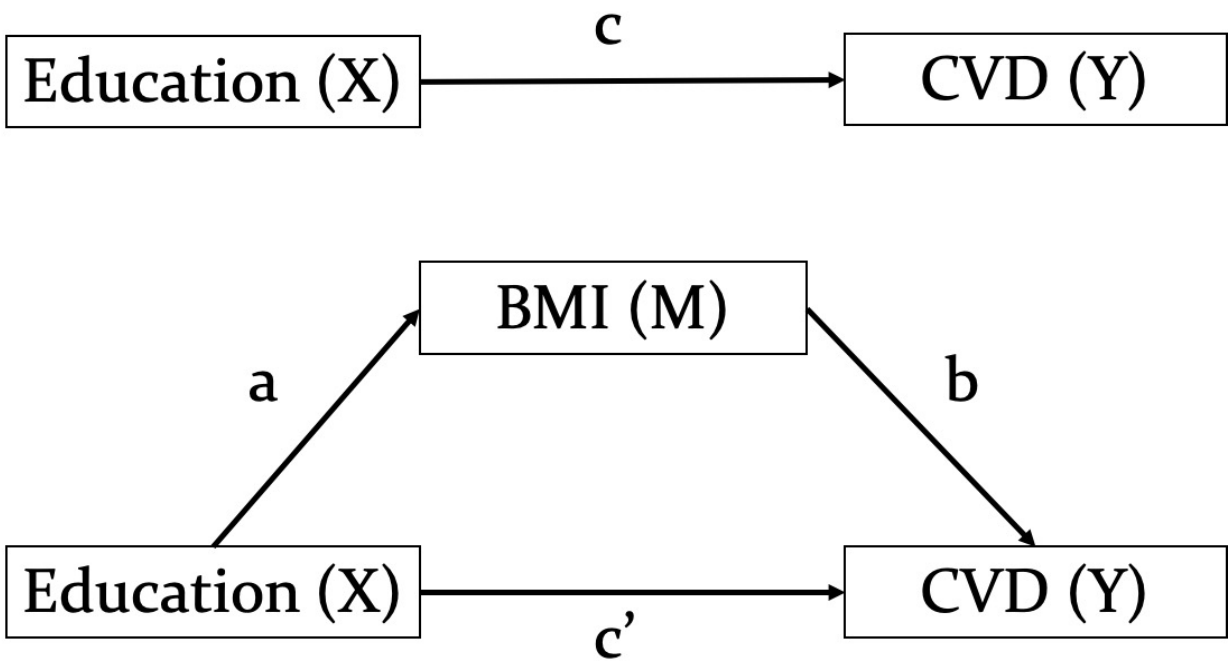

*Figure 2.5: Schematic of total and mediated effects. Path c represents the total effect of the exposure on the outcome. Path c' is the direct effect; that is the effect of X on Y not mediated by M. The indirect effect can be estimated by i) a*b known as the product of coefficients or ii) c-c' known as the difference method.*

A number of assumptions are required for these estimators to be unbiased. Firstly, there should be no unmeasured confounding between the exposure, the mediator and outcome (228-230). A second, strong assumption required, is that of no exposure-mediator interaction. A third assumption is that confounders of the mediator and outcome are not themselves

caused by the exposure, which if present can induce collider bias (231-233). Causal inference methods, including G-computation (234) and inverse probability weighting (235) can be used to estimate the unbiased direct effect, whilst also controlling for the confounder.

In order to relax some of these assumptions, a focus of methods development in mediation analysis is to identify methods that can account for some of these strong assumptions and clearly specify how causal interpretations are made (207, 228, 236). Robins and Greenland proposed that counterfactual conditional statements were fundamental to causal models. In this situation, each individual is observed under a given circumstance, but analyses consider what would happen to the same individual if they were observed under the counterfactual circumstance, that is the effect that is counter-to-fact (did not occur) (237). Counterfactual (counter to fact) theory prescribes that an event (effect) has only occurred because of a prior fact (cause); for example, "If I had not gone to sleep, I would not have woken up". In order for one to wake up, one must first go to sleep. Whereby, for the effect to happen, there must have been a cause (and in turn an intermediate cause, the mediator cause). Similarly, if that cause was not present, the effect would also not be present (207).

The parameters estimated in counterfactual mediation methods differ somewhat from traditional methods, where these assumptions are explicitly stated. These parameters are, the total causal effect (either as the individual causal effect or the population causal effect) the controlled direct effect (CDE), the natural direct effect (NDE) and the natural indirect effect (NIE) (208). Where interactions exist between an exposure and mediator the natural indirect effect reflects the portion of the total effect attributable to mediation only, as opposed to interaction only or mediation and interaction combined (mediated-interaction) (238-240).

However, counterfactual methods for mediation analysis still require strong, unverifiable, assumptions to be made around unmeasured confounding; these are that there are i) no unmeasured confounders of the exposure and outcome ii) no unmeasured confounders of the mediator and outcome and iii) no measured, or unmeasured, confounders of the mediator and outcome that are themselves affected by the exposure (intermediate confounders) (233, 241). In addition to these confounding assumptions, all mediation models also assume appropriate temporal ordering of the exposure, mediator and outcome and can be biased by measurement error in either the exposure or the mediator (238, 241).

In this thesis (Chapter 3 and Chapter 4), these methods are described as phenotypic mediation methods, to distinguish from genetic MR mediation methods.

2.7.3 Mendelian randomisation

Mendelian randomisation is an application of instrumental variable (IV) analysis, which uses genetic variants as an instrument for an exposure of interest and often described as natures' RCT (18, 242). MR relies on the principle that offspring randomly inherit their DNA from their parents during meiosis and at conception and that germline DNA is not modified by lifestyle factors later in life. For this reason, effect estimates from an MR analysis can be more robust to confounding and are not affected by reverse causality, two of the major pitfalls of traditional phenotypic epidemiology. MR can therefore offer a more robust form of causal inference in analyses (18).

There are three core assumptions that need to be satisfied for MR results to be valid (Figure 2.6):

1. The instrument (genotype) is associated with the exposure of interest (relevance assumption)
2. There are no common cause of the instrument and the outcome (the independence assumption)
3. The instrument only affects the outcome via the exposure of interest – i.e. no pleiotropic pathways (the exclusion restriction criteria)

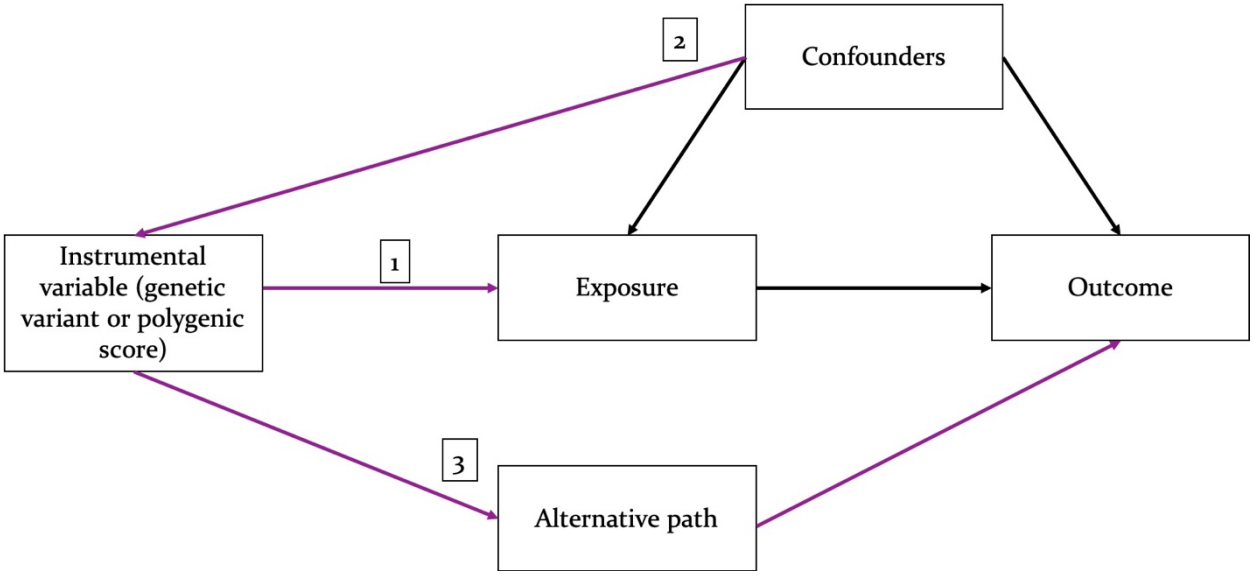

Figure 2.6:Schematic of Mendelian Randomisation and the assumptions that must be satisfied for the results to be valid Violations of the Mendelian randomisation assumptions are demonstrated in purple

Two types of data can be used for an MR analysis. This first is individual level data where information is available for the instrument (genotype), exposure and outcome for every individual the analysis is to be carried out on (243). This is also known as one-sample, single

sample or individual level MR. This can involve either individual single nucleotide polymorphisms (SNPs) or a PGS for each individual where the SNPs which are included in the score are identified from a GWAS of the exposure of interest.

The second approach is known as two-sample MR or summary MR (244-247). In this approach, the estimates of the instrument to exposure association and the instrument to outcome association come from two separate GWAS. The Wald ratio can then be estimated by dividing the summary statistics for the instrument-outcome association by the instrument-exposure association summary statistics. This method does not require access to individual level data.

Methods for MR have rapidly expanded in recent years (248) including the development of methods to extend the applications of MR to understand more complex causal mechanisms. For example, factorial MR can be used to investigate the joint effects of two risk factors on a single outcome (249-252).

Recently developed MR methods can be used in mediation analyses. These are two-step, or network, MR and multivariable MR (MVMR) (19, 20, 253). These methods are discussed in detail in Chapter 3 and Chapter 4.

Sources of bias in Mendelian randomisation

Bias can arise when any one of the three instrumental variable assumptions are violated. Of particular concern, and indeed one of the most common sources of bias in MR, is that of assumption (3), where the genetic variant is associated with the outcome via an alternative path to the exposure of interest. Where a genetic variant influences more than one trait this is termed 'pleiotropy' (18). Pleiotropy can take two forms; vertical pleiotropy and horizontal pleiotropy. Vertical pleiotropy occurs when the genetic variant influences a trait downstream of the exposure i.e. a mediator of the exposure of interest, but any effect of the genetic variant on this mediator is via the exposure of interest. This does not violate the instrumental variable assumption (3) (248, 254). Conversely, horizontal pleiotropy occurs when the effect of the genetic variant independently affects a second phenotype, here the instrumental variable assumption (3) is violated (248, 254). A number of methods have been developed in recent years to test for, and minimise bias due to horizontal pleiotropy, such as MR-Egger (255) and the weighted median estimator (256).

Due to the often limited explanatory power of the genetic instrumental variables used in MR, sample size for MR are typically required to be much larger than conventional phenotypic epidemiology methods to achieve appropriate statistical power (257, 258).

2.7.4 Triangulation

Triangulation has long been described and used in many research fields, but the use of triangulation in epidemiology is emerging. Triangulation aims to obtain more reliable research answers by integrating results from several different approaches, such as the methods or data used, which have different key sources of bias (259). This is distinct from replication or validation analyses, which aim to use the same method or data to compare results from the same study design. Where these different approaches are carried out, and provide consistent effect estimates or directions of effect, the evidence of causality can be strengthened. Importantly for this causal inference, the key sources of bias should be explicitly acknowledged, as well as the expected direction of effect that these biases would result in (260).

In this thesis, triangulation is carried out, where possible, by comparing results from different methods (Chapter 3 and Chapter 4), such as phenotypic and MR methods for mediation analysis. In Chapter 4 results from individual level and summary MR are compared, and in Chapter 5 analyses carried out using data from UK Biobank baseline assessment centres to derive QRISK3 cardiovascular risk scores are compared with analyses using QRISK3 cardiovascular risk scores recorded in primary care data.

2.8 Applying genetic epidemiology to social epidemiology

This thesis integrates genetic epidemiology methods, such as MR and PGS analysis to understand the social causes and consequences of CVD. In recent years, social epidemiology has expanded to consider the biological interplay with social exposures, acknowledging that omitting either biology or sociology would likely lead to incomplete conclusions (261-263). In some instances, the biological state is the outcome. For example in a study by Fraga *et al,* the authors look to explore the effect of socioeconomic position on the inflammatory markers C-reactive protein, Interleukin-6 and tumour necrosis factor-α (264).

In other examples, the biological measure may be used in the context of an exposure, such as the MR analyses by Tyrell and colleagues demonstrating potential causal effects of genetically instrumented BMI and height on socioeconomic outcomes, including education and income (265). As studies with genetic data become more plentiful and can be used in in GWAS to

identify genetic contributions to phenotypes, the opportunities for identifying genetic variants for social exposures increase (149, 266).

2.8.1 Genome wide association study of educational attainment

The first GWAS for educational attainment was carried out in 2013 and included 101 069 individuals. Three independent genetic variants were identified to associate with education (148). The effect size of these three variants was small, equating to approximately one additional month of schooling per risk allele or about 2% of the variance in educational attainment. There have since been a further two GWAS of educational attainment since, each increasing in sample size (N = 293 723 and 1.1 million respectively), number of genetic variants identified (74 SNPs and 1271 SNPs respectively) and variance explained (3.2% and 11% respectively) (17, 149). Importantly for this thesis, these GWAS allow for opportunities to triangulate results with different analytical methods and types of data to infer causality (259). A number of epidemiological analyses have now been carried out using the results from these GWAS, including investigating the causal effects of educational attainment on CVD and cardiovascular risk factors.

Davies and colleagues compared instrumental variables estimates from the Raising of School Leaving age (RoSLA), a natural experiment, and MR estimates from the Okbay educational attainment GWAS (17). Both methods rely on the same instrumental variable assumptions as discussed in section 2.7.3, but the different data sources may be biased by different mechanisms through violations of these assumptions. Here, it was demonstrated that the effect of a 1-year increase in educational attainment, instrumented either via the RoSLA or the polygenic score, associated similarly with adverse effects on health (267). This work demonstrates the validity of the PGS in an MR approach, compared with a widely accepted natural experiment instrument. Although it should be acknowledged that the MR estimates may be biased by family level confounding (dynastic effects), which would not bias RoSLA estimates.

Tillmann and colleagues demonstrated how genetic variants for educational attainment could be used to instrument the effect of educational attainment on cardiovascular outcomes. Here, it was found that each 3.6 years of genetically instrumented higher educational attainment was associated with a 33% reduction in CHD (9). Additionally, it was demonstrated that this may be partially mediated by health-related behaviours such as smoking and BMI. Further MR analyses have demonstrated potentially causal effects of higher educational attainment on

lower BMI (118), lower rates of smoking initiation, heaviness and cessation (268) and reduced binge drinking but increased alcohol intake frequency (269).

Given that education and intelligence are highly correlated, it is difficult to know whether genetically predicted educational attainment is capturing phenotypic education, phenotypic intelligence or both (270). Using MVMR the causal effect of educational attainment independent of intelligence, on a number of outcomes have been demonstrated, including independent effects on smoking (271, 272), BMI (272), sedentary behaviour (272) and CHD (273).

2.9 Chapter summary

In this chapter I have explored the historical context of social causes of CVD, and how SEP can be considered and defined in epidemiology. I then considered causal inference in Epidemiology, including the sources of bias hindering causality and methods aimed at improving causality which are used throughout this thesis. As part of this, I introduced existing methods typically used in mediation analyses, genetic epidemiology, MR methods and the concept of triangulation. I then went on to explore potential mediators downstream of educational attainment, an early life measure of SEP, in the aetiology of CVD. I considered how phenotypic and genetic risk scores can be used in disease prediction. Finally, I considered the role that genetic epidemiology can play in interrogating social epidemiological research questions.

Chapter 3. Mendelian randomisation for mediation analysis: current methods and challenges for implementation

3.1 Author list and contributions

Alice R Carter [1,2], Eleanor Sanderson [1,2], Gemma Hammerton [1,2,3], Rebecca C Richmond [1,2], George Davey Smith [1,2,4], Jon Heron [1,2,3], Amy E Taylor [1,2,4], Neil M Davies [1,2,5], Laura D Howe [1,2]

All affiliations are presented in Appendix 1

ARC devised the project, analysed and cleaned the data, interpreted results, wrote and revised the manuscript. ES devised the project, generated and analysed simulated data, interpreted results and critically revised the manuscript. GH, RCR, GDS, KT, JH AET, NMD and LDH devised the project, interpreted the results, and critically revised the manuscript. All authors had full access to the data in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis. ARC and LDH are the guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

3.2 Summary of personal contributions

This chapter uses simulations and applied examples to demonstrate how Mendelian randomisation mediation methods, multivariable Mendelian randomisation and two-step Mendelian randomisation, can be used to minimize bias in mediation analysis. I applied methods previously developed to decompose mediation effects. Prior to this work, multivariable Mendelian randomisation had largely been used to estimate direct effects, either in the presence of pleiotropy or potential mediation. Two-step, or network Mendelian randomisation had predominantly been used to infer causal mediation pathways, but had not decomposed the direct effect, indirect effect or proportion mediated.

These analyses were primarily carried out with support from Dr. Eleanor Sanderson who advised on and carried out simulation analyses. A version of this work has been posted to the preprint server BioRxiv (doi: https://doi.org/10.1101/835819) and is currently under peer review.

My role in this work was to assist in developing simulations to include, such as deciding which sources of bias to include, which methods to simulate and which simulations would contribute most to the existing literature. Additionally, I was responsible for carrying out applied phenotypic and Mendelian randomisation analyses in UK Biobank. I was responsible

for collating all results and creating publication quality tables and figures. I drafted the manuscript, which was advised and informed by comments from all co-authors.

Due to journal word limits it was agreed by all co-authors to focus the manuscript on the main conclusions from the simulation scenarios, with the full results available as supplementary material. The applied examples were presented as a standalone supplementary material. In this thesis chapter, the applied example has been integrated with the manuscript. Some simulation results are presented within the chapter, however due to the volume of tables and results, some are included in the appendix and referenced to throughout the chapter. To enhance readability, this chapter does not follow a standard IMRAD structure.

Full contributions from myself include devising the project, writing and circulating the analysis plan, cleaning the UK Biobank data, analysis and interpreting the results, writing and drafting the manuscript, submitting the manuscript, responding to and revising according to peer review comments.

3.3 Abstract

Background

Mendelian randomisation uses genetic variants randomly allocated at conception as instrumental variables for a modifiable exposure of interest. Recent methodological advances allow for mediation analysis to be carried out using Mendelian randomisation. When genetic instruments are available for both an exposure and mediator, both multivariable and two-step Mendelian randomisation may be applied.

Methods

I use simulations and an applied example to demonstrate when multivariable Mendelian randomisation and two-step Mendelian randomisation methods are valid and how they relate to traditional phenotypic regression-based approaches to mediation. I demonstrate how Mendelian randomisation methods can relax assumptions required for causal inference in phenotypic mediation, as well as which Mendelian randomisation specific assumptions are required. I illustrate these methods in data from UK Biobank, estimating the role of body mass index and low-density lipoprotein cholesterol mediating the association between education and cardiovascular outcomes.

Results

Both multivariable Mendelian randomisation and two-step Mendelian randomisation are unbiased when estimating the total effect, direct effect, indirect effect and proportion mediated when both confounding, and measurement error are present. Where both the exposure and mediator are continuous, in the presence of a rare or common binary outcome, we found little evidence of bias from non-collapsibility of the odds ratio.

Conclusion

Phenotypic mediation methods require strong, often untestable, assumptions. Mendelian randomisation provides an opportunity for improving causal inference in mediation analysis. Although Mendelian randomisation specific assumptions apply, such as no weak instrument bias and no pleiotropic pathways, strong assumptions of no confounding and no measurement error can be relaxed.

3.4 Introduction

Mediation analysis can improve aetiological understanding and identify intermediate variables as potential intervention targets when intervening on an exposure is not feasible. However, in order to make causal inferences, phenotypic mediation analysis requires strong assumptions. Mendelian randomisation (MR) is an alternative causal inference approach using genetic variants as instrumental variables (IV) for a phenotype (245). In this chapter phenotypic regression-based methods for mediation analysis are compared with MR methods for mediation analysis, and the assumptions required for MR mediation methods to make valid causal inferences are described.

3.4.1 Mediation analysis

Methods for mediation analysis emerged in the early twentieth-century, although often not described as such at the time, with formal methods developed by Baron and Kenny in the 1980s (225, 227). More recently, a large amount of research has built on and improved mediation methods for better causal inference (241).

Three parameters are typically estimated in traditional mediation analysis i) the total effect (the effect of the exposure on the outcome through all potential pathways) ii) the direct effect (the remaining effect of the exposure on the outcome that acts through pathways other than the specified mediator or set of mediators) and iii) the indirect effect (the path from exposure to outcome that acts through the mediator(s)). In situations where the total effect, direct effect and indirect effect all act in the same direction, an estimate of the "proportion mediated" (i.e. proportion of the total effect explained by the mediator) can be calculated. Two common approaches to estimate the indirect effect are; the product of coefficients method and the difference in coefficients method (274) (see Figure 3.1 A).

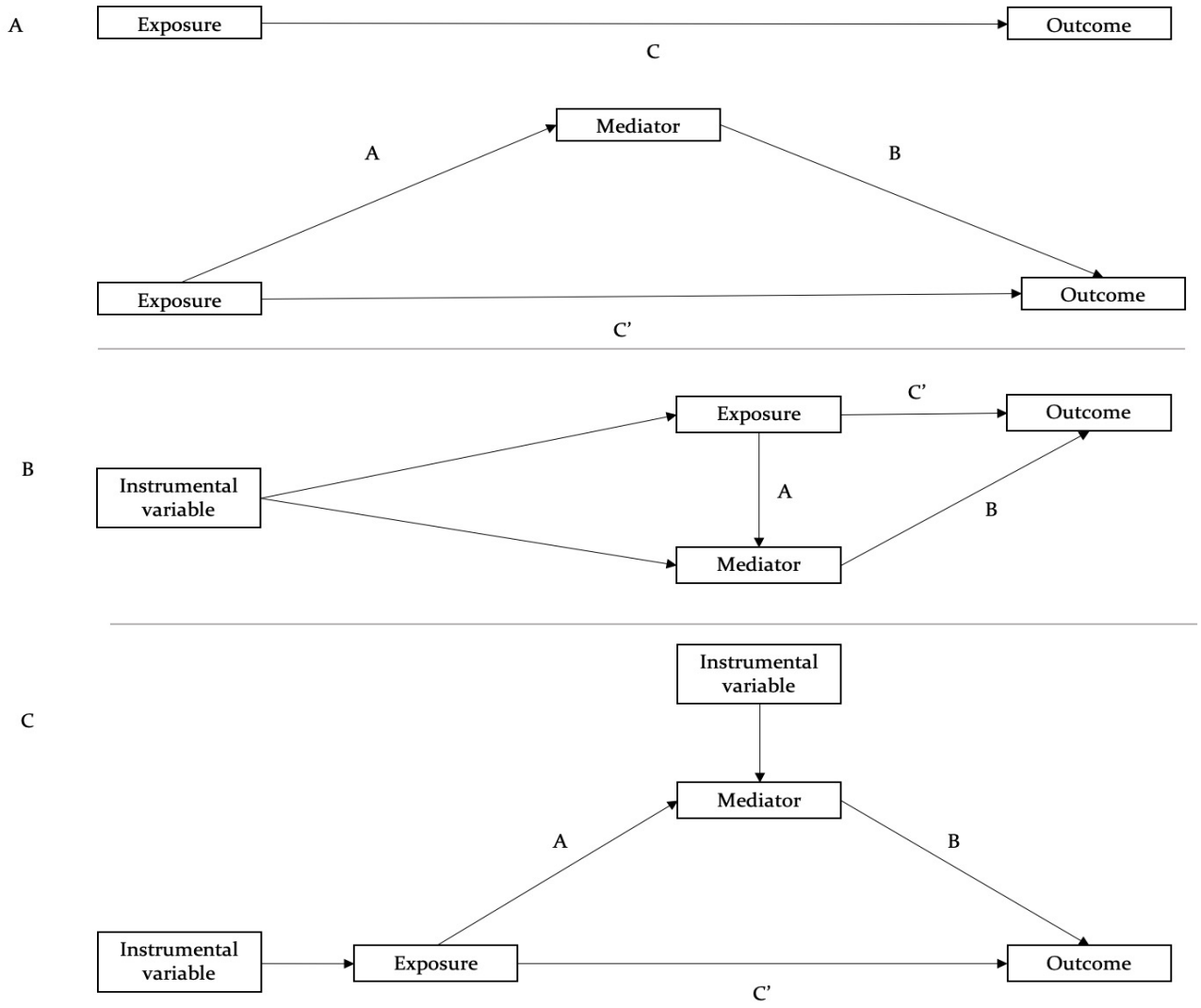

*Figure 3.1: The decomposed effects in A) phenotypic regression-based mediation analysis where C represents the total effect, C' represents the direct effect and the indirect effect can be calculated by subtracting C' from C (difference method) or multiplying A times B (product of coefficients method) B) multivariable MR, using a combined genetic instrument for both the exposure and mediator of interest, to estimate the direct effect (C') of the exposure and C) two-step Mendelian randomisation, where the effect of the exposure on the mediator (A) and mediator on the outcome (B) are estimated separately, using separate genetic instrumental variables for both the exposure and mediator. These estimates are then multiplied together to estimate the indirect effect of the mediator (A*B)*

Traditional mediation methods, such as Baron and Kenny methods, rely on several strong, untestable assumptions including, among others i) a causal effect of the exposure on the outcome, exposure on the mediator and mediator on the outcome ii) no unmeasured confounding between the exposure, mediator and outcome iii) no exposure-caused confounders of the mediator and outcome (intermediate confounders, see Figure 3.2 A) and iv) no exposure-mediator interaction (111, 229, 241). Furthermore, measurement error in either the exposure or mediator can introduce bias (275).

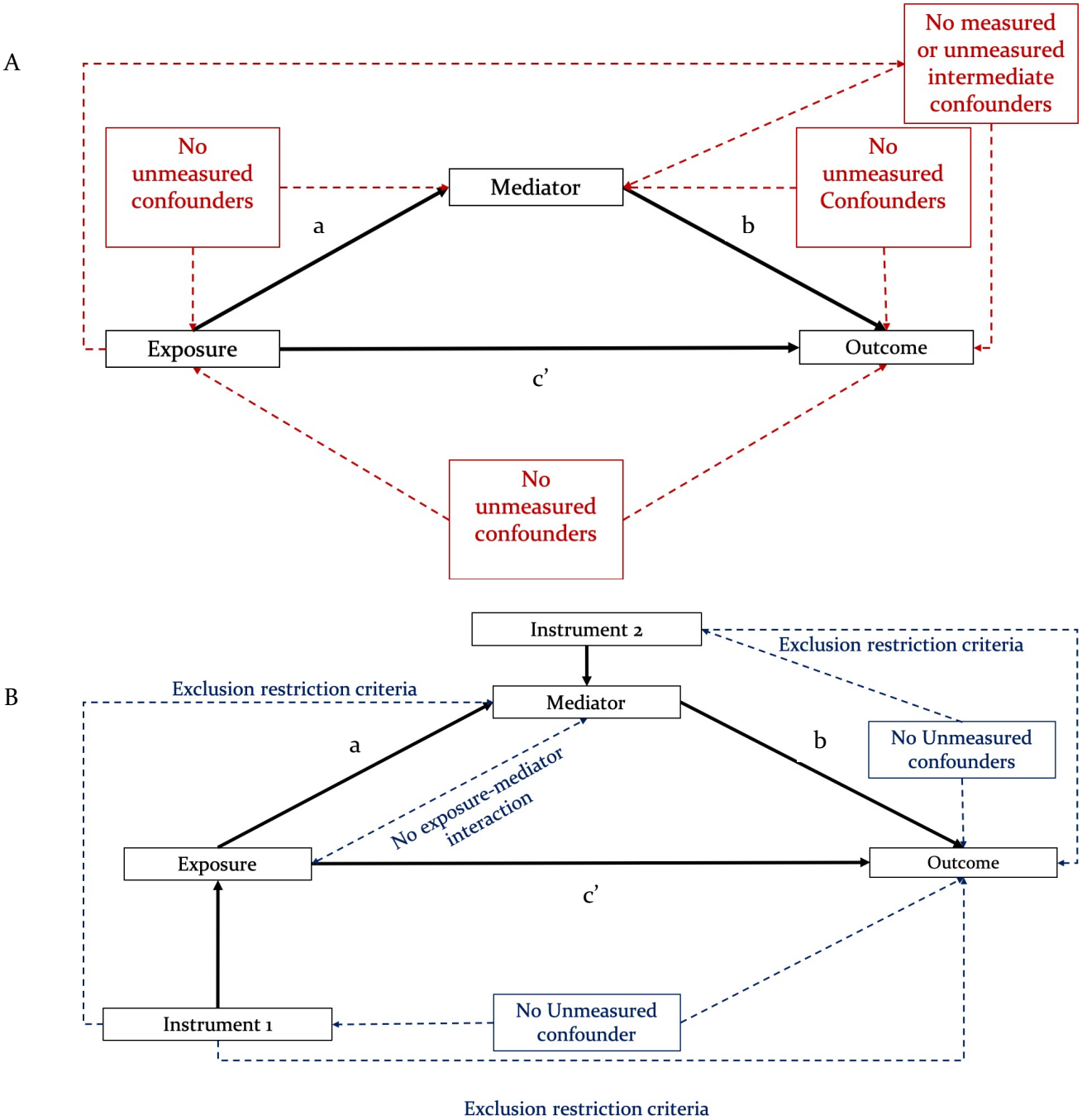

Figure 3.2: Schematic diagram illustrating the causal assumptions (dashed lines) in A) phenotypic regression-based mediation methods and B) Mendelian randomisation mediation analysis with the measured associations in solid black lines.

Additional assumption for phenotypic mediation is that of no measurement error in the exposure or mediator

In Mendelian randomisation, the exclusion restriction criteria mean there are no alternative pathways from the instrument to the outcome other than via the exposure (or mediator) of interest.

Baron and Kenny methods were introduced to estimate mediation with a continuous exposure, outcome and mediator, although they are also now often applied to binary variables. In the presence of a continuous or rare binary outcome the estimates from the difference in coefficients and the product of coefficients method should coincide (228, 241). Where effects are estimated on the odds ratio scale, the causal effects are only approximated

due to non-collapsibility of odds ratios, where the association between an exposure and outcome would not be constant by strata of categorical covariate. This is a major limitation as binary disease status is often of interest as an outcome.

Counterfactual reasoning has been used to develop methods that can address some of the previously described strong assumptions (228, 238, 240, 276, 277). These methods can estimate mediation in the presence of exposure-mediator interactions and account for measured intermediate confounders. Additionally, these more flexible counterfactual methods can allow for binary mediators and outcomes. However, these methods remain biased in the presence of unmeasured confounding, measurement error in the exposure or mediator, or in a mis-specified model with reverse causality (241, 278). In counterfactual methods, the estimated direct effect is described as being a "controlled direct effect" (CDE) if the value of the mediator is controlled at a certain value for all individuals in the population, or a "natural direct effect" (NDE), when the value of the mediator is allowed to take the value for each person that it would have taken naturally had they been unexposed, in a counterfactual scenario. The "natural indirect effect" (NIE) represents the average change in an outcome if the value of the exposure was fixed, but the value of the mediator changes from its natural value when exposed to its natural value when unexposed. If there is no interaction between the exposure and mediator, the estimate of the natural direct effect is equivalent to the controlled direct effect, and indeed would align with estimates from Baron and Kenny approaches to mediation (228, 237, 241).

3.4.2 Mendelian randomisation

In MR, randomly allocated genetic variants are used as IVs for a phenotype (18, 245, 279). Given the random allocation of genetic variants at meiosis and conception, MR estimates are robust to bias from confounding, reverse causation and non-differential measurement error (279). Three core assumptions are required for a genetic variant to be a valid instrumental variable, these are i) the genetic variants are robustly associated with the exposure (the relevance assumption) ii) the genetic instruments are exchangeable with the outcome (the independence assumption) and iii) the genetic variants do not affect the outcome via any variable other than the exposure (the exclusion restriction criteria) (Figure 3.3) (245).

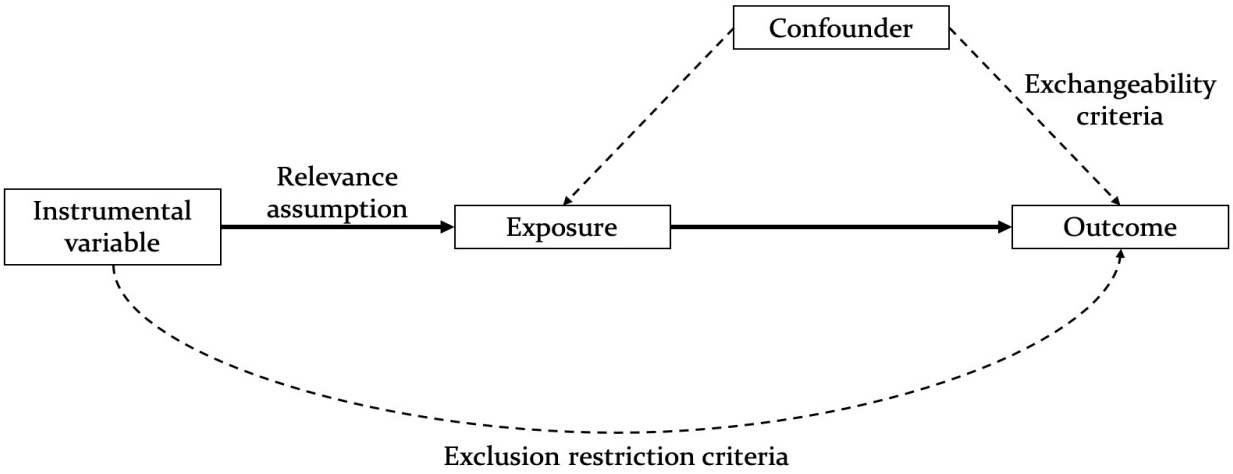

Figure 3.3: Directed acyclic graph illustrating Mendelian randomisation and the instrumental variable assumptions required for valid inference

3.4.3 Rationale for using Mendelian randomisation in mediation analysis

Mendelian randomisation can be used to overcome some of the previously described strong assumptions required for causal inference in mediation analysis. For example, estimates can be robust to bias from specific forms of unmeasured confounding, including that of intermediate confounding, and estimates cannot be biased by reverse causality.

In mediation terms, a univariable MR estimates the total effect of the exposure on the outcome. Two differing MR approaches can then be used which broadly mirror traditional phenotypic regression-based approaches to mediation to decompose the direct and indirect effects: multivariable MR (MVMR) (253, 280) and two-step MR (19, 20, 281).

In MVMR the controlled direct effect of the exposure on the outcome, controlling for the mediator, is estimated (20, 253). The genetic instruments for both the primary exposure and the second exposure (mediator) are included as instruments in the analysis (Figure 3.1 A) (282, 283). The indirect effect can then be estimated by subtracting the direct effect from the total effect (akin to the difference in coefficients method). MVMR assumes no interaction between the exposure and the mediator; therefore, the CDE estimated is equivalent to the NDE where this assumption holds true. As such, this is referred to as the direct effect, without further distinction, throughout this chapter.

Two-step Mendelian randomisation (also known as network MR) is akin to the product of coefficient methods. Two MR estimates are calculated i) the causal effect of the exposure on the mediator and ii) the causal effect of the mediator on the outcome (Figure 3.1 B) (19, 20,

248). These two estimates can then be multiplied together to estimate the indirect effect. Two-step MR also assumes no interaction between the exposure and the mediator.

These MR methods are increasingly being used in mediation analysis (253, 284-287). In this chapter, I demonstrate how MVMR and two-step MR can be used to estimate the direct effect, indirect effect and the proportion mediated, and which assumptions are required for the resulting estimates to be unbiased (20, 282, 283). I provide guidance about how to carry out each method, with code provided, and illustrate each method using both simulated and real data, applied to an individual level MR analysis.

3.5 Methods

3.5.1 Simulation study

Data were simulated under the model illustrated in Figure 3.1 with continuous, rare binary (5% prevalence) and common binary (25% prevalence) outcomes. The size of the total effect of the exposure, direct effect of the exposure and proportion mediated were varied. Additionally, results were simulated where the total effect of the exposure on the outcome is small, and where each of the exposure and mediator were subject to non-differential measurement error. Finally, simulations were used to show how MR methods can estimate mediation in the presence of multiple mediators, these simulations are illustrated in Figure 3.4. The full range of scenarios simulated are presented in Table 3.1. Simulation analyses were carried out using R version 3.5.1 and the corresponding code for the simulation studies can be found at https://github.com/eleanorsanderson/MediationMR.

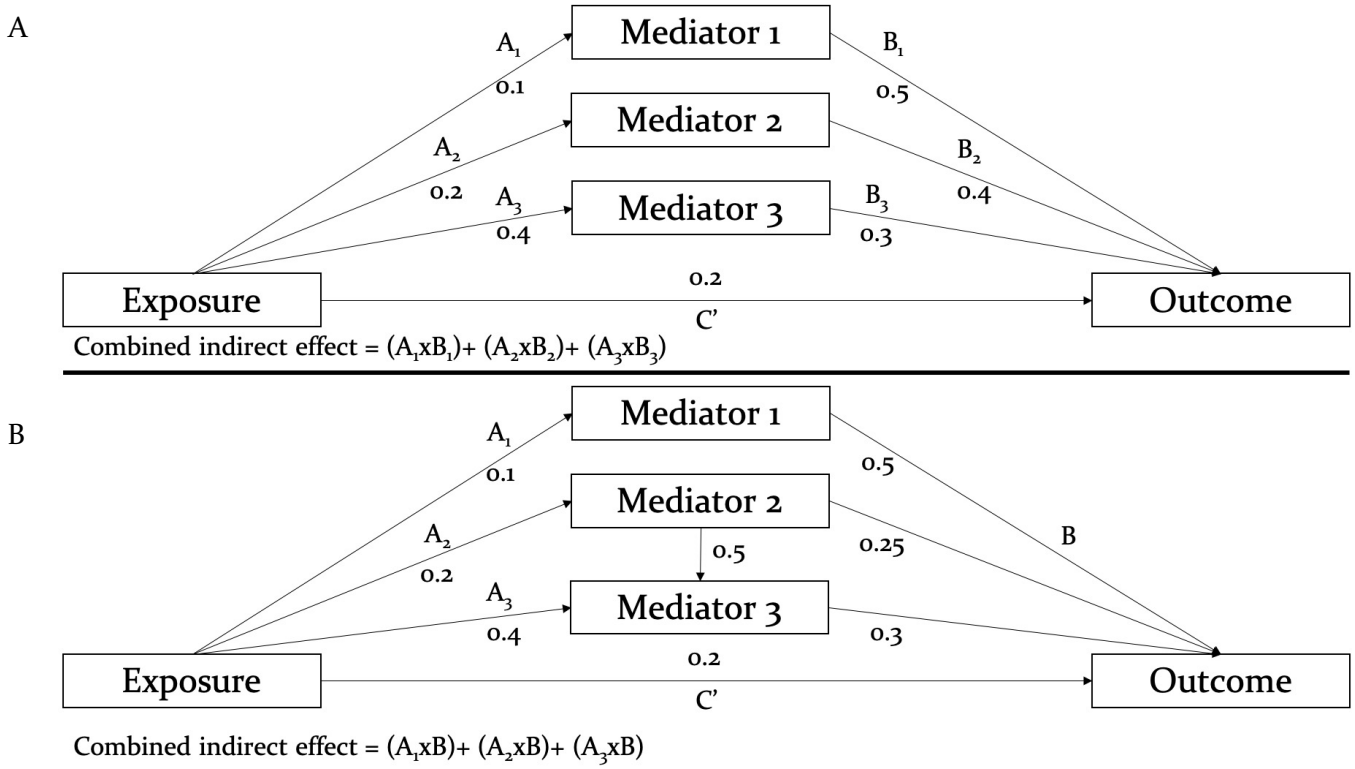

Figure 3.4: Directed acyclic graphs depicting simulation scenarios considering the role of multiple mediators where in A) all three mediators are independent and in B) there is covariance between two of the three mediators

Table 3.1: Simulation scenarios
*In all simulations the effect of the mediator on the outcome is set to 0.2. All simulations undergo 1000 replications. Confounding is simulated as residual covariance between the exposure, mediator and outcome in all scenarios except **

	Total effect	**Proportion mediated**					**Sample Size**	**Measurement error**	**Weak instrument**
No Mediation	0.5	0					5000		
Inconsistent mediation	0.5	-0.5					5000		
Varying total effect	0	0.05	0.25	0.75			5000		
	0.2	0.05	0.25	0.75			5000		
	0.5	0.05	0.25	0.75			5000		
	1	0.05	0.25	0.75			5000		
Small total effect	0.01	0.05	0.25	0.75			5000		
	0.05	0.05	0.25	0.75			5000		
	0.1	0.05	0.25	0.75			5000		
Imprecise total effect	0.2	0.05	0.25	0.75			1000		
Measurement error	0.5	0.25					5000	Exposure	
	0.5	0.25					5000	Mediator	
Weak instrument bias	0.5	0.25					5000		Exposure
	0.5	0.25					5000		Mediator
No confounding*	0	0.05	0.25	0.75			5000		
Multiple mediators		Joint	M1	M2	M3	M3 via M2	5000		
	0.45	0.56	0.11	0.18	0.12	0			
		0.56	0.11	0.18	0.27	0.06			

3.5.2 Applied example

Using data from UK Biobank (N = 184 778) (see Figure 3.5: Flow chart for exclusions made in UK Biobank for resultant sample for mediation analysis), I investigated the role of body mass index (BMI) and low-density lipoprotein cholesterol (LDL-C) in mediating the associations of education with systolic blood pressure, cardiovascular disease (CVD) and hypertension (continuous, rare binary and common binary outcomes, respectively). The effects on binary outcomes (hypertension and incident CVD) were estimated on risk difference, log odds ratio, and odds ratio scales. Applied analyses were performed using Stata version 15 (StataCorp LP, Texas) and corresponding code is available at https://github.com/alicerosecarter/MediationMR.

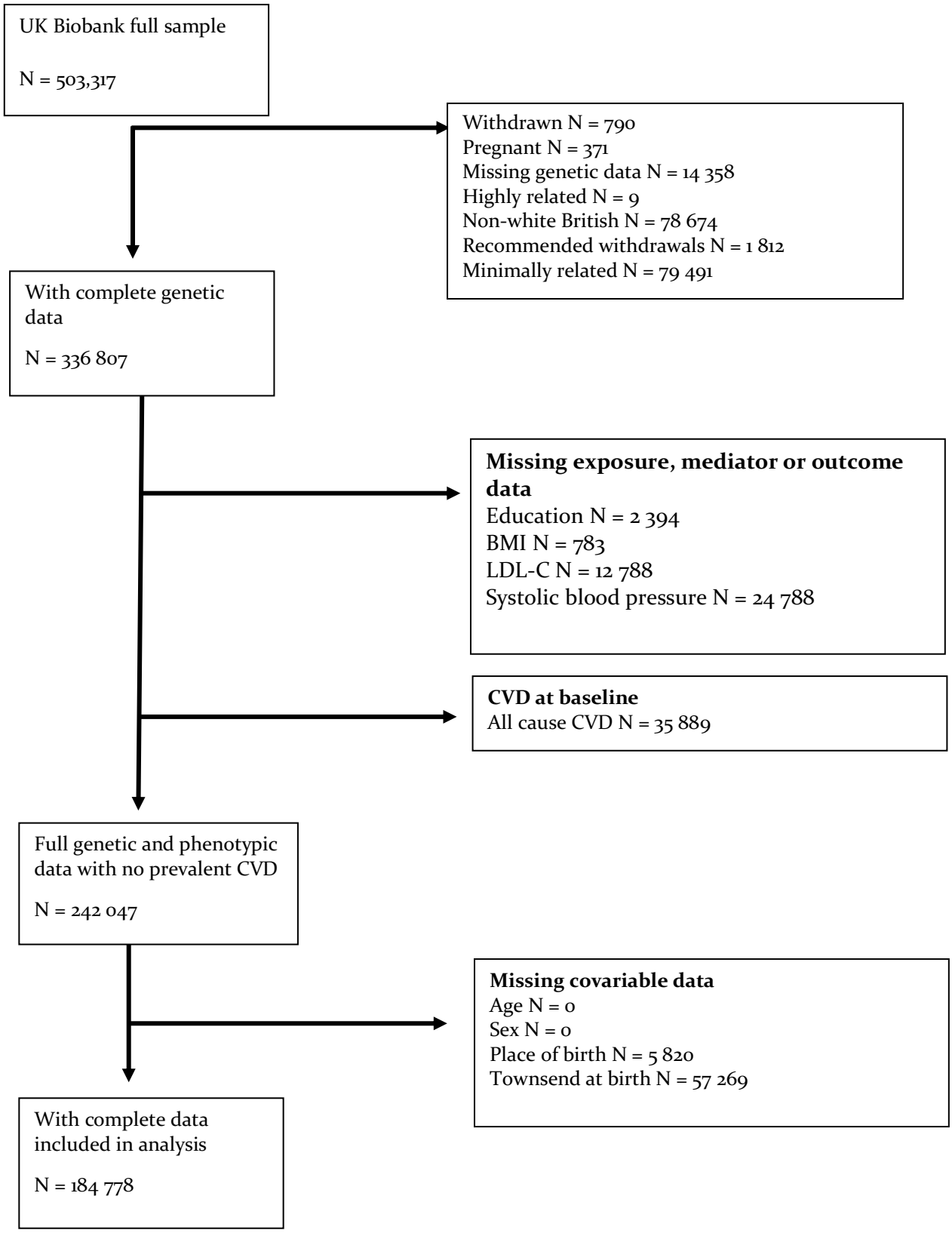

Figure 3.5: Flow chart for exclusions made in UK Biobank for resultant sample for mediation analysis

Note: At each stage the same participant could have missing data for multiple variables, therefore overlap is present between the variables. The total excluded may be less than the sum of individuals at each stage.

3.5.2.1 UK Biobank

At baseline, UK Biobank participants (N = 503 317) took part in questionnaires, interviews, anthropometric, physical and genetic measurements (15, 16). A total of 184 778 individuals of White British ancestry, with complete data on genotypes, age, sex, educational attainment, cardiovascular outcomes, BMI, LDL-C, blood pressure (including hypertension), socioeconomic position (as measured by Townsend Deprivation Index at birth [TDI]) and place of birth. Individuals of White British descent were defined using both self-reported questionnaire data and having similar genetic ancestry (principal components [PCs]) to the European reference panel (from 10,000 genomes panel derived by UK Biobank) (288).

Data from the baseline assessment centre on highest qualifications completed, BMI, LDL-C, systolic blood pressure, hypertension, and all covariate measures (age, sex, place of birth and Townsend deprivation index at birth) were used for the analyses.

3.5.2.2 Genetic exclusion criteria

Individuals were excluded if their genetic sex differed to their gender reported at the assessment centre or for having aneuploidy of their sex chromosomes. Further individuals were excluded for being outliers for their heterozygosity and any missing genetic data. Related individuals were also excluded from analyses, and the remaining subset was a maximal set of unrelated individuals. This exclusion list was derived in-house using an algorithm applied to the list of all the related pairs provided by UK Biobank (3rd degree or closer) (Figure 3.5). It preferentially removes the individuals related to the greatest number of other individuals until no related pairs remain (288).

3.5.2.3 Education

Participants reported their highest qualification at the baseline assessment centre ranging from no qualifications (equivalent to leaving school after 7 years) up to degree level (equivalent to 20 years of schooling). These were converted to the International Standard Classification for Education (ISCED) coding of educational attainment (Table 3.2) (149).

Table 3.2 International Standard for Classification of Education codes mapped to UK Biobank self-report highest qualification to estimate years of education

Qualification (as reported in UK Biobank)	ISCED	Years of education	N
College or University degree	5	20	61 037
NVQ or HND or HNC or equivalent	5	19	11 775
Other prof. qual. e.g.: nursing, teaching	4	15	9 154
A levels/AS levels or equivalent	3	13	22 190
O levels/GCSEs or equivalent	2	10	42 194
CSEs or equivalent	2	10	10 662
None of the above	1	7	27 806
Prefer not to answer	Excluded		

Mendelian randomisation studies require independent samples for the genetic variant-exposure discovery genome-wide association study (GWAS) and analysis sample. If samples overlap MR estimates can be overestimated (289). Therefore, in this analysis, genetic variants were selected from GWAS that did not include UK Biobank, and as such, are not always the most recent GWAS (244, 289).

To estimate the education polygenic score (PGS), 72 independent single-nucleotide polymorphisms (SNPs) that attained genome-wide significance ($P<5x10^{-8}$) for education reported in main results from an earlier 2016 SSGAC GWAS meta-analysis of 293,723 individuals and were available in the UK Biobank genotyping platform, to create a weighted allele score (17). Alleles were harmonised to all reflect education increasing single nucleotide polymorphisms (SNPs) and individual variants were recoded as 0, 1 or 2 according to the number of education increasing alleles. A genetic score for education was created by weighting each SNP by its relative effect size in the GWAS and summing all variants together in an additive model. Five instruments for education were not available in UK Biobank and proxy SNPs in perfect LD ($r^2=1$) were used (Table 3.3).

Table 3.3: Proxy SNPs for education instrument used in one-sample MR analysis

GWAS SNP (Okbay)	SNP in LD used (UKBB)
rs114598875	rs17538393
rs148734725	rs9878943
rs9320913	rs1487445
rs8005528	rs8008779
rs192818565	rs55943044

For phenotypic sensitivity analyses to further test the non-collapsibility of odds ratios, a binary measure of low and high education was created. Individuals who left school with 10 years or less of education (equivalent to a highest qualification of GCSE, or equivalent) were classed as low education. Individuals who completed further education after GCSEs were classed as having high education.

3.5.2.4 Body mass index

Clinic nurses at baseline assessment centres measured participants' height (m) and weight (kg), which was used to calculate BMI (kg/m^2).

To estimate the BMI PGS, 69 independent SNPs, available on the UK Biobank genotyping platform, which had attained genome-wide significance ($P<5\times10^{-8}$) for BMI in both males and females of European ancestry in the Genetic Investigation of ANthropometric Traits (GIANT) Consortium GWAS, which did not include UK Biobank participants (290). Alleles were harmonised to all reflect BMI increasing SNPs and individual variants were recoded as 0, 1 or 2 according to the number of BMI increasing alleles. A genetic score for BMI was created by weighting each SNP by its relative effect size in the GWAS and summing all variants together in an additive model.

In phenotypic sensitivity analyses a binary measure of BMI was created. Individuals with a BMI of less than $25kg/m^2$ were grouped together as normal or underweight individuals. Those with a BMI of $25kg/m^2$ or higher were grouped together as overweight or obese individuals.

3.5.2.5 Low density lipoprotein cholesterol

Direct low-density lipoprotein cholesterol (LDL-C) was measured from serum samples collected at baseline, using the Enzymatic Selective Protection Method.

To estimate the LDL-C PGS, 9 independent SNPs (Table 3.4) which had attained genome-wide significance ($P<5x10^{-8}$) for LDL-C in both males and females of predominantly European ancestry from the Global Lipids Genetics consortium (291). Genetic variants for LDL-C are often also associated with high density lipoprotein cholesterol and triglycerides. To avoid bias from pleiotropy, any SNP that was associated with LDL-C and at least one other lipid trait (as reported by Willer *et al*) was excluded from MR analysis ($N_{SNP}=51$). Alleles were harmonised to all reflect LDL-C increasing SNPs and individual variants were recoded as 0, 1 or 2 according to the number of LDL-C increasing alleles. The PGS was weighted by each SNP by its relative effect size in the GWAS and summing all variants together in an additive model.

Table 3.4: Independent SNPs used as instruments for LDL-C

SNP (RSID)	Effect allele	Other allele	Chromosome	hg19 Position (Mb)	Beta	P value
rs267733	G	A	1	150.96	-0.0331	$5.29x10^{-09}$
rs2710642	A	G	2	63.15	0.0239	$6.09x10^{-09}$
rs1250229	C	T	2	216.3	0.0243	$3.13x10^{-08}$
rs4942486	C	T	13	32.95	-0.0243	$2.26x10^{-11}$
rs8017377	A	G	14	24.88	0.0303	$2.52x10^{-15}$
rs1801689	C	A	17	64.21	0.1028	$9.81x10^{-12}$
rs364585	G	A	20	12.96	0.0249	$4.28x10^{-10}$
rs2328223	C	A	20	17.85	0.0299	$5.63x10^{-09}$
rs5763662	T	C	22	30.38	0.0767	$1.19x10^{-08}$

3.5.2.6 Blood pressure

Systolic and diastolic blood pressure were both recorded automatically and manually at the baseline assessment centre. All participants had an automatic reading, but manual readings were only taken for a subset. Each reading was taken twice, two minutes apart. This analysis uses the second reading of the automated blood pressure, where missing data were replaced with the first measure.

A binary measure of hypertension was created according to the World Health Organization's standard classification for hypertension (SBP ≥ 140 mm Hg and DBP ≥ 90 mm Hg) or if an individual was taking antihypertensive medication as recorded at the nurse's interviews.

3.5.2.7 Cardiovascular disease

Cardiovascular disease diagnoses were ascertained through linkage mortality data and hospital episode statistics (HES) and Scottish morbidity records (SMR) (referred to jointly as hospital inpatient records), with cases (all subtypes) defined according to ICD-9 (390-459) and ICD-10 codes (all I codes and G45) (292). Individuals who had experienced a CVD event prior to the baseline assessment (prevalent cases) were excluded and only first event, incident cases following the assessment centre were considered. Date of diagnoses are provided by hospital inpatient records, which was linked with the date of assessment centre provided by UK Biobank to identify incident and prevalent cases. All UK Biobank participants are linked to either HES data or SMRs, with data available from 1997 in England, 1998 in Wales and 1981 in Scotland (293), with the most recent entry recorded in this analysis in May 2017.

3.5.2.8 Covariates

Variables considered as confounders were measured at the baseline assessment centres through interviews. Covariates considered were age, sex, place of birth (northing and easting co-ordinates), birth distance from London, and TDI at birth. Sex and ethnicity were confirmed according to genetic data. Place of Birth was adjusted for by the northing and easting birth location coordinates. Although the TDI of historic birth location is not recorded in UK Biobank, this has been estimated from the index of multiple deprivation indices using the current TDI of birth location as a proxy for historic birthplace TDI. Mendelian randomisation models were also adjusted for the same confounders. Although a core assumption of MR is that the genetic variants are unrelated to confounders, there is some evidence of associations with place of birth for the educational attainment variants in UK Biobank (8).

3.5.3 **Statistical analysis**

The following approaches were applied to both applied analyses and simulated data.

Using the notation X = exposure, M = mediator, M1 = mediator 1, M2 = mediator 2, M3 = mediator 3, Y = outcome, G = genetic instruments, C = measured confounders, V = uncorrelated error term, μ = uncorrelated error term, four methods are compared (figure 1). Notation and equations for the difference method and product of coefficients method are adapted from Vanderweele, 2015, where full details of the equations and notations are available (294). Variables and parameters given in bold indicate the main coefficient(s) of interest in each case.

All phenotypic analyses were adjusted for potential confounders; age, sex, place of birth, birth distance from London, and Townsend deprivation index at birth. Mendelian Randomisation analyses were adjusted for the same confounders, in addition to the 40 genetic principal components (derived by UK Biobank) to account for population structure.

3.5.3.1 Difference in coefficients method

Each outcome was regressed on the exposure adjusting for the mediator to estimate the direct effect of the exposure. The direct effect was subtracted from the total effect, estimated using multivariable regression adjusting for potential confounders, to estimate the indirect effect. In all simulation scenarios the standard deviation of the regression coefficients was calculated across repeats to evaluate precision.

Total: $Y = \theta^{+}_{0} + \boldsymbol{\theta^{+}_{1}X} + \theta^{+}_{3}C$

Direct: $Y = \theta_{0} + \boldsymbol{\theta_{1}X} + \theta_{2}M + \theta_{4}C$

Indirect: $\theta^{+}_{1} - \theta_{1}$

3.5.3.2 Product of coefficients method

Two regression models were estimated. Firstly, the mediator was regressed on the exposure. Secondly, the outcome was regressed on the mediator, adjusting for the exposure. These two estimates were multiplied together to estimate the indirect effect. In applied analyses, confidence intervals for the indirect effect were derived from bootstrapping with 100 replications.

Exposure-Mediator: $M = \beta_{0} + \boldsymbol{\beta_{1}X} + \beta_{3}C$

Direct: $Y = \theta_{0} + \boldsymbol{\theta_{1}X} + \theta_{2}M + \theta_{4}C$

Indirect: $\beta_{1}\theta_{2}$

3.5.3.3 Multivariable Mendelian randomisation

Using MVMR to estimate the direct effect, in the first stage regression, the weighted allele score for the exposure and the weighted allele score for the mediator are used to predict each exposure respectively, conditional on each other. In the second stage regression, the outcome was regressed on the predicted values of each exposure. The direct effect was then subtracted from the total effect, estimated using two-stage least squares regression, to estimate the indirect effect.

Total:

$$X = \pi_{0} + \pi_{1}G_{x} + v_{1}$$

$$Y = \beta_{0} + \boldsymbol{\beta_{XT}X} + \mu_{1}$$

Direct:

$$X = \pi_{0} + \pi_{1x}G_{x} + \pi_{2x}G_{M} + v_{1}$$

$$M = \pi_{0} + \pi_{1z}G_{x} + \pi_{2z}G_{M} + v_{2}$$

$$Y = \beta_{0} + \boldsymbol{\beta_{X}X} + \beta_{M}M + \mu_{2}$$

Indirect:

$$\boldsymbol{\beta_{XT} - \beta_{X}}$$

3.5.3.4 Two-step Mendelian randomisation

A univariable MR model was carried out to estimate the effect of the exposure on the mediator. A second model estimating the effect of the mediator on each outcome was carried

out using MVMR. Both the genetic variants for the mediator and the exposure were included in the first and second stage regressions in MVMR. Previous approaches in the literature have not used MVMR for this second step (19, 20) and propose carrying out a univariable MR of the effect of the mediator on the outcome. However, using MVMR ensures any effect of the mediator on the outcome is independent of the exposure. Additionally, this method provides an estimate of the direct effect of the exposure on the outcome. The two regression estimates from the second stage regression are multiplied together to estimate the indirect effect. In applied analyses, confidence intervals for the indirect effect were derived from bootstrapping with 100 replications.

Exposure-Mediator:

$$X = \pi_0 + \pi_1 G_X + v_1$$

$$M = \beta_0 + \boldsymbol{\beta_{XM} X} + \mu_1$$

Direct:

$$X = \pi_0 + \pi_{1x} G_X + \pi_{2x} G_M + v_1$$

$$M = \pi_0 + \pi_{1z} G_X + \pi_{2z} G_M + v_2$$

$$Y = \beta_0 + \boldsymbol{\beta_X X} + \beta_M M + \mu_2$$

Indirect:

$$\boldsymbol{\beta_{XM}\, \beta_M}$$

3.5.4 Multiple mediators

In phenotypic analyses, to estimate the direct effect attributable to multiple mediators, the outcome was regressed on the exposure, controlling for all mediators, using multivariable regression. Here, the coefficient for the exposure reflects the direct effect (295). This direct effect was then subtracted from the total effect to estimate the indirect effect. Secondly, the product of coefficients method was used to estimate the indirect effect of each mediator individually. The combined effect of all three mediators was then estimated by summing together each individual effect.

In MR analyses, the direct effect attributable to multiple mediators was assessed using MVMR, controlling for all mediators. This direct effect was then subtracted from the total effect to estimate the combined indirect effect. Secondly two-step MR was used, as previously described, considering each mediator individually and summing the effects together to obtain the indirect effect of all mediators combined.

Corresponding equations for these methods can be found in Appendix 1.

3.5.5 Proportion mediated

The proportion mediated is calculated by dividing the indirect effect by the total effect. In individual-level MR, the confidence intervals were estimated via bootstrapping with 100 replications.

3.5.6 Applied sensitivity analyses

In the applied phenotypic analysis, sensitivity analyses were carried out dichotomising education and/or BMI to a binary variable to further test non-collapsibility, where analyses were carried out on the log odds ratio scale. See supplementary methods for details.

Bidirectional univariable MR analysis was carried out to test whether mediator-mediator associations exist between BMI and LDL-C.

Instrument strength was assessed by calculating F-statistics for univariable MR and conditional F-statistics for MVMR (296). Sensitivity analyses for MR methods included using MR-Egger and MVMR-Egger to test for pleiotropy in the applied example (255, 297).

3.6 Applied analysis results

3.6.1 Participant characteristics

Descriptive characteristics of UK Biobank participants included in the real data example are shown in Table 3.5. To summarise, participants were more likely to be more highly educated, with 32% of participants leaving school after 20 years of education compared with 16% leaving with 7 years of education. Participants eligible for analyses were comparable to the full UK Biobank sample, although in the analysis sample hypertension was less prevalent. The prevalence of hypertension was 33% in the analysis sample compared with 40% in all participants.

Table 3.5: UK Biobank cohort descriptive statistics

Variable		Eligible Sample N = 184 778		All UK Biobank (excluding withdrawals) N = 502 527	
		Mean (SD) or N (%)		N	Mean (SD) or N (%)
Sex	Female	101 757 (55%)		502 527	273 396 (54%)
Age (at baseline)		56.21 (8.04)		502 527	56.53 (8.10)
Educational attainment (years)	7 10 13 15 19 20	27 806 (15%) 52 816 (29%) 22 190 (12%) 9 154 (5%) 11 775 (6%) 61 037 (33%)		492 393	85 275 (17%) 132 087 (11%) 55 325 (11%) 25 805 (5%) 32 730 (7%) 161 171 (33%)
Body mass index		27.07 (4.55)		499 422	27.43 (4.80)
Low-density lipoprotein cholesterol		3.63 (0.85)		468 727	3.56 (0.87)
Systolic blood pressure		137.89 (18.53)		456 985	137.78 (18.63)
Incident cardiovascular disease	Control Case	141 909 (77%) 42 869 (23%)		418 781	321 633 (77%) 97 148 (23%)
Hypertension	Control Case	124 119 (67%) 60 659 (33%)		467 429	282 816 (61%) 184 613 (40%)

3.6.2 Effect of education on systolic blood pressure, CVD and hypertension

Both multivariable regression and univariable MR provided evidence to support a causal effect of education on systolic blood pressure, as well as for a role of BMI mediating this effect on the risk difference scale. Phenotypically, the difference method estimated the indirect effect for a one standard deviation increase in education on systolic blood pressure mediated via a one standard deviation increase in BMI to be -0.33 mmHg (95% CI: -0.35 to -0.32) and the proportion mediated to be 27.7% (95% CI: 25.6% to 29.9%) (Table 3.6). Using MVMR the indirect effect estimated was -0.55 mmHg (95% CI: -0.83 to -0.28). Despite the MVMR indirect effect and total effect being larger than the phenotypic difference estimate, this corresponded to a smaller proportion mediated of 16.9% (95% CI: 8.6% to 25.2%).

Using the phenotypic product of coefficients method, the indirect effect of a one standard deviation increase in education via a one standard deviation (SD) increase in BMI on systolic blood pressure was -0.33 mmHg (95% CI: -0.35, -0.32) with a proportion mediated of 27.7% (95% CI: 25.7% to 29.8%) (Table 3.6). Comparatively, using two-step MR, the indirect effect was estimated to be -0.55 mmHg (95% CI: -0.85 to -0.26) corresponding to a proportion mediated of 16.9% (95% CI: 7.3% to 26.5%).

Both multivariable regression and univariable MR provided evidence to support a causal effect of education on CVD, including for a mediating role of BMI. For example, the indirect effect via a one SD increase in BMI on the effect of a standard deviation increase in education on incident CVD, was estimated to be reduce the risk of CVD by -0.02 (95% CI for MVMR and two-step MR: -0.02 to -0.01). The estimate of the proportion mediated via both MVMR and two-step MR was 21.0% (95% CI for MVMR: 11.0% to 30.9%; 95% CI for two-step MR: 10.3% to 31.6%). The estimates of the decomposed mediated effects were similar when analysed using the log odds ratio scale, however estimates had wider confidence intervals (Table 3.6).

Mendelian randomisation suggested more education reduced risk of hypertension; however, estimates were imprecise and confidence intervals were consistent with an increased risk. This led to large confidence intervals around the estimate of the proportion mediated by BMI. On the risk difference scale, the proportion mediated that was estimated by MVMR was 21.0% (95% CI: 3.7% to 38.2%) and by two-step MR was 22.7% (95% CI: 1.7% to 43.7%). Similar values were obtained using the log odds ratio scales (Table 3.6).

For both CVD and hypertension, the decomposed mediated effects estimated on the odds ratio scale were discordant compared with those on either the risk difference or log odds ratio scale.

Table 3.6: Real-data example estimating the mediating role of BMI independently between education and systolic blood pressure, cardiovascular disease and hypertension, using multivariable observational methods and mendelian randomisation methods

Outcome	Scale	Method	Total Effect (95% CI)	Direct effect (95% CI)	Difference method or MVMR		Product method or two-step MR	
					Indirect effect (95% CI)	Proportion mediated (95% CI)	Indirect effect (95% CI)	Proportion mediated (95% CI)
Systolic blood pressure	Mean difference	Phenotypic	-1.20 (-1.28, -1.12)	-0.87 (-0.95, -0.79)	-0.33 (-0.35, -0.32)	27.73 (25.62, 29.85)	-0.33 (-0.35, -0.32)	27.73 (25.70, 29.76)
		MR	-3.28 (-4.19, -2.37)	-2.73 (-3.68, -1.78)	-0.55 (-0.83, -0.28)	16.90 (8.60, 25.20)	-0.55 (-0.85, -0.26)	16.90 (7.29, 26.51)
Cardiovascular disease	Risk difference	Phenotypic	-0.03 (-0.03, -0.03)	-0.02 (-0.02, -0.02)	-0.01 (-0.01, -0.01)	22.24 (20.19, 24.28)	-0.01 (-0.01, -0.01)	22.24 (20.29, 24.19)
		MR	-0.08 (-0.11, -0.06)	-0.07 (-0.09, -0.04)	-0.02 (-0.02, -0.01)	20.97 (11.03, 30.91)	-0.02 (-0.02, -0.01)	20.97 (10.3, 31.63)
	Log odds ratio	Phenotypic	-0.16 (-0.17, -0.15)	-0.13 (-0.14, -0.11)	-0.03 (-0.03, -0.03)	20.49 (18.47, 22.51)	-0.04 (-0.04, -0.03)	22.84 (20.87, 24.82)
		MR	-0.50 (-0.63, -0.37)	-0.4 (-0.53, -0.26)	-0.11 (-0.14, -0.07)	21.11 (10.54, 31.68)	-0.11 (-0.14, -0.07)	21.15 (11.56, 30.75)
	Odds ratio	Phenotypic	0.85 (0.84, 0.86)	-0.13 (0.87, 0.89)	-0.03 (-0.03, -0.03)	-3.31 (-3.5, -3.11)	-0.15 (-0.15, -0.14)	-17.22 (-18.01, -16.44)
		MR	0.61 (0.53, 0.69)	0.67 (0.59, 0.77)	-0.07 (-0.09, -0.05)	-11.18 (-15.08, -7.27)	-0.49 (-0.57, -0.41)	-81.4 (-97.06, -65.73)
Hypertension	Risk difference	Phenotypic	-0.02 (-0.03, -0.02)	-0.02 (-0.02, -0.02)	-0.01 (-0.01, 0.01)	22.73 (20.54, 24.93)	-0.01 (-0.01, 0.01)	22.73 (20.14, 25.32)
		MR	-0.05 (-0.07, -0.02)	-0.04 (-0.06, -0.01)	-0.01 (-0.02, 0.01)	20.97 (3.73, 38.20)	-0.01 (-0.02, 0.01)	22.71 (1.69, 43.74)
	Log odds ratio	Phenotypic	-0.11 (-0.12, -0.10)	-0.09 (-0.1, -0.08)	-0.03 (-0.03, -0.02)	22.64 (20.37, 24.91)	-0.03 (-0.03, -0.03)	24.61 (22.16, 27.06)
		MR	-0.24 (-0.35, -0.12)	-0.18 (-0.31, -0.06)	-0.05 (-0.09, -0.02)	22.82 (-4.81, 50.45)	-0.05 (-0.08, -0.02)	22.87 (1.48, 44.26)
	Odds ratio	Phenotypic	0.89 (0.88, 0.90)	-0.09 (0.91, 0.93)	-0.02 (-0.02, -0.02)	-2.56 (-2.72, -2.40)	-0.13 (-0.14, -0.13)	-15.09 (-15.82, -14.37)
		MR	0.79 (0.7, 0.89)	0.83 (0.74, 0.95)	-0.04 (-0.07, -0.02)	-5.53 (-9.13, -1.93)	-0.43 (-0.5, -0.35)	-54.18 (-64.94, -43.42)

Difference = difference in coefficients method; MVMR; multivariable MR; product = product of coefficient method; MR = Mendelian randomisation; CI = confidence interval

There was little evidence that LDL-C mediates the effect of education on systolic blood pressure, hypertension and CVD (Table 3.7). Phenotypically, both the difference in coefficients method and product of coefficients method estimated 0.9% (95% CI (product method): 0.3% to 1.6%) of the effect of education on systolic blood pressure was mediated by LDL-C. In MR, both MVMR and two-step MR estimated the proportion mediated to be -1.8% (95% CI (two-step MR): -6.6 to 2.4). In both phenotypic and MR analyses, there was limited evidence that LDL-C mediated the effect of education on CVD or hypertension.

Table 3.7: Real-data example estimating the mediating role of low-density lipoprotein cholesterol independently between education and systolic blood pressure, cardiovascular disease and hypertension, using multivariable observational methods and mendelian randomisation methods

Outcome	Scale	Method	Total Effect (95% CI)	Direct effect (95% CI)	Difference method or MVMR		Product method or two-step MR	
					Indirect effect (95% CI)	Proportion mediated (95% CI)	Indirect effect (95% CI)	Proportion mediated (95% CI)
Systolic blood pressure	Mean difference	Phenotypic	-1.20 (-1.28, -1.12)	-1.18 (-1.26, -1.10)	-0.02 (-0.03, -0.01)	1.78 (1.14, 2.42)	-0.02 (-0.03, -0.01)	1.78 (1.09, 2.47)
		MR	-3.28 (-4.19, -2.37)	-3.37 (-4.30, -2.45)	0.09 (-0.09, 0.27)	-2.78 (-8.94, 3.38)	0.09 (-0.07, 0.25)	-2.78 (-8.75, 3.19)
Cardiovascular disease	Risk difference	Phenotypic	-0.03 (-0.03, -0.03)	-0.03 (-0.03, -0.03)	$2.86x10^{-4}$ ($1.78x10^{-4}$, $3.93x10^{-4}$)	-1.04 (-1.49, -0.58)	$2.86x10^{-4}$ ($1.81x10^{-4}$, $3.90x10^{-4}$)	-1.04 (-1.47, -0.61)
		MR	-0.08 (-0.11, -0.06)	-0.09 (-0.11, -0.06)	$6.43x10^{-4}$ ($-2.76x10^{-3}$, $4.04x10^{-3}$)	-0.76 (-4.57, 3.05)	$6.43x10^{-4}$ ($-2.49x10^{-3}$, $3.77x10^{-3}$)	-0.76 (-4.58, 3.06)
	Log odds ratio	Phenotypic	-0.16 (-0.17, -0.15)	-0.16 (-0.17, -0.15)	$9.88x10^{-4}$ ($4.15x10^{-4}$, $1.56x10^{-3}$)	-0.62 (-1.03, -0.22)	$1.48x10^{-3}$ ($8.68x10^{-4}$, $2.08x10^{-3}$)	-0.93 (-1.25, -0.61)
		MR	-0.50 (-0.63, -0.37)	-0.51 (-0.63, -0.38)	$3.94x10^{-3}$ (-0.01, 0.02)	-0.78 (-3.88, 2.31)	$3.93x10^{-3}$ (-0.02, 0.02)	-0.78 (-4.9, 3.33)
	Odds ratio	Phenotypic	0.85 (0.84, 0.86)	-0.16 (0.84, 0.86)	$8.43x10^{-4}$ ($3.72x10^{-4}$, $1.31x10^{-3}$)	0.10 (0.04, 0.15)	-0.01 (-0.01, -0.01)	-1.26 (-1.80, -0.72)
		MR	0.61 (0.53, 0.69)	0.60 (0.53, 0.69)	-0.05 (-0.09, 0.01)	0.39 (-1.28, 2.07)	-0.05 (-0.09, 0.01)	-7.46 (-15.31, 0.40)
Hypertension	Risk difference	Phenotypic	-0.02 (-0.03, -0.02)	-0.02 (-0.03, -0.02)	$4.78x10^{-4}$ ($3.14x10^{-4}$, $6.42x10^{-4}$)	-2.06 (-2.93, -1.19)	$4.78x10^{-4}$ ($2.91x10^{-4}$, $6.65x10^{-4}$)	-2.06 (-2.85, -1.27)
		MR	-0.05 (-0.07, -0.02)	-0.05 (-0.07, -0.03)	$1.82x10^{-3}$ ($-2.27x10^{-3}$, 0.01)	-3.91 (-14.18, 6.36)	$1.82x10^{-3}$ ($-2.21x10^{-3}$, 0.01)	-3.91 (-16.79, 8.97)
	Log odds ratio	Phenotypic	-0.11 (-0.12, -0.10)	-0.11 (-0.12, -0.10)	$1.87x10^{-3}$ ($1.00x10^{-3}$, $2.74x10^{-3}$)	-1.67 (-2.44, -0.91)	$2.24x10^{-3}$ ($1.50x10^{-3}$, $2.98x10^{-3}$)	-2.00 (-2.76, -1.25)
		MR	-0.24 (-0.35, -0.12)	-0.25 (-0.36, -0.13)	0.01 (0.01, 0.02)	-4.01 (-11.86, 3.84)	-0.01 (-0.01, 0.03)	-4.01 (-14.69, 6.68)
	Odds ratio	Phenotypic	0.89 (0.88, 0.90)	-0.09 (0.91, 0.93)	-0.02 (-0.02, -0.02)	-2.56 (-2.72, -2.40)	-0.13 (-0.14, -0.13)	-15.09 (-15.82, -14.37)
		MR	0.79 (0.70, 0.89)	0.78 (0.69, 0.88)	-0.04 (-0.08, 0.01)	0.94 (-0.60, 2.48)	-0.04 (-0.08, 0.01)	-5.11 (-9.41, -0.81)

Difference = difference in coefficients method; MVMR; multivariable MR; product = product of coefficient method; MR = Mendelian randomisation; CI = confidence interval

Table 3.8: Effect of a one standard deviation increase of body mass index (BMI) on low-density lipoprotein cholesterol (LDL-C) and a one standard deviation increase in LDL-C on BMI in a Mendelian randomisation analysis

Exposure	Outcome	Beta (95% CI)
BMI	LDL-C	0.51 (0.48, 0.55)
LDL-C	BMI	1.91 (1.62, 2.19)

3.6.3 Joint mediation by BMI and LDL-C

Considering BMI and LDL-C jointly, in phenotypic mediation using the difference method on the risk difference scale 28.4% (95% CI: 26.3% to 30.5%) of the association between education and systolic blood pressure was explained (Table 3.9), compared with 27.7% (95% CI: 25.62% to 29.9%) by BMI individually and 1.8% (95% CI: 1.1% to 2.4%) (Table 3.6) by LDL-C individually (Table 3.7). When considering CVD as the outcome 21.7% (95% CI: -20.1% to 23.4%) was explained by BMI and LDL-C jointly (Table 3.9), similar to the amount explained by BMI individually (22.2% [95% CI: 20.2 to 24.3%]) (Table 3.6). BMI and LDL-C jointly explained 21.8% (95% CI: 19.7% to 23.9%) of the association between education and hypertension (Table 3.9), again, this was similar to the amount explained by BMI individually (22.7% [95% CI: 20.5% to 24.9%]) (Table 3.6).

In MR analyses, using MVMR to estimate the combined proportion mediated on the risk difference scale, 12.6% (95% CI: 1.95% to 23.1%) was explained by BMI and LDL-C on the association between education and SBP (Table 3.9). This was less than the amount explained by BMI individually (16.9% [95% CI:8.6% to 25.2%]) (Table 3.6). BMI and LDL-C jointly explained 20.3% (95% CI: 18.5% to 22.0%) of the association between education and CVD (Table 3.9), similar to the amount explained by BMI individual (21.0% [95% CI:11.0% to 30.95]) (Table 3.6).Considering hypertension as the outcome, BMI and LDL-C jointly explained 21.9% (95% CI: 19.1% to 24.6%) of the association (Table 3.9). Again, this was similar to the amount explained by BMI individually (21.0% [95% CI: 3.7% to 38.2%]) (Table 3.6).

Table 3.9: Real-data example estimating the joint mediating role of BMI and LDL-C between education and systolic blood pressure (SBP), hypertension and cardiovascular disease (CVD) using multivariable observational methods and mendelian randomisation methods, where the joint direct effect was estimated using the difference in coefficients method, or multivariable mendelian randomisation method

Outcome	Scale	Method	Total Effect (95% CI)	Direct effect (95% CI)	Difference method or MVMR	
					Indirect effect (95% CI)	Proportion mediated (95% CI)
Systolic blood pressure	Mean difference	Phenotypic	-1.20 (-1.28, -1.12)	-0.86 (-0.94, -0.78)	-0.34 (-0.36, -0.32)	28.39 (26.29, 30.48)
		MR	-3.28 (-4.19, -2.37)	-2.87 (-3.84, -1.9)	-0.41 (-0.71, -0.12)	12.55 (1.95, 23.14)
Cardiovascular disease	Risk difference	Phenotypic	-0.03 (-0.03, -0.03)	-0.02 (-0.02, -0.02)	-0.01 (-0.01, -0.01)	21.73 (20.05, 23.42)
		MR	-0.16 (-0.17, -0.15)	-0.13 (-0.14, -0.12)	-0.03 (-0.03, -0.03)	20.25 (18.46, 22.03)
	Log odds ratio	Phenotypic	-0.08 (-0.11, -0.06)	-0.07 (-0.09, -0.05)	-0.02 (-0.02, -0.01)	19.50 (8.70, 30.29)
		MR	-0.50 (-0.63, -0.37)	-0.40 (-0.54, -0.27)	-0.10 (-0.14, -0.06)	19.59 (9.18, 29.99)
Hypertension	Risk difference	Phenotypic	-0.02 (-0.03, -0.02)	-0.02 (-0.02, -0.02)	-0.01 (-0.01, 0.01)	21.8 (19.70, 23.91)
		MR	-0.11 (-0.12, -0.10)	-0.09 (-0.10, -0.08)	-0.02 (-0.03, -0.02)	21.88 (19.13, 24.62)
	Log odds ratio	Phenotypic	-0.05 (-0.07, -0.02)	-0.04 (-0.06, -0.01)	-0.01 (-0.02, 0.01)	19.50 (-5.2, 44.19)
		MR	-0.24 (-0.35, -0.12)	-0.20 (-0.32, -0.07)	0.17 (-0.05, 0.39)	-71.08 (-223.18, 81.03)

Difference = difference in coefficients method; MVMR; multivariable MR; product = product of coefficient method; MR = Mendelian randomisation; CI = confidence interval

3.6.4 **Sensitivity analyses**

Applied examples using phenotypic mediation methods were extended to examine the role of binary exposures or mediators on non-collapsibility. In both rare and common binary outcomes, where the education exposure was dichotomized to low (10 years of education or less) compared with high education (greater than 10 years of education) the difference in coefficients method and product of coefficients method estimated similar mediating roles by a continuous standard deviation increase in BMI. For example, the proportion mediated by BMI on the association between education (high vs low) and CVD was 19.6% (95% CI: 17.7% to 21.4%) for the difference method and 22.0% (95% CI: 20.0% to 24.1%) for the product of coefficients method. Where the mediator was binary (normal and underweight vs overweight and obese) the two methods diverged. For example, the proportion mediated by high versus low BMI on the association between a one SD increase in education and incident CVD was 11.7% (95% CI: 10.5% to 12.8%) for the difference in coefficients method and 62.7% (95 % CI: 57.2% to 68.1%) for the product of coefficients method. This was similar when both the exposure and outcome were considered as binary. Similar results were also seen when considering common hypertension as the outcome (Table 3.10). Where both the mediator and outcome are binary, counterfactual methods for mediation analysis should be considered.

All instruments had strong F statistics (215 to 3094) and conditional F statistics (214 to 2457) (Table 3.11).

Both MR-Egger and MVMR-Egger provide little evidence to support pleiotropic effects of the instruments biasing results (Table 3.12)

Table 3.10: Evaluating non-collapsibility in real-data example with binary exposures and/or binary mediators with a rare binary and common binary outcome on the log odds ratio scale using phenotypic mediation methods

Outcome	Exposure (education)	Mediator (BMI)	Total effect (95% CI)	Difference in coefficients		Product of coefficients	
				Indirect effect (95% CI)	Proportion mediation (95% CI)	Indirect effect (95% CI)	Proportion mediation (95% CI)
CVD (rare)	Education low vs high	Continuous BMI	-0.29 (-0.31, -0.26)	-0.06 (-0.06, -0.05)	19.55 (17.65, 21.44)	-0.06 (-0.07, -0.06)	22.04 (20.01, 24.07)
	Continuous education	Normal / underweight vs. obese/overweight	-0.16 (-0.17,-0.15)	-0.02 (-0.02, -0.02)	11.68 (10.53, 12.83)	-0.10 (-0.11, -0.09)	62.66 (57.21, 68.11)
	Education low vs high	Normal / underweight vs obese/overweight	-0.29 (-0.31,-0.26)	-0.03 (-0.04,-0.03)	11.45 (10.03, 12.87)	-0.17 (-0.19,-0.16)	61.20 (54.78, 67.61)
Hypertension (common)	Education low vs high	Continuous BMI	-0.19 (-0.21, -0.17)	-0.04 (-0.05,-0.04)	22.61 (19.75, 25.47)	0.05 (-0.05,-0.04)	24.62 (22.14, 27.09)
	Continuous education	Normal / underweight vs obese/overweight	-0.13 (-0.14,-0.12)	-0.02 (-0.02,-0.01)	13.77 (12.00, 15.54)	-0.08 (-0.08,-0.07)	70.05 (62.02, 78.08)
	Education low vs high	Normal / underweight vs obese/overweight	-0.19 (-0.21,-0.17)	-0.03 (-0.03,-0.02)	14.01 (12.10, 15.91)	-0.14 (-0.15,-0.12)	70.91 (61.58, 80.22)

Low education defined as 10 years or less years of education, equivalent to a highest qualification of GCSE/CSE or equivalent. High education defined as more than 10 years of education equivalent to post-secondary qualifications. Low education had a prevalence of 34%

BMI = body mass index ; CI = confidence interval

Normal or underweight defined as a BMI below 25 Kg/m^2. Overweight or obese defined as a BMI of 25Kg/m^2 of above. Normal weight or underweight had a prevalence of 34%.

Table 3.11: F statistics to test instrument strength in real-data Mendelian randomisation

	Education	BMI	LDL-C
Conditional variable			
Education	**1452.99**	2456.54	216.92
BMI	1296.39	**3093.68**	213.46
LDL-C	1406.35	1866.21	**214.63**
Education, BMI and LDL-C	1189.71	1697.33	216.92

F statistics for univariable MR analyses are in bold, all other estimates are conditional F statistics

BMI = Body mass index ; LDL-C = low-density lipoprotein cholesterol

Table 3.12: MR-Egger and MVMR-Egger results for the applied example examining the mediating role of body mass index (BMI) and low-density lipoprotein cholesterol (LDL-C) on the association between education and systolic blood pressure, cardiovascular disease and hypertension, estimated on the mean or risk difference scale

	Systolic blood pressure		**Cardiovascular disease**		**Hypertension**	
	Univariable MR-Egger	Multivariable MR-Egger*	Univariable MR-Egger	Multivariable MR-Egger*	Univariable MR-Egger	Multivariable MR-Egger*
Education						
Constant	-2.28×10^{-6}		-1.73×10^{-8}		2.13×10^{-9}	
95% confidence interval	-2.67×10^{-5} to 2.21×10^{-5}		-2.58×10^{-7} to 2.93×10^{-7}		-4.08×10^{-7} to 4.12×10^{-7}	
P Value	0.853		0.901		0.992	
Body mass index						
Constant	-1.72×10^{-6}	-1.65×10^{-6}	-2.87×10^{-8}	-2.84×10^{-8}	-1.43×10^{-8}	-1.31×10^{-8}
95% confidence interval	-2.75×10^{-5} to 2.40×10^{-5}	-1.97×10^{-5} to 1.64×10^{-5}	-2.73×10^{-7} to 3.30×10^{-7}	-1.76×10^{-7} to -2.32×10^{-7}	-4.67×10^{-7} to 4.95×10^{-7}	-3.02×10^{-7} to 3.28×10^{-7}
P Value	0.895	0.857	0.851	0.784	0.953	0.934
Low-density lipoprotein cholesterol						
Constant	-2.32×10^{-6}	-2.23×10^{-6}	-2.07×10^{-8}	-2.13×10^{-8}	-8.80×10^{-9}	-9.02×10^{-9}
95% confidence interval	-4.30×10^{-5} to 3.84×10^{-5}	-2.56×10^{-5} to 2.11×10^{-5}	-3.38×10^{-7} to 3.79×10^{-7}	-2.08×10^{-7} to -2.51×10^{-7}	-4.89×10^{-7} to 5.06×10^{-7}	-3.09×10^{-7} to 3.27×10^{-7}
P Value	0.909	0.850	0.908	0.855	0.972	0.955

Education adjusted for either body mass index or low-density lipoprotein cholesterol

MR = Mendelian randomisation

3.7 Testing the assumptions of mediation analysis

In this analysis, a number simulations were carried out to demonstrate scenarios where phenotypic or MR methods for mediation analysis may provide biased answers. In this section I outline these results and any implications for analyses.

3.7.1 Unmeasured confounding

Many of the key assumptions in phenotypic mediation analysis relate to assumptions of no unmeasured confounding between all of the exposure, mediator and outcome, including where confounders of the mediator and outcome are descendants of the exposure (intermediate confounding). Multivariable regression analyses often suffer from residual confounding because it is generally impossible to measure a sufficient set of confounders, and frequently those that are measured are measured with error.

Indeed, in simulations where residual covariance was simulated to reflect confounding, both the phenotypic difference method and phenotypic product of coefficients method were equally biased (Figure 3.6 and Appendix 1 Table 1). Where no confounding was simulated in the case of no true total effect, estimates from phenotypic approaches were free from bias (Appendix 1 Table 2). In simulations both with and without residual covariance to reflect confounding, MVMR and two-step MR estimated the direct effect, indirect effect and proportion mediated with no bias (Figure 3.6 and Appendix 1 Table 3-Appendix 1 Table 4).

Collider bias can be introduced by adjusting for the mediator in the presence of un- or mis-measured mediator-outcome confounders, where a backdoor path opens up between the exposure and the confounder (Figure 3.7) (229, 231, 232). Given that MR estimates are unbiased by unmeasured confounding of the exposure-outcome and mediator-outcome relationships (245, 279), this means that within MR analyses, adjusting for the mediator does not result in collider bias.

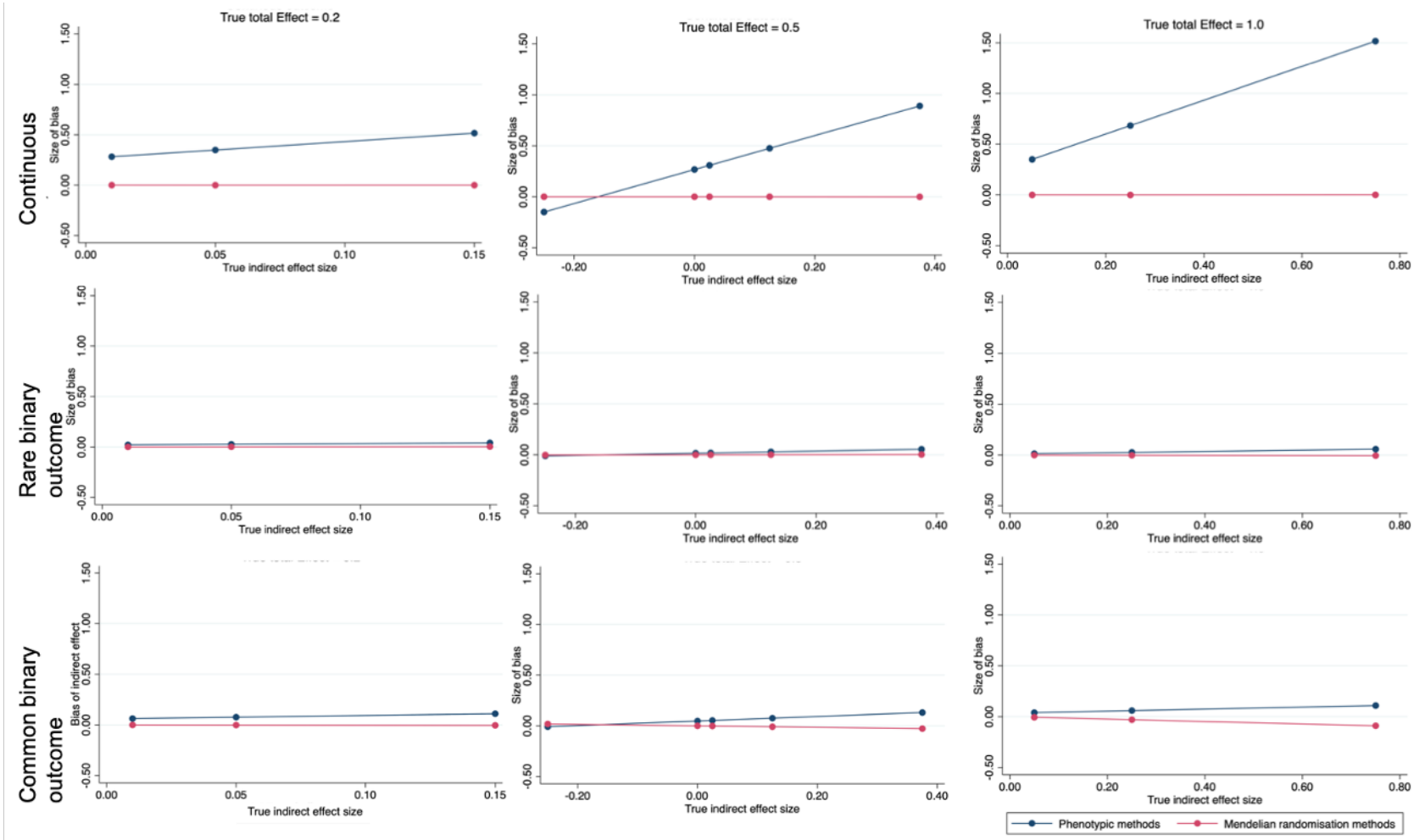

Figure 3.6: Size of absolute bias for the indirect effect of an exposure on range of outcomes through a continuous mediator, for a range of fixed true total effect sizes (0.2, 0.5 and 1.0) and range of true indirect effect sizes using phenotypic mediation methods or Mendelian randomisation, on the risk difference scale (simulated N = 5000)

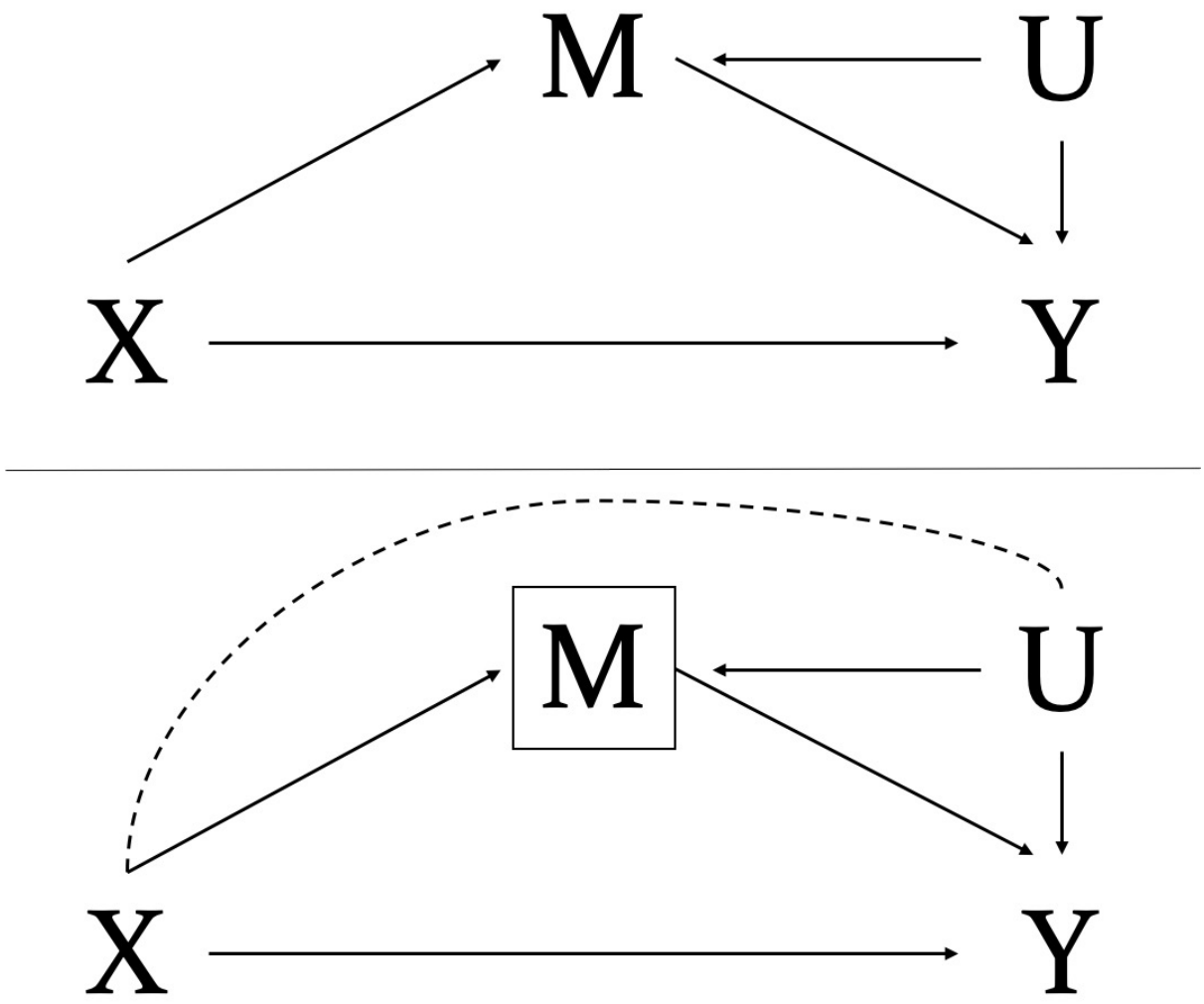

Figure 3.7: Directed acyclic graphs depicting how collider bias can be introduced in phenotypic mediation analysis when conditioning on a mediator in the presence of un- or mis- measured mediator-outcome confounders

3.7.2 Analysis of binary outcomes

Mediation analysis of binary outcome is challenging because of the non-collapsibility of odds ratios. This means the association between an exposure and outcome would not be constant on the odds-ratio scale by strata of categorical covariate (298, 299). In mediation analysis, including the mediator in the model estimating the direct effect, means the model is no longer comparable with that for the total effect.

The mediation literature indicates that to estimate the direct and indirect effects of a binary outcome, the outcome must be rare (less than 10% prevalence), so the odds ratio approximates the risk ratio, and the product of coefficients method should be used for phenotypic data (228). In the presence of a common binary outcome, estimates from the product of coefficients method and difference method are unlikely to align (and indeed the literature suggests both are likely biased) (241).

In simulations, both the difference in coefficients and the product of coefficients phenotypic methods, with common and rare binary outcomes on a linear relative scale were biased as expected (Figure 3.6 and Appendix 1 Table 5 to Appendix 1 Table 8). In simulated MR scenarios with common and rare binary outcomes on a linear relative scale, estimated effects were concordant between MVMR and two-step MR, with little to no bias (Figure 3.6 and Appendix 1 Table 9 to Appendix 1 Table 12).

In the scenarios simulated, there was some bias when analysing binary outcomes on the log odds ratio scale using both MVMR and two-step MR, for both common and rare binary outcomes (Appendix 1 Table 13 and Appendix 1 Table 14). This bias was small and typically would not alter conclusions made, although typically the size of absolute bias increased as the size of the true proportion mediated increased. However, the exact bias from non-collapsibility will be unique to each scenario, including depending on the strength of the mediators. Analyses in individual level MR can be conducted on the risk difference scale, which reduced bias due to non-collapsibility.

In simulation scenarios explored, neither MVMR nor two-step MR were able to estimate the mediated effects without bias when using the odds ratio scale (Appendix 1 Table 15 and Appendix 1 Table 16).

3.7.3 Measurement error in the exposure or mediator

These results show that in phenotypic approaches, with a continuous exposure and mediator, non-differential measurement error in the mediator leads to an underestimate of the mediated

effect. This is consistent with previous methodological and applied work (275). Where non-differential measurement error was simulated in the exposure, the mediated effect was over estimated (Appendix 1 Table 17).

In Mendelian randomisation simulations, both MVMR and two-step MR estimated the mediated effects with little bias when non-differential measurement error was simulated either in the exposure or the mediator (Appendix 1 Table 18). This is consistent with the previous literature demonstrating that MR estimates are less prone to bias by non-differential measurement error than conventional phenotypic analyses (245, 279).

3.7.4 **Weak instrument bias**

In order to obtain valid causal inference for mediation, all standard MR assumptions must be met. This includes having strong instruments, typically determined through an F-statistic or conditional F-statistic of greater than 10. When the instruments in the simulation were weakly associated with the exposure both MVMR and two-step MR estimates of the indirect effect and proportion mediated were biased. The size of bias was greatest for a common binary outcome. When weak instruments were simulated for the mediator, estimates of the indirect effect and proportion mediated from both MVMR and two-step MR were biased (Figure 3.8). Bias due to weak instruments have been discussed extensively in the literature (244, 300, 301), and some methods are now available for testing for weak instrument bias in MVMR (302).

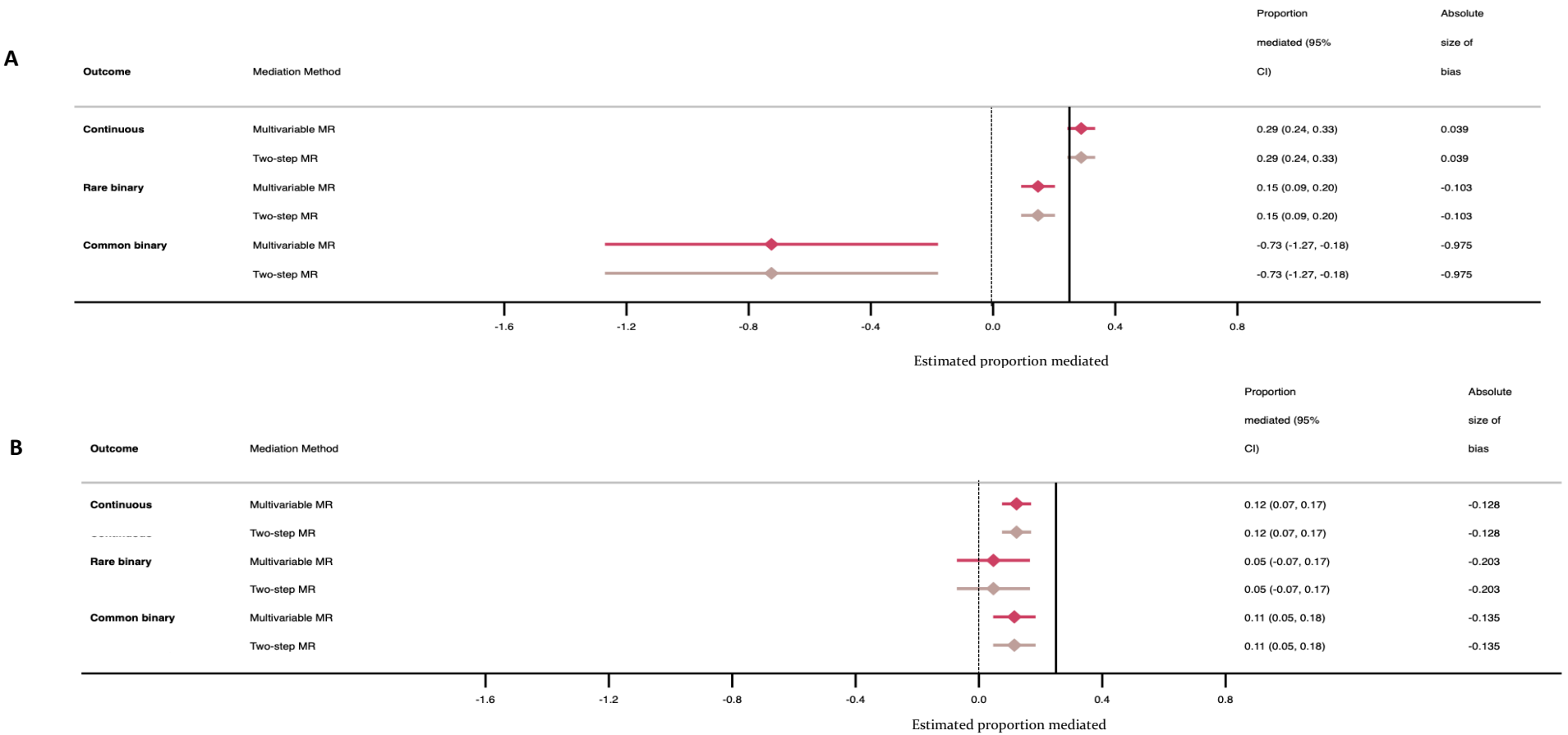

Figure 3.8: Estimates of the proportion mediated and size of absolute bias when weak instrument bias is simulated in A) the exposure and B) the mediator for a true proportion mediated of 0.25 (solid line) (simulated N = 5000)

3.7.5 **Small total effects**

In simulation studies with no true total effect the MR estimate of the proportion mediated is implausible (Appendix 1 Table 3). Where there is no evidence of a total effect, consideration should be given as to whether it is appropriate to continue with mediation analyses. Although an indirect effect can be estimated in the absence of a significant total effect, or absence of total effect when the indirect effect and direct effect act in opposing directions and cancel each other out, these estimates are prone to inflated type 1 errors (i.e. false positive results) (303).

Where the total effect is weak or estimated imprecisely (with confidence intervals crossing the null) simulations show the indirect effect and the proportion mediated using MR can be estimated but have large standard deviations (Appendix 1 Table 19 to Appendix 1 Table 22). In this case, results should be interpreted with caution, especially considering the bounds of error.

3.7.6 **Analysis of multiple mediators**

The direct effect of an exposure controlling for multiple mediators in a single model can be assessed using MVMR, with no evidence of bias (Appendix 1 Table 23). Here, non-overlapping SNPs for all exposures and mediators are included in one set of instruments. The estimated direct effect attributable to multiple mediators is unbiased, even in the presence of mediator-mediator relationships. In simulations presented here, this relationship was demonstrated by M2 causing M3 (Figure 3.4).

Where there are no mediator-mediator relationships, estimates of the indirect effects and proportion mediated from both MVMR (mutually adjusting for all mediators) and two-step MR (considering each mediator individually and summing together) will coincide (Appendix 1 Table 23). In simulations, both MR methods estimated the indirect effect of each mediator, and the three mediators jointly, with no bias (Appendix 1 Table 23). This is consistent with the existing literature on phenotypic multiple mediators (295).

Where mediator-mediator relationships are present, the indirect effect estimated via two-step MR captures both the amount of the association explained by the mediator of interest, and the amount of the mediator-outcome association captured by related mediators. In the simulated example, this means that the effect of M3 is estimated twice, once directly and once via M2. As such, the estimate for the proportion mediated summing all three mediators together will likely be an overestimate of the combined proportion mediated, but the estimated direct effect remains unbiased. In my simulations, the combined proportion mediated was over-estimated

by 6% (Appendix 1 Table 23), which is equivalent to the proportion explained by M2 through M3. The indirect effect of M2 therefore reflects both the direct effect of M2 on the outcome and the indirect effect via M3 (Figure 3.4).

3.8 Applied results in context

The results from the applied example demonstrate a causal total effect of education on systolic blood pressure, supporting results in the wider literature (304-306) and shows that BMI is a mediator of the association between education and systolic blood pressure. Given my analyses showing that systolic blood pressure is itself a mediator of the associations between education and CVD (Chapter 4), this work suggests systolic blood pressure is downstream of BMI on the causal pathway, although bi-directional associations were not explored in this analysis.

Despite LDL-C being a major, modifiable risk factor for cardiovascular outcomes (50, 307), and there being some evidence from non-genetic instrumental variable analyses that levels of LDL-C decrease with increased education (308), it does not appear to explain any of the of educational inequalities in these outcomes. Although the instrument for LDL-C only comprised of 9 SNPs which did not have pleiotropic effects on either high-density lipoprotein cholesterol or triglycerides, the F-statistics and conditional F-statistics for these analyses remained high, and results are unlikely to be biased due to weak instrument bias.

Considering BMI and LDL-C jointly, the proportion mediated between education and CVD increased by 9% compared with the BMI individually. However, for each of the individual estimates of the proportion mediated and the joint estimate the confidence intervals were wide. Considering systolic blood pressure and hypertension as the outcomes, the proportion mediated decreased.

3.9 Limitations of Mendelian randomisation applied to mediation analysis

3.9.1 Instrument selection

Instruments associated with multiple exposures can be included in a MVMR analysis when MVMR is being used to test for potential pleiotropic pathways (22, 282, 309). However, when MVMR is used to test for mediation, these overlapping instruments should not be included. If overlapping instruments were included and an attenuation of the direct effect compared with the total effect was observed, it would not be possible to distinguish whether this were attributable to mediation or pleiotropy (i.e. an effect of the SNP on the outcome via the mediator that is not due to the exposure). In a two-step MR mediation analysis, the mediator

is considered as both an exposure (of the outcome) and as an outcome (of the exposure) and therefore any instruments for the exposure that are also instruments for the mediator are pleiotropic in the estimation of the effects of the exposure on the mediator and should be excluded. Where there are no independent SNPs, or the SNPs had a perfectly proportional effect on both the exposure and the mediator, then it would not be possible to use MR methods to estimate mediation.

The exclusion restriction criteria assuming no pleiotropic pathway is an important assumption of standard univariable MR, which applies equally when MR is used for mediation analysis. Some methods are available to assess pleiotropy including for the use of MVMR (255, 256, 297).

3.9.2 Binary exposures and/or mediators

Very few binary exposures will be truly binary and are likely a dichotomization of an underlying liability, changing the interpretation of an MR analysis (310). For example, smoking is often defined as ever versus never smokers, when the underlying exposure is a latent continuous variable reflecting smoking heaviness and duration. As a result, the exclusion restriction criteria are violated, where the genetic variant can influence the outcome via the latent continuous exposure, even if the binary exposure does not change (310). In a mediation setting, the same would apply to a binary mediator. In these scenarios, two-step MR could be used to test whether there is evidence of a causal pathway between the binary exposure and/or mediator. However, the estimates of mediation would likely be biased.

3.9.3 Interactions between the exposure and mediators

Within phenotypic analysis, exposure-mediator interactions can be accommodated when estimating mediation parameters. This is not possible in either MVMR or two-step MR. Methods are available for estimating interactions in an MR framework with individual level data, but these do not currently extend to estimating mediation in the presence of exposure-mediator interactions (240, 251, 252). Estimates of mediation from MR mediation methods will require assuming effect homogeneity of both the exposure on the mediator and outcome, and mediator on the outcome. This means that the effects of the exposure and the mediator are the same for all individuals. Where interactions between the exposure and mediator are hypothesised this assumption may not hold true. Developing MR methods which can account for these interactions will be important areas of future research.

3.9.4 Power

Mendelian randomisation studies require very large sample sizes to achieve adequate statistical power. Conditional F-statistics in MVMR are typically weaker than standard F-statistics, and indeed are likely to become weaker with each additional mediator included, further decreasing the power of complex analyses. Therefore, to achieve adequate statistical power, or precision, sample sizes for mediation analysis likely need to be even larger than those needed in a univariable MR analyses.

In the absence of formal power calculators for complex MR scenarios, the power of these analyses can be considered by evaluating the precision of the confidence intervals for all of the total, direct and indirect effects, as well as assessing the conditional instrument strength.

3.9.5 Confounding

Although assumptions about unmeasured confounding in MR can be relaxed compared with traditional phenotypic analyses, confounding can be introduced through population stratification, assortative mating, and dynastic effects (186). Adjusting for genetic principal components and other explanatory variables that capture population structure or within family analyses can minimise bias.

3.9.6 Mediation analysis with summary sample Mendelian randomisation

Methods applied in this paper can be used with summary data MR (see Box 1). Similar considerations will apply for both individual level MR, as presented here, and summary data MR. Importantly, all sources of summary statistics for the exposure, mediator and outcome should be non-overlapping (289). As the mediator is considered an outcome in the exposure-mediator model, sample overlap can introduce bias (289). As individual level data is not available in summary data MR, bootstrapping cannot be used to estimate the confidence intervals for the indirect effect or proportion mediated, but the delta method can be used to approximate these confidence intervals if samples are independent (287). Analyses will also be restricted to the scale reported by the GWAS used, so consideration will need to be given for binary outcomes where sensitivity analyses to test potential non-collapsibility are limited.

Box 1: Summary of Mendelian randomisation

Individual level data Mendelian randomisation
Individual SNPs or polygenic risk scores are created for each individual in a study, where all study information and genetic information is provided for each individual.
Both the gene-exposure and gene-outcome estimates are calculated in the same sample
Analyses can be carried out on either a binary (log odds ratio) or continuous scale
The F-statistic and Sanderson-Windmeijer F-statistic can be used to assess instrument strength in univariable and multivariable MR respectively.
Summary data Mendelian randomisation
Summary estimates of the gene-exposure and gene-outcome association are estimated in separate samples
Analyses must be carried out on the scale reported by the outcome genome wide association study
Provides an opportunity to maximise statistical power by using multiple data sources
MR-Egger can be extended to investigate pleiotropy in MVMR **(297)**

3.10 Which method and when

Although MR is robust to many of the untestable causal assumptions in phenotypic mediation analysis, these are replaced with a set of MR specific causal assumptions (Figure 3.2), and careful consideration should be given to which assumptions are most plausible. Additionally, the data available, or research question of interest may not be suitable to test in an MR framework. For example, where the research question is primarily interested in time varying exposures or mediators, MR becomes increasingly complex (311). Mediation estimates from MR assume a time-fixed effect of the exposure and mediator, representing long-term relationships between the exposure and mediator (20). In some unique cases instruments may be available for an exposure at different time points (e.g. childhood and adulthood BMI), but using these instruments come with additional methodological challenges (312).

Mendelian Randomisation has specific advantages compared with phenotypic methods where causal assumptions are required. The causal effect of the exposure on the outcome, the exposure on the mediator and the mediator on the outcome can all be tested. Additionally, bi-directional MR could be used to determine which of two variables is the causal exposure and causal mediator, where this is not known.

These results demonstrate that both MVMR (akin to the difference in coefficients method) and two-step MR (akin to the product of coefficients method) can estimate the mediating

effects for both continuous and binary outcomes, with little evidence of bias. However, caution is required in some instances, for example where total effects are weak. Where all exposures, mediators and outcomes are continuous, MVMR may confer an advantage of power, where the standard deviations for the simulated effects estimated in MVMR were smaller compared with the same effects estimated using two-step MR (313).

If an analysis is interested in estimating the effects of multiple mediators, consideration should be given to the causal question of interest when deciding which method to use to analyse multiple mediators. Where the causal question specifically relates to identifying the combined effects of multiple mediators, MVMR is likely to be the most appropriate method. Where the causal question aims to estimate the effect of multiple mediators individually, and potentially any impact of intervening on a mediator, two-step MR is likely to be most appropriate. However, it is important to note, that as the number of mediators included in an MVMR model increases, the power of the analysis would likely decrease. Additionally, future research should be carried out to determine if including increasing numbers of exposures in an MVMR model further violates any of the MR assumptions.

Although a range of simulation scenarios were included, including both continuous and binary outcomes, this is not an exhaustive range of scenarios and there may be further scenarios where MR methods are biased.

The flow chart in Figure 3.9 aims to help with the decision-making process, based on practical limitations of MR. Key recommendations for these analyses are reviewed in Box 2. However, best practice would always be to triangulate across phenotypic and genetic approaches, and across multiple data sources wherever possible (259).

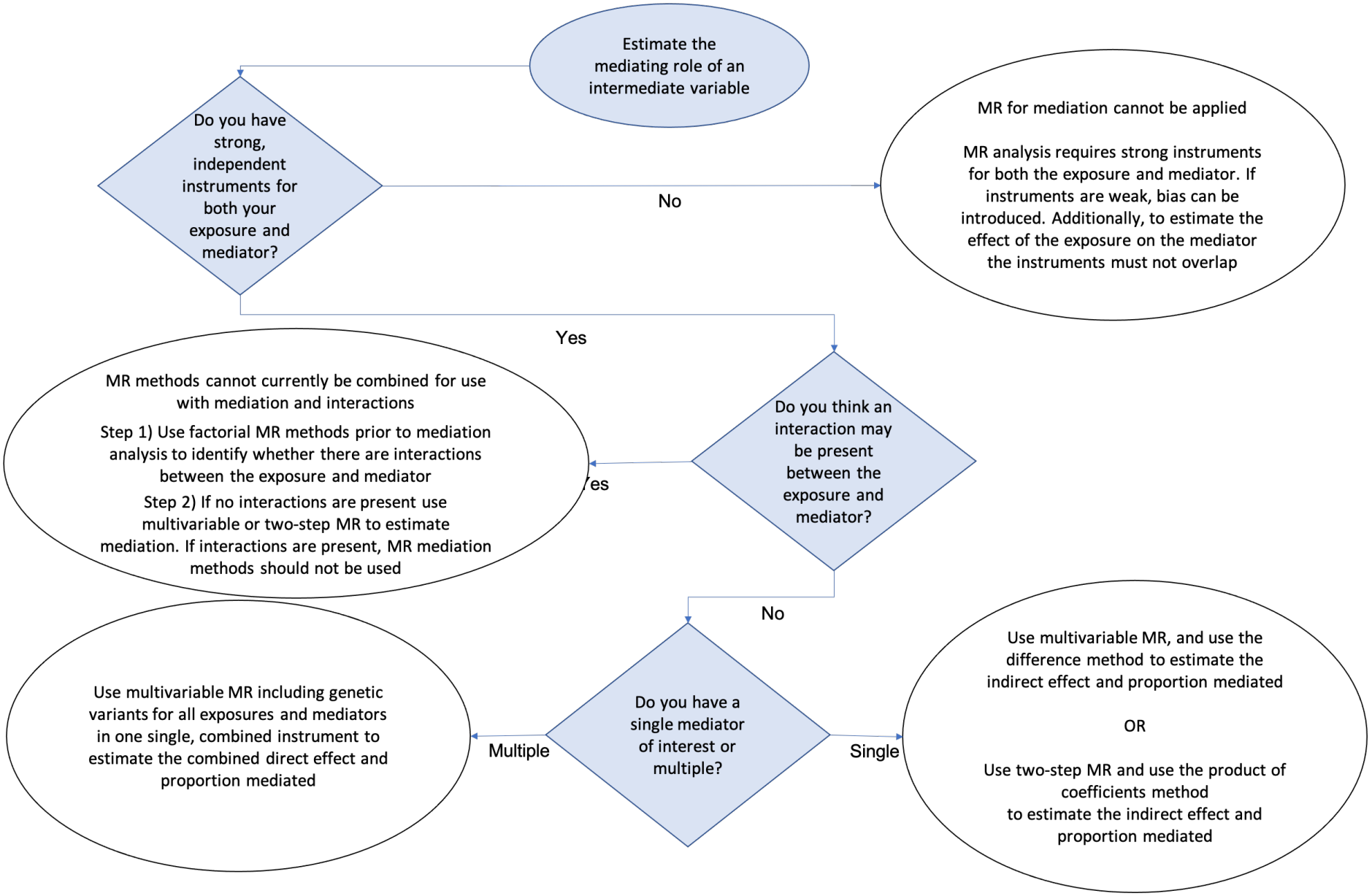

Figure 3.9: Decision flow chart to determine most appropriate mediation method

MR = Mendelian randomisation

Box 2: Key recommendations when using Mendelian randomisation for mediation analysis

- Ensure strong instruments are available for exposures and mediators and test instrument strength using the F-statistic. Test the conditional instrument strength for multivariable MR using the Sanderson-Windmeijer F-statistic (296)
- Instruments for the exposure and mediator must be independent for both multivariable MR and two-step MR methods
- The instruments must not have a pleiotropic effect on the mediator or outcome
- Current MR methods are optimised for use with continuous exposures and mediators. Binary exposures or mediators which are a reflection of a true underlying continuous measure can lead to violation of the exclusion restriction criteria
- Use univariable MR to test for evidence of causal association in each step of the mediation path, from the exposure to the outcome, exposure to the mediator and mediator to the outcome
- Where individual-level data are being used and outcomes are binary, estimate effects on a linear scale to alleviate potential bias from non-collapsibility of odds ratios
- If using summary level data with a binary outcome, estimate effects on the log odds ratio scale and transform after analysis if odds ratios are required

3.11 Conclusions

Mendelian randomisation can be extended to estimate direct effects, indirect effects and proportions mediated. MR estimates are robust to violations of the often-untestable assumptions of phenotypic mediation analysis, including unmeasured confounding, reverse causality and measurement error. MR analysis makes its own strong, but distinct assumptions, especially relating to instrument validity. To estimate mediation using MR, large sample sizes are required, and strong instruments are needed for both the exposure and mediator.

Chapter 4. Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study.

4.1 Publication details

Carter AR, Gill D, Davies NM, Taylor AE, Tillmann T, Vaucher J, *et al*. Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study. BMJ. 2019;365:l1855.

4.2 Author list and contributions

Alice R Carter [1,2]*, Dipender Gill [3]*, Neil M Davies [1,2], Amy E Taylor [2,4], Taavi Tillmann [5], Julien Vaucher [6,7], Robyn E Wootton [1,8], Marcus R Munafò [1,8,9], Gibran Hemani [1,2], Rainer Malik [10], Sudha Seshadri [11-14], Daniel Woo [15], Stephen Burgess [16, 17, 1], George Davey Smith [1,2], Michael V Holmes [18-20], Ioanna Tzoulaki [3,21,224], Laura D Howe [1,2]† & Abbas Dehghan [3,21]†

All affiliations are presented in Appendix 2.

ARC and DG contributed equally to this project and are joint first authors. AD and LDH contributed equally to this project and are joint senior authors. ARC and DG devised the project, analysed and cleaned the data, interpreted results, wrote and revised the manuscript. ARC primarily carried out analyses using the UK Biobank. DG primarily carried out two-sample mendelian randomisation analyses. AET, NMD, TT, JV, SB, GDS, MVH, IT, LDH, and AD devised the project, interpreted the results, and revised the manuscript. TT, JV, GH, SB, and GDS contributed to the design of the project and critically revised the manuscript. REW, MRM, RM, SS, and DW provided data, and critically reviewed and revised the manuscript. All authors had full access to the data in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis. ARC and DG are the guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

4.3 Summary of personal contributions

This chapter uses phenotypic mediation methods and Mendelian randomisation mediation methods, established in the previous chapter (Chapter 3), to explore the role of body mass index,

systolic blood pressure and lifetime smoking behaviour in mediating the association between educational attainment and cardiovascular outcomes. Although phenotypic mediation analyses have previously been carried out to identify downstream factors from education which may be responsible for cardiovascular outcomes, as previously discussed, these methods can be prone to a number of biases, particularly from unmeasured confounding, measurement error and reverse causality. This chapter aimed to use Mendelian randomisation for causal mediation analysis.

The analyses in this chapter were carried out collaboratively with Dr. Dipender Gill, based at Imperial College London, with support from a number of other researchers across institutions within the UK. A version of this manuscript has been published in the British Medical Journal (https://doi.org/10.1136/bmj.l1855).

My role in this work was to act as the joint first author with Dr. Gill and as the corresponding author with the journal. I was responsible for carrying out all UK Biobank analyses, including analyses using phenotypic mediation methods and individual level Mendelian randomisation analyses. Additionally, I was responsible for creating journal quality figures and tables. Dr. Gill was responsible for carrying out summary data Mendelian randomisation using summary statistics from genome-wide association study consortia. Jointly, Dr. Gill and I drafted the manuscript, which was advised and informed by comments from all co-authors.

Due to word limits in the journal all authors agreed to present the phenotypic estimates from UK Biobank and the summary data Mendelian randomisation analyses in the main manuscript, whilst presenting the results of the individual level Mendelian randomisation analyses in the supplementary material. Although both Mendelian randomisation methods indicated the same conclusions, the results from summary data Mendelian randomisation were estimated with greater precision, due to the larger sample sizes included in the analyses. For this thesis, I have included my contributions (the phenotypic analyses and individual level Mendelian Randomisation analyses) in the main chapter and reduced the emphasis on the summary data Mendelian randomisation analyses and results.

Full contributions from myself include devising the project, establishing collaborations with other researchers, writing and circulating the analysis plan, cleaning the UK Biobank data, analysis and interpreting the results, writing and drafting the manuscript, submitting the manuscript, responding to and revising according to peer review comments.

4.4 Abstract

Background:

Studies have demonstrated causal effects of educational attainment on cardiovascular disease (CVD). We aimed to investigate the role of body mass index, systolic blood pressure and smoking in explaining the effect of education on risk of CVD outcomes triangulating across multivariable regression analysis of observational data and one- and two-sample Mendelian randomisation (MR) analysis; an instrumental variable approach more robust to bias from confounding and reverse causation.

Methods:

Individual level data from UK Biobank (N = 217 013) was used for multivariable analyses and individual level Mendelian randomisation. Summary statistics from genome-wide association studies were used in summary data MR.

The total effect of education on risk of coronary heart disease, CVD (all subtypes), myocardial infarction and stroke (all measured in odds ratio, OR) was assessed using multivariable regression and univariable MR.

The degree to which this effect is mediated through body mass index, systolic blood pressure and smoking respectively (the indirect effect and proportion mediated) was estimated using the product of coefficients method, where the effect of education on each mediator, and each mediator on each outcome was assessed using multivariable regression and two-step MR. The joint contribution of all three risk factors was assessed via the difference method, using multivariable regression or multivariable MR.

Results:

Each additional standard deviation of education associated with 13% lower risk of coronary heart disease (OR 0.87, 95% confidence interval [CI] 0.84 to 0.89) in observational analysis and 37% lower risk (OR 0.63, 95% CI 0.60 to 0.67) in MR analysis. As a proportion of the total risk reduction, body mass index mediated 15% (95% CI 13% to 17%) and 18% (95% CI 14% to 23%) in the observational and MR estimates respectively. Corresponding estimates for systolic blood pressure were 11% (95% CI 9% to 13%) and 21% (95% CI 15% to 27%), and for smoking, 19% (15% to 22%) and 34% (95% CI 17% to 50%). All three risk factors combined mediated 42% (95% CI 36% to 48%) and 36% (95 % CI 16% to 63%) of the effect of education on coronary heart disease in

observational and MR respectively. Similar results were obtained for risk of stroke, myocardial infarction and all-cause CVD.

Conclusions:

Body mass index, systolic blood pressure and smoking mediate a substantial proportion of the protective effect of education on risk of cardiovascular outcomes. Intervening on these would reduce cases of CVD attributable to lower education. However, more than half of the protective effect of education remains unexplained.

4.5 Introduction

Cardiovascular disease (CVD) is the leading cause of mortality worldwide, accounting for over 17 million deaths annually (1). Recent studies have suggested that socioeconomic risk factors such as education play a causal role in the aetiology of CVD (8, 9, 126). Tillmann and colleagues found that an additional 3.6 years of education reduced the risk of coronary heart disease by approximately one third (9). However, educational opportunities are not equitable throughout populations and education is inherently difficult to intervene on. Therefore, understanding the risk factors that may be driving the adverse later life outcomes associated with lower levels of education would provide the opportunity for interventions to reduce inequalities.

Existing studies suggest that body mass index (BMI), systolic blood pressure and smoking behaviour at least partly explain differences in CVD risk related to educational attainment (12-14). However, these studies have relied on phenotypic mediation analyses, that may suffer from bias. Additionally, many phenotypic mediation methods use a single snapshot of a risk factor, which may incompletely capture a person's lifetime exposure (275). For example, blood pressure measured at a single time point will suffer from measurement error due to day-to-day fluctuations and will not capture changes across the life course. This measurement error can lead to an underestimation of mediation (275). Furthermore, other biases such as unmeasured confounding cannot be addressed using phenotypic methods (229).

Mendelian randomisation (MR) uses genetic variants as instrumental variables (IVs) to estimate the effect of an exposure on an outcome of interest (18). During meiosis, genetic variants are randomly allocated from parents to offspring, which remain fixed from the point of conception and are not altered during the life course. This random allocation of genetic variants can be exploited to infer causal effects that are potentially robust to non-differential measurement error and confounding of the exposure-outcome relationship (18). Two-step MR for mediation analysis, unlike phenotypic mediation analysis approaches, can be used to estimate the causal effects of the mediator, even if the phenotypes are measured with error (19). Recent genome-wide association study (GWAS) meta-analyses have identified a number of genetic variants for educational attainment and the other mediators of interest that may be used as IVs (17, 149).

Mendelian randomisation has previously been used to demonstrate the causal effects of education on BMI, systolic blood pressure and smoking and also the effects of BMI and smoking on CVD (45, 59, 118, 119, 135, 268). While the results from these studies suggest that BMI, systolic blood

pressure and smoking are likely to explain some of the protective mechanisms of education on CVD, they alone do not quantify the mediated effect. In this study, I investigated the role of BMI, systolic blood pressure and lifetime smoking in mediating the causal effect of educational attainment on CVD risk using three complementary approaches: multivariable regression, individual level MR and summary data MR. BMI, systolic blood pressure and smoking were selected as intermediate risk factors based on previous literature implicating them as both being affected by education and as risk factors for CVD, with availability of data across all three complementary methods. I consider the three risk factors both individually and simultaneously. Understanding the mechanisms by which education affects cardiovascular health could have powerful applications, such as for public health policy. For this, it is important to understand the population-level implications of changes to BMI, smoking and systolic blood pressure on inequalities in CVD risk.

4.6 Methods

4.6.1 Overall study design

This study used multivariable regression of phenotypic data, one-sample MR of individual level genetic data and two-sample MR of summary level genetic data to investigate whether lower BMI, systolic blood pressure and lifetime smoking explain the protective effect of education on risk coronary heart disease (CHD) myocardial infarction (MI), stroke risk and CVD (all subtypes).

4.6.2 UK Biobank

UK Biobank recruited 503,317 UK adults between 2006 and 2010. Participants attended assessment centres involving questionnaires, interviews, anthropometric, physical and genetic measurements (15, 16). In the phenotypic analysis, 217 013 White British individuals, with complete data on genotypes, age, sex, educational attainment, cardiovascular outcomes, BMI, smoking status, blood pressure, socioeconomic status (as measured by Townsend Deprivation Index at birth [TDI]) and place of birth were included.

4.6.2.1 Exclusion criteria

Individuals were excluded if their genetic sex differed to their gender reported at the assessment centre or for having aneuploidy of their sex chromosomes. Further individuals were excluded for being outliers for their heterozygosity and any missing genetic data. Related individuals were also excluded from analyses, and the remaining subset was a maximal set of unrelated individuals.

This exclusion list was derived in-house using an algorithm applied to the list of all the related pairs provided by UK Biobank (3rd degree or closer). It preferentially removes the individuals related to the greatest number of other individuals until no related pairs remain (288). Individuals of White British descent were defined using both self-reported questionnaire data and similar genetic ancestry to the European ancestry principal components (PCs) computed from the 1000 genomes project (288). Available follow-up data were used where baseline data were missing. For the sample used in mediation analyses, individuals were excluded if there were missing data at baseline and no available follow up data for education, BMI, systolic blood pressure, smoking, CVD status or for any of the variables considered as confounders. Figure 4.1 illustrates the exclusion criteria and number of individuals excluded at each stage for analyses in UK Biobank, demonstrating how the final sample size of 217 013 individuals was achieved.

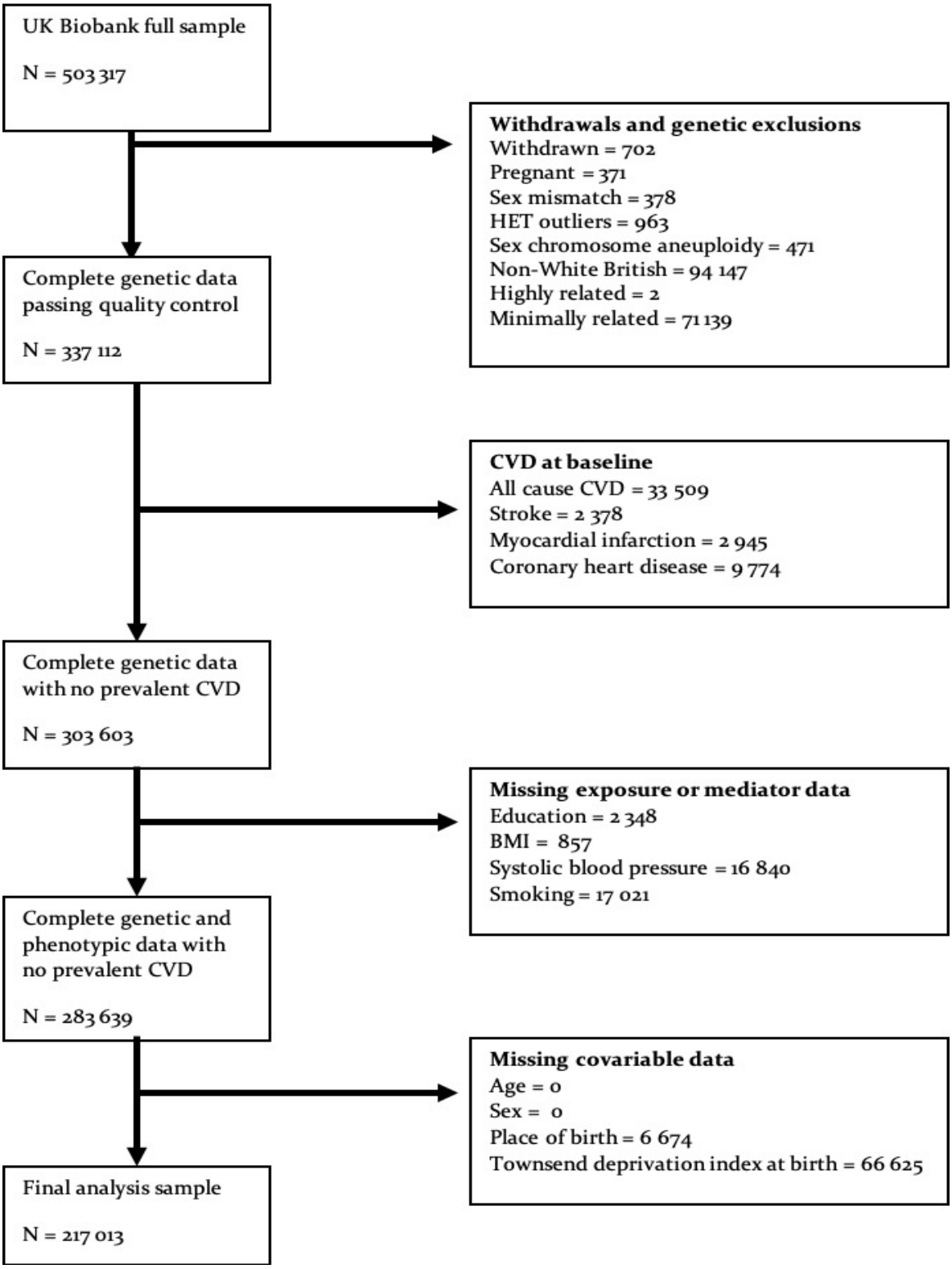

Figure 4.1: Flow chart illustrating exclusions made in UK Biobank for the analysis sample for mediation analyses

Note: At each stage the same participant could have missing data for multiple variables, therefore overlap is present between the variables. The total excluded may be less than the sum of individuals at each stage.

CVD = cardiovascular disease; BMI = body mass index

4.6.2.2 Educational Attainment

Participants reported their highest qualification and age of leaving school if they did not have a degree. These were converted to the International Standard Classification for Education (ISCED) coding of educational attainment (Table 4.1)(149).

For the individual level MR analysis, instruments were selected from analysis of populations that did not overlap with those considered in the outcome estimates. Accordingly, 74 independent single-nucleotide polymorphisms (SNPs) that attained genome-wide significance ($P<5x10^{-8}$) for education reported in main results from the 2016 SSGAC GWAS meta-analysis of 293,723 individuals that did not include UK Biobank participants were used, to create a polygenic score (17). Five instruments for education were not available in UK Biobank and proxy synonymous SNPs in perfect LD ($r^2=1$) were used (Table 4.2). The SNPs were clumped for linkage disequilibrium with an r^2 threshold 0.001 and within a distance of 10 000kB.

Table 4.1: International Standard for Classification of Education codes mapped to UK Biobank self-report highest qualification to estimate years of education

Qualification (As reported in UK Biobank)	ISCED	Years of education	N
College or University degree	5	20	69 935
NVQ or HND or HNC or equivalent	5	19	14 017
Other prof. qual. e.g.: nursing, teaching	4	15	10 986
A levels/AS levels or equivalent	3	13	25 590
O levels/GCSEs or equivalent	2	10	49 349
CSEs or equivalent	2	10	12 288
None of the above	1	7	34 849
Prefer not to answer	Excluded		

Table 4.2: Proxy single nucleotide polymorphisms for educational attainment instrument used in individual level Mendelian Randomisation analyses

GWAS SNP (Okbay)	SNP in LD used (UKBB)
rs114598875	rs17538393
rs148734725	rs9878943
rs9320913	rs1487445
rs8005528	rs8008779
rs192818565	rs55943044

4.6.2.3 Body mass index

Measures of height and weight taken by UK Biobank study nurses at baseline assessment centres were used to calculate BMI (kg/m^2).

In individual level MR analysis, 77 SNPs which had attained genome-wide significance ($P<5x10^{-8}$) for BMI in the Genetic Investigation of ANthropometric Traits (GIANT) Consortium genome-wide association study (GWAS) analysis of individuals with European ancestry were used as instruments (290).The SNPs were clumped for linkage disequilibrium with an r^2 threshold 0.001 and within a distance of 10 000kB. Alleles were harmonised to all reflect BMI increasing SNPs and individual variants were recoded as 0, 1 or 2 according to the number of BMI increasing alleles. A genetic score for BMI was created by weighting each SNP by its relative effect size in the GWAS and summing all variants together in an additive model.

4.6.2.4 Systolic blood pressure

Systolic and diastolic blood pressure were recorded both automatically and manually at the baseline assessment centre for all participants. Each reading was taken twice, two minutes apart. This analysis uses the second reading of the automated blood pressure, where missing data were replaced with the first measure or any follow up assessment centre measures.

Participants were required to take all medication they are currently using to the assessment centre, details of which were recorded by nurses. A variable for antihypertensive use was generated based on the treatments recorded and 10mmHg was added to systolic blood pressure measurements for these individuals, consistent with previous studies to account for treatment effects (314).

Mendelian randomisation studies require the SNP-exposure and the SNP-outcome associations to be estimated in independent samples, otherwise estimates can be overestimated (244, 289). Existing systolic blood pressure and lifetime smoking GWASs have been estimated using UK Biobank data (315-317). To avoid participant overlap for exposure and outcome genetic estimates in the UK Biobank (289), a split sample GWASs of systolic blood pressure and smoking respectively were performed using the University of Bristol MRC Integrative Epidemiology Unit GWAS Pipeline (318). A total of 318,147 unrelated UK Biobank participants were eligible for inclusion in the GWAS (

Figure 4.2). All the eligible participants were randomly allocated into one of two halves (sample 1 and sample 2). A GWAS was performed on both samples 1 and 2 separately, adjusted for age, sex and the first 40 principle components in UK Biobank. BOLT-LMM method was used to account for population stratification. The top hit SNPs were determined using the 'clump_data' command in the Summary data MR R package ($r^2 > 0.001$, distance >10,000kb) (default settings of the 'clump_data' command) (319). This process was carried out for both systolic blood pressure and lifetime smoking phenotypes.

The genetic score was created for each sample independently, by weighting each SNP by its relative effect size from the GWAS results of the opposing sample (i.e. the genome-wide significant SNPs and betas identified in the GWAS of sample 1 were used to generate the genetic score in sample 2 individuals). All genetic variants were summed together in an additive model.

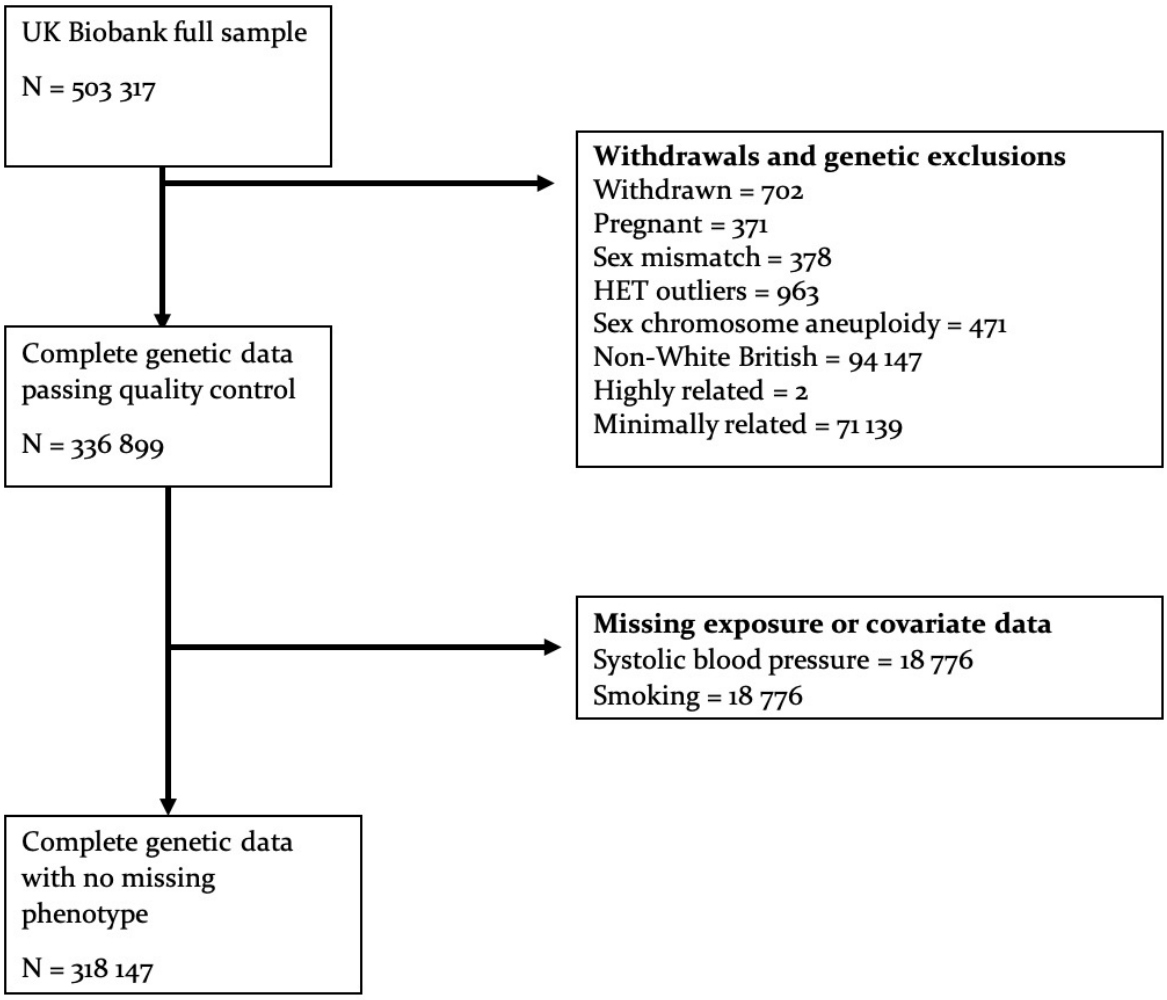

Figure 4.2: Flow chart for exclusions made in UK biobank for use in systolic blood pressure and smoking GWAS analyses

4.6.2.5 Smoking

A measure of lifetime smoking was constructed in the UK Biobank from self-reported age at initiation, age at cessation and cigarettes per day. From this information, smoking duration and time since cessation were calculated. The lifetime smoking measure further includes a simulated constant (half-life) which captures the exponentially decreasing effect of cigarettes on health over time. Aspects of smoking behaviour were combined into one score ranging from 0 (for non-smokers) to 4.17. The mean lifetime smoking score was 0.35 (standard deviation = 0.69). Full details of score construction have been published previously (320). As described previously for systolic blood pressure, a split sample GWAS was carried out using the UK Biobank GWAS pipeline hosted by the MRC-IEU to identify suitable instruments from non-overlapping samples in UK Biobank.

4.6.2.6 Covariates

Variables considered as covariates were measured at the baseline assessment centres through interviews. Sex and ethnicity were confirmed according to genetic data. Place of Birth was adjusted for by the northing and easting birth location coordinates. Although the Townsend Deprivation Index (TDI) of historic birth locations are not recorded in UK Biobank, this has been estimated from the index of multiple deprivation indices using the current TDI of birth location as a proxy for historic birthplace TDI. Mendelian randomisation models were also adjusted for the same confounders. Although a core assumption of MR is that the genetic variants are unrelated to confounders, there is some evidence of small associations with place of birth for the educational attainment variants in UK Biobank (8, 185). MR models were additionally adjusted for the first 10 genetic PCs, as derived by UK Biobank, to help control for population stratification. These were only considered in phenotypic and individual level MR analyses, where individual level data were available.

4.6.2.7 Cardiovascular disease outcomes

Cardiovascular diagnoses (including diagnoses of stroke, MI and CHD) and events were ascertained through linkage to mortality data and hospital inpatient records, with cases defined according to ICD-9 and ICD-10 codes (Table 4.3) (292). Individuals who had experienced a CVD event prior to the baseline assessment (prevalent cases) were excluded and only first event, incident cases following the assessment centre were considered). Hospital inpatient records were available from 1997 in England, 1998 in Wales and 1981 in Scotland (293), with the most recent entry recorded in this analysis in February 2016.

Table 4.3: ICD 9 and ICD 10 codes used to identify incident cases of cardiovascular disease and cardiovascular subtypes from hospital inpatient records in UK Biobank

Outcome	ICD-9 code	ICD-10 code
CVD (all subtypes)	390-459	I, G45
Stroke	434.91	I6, G45
MI	410.9, 412.9	I21, I22
CHD	410-414	I20-I25

4.6.3 GWAS meta-analyses used for summary data Mendelian randomisation

In the summary data MR analysis, summary genetic associations from GWAS data for each respective phenotype were obtained. For education, this was the Social Science Genetic

Association Consortium (SSGAC) GWAS meta-analysis of years of schooling in 1,131,881 individuals of European ancestry (149), with summary data made available for 766,345 of these participants. Instruments were selected as the 1,271 independent genome-wide significant SNPs (pairwise r^2<0.1) from the full discovery sample (149). Genetic estimates for BMI were obtained from the GIANT consortium's 2018 GWAS meta-analysis of 681,275 individuals of European decent (61). Genetic association estimates for systolic blood pressure and smoking were estimated from a GWAS of 318,417 White British individuals in UK Biobank. Instruments for BMI, systolic blood pressure and smoking were identified as the lead SNPs in loci reaching genome-wide significance after clumping summary estimates from the largest available GWAS for linkage disequilibrium (LD) threshold r^2<0.001 and distance >10,000kb, using a 1000 genomes European reference panel through the TwoSampleMR package (default settings of the 'clump_data' command) in the statistical software R (321). For CHD, publicly available genetic association estimates from the CARDIoGRAMplusC4D 1000 Genomes-based GWAS meta-analysis of 60,801 cases and 123,504 controls were used (322). The definition for CHD was broad and inclusive, considering acute coronary syndrome, myocardial infarction, angina with one or angiographic stenoses of greater than 50%, and chronic stable angina. A summary of all phenotypes and GWAS data used are presented in Table 4.4.

Table 4.4: Summary of phenotypes and GWAS data used as instrumental variables across analyses

	Multivariable phenotypic analysis (all in UK Biobank)	**Individual level Mendelian Randomisation**	**Summary data Mendelian Randomisation**
Educational attainment	Self-reported highest qualification mapped to ISCED years of schooling	Polygenic score, using genome-wide significance SNPs (N_{SNPs} = 74) and beta weights from Okbay *et al,* 2016 (323)	Individual SNPs from Lee *et al*, 2018 (149) (N_{SNPs} = 1271)
Body mass index	Measured weight and height	Polygenic score, using genome-wide significance SNPs (N_{SNPs} = 77) and beta weights from Locke *et al,* 2015 (290)	Individual SNPs from Yengo *et al*, 2018 (61) (N_{SNPs} = 360)
Systolic Blood pressure	Median of two automated blood pressure measurements	Polygenic score, using genome-wide significance SNPs (N_{SNPs} = 65 and 55 sample 1 and 2 respectively) from a split sample GWAS in UK Biobank	Individual SNPs from systolic blood pressure GWAS carried out as part of this work on full UK Biobank sample (N_{SNPs} = 191)
Smoking	Estimate of lifetime smoking using self-report data on smoking behaviours	Polygenic score, using genome-wide significance SNPs (N_{SNPs} = 18 and 15 sample 1 and 2 respectively) from a split sample GWAS in UK Biobank (317)	Individual SNPs from Wootton *et al*, 2018 using full UK Biobank sample (317) (N_{SNPs} = 126)

4.6.4 Statistical Analysis

4.6.4.1 Effect of education on cardiovascular disease

In phenotypic analyses of UK Biobank data, multivariable logistic regression was used to estimate the association of education with CVD and its subtypes. All analyses using UK Biobank were adjusted for potential confounders; age, sex, place of birth, birth distance from London, and TDI at birth. These confounders were determined *a priori*, with place of birth and birth distance from London included to control for population structure in UK Biobank (8, 185).

In the individual level MR of UK Biobank data, the total effect of education on cardiovascular outcomes was investigated using two-stage least squares regression. In the first regression, the effect of the education polygenic score on self-reported educational attainment was estimated. This estimate was used to generate a prediction of educational attainment. In the second stage,

the effect of predicted educational attainment on the CVD outcome using robust standard errors in a logistic model was estimated (324). Both regression stages were adjusted for adjusted for age, sex, place of birth, birth distance from London, and TDI as well as the first ten genetic PCs.

In summary data MR analysis, the effects of education on CVD subtypes were investigated using ratio method MR with standard errors derived using the delta method (325). Fixed-effect inverse-variance weighted (IVW) meta-analysis was used to pool MR estimates across individual SNPs (326).

4.6.4.2 Mediation by body mass index, systolic blood pressure and smoking

In multivariable phenotypic and individual level MR analyses, when investigating the degree to which the effects of education on CVD and its subtypes are mediated through each risk factor (BMI, systolic blood pressure and smoking) individually, the product of coefficients method was used to estimate the indirect effect (i.e. the effect of education on CVD that goes through the risk factor) (20).

In the phenotypic analysis, multivariable linear regression was used to estimate the association of education with each risk factor after adjusting for confounders (as in the total effects models). The effect of each risk factor on the individual CVD subtypes was then estimated using multivariable logistic regression with the additional adjustment for self-reported educational attainment (241). The two estimates were multiplied together to estimate the indirect effect (of education, through the risk factor).

In individual level MR analyses two-stage least squares regression using the Stata IVREG2 package was used to estimate the effect of education on each mediator individually.

Two-stage least squares multivariable Mendelian randomisation (MVMR) was then used to estimate the effect of each mediator on each outcome, additionally controlling for the polygenic score for education. The second stage of this regression was estimated on the log odds ratio scale. This additionally provided an estimate of the direct effect of education on each outcome. All analyses were adjusted for covariates and PCs as above.

These estimate of i) education on the mediator and ii) the mediator on the cardiovascular outcome controlling for education, were then multiplied to estimate the indirect effect, which is the amount of the association between education and CVD going via each of the three risk factors individually.

Where split sample GWAS estimates were used to create the allele score in systolic blood pressure and smoking the MR analyses were run separately for each 50% sample and meta-analysed to estimate an overall effect.

For the summary data MR, the IVW MR approach was used to estimate the effect of education on each risk factor and regression-based multivariable MR was used to estimate the effect of each risk factor on risk of the considered CVD subtypes, adjusting for genetic effect of the instruments on education (255). The indirect effect of education on risk of each CVD subtype through the considered risk factor was estimated by multiplying results from these two MR analyses.

4.6.4.3 Investigating all three risk factors combined

When investigating the role of all three risk factors together on the association between education and CVD, the difference method of estimating the indirect effect was used (241). This involved estimating the total effect of education on each CVD subtype, as described in section 4.6.4.2. The direct effect of education on each CVD subtype controlling for all three risk factors together was estimated, using either multivariable regression or multivariable MR, in phenotypic and MR analyses respectively. To estimate the total effect of education mediated indirectly through all three risk factors collectively using summary data MR, the direct effect of education after adjusting for the three risk factors together was estimated using MVMR, with this estimate divided by the total effect and then subtracted from one. In phenotypic analyses, a multivariable logistic model for the effect of education on CVD (and subtypes) adjusting for all three risk factors was used to estimate the direct effect of education independently of the risk factors. This was subtracted from the total effect to estimate the indirect effect of education through the three risk factors collectively. In individual level MR, the direct effect of education after adjusting for the three risk factors together was estimated using MVMR. This was subtracted from the total effect to estimate the indirect effect. The indirect effect was subsequently divided by the total effect to estimate the proportion mediated. Confidence intervals for the indirect effect and proportion mediated were estimated using bootstrapping.

4.6.4.4 Sensitivity analyses

A range of MR sensitivity analyses were carried out. Mendelian randomisation estimates are prone to bias if the underlying assumptions of the analysis are violated. Horizontal pleiotropy, where a genetic variant is associated to the outcome of interest via an alternative pathway, can potentially bias the MR estimates (254). MR-Egger allows for directional (unbalanced) horizontal pleiotropy

under the assumption that the size of the variants on the exposure are independent of the size of the direct effects on the outcome (i.e. there is no dose-response confounding) (253). The weighted median estimator is able to provide robust MR estimates when more than half of the information for the analysis comes from valid instruments (256). In the MR analysis of the total effect of education on CVD outcome risk, and the effect of education on each risk factor, we also performed these techniques to investigate the robustness of our findings when relaxing assumptions on horizontal pleiotropy. These pleiotropy robust techniques are not yet developed for application in MR mediation analysis.

For all analyses in UK Biobank, models were replicated on the risk difference scale using multivariable linear regression to assess whether the mediation estimates were biased by the non-collapsibility of odds ratios. For the individual level MR analyses, the IVREG2 Stata package was used for this (327). Additionally, to test for sex differences and age differences, all analyses were replicated using unadjusted models, models adjusted for age and sex only, and models stratified by sex and age dichotomized at the median (39-57 years compared with 58-72 years). On a subsample of UK Biobank participants with dietary recall questionnaires (including protein, carbohydrate, total fat, saturated fat, polyunsaturated fat, total sugar and fibre consumption) and exercise (weekly duration of moderate and vigorous physical activity) measures (N = 20,298 with dietary recall measures), a phenotypic multivariable multiple mediator model was analysed. This could not be completed using MR analyses as there are not suitable instruments for diet and exercise phenotypes. This analysis, and those stratified by age and sex, were carried out for the association between education and CVD (all subtypes) only, due to limited outcome events.

4.6.5 **Statistical software and ethical approval**

Analysis was performed using Stata version 14 (StataCorp LP) and R version 3.4.3 (The R Foundation for Statistical Computing). The mrrobust package for Stata and the TwoSample MR package for R were used to facilitate MR analyses (321, 328). Ethical approval was not sought for publicly available data because all participating studies had already obtained relevant authorisation. Project approval was obtained from UK Biobank (study ID: 10953) and data will be returned to them for archiving. Analysis code for one-sample MR analyses are available from https://github.com/alicerosecarter/EducationMediators.

4.6.6 Patient and public involvement

Neither patients nor the public were involved in the initial design or implementation of this study. Feedback from a lay reviewer was incorporated in the revision stages.

4.7 Results

4.7.1 UK Biobank Cohort Description

The UK Biobank sample used in the phenotypic and individual level MR analysis was comparable to the participants in UK Biobank as a whole, although UK Biobank is not representative of the wider UK population (participants are typically more educated and of a higher socioeconomic status as compared to the general population) (16). In the analysis sample, 32% of individuals had over 20 years of education, equivalent to a vocational qualification or degree. Comparatively, only 16% of individuals left school with no formal qualifications after seven years (Table 4.1). The standard deviation (SD) of educational attainment was 3.6 years, BMI was 4.69 kg/m^2 and systolic blood pressure was 18.68 mm Hg. For lifetime smoking, one SD increase is equivalent to, for example, an individual smoking 20 cigarettes a day for 15 years and stopping 17 years ago, or an individual smoking 60 cigarettes a day for 13 years and stopping 22 years ago (317). A total of 65 (R^2 = 0.0035) and 55 (R^2 = 0.0027) genome-wide significant SNPs were identified for systolic blood pressure (with 10mm Hg added for antihypertensive use) for sample 1 and sample 2 respectively (Appendix 2 Table 1). In the split-sample GWAS for smoking, 18 (R^2 = 0.0012) and 15 (R^2 = 0.0014) genome-wide significant SNPs were identified in sample 1 and sample 2 respectively (Appendix 2 Table 2).

Table 4.5: Cohort Characteristics for the UK biobank analysis sample used in phenotypic analyses and individual level MR analyses and comparisons with the full UK Biobank cohort

Variable	**Level**	**N Analysis Sample (N = 217 013)**	**% Analysis Sample**		**N Full UKBB (N = 502 240)**	**% Full UKBB**
Sex	Female	119 198	54.93		273 076	54.37
Age	<40	2 260	1.04		5 424	1.08
	41-50	54 234	24.99		126 426	25.15
	51-60	77 071	35.51		177 264	35.27
	61-70	83 444	38.45		193 119	38.42
	71+	4	<0.01		422	0.08
Years of education	7 years	34 637	15.96		84 895	17.23
	10 years	38 326	17.66		82 757	16.79
	13 years	11 865	5.47		27 008	5.48
	15 years	26 822	12.36		58 680	11.91
	19 years	34 934	16.10		32 725	6.65
	20 years	70 429	32.45		160 982	32.71
Body mass index	Underweight	1 106	0.51		2 624	0.52
	Normal	73 037	33.66		162 261	32.28
	Overweight	92 742	42.74		212 071	42.19
	Obese	50 128	23.10		125 699	25.01
Systolic blood pressure	Mean (SD)	136.51 (18.68)			135.95 (18.72)	
Smoking initiation	Never	86 999	40.20		200 747	40.2
	Ever	129 391	59.80		298 665	59.8
Cardiovascular disease (all subtypes)	Control	200 787	92.52		418 126	92.38
	Case	16 225	7.48		34 513	7.62
Stroke	Control	200 787	99.18		418 126	99.09
	Case	1 776	0.88		3 840	0.91
Acute Myocardial Infarction	Control	200 787	99.37		418 126	99.32
	Case	1 343	0.66		2 860	0.68
Coronary Heart Disease	Control	200 787	97.82		418 126	97.74
	Case	4 582	2.23		9 677	2.26

4.7.2 Effect of education on risk of cardiovascular outcomes

In phenotypic analyses, a 1-SD higher education was associated with a 14% lower risk of CHD with an odds ratio (OR) of 0.86 (95% CI 0.84 to 0.89). Individual level MR analysis indicated a stronger protective effect, with an OR of 0.38 (95% CI 0.24 to 0.59) (Figure 4.3).

Similar protective associations were found for the effect of education on other CVD subtypes (Figure 4.3). In phenotypic analyses, a one SD higher education was associated with an 11% lower risk of stroke, with an OR of 0.89 (95% CI 0.85 to 0.93). In individual level MR analyses the protective effect was stronger, although estimated with less precision, with an OR of 0.53 (0.26 to 1.07) (Figure 4.3). All three approaches (phenotypic, individual-level MR and summary-sample MR) provided consistent evidence for a protective effect of education with CVD risk and its subtypes.

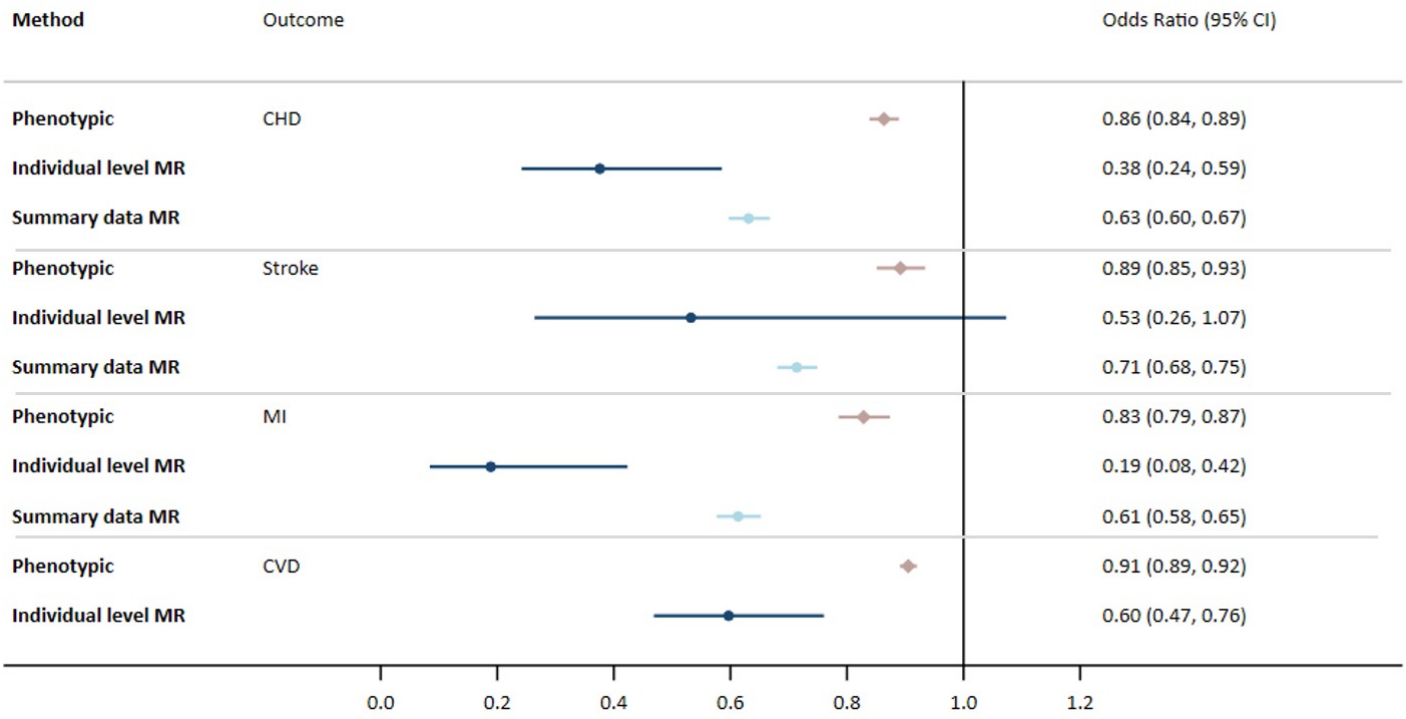

Figure 4.3: The effect of a 1-SD increase in education on the risk of cardiovascular disease and its subtypes. Phenotypic multivariable estimates are plotted in pink and individual level Mendelian randomisation (MR) estimates plotted in navy and summary data MR estimates in light blue. Multivariable analyses and individual level MR analyses adjusted for: age, sex, place of birth and Townsend deprivation index at birth. Body mass index (BMI), systolic blood pressure (SBP) and smoking were measured in one SD units. Cardiovascular disease (CVD) (All subtypes) was not available for analysis in summary data MR analysis.

CHD = coronary heart disease; MI = myocardial infarction; CI = confidence interval

4.7.3 Effect of education on BMI, systolic blood pressure and smoking

In all methods, a longer time in education was associated with lower BMI, systolic blood pressure and smoking (Figure 4.4).

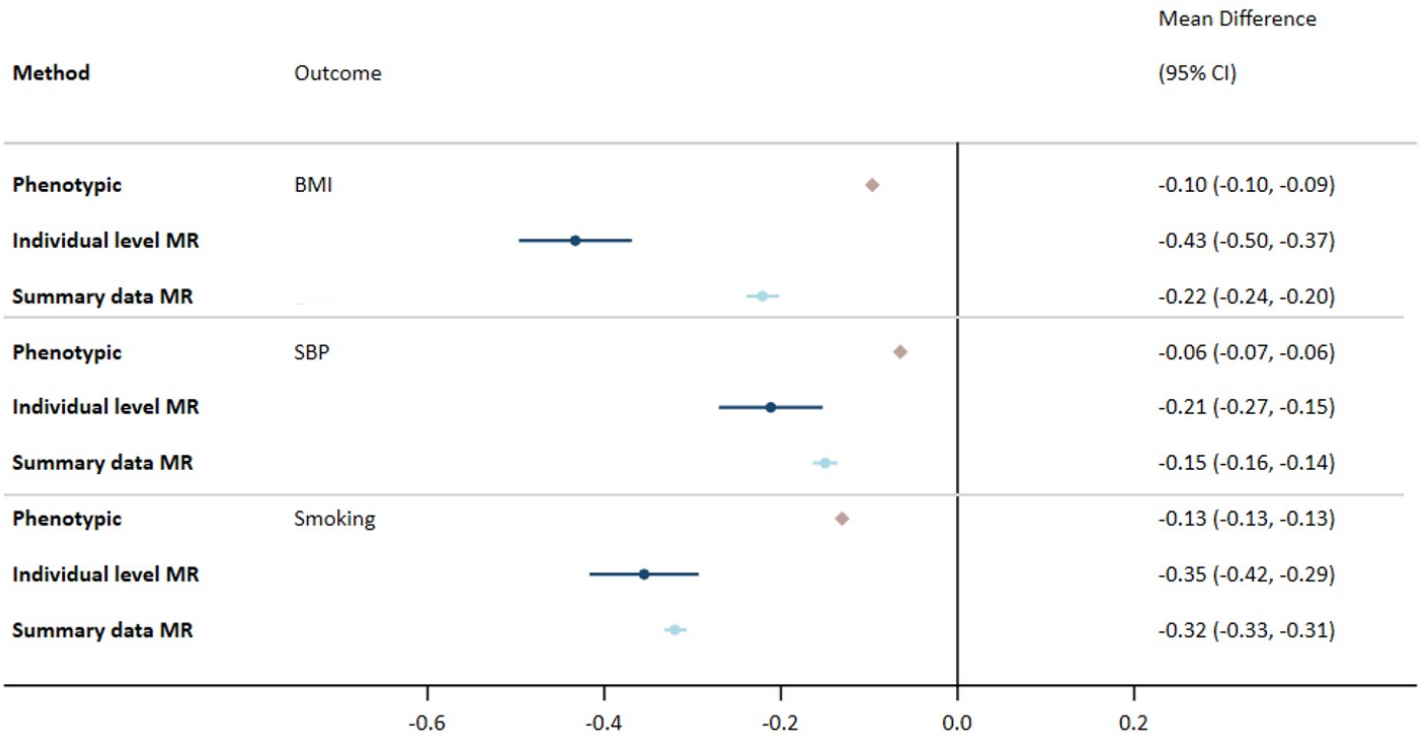

Figure 4.4: Phenotypic and summary data MR estimates for the association between one SD higher education and body mass index (BMI), systolic blood pressure (SBP) and lifetime smoking respectively. All outcomes are in one SD units. Phenotypic multivariable results are plotted in pink, with individual level Mendelian randomisation (MR) estimates plotted in navy and summary data MR estimates in light blue.

4.7.4 Effect of BMI, systolic blood pressure and smoking on risk of cardiovascular outcomes

Both phenotypic and summary data MR analyses consistently found evidence to support an increased risk of CHD with higher BMI, systolic blood pressure and smoking, after adjusting for education (Figure 4.5). Although in some instances the point estimates from individual level MR indicated a protective effect of the risk factors, such as for BMI to MI, these estimates were imprecise with confidence intervals spanning the null value.

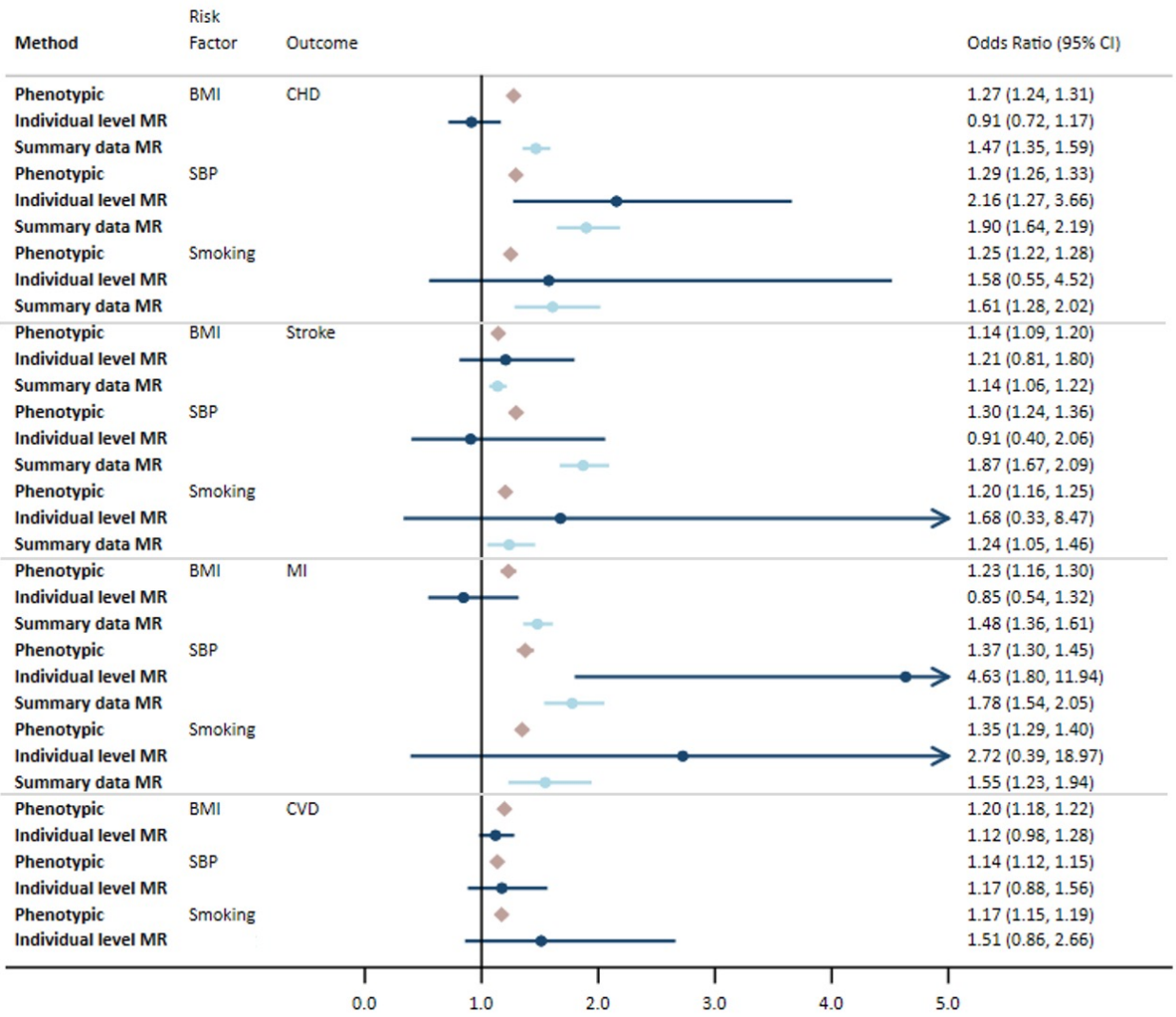

Figure 4.5: Phenotypic, individual level and summary data Mendelian randomisation (MR) associations of a one SD higher body mass index (BMI), systolic blood pressure (SBP) and lifetime smoking on the risk of cardiovascular disease (CVD) and its subtypes. Phenotypic multivariable results are plotted in pink, with individual level MR estimates plotted in navy and summary data MR estimates in light blue.

CHD = coronary heart disease; MI = myocardial infarction; CI = confidence interval

4.7.5 Mediation by BMI, systolic blood pressure and smoking

In the phenotypic analysis, the proportion of the effect of education on CHD risk mediated by BMI was 15% (95% CI: 13% to 17%), 11% for systolic blood pressure (95% CI: 9% to 13%) and 19% for smoking (95% CI: 15% to 22%) (Figure 4.6). In the individual level MR analysis, the proportion mediated by BMI was -4% (95% CI: -13% to 4.5%), 17% by systolic blood pressure (95% CI: -20% to 53%) and 17% by smoking (95% CI: -49% to 82%). In the summary data MR analysis, the percentage mediated by BMI was 18% (95% CI: 14% to 23%), 21% by systolic blood pressure (95% CI: 15% to 26%) and 33% by smoking (95% CI: 17% to 49%) (Figure 4.6).

In phenotypic analyses, combining all three risk factors together explained 42% (95% CI: 36% to 48%) of the effect of education on risk of CHD (Figure 4.6). In the individual level MR analyses, all three risk factors estimated 35% (95% CI: 15% to 56%) of the effect of BMI on CHD. In summary data MR the combined effect of all three risk factors on CHD as 36% (95% CI: 16% to 63%).

Similar results were found for other CVD subtypes in multivariable phenotypic analyses. Smoking consistently mediated around 20% of the association. BMI explained between 10% and 17% of the association between education and CVD and its subtypes, whilst systolic blood pressure explained between 8% and 18%. In summary data MR analyses, smoking explained up to 34% of the association between education and CVD subtypes, whilst BMI estimated up to 18% and systolic blood pressure up to 28% of the association. Individual level MR analyses were consistent with the main conclusions from phenotypic and summary data MR analyses, although estimates were more imprecise, and the confidence intervals spanned the null value for some risk factors (Figure 4.6).

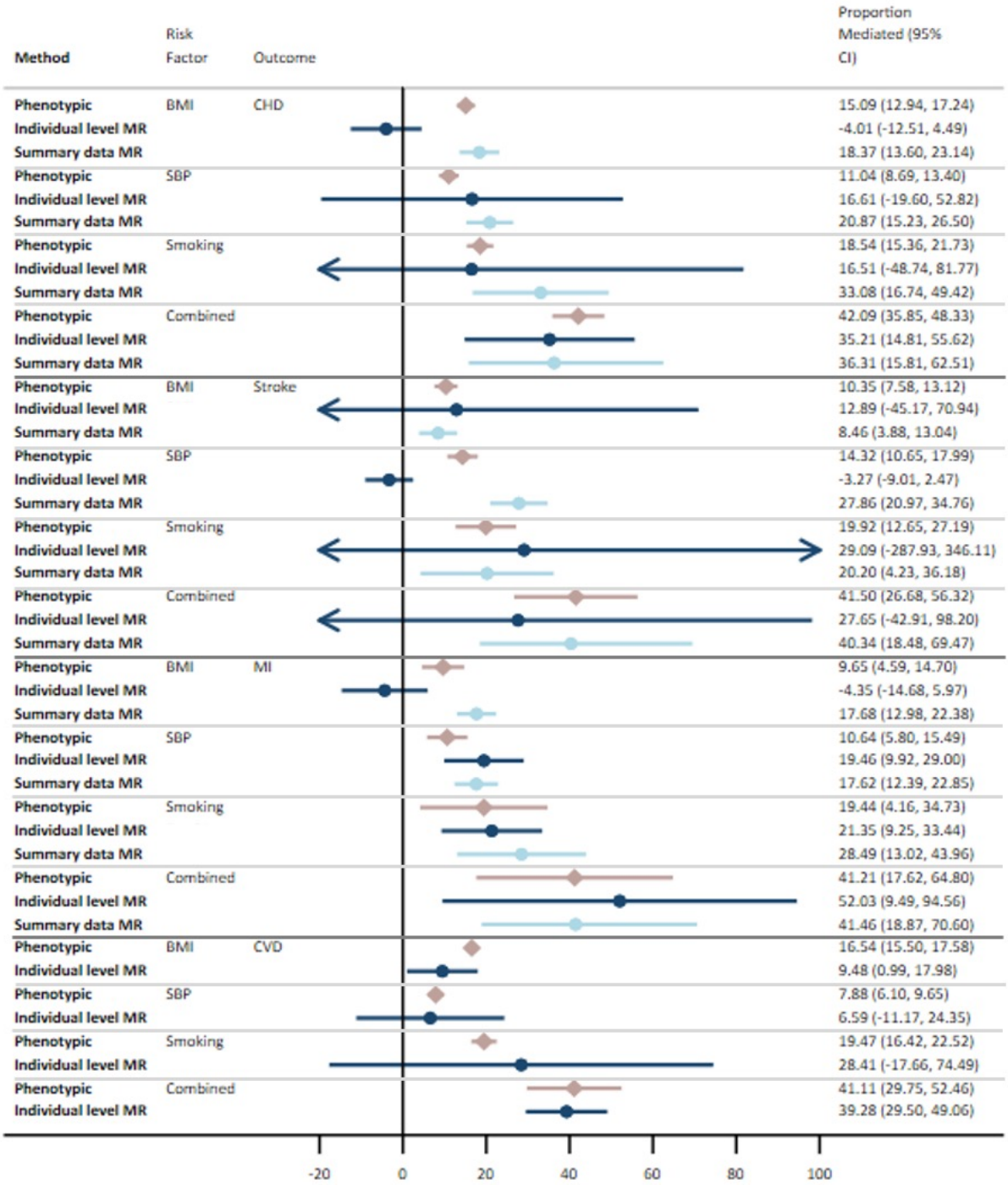

Figure 4.6: Estimates for the effect of education on cardiovascular disease (CVD) and its subtypes explained by body mass index (BMI), systolic blood pressure (SBP) and smoking respectively estimated on the odds ratio scale. Results are provided for the multivariable phenotypic analysis (plotted in pink) and individual level Mendelian randomisation (MR) (plotted in navy) and summary data MR (plotted in light blue). Combined estimates refer to the effect of BMI, systolic blood pressure and smoking considered together in a single mode. Phenotypic and individual level MR analyses are adjusted for age, sex, place of birth and Townsend deprivation index at birth. BMI, systolic blood pressure and smoking were measured in 1-SD units.

CHD = coronary heart disease; MI = myocardial infarction; CI = confidence interval

4.7.6 **Sensitivity analyses**

Results from MR-Egger sensitivity analyses were comparable to the main results but produced less precise estimates with wider confidence intervals, indicating results are unlikely to be biased by pleiotropy (Table 4.6 and Table 4.7).

Unadjusted and age and sex adjusted models were also consistent with the main fully adjusted models for multivariable and individual level MR analyses (Table 4.8 and Table 4.9).

Analyses stratified by age and separately by sex were consistent with the non-stratified main results, although confidence intervals were wide in MR (Figure 4.7 and Figure 4.8).

The effects of each mediator individually, and combined, estimated on the risk difference scale and using the difference method in individual data were consistent with main analyses on the log odds ratio scale suggesting results are unlikely to be biased due to the non-collapsibility of the odds ratio (Figure 4.9).

Including diet and exercise measures in addition to BMI, systolic blood pressure and smoking did not change the amount of the education to CVD (all subtypes) association explained (Figure 4.10).

Table 4.6: Mendelian Randomisation sensitivity analyses for the association between education and mediators, using MR-Egger and Weighted median analyses, in standard deviation units.

	Summary data MR		**Individual level MR**	
	Estimate (95% CI)	**P Value**	**Estimate (95% CI)**	**P Value**
Education-body mass index				
IVW	-0.22 (-0.24, -0.20)	$1.10x10^{-123}$	-0.36 (-0.49, -0.23)	$4.18x10^{-8}$
MR-Egger	-0.28 (-0.49, -0.07)	0.009	-0.15 (-0.41, 0.12)	0.292
MR-Egger intercept		0.989		0.081
Weighted median	-0.27 (-0.30, -0.23)	$5.33x10^{-53}$	-0.51 (-0.62, -0.39)	$1.62x10^{-18}$
Education-Systolic blood pressure				
IVW	-0.15 (-0.17, -0.14)	$3.59x10^{-105}$	-0.14 (-0.24, -0.04)	0.005
MR-Egger	-0.13 (-0.21, -0.05)	0.002	-0.10 (-0.30, 0.11)	0.366
MR-Egger intercept		0.325		0.646
Weighted median	-0.18 (-0.21, -0.16)	$1.14x10^{-52}$	-0.12 (-0.21, -0.03)	0.008
Education-Smoking				
IVW	-0.32 (-0.33, -0.31)	$<1x10^{-300}$	-0.37 (-0.47, -0.28)	$1.243x10^{-14}$
MR-Egger	-0.29 (-0.36, -0.22)	$1.43x10^{-16}$	-0.40 (-0.60, -0.20)	$7.50x10^{-5}$
MR-Egger intercept				0.734
Weighted median	-0.35 (-0.37, -0.33)	$4.05x10^{-236}$	-0.37 (-0.46, -0.29)	$6.765x10^{-18}$

MR = Mendelian randomisation; CI = confidence interval; IVW = inverse variance weighted

Table 4.7: Mendelian Randomisation sensitivity analyses for the association between education and cardiovascular outcomes, using MR-Egger and Weighted median analyses, in OR units. In individual level analyses the weighted median was estimated on the risk difference scale and converted to OR using linear combinations.

	Summary data MR		**Individual level MR**	
	Estimate (95% CI)	**P Value**	**Estimate (95% CI)**	**P Value**
Education - Coronary heart disease				
IVW	0.63 (0.60, 0.67)	$1.77\text{-x}10^{-59}$	0.51 (0.26, 1.00)	0.051
MR-Egger	0.68 (0.54, 0.85)	0.001	0.54 (0.12, 2.34)	0.406
MR-Egger intercept		0.37		0.934
Weighted median	0.62 (0.57, 0.67)	$3.48\text{x}10^{-31}$	0.98 (0.96, 0.99)	0.33
Education - Stroke				
IVW	0.71 (0.68, 0.75)	$6.22\text{x}10^{-44}$	0.46 (0.30, 0.71)	$3.58\text{x}10^{-4}$
MR-Egger	0.72 (0.60, 0.87)	0.001	0.57 (0.22, 1.47)	0.245
MR-Egger intercept		0.757		0.601
Weighted median	0.71 (0.66, 0.76)	$5.79\text{x}10^{-22}$	0.99 (0.98, 1.00)	0.002
Education - myocardial infarction				
IVW	0.61 (0.58, 0.65)	$5.63\text{x}10^{-55}$	0.18 (0.08, 0.38)	$1.31\text{x}10^{-5}$
MR-Egger	0.67 (0.52, 0.85)	0.001	0.20 (0.04, 1.03)	0.054
MR-Egger intercept		0.384		0.883
Weighted median	0.59 (0.54, 0.65)	$2.09\text{x}10^{-28}$	0.99 (0.98, 1.00)	0.002
Education - Cardiovascular disease (all subtypes)				
IVW			0.64 (0.51, 0.82)	$1.98\text{x}10^{-4}$
MR-Egger			0.57 (0.34, 0.95)	0.031
MR-Egger intercept				0.591
Weighted median			0.96 (0.94, 0.98)	$1.05\text{x}10^{-4}$

MR = Mendelian randomisation; CI = confidence interval; IVW = inverse variance weighted

Table 4.8: Unadjusted estimates for the proportion mediated by body mass index (BMI), systolic blood pressure (SBP) and smoking on the association between education and cardiovascular outcomes using phenotypic logistic regression and individual level Mendelian randomisation (MR) analyses in UK Biobank

Outcome	Method	Mediator	Proportion Mediated (%) (95% CI)
Cardiovascular disease (all subtypes)	Multivariable phenotypic	BMI	10.93 (9.52, 12.34)
		SBP	19.18 (16.96, 21.40)
		Smoking	12.66 (10.80, 14.52)
	Individual level MR	BMI	9.47 (-8.18, 21.12)
		SBP	4.62 (-9.63, 18.88)
		Smoking	20.96 (-23.20, 65.12)
Stroke	Multivariable phenotypic	BMI	5.76 (3.36, 8.16)
		SBP	21.68 (15.32, 28.05)
		Smoking	10.23 (6.71, 13.75)
	Individual level MR	BMI	13.58 (-251.31, 278.47)
		SBP	-4.88 (-66.78, 57.00)
		Smoking	24.00 (-2634.32, 2682.32)
Myocardial infarction	Multivariable phenotypic	BMI	9.15 (6.03, 12.28)
		SBP	24.56 (17.42, 31.70)
		Smoking	18.35 (13.10, 23.59)
	Individual level MR	BMI	-4.93 (-18.47, 8.61)
		SBP	16.07 (-21.86, 54.00)
		Smoking	20.11 (-4.06, 44.27)
Coronary heart disease	Multivariable phenotypic	BMI	10.64 (8.66, 12.62)
		SBP	23.68 (20.42, 26.95)
		Smoking	13.88 (11.60, 16.19)
	Individual level MR	BMI	-4.96 (-16.75, 6.84)
		SBP	13.76 (-15.49, 43.01)
		Smoking	11.75 (-19.50, 43.00)

Table 4.9: Minimally adjusted (age and sex only) estimates for the proportion mediated by body mass index (BMI), systolic blood pressure (SBP) and smoking on the association between education and cardiovascular outcomes using phenotypic logistic regression and individual level Mendelian randomisation (MR) analyses in UK Biobank

Outcome	Method	Mediator	Proportion Mediated (%)
Cardiovascular disease (all subtypes)	Multivariable phenotypic	BMI	16.75 (13.73, 19.77)
		SBP	8.10 (6.50, 9.69)
		Smoking	19.36 (15.67, 23.05)
	Individual level MR	BMI	9.18 (-5.97, 24.34)
		SBP	5.95 (-10.04, 21.93)
		Smoking	23.26 (-6.26, 52.78)
Stroke	Multivariable phenotypic	BMI	10.59 (3.05, 18.13)
		SBP	14.47 (5.77, 23.17)
		Smoking	19.83 (9.15, 30.51)
	Individual level MR	BMI	12.43 (-233324.91, 23349.77)
		SBP	-4.30 (-62.25, 53.66)
		Smoking	25.25 (-145.56, 196.07)
Myocardial infarction	Multivariable phenotypic	BMI	9.85 (5.88, 13.83)
		SBP	10.78 (6.98, 14.58)
		Smoking	19.24 (12.99, 25.84)
	Individual level MR	BMI	-4.64 (-18.07, 8.79)
		SBP	20.23 (-26.15, 66.61)
		Smoking	18.95 (-9.84, 47.73)
Coronary heart disease	Multivariable phenotypic	BMI	15.37 (11.55, 19.18)
		SBP	11.26 (8.92, 13.60)
		Smoking	18.50 (14.45, 22.56)
	Individual level MR	BMI	-4.24 (-15.23, 6.74)
		SBP	16.97 (-17.84, 51.79)
		Smoking	11.75 (-19.50, 43.00)

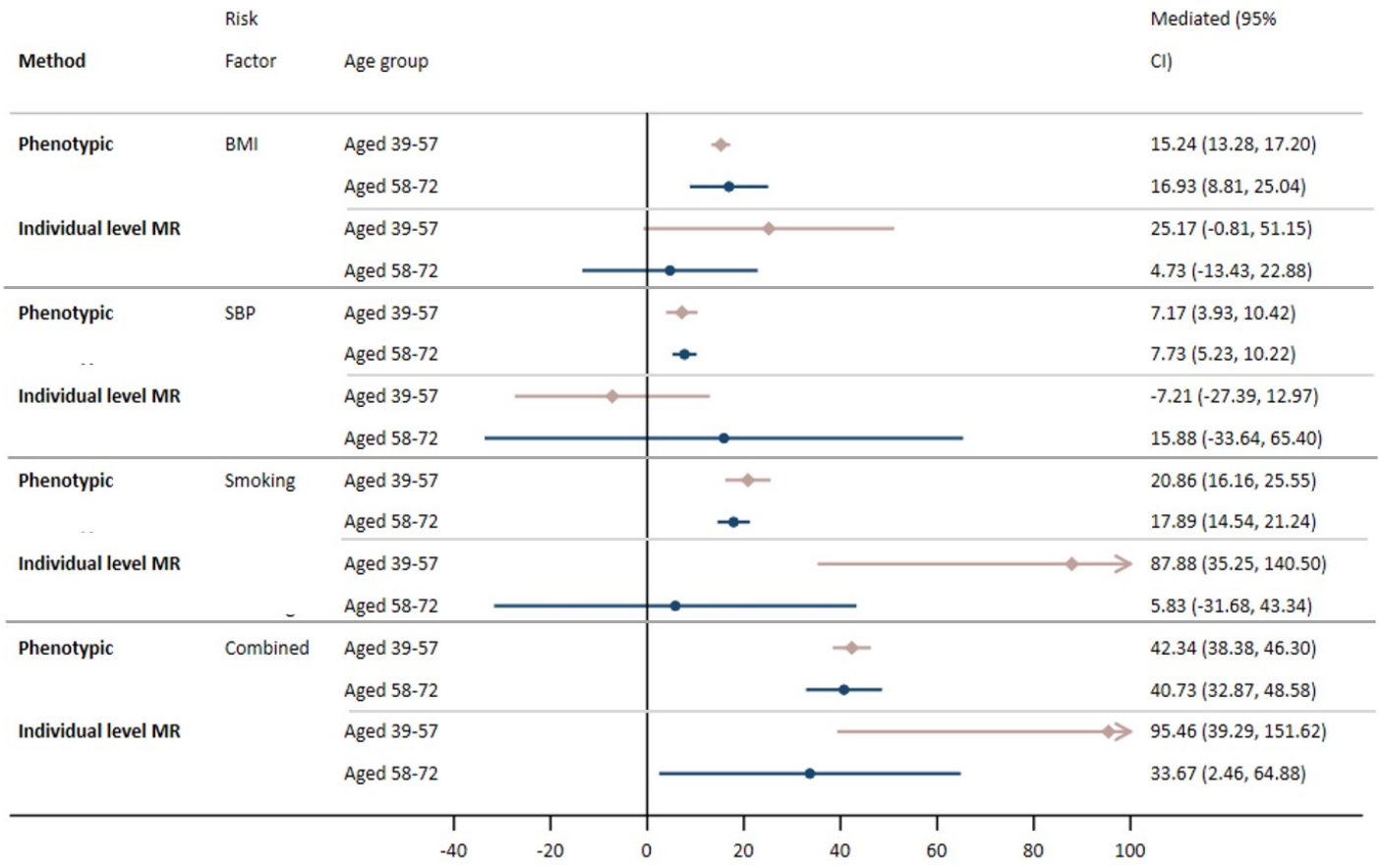

Figure 4.7: Estimates of the proportion mediated between education and cardiovascular disease (all subtypes) by body mass index (BMI), systolic blood pressure (SBP) and smoking in phenotypic multivariable analyses and individual level Mendelian randomisation (MR) analyses stratified by below the median value for age (39-57 years in pink) and above the median value for age (58-72 years in Navy). Analyses are adjusted for age, sex, place of birth and Townsend deprivation index at birth. BMI, systolic blood pressure and smoking were measured in one SD units.

CI = confidence interval

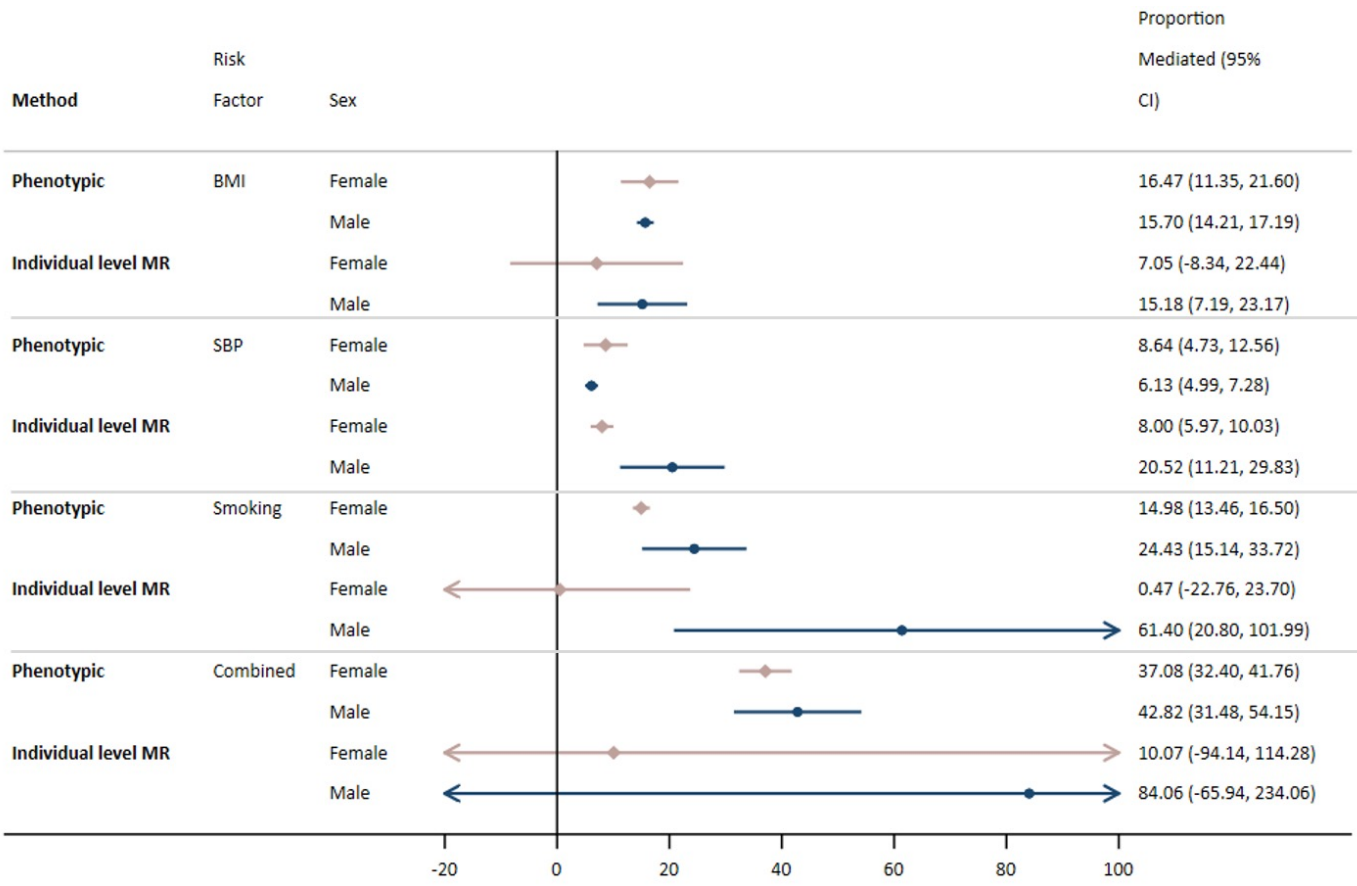

Figure 4.8: Estimates of the proportion mediated between education and cardiovascular disease (CVD) by body mass index (BMI), systolic blood pressure (SBP) and smoking in phenotypic multivariable analyses and individual level Mendelian randomisation (MR) analyses stratified by sex. Analyses are adjusted for age, sex, place of birth and Townsend deprivation index at birth. BMI, systolic blood pressure and smoking were measured in one SD units.

CI = confidence interval

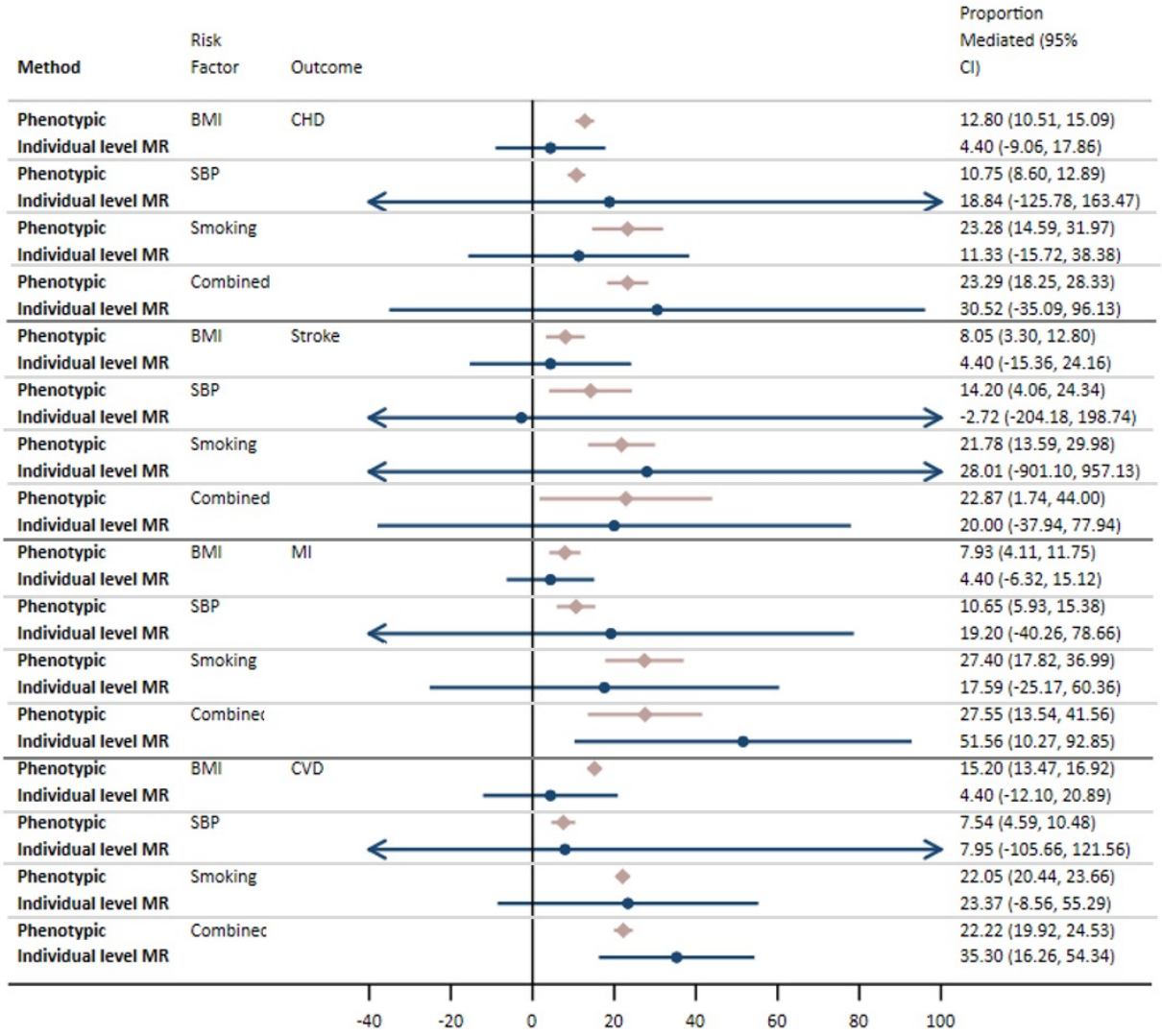

Figure 4.9: Estimates for the effect of education on cardiovascular disease (CVD) and its subtypes explained by body mass index (BMI), systolic blood pressure (SBP) and smoking respectively, estimated on the risk difference scale *Results are provided for the multivariable phenotypic analysis (plotted in pink) and individual level Mendelian randomisation (MR) (plotted in navy). Combined estimates refer to the effect of BMI, SBP and smoking considered together in a single model Analyses are adjusted for age, sex, place of birth and Townsend deprivation index at birth. BMI, systolic blood pressure and smoking were measured in one SD units.*

CHD = coronary heart disease; MI = myocardial infarction; CI = confidence interval

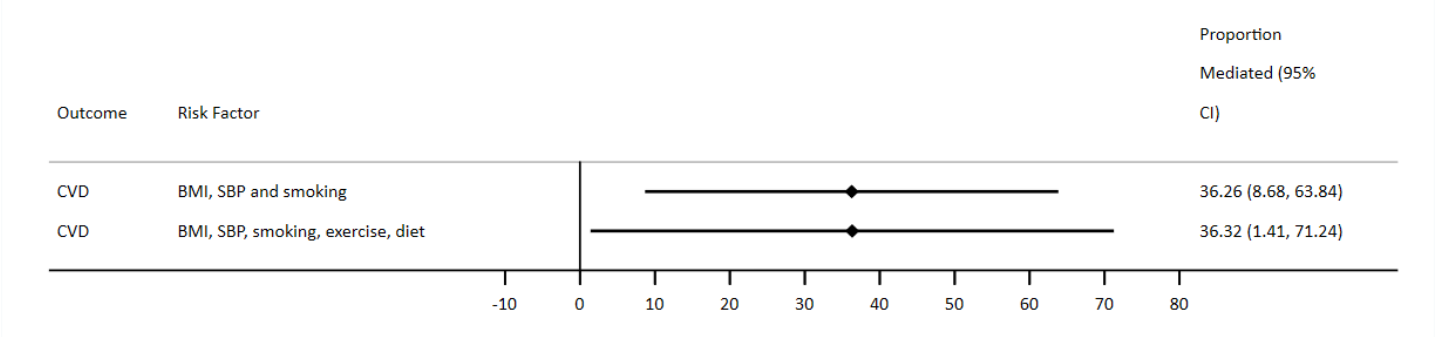

Figure 4.10: Estimate of the additional proportion mediated by exercise and diet compared with body mass index (BMI), systolic blood pressure (SBP) and smoking in multivariable phenotypic multiple mediator models (N=20 298). Both models additionally adjusted for covariates, including age, sex, place of birth and Townsend deprivation index at birth. BMI, systolic blood pressure and smoking were measured in one SD units.

4.8 Discussion

In this chapter, I use phenotypic and genetic analyses to provide complementary evidence that the effect of education on risk of CVD is mediated by approximately up to one third through any of BMI, systolic blood pressure or smoking. When investigating all three risk factors together, around 40% of the association between education and CVD was explained by the three risk factors combined, both in phenotypic and MR analyses. It is important to note that over half of the effect of education remained unexplained in these analyses. The main analyses did not consider the contributions of exercise, diet, health system factors, lipid profile and glycaemic traits (329-335). However, these risk factors are likely to be inter-related with the three main risk factors considered in our analysis. For example, much of the effect of diet and activity on CVD is likely to act through BMI and systolic blood pressure, and therefore the cumulative effect of BMI, systolic blood pressure and smoking together is likely to be capturing some of their effects. Indeed, in a phenotypic sensitivity analysis including diet and exercise alongside BMI, systolic blood pressure and smoking, no more of the association between education and CVD was explained compared with just looking at BMI, systolic blood pressure and smoking.

In this analysis, I have triangulated evidence across three distinct approaches. Although the point estimates vary, along with the mediation results, all three approaches indicate the same conclusions. The MR estimates are much larger in magnitude than the phenotypic results. In MR, the genetic instruments used to proxy the exposure and mediators estimate a lifetime effect, rather than a single snapshot, which may explain the larger estimates in MR. Additionally, this may be due to bias from negative confounding or measurement error in phenotypic analyses. Cases recruited to the case-control studies included in summary data analyses may represent a more extreme phenotype than in cohort studies such as UK Biobank. The summary data MR estimates are more precise than the individual level MR results from UK Biobank, likely related to the larger sample sizes and number of cases.

4.8.1 Findings in context

Mendelian randomisation studies have previously investigated the causal effects of education on CHD, BMI, systolic blood pressure and smoking (9, 118, 119, 268), with others further estimating the effects of BMI and smoking on CVD (45, 135). The current study makes a number of notable advances. The most recent GWAS of educational attainment was used to optimise the power of the summary data MR analysis. With the larger sample size, the instruments selected from this study explained approximately 12% of the variance in

education, as compared with the 3% accounted for in the previous studies of education and CVD (17, 149). Similarly, by leveraging the power of the UK Biobank and recent large-scale GWAS meta-analyses, it was possible to study additional cardiovascular outcomes, including stroke and MI. In addition to the overall effects of the considered risk factors on CVD, I have been able to estimate the proportion of the effect of education that they mediate using a recently developed method (19, 20). To date, genetic instruments for smoking have been limited and are typically related to binary measures that would introduce severe bias in MR (310). The development of a GWAS for the continuous measure of lifetime smoking allowed me to include this in a mediation model (317).

A number of studies have used phenotypic multivariable regression methods to support mediating roles of BMI, systolic blood pressure and smoking in the pathway between education and CVD risk (13, 14, 114, 115), with consistent results obtained using various measures of education, including time spent in schooling and academic qualifications. In an analysis of Dutch individuals, Kershaw *et al*, attributed almost 27% of the association between education and CHD to be due to smoking, with 10% and 5% attributed to obesity and hypertension respectively (114). Similarly, Dégano *et al* found 7% and 14% of the association between education and CVD could be explained by BMI and hypertension respectively (14). However, they found little evidence that smoking mediated the association. Veronesi *et al* analysed their data stratified by sex, but consistently found that systolic blood pressure and smoking mediated the effects of education in both males and females (115). The findings in this study show that phenotypic estimates likely underestimate how much of the effect of education is mediated via smoking, BMI and systolic blood pressure compared to estimates from MR, likely due to measurement error in the mediators that bias phenotypic estimates towards the null, which is likely to have less impact on MR analyses (19). Given the importance of measurement error as a source of bias in mediation analysis (275), MR is potentially a useful tool for understanding mediation.

4.8.2 Strengths and limitations

The major strength of this work is that it allowed for assessment of the causal role of mediators using MR, an approach that is robust to non-differential measurement error in the mediator. I have used multiple data sources and approaches, each with different potential sources of biases, to thus improve the reliability of our findings through triangulation (259). Furthermore, the mediated effects estimated were consistent across the two MR approaches and in statistical sensitivity analyses. The imprecision in the individual level MR analysis demonstrates the need for very large sample sizes to achieve sufficient statistical power when

estimating mediation in an MR framework. The results were complemented by the summary data MR approach, which had greater statistical power, but may be susceptible to alternative sources of bias, including those related to participant overlap in the samples used to obtain genetic association estimates for the exposures and outcomes (289). Existing systolic blood pressure GWAS meta-analyses have adjusted for BMI as a covariate, which could introduce collider bias (315, 316), and for this reason I performed a GWAS of systolic blood pressure in UK Biobank to select instruments, without adjusting for BMI. I also applied a 'split sample' systolic blood pressure GWAS approach on unrelated individuals in UK Biobank for use in individual-level data MR to avoid overlapping populations in the genetic association estimates for the exposure and outcome (336), and any associated bias (244, 289). To this end, the individual level MR entirely avoided any population overlap when obtaining genetic estimates for the exposures and the outcomes.

For all CVD subtypes and individual risk factors considered, the largest effects of education were consistently seen with the MR approaches, with smaller effects seen in the analysis of phenotypic data. Measurement error in a mediator leads to an underestimation of the proportion mediated, so the discrepancy between the phenotypic and MR analyses may in part be attributable to MR analyses suffering less bias from measurement error (275). BMI is accurately measured and has little daily variation – and correspondingly the estimates of the proportion of effect mediated by BMI in the phenotypic and MR analyses are similar (15% and 18% respectively). In contrast, systolic blood pressure and lifetime smoking are difficult to measure accurately – and the estimated proportion mediated is smaller in the phenotypic analysis than the MR (11% vs. 21% for systolic blood pressure, and 19% vs. 33% for smoking). Measurement error could also be introduced by participants over-reporting traits perceived to be 'desirable' such as education and underreporting traits perceived to be 'undesirable' such as smoking (337). The estimates for all three risk factors together were more similar between phenotypic and MR estimates, although for all models, the confidence intervals were wide. It is important to note that while MR is more robust to measurement error, the instruments may not necessarily be capturing all aspects of the exposure phenotype under consideration. For example, the instruments for systolic blood pressure capture average systolic blood pressure but may not necessarily reflect variability in blood pressure.

Estimates from MR analyses are robust to reverse causation bias, due to the random allocation of genetic instruments from parents at conception (and thus prior to development of the outcome under consideration). Tyrrell and colleagues have previously used MR to estimate the effects of BMI on education (265), and it is possible that education affects BMI. In these

analyses I only focused on one direction of effect – i.e. that from education to BMI. However, the results presented here are unlikely to be due to reverse causality. As the genetic variants used as instruments are set at conception, they are not influenced by later life exposures. Additionally, we used a large number of strong instruments for education.

Another limitation of the MR approach is that estimates can be biased by pleiotropic pathways where the instrument is associated with the outcome via a phenotype independent of the exposure under consideration. To investigate this possibility, we additionally performed MR-Egger and weighted median sensitivity analyses were performed that are more robust to such pleiotropy (255, 256, 338), which produced results consistent with those from the main MR analyses. If the assumption is made that the genetic variants have a monotonic effect on the exposure, MR estimates will reflect the local average effect of the exposure on the outcome for all individuals whose exposure was affected by the genetic instrument. Little evidence of heterogeneity in the effect of the exposures was found. This suggests the effects of the SNPs on the exposure may be similar across the population, in which case the MR estimate may provide a reliable estimate of the average effect in the population.

Analyses in UK Biobank were carried out on white, European individuals, potentially limiting the generalisability of these results to other populations and ethnicities. However, summary data MR analyses were not exclusive to white European individuals (although proportions were low for other populations) and produced consistent results to individual level MR analyses. UK Biobank is not representative of the UK population as a whole and is subject to healthy volunteer bias. Therefore, these results may be biased by selection bias (339).

When estimating the indirect effects of a mediator on a binary outcome, the product of coefficients method (two-step MR) results in the least amount of bias (241), and as such this approach was used to estimate the effects of education through each risk factor individually. However, this method cannot currently be used to consider multiple mediators simultaneously in an MR analyses. For this reason, the difference method (MVMR) was used to estimate the effect of education through the three considered risk factors collectively with MR. Although such an approach may introduce theoretical bias due to the non-collapsibility of an odds ratio when investigating a binary outcome, individual level data analyses in UK Biobank were also carried out on a linear risk difference scale to identify whether results on the odds ratio scale may be biased in this way. Estimates for the effect of education through the risk factors collectively were consistent between different scales in these analyses, and as such we would not expect any potential biases to alter the interpretation of our results.

4.8.3 Clinical and public health implications

Past policies that increase the duration of compulsory education have improved health and such endeavours must continue (4). However, intervening directly on education is difficult to achieve without social and political reforms. The findings of this study have notable implications for policymakers as they identify potential strategies for reducing education inequalities in health. Furthermore, they also produce quantitative estimates of this, allowing specific consideration of potential public health impact. It is an important finding of this work that BMI, systolic blood pressure and smoking together explain less than half of the overall effect of education. Further research identifying the other related factors and the interplay between them will be key to reducing social inequalities in cardiovascular disease. Furthermore, work investigating more diverse populations will be necessary to support the extrapolation of these findings outside of the considered contexts.

4.8.4 Conclusion

Using distinct analytical methodologies, including genetic approaches that are able draw causal inference, these results suggest that interventions aimed at reducing BMI, systolic blood pressure and smoking in European populations would lead to reductions in cases of CVD attributable to lower levels of education. Importantly, over half of the effect of education on risk of cardiovascular disease is not mediated through these risk factors and further work is required towards investigating this.

Chapter 5. Educational inequalities in statin treatment for preventing cardiovascular disease: cross-sectional analysis of UK Biobank

5.1 Author list and contributions

Alice R Carter[1,2*], Dipender Gill[3-7], Richard Morris[2,8], George Davey Smith[1,2,9], Amy E Taylor[2,9], Neil M Davies[1,2,10†], Laura D Howe[1,2†]

[†]NMD and LDH contributed equally

ARC designed the study, cleaned and analysed the data, interpreted the results, wrote and revised the manuscript. DG advised on defining medications, interpreted the results and critically reviewed and revised the manuscript. RM advised on analyses, interpreted the results and critically reviewed and revised the manuscript. GDS, AET, NMD and LDH all designed the study, interpreted the results, critically reviewed and revised the manuscript and provided supervision for the project. NMD and LDH contributed equally and are joint senior authors on this manuscript. ARC and LDH serve as guarantors of the paper. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

5.2 Summary of personal contributions

In this chapter I triangulate data from UK Biobank baseline assessment centres, linked hospital inpatient records, linked primary care data and linked mediation records to investigate educational inequalities in statin prescribing.

I was sole lead author for the work in this chapter. I was responsible for deriving QRISK3 cardiovascular risk scores using data from baseline assessment centres. I carried out all analyses, following an analysis plan agreed upon by all co-authors, and created publication quality tables and figures. I was responsible for writing the manuscript and revising in accordance to co-authors advice.

A version of this manuscript has been published on the MedRxiv preprint server and it is currently under review (doi: https://doi.org/10.1101/2020.06.11.20128116). In this thesis chapter I have incorporated supplementary figures with the main text.

5.3 Abstract

Background:

The most socioeconomically deprived individuals remain at the greatest risk of cardiovascular disease. Differences in risk adjusted use of statins between educational groups may contribute to these inequalities. I explore whether people with lower levels of educational attainment are less likely to take statins for a given level of cardiovascular risk.

Methods:

Using data from a large prospective cohort study, UK Biobank, I calculated a QRISK3 cardiovascular risk score for 472 097 eligible participants with complete data on self-reported educational attainment and statin use (55% female; mean age, 56). I used logistic regression to explore the association between i) QRISK3 score and self-report statin use and ii) educational attainment and self-report statin use. I then stratified the association of QRISK3 score, and statin use by strata of educational attainment to test for an interaction.

Results:

There was evidence of an interaction between QRISK3 and education, such that for the same QRISK3 score, people with more education were more likely to report taking statins. For example, in women with 7 years of schooling, equivalent to leaving school with no formal qualifications, a one unit increase in QRISK3 score was associated with a 7% higher odds of statin use (odds ratio (OR) 1.07, 95% CI 1.07, 1.07). In contrast, in women with 20 years of schooling, equivalent to obtaining a degree, a one unit increase in QRISK3 score was associated with an 14% higher odds of statin use (OR 1.14, 95% CI 1.14, 1.15). Comparable ORs in men were 1.04 (95% CI 1.04, 1.05) for men with 7 years of schooling and 1.08 (95% CI 1.08, 1.08) for men with 20 years of schooling. Linkage between UK biobank and primary care data meant we were able to carry out a number of sensitivity analyses to test the robustness of our findings. However, a limitation of our study is that a number of assumptions were made when deriving QRISK3 scores which may overestimate the scores.

Conclusions:

For the same level of cardiovascular risk, individuals with lower educational attainment are less likely to receive statins, likely contributing to health inequalities.

5.4 Introduction

Despite reductions in cardiovascular morbidity and mortality in high income countries, the most socioeconomically deprived groups remain at the highest risk of cardiovascular disease (CVD) (3, 76). There is evidence that education is a causal risk factor for CVD (8, 9, 287). I have previously demonstrated that part of this association acts through three modifiable risk factors; body mass index (BMI), systolic blood pressure and lifetime smoking behaviour (Chapter 4) (287). However, as much as 60% of the effect of education on CVD remains unexplained.

Previous studies have assessed the association of socioeconomic position (SEP) with primary (prescribed prior to a cardiovascular event) and secondary (prescribed as a result of a cardiovascular event) CVD preventative treatment rates; however, the direction of effect has been mixed (142, 143, 146, 147, 340). In an analysis of the Whitehall II cohort study (141), and in the British Regional Heart Study (341) there was no evidence of socioeconomic differences in statin prescribing. In other studies it has been reported that those with lower socioeconomic position are more likely to be prescribed statins (36, 142-144). Conversely, some studies have found that individuals of lower socioeconomic position are less likely to be prescribed statins (87, 145-147).

One key challenge in trying to unpick the role of education in statin prescribing (or other primary or secondary prevention mechanisms) is that lower education is associated with higher levels of cardiovascular risk factors. For example, lower education is associated with higher BMI, smoking, higher blood pressure, and lower levels of physical activity (115, 117, 287). Therefore, individuals with low education likely have a greater underlying risk of CVD and therefore potentially have a greater need for statins. However, it is possible that educational differences in health-seeking behaviour or interactions between patients and healthcare professionals may result in those with higher educational levels being prescribed preventative medication at a lower level of clinical 'need' (342, 343). Consequently, it is more informative to test whether there are educational differences in statin use dependent upon cardiovascular risk, rather than to look at the crude association of education and statin use.

Using the UK Biobank cohort, I investigated whether for a given level of cardiovascular risk, measured using the QRISK3 cardiovascular risk score, people with lower education were less likely to report taking statins as primary prevention than those with higher education (23, 344, 345). In secondary analyses I identify whether there are inequalities in the type of statin

(Atorvastatin compared with Simvastatin) prescribed, given that Atorvastatin has greater efficacy than Simvastatin but is more costly (346-349).

5.5 Methods

5.5.1 UK Biobank

The UK Biobank study recruited 503 317 UK adults between 2006 and 2010. Participants attended baseline assessment centres involving questionnaires, interviews, anthropometric, physical and genetic measurements (15, 16). All UK Biobank participants are linked to hospital episode statistics (HES) or Scottish morbidity records (SMR) (referred to jointly as hospital inpatient records), with data available from 1997 in England, 1998 in Wales and 1981 in Scotland (293), with the most recent entry recorded in this analysis in May 2017. Additionally, a subset of participants (approximately 230,000 participants) are linked with primary care data and prescribing data (350). In this chapter, I use data from baseline assessment centres, hospital inpatient records, and linked primary care data where available.

5.5.2 QRISK risk score and included variables

A CVD risk score was created using the QRISK3 algorithm (23). The QRISK3 score is currently used in primary care systems in England and Wales to define the treatment threshold for statin prescriptions. Current guidelines recommend prescribing statins to individuals with a 10% or greater risk of having a cardiovascular event within 10 years (24, 25). QRISK3 scores were derived for all participants with complete data for educational attainment and reported statin use (N= 472 097) (Figure 5.1). For individuals with missing data in any of the QRISK3 variables multiple imputation was used (see statistical analysis section). Scores were derived according to the publicly available QRISK3 algorithm https://qrisk.org/three/index.php.

Where measures were recorded in baseline assessment centres, such as BMI, Townsend deprivation index (TDI) or systolic blood pressure, these values were used. With the exception of systolic blood pressure variability (standard deviation of repeated values) and coronary heart disease (CHD) in a first-degree relative under 60 years of age, all QRISK3 variables were available in UK Biobank. All variables used and assumptions made when deriving QRISK3 scores are available in Table 5.1.

5.5.2.1 Diagnoses of disease

Diagnoses of disease including arthritis, diabetes (type I and type II), systemic lupus erythromatosus, atrial fibrillation, chronic kidney disease, migraine, HIV/AIDS, severe mental illness and erectile dysfunction were ascertained via linked hospital inpatient records or via

linked medication data. UK Biobank treatment codes used to identify cases and ICD-9 and ICD-10 codes are presented in Appendix 3 Table 1.

5.5.2.2 Treatments

Use of drugs at baseline (antihypertensives, corticosteroids and atypical antipsychotics) were defined by self-reported medication use to clinic nurses at baseline. Individuals were coded as using medication if they reported any medication included in the QRISK3 score. In the QRISK3 derivation cohort individuals were required to have at least two prescriptions representing long term use (23). It was not possible to ascertain the number of prescriptions in UK Biobank; however, UK Biobank participants were asked to record regular treatments, rather than short term medication or over the counter medication. All treatment codes used to define these variables in UK Biobank are available in Appendix 3 Table 2.

5.5.2.3 Ethnicity

Ethnicity was reported by participants to study nurses at UK baseline assessment centres. Ethnicity was categorised according to the categories used in the QRISK3 algorithm (23).

5.5.2.4 Townsend deprivation index

Townsend deprivation index of current location was recorded by UK Biobank at baseline .

5.5.2.5 BMI

Height (m) and weight (kg) were measured by UK Biobank study nurses ate baseline assessment centres which were used to calculate BMI (kg/m^2).

5.5.2.6 Smoking

Smoking status (never, former or current) was determined by self-reported data at baseline assessment centres. The number of cigarettes smoked per day in current smokers was reported at baseline assessment centres and categorised according to QRISK3 categories of light (1-9/day), moderate (10-19/day) and heavy smokers (≥20/day) (23).

5.5.2.7 Systolic blood pressure

The mean from two resting automated measures of systolic blood pressure, measured using an Omron HEM-7105IT digital blood pressure monitor, was used in the QRISK3 score.

5.5.2.8 Systolic blood pressure variability

In the absence of longitudinal data on repeated measures of systolic blood pressure in UK Biobank a measure of systolic blood pressure variability was derived from the standard deviation of the two recorded measurements of systolic blood pressure at the baseline assessment centre.

5.5.2.9 Total cholesterol:HDL cholesterol ratio

Non-fasting measures of total serum cholesterol and high-density lipoprotein (HDL)-cholesterol were measured using enzymatic assays (Backman Coulter AU5800) and the ratio of the two values was calculated. UK Biobank corrected serum data for laboratory dilution effects and were excluded if they did not pass UK Biobank quality control (351).

5.5.2.10 Coronary heart disease in a first degree relative under 60 years of age

A measure of family history of CHD was proxied from reported CVD in mothers, fathers and siblings of UK Biobank participants, however age of diagnosis, nor type of CVD, could not be determined.

5.5.2.11 Primary care QRISK score

In a subset of individuals with linked primary care data, QRISK (read 2 code: 38DF.) (N=1 495) (344), or QRISK2 scores (read 2 code: 39DP.) (N = 10 633) (345) were recorded from 2007 onwards. Where more than one QRISK score was recorded for an individual, the first recorded value was used in analysis.

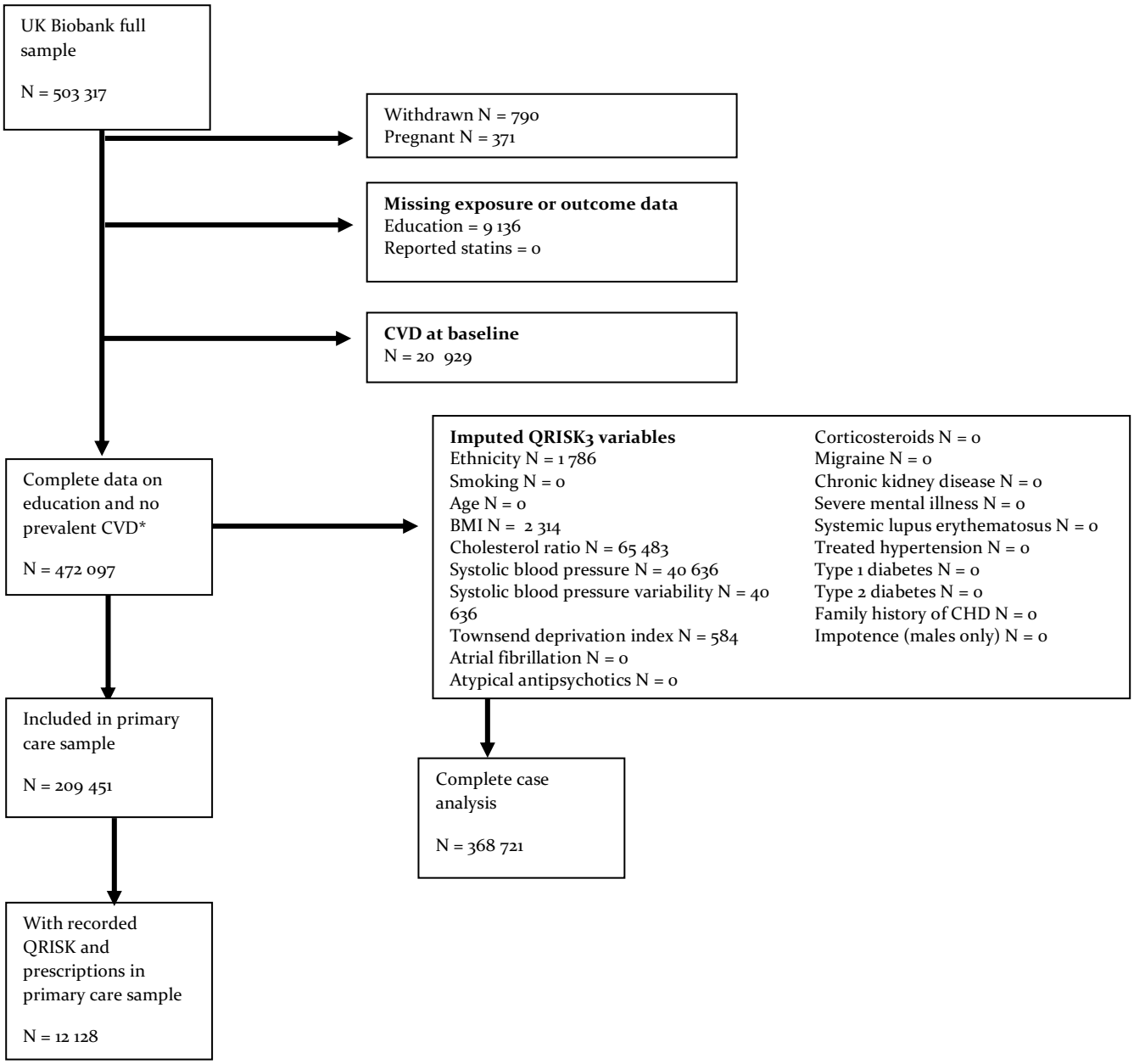

Figure 5.1:Study flow chart identifying eligible participants for analysis

Note: At each stage the same participant could have missing data for multiple variables, therefore overlap is present between the variables. The total excluded may be less than the sum of individuals at each stage.

CVD = cardiovascular disease; BMI = body mass index; CHD = coronary heart disease

Table 5.1: Variables used, and assumptions made when generating QRISK3 scores in UK Biobank participants at baseline

Variable included in QRISK3 algorithm	Measured in UK Biobank by	ICD Code	UK Biobank Variable	Assumptions/limitations to the UK Biobank variables
Diagnoses				
Arthritis	HES data and SMR	M05		
Diabetes (Type I and II)	HES data and SMR	E10-E14		
Systemic lupus erythematosus	HES data and SMR	M32.9		
Atrial fibrillation	HES data and SMR	I48		
Chronic kidney disease	HES data and SMR	N18.3-N18.5		
Migraine	HES data and SMR	G43		
HIV/AIDS	HES data and SMR	B20		
Severe mental illness	HES data and SMR	F20, F23, F31, F32, F33		
Erectile dysfunction	Nurses interview treatment data	N52	n_20003_0	
Treatments				
Antihypertensives	Nurses interview treatment data		n_20003_0	Original QRISK3 derivation specifies that use of drugs at baseline was defined as at least two prescriptions, with the most recent one no more than 28 days before the date or cohort entry. This cannot be ascertained in UK Biobank baseline data, and
Corticosteroids	Nurses interview treatment data		n_20003_0	
Second generation atypical Psychotics	Nurses interview treatment data		n_20003_0	
Lifestyle and biological factors				
Ethnicity	Self-report/ Genetic confirmation		n_21000_0_0	
TDI	Postcode at baseline		n_189_0_0	
BMI	Baseline clinic		n_21001_0_0	
Smoking	Self-report at baseline		n_20116_0_0 n_3456_0_0	Calculated from derived variable for cigarettes per day
Age	Baseline clinic		n_21003_0_0	
Systolic blood pressure	Baseline clinic		n_4080_0_1 n_4080_0_0	
Systolic blood pressure variability	Baseline clinic		n_4080_0_1n_4080_0_0	The QRISK3 algorithm uses the standard deviation of repeated values of blood pressure. This was not available in UK Biobank; therefore, systolic blood pressure variability was derived from the standard deviation between two baseline automated readings of systolic blood pressure
Total cholesterol: HDL ratio	Baseline clinic serum metabolomics		n_30690_0_0 n_30760_0_0	
CHD in first degree relative (<60 years)	Self-report		n_20107_0_0 n_20110_0_0 n_20111_0_0	Includes all reported family history of CVD, not restricted to cases under 60 or specific subtypes

HES = hospital episode statistics; SMR = Scottish morbidity records; TDI = Townsend deprivation index; BMI = body mass index; HDL = high-density lipoprotein; CHD = coronary heart disease; CVD = cardiovascular disease

5.5.3 Measuring educational attainment

UK Biobank participants reported their highest qualification achieved at baseline assessment centres, which was converted to the International Standard Classification for Education (ISCED) coding of years of education (Table 5.2) (323).

Table 5.2: International Standard for Classification of Education codes mapped to UK Biobank self-report highest qualification to estimate years of education

Qualification (As reported in UK Biobank)	ISCED	Years of education	N
College or University degree	5	20	157 109
NVQ or HND or HNC or equivalent	5	19	30 919
Other prof. qual. e.g.: nursing, teaching	4	15	24 550
A levels/AS levels or equivalent	3	13	53 456
O levels/GCSEs or equivalent	2	10	101 222
CSEs or equivalent	2	10	25 999
None of the above	1	7	78 422
Prefer not to answer	Excluded		

5.5.4 Measuring statin use

Participants were asked about regular medication they were taking, details of which were recorded by UK Biobank study nurses. From this, a primary variable for any reported statin use was generated. The type of statin used (Atorvastatin, Simvastatin, Fluvastatin, Pravastatin and Rosuvastatin) was recorded by study nurses and was used to derive a variable for type of statin.

In individuals with linked primary care data, statin prescriptions were recorded in prescription data. In these individuals, a measure of validated statin use was created, defined by a prescription in both the 3 months before and 3 months after baseline. For sensitivity analyses in individuals with a QRISK or QRISK2 score recorded in primary care data, statin use was defined as any statin prescription after a QRISK score was recorded.

5.5.5 Exclusion criteria

Individuals with prevalent CVD at baseline, which would result in a statin prescription according to NICE guidelines (24-26, 188), were excluded from analyses. These cardiovascular diagnoses and events were ascertained through linkage to hospital inpatient records, with cases defined according to ICD-9 and ICD-10 codes (Appendix 3 Table 1). Individuals were excluded if they had experienced at least one diagnosis of myocardial infarction, angina, stroke, transient ischaemic attack, peripheral arterial disease, type 1 diabetes, chronic kidney disease or familial hypercholesterolaemia (24, 26). The date for each diagnosis is provided in

the hospital inpatient records, which was linked with the date of assessment centre visit provided by UK Biobank.

Complete case analyses were carried out on 368 721 individuals, with complete data on age, sex, educational attainment, self-reported statin (medication) use and all variables required for the QRISK3 score (Figure 5.1).

5.5.6 Code and data availability

The derived variables have been returned to UK Biobank for archiving. The code used to derive QRISK3 scores and carry out analyses is available at github.com/alicerosecarter/statin_inequalities.

5.5.7 Patient and public involvement

Ethical approval for this study was sought from the UK Biobank (project 10953). No patients or participants were involved in setting the research question or the outcome measures, nor were they involved in developing plans for design or implementation of the study. No patients were asked to advise on the interpretation or writing up of results.

5.5.8 Statistical analyses

To maximise power and potentially reduce bias, multivariable multiple imputation by chained equations (352) was used to impute variables included in the QRISK3 score with missing data, under the missing at random assumption. The sample for imputation was defined as all individuals with complete data on educational attainment and reported statin use. The proportion of missing data ranged from 0% to 15% (Table 5.3). In total, 77% of participants had no missing data, 13% of participants were missing data for one QRISK3 variable, 8% of participants were missing data for two QRISK3 variables and 2% of participants were missing data for three, four or five variables. A total of 25 imputed datasets were generated (353). Imputation was carried out separately within strata of years of education and sex to preserve interactions tested in the statistical analyses (354). The mean and standard deviation of continuous variables or proportion and standard error of categorical variables in the imputed data were compared with those from the complete data. All analyses were then carried out in each imputed dataset, with results combined according to Rubin's rules.

It was determined *a priori* to carry out all analyses stratified by sex given the known differences in cardiovascular risk profiles for males and females (355, 356), as well as the QRISK3 score being derived separately by sex (23).

To confirm the validity of the derived QRISK3 score, a univariable logistic regression model was used to assess the association of the risk score with i) self-reported statin use and ii) incident CVD.

I estimated the associations of years of education with i) QRISK3 score (using linear regression) and ii) statin use (using logistic regression).

Table 5.3: Proportion of missing data in QRISK3 variables

Variable	Female	Male
	% missing	
QRISK	24%	22%
Age	0%	0%
BMI	0.5%	0.7%
Systolic blood pressure	9%	9%
Townsend deprivation index	0.1%	0.1%
Total cholesterol:HDL cholesterol	15%	13%
Years of education	2%	2%
Ethnicity	0.5%	0.7%
Smoking	0%	0%
Family history of CVD	0%	0%
Statin (reported)	0%	0%
Statin type	0%	0%

5.5.8.1 Testing for interaction between QRISK3 score and educational attainment on statin use

Logistic regression was used to estimate the association of QRISK3 score with self-reported statin use, stratified by years of education, providing an estimate of interaction on the multiplicative scale (Figure 5.2, Route 1). These analyses were not adjusted for any other covariates, assuming all relevant variables are incorporated into the QRISK3 score. Evidence of an interaction between QRISK3 score and years of education was evaluated in a linear model where the interaction term QRISK3*educational attainment was included in the regression model.

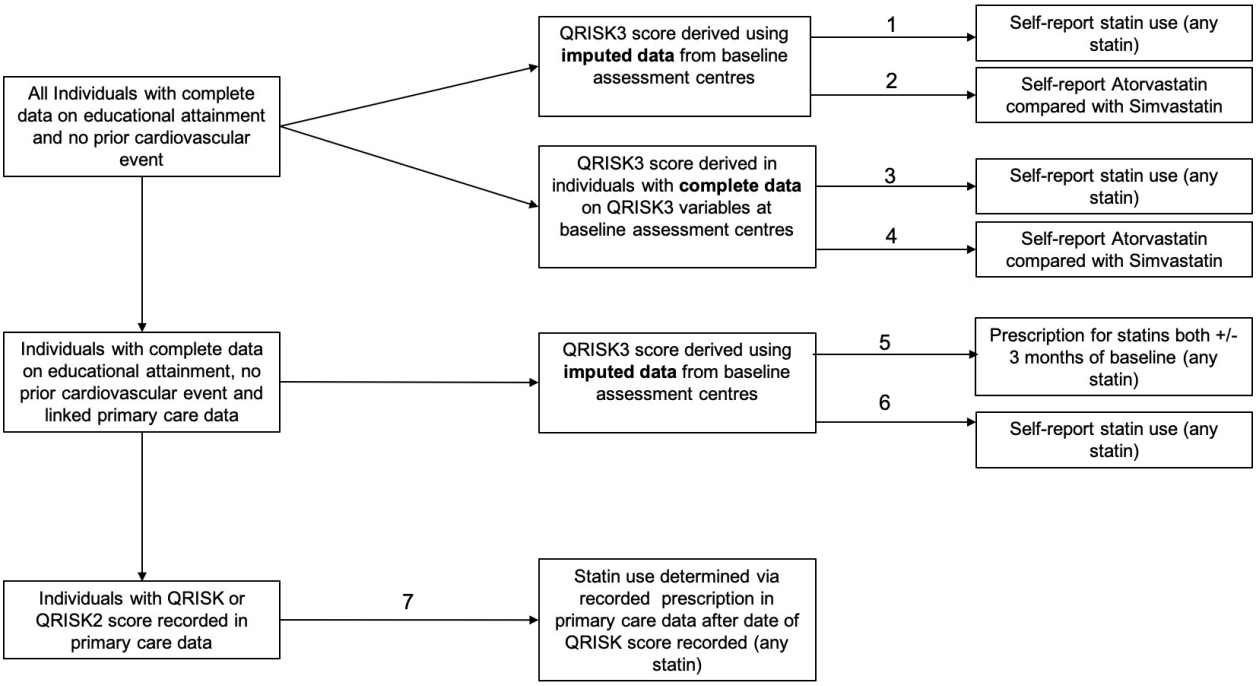

Figure 5.2: Schematic of primary and secondary analyses carried out

5.5.8.2 Secondary analyses

To test the hypothesis that there may be educational inequalities in the type of statin prescribed, in individuals who reported using statins to baseline study nurses, I assessed i) whether there was an association of QRISK3 score and years of education independently with self-reported Atorvastatin, which has been suggested to have a greater efficacy, compared with self-reported Simvastatin (baseline) (346-348) and ii) whether there was any evidence of an interaction between QRISK3 score and years of education on type of statin prescribed (Figure 5.2, Route 2).

Analyses testing the association between QRISK3 and years of education on statin use and statin type independently, as well as for any interaction between QRISK3 score and educational attainment, on statin use were replicated on the additive interaction scale. Additionally main analyses for statin use and type of statin prescribed were replicated using complete case data (Figure 5.2, Route 3 and 4).

To test whether the self-reported statin use data affected the results, I repeated analyses with statin use defined as a prescription both 3 months before and after baseline from linked primary care prescription data (Figure 5.2, Route 5), and also repeated main analyses with self-reported statin use in the subset of participants with the linked prescription data (Figure 5.2, Route 6).

In the subsample of primary care individuals with a QRISK or QRISK2 score recorded, analyses were replicated to test for evidence of an interaction between QRISK score and incident statin prescribing. This was defined as any prescription for a statin recorded in primary care data, excluding individuals who reported using statins to study nurses at the baseline assessment centre (Figure 5.2, Route 7). QRISK scores were included if they were recorded on or prior to the date of first statin prescription, but consideration was not given to the time between both events.

Two further estimates of QRISK3 were derived excluding i) variability of systolic blood pressure and ii) family history of CVD from QRISK3 scores. The pairwise correlation between scores with and without these variables was tested.

5.6 Results

5.6.1 UK Biobank sample

In the main analyses (N = 472 097) 55% of participants were female with a mean age of 56. In females, the QRISK3 score implied a mean 10-year risk of a cardiovascular event of 6.9% (standard deviation (SD) = 5.5). In males, the QRISK3 score implied mean a 10-year risk of a cardiovascular even of 13.1% (SD = 8.4). Participants were more likely to have completed 20 years of education (female = 35%, male = 38%) than 7 years of education (female = 14%, male = 14%). 10% of females and 17% of males reported using statins.

The distribution of variables was similar between the multiply imputed dataset, complete case data, and in the subset of participants with linked primary care data (Table 5.4).

Table 5.4: Descriptive characteristics of UK Biobank participants in i) the full eligible sample analysed ii) the full eligible sample who also have linked primary care data and iii) participants with linked primary care data and a recorded QRISK score

Variable		Imputed analysis sample		Primary care analysis sample (imputed)		Primary care analysis sample with recorded QRISK		Complete case analysis sample	
		(N = 472 097)		(N = 209 451)		(N = 12 128)		(N = 368 721)	
		Female	**Males**	**Female**	**Males**	**Female**	**Male**	**Female**	**Male**
		(N = 261 147)	(N = 210 950)	(N = 117 038)	(N = 92 413)	(N = 7 338)	(N = 4 790)	(N = 201 532)	(N = 167 189)
Continuous variables		**Mean (SD)**							
QRISK*	QRISK3 (baseline)	6.87 (5.54)	12.98 (8.34)	6.94 (5.57)	13.11 (8.35)	6.21 (4.68)	11.44 (7.1)	6.84 (5.5)	12.97 (8.32)
	QRISK3 excluding 'non-validated' statin users	NA	NA	6.09 (4.98)	11.54 (7.82)	NA	NA	NA	NA
	Recorded value of QRISK in primary care	NA	NA	NA	NA	10.17 (6.94)	16.11 (9.2)	NA	NA
Age		56.23 (7.98)	56.44 (8.2)	56.26 (7.94)	56.5 (8.15)	56.28 (7.98)	56.45 (8.2)	56.28 (7.98)	56.45 (8.2)
BMI		27.02 (5.15)	27.75 (4.2)	27.14 (5.18)	27.86 (4.23)	26.96 (5.08)	27.74 (4.18)	26.96 (5.08)	27.74 (4.18)
Systolic blood pressure		135.14 (19.18)	140.94 (17.35)	135.46 (19.17)	141.31 (17.39)	135.15 (19.15)	141 (17.31)	135.15 (19.15)	141 (17.31)
TDI		-1.38 (3.2)	-1.31 (3.12)	-1.41 (2.95)	-1.36 (3.05)	-1.4 (2.99)	-1.34 (3.09)	-1.4 (2.99)	-1.34 (3.09)
Total cholesterol:HDL cholesterol		3.86 (1)	4.48 (1.15)	3.88 (1.01)	4.49 (1.15)	3.84 (1)	4.49 (1.15)	3.84 (1)	4.49 (1.15)
Categorical variables		**Percent of Sample (SE)**				**Frequency (%)**			
Years of education	7 years	14.21 (0.08)	13.83 (0.09)	15.29 (0.12)	14.67 (0.14)	1 034 (14)	601 (13)	32 785 (16)	26 874 (16)
	10 years	19.4 (0.09)	13.52 (0.09)	19.1 (0.13)	13.36 (0.13)	1 520 (21)	649 (14)	39 795 (20)	22 945 (14)
	13 years	6.06 (0.05)	5.27 (0.06)	5.81 (0.08)	5.05 (0.09)	436 (6)	285 (6)	11 729 (6)	8 449 (5)
	15 years	12.83 (0.07)	10.04 (0.08)	12.69 (0.11)	10.16 (0.12)	961 (13)	497 (10)	26 936 (13)	17 161 (10)
	19 years	12.88 (0.07)	19.67 (0.1)	13.13 (0.11)	20.17 (0.16)	911 (12)	944 (20)	25 653 (13)	32 940 (20)
	20 years	34.62 (0.11)	37.67 (0.12)	33.98 (0.16)	36.58 (0.19)	2 476 (34)	1 814 (38)	64 634 (32)	58 820 (35)
Ethnicity	White	94.96 (0.05)	94.7 (0.06)	95.75 (0.07)	95.33 (0.08)	7 026 (96)	4 600 (96)	190 903 (95)	158 386 (95)
	Indian	0.98 (0.02)	1.2 (0.03)	1.04 (0.03)	1.3 (0.04)	66 (1)	49 (1)	2 082 (1)	2 108 (1)
	Pakistani	0.23 (0.01)	0.42 (0.02)	26.52 (0.02)	0.46 (0.03)	21 (0)	11 (0)	462 (0)	717 (0)
	Other Asian	0.48 (0.02)	0.6 (0.02)	0.4 (0.02)	0.58 (0.03)	25 (0)	22 (0)	982 (0)	979 (1)
	Black Caribbean	10.73 (0.02)	0.81 (0.02)	0.77 (0.03)	0.64 (0.03)	55 (1)	18 (0)	2 464 (1)	1 408 (1)
	Black African	0.68 (0.02)	0.86 (0.02)	0.46 (0.02)	0.54 (0.03)	40 (1)	21 (0)	1 435 (1)	1 406 (1)
	Chinese	0.38 (0.01)	0.28 (0.01)	0.32 (0.02)	0.23 (0.02)	26 (0)	26 (0)	719 (0)	719 (0)
	Other	1.22 (0.02)	1.12 (0.03)	1.01 (0.03)	0.92 (0.04)	70 (1)	70 (1)	2 485 (1)	2 485 (1)

Smoking	Never	60.54 (0.11)	52.29 (0.13)	60.79 (0.16)	52.33 (0.19)	4 388 (60)	2 536 (53)	120 335 (60)	83 129 (50)
	Former	30.39 (0.1)	35.02 (0.12)	30.05 (0.15)	35.16 (0.19)	2 346 (32)	1 715 (36)	63 059 (31)	63 033 (38)
	Light (1-9/day)	1.66 (0.03)	1.29 (0.03)	1.59 (0.04)	1.24 (0.04)	128 (2)	57 (1)	3 287 (2)	2 056 (1)
	Moderate (10-19/day)	2.99 (0.04)	2.96 (0.04)	3.16 (0.06)	3.01 (0.07)	176 (2)	102 (2)	6 094 (3)	4 931 (3)
	Heavy (>20/day)	4.42 (4.42)	8.45 (0.07)	4.42 (0.07)	8.26 (0.11)	300 (4)	380 (8)	8 757 (4)	14 040 (8)
Family history of CVD	Control	72.37 (0.1)	78.22 (0.11)	71.5 (0.15)	77.57 (0.16)	5 242 (71)	3 749 (78)	142 641 (71)	128 314 (77)
	Case	27.63 (0.1)	21.78 (0.11)	28.5 (0.15)	22.43 (0.16)	2 096 (29)	1 041 (22)	58 891 (29)	38 875 (23)
Statin (reported)	Control	90.27 (0.06)	82.99 (0.08)	90.14 (0.09)	82.39 (0.13)	NA	NA	181 903 (90)	138 619 (83)
	Case	9.73 (0.06)	17.01 (0.08)	9.86 (0.09)	17.61 (0.13)	NA	NA	19 629 (10)	28 570 (17)
Statin type	No statin	90.27 (0.06)	82.99 (0.08)	90.14 (0.09)	82.39 (0.13)	NA	NA	181 903 (90)	138 619 (83)
	Atorvastatin	1.64 (0.02)	2.87 (0.04)	1.68 (0.04)	2.9 (0.06)	NA	NA	19 629 (10)	28 570 (17)
	Fluvastatin	0.02 (0)	0.06 (0.01)	0.03 (0)	0.06 (0.01)	NA	NA	181 903 (90)	138 619 (83)
	Pravastatin	0.3 (0.01)	0.47 (0.01)	0.29 (0.02)	0.44 (0.02)	NA	NA	3 281 (2)	4 750 (3)
	Rosuvastatin	0.39 (0.01)	0.61 (0.02)	0.38 (0.02)	0.65 (0.03)	NA	NA	49 (0)	96 (0)
	Simvastatin	7.37 (0.05)	13.01 (0.07)	7.49 (0.08)	13.56 (0.11)	NA	NA	617 (0)	787 (0)
Statin (validated)	Control	NA	NA	97.62 (0.05)	95.40 (0.08)	6 345 (86)	3 878 (81)	NA	NA
	Case	NA	NA	2.38 (0.05)	4.60 (0.08)	993 (14)	912 (19)	NA	NA
Reported statin with no prescription*	Control	NA	NA	92.90 (0.08)	86.01 (0.13)	NA	NA	NA	NA
	Case	NA	NA	7.10 (0.08)	13.99 (0.13)	NA	NA	NA	NA
Incident CVD	Control	79.63 (0.08)	0.08 (73.66)	79.85 (0.13)	0.13 (73.57)	5 379 (82)	3 439 (80)	140 753 (79)	106 032 (74)
	Case	20.37 (0.08)	0.08 (26.34)	20.15 (0.13)	0.13 (26.43)	1 179 (18)	885 (20)	36 401 (21)	38 171 (26)

Derived QRISK3 variable from baseline measured in UK Biobank for the full analysis sample and primary care analysis sample, recorded QRISK or QRISK2 scores in primary care data for the primary care analysis sample with recorded QRISK.

**Proportion of individuals excluding individuals with validated prescriptions*

BMI = body mass index; TDI = Townsend deprivation index; HDL = high-density lipoprotein cholesterol; SE = standard error; CVD = cardiovascular disease

5.6.2 Association of QRISK3 score with statins and cardiovascular disease

For a one unit increase in QRISK3 score (i.e. a 1% increase in the 10-year risk of experiencing a cardiovascular event) in females, the odds ratio (OR) for reporting statin use to study nurses was 1.12 (95% confidence interval (CI): 1.12 to 1.13) and the OR for an incident cardiovascular event was 1.12 (95% CI: 1.12 to 1.12) (Figure 5.3 and Table 5.5). Females with a QRISK3 score of 10 or greater were 1.34 (95% CI: 1.31 to 1.36) times more likely to report using statins than those with a QRISK score of less than 10. In males, the OR for statin use was 1.07 (95% CI: 1.07 to 1.07) and for an incident cardiovascular event the OR was 1.08 (95% CI: 1.08 to 1.08) per unit higher QRISK3 score (Figure 5.3 and Table 5.5). Males with a QRISK3 score of 10 or greater were 1.49 (95% CI: 1.46 to 1.52) times more likely to report using statins than those with a QRISK score of less than 10.

Exposure	Sex	Years of Education	Odds Ratio (95% CI)	N	N statins (%)
QRISK (per unit)	Female	No education interaction	1.12 (1.12, 1.13)	261 147	25 409 (10)
		7 years	1.07 (1.07, 1.07)	43 175	7 736 (18)
		10 years	1.12 (1.12, 1.13)	50 709	4 916 (10)
		13 years	1.13 (1.12, 1.15)	15 219	1 185 (8)
		15 years	1.12 (1.11, 1.13)	34 291	3 577 (10)
		19 years	1.14 (1.13, 1.15)	33 308	2 643 (8)
		20 years	1.14 (1.14, 1.15)	84 445	5 352 (6)
QRISK (per unit)	Male	No education interaction	1.07 (1.07, 1.07)	210 950	35 893 (17)
		7 years	1.04 (1.04, 1.05)	34 550	8 653 (25)
		10 years	1.08 (1.07, 1.08)	28 574	4 653 (16)
		13 years	1.07 (1.06, 1.07)	10 688	1 629 (15)
		15 years	1.06 (1.06, 1.07)	21 578	4 053 (19)
		19 years	1.08 (1.07, 1.08)	41 104	6 511 (16)
		20 years	1.08 (1.08, 1.08)	74 456	10 394 (14)

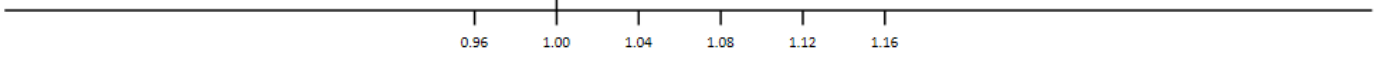

Figure 5.3: Odds ratio of self-report statin use per unit increase in baseline QRISK3 score with no education interaction and stratified by years of education in females and males

Analyses stratified by years of education provide an estimate of interaction on the multiplicative scale

P value for interaction in females = $1.385x10^{-85}$ and males = $1.551x10^{-48}$

CI = confidence interval

Table 5.5: Odds ratio of i) statin use and ii) incident cardiovascular disease per unit increase in QRISK3 score and unit increase in years of education

Exposure	Outcome	Females		Males	
		Complete Case Odds ratio (95% CI) (N = 201 532)	**Imputed sample Odds ratio (95% CI)** (N = 261 147)	**Complete Case Odds ratio (95% CI)** (N = 167 189)	**Imputed sample Odds ratio (95% CI)** (N = 210 950)
QRISK3	Statins (any)	1.12 (1.12, 1.13)	1.12 (1.12 1.13)	1.07 (1.07, 1.07)	1.07 (1.07, 1.07)
	Incident cardiovascular event	1.14 (1.14, 1.15)	1.12 (1.12, 1.12)	1.09 (1.09, 1.09)	1.08 (1.08, 1.08)
Education	Statins (any)	0.93 (0.93, 0.93)	0.93 (0.93, 0.93)	0.96 (0.96, 0.96)	0.96 (0.96, 0.96)
	Incident cardiovascular event	0.95 (0.95, 0.95)	0.95 (0.95, 0.95)	0.93 (0.93, 0.93)	0.96 (0.95, 0.96)

5.6.3 Association of education with QRISK3 score and statin prescribing

Per year increase in educational attainment was associated with a -0.30 (95% CI: -0.30 to -0.29) reduction in mean QRISK3 score in females and a -0.35 (95% CI: -0.35 to -0.34) reduction in mean QRISK3 score in males (Table 5.6 and Figure 5.4).

The prevalence of statin use was highest in those in the lowest strata of educational attainment (equivalent to leaving school after 7 years, with no formal qualifications). Not accounting for cardiovascular risk, each additional year of education was associated with a lower odds of being prescribed statins (all types), (OR in females: 0.93; 95% CI: 0.93 to 0.93, OR in males: 0.96; 95% CI: 0.96 to 0.96) (Table 5.5 and Figure 5.6).

Table 5.6: Mean difference in QRISK3 score per unit increase in between educational attainment

Outcome	Females		Males	
	Complete Case Mean difference (95% CI) (N = 201 532)	**Imputed Sample Mean difference (95% CI)** (N = 261 147)	**Complete Case Mean difference (95% CI)** (N = 167 189)	**Imputed Sample Mean difference (95% CI)** (N = 210 950)
QRISK3	-0.29 (-0.30, -0.29)	-0.30 (-0.30, -0.29)	-0.34 (-0.35, -0.33)	-0.35 (-0.35, -0.34)

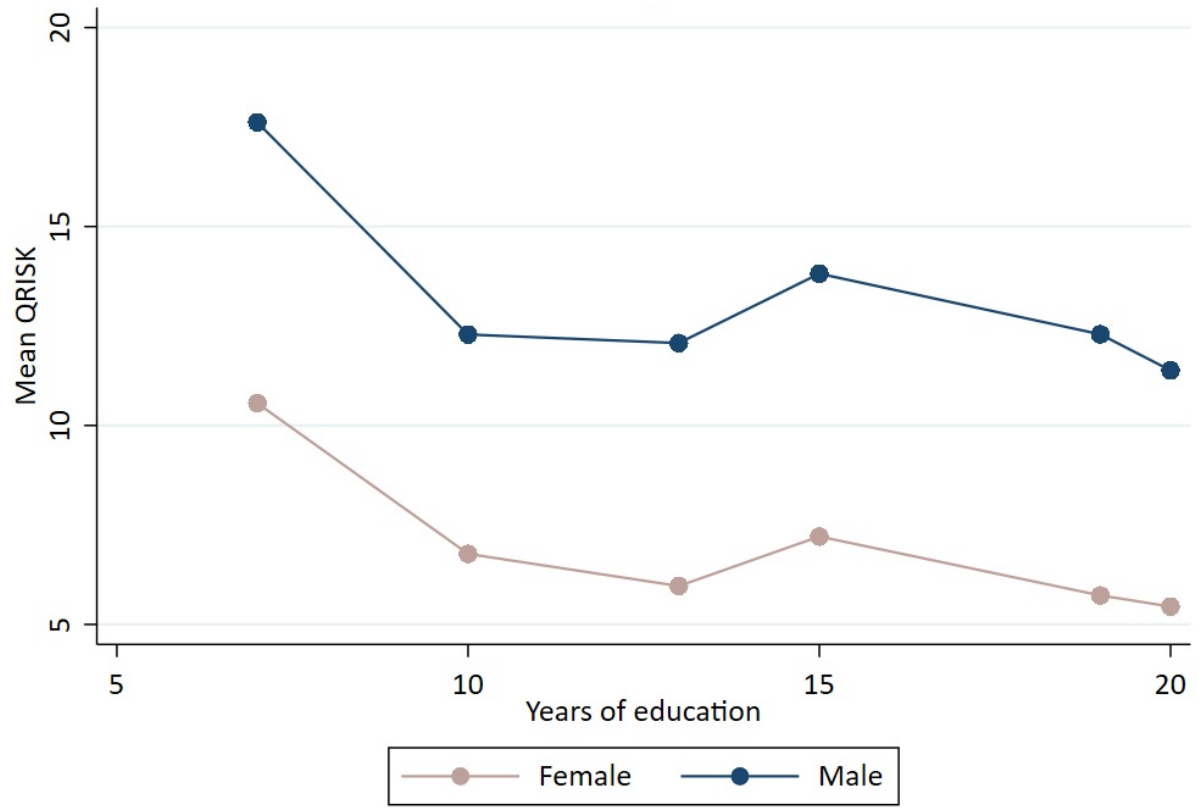

Figure 5.4: Mean value of QRISK3 score on those with complete data, by years of education for females and males

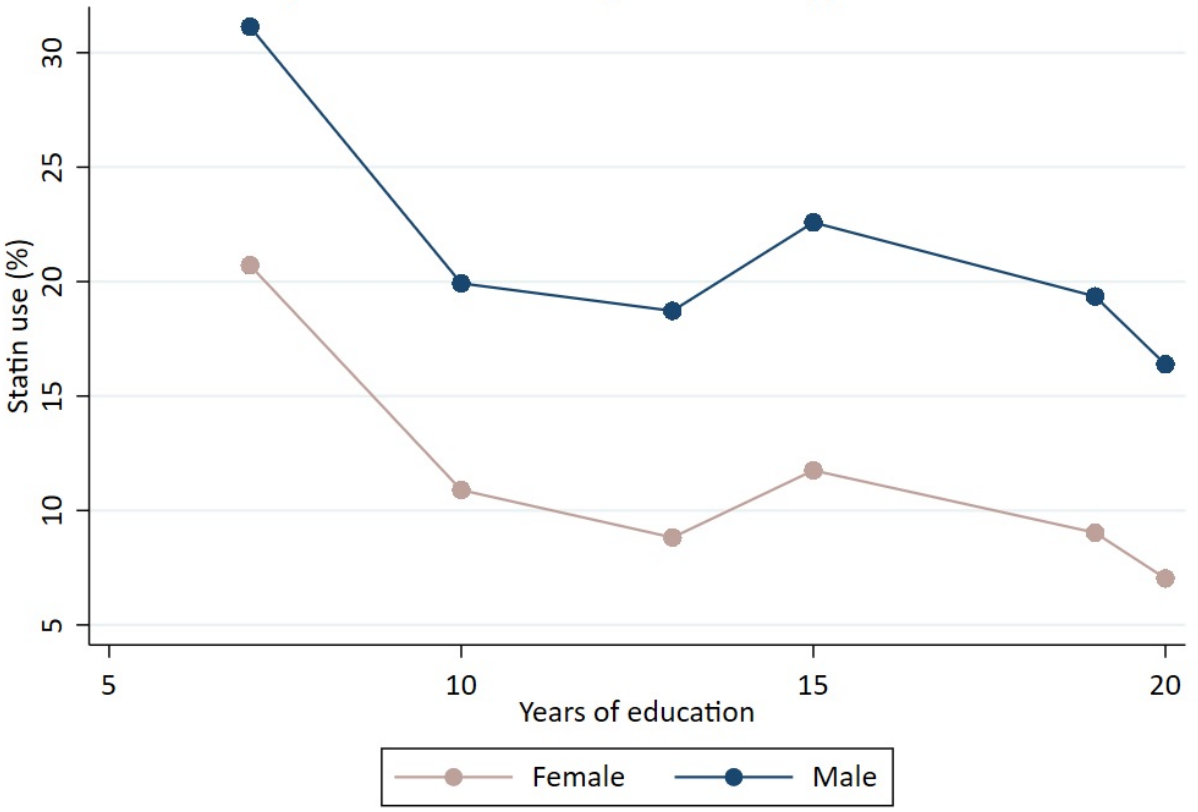

Figure 5.5: Prevalence of statin prescribing by years of education in females and males in individuals with complete data

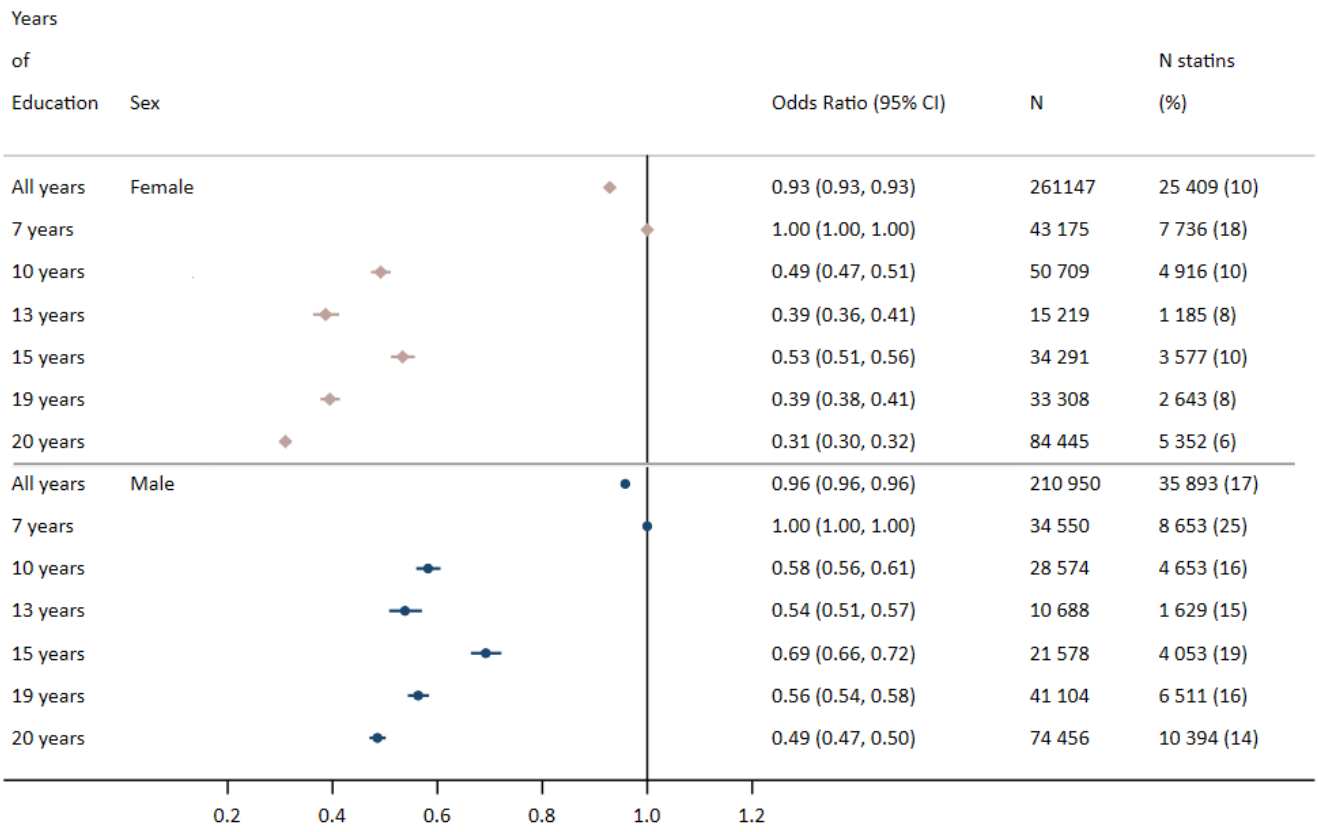

Figure 5.6: Odds ratio of statin use per year unit increase in educational attainment (all years) and per strata of educational attainment

CI = confidence interval

5.6.4 Interaction between education and QRISK3 score in relation to statin prescribing

In both females and males, there was evidence of an interaction between QRISK3 score and years of education on statin use, such that for the same increase in QRISK3 score, the likelihood of statin use increased more for those of high educational attainment. In females, per unit increase in QRISK3, the OR for reporting statin use in those with the greatest years of education (20 years, equivalent to obtaining a degree) was 1.14 (95% CI: 1.14 to 1.15) compared with an OR of 1.07 (95% CI: 1.07 to 1.07) for those with the least years of education (7 years, equivalent to leaving school with no formal qualifications) (Figure 5.3). In males, the OR for statin use per unit increase in QRISK3 score in those with 20 years of education was 1.08 (95% CI: 1.08 to 1.08) compared with an OR of 1.04 (95% CI: 1.04 to 1.05) for those with 7 years of education (Figure 5.3). The P value for interaction in females was $1.385x10^{-85}$ and in males the P value for interaction was $1.551x10^{-48}$.

5.6.5 Secondary analyses

Among individuals prescribed with either atorvastatin or simvastatin, those with higher QRISK3 scores were more likely to have been prescribed the more effective Atorvastatin. The OR for a one-unit higher QRISK3 and reporting Atorvastatin use was, 1.02 (95%CI: 1.02 to 1.03) (Table 5.7). This was similar in males; OR: 1.02 (95% CI: 1.01 to 1.02). Females, but not males, were less likely to have been prescribed Atorvastatin if they had more years of education; e.g. the OR for Atorvastatin prescription for 20 years of education versus 7 years of education was 0.92 in females (95% CI 0.83 to 1.01) and 1.02 in males (95% CI 0.94 to 1.11). There was little evidence of an interaction between QRISK3 score and educational attainment on statin type in females and males (P value for interaction in females = 0.4; P value for interaction in males = 0.9) (Figure 5.7).

Table 5.7: Odds ratio of Atorvastatin use compared with Simvastatin (baseline) use per unit increase in QRISK3 score and by strata of educational attainment (not controlling for QRISK3 score)

Exposure		Females		Males	
		Complete Case Odds ratio (95% CI) (N = 18 180)	**Imputed sample Odds ratio (95% CI)** (N = 23 538)	**Complete Case Odds ratio (95% CI)** (N = 26 633)	**Imputed sample Odds ratio (95% CI)** (N = 33 499)
QRISK3		1.02 (1.02, 1.03)	1.02 (1.02, 1.03)	1.02 (1.01, 1.02)	1.02 (1.01, 1.02)
Education	All years	1.00 (0.99, 1.00)	0.99 (0.99, 1.00)	1.00 (1.00, 1.01)	1.00 (1.00, 1.01)
	7 years	Baseline		Baseline	
	10 years	1.02 (0.92, 1.14)	0.99 (0.90, 1.09)	1.02 (0.92, 1.14)	0.99 (0.90, 1.09)
	13 years	1.14 (0.95, 1.37)	1.08 (0.92, 1.26)	0.97 (0.83, 1.14)	1.00 (0.87, 1.15)
	15 years	1.14 (1.01, 1.28)	1.07 (0.97, 1.19)	1.00 (0.89, 1.12)	0.98 (0.89, 1.09)
	19 years	0.98 (0.85, 1.12)	0.93 (0.82, 1.05)	1.02 (0.93, 1.13)	0.99 (0.91, 1.08)
	20 years	0.92 (0.83, 1.03)	0.92 (0.83, 1.01)	1.06 (0.97, 1.15)	1.02 (0.95, 1.11)

Note: Atorvastatin is generally regarded as more efficacious than Simvastatin. Simvastatin is available to purchase over the counter

CI = confidence interval

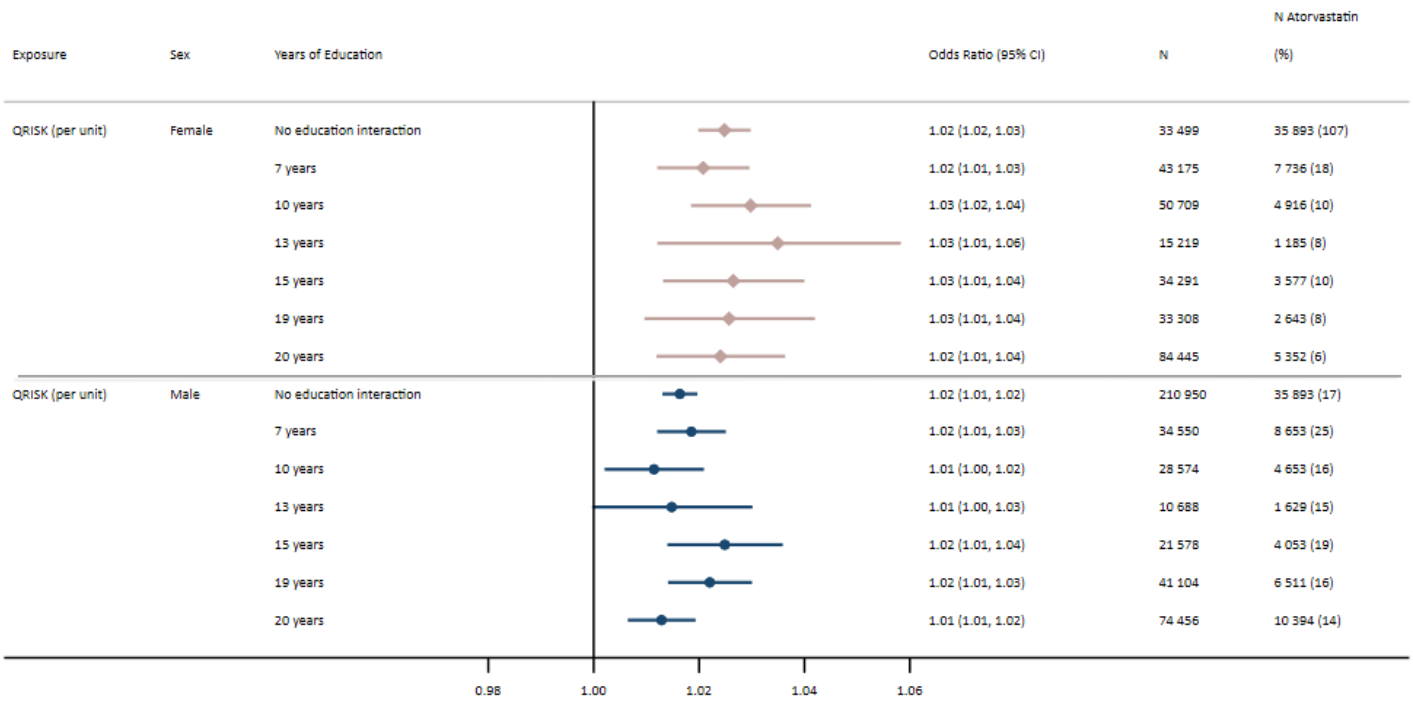

Figure 5.7: Odds ratio of Atorvastatin prescribing compared to Simvastatin, per unit increase in QRISK3 score with no education interaction and stratified by years of education in females and males to test for evidence of an interaction

P value for interaction in females = 0.441 and males = 0.872

CI = confidence interval

When interaction analyses were replicated using eligible participants with linked primary care data using i) baseline measures of QRISK3 and self-report statin use, ii) baseline measures of QRISK3 with statin use validated by a prescription and iii) QRISK or QRISK2 score recorded in primary care data with a statin prescription, the evidence for interaction between QRISK3 and educational attainment on statin use remained in females (Figure 5.8 and Figure 5.9). In males, the interaction between baseline QRISK3 scores and educational attainment on self-report statin and validated prescription remained. However, there was less evidence of an interaction between the primary care recorded QRISK scores and educational attainment on statin prescriptions (P=0.09), although the direction of effect was similar where males with 20 years of education were more likely to be prescribed statins (OR: 1.08; 95% CI: 1.07 to 1.10) than those with 7 years of education (OR: 1.05; 95% CI: 1.03 to 1.08) (Figure 5.8 and Figure 5.9).

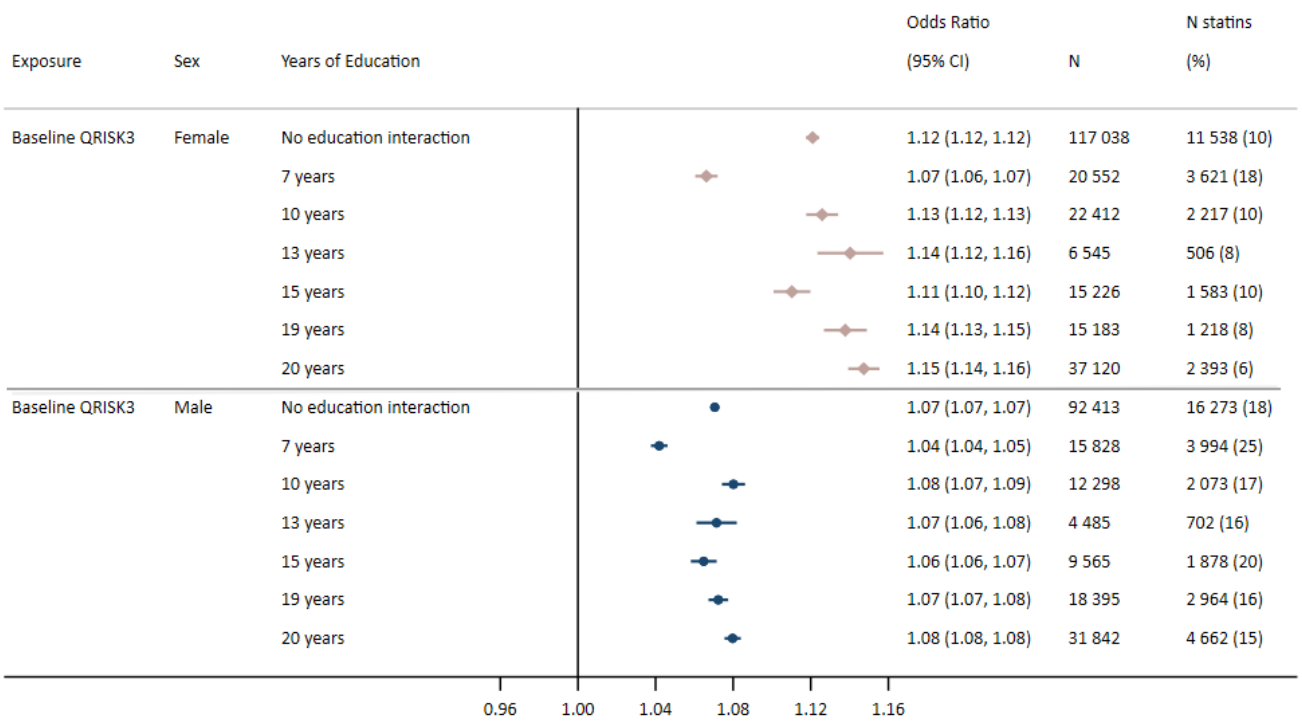

Figure 5.8: Odds ratio of self-report statin use per unit increase in baseline QRISK3 score with no education interaction and stratified by years of education to test for evidence of an interaction in females and males with linked primary care data

P value for interaction in females = $4.76x10^{-48}$ and males = $4.25x10^{-21}$

CI = confidence interval

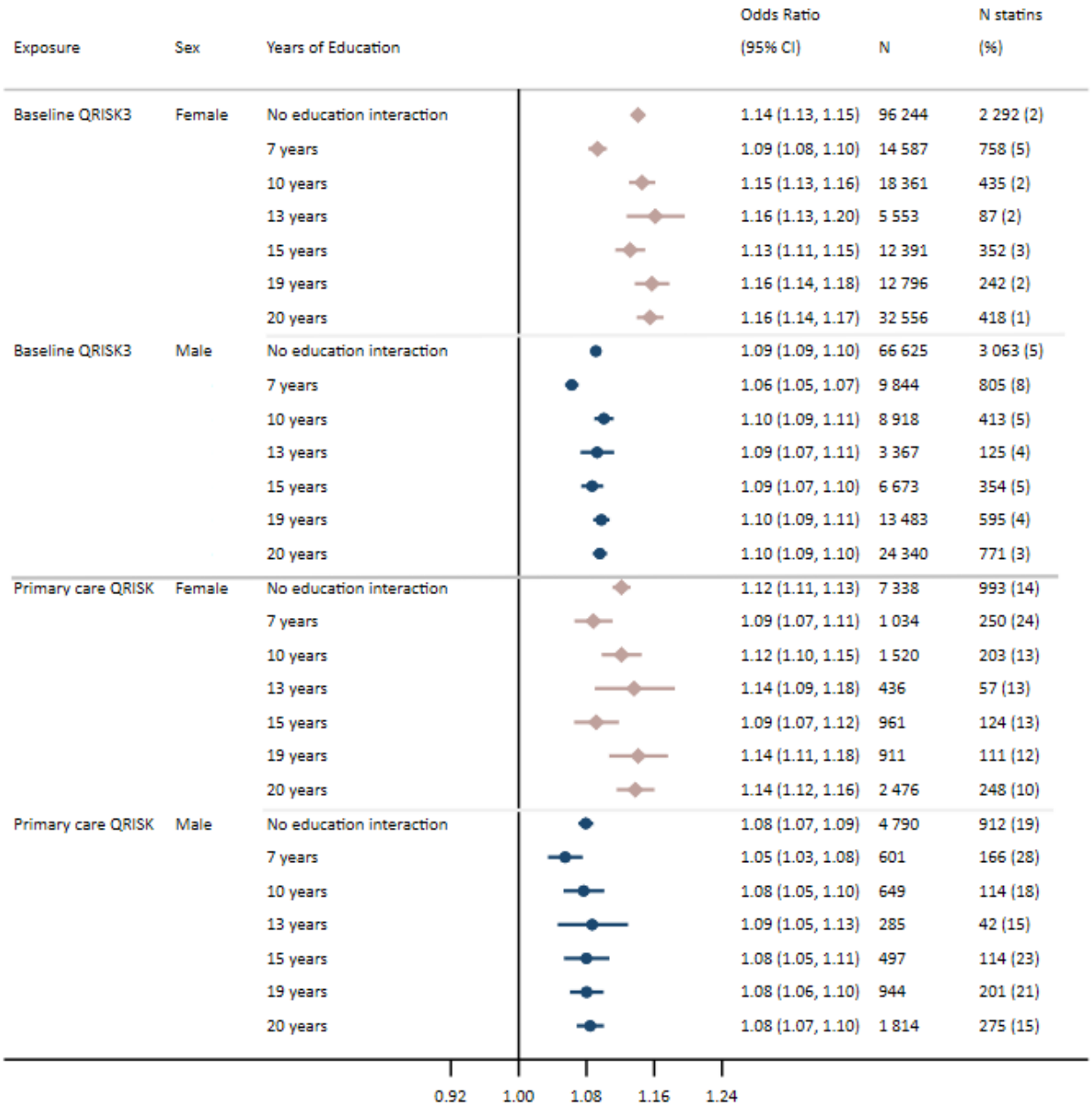

Figure 5.9: Odds ratio of statin use recorded in primary care prescription data per unit increase in A) baseline QRISK3 score and B) QRISK or QRISK2 score recorded in primary care, in females and males. Analyses stratified by years of education provide an estimate of interaction on the multiplicative scale

Baseline QRISK3: P value for interaction in females = 4.27×10^{-10} and males = 3.26×10^{-7}

QRISK score recorded in primary care: P value for interaction in females = 0.034 and males = 0.091

CI = confidence interval

In analyses on the additive scale, there was evidence of an interaction between QRISK3 score and education in both females and males, although the strength of the interaction was smaller compared with analyses on the multiplicative scale, particularly in females (Figure 5.10).

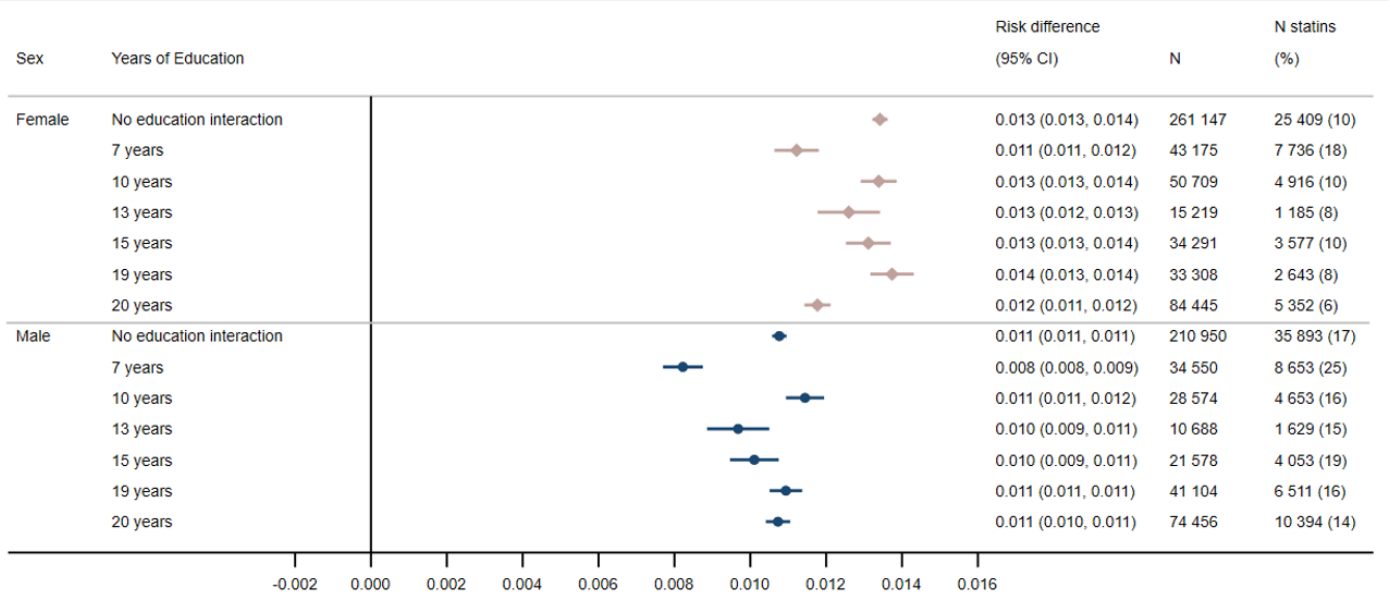

Figure 5.10: Odds ratio of self-report statin use per unit increase in baseline QRISK3 score with no education interaction and stratified by years of education to test for evidence of an interaction in females and males with linked primary care data

P value for interaction in females = 0.064 and males = $9.41x10^{-7}$

CI = confidence interval

In the complete case sample, there was evidence of an interaction between QRISK3 and education in both males and females considering reported statin use as the outcome, where the P value for interaction in females was $4.36x10^{-69}$ and in males it was $3.06x10^{-37}$. However, there was little evidence of an interaction between QRISK3 and education on statin type (Table 5.8).

Pairwise correlation between the baseline derived QRISK3 score and QRISK3 scores derived excluding i) systolic blood pressure variability estimated from the difference between two baseline measures and ii) self-report of any CVD in a mother, father or sibling, were high (all >0.97) (Table 5.9).

Table 5.8: Odds ratio of i) statin use and ii) Atorvastatin use compared with Simvastatin (baseline) use per unit increase in QRISK3 score stratified by educational attainment in the complete case sample to test for evidence of an interaction

Outcome	Years of education	Females		Males	
		Complete Case Odds ratio (95% CI) (N = 201 532)	**P Value for interaction**	**Complete Case Odds ratio (95% CI)** (N = 167 189)	**P Value for interaction**
Statins (self-report)	7	1.07 (1.06, 1.07)	$4.36x10^{-69}$	1.04 (1.04, 1.05)	$3.06x10^{-37}$
	10	1.12 (1.12, 1.13)		1.08 (1.07, 1.08)	
	13	1.13 (1.12, 1.14)		1.06 (1.06, 1.07)	
	15	1.12 (1.11, 1.13)		1.06 (1.06, 1.07)	
	19	1.14 (1.13, 1.15)		1.07 (1.07, 1.08)	
	20	1.14 (1.14, 1.15)		1.08 (1.08, 1.08)	
Statin type (atorvastatin vs simvastatin)	7	1.02 (1.01, 1.03)	0.707	1.02 (1.01, 1.03)	0.783
	10	1.03 (1.01, 1.04)		1.01 (1.00, 1.02)	
	13	1.04 (1.01, 1.07)		1.01 (1.00, 1.03)	
	15	1.02 (1.01, 1.04)		1.03 (1.02, 1.04)	
	19	1.01 (1.00, 1.03)		1.02 (1.01, 1.03)	
	20	1.01 (1.00, 1.02)		1.01 (1.00, 1.02)	

CI = confidence interval

Table 5.9: Pairwise correlation for QRISK3 scores derived from baseline measures in UK Biobank including all variables and excluding i) family history of CVD and iii) systolic blood pressure variability

QRISK3 score	Pairwise correlation with complete score
Female	
Excluding reported family history of any cardiovascular disease at any age	0.9799
Excluding systolic blood pressure from two baseline measures of systolic blood pressure	0.9991
Male	
Excluding reported family history of any cardiovascular disease at any age	0.9736
Excluding systolic blood pressure from two baseline measures of systolic blood pressure	0.9984

5.7 Discussion

Despite there being a higher prevalence of statin prescribing overall in those with lower levels of education, at a given level of QRISK3 score as a measure of clinical assessment of cardiovascular risk, less educated individuals were less likely to receive statin treatment compared to more highly educated individuals.

5.7.1 Results in context

Lifestyle and behavioural factors, such as BMI, diet, smoking, risky drinking and exercise have previously been implicated as mediators of the association between education and CVD (12-14, 112-117). Indeed, the higher overall prevalence of statin use in lower educated individuals is likely due to the greater prevalence of these intermediate risk factors, compared with those of greater education (114, 117, 287). However, much of the association between education and CVD remains unexplained. The results presented in this analysis suggest that access to preventative medication for CVD may be contributing to persisting socioeconomic inequalities.

It has previously been reported that inequalities exist in favour of those with higher SEP when accessing preventative healthcare (357). In the UK, National Health Service (NHS) health checks are offered to all residents aged between 40 and 74 without pre-existing conditions every 5 years, with the aim of preventing a number of diseases including CVD (such as by calculating QRISK scores), kidney disease and dementia (358). In a recent systematic review by Bunten and colleagues, seven studies were identified that indicated uptake of these health checks is lower in more socioeconomically deprived groups (359). Additionally one study included in this systematic review identified a trend towards lower uptake in smokers; an important risk factor for CVD that is also socially patterned (359, 360). Similar findings were also reported by Wilson and colleagues (361). These reasons for non-uptake of health checks, in combination with the inequalities identified in this study, indicate that methods to improve engagement with NHS health checks and preventative screening methods may reduce inequalities in cardiovascular outcomes.

Indeed, differences in health seeking behaviours may be driving some of the inequalities in statin use identified in this study. However, when interaction analyses were repeated using QRISK or QRISK2 scores recorded in primary care data and primary care records of prospective statin prescriptions, these inequalities remained. Therefore, attendance to primary care clinics cannot be the sole driver of these inequalities.

The literature is mixed in the direction to which inequalities in access to statins exist, where some studies suggest that individuals with lower SEP are less likely to be prescribed statins (87, 145-147) and other studies finding the opposite or no differences (36, 141-144). Of these previous studies, there was limited consideration for underlying cardiovascular risk in the analyses. Some studies adjusted for cardiovascular comorbidities (87, 143-145) such as cholesterol level, diabetes status or prevalent cardiovascular events. However, only one previous study was identified that comprehensively adjusted for cardiovascular risk (141). Forde and colleagues established risk status via 10-year absolute risk of coronary heart disease determined using the Framingham study (141, 362) and assessed SEP by British civil service grade of employment. In contrast to the results presented here using educational attainment as a measure of SEP, they did not find evidence of inequalities in statin use. The differences in my results compared with Forde and colleagues could be the different measure of SEP used (income vs education) or due to cohort differences, where Forde and colleagues used an occupational cohort study and here, this analysis uses a population-based cohort. Additionally, it has been demonstrated that the QRISK score has a greater predictive power compared with the Framingham score (363). Therefore, these analyses may better account for underlying differences in cardiovascular risk.

Currently, the QRISK3 scores captures the prevalence of key risk factors in individuals, such as BMI, blood pressure and smoking, but these results show that accounting for these factors alone is not enough to address cardiovascular inequalities. Cardiovascular risk scores may need to be adapted to pay greater attention to SEP; something that has been described previously in the literature (364-366) These risk scores should be in principle, easy to use and clear for clinicians, where it has previously been reported that the use of risk scores in general practice is a source of confusion (367).

Despite there being almost 30 000 first instances of statin prescriptions after 1st January 2008 (where QRISK scores were first introduced in 2007), in the primary care data linked with UK Biobank, there were only around 14 000 individuals with a recorded QRISK or QRISK2 scores in the same data. This is higher than in previous research by Finnikin and colleagues, where they identified using primary care records, that only 27% of patients prescribed statins had a recorded QRISK2 score (368). However, the lack of recorded QRISK scores, suggests the decision to prescribe statin treatment may be independent of an objective measure of cardiovascular risk, and potentially prescribed based on more subjective measures by the clinician or the patient.

In individuals with linked primary care data, 14% of eligible participants reported using statins to study nurses, however only 3% of participants had a linked prescription in the three months before and after baseline. These individuals without a linked prescription are likely a combination of individuals who are purchasing statins over the counter, have received a prescription from a private clinician, or are no longer prescribed statins. The majority (91%) of those without a linked prescription reported taking Simvastatin (currently the only statin available as an over the counter medicine). Although the reason these individuals did not have a prescription cannot definitively be discerned, it is possible that accessing statins through non-NHS GPs (i.e. through private practices) or over the counter is further contributing to inequalities in cardiovascular outcomes. There is, to date, little freely available data on the prevalence of purchasing statins over the counter, rather than via attending a primary care clinic. However, data used here suggests it could indeed be highly prevalent in the population.

5.7.2 Strengths and limitations

The major strength of this work is the large sample size and array of data available. Given the age range of participants (45-76 years) reported statin use is highly prevalent (10% in females and 17% in males). Additionally, the linked primary care data for 44% of the eligible sample allowed us to i) validate self-reported statin use and ii) compare different mechanisms through which inequalities may be arising. Where inequalities are present in primary care recorded QRISK scores, inequalities are unlikely to be due to health seeking behaviour and more likely due to factors arising within clinic settings. Conversely, where data is used from UK Biobank baselines assessment, inequalities may be due to either differences in health seeking behaviour (i.e. attending NHS health checks) or factors that arise within the healthcare setting.

Lifestyle and behavioural characteristics, which are incorporated in to the QRISK3 score, are likely to be captured much more accurately and completely in UK Biobank compared with a primary care setting. However, there may be some settings where UK Biobank variables may have been measured differently than they would in primary care (369), such as non-fasting blood biomarker measurements. However, the magnitude to which these measurements differ is unlikely to introduce much bias to estimates of the QRISK3 score. Additionally, selection bias is present in UK Biobank, where participants are generally of a higher SEP and healthier than the general population (16). Those who are of a lower SEP in the UK Biobank potentially differ to those of an equivalent SEP (or level of educational attainment) in the general population, where UK Biobank participants may be more health conscious and health aware.

Therefore, it is possible that the inequalities in the wider population are greater than the inequalities reported here.

Despite the large sample size and wealth of data, a number of assumptions were made when generating the QRISK3 scores. For example, in the QRISK3 algorithm (23), the study authors specify medications should be considered if the individual has two or more prescriptions for each class of medication (e.g. corticosteroid or atypical antipsychotic). The number of prescriptions was not available at baseline and therefore relied on a single self-report measure of medication use. Therefore, medication use may be overestimated in this sample, which would result in an overestimate of the QRISK3 score. Additionally, some measures, such as systolic blood pressure variability and coronary heart disease in a first degree relative under the age of 60, are not available in the UK Biobank data. Although some proxy measures were included which would likely capture these risk factors, this may introduce bias to the QRISK3 estimate in UK Biobank compared with a primary care setting.

In this analysis, primary analyses have been carried out on the multiplicative scale for interaction. Where there is evidence of multiplicative interaction it means the effect of the combined association between education and QRISK3 score on statin use is greater than the product of the individual associations between education and QRISK3 separately on statin use (370). On the additive scale for interaction, the joint effect of the two risk factors is greater than the sum of the individual associations. This additive scale can be considered as more relevant for public health interventions, where . Here, I found evidence of an interaction between education and QRISK3 score on statin use on both the multiplicative and additive scale for an interaction. In practice, this means consideration should be given to both education and QRISK3 score when determining whether a statin prescription should be administered.

The ISCED definitions of educational attainment (years in schooling) can differ with respect to other measures of socioeconomic position. For example, using ISCED definitions, individuals who left school with a vocational qualification are given a high number of years of schooling (19 years) but will typically go into manual labour jobs. This is likely to explain some of the non-linearities in effects stratified by educational attainment.

5.7.3 Clinical implications

The results presented here highlight inequalities in statin use by educational attainment. Given the persisting inequalities in CVD, addressing the contribution of differences in statin prescription provides a clear policy target. The two complimentary data sources used in this

analysis, UK Biobank baseline data and linked primary care data, indicate two potential mechanisms for these inequalities. Firstly, there are likely to be differences in health seeking behaviour such as in attending NHS health checks as previously evidenced in the literature. Secondly, the inequalities present in the primary care data suggest there are important interactions between the healthcare practitioner and patient that result in unequal prescribing of statins.

Healthcare professionals should consider potential biases in prescribing preventative treatments, or in carrying out risk assessments, such as calculating a QRISK score. Additionally, patient preference for treatment may be socially patterned (371). However, addressing these inequalities requires systemic change and different interventions may be required to address the different mechanisms of inequalities. For example, policy makers and healthcare professionals should consider how they can improve the uptake of NHS health checks, where these risk assessments are carried out, in those who are socioeconomically disadvantaged.

5.7.4 Conclusions

These analyses demonstrate that at a given level of cardiovascular risk, people with lower levels of educational attainment are less likely to be prescribed statins than people with higher educational attainment, meaning differences in statin prescribing likely contribute to inequalities in cardiovascular disease. Policies should consider how these inequalities can be minimised.

Chapter 6. Educational attainment as an effect modifier of polygenic scores for cardiovascular risk factors: cross-sectional and prospective analysis of UK Biobank

6.1 Author list and contributions

Alice R Carter[1,2*], Sean Harrison[1,2], Dipender Gill[3-6], Richard Morris[2,7], George Davey Smith[1,2,8], Amy E Taylor[1,2,8], Laura D Howe[1,2†], Neil M Davies[1,2,9†]

[†]LDH and NMD contributed equally

ARC designed the study, cleaned and analysed the data, interpreted the results, wrote and revised the manuscript. SH assisted with data analysis, interpreted the results and critically reviewed and revised the manuscript. DG advised on defining medications, interpreted the results and critically reviewed and revised the manuscript. RM advised on analyses, interpreted the results and critically reviewed and revised the manuscript. GDS, AET, NMD and LDH all designed the study, interpreted the results, critically reviewed and revised the manuscript and provided supervision for the project. NMD and LDH contributed equally and are joint senior authors on this manuscript. ARC and NMD serve as guarantors of the paper. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

6.2 Summary of personal contributions

In this chapter I use data from UK Biobank baseline assessment centres and linked hospital inpatient records to investigate how educational attainment acts as an effect modifier of polygenic scores for a number of cardiovascular risk factors and outcomes.

I was the sole lead author for the work in this chapter. I carried out all analyses following an analysis plan agreed upon by all co-authors and created publication quality figures. I wrote and revised the manuscript in according to comments from co-authors. This manuscript has not yet been published, nor posted to a preprint server.

6.3 Abstract

Background:

The mechanisms relating socioeconomic position to cardiovascular disease is largely unknown. Understanding the interplay between socioeconomic position and genetic predictors of cardiovascular risk in this relationship may improve our understanding of underlying pathways.

Methods:

In 320 120 UK Biobank participants of White British ancestry (mean age = 57, female 54%), I created polygenic scores for nine cardiovascular risk factors or diseases; alcohol consumption, body mass index (BMI), low-density lipoprotein cholesterol (LDL-C), lifetime smoking behaviour, systolic blood pressure, atrial fibrillation, coronary heart disease, type 2 diabetes and stroke. I then estimated the extent to which educational attainment modified genetic susceptibility to these risk factors on the observed trait.

Results:

On the additive scale, higher educational attainment protected against genetic susceptibility to higher BMI, smoking, atrial fibrillation and type 2 diabetes. However, on the same scale, higher educational attainment increased genetic susceptibility to higher LDL-C and higher systolic blood pressure.

On the multiplicative scale, there was evidence that higher educational attainment increased genetic susceptibility to atrial fibrillation and coronary heart disease, but no evidence of effect modification was found for other traits on the multiplicative scale.

Conclusions:

Educational attainment modifies the genetic susceptibility to some cardiovascular risk factors and diseases. The direction of this effect was mixed, suggesting modification of the effect of genetic susceptibility to cardiovascular risk factors or cardiovascular disease by education attainment are unlikely to contribute to the mechanisms driving inequalities in cardiovascular risk.

6.4 Introduction

Cardiovascular disease (CVD) remains the leading cause of death globally (27). Although rates of CVD have reduced in high income countries, individuals who are more socioeconomically deprived remain at the greatest risk of disease (92). Although some cardiovascular outcomes are monogenic in risk, such as familial hypercholesterolaemia (189), most cardiovascular outcomes are complex multifactorial diseases with both environmental and genetic aetiology (30, 158, 372). Therefore, it is plausible that socioeconomic position (SEP) may interact with, or modify, genetic susceptibility for CVD.

Many previous studies have studied gene*environment interactions with single genetic variants, known as a candidate gene approach (203, 373-376). For example, using this approach , Schmidt and colleagues identified an interaction between income and a genetic polymorphisms in the *CDKN2B-AS1* increasing the risk of experiencing coronary artery calcification (373). However, many of these studies have failed to replicate and results have been demonstrated to be spurious (204). Therefore, it is important to i) carry out gene*environment interaction studies in large sample sizes and, where possible, with replication in multiple independent studies, and ii) consider a polygenic approach to gene*environment interaction.

Using a polygenic approach, Tyrrell and colleagues demonstrated in 120 000 UK Biobank participants, that individuals with a higher Townsend deprivation index have an accentuated risk of obesity in genetically susceptible adults (377). Rask-Anderson and colleagues replicated this association in the second release of genetic data for UK Biobank participants (175). However, in the same analysis, they did not find evidence that education modified the effect of genetic BMI risk on observed BMI (175). Amin and colleagues found similar results for the effect of education on BMI susceptibility in a study using data from the UK and Finland (378).

Whilst educational attainment and has been shown to modify the association of cardiovascular risk factors on CVD (92, 379) it is unclear whether educational attainment modifies the effect of genetic susceptibility to a wide range of cardiovascular risk factors. Understanding the gene-environment interplay in relation to education and cardiovascular risk factors may improve our understanding of the mechanisms underlying educational inequalities in cardiovascular disease (380). Here, I ask whether educational attainment modifies the effect of polygenic susceptibility to multiple cardiovascular risk factors. Previous research has often framed this as a gene*environment interaction. I will describe the interplay between education and polygenic

susceptibility to CVD as effect modification, where I hypothesise that education specifically changes the effect of the polygenic score on the phenotype.

6.5 Methods

6.5.1 UK Biobank

The UK Biobank recruited 503 317 adults from around the UK between 2006 and 2010, aged 37 to 73 (16). Participants attended baseline assessment centres involving questionnaires, interviews, anthropometric, physical and genetic measurements (15, 16). In this analysis, I use up to 320 120 individuals of White British ancestry (Figure 6.1).

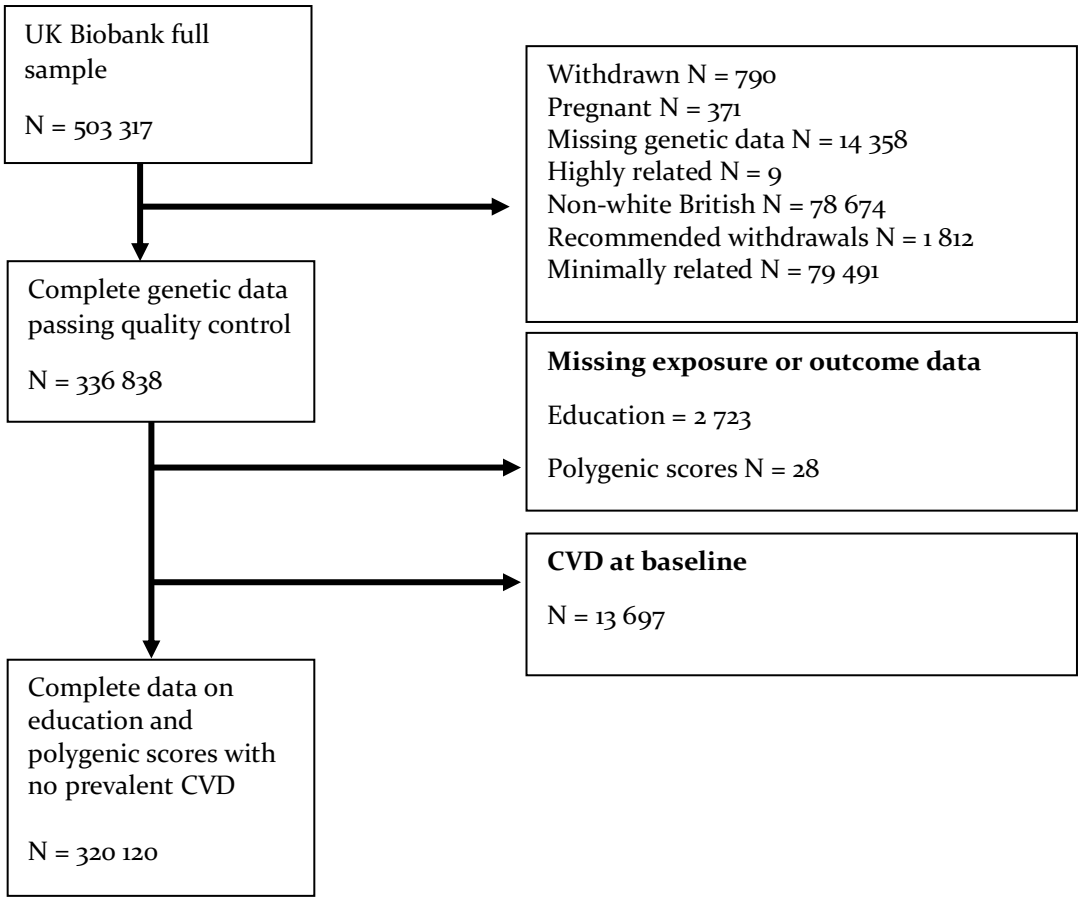

Figure 6.1: Study flow chart of eligible participants

Note: At each stage the same participant could have missing data for multiple variables, therefore overlap is present between the variables. The total excluded may be less than the sum of individuals at each stage.

CVD = cardiovascular disease

6.5.2 **Educational attainment**

UK Biobank participants reported highest qualification achieved at baseline assessment centres, which was converted to the International Standard Classification for Education (ISCED) coding of educational attainment (Table 6.1) (17).

Table 6.1: International Standard for Classification of Education definition of educational attainment

Qualification (As reported in UK Biobank)	ISCED	Years of education	N
College or University degree	5	20	104 037
NVQ or HND or HNC or equivalent	5	19	20 892
Other prof. qual. e.g.: nursing, teaching	4	15	16 481
A levels/AS levels or equivalent	3	13	37 235
O levels/GCSEs or equivalent	2	10	71 424
CSEs or equivalent	2	10	17 551
None of the above	1	7	52 500
Prefer not to answer	Excluded		

6.5.3 **Cardiovascular risk factors and cardiovascular disease**

Cardiovascular risk factors were included in my study if there is causal evidence from either Mendelian randomisation studies or randomised controlled trials that they are a causal risk factor for CVD, and suitable genome wide association study (GWAS) summary statistics available. Additionally, I included a number of cardiovascular diseases for which PGS are available. In total, 9 risk factors or diseases were included in my analyses; 6 risk factors (alcohol consumption (41), body mass index (BMI), diabetes (type 2) (381), low density lipoprotein cholesterol (LDL-C) (382), lifetime smoking behaviour (287, 383), systolic blood pressure (384)) and three diseases (atrial fibrillation, coronary heart disease (CHD) and stroke). Cardiovascular risk factors were measured at baseline assessment centres, whilst incident cases of cardiovascular diseases were determined by linked hospital episode statistics (HES) and Scottish Morbidity records (SMR) (referred to as hospital inpatient records) (see Table 6.2).

Table 6.2: International Classification for Disease codes used in cardiovascular case definition

Diagnosis	ICD9	ICD10
Atrial Fibrillation	42731	I48
Coronary heart disease	4100 - 4149	I20-I25
Stroke	4300 - 4389	I6, G45
Type 2 diabetes	4359	G45

6.5.3.1 Alcohol consumption

Alcohol consumption was defined as the number of drinks consumed per week. At baseline assessment centres, participants were asked to described current drinking status (current, former or never) and estimate their current alcohol intake. Of those reporting a current frequency of at least once or twice a week, they were asked to estimate their current average weekly intake of different alcohol beverages. These were summed together to estimate an average number of drinks per week. Never drinkers and individuals reporting a current intake of "one to three times a month" or less frequently, were assumed to have a weekly intake of 0. This variable has been described in detail previously (385).

Summary statistics from the GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) GWAS of drinks per week were used for the PGS (386). This GWAS included predominantly European participants, excluding participants from UK Biobank to avoid overlapping samples for the discovery and analysis dataset, which can lead to inflated effect estimates.

6.5.3.2 BMI

Baseline measures of height and weight were used to calculate BMI (kg/m^2).

Summary statistics for use in the PGS for BMI came from the Genetic Investigation of Anthropometric Traits (GIANT) Consortium GWAS analysis of 339 224 individuals with European ancestry (290). This is the most recent GWAS of BMI not including UK Biobank, to avoid sample overlap.

6.5.3.3 Low density lipoprotein cholesterol

Non-fasting measures of LDL-C were measured using enzymatic assays (Backman Coulter AU5800). UK Biobank corrected serum data for laboratory dilution effects and were excluded if they did not pass UK Biobank quality control (351).

Summary statistics for use in the PGS came from the Global Lipids Genetics consortium, which included 188 577 males and females of predominantly European ancestry (291).

6.5.3.4 Smoking

A measure of lifetime smoking was constructed in the UK Biobank from self-reported age at initiation, age at cessation and cigarettes per day. From this information, smoking duration and time since cessation were calculated. The lifetime smoking measure further includes a simulated

constant (half-life) which captures the exponentially decreasing effect of cigarettes on health over time. Aspects of smoking behaviour were combined into one score ranging from 0 (for non-smokers) to 4.00 (mean = 0.33, standard deviation = 0.67). Full details of score construction can be found elsewhere (320). The main advantage of using this measure of smoking is that it is a continuous measure, improving statistical power, and it considers all aspects of smoking which may affect health, e.g. duration of smoking and smoking heaviness.

I carried out a split sample GWAS of lifetime smoking in UK Biobank to identify genetic variants associated with lifetime smoking to use in a PGS. I included 318 147 participants with White British ancestry, who were randomly assigned to one of two samples. In each half of the eligible participants, the GWAS was conducted, which was used to derive the PGS in the opposing sample so as to avoid sample overlap which can inflate genetic estimates. This split sample GWAS of lifetime smoking has previously been used in a PGS and described in detail (287). The estimates from each sample were meta-analysed using the *metan* command to create a single estimate (387).

6.5.3.5 Systolic blood pressure

The mean from two resting automated measures of systolic blood pressure, measured using an Omron HEM-7105IT digital blood pressure monitor at baseline assessment centres was used for phenotypic measurements.

I carried out a split sample GWAS of systolic blood pressure in UK Biobank to identify genetic variants for use in the PGS. For individuals who reported taking antihypertensive medication to UK Biobank study nurses, I added 10mm Hg to the phenotypic measurement of systolic blood pressure (314). This GWAS was conducted as described previously for smoking and has been described in detail previously (287). The estimates from each sample were meta-analysed to create a single estimate of effect modification for systolic blood pressure.

6.5.3.6 Atrial Fibrillation

Atrial fibrillation events were ascertained through linkage to mortality data and hospital inpatient records, with cases defined according to ICD-9 and ICD-10 codes (see Table 6.2 for ICD codes used in case definition). Date of diagnoses are provided by hospital inpatient records, which was linked with the date of assessment centre provided by UK Biobank to identify incident and prevalent cases.

Summary statistics for use in the PGS were from a 2012 GWAS of 59 133 individuals (6 707 cases) of European ancestry (388).

6.5.3.7 Coronary heart disease

Coronary heart disease (CHD) events were ascertained through linkage to mortality data and hospital inpatient records, with cases defined according to ICD-9 and ICD-10 codes (see Table 6.2 for ICD codes used in case definition) (292). Date of diagnoses are provided by hospital inpatient records, which was linked with the date of assessment centre provided by UK Biobank to identify incident and prevalent cases.

Summary statistics from the most recent GWAS for CHD not including UK Biobank were used for deriving the PGS (322). A total of 184 305 individuals (60 801 cases) were included in this GWAS of predominantly European descent.

6.5.3.8 Diabetes

Type 2 diabetes was ascertained by linkage to hospital inpatient records (see Table 6.2), with date of diagnosis defined by hospital inpatient records. Additionally, individuals were defined as diabetic if they had reported to UK Biobank study nurses that they had ever had diabetes diagnosed by a doctor (variable 2443). This variable does not distinguish between type 1 and type 2 diabetes, however individuals with a hospital inpatient recordsfor type 1 diabetes were excluded from analyses and in this adult population new diagnoses are more likely to be type 2 diabetes. Individuals were defined as a prevalent case if they reported a diagnosis at baseline assessment centres (variable n_2443_0_0). Incident cases were defined as those who reported a diagnosis at follow up clinics (variable n_2443_1_0 and variable n_2443_2_0), with no previous diagnosis reported (although only a subset of individuals have follow up measures).

Summary statistics of 158 808 European individuals (26 276 Cases) from the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium GWAS of type 2 diabetes were used for the PGS (389).

6.5.3.9 Stroke

Stroke events (all subtypes) were ascertained through linkage to mortality data and hospital inpatient records, with cases defined according to ICD-9 and ICD-10 codes (see Table 6.2) (292). Date of diagnoses are provided by hospital inpatient records, which was linked with the date of assessment centre provided by UK Biobank to identify incident and prevalent cases.

For the PGS, summary statistics for all subtypes of stroke were obtained from the MEGASTROKE consortium, consisting of 521 612 males and females (67 162 cases) of predominantly European ancestry (390).

6.5.4 Deriving polygenic scores

Summary statistics for single nucleotide polymorphisms (SNPs) associated with each cardiovascular trait were downloaded from each of the relevant GWAS. Relevant GWAS were the most recent GWAS for each specified trait excluding UK Biobank participants to avoid bias by sample overlap. The 1000 genomes project was used to find proxy SNPs in LD with SNPs not found in UK Biobank. Pruning of SNPs was carried out using the clump command in PLINK using an r^2 parameter of 0.25 and a physical distance threshold for clumping of 500kB. The pruned SNPs from each GWAS were harmonised with the SNPs from UK Biobank, aligning the effect estimates and alleles. Any SNPs that could not be harmonised, palindromic SNPs or triallelic SNPs were excluded from PGSs. The PGSs were created by multiplying the number of effect alleles for each participant in UK Biobank by the effect estimate of the SNP from summary statistics from each GWAS, then summing across all SNPs associated with each trait. For continuous traits, the PGSs represent a unit increase and for binary traits they represent a log odds ratio increase. All PGSs were standardized for use in analyses so that coefficients reflect a one standard deviation (SD) change.

Polygenic scores were constructed using a range of p-value thresholds $p \leq 5 \times 10^{-8}$ (genome-wide significant), 0.05, and 0.5). As the p-value threshold increases, the variance explained by the PGS typically increases. However, increasing the numbers of SNPs increases the risk of pleiotropy and false positive effects. Main analyses are presented using PGSs derived at the genome-wide significance threshold with other PGSs presented in the supplement. See Appendix 4 Table 1 to Appendix 4 Table 11 for SNPs included in PGSs at the genome-wide significance level.

6.5.5 Exclusion criteria

As studies of effect modification can be biased by reverse causality, individuals who had experienced a cardiovascular event prior to baseline were excluded from analyses. These diagnoses and events were ascertained through linkage to mortality data and hospital inpatient records, with cases defined according to ICD-9 and ICD-10 codes (Table 6.3). Individuals were excluded if they had experienced at least one diagnosis of any of the outcomes considered in analyses (atrial fibrillation, CHD, stroke and type 2 diabetes) or any one of myocardial infarction,

angina, stroke, transient ischaemic attack, peripheral arterial disease or familial hypercholesterolaemia. Exclusions were also made for prevalent cases of type 1 diabetes and chronic kidney disease, which can result in statins being prescribed to prevent cardiovascular diseases (26) and therefore may affect behaviours and subsequently the observed effect modification. The date for each diagnosis is provided by hospital inpatient records, which was linked with the date of assessment centre visit provided by UK Biobank to determine prevalent cases of disease.

Table 6.3: International classification for disease codes used for cardiovascular exclusions

Cardiovascular event	ICD9	ICD10
Myocardial infarction	4100-4109, 4120-4129	I21, I22
Angina	4139	I20
Transient ischaemic attack	4359	G45
Peripheral arterial disease	4439	I73.9
Stroke	4349	I6, G45
Type 1 diabetes	2500- 25011, 25013, 2504-25041, 25043, 2505-25051, 25053, 2506-25061, 25063, 2507-25071, 25073, 2509-25091, 25093	E10
Chronic kidney disease	5383, 5384, 5385	N183, N184, N185
Familial hypercholesterolaemia	2720	I78.0

Quality control of the genetic data was carried out according to the MRC Integrative Epidemiology Unit quality control pipeline, described in full previously (391). In brief, individuals were excluded if their genetic sex differed to their gender reported at the assessment centre or for having aneuploidy of their sex chromosomes (non-XX or -XY chromosomes). Further individuals were excluded for extreme heterozygosity or any missing genetic data. Related individuals were excluded based on an in-house algorithm removing those related (3^{rd} degree or closer) to the greatest number of other participants, until no related pairs were left (391). This exclusion list was derived in-house using an algorithm applied to the list of all the related pairs provided by UK Biobank (3^{rd} degree or closer) (Figure 6.1). In addition, individuals were excluded if they had withdrawn from UK Biobank or were, or may be, pregnant.

Additionally, individuals were excluded if there were any missing data for education, age and sex. Individuals were excluded from specific analyses if they were missing phenotypic measurements of the trait under consideration (see Figure 6.1).

6.5.6 Statistical Analysis

6.5.6.1 Association of educational attainment with outcomes

Multivariable linear regression (adjusting for age and sex) was carried out to estimate the association between educational attainment and cardiovascular risk factors.

6.5.6.2 Association between each polygenic score and observed phenotype

For each of the cardiovascular risk factors or diseases, we estimated the association between each polygenic score and the phenotypic measure of the risk factor or outcome using multivariable linear regression. Analyses were adjusted for age, sex, educational attainment and 40 genetic principal components to control for population structure. For continuous cardiovascular risk factor, measures were standardised, so estimates reflect the mean difference in SD of the phenotype for a one SD higher polygenic score. For binary outcomes, estimates reflect the risk difference or log odds ratio of outcome for a one SD higher polygenic score.

6.5.6.3 Effect modification by educational attainment on polygenic scores for cardiovascular risk

To test for effect modification, the linear model was stratified by years of educational attainment. To estimate the magnitude and direction of the effect modification, an interaction term was included in the linear model (e.g, polygenic score*education [continuous]). Analyses were adjusted for age, sex and 40 genetic principal components. Continuous phenotypic measures were used to limit spurious results, where categorical variables can lead to inflations in the gene-exposure estimates (377). Tests of effect modification were carried out on both the additive and multiplicative scale (370).

6.5.7 Secondary Analyses

All analyses were replicated for polygenic scores at P value thresholds of 0.05 and 0.5.

6.5.8 Data and code availability

The data used in this study has been archived with the UK Biobank study. The analysis code used is available at github.com/alicerosecarter/gxe_cv_riskfactors.

6.6 Results

6.6.1 UK Biobank cohort

Eligible UK Biobank participants (55% female) had a mean age of 57 (standard deviation [SD] = 8.00). A higher proportion of participants (33%) left school after 20 years (equivalent to obtaining a degree), compared with those who left school after 7 years (equivalent to no formal qualifications) (16%) (Table 6.4).

For a P value of $<5x10^{-8}$, the PGSs explained between 0.06% (atrial fibrillation) and 14% (systolic blood pressure) of variance in the phenotypes (Table 6.5).

Table 6.4: Descriptive characteristics of the main analysis sample compared with all individuals in UK Biobank at baseline

Variable	**Analysis sample**			**Full UK Biobank***	
	(N = 320 120)			**(N = 502 156)**	
Continuous variables	N		Mean (SD)	N	Mean (SD)
Age	320 120		56.66 (8.00)	502 156	56.54 (8.09)
Drinks per week	318 300		8.17 (9.05)	497 917	7.79 (9.05)
BMI	319 201		27.3 (4.72)	499 065	27.43 (4.8)
LDL-C	304 700		3.61 (0.86)	468 390	3.56 (0.87)
Systolic blood pressure	292 277		138.16 (18.58)	456 647	137.79 (18.62)
Smoking (lifetime behaviour)	301 684		0.32 (0.66)	318 112	0.34 (0.67)
Categorical variables	**N**		**Frequency (%)**	**N**	**Frequency (%)**
Sex	Female	320 120	175 108 (55)	502 156	273 025 (54)
Years of education	7 years	320 120	52012 (16)	493 033	84648 (17)
	10 years		54899 (17)		82357 (17)
	13 years		17355 (5)		26857 (5)
	15 years		39144 (12)		58271 (12)
	19 years		51418 (16)		77668 (16)
	20 years		105292 (33)		163232 (33)
Atrial fibrillation (incident)	Control	316 912	307352 (97)	495 772	480007 (97)
	Case		9560 (3)		15765 (3)
Coronary heart disease (incident)	Control	317 055	302574 (95)	481 533	458689 (95)
	Case		14481 (5)		22844 (5)
Type 2 diabetes (incident)	Control	316 406	305327 (96)	492 726	472098 (96)
	Case		11079 (4)		20628 (4)
Stroke (incident)	Control	320 120	314191 (98)	497 151	487084 (98)
	Case		5929 (2)		10067 (2)

**Excluding withdrawn participants; BMI = body mass index; LDL-C = low-density lipoprotein cholesterol*

Table 6.5: Number of single nucleotide polymorphisms (SNPs) and variance explained (R^2) by polygenic scores for cardiovascular risk factors and outcomes

	$P=5x10^{-8}$		P=0.05		P=0.5	
	N_{SNPs}	R^2	N_{SNPs}	R^2	N_{SNPs}	R^2
Alcohol (drinks per week)	14	0.0840	72 962	0.0857	449 080	0.0857
Body mass index	127	0.0276	20 542	0.0646	139 582	0.0674
Low density lipoprotein cholesterol	398	0.0540	23 724	0.0144	13 337	0.0136
Systolic blood pressure (sample 1 GWAS)	126	0.1407	77 709	0.1579	373 402	0.1574
Systolic blood pressure (sample 2 GWAS)	112	0.1426	76 557	0.1633	372 715	0.1606
Smoking (sample 1 GWAS)	23	0.0097	67 741	0.0200	391 104	0.0206
Smoking (sample 2 GWAS)	21	0.0113	66 909	0.0222	390 557	0.0223
Atrial fibrillation	3431	0.0061	60 738	0.0655	361 969	0.1064
Coronary heart disease	75	0.0654	49 098	0.0661	345 040	0.0656
Type 2 diabetes	18	0.0411	5 137	0.0366	134 673	0.0350
Stroke	11	0.0474	63 025	0.0480	373 240	0.0472

6.6.2 **Association between educational attainment and cardiovascular risk factors use**

Educational attainment was associated with all cardiovascular risk factors, except for LDL-C (Table 6.6). For all risk factors, except for higher alcohol consumption, higher educational attainment led to a reduction in the mean difference of the trait (Table 6.6).

Table 6.6: Association between educational attainment and observed phenotypic trait adjusted for age and sex

Trait	Mean difference in SD of phenotypic trait per unit increase in education (95% CI)
Alcohol	0.01 (0.01, 0.01)
BMI	-0.02 (-0.02, -0.02)
Low density lipoprotein cholesterol	$5.1x10^{-4}$ ($-2.0x10^{-4}$, $1.2x10^{-3}$)
Smoking (lifetime behaviour)	-0.03 (-0.03, -0.03)
Systolic blood pressure	-0.01 (-0.01, -0.01)
	Risk difference of outcome per unit increase in education (95% CI)
Atrial fibrillation	$-5.1x10^{-4}$ ($-6.3x10^{-4}$, $-3.9x10^{-4}$)
Coronary artery disease	$-1.6x10^{-3}$ ($-1.7x10^{-3}$, $-1.5x10^{-3}$)
Diabetes (type 2)	$-1.7x10^{-3}$ ($-1.8x10^{-3}$, $-1.6x10^{-3}$)
Stroke	$-5.7x10^{-4}$ ($-6.6x10^{-4}$, $-4.7x10^{-4}$)

6.6.3 **Effect modification by educational attainment on genetic susceptibility to cardiovascular risk factors**

For most polygenic scores, there was evidence that educational attainment modified the effect of the polygenic score on either the additive or multiplicative scale. There was little evidence that educational attainment modified genetic susceptibility to alcohol consumption on either scale (Figure 6.1-Figure 6.4 and Table 6.7 and Table 6.8).

On the additive scale, higher educational attainment protected against genetic susceptibility to higher BMI, smoking, atrial fibrillation and type 2 diabetes (Figure 6.2 and Figure 6.3 and Table 6.7). For example, a one SD increase in polygenic score for smoking increased mean difference in lifetime smoking by 0.05 SD (95% CI: 0.04 to 0.06) for those with 7 years education and by 0.03 SD (95% CI: 0.02 to 0.03) for 20 years of education (Figure 6.2 and Figure 6.3 and Table 6.7) ($P_{effect\ modification} = 0.001$).

On the same scale, higher educational attainment increased genetic susceptibility to LDL-C and systolic blood pressure. For example, for those with 7 years of education an increase of one SD in the polygenic score for LDL-C increased mean phenotypic LDL-C by 0.19 SD (95% CI: 0.18 to 0.19). However, for those with 20 years of education, mean LDL-C increased by 0.22 SD (95% CI: 0.22 to 0.23) ($P_{effect\ modification} = 1.12x10^{-4}$) per SD increase in polygenic score (Figure 6.3 and Table 6.7).

On the multiplicative scale, there was evidence that higher educational attainment increased genetic susceptibility to atrial fibrillation and CHD. For example, for a one SD increase in atrial fibrillation polygenic score, the odds ratio for atrial fibrillation in individuals with 7 years of education was 1.59 (95% CI: 1.45 to 1.57) and for 20 years of educational attainment the odds ratio was 1.65 (95% CI: 1.59 to 1.71) ($P_{\text{effect modification}} = 9.03 \times 10^{-8}$) (Figure 6.2 and Figure 6.4 and Table 6.8). There was little evidence of a effect modification by education on the multiplicative scale for all other PGSs.

For all outcomes, the size of the coefficient for effect modification small. Where outcomes were binary, the coefficient was larger on the multiplicative scale, compared with the additive scale. However, for continuous outcomes, the coefficient was larger on the additive scale. For all outcomes, estimates on the multiplicative scale had greater uncertainty (Figure 6.2).

Non-linear effects by strata of educational attainment were observed (Figure 6.3 and Figure 6.4). For example, considering the additive scale between BMI PGS and educational attainment, a one SD increase in PGS increased mean difference in BMI by 0.13 SD (95% CI: 0.12 to 0.14) for people with 7 years education, 0.13 SD (95% CI: 0.12 to 0.14) for 10 years education, 0.14 SD (95% CI: 0.13 to 0.14) for 19 years education and by 0.12 SD (95% CI: 0.11 to 0.12) for 20 years of education (Figure 6.3 and Table 6.7) ($P_{\text{effect modification}} = 0.036$).

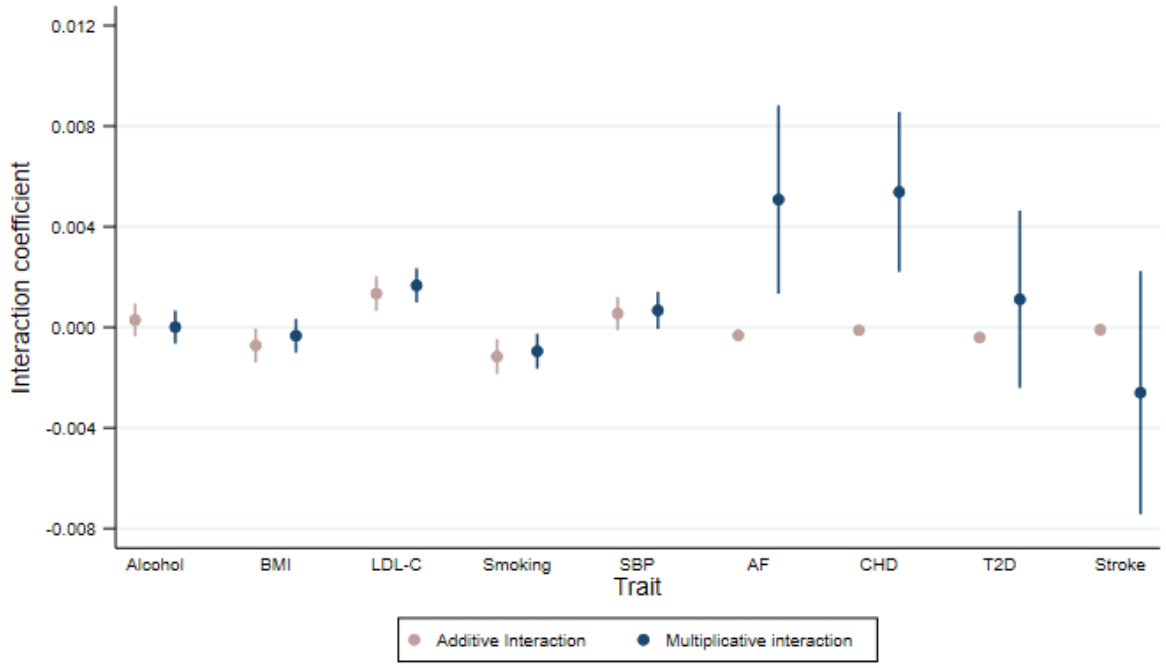

Figure 6.2: Coefficient for educational attainment as an effect modifier of polygenic susceptibility to cardiovascular risk factors or diseases on the additive and multiplicative scale

Analyses adjusted for age, sex and 40 genetic principal components

Alcohol = drinks per weekly BMI = body mass index; LDL-C = Low density lipoprotein cholesterol; smoking = lifetime smoking behaviour; SBP = systolic blood pressure; AF = Atrial fibrillation; CHD = Coronary heart disease; T2D = Type 2 diabetes

Note: coefficients for binary outcomes are on the log odds scale, rather than exponentiated odds ratio scale as in following figures to allow for direct comparisons in the direction of effect modification between the additive and multiplicative scales

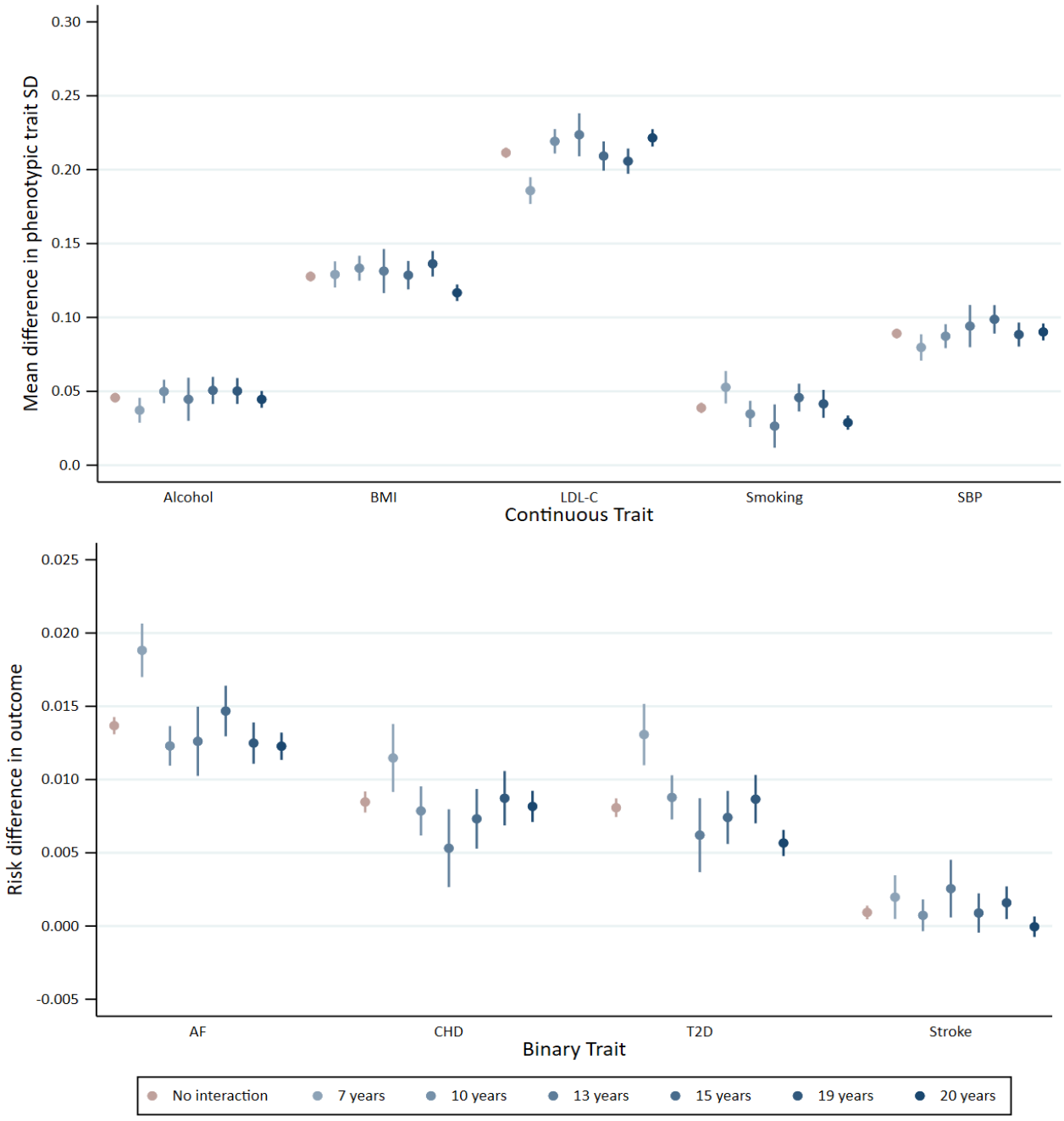

Figure 6.3: Association between polygenic scores for susceptibility to cardiovascular risk and phenotypic measure of each risk factor, stratified by educational attainment demonstrating effect modification on the additive scale

Analyses adjusted for age, sex and 40 genetic principal components

Alcohol (drinks per week) P_{EM} = 0.384; body mass index (BMI) P_{EM} = 0.036; low-density lipoprotein cholesterol (LDL-C) P_{EM} = $1.12x10^{-4}$; lifetime smoking behaviour P_{EM} = 0.001; systolic blood pressure (SBP) P_{EM} = 0.104

Atrial fibrillation (AF) P_{EM} = $9.03x10^{-8}$; coronary heart disease (CHD) P_{EM} = 0.103; type 2 diabetes (T2D) P_{EM} = $3.23x10^{-10}$; stroke P_{EM} = 0.036

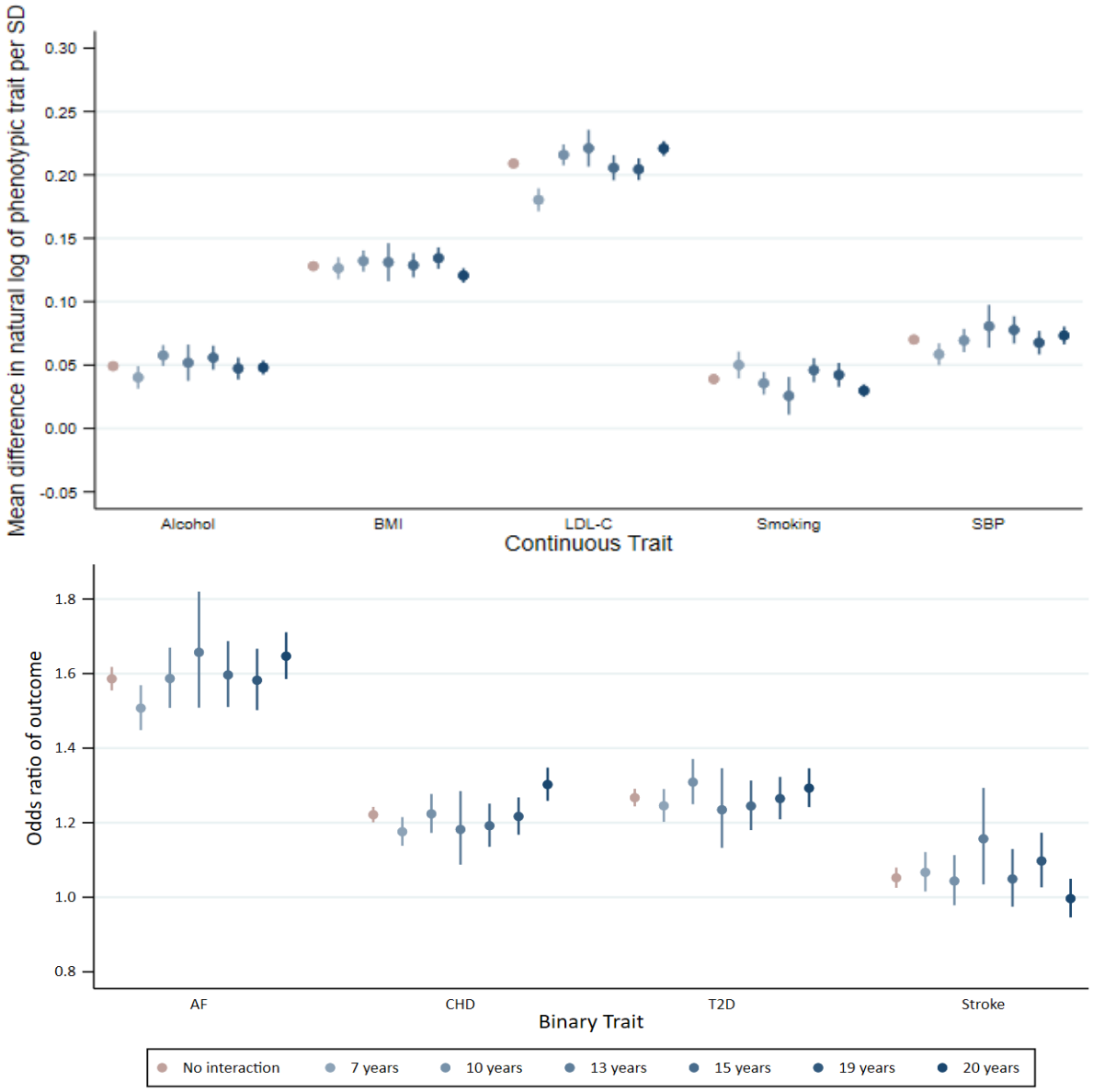

Figure 6.4: Association between polygenic scores for susceptibility to cardiovascular risk and phenotypic measure of each risk factor, stratified by educational attainment demonstrating effect modification on the multiplicative scale

Analyses adjusted for age, sex and 40 genetic principal components

Alcohol (drinks per week) P_{EM} = 0.976; body mass index (BMI) P_{EM} = 0.330; low-density lipoprotein cholesterol (LDL-C) P_{EM} = $1.63x10^{-6}$; lifetime smoking behaviour P_{EM} = 0.008; systolic blood pressure (SBP) P_{EM} = 0.076

Atrial fibrillation (AF) P_{EM} = 0.008; coronary heart disease (CHD) P_{EM} = $8.94x10^{-4}$; type 2 diabetes (T2D) P_{EM} = 0.537; stroke P_{EM} = 0.292

Table 6.7: Association between polygenic scores for susceptibility to continuous cardiovascular risk factors and phenotypic measure of each risk factor, stratified by educational attainment demonstrating effect modification

Trait	Years of education	N	Additive scale		Multiplicative scale	
			Mean difference in SD of phenotypic trait (95% CI)	P value for effect modification	Mean difference in SD of log phenotypic trait (95% CI)	P value for effect modification
Alcohol	All	318 300	0.05 (0.04, 0.05)		0.05 (0.05, 0.05)	
	Education 7	51 509	0.04 (0.03, 0.05)	0.384	0.04 (0.03, 0.05)	0.976
	Education 10	54 567	0.05 (0.04, 0.06)		0.06 (0.05, 0.07)	
	Education 13	17 267	0.04 (0.03, 0.06)		0.05 (0.04, 0.07)	
	Education 15	38 974	0.05 (0.04, 0.06)		0.06 (0.05, 0.07)	
	Education 19	51 095	0.05 (0.04, 0.06)		0.05 (0.04, 0.06)	
	Education 20	104 888	0.04 (0.04, 0.05)		0.05 (0.04, 0.05)	
	Effect modification beta		$2.93x10^{-4}$ ($-3.66x10^{-4}$, $9.51x10^{-4}$)		$9.96x10^{-4}$ ($-6.51x10^{-4}$, $6.71x10^{-4}$)	
BMI	All	319 201	0.13 (0.12, 0.13)		0.13 (0.12, 0.13)	
	Education 7	51 773	0.13 (0.12, 0.14)	0.036	0.13 (0.12, 0.14)	0.330
	Education 10	54 739	0.13 (0.12, 0.14)		0.13 (0.12, 0.14)	
	Education 13	17 319	0.13 (0.12, 0.15)		0.13 (0.12, 0.15)	
	Education 15	39 041	0.13 (0.12, 0.14)		0.13 (0.12, 0.14)	
	Education 19	51 309	0.14 (0.13, 0.14)		0.13 (0.13, 0.14)	
	Education 20	105 020	0.12 (0.11, 0.12)		0.12 (0.11, 0.13)	
	Effect modification beta		$-7.20x10^{-4}$($-1.40x10^{-3}$, $-4.90x10^{-5}$)		$-3.34x10^{-4}$ ($-1.01x10^{-3}$, $3.37x10^{-4}$)	
Low density lipoprotein cholesterol	All	304 700	0.21 (0.21, 0.21)		0.21 (0.21, 0.21)	
	Education 7	49 435	0.19 (0.18, 0.19)	$1.12x10^{-4}$	0.18 (0.17, 0.19)	$1.63x10^{-6}$
	Education 10	52 311	0.22 (0.21, 0.23)		0.22 (0.21, 0.22)	
	Education 13	16 521	0.22 (0.21, 0.24)		0.22 (0.21, 0.24)	
	Education 15	37 257	0.21 (0.2, 0.22)		0.21 (0.2, 0.22)	
	Education 19	48 942	0.21 (0.2, 0.21)		0.2 (0.2, 0.21)	
	Education 20	100 234	0.22 (0.22, 0.23)		0.22 (0.21, 0.23)	
	Effect modification beta		$1.34x10^{-3}$ ($6.62x10^{-4}$, $2.02x10^{-3}$)		$1.67x10^{-3}$ ($9.86x10^{-4}$, $2.35x10^{-3}$)	
Smoking (lifetime behaviour)	All	301 684	0.04 (0.04, 0.04)		0.04 (0.04, 0.04)	
	Education 7	48 470	0.05 (0.04, 0.06)	0.001	0.05 (0.04, 0.06)	0.008
	Education 10	52 292	0.03 (0.03, 0.04)		0.04 (0.03, 0.04)	
	Education 13	16 261	0.03 (0.01, 0.04)		0.03 (0.01, 0.04)	
	Education 15	37 149	0.05 (0.04, 0.06)		0.05 (0.04, 0.06)	
	Education 19	48 639	0.04 (0.03, 0.05)		0.04 (0.03, 0.05)	
	Education 20	98 873	0.03 (0.02, 0.03)		0.03 (0.02, 0.03)	
	Effect modification beta		$-1.08x10^{-3}$ (-0.02, $-3.70x10^{-4}$)		$-9.49x10^{-4}$ ($-1.64x10^{-3}$, $-2.53x10^{-4}$)	
Systolic blood pressure	All	292 277	0.09 (0.09, 0.09)		0.07 (0.07, 0.07)	
	Education 7	46 726	0.08 (0.07, 0.09)	0.104	0.06 (0.05, 0.07)	0.076
	Education 10	50 789	0.09 (0.08, 0.1)		0.07 (0.06, 0.08)	
	Education 13	15 772	0.09 (0.08, 0.11)		0.08 (0.06, 0.1)	
	Education 15	36 033	0.1 (0.09, 0.11)		0.08 (0.07, 0.09)	
	Education 19	47 177	0.09 (0.08, 0.1)		0.07 (0.06, 0.08)	
	Education 20	95 780	0.09 (0.08, 0.1)		0.07 (0.07, 0.08)	
	Effect modification beta		$6.97x10^{-4}$ ($1.03x10^{-5}$, $1.38x10^{-3}$)		$6.74x10^{-4}$ ($-7.03x10^{-5}$, $1.42x10^{-3}$)	

Table 6.8: Association between polygenic scores for susceptibility to cardiovascular risk factors and diseases and phenotypic measure of each risk factor or disease, stratified by educational attainment demonstrating effect modification

Trait	Years of education	N	N cases	Additive scale		Multiplicative scale	
				Risk difference (95% CI)	P value for effect modification	OR (95% CI)	P value for effect modification
Atrial fibrillation	All	316 912	9 560	0.0137 (0.0131, 0.0143)		1.59 (1.55, 1.62)	
	Education 7	51 246	2 438	0.0188 (0.017, 0.0206)	$9.03x10^{-08}$	1.51 (1.45, 1.57)	0.008
	Education 10	54 460	1 466	0.0123 (0.0109, 0.0136)		1.59 (1.51, 1.67)	
	Education 13	17 213	446	0.0126 (0.0102, 0.015)		1.66 (1.51, 1.82)	
	Education 15	38 694	1 232	0.0147 (0.0129, 0.0164)		1.60 (1.51, 1.69)	
	Education 19	50 942	1 396	0.0125 (0.0111, 0.0139)		1.58 (1.5, 1.67)	
	Education 20	104 357	2 582	0.0123 (0.0113, 0.0132)		1.65 (1.59, 1.71)	
	Effect modification beta			$-3.20x10^{-4}$ ($-4.30x10^{-4}$, $-2.00x10^{-4}$)		1.00 (1.00, 1.01)	
Coronary heart disease	All	317 055	14 481	0.0085 (0.0077, 0.0092)		1.22 (1.2, 1.24)	
	Education 7	51 061	3 989	0.0115 (0.0091, 0.0138)	0.103	1.18 (1.14, 1.21)	0.001
	Education 10	54 483	2 292	0.0079 (0.0062, 0.0095)		1.22 (1.17, 1.28)	
	Education 13	17 220	581	0.0053 (0.0027, 0.008)		1.18 (1.09, 1.28)	
	Education 15	38 740	1 733	0.0073 (0.0053, 0.0094)		1.19 (1.14, 1.25)	
	Education 19	50 912	2 477	0.0087 (0.0069, 0.0106)		1.22 (1.17, 1.27)	
	Education 20	104 639	3 409	0.0082 (0.0071, 0.0092)		1.30 (1.26, 1.35)	
	Effect modification beta			$-1.20x10^{-4}$ ($-2.60x10^{-4}$, $2.39x10^{-5}$)		1.00 (1.00, 1.01)	
Diabetes (Type 2)	All	316 406	11 079	0.0081 (0.0074, 0.0087)		1.27 (1.24, 1.29)	
	Education 7	50 904	3 175	0.0131 (0.011, 0.0152)	$3.23x10^{-10}$	1.25 (1.2, 1.29)	0.537
	Education 10	54 261	1 809	0.0088 (0.0073, 0.0103)		1.31 (1.25, 1.37)	
	Education 13	17 190	512	0.0062 (0.0037, 0.0087)		1.23 (1.13, 1.35)	
	Education 15	38 683	1 336	0.0074 (0.0056, 0.0092)		1.24 (1.18, 1.31)	
	Education 19	50 814	1 914	0.0087 (0.007, 0.0103)		1.26 (1.21, 1.32)	
	Education 20	104 554	2 333	0.0057 (0.0048, 0.0066)		1.29 (1.24, 1.35)	
	Effect modification beta			$-4.00x10^{-4}$ ($-5.30x10^{-4}$, $-2.80x10^{-4}$)		1.00 (1.00, 1.00)	
Stroke	NONE	320 120	5 929	0.0009 (0.0005, 0.0014)		1.05 (1.03, 1.08)	
	Education 7	52 012	1 620	0.002 (0.0005, 0.0035)	0.036	1.07 (1.02, 1.12)	0.292
	Education 10	54 899	948	0.0007 (-0.0004, 0.0018)		1.04 (0.98, 1.11)	
	Education 13	17 355	311	0.0026 (0.0006, 0.0045)		1.16 (1.03, 1.29)	
	Education 15	39 144	731	0.0009 (-0.0005, 0.0022)		1.05 (0.97, 1.13)	
	Education 19	51 418	874	0.0016 (0.0005, 0.0027)		1.10 (1.03, 1.17)	
	Education 20	105 292	1 445	$-4.83x10^{-5}$ (-0.0007, 0.0007)		1.00 (0.95, 1.05)	
	Effect modification beta			$-9.80x10^{-5}$ ($-1.90x10^{-4}$, $-6.40x10^{-6}$)		1.00 (0.99, 1.00)	

OR = odds ratio; SD = standard deviation; CI= confidence interval

6.6.4 **Secondary analyses**

Analyses using more liberal P-value thresholds to generate the PGS were broadly consistent with the main genome-wide results. Similar directions of effect and magnitudes of effect modification were observed, for example on the additive scale higher educational attainment protected against genetic susceptibility to BMI and lifetime smoking behaviour at the P-value threshold P=0.05. However, at the P=0.5 threshold, there was no longer evidence of an effect modification by education with BMI. Consistent with results using the genome-wide significant PGS, genetic susceptibility to LDL-C and systolic blood pressure were increased for PGSs derived using both the P-value threshold P=0.05 and P=0.5 (Table 6.9). Similar associations were observed for atrial fibrillation and coronary heart disease where a one unit increase in educational attainment increased susceptibility to these traits (Table 6.9).

Table 6.9: Education as an effect modifier of genetic susceptibility to cardiovascular risk factor on observed phenotypic cardiovascular risk factor for continuous traits (per SD), on the additive scale using polygenic scores at a range of P value thresholds

Exposure	Educational attainment	N	P=0.05				P=0.5			
			Additive scale		Multiplicative scale		Additive scale		Multiplicative scale	
			Mean difference in SD of phenotypic trait (95%CI)	P value	Mean difference in SD of log phenotypic trait (95%CI)	P value	Mean difference in SD of phenotypic trait (95%CI)	P value	Mean difference in SD of log phenotypic trait (95%CI)	P value
Alcohol	All years	318,300	0.08 (0.07, 0.08)		0.06 (0.06, 0.07)		0.08 (0.07, 0.08)		0.07 (0.06, 0.07)	
	Education 7	51,509	0.08 (0.06, 0.09)	0.694	0.06 (0.05, 0.07)	0.108	0.06 (0.05, 0.07)	0.669	0.06 (0.06, 0.07)	0.04
	Education 10	54,567	0.07 (0.06, 0.08)		0.07 (0.06, 0.08)		0.07 (0.06, 0.08)		0.07 (0.06, 0.08)	
	Education 13	17,267	0.08 (0.06, 0.09)		0.08 (0.06, 0.09)		0.07 (0.06, 0.09)		0.07 (0.06, 0.09)	
	Education 15	38,974	0.06 (0.05, 0.07)		0.06 (0.06, 0.07)		0.07 (0.06, 0.08)		0.07 (0.06, 0.08)	
	Education 19	51,095	0.06 (0.05, 0.06)		0.06 (0.05, 0.07)		0.06 (0.05, 0.07)		0.06 (0.05, 0.07)	
	Education 20	104,888	0.06 (0.06, 0.07)		0.06 (0.05, 0.07)		0.06 (0.06, 0.07)		0.06 (0.05, 0.06)	
	Effect modification beta		$-1.31x10^{-4}$ ($-7.85x10^{-4}$, $5.23x10^{4}$)		-5.39x10-4 (-1.19x10-3, 1.18x10-4)		$-1.43x10^{-4}$ ($-7.95x10^{-4}$, $5.10x10^{-4}$)		$-6.68x10^{-4}$ ($-1.32x10^{-3}$, $-1.25x10^{-5}$)	
Body mass index	All years	319,201	0.23 (0.23, 0.23)		0.23 (0.23, 0.23)		0.23 (0.23, 0.24)		0.24 (0.23, 0.24)	
	Education 7	51,773	0.22 (0.22, 0.23)	0.005	0.22 (0.21, 0.23)	0.453	0.23 (0.22, 0.24)	0.215	0.22 (0.22, 0.23)	0.396
	Education 10	54,739	0.24 (0.23, 0.25)		0.24 (0.23, 0.25)		0.24 (0.23, 0.25)		0.24 (0.23, 0.25)	
	Education 13	17,319	0.24 (0.22, 0.25)		0.24 (0.22, 0.25)		0.25 (0.23, 0.26)		0.25 (0.23, 0.26)	
	Education 15	39,041	0.23 (0.22, 0.24)		0.24 (0.23, 0.24)		0.23 (0.23, 0.24)		0.24 (0.23, 0.25)	
	Education 19	51,309	0.23 (0.23, 0.24)		0.23 (0.22, 0.24)		0.24 (0.23, 0.25)		0.24 (0.23, 0.25)	
	Education 20	105,020	0.21 (0.2, 0.22)		0.22 (0.21, 0.22)		0.22 (0.22, 0.23)		0.23 (0.22, 0.23)	
	Effect modification beta		$-9.56x10^{-4}$ ($-1.62x10^{-3}$, $-2.96x10^{-4}$)		-2.52x10-4 (-9.10x10-4, 4.06x10-4)		$-4.17x10^{-4}$ ($-1.07x10^{-3}$, $2.42x10^{-4}$)		$2.85x10^{-4}$ ($-3.72x10^{-4}$, $9.42x10^{-4}$)	
Low density lipoprotein cholesterol	All years	304,700	0.07 (0.07, 0.07)		0.07 (0.07, 0.07)		0.07 (0.06, 0.07)		0.06 (0.06, 0.07)	
	Education 7	49,435	0.06 (0.05, 0.07)	0.148	0.06 (0.05, 0.07)	0.056	0.05 (0.04, 0.06)	0.072	0.05 (0.04, 0.06)	0.033
	Education 10	52,311	0.08 (0.07, 0.09)		0.08 (0.07, 0.09)		0.07 (0.06, 0.08)		0.07 (0.06, 0.08)	
	Education 13	16,521	0.07 (0.05, 0.08)		0.07 (0.05, 0.08)		0.07 (0.05, 0.08)		0.06 (0.05, 0.08)	
	Education 15	37,257	0.07 (0.06, 0.08)		0.07 (0.06, 0.08)		0.06 (0.05, 0.07)		0.05 (0.04, 0.06)	
	Education 19	48,942	0.07 (0.06, 0.08)		0.07 (0.06, 0.08)		0.07 (0.06, 0.08)		0.06 (0.06, 0.07)	
	Education 20	100,234	0.08 (0.07, 0.08)		0.07 (0.07, 0.08)		0.07 (0.06, 0.08)		0.07 (0.06, 0.07)	
	Effect modification beta		$5.12x10^{-4}$ ($-1.82x10^{-4}$, $1.21x10^{-3}$)		6.79x10-4 (-1.61x10-5, 1.37x10-3)		$6.38x10^{-4}$ ($-5.80x10^{-5}$, $1.33x10^{-3}$)		$7.59x10^{-4}$ ($6.21x10^{-5}$, $1.46x10^{-3}$)	

Smoking (lifetime behaviour)	All years	301,684	0.13 (0.13, 0.14)		0.13 (0.13, 0.14)		0.15 (0.14, 0.15)		0.15 (0.15, 0.16)	
	Education 7	48,470	0.18 (0.17, 0.2)	$2.16x10^{-52}$	0.18 (0.16, 0.19)	$6.10x10^{-40}$	0.2 (0.18, 0.21)	$8.85x10^{-50}$	0.19 (0.18, 0.21)	$9.83x10^{-38}$
	Education 10	52,292	0.14 (0.12, 0.15)		0.14 (0.13, 0.15)		0.16 (0.14, 0.17)		0.16 (0.14, 0.17)	
	Education 13	16,261	0.11 (0.1, 0.13)		0.12 (0.1, 0.14)		0.13 (0.11, 0.15)		0.13 (0.11, 0.15)	
	Education 15	37,149	0.12 (0.11, 0.13)		0.12 (0.11, 0.13)		0.14 (0.12, 0.15)		0.14 (0.12, 0.15)	
	Education 19	48,639	0.14 (0.13, 0.15)		0.14 (0.13, 0.15)		0.16 (0.15, 0.18)		0.17 (0.15, 0.18)	
	Education 20	98,873	0.09 (0.08, 0.1)		0.1 (0.09, 0.1)		0.10 (0.09, 0.11)		0.11 (0.1, 0.12)	
	Effect modification beta		$-6.21x10^{-3}$ ($-7.01x10^{-3}$, $-5.41x10^{-3}$)		-5.39x10-3 (-6.19x10-3, -4.59x10-3)		$-6.38x10^{-3}$ ($-7.22x10^{-3}$, $5.54x10^{-3}$)		$-5.52x10^{-3}$ ($-6.37x10^{-3,}$ $-4.68x10^{-3}$)	
Systolic blood pressure	All years	292,277	0.19 (0.19, 0.2)		0.16 (0.15, 0.16)		0.2 (0.2, 0.21)		0.16 (0.16, 0.17)	
	Education 7	46,726	0.18 (0.17, 0.19)	0.160	0.14 (0.13, 0.15)	0.127	0.19 (0.18, 0.2)	0.124	0.14 (0.13, 0.15)	0.191
	Education 10	50,789	0.2 (0.19, 0.21)		0.17 (0.15, 0.18)		0.21 (0.2, 0.22)		0.18 (0.16, 0.19)	
	Education 13	15,772	0.21 (0.19, 0.22)		0.17 (0.15, 0.19)		0.23 (0.21, 0.24)		0.18 (0.16, 0.2)	
	Education 15	36,033	0.19 (0.18, 0.2)		0.15 (0.13, 0.16)		0.2 (0.19, 0.21)		0.15 (0.14, 0.17)	
	Education 19	47,177	0.18 (0.17, 0.19)		0.14 (0.13, 0.15)		0.19 (0.18, 0.21)		0.15 (0.14, 0.16)	
	Education 20	95,780	0.2 (0.19, 0.2)		0.16 (0.16, 0.17)		0.21 (0.2, 0.21)		0.17 (0.16, 0.18)	
	Effect modification beta		$5.62x10^{-4}$ ($-2.22x10^{-4}$, $1.35x10^{-3}$)		6.90x10-4 (-1.97x10-4, 1.58x10-3)		$6.73x10^{-4}$ ($-1.84x10^{-4}$, $1.53x10^{-3}$)		$6.45x10^{-4}$ ($-3.24x10^{-4,}$ $1.61x10^{-3}$)	

P value = P value for effect modification; SD = standard deviation; CI = confidence interval

Table 6.10: Education as an effect modifier of genetic susceptibility to cardiovascular risk factor on observed phenotypic cardiovascular risk factor for binary traits , on the additive scale using polygenic scores at a range of P value thresholds

Exposure	Educational attainment	N	P=0.05				P=0.5			
			Additive scale		Multiplicative scale		Additive scale		Multiplicative scale	
			Risk difference of phenotypic trait (95%CI)	P value	Odds ratio of phenotypic trait (95%CI)	P value	Risk difference of phenotypic trait (95%CI)	P value	Odds ratio of phenotypic trait (95%CI)	P value
Atrial fibrillation	All years	316,912	0.05 (0.04, 0.05)		4.67 (4.56, 4.79)		0.06 (0.06, 0.06)		5.88 (5.73, 6.03)	
	Education 7	51,246	0.07 (0.06, 0.07)	$4.52x10^{-112}$	4.44 (4.22, 4.67)	0.004	0.08 (0.08, 0.08)	$4.97x10^{-167}$	5.51 (5.23, 5.8)	$2.87x10^{-4}$
	Education 10	54,460	0.04 (0.04, 0.04)		4.5 (4.23, 4.79)		0.05 (0.05, 0.05)		5.58 (5.23, 5.95)	
	Education 13	17,213	0.04 (0.04, 0.04)		4.88 (4.34, 5.48)		0.05 (0.05, 0.05)		6.49 (5.74, 7.35)	
	Education 15	38,694	0.05 (0.05, 0.05)		4.71 (4.4, 5.05)		0.06 (0.06, 0.06)		5.77 (5.37, 6.2)	
	Education 19	50,942	0.04 (0.04, 0.04)		4.92 (4.61, 5.26)		0.06 (0.05, 0.06)		6.4 (5.97, 6.86)	
	Education 20	104,357	0.04 (0.04, 0.04)		4.83 (4.6, 5.06)		0.05 (0.05, 0.05)		6.15 (5.85, 6.46)	
	Effect modification beta		$-1.31x10^{-3}$ ($-1.42x10^{-3}$, $-1.19x10^{-3}$)		1.01 (1.00, 1.01)		$-1.57x10^{-3}$ ($-1.69x10^{-3}$, $-1.46x10^{-3}$)		1.01 (1.00, 1.01)	
Coronary heart disease	All years	317,055	0.01 (0.01, 0.01)		1.23 (1.21, 1.25)		0.01 (0.01, 0.01)		1.22 (1.2, 1.24)	
	Education 7	51,061	0.01 (0.01, 0.02)	$8.79x10^{-06}$	1.2 (1.16, 1.24)	0.372	0.01 (0.01, 0.02)	$3.03x10^{-4}$	1.21 (1.17, 1.25)	0.140
	Education 10	54,483	0.01 (0.01, 0.01)		1.24 (1.19, 1.3)		0.01 (0.01, 0.01)		1.19 (1.14, 1.24)	
	Education 13	17,220	0.01 (0.01, 0.01)		1.28 (1.18, 1.4)		0.01 (0, 0.01)		1.23 (1.13, 1.34)	
	Education 15	38,740	0.01 (0.01, 0.01)		1.22 (1.16, 1.28)		0.01 (0.01, 0.01)		1.2 (1.14, 1.26)	
	Education 19	50,912	0.01 (0.01, 0.01)		1.21 (1.16, 1.26)		0.01 (0.01, 0.01)		1.19 (1.14, 1.24)	
	Education 20	104,639	0.01 (0.01, 0.01)		1.26 (1.21, 1.3)		0.01 (0.01, 0.01)		1.27 (1.23, 1.31)	
	Effect modification beta		$-3.22x10^{-4}$ ($-4.64x10^{-4}$, $1.80x10^{-4}$)		1.00 (1.00, 1.00)		$-2.61x10^{-4}$ ($-4.03x10^{-4}$, $1.19x10^{-4}$)		1.00 (1.00, 1.01)	
Diabetes (Type 2)	All years	316,406	0.004 (0.004, 0.005)		1.14 (1.11, 1.16)		0.0016 (0.0009, 0.0022)		1.05 (1.03, 1.07)	
	Education 7	50,904	0.006 (0.004, 0.008)	0.011	1.11 (1.07, 1.15)	0.273	0.0017 (-0.0004, 0.0037)	0.317	1.03 (0.99, 1.07)	0.705
	Education 10	54,261	0.005 (0.003, 0.006)		1.15 (1.1, 1.21)		0.0024 (0.0009, 0.0039)		1.08 (1.03, 1.13)	
	Education 13	17,190	0.003 (0, 0.005)		1.11 (1.01, 1.21)		0.0021 (-0.0004, 0.0047)		1.07 (0.98, 1.17)	
	Education 15	38,683	0.004 (0.002, 0.006)		1.13 (1.07, 1.19)		0.0009 (-0.0009, 0.0027)		1.03 (0.97, 1.09)	
	Education 19	50,814	0.006 (0.004, 0.008)		1.18 (1.13, 1.24)		0.0018 (0.0002, 0.0035)		1.05 (1.01, 1.1)	
	Education 20	104,554	0.003 (0.002, 0.004)		1.14 (1.09, 1.18)		0.0011 (0.0002, 0.002)		1.05 (1.01, 1.1)	
	Effect modification beta		$-1.64x10^{-4}$ ($-2.90x10^{-4}$, $-3.81x10^{-5}$)		1.00 (1.00, 1.01)		$-6.41x10^{-5}$ ($-1.90x10^{-4}$, $6.15x10^{-5}$)		1.00 (1.00, 1.00)	

Stroke	All years	320,120	0.002 (0.002, 0.002)		1.1 (1.07, 1.13)		0.0003 (-0.0001, 0.0008)		1.02 (0.99, 1.05)	
	Education 7	52,012	0.003 (0.003, 0.005)	0.015	1.12 (1.06, 1.17)	0.538	0.0007 (-0.0008, 0.0022)	0.378	1.02 (0.97, 1.08)	0.666
	Education 10	54,899	0.001 (0.001, 0.002)		1.06 (0.99, 1.13)		0.0005 (-0.0006, 0.0016)		1.03 (0.97, 1.1)	
	Education 13	17,355	0.001 (0.001, 0.003)		1.03 (0.92, 1.16)		0.0002 (-0.0018, 0.0022)		1.01 (0.9, 1.14)	
	Education 15	39,144	0.002 (0.002, 0.004)		1.15 (1.06, 1.24)		0.0004 (-0.0009, 0.0018)		1.02 (0.95, 1.1)	
	Education 19	51,418	0.001 (0.001, 0.002)		1.07 (1, 1.15)		-0.0006 (-0.0017, 0.0005)		0.97 (0.9, 1.03)	
	Education 20	105,292	0.001 (0.001, 0.002)		1.1 (1.04, 1.16)		0.0005 (-0.0002, 0.0012)		1.04 (0.99, 1.1)	
	Effect modification coefficient		$-1.31x10^{-4}$ ($-2.05x10^{-4}$, $-2.17x10^{-5}$)		1.00 (0.99, 1.00)		$-4.13x10^{-5}$ ($-1.33x10^{-4}$, $5.05x10^{-5}$)		1.00 (0.99, 1.00)	

P value = P value for effect modification; SD = standard deviation; CI = confidence interval

6.7 Discussion

In this analysis of UK Biobank participants, I found evidence that educational attainment modified the risk of genetic susceptibility to a number of cardiovascular risk factors and outcomes. The direction of this effect was mixed the size of the coefficient for effect modification was small. For some risk factors such as BMI and smoking behaviours higher educational attainment mitigated genetic risk. However, for some risk factors and diseases, such as LDL-C, atrial fibrillation and CHD, higher educational attainment increased genetic susceptibility. These results suggest that modification of the effect of polygenic scores by educational attainment is unlikely to play a clinically meaningful role in the aetiology of cardiovascular inequalities.

Where educational attainment increased genetic susceptibility to cardiovascular disease events and diagnoses it is possible these differences are observed due to differences in rates of diagnosis, which may independently contribute to cardiovascular inequalities.

6.7.1.1 Results in context

A number of studies have sought to identify the interplay between genetic susceptibility to cardiovascular risk factors with a range of lifestyle and environmental factors. For example, a number of studies have demonstrated interactions between genetic susceptibility to BMI and diet and with physical activity (392-395). Gene*environment interactions have been identified between the *APOE* genotype (increasing susceptibility to cardiovascular disease) and smoking (374, 375), the PPAR-γ2 gene increasing susceptibility to type 2 diabetes risk with diet and exercise (376), polygenic score for type 2 diabetes and healthy lifestyle and between the 9p21 genetic variant (increasingly susceptibility to CHD) and smoking (203). Most, although not all, of these previous studies have employed candidate gene approaches and few have considered the role of socioeconomic position interacting with genetic risk.

Two recent studies using UK Biobank have demonstrated that a greater Townsend deprivation index accentuated the genetic risk of obesity (175, 377). However, the previous literature has not found evidence that education modifies the genetic risk of obesity (175, 378). We have expanded on this here by exploring the extent to which education modifies polygenic susceptibility to a wide range of cardiovascular risk factors, rather than focussing on one risk factor. In contrast to the previous literature, we found evidence that educational attainment modifies genetic susceptibility to BMI.

One explanation as to why I found evidence of education as an effect modifier of cardiovascular risk may be because of the education definition used. In my research I have

converted highest educational qualification to ISCED years of schooling, however previous research has used slightly different definitions of education. In one study using UK Biobank, age at which full time education was completed was used (175). In a study using the Understanding Society dataset, highest qualification was used to define education (378). My definition of education has previously been used to demonstrate causal effects of education on i) BMI and ii) CVD (287).

A recent source of much debate in the genetic epidemiology literature is whether the addition of a polygenic risk score in clinical practice adds little predictive power over and above that of a phenotypic risk score (195-197). Whilst phenotypic cardiovascular risk would be known by a clinician, currently, genetic risk is typically unknown to both clinician and patient. My research demonstrates that at the individual level, understanding genetic susceptibility to cardiovascular risk factors or outcomes may help elucidate mechanisms in cardiovascular aetiology, but these are unlikely to explain a substantial proportion of socioeconomic inequalities.

6.7.1.2 Strengths and weaknesses

There are a number of strengths in this study. Much of the previous literature on gene *environment interactions in cardiovascular disease rely on candidate gene style studies (373, 375, 376), which are often criticised for a failure to replicate (204). Here, I have created PGSs for nine phenotypic measures of cardiovascular risk factors or diseases. Whilst candidate gene studies typically focus on a (rare) single genetic variant, or small group of (common) genetic variants that individually explain a large(r) amount of the variance in the trait, PGSs include a large number of genetic variants which each explain a small amount of the variation, but cumulatively explain a large amount (170, 396). For most diseases, including CVD, polygenic inheritance of these common variants plays a greater role than rare monogenic mutations (170, 397). Therefore, the broad measure of genetic susceptibility used here is likely to represent a greater number of biological pathways for the aetiology of cardiovascular disease.

Additionally, I created PGSs at a range of stringent and liberal P value thresholds. At a more stringent threshold (e.g. $P=5x10^{-8}$) the genetic variants included are less likely to be pleiotropic (i.e. also associated with different phenotypes), but the variance explained by the PGS may be lower than with a more liberal threshold (e.g. P=0.5).

Identifying whether the modifying effect of education acts in same direction for each risk factor (i.e. if. education decreased genetic susceptibility to all cardiovascular risk factors) would be of public health importance in identifying opportunities to mitigate cardiovascular

inequalities. With the exception of genetic susceptibility to alcohol consumption, educational attainment was found to modify the effect of all polygenic scores on at least one scale. However, the effect of education did not have the same direction of effect for all risk factors. In the case of BMI and smoking, higher education mitigated polygenic susceptibility to the phenotypes, however for LDL-C and systolic blood pressure, higher education resulted in higher phenotypic measures for a given value of the polygenic score. This means the results identified here are unlikely to explain persisting inequalities in CVD.

The lack of effect modification for alcohol consumption could be due to insufficient power to detect an effect modification or because of the way the variable was defined. For example, alcohol consumption was defined as drinks per week, but type of alcohol consumed may be an important factor which was not accounted for. This work should be replicated in large independent samples to verify the validity of this effect modification.

Studies of effect modification can be biased by reverse causality and confounding. Where possible, for example with genetic susceptibility to cardiovascular diseases, I restricted analyses to incident cases. As education is an early life measure of socioeconomic position many risk factors for disease would be acting as mediators (i.e. on the causal pathway between education and CVD) rather than as confounders (287). Similarly, genetic variants are determined at conception, and therefore not biased by unmeasured later life confounding. However, they can be confounded by population structure (185). In this analysis, I controlled for genetic principal components to minimise bias due to this.

One limitation is the generalisability of these results to other populations. UK Biobank is not representative of the wider UK population, particularly with respect to SEP (16). UK Biobank participants are typically more highly educated and of a higher SEP. Therefore, the absence of effect modification in this sample may be due collider bias caused by non-random selection into the study (221).

Although I have identified education modifies the effect of polygenic scores for some cardiovascular risk factors, these effects may differ (e.g. be larger in magnitude), should measures of adult socioeconomic position be considered. This may also explain some of the non-linearities observed when stratifying by years of educational attainment, as the ISCED definitions of educational attainment used here, assign a high number of years of education to those who attain a vocational qualification and likely enter manual labour.

I used the summary statistics from the largest available GWAS for each trait (not including UK Biobank), however the PGSs explain small amounts of the phenotypes. As GWAS become larger and explain more variance in phenotypic traits, it may be possible to detect smaller effect modification.

6.7.1.3 Public health implications

In this analysis I have demonstrated that educational attainment modifies genetic susceptibility to a number of cardiovascular risk factors and outcomes. However, the direction of these effects was not consistent. These results do not specifically say what it is about educational attainment that modified genetic susceptibility to cardiovascular risk factors and outcomes. Additionally, it is possible that differences in cardiovascular diseases are due to differences in rates of diagnosis. Although this works begins to allude to risk stratified interventions based on genetics, it will be important to understand more specifically what it is about education that leads to these more adverse consequences. For example, remaining in education may protect an individual from starting to smoke due to social pressure or increased knowledge of the harms, even if they have genetic variants increasing their susceptibility to heavier smoking. However, it will be important to identify what factors may explain the differences in the directions of effects.

6.7.1.4 Conclusions

In this study I have found that educational attainment modifies the genetic susceptibility to a number of cardiovascular risk factors. The direction of this effect was mixed, and the sizes of the effect modification coefficients were small, suggesting modification of the effect of genetic susceptibility to cardiovascular risk factors or cardiovascular disease by education attainment are unlikely to contribute to the mechanisms driving inequalities in cardiovascular risk.

Chapter 7. Discussion

In Chapter 3 to Chapter 6 the main findings of each analysis were presented alongside a discussion of the strengths and limitations from each of the analyses, including the methods and data. In this chapter, I summarise the key findings of each analysis chapter. I consider the contribution this thesis makes to our understanding of cardiovascular inequalities and the causal inference literature. I discuss the strengths and limitations of the thesis as a whole. Finally, I examine the public health and policy implications of my thesis and make recommendations for future research.

7.1 Summary of key findings

Despite reductions in the rates of morbidity and mortality from cardiovascular disease (CVD) in high income countries (28, 30, 32), individuals who are the most socioeconomically deprived remain at the greatest risk of disease (3, 76). In this thesis, I aimed to understand what processes may be driving socioeconomic inequalities in CVD, focussing on educational attainment as a measure of socioeconomic position (SEP) using causal inference methods.

Multivariable Mendelian randomisation (MVMR) and two-step Mendelian randomisation (MR) methods were not new to this thesis (19-21, 283). Previous literature had used these methods to estimate direct effects (253, 398), including in the presence of pleiotropy (21, 282), and to infer causation (19, 399). However, they had not been used to decompose the direct effect indirect effect and proportion mediated. Additionally, there was no guidance in the literature about the two approaches to using MR for mediation, whether and when they differed, and whether there were situations in which one method was more appropriate than the other. In Chapter 3, using simulations and an applied example, I demonstrated how these two MR methods could be used in mediation analysis, to estimate direct effects, indirect effects and the proportion mediated. I presented a number of methodological considerations, including current limitations and sources of bias in these analyses. This work has been designed to reach different audiences, both applied and methodological researchers. In the motivating example for this chapter I demonstrated that body mass index (BMI) likely mediated the association between educational attainment and i) systolic blood pressure, ii) hypertension and iii) CVD (all subtypes combined). However, there was little evidence that low-density lipoprotein cholesterol (LDL-C) mediated these associations.

In Chapter 4 I have identified that educational inequalities in CVD may occur via a number of mediating pathways. Individually, BMI explained up to 18% of the association between

education and coronary heart disease (CHD), smoking explained up to 21% and systolic blood pressure explained up to 33%. When considered together, BMI, smoking and systolic blood pressure were estimated to explain up to 36% of the effect of education on CHD. For the association between education and myocardial infarction the three risk factors explained up to 41%. Considering stroke as the outcome up to 52% of the association was explained and for all CVD as the outcome up to 41% was explained.

I have identified that educational differences in statin use for primary prevention are likely to contribute to educational inequalities in CVD. I identified an interaction between educational attainment and cardiovascular risk (via QRISK3 score), such that for a given QRISK3 score individuals who leave education after 7 years (equivalent to compulsory education) are less likely to report using statin medication compared with those who leave education after 20 years (equivalent to obtaining a degree) (Chapter 5).

Finally, I have demonstrated that effect modification of cardiovascular risk by educational attainment is unlikely to substantially contribute to the development of inequalities in cardiovascular risk. My research showed that educational attainment mitigates genetic susceptibility to BMI and lifetime smoking, but accentuated genetic susceptibility to LDL-C, atrial fibrillation and CHD (Chapter 6). For example, for a given level of genetic risk of smoking, individuals with lower educational attainment, had a higher lifetime exposure to smoking compared with higher educated individuals with an equivalent genetic risk. However, for a given genetic risk of LDL-C, individuals with higher educational attainment were more likely to have higher observed levels of LDL-C, compared with lower educated individuals with an equivalent genetic risk. Where effect modification was observed, the size of the effect modification coefficients was typically small.

7.2 Contributions to the literature

Although each analysis chapter features a full discussion of the results in context, here I make a general discussion of the contributions made. My results indicate that BMI, smoking, systolic blood pressure and statin use contribute to the accumulation of CVD in individuals with low educational attainment, but not LDL-C or effect modification of polygenic scores by education on cardiovascular risk.

Previous mediation analyses have implicated BMI, smoking and systolic blood pressure as mediators of education and CVD. For example, Kershaw and colleagues, identified that almost 27% of the association between education and CHD was mediated by smoking, with 10% and 5% attributed to obesity and hypertension respectively (114). My work in Chapter 4 builds on

the previous literature by using novel MR methods to investigate the *causal* role of BMI, smoking and systolic blood pressure as mediators. I demonstrated in Chapter 3 that MR could be used to overcome confounding and measurement error in mediation analyses, improving the causal inference that can be made.

Whilst a number of studies have identified associations between SEP and statin use, the direction of this effect has been mixed (36, 87, 142, 143, 145-147, 340). Although some previous studies adjusted for some cardiovascular comorbidities, few previous studies have comprehensively accounted for underlying cardiovascular risk (87, 143, 145). As demonstrated in Chapter 4, individuals with lower educational attainment have a higher prevalence of cardiovascular risk factors. Therefore, it would be expected that individuals with lower educational attainment have a higher prevalence of statin use. One previous study was identified that accounted for underlying cardiovascular risk assessed by Framingham score; however, this study did not test for interaction between SEP and cardiovascular risk. Here, it was found that the use of statins was not associated with SEP (141). My research in Chapter 5 builds on these previous studies by investigating interactions between educational attainment and underlying cardiovascular risk. Here, I found that for an equivalent cardiovascular risk (assessed via QRISK3 score) higher educated individuals were more likely to report using statins, compared with lower educated individuals.

In Chapter 6, I investigated effect modification between educational attainment and polygenic scores (PGS) for a number of cardiovascular risk factors and diseases. Although some previous analyses had sought to study similar research questions, no previous analysis had examined multiple cardiovascular risk factors using a polygenic (as opposed to candidate gene) framework. For example, Tyrrell and Colleagues investigated gene*environment interactions between a BMI PGS and a number of indicators of the obesogenic environment (377). Here, an interaction between BMI PGS and Townsend deprivation index (TDI) (a population indicator of SEP) on observed BMI was found. Conversely, Amin and colleagues did not identify an interaction between BMI PGS and educational attainment in both Finnish and UK cohorts (378). I identified that educational attainment mitigated the risk of genetic susceptibility to BMI and smoking, but increased genetic susceptibility to LDL-C, atrial fibrillation and CHD. Although this work identified that educational attainment may be an important effect modifier for some cardiovascular risk factors, this is unlikely to strongly contribute to inequalities in CVD.

7.3 Strengths and limitations of this research

There are a number of strengths and limitations of the work presented in this thesis. Understanding what these are and how they might affect the interpretation of the results presented is important for understanding the wider contribution of this work and the causal inference that can be made.

7.3.1 UK Biobank

UK Biobank is an incredibly rich data source, including phenotypic, genetic, metabolomic data and linked health outcomes in over 500 000 individuals. Few studies, if any, have the extensiveness of data like UK Biobank. The breadth of data collected has allowed for thorough and robust interrogation of the research aims addressed in this thesis. The adult population of UK Biobank (age range 36-75) makes this is an ideal cohort study for exploring cardiovascular outcomes, where CVD is most common in older individuals (28). Although it should be noted that as the length of follow up for participants is still relatively short (maximum follow up 11 years) there are still relatively small numbers of some cardiovascular outcomes, and therefore MR mediation analyses in Chapter 4 were complemented with summary data MR analyses.

Despite educational attainment being measured retrospectively in UK Biobank participants, individuals were asked to report their highest qualification achieved, which is unlikely to be subject to recall bias. The long latent period between educational attainment and CVD the study design of UK Biobank makes it a suitable cohort study for the research questions addressed here.

There are a number of specific strengths and limitations of using these data within my thesis, which are discussed throughout this section.

7.3.2 Statistical power

The methods used in this thesis, such as individual level MR, mediation analysis and interaction analyses typically require very large sample sizes to achieve statistical power. Therefore, the number of participants in UK Biobank is important in going some way to achieve adequate power for analyses. Notably in individual level MR mediation results in Chapter 3 and Chapter 4 estimates are imprecise, suggesting power may not be sufficient to detect the effects of interest, particularly when estimating the proportion mediated.

In Chapter 6, education as an effect modifier of PGSs for a number of cardiovascular risk factors or diseases was assessed. Risk factors were selected based on evidence of causal effects on CVD. Additionally, it has been demonstrated, including from my own research, that

education is a cause of CVD. Therefore, effect modification should be present for all risk factors on either the additive or multiplicative scale. For some risk factors, including alcohol consumption and stroke, there was no evidence of effect modification. It is possible this was due to there not being sufficient power to detect these associations.

7.3.3 Reverse causality

A key assumption throughout this work, is that the temporality between the exposures, mediators and outcomes have been correctly specified and the results are not biased by reverse causality. However, there is some evidence that high BMI is causally associated with lower SEP, including educational attainment (265, 400), as well as some evidence that smoking initiation and lifetime smoking are associated with lower educational attainment (401).

Bias by reverse causality has been mitigated in two main ways in this thesis. Firstly, MR has been used where appropriate for mediation analyses. Due to the properties of genetic variants being i) randomly allocated at meiosis and ii) stable throughout the life course, MR estimates are robust to bias by reverse causality (see a full discussion of these methods in 7.3.4). These methods can also be used to identify the direction of causality and to test for evidence of bi-directional effects (313). Secondly, wherever possible, the temporality of data has been maintained, particularly considering cardiovascular outcomes. For example, individuals with prevalent (at baseline) cases of CVD have been excluded from main analyses.

Given the linkage to hospital inpatient records (hospital episode statistics in England and Wales and Scottish morbidity records), primary care data and prescription data, incident and prevalent cases of disease could be ascertained. This means the association between exposures, mediators (where applicable) and outcomes could be assessed prospectively. Although MR analyses (Chapter 3 and Chapter 4) are not biased by reverse causality, phenotypic analyses and studies of effect modification (Chapter 3 to Chapter 6) can be biased in this way. This temporality means bias due to reverse causality is unlikely to be present.

In Chapter 5, the primary outcome considered was statin use. Due to the large sample sizes required to achieve statistical power in interaction analyses, primary analyses were carried out using cross sectional analyses, i.e. cardiovascular risk factors were assessed at the same time as statin use. These analyses were replicated using primary care and prescription data, available for about half of the eligible participants. Here, temporality could be maintained between assessing cardiovascular risk at baseline with prospective prescriptions for statins. These

prospective results were comparable to results using cross-sectional data, suggesting bias by reverse causality is not present in the main analyses.

However, it should be noted that these linked sources of data (hospital inpatient records and prescription data for example) have their own limitations, particularly that of missing data both in terms of data collected in healthcare and because measurements are only available for individuals who present at a healthcare setting. (402).

7.3.4 Assumptions of Mendelian randomisation

In Chapter 3 and Chapter 4, MR methods were used to estimate the role of intermediate risk factors mediating the association between education and CVD. Mendelian randomisation studies have been described as natures Randomised controlled trial (RCT) (254, 403), where genetic variants are randomly allocated at conception and not influenced by later life factors (i.e. confounders) (18, 248). These properties mean estimates are unbiased by unmeasured confounding (with some exceptions, see section 7.3.6.1) and reverse causality. Additionally, non-differential measurement error is less of an issue (see section 7.3.7) in MR studies (248, 279, 404).

However, estimates from MR can be biased from a number of different sources (405). One important potential source of bias is through invalid instruments, which may have pleiotropic effects on pathways independent of the exposure of interest (254, 255). When carrying out MR mediation analysis, these pleiotropic pathways may also be present for the instrument for the mediator of interest. Limited methods are available for testing for pleiotropy in MR mediation (302) and indeed MVMR was introduced as a method for dealing with pleiotropy (282, 283). In the absence of specific tests, MR-Egger was used to test for evidence of pleiotropy by the instruments for the exposure and each of the mediators (255). The estimates from MR-Egger were consistent with those from the main IV regression analyses, suggesting that estimates are unlikely to be biased due to pleiotropy.

Estimates from MR can also be biased by weak instrument bias (300). This was assessed by F-statistics and conditional F-statistics in MR mediation analyses (296). For all exposures and mediators, the instruments had high F-statistics, indicating that the effect estimates were unlikely to be biased by weak instruments. As demonstrated in Chapter 3, weak instrument bias in the exposure and mediator introduced bias to the estimates of mediation, with the size of bias greatest when a common binary outcome was considered.

7.3.5 Lifetime exposure in Mendelian randomisation

Estimates from MR are said to be estimating effects of lifetime exposure to a trait (18, 406). Whilst a trait such as educational attainment is likely to be stable across much of the life course, other traits, including BMI, smoking and systolic blood pressure are likely to be less stable. Indeed, systolic blood pressure is subject to daily, or even context dependent, variation (407). It may indeed be these variations that are more important to disease aetiology, rather than a lifetime exposure. Although methods for accounting for time-varying exposures in MR are emerging, they are not yet widespread and only available for a limited number of traits (311, 406, 408). It is important to consider the results presented in this thesis in the context of lifetime exposure to the traits considered. In Chapter 4 considering the mediating role of BMI, smoking and systolic blood pressure between education and a number of cardiovascular outcomes, estimates of the proportion mediated were typically larger in MR mediation analyses (in particular using summary data MR) compared with phenotypic analyses. For example, in phenotypic analyses systolic blood pressure mediated 19% of the association between education and CHD, in individual level MR the proportion mediated was 17% and in summary data MR the proportion mediated was 21%.

7.3.6 Confounding

Not all work in this thesis has used MR and therefore may be biased by unmeasured confounding (409). In Chapter 3 and Chapter 4,phenotypic mediation analysis was also used. As educational attainment (the exposure considered) is an early life measure of SEP, many individual level risk factors for disease are more likely to be acting as mediators and not confounders of the association between educational attainment and CVD. However, based on the previous literature, some familial level factors may still be acting as confounders, such as parental SEP (8). If not appropriately controlled for, confounding can introduce bias by inducing an association between an exposure and an outcome that does not truly exist or over/under-estimate the effect of the exposure on the outcome. To control for this, a proxy measure for Townsend deprivation index (TDI) at birth was estimated based on birth location and current TDI for the location. Analyses were additionally adjusted for birth location. However, these proxy measures are poor measures of true family level SEP and therefore residual confounding may remain. It is important to note that mediators of an association should not be controlled for (if not in a specific mediation analysis) as this can underestimate the effect of the exposure on the outcome, therefore careful consideration was given in these analyses to not over-adjust (410).

The QRISK3 cardiovascular risk score estimated in Chapter 5 is a type of prediction model. The goal here is to estimate the 10 year risk of future cardiovascular disease, conditional on the values of multiple risk factors (411). These models represent a comprehensive assessment of cardiovascular risk, capturing risk factors that may be considered as either confounders or mediators of the association in addition to area level SEP measured by TDI of current location. Therefore, no further covariates were included in analyses in Chapter 5.

7.3.6.1 Population stratification, assortative mating and dynastic effects

Genetic associations are assumed to reflect direct genetic effects. However, family level effects including assortative mating (412, 413) and dynastic effects (414), or fine-scale population structure (185), are all potential sources of confounding in genetic studies of unrelated individuals (186, 415). Without accounting for, or controlling for these effects, confounding can be introduced. Indeed, even in non-genetic analyses dynastic effects, which occur when the parent phenotype directly influences the phenotype of their offspring, can lead to confounding due to family level SEP (186, 416).

When considering social exposures, such as educational attainment, bias caused by these effects can be pertinent (185, 415). It has been demonstrated that after controlling for family effects, the heritability of educational attainment is reduced i.e. there is a strong indirect effect of parental education on offspring education (414). However, importantly for the work presented here, even after controlling for family level factors (in the form of twin-studies) causal effects of education on health remain (417).

Methods are emerging for within-family MR analyses, and the results presented here would be an ideal candidate for replication with these analyses (see 7.5.1). Through this design, confounding by dynastic effects, assortative mating and population structure are controlled for. These designs can either account for family structure using sibling data or parent-offspring trio data. In sibling studies, the difference between phenotype and genotype within siblings can be estimated or family level means can be estimated and controlled for (186, 418). In parent-offspring trio designs, parental genotype can simply be adjusted for, or MVMR methods can be used to estimate the direct and indirect effects of the parents genotypes (414, 418). However, achieving adequate statistical power is challenging; particularly for analyses such as mediation and interactions. In UK Biobank, about 20 000 sibling pairs are available, but this sample size is unlikely to be large enough to provide adequate statistical power for the analyses carried out in this thesis.

In this thesis I adjusted genetic analyses for genetic principal components, as a method of controlling for wider population stratification. Although it should be noted that this often is not enough to account for population stratification (185, 419). In an effort to control for family level effects, TDI at birth and location of birth were adjusted for, which would capture some family level SEP. However, these family effects were not controlled for using genetic methods, e.g. within family genome-wide association studies (GWAS) (420).

Despite these limitations, the results presented in this thesis remain important for considering potential opportunities to reduce inequalities in CVD. Understanding the role of family effects in the aetiology of cardiovascular inequalities will be important for improving the health of future generations. However, understanding the effect of educational attainment, without accounting for family effects, provides an opportunity to identify why and how inequalities exist in those (adults) most at risk of disease in the short term.

7.3.7 Measurement error

Throughout this work, careful consideration has been given to minimising measurement error in all risk factors (i.e. exposures and mediators) and outcomes considered. Although one of the main strengths of this thesis is the triangulation of different methods, where MR is robust to measurement error, phenotypic analyses analyses can still be biased by measurement error and indeed measurement error in GWAS can introduce bias to MR estimates.

As previously discussed, (7.3.1), hospital inpatient records were used to identify cases of CVD. The end points considered in Chapter 3, Chapter 4 and Chapter 6 of this thesis, were all serious cardiac events that would likely result in a hospital admission. This is a more objective measurement of CVD, compared with self-report data for example, where misclassification of the outcome may occur due to misreporting or recall bias by an individual.

UK Biobank baseline assessment centres followed clear protocols for all data collected. Risk factors considered in this thesis, such as BMI, systolic blood pressure and LDL-C (Chapter 3, Chapter 4 and Chapter 6) were measured objectively by trained study nurses using calibrated machinery, minimising risk of measurement error. Risk factors such as smoking and alcohol consumption (Chapter 4 and Chapter 6) may be subject to misclassification, or measurement error, due to self-reporting bias (421), possibly due to social desirability (422). As UK Biobank is typically healthier than the general population it can be difficult to compare with population estimates of smoking or alcohol consumption for example.

In Chapter 5 where a number of different disease diagnoses were included in QRISK3 scores. A combination of self-report disease status, assessment centre medication data and hospital inpatient records was used to code each disease. Typically, hospital inpatient records will only capture severe cases of disease, resulting in a hospital admission. Whilst this is highly likely to occur for a serious cardiac event (e.g. stroke or myocardial infarction) this is less likely to be the case for diabetes for example. Including self-report of a diabetes diagnosis therefore reduces the potential misclassification of a case as a control but does introduce potential recall bias. Similarly, for a diagnosis such as impotence (a variable included in the QRISK3 cardiovascular risk score), a diagnostic ICD code is available, but cases are unlikely to attend a hospital for this condition. Therefore, cases could more reliably be ascertained via medication data. However, it should be acknowledged that for many conditions, including impotence, participants may choose not to present at a clinical setting for this reason and therefore would not have relevant medication data and nor would they self-report the condition to study nurses. The breadth of data available in UK Biobank therefore reduces potential bias by measurement error or misclassification as case status can be ascertained in a number of ways, although some small bias may still be present.

One area where measurement error has been difficult to quantify is in reported statin use (Chapter 5). UK Biobank participants were asked to report to study nurses any regular medication they were taking, but not those purchased over the counter without a prescription, medications prescribed but not taken, or any supplements or vitamins the participants used. Participants were asked to take the packets of medication so study nurses could exactly record the medication and dosage. This is a clear strength; study nurses are likely to accurately record the medication and it is not reliant on participant recall. However, this method does rely on participant taking medication to assessment centres. Through triangulating with primary care prescription records there was a large amount of discrepancy between those reporting statin use to study nurses and those having a statin prescription (see 7.5.1 and 7.6). I was able to identify that a number of these individuals had a statin prescription prior to baseline, but no current prescription (defined as a prescriptions 3 months before and after baseline). Similarly, I was able to identify a number of participants who either never received a prescription, or only received a prescription after baseline. There were also individuals who had a prescription during the 3 months before and after baseline but did not report using them to study nurses. For this reason, care was taken to describe statins inequalities in terms of reporting, rather than prescriptions or use. Reassuringly, in analyses replicated with different definitions of statin use (e.g. self-report or prescription only) results were comparable.

Although MR estimates are robust to bias by non-differential measurement error, differential measurement error can induce bias in the GWAS used to identify instruments. This has been demonstrated using behavioural traits (alcohol consumption and smoking) in UK Biobank (423), where misreporting or longitudinal changes in a phenotype result in reduced power of a GWAS to detect true signals or the inducement of spurious signals. This would have implications for the validity of the instruments included in MR analyses.

7.3.8 Selection bias and generalisability of results

One of the key limitations of using UK Biobank, is that it is not representative of the general population (16). Participants are typically more highly educated, have a higher SEP and exhibit greater health seeking behaviours than the general population (16, 424). Despite the large sample size of UK Biobank, this sample represents a response rate of only 5.5% of the 9.2 million individuals invited to take part (16, 425). Selection bias and unequal distributions of SEP is not unique to UK Biobank and indeed a number of studies report similar differences (426-430). This selection bias means the results presented here may not be generalisable to the general UK population (and indeed non-UK populations). This is particularly important when using these data to study health inequalities, as I have done in this thesis. It would be expected that the associations presented here are likely to be larger in the general population, where there are greater socioeconomic disparities (431).

A number of studies have sought to identify predictors of this study participation and quantify the size and effect this selection bias may have on the generalisability of results, both in genetic and non-genetic analyses. Selection bias can occur if the exposure, outcome, or causes of the exposure and outcome are associated with participation (432). For this reason, Tyrrell and colleagues sought to identify causal factors that influenced participation into option UK Biobank follow-up assessments. Using summary data MR, it was found, that among other factors, higher educational attainment increased participation, whilst higher adiposity decreased participation. In phenotypic analyses lower deprivation and never smoking increased participation (433). Although these results are exploring participation for optional follow-up assessments in UK Biobank, not baseline participation, it is likely that similar factors will be involved in initial participation. Given that educational attainment, BMI and smoking have all been considered as exposures (or mediators) in analyses presented in this thesis and have previously been implicated as risk factors for cardiovascular outcomes, results presented here may be biased by selection bias.

It has often been said that selection bias should not have an effect on the observed associations between an exposure and outcome, but may lead to stronger biases in estimates of prevalence (434). This was recently assessed by Batty and colleagues comparing risk factor associations in UK Biobank with the more representative, general population based studies, Health Surveys for England and Scottish Health Surveys (HSE-SHS) (424). Both studies had similar age and sex distributions, although UK Biobank participants were more educated, more likely to be physically active and less likely to smoke compared with HSE-SHS participants. It was found that some, but not all, risk factor associations in UK Biobank were comparable to those from with HSE-SHS. For example, in UK Biobank the hazard ratio for the association of baseline biomedical characteristics (including among other characteristics age, sex, total cholesterol and systolic blood pressure) on self-report CVD was 4.92 (95% CI: 4.50 to 5.39), whilst the comparative hazard ratio in HSE-SHS was 2.61 (95% CI: 2.35 to 2.90). However, the association between baseline biomedical characteristics and obesity was more similar, where the hazard ratio in UK Biobank was 1.68 (95% CI: 1.55 to 1.83) and in HSE-SHS the hazard ratio was 1.47 (95% CI: 1.31 to 1.61). The study authors concluded that association estimates in UK Biobank were likely generalisable to the UK general population (424).

However, this view that selection bias does not bias observed associations has been criticised for not considering the effect of selection bias on collider bias (221). Munafò and colleagues demonstrated in a simulation based on UK Biobank data, that analyses using PGSs are particularly vulnerable to collider bias caused by selection. Here, the association between the phenotypic of interest and participation will result in the PGS being more strongly related to participation, compared with individual genetic variants alone (221). Given the use of PGSs in this thesis there is potential that the results are affected by collider bias.

Further to the issue of selection into the study, is selection out of a study, such as via non-participation, loss to follow up or participant withdrawal (435, 436). There have been some selective follow up clinics of UK Biobank, however these have not sought to engage all participants. The primary source of follow up data used in this thesis is linked hospital inpatient records. Although participants are free to withdraw from the study, and indeed small numbers of participants have, this follow up process is automated through linkage processes. Therefore, in the context of these analyses, bias due to attrition (whereby those who remain in a study differ from those who remain in a study) should be limited (437).

Further selection bias and limitations of generalisability may have been introduced when individuals were excluded from analyses based on missing data, previous CVD or due to

ancestry. Where genetic analyses have been carried out, analyses have been restricted to participants of White British ancestry (Chapter 3, Chapter 4 and Chapter 6). This restriction has been made due to i) the potential confounding that can be introduced by population stratification (185) and ii) because genome-wide association studies are typically carried out in European participants, where results are often not generalisable across different ancestries (438). Therefore, to ensure the validity of the instruments used, the analyses are carried out on a similarly restrictive sample of participants. Given known ethnic differences in CVD and SEP, it is likely that the association between education (or SEP more widely) and CVD differs in different ancestries. For example, different mediators may partly explain the association, or different interactions between the genes and observed environment may be important. Therefore, it may not be possible to generalise the results presented here to other populations (439).

In Chapter 4, individual level MR estimates were triangulated to summary data MR estimates. Although the summary statistics for summary data MR largely came from GWAS of European participants, the populations considered represented greater population variation than in UK Biobank. Estimates from summary sample MR were comparable (albeit with greater precision) to the individual level MR analyses, suggesting despite this selection bias into UK Biobank the results were still largely generalisable to other (predominantly) European populations.

Where exclusions to the UK Biobank sample were made, I compared the distribution of the characteristics between those included in analyses and those who were excluded either based on ancestry or other factors such as missing data. The characteristics were comparable across the participants included/excluded, suggesting this internal selection is unlikely to be a source of bias (with the exception of generalisability across ancestries).

7.3.9 Missing data

Although data from UK Biobank baseline assessment centres is largely complete, there is still some missing data. This is particularly pertinent for the biochemical assays, including measures of LDL-C and total cholesterol, as well as systolic blood pressure. To minimise bias due to missing data in Chapter 5 I carried out multivariate multiple imputation to impute data for any variable included in the QRISK3 cardiovascular risk score with missing data. This improved power to test for interactions by not reducing the sample size. Provided the assumption that data are missing not and random and that the regression model used to impute the data is correctly specified, analyses using multiple imputation are less prone to bias than complete case analyses (440). Importantly, for the analyses in Chapter 5, the

imputation model was specified to include interactions between i) sex and ii) educational attainment to preserve the interactions being tested in these analyses (354). Although there are no methods to directly test that the multiple imputation assumptions hold, when comparing the analyses using imputed data with the complete case data there was little difference between the observed effects, suggesting that missing data is not an important source of bias.

7.3.10 Genome wide association study of educational attainment

Educational attainment has long been described as a result of shared genetic and environmental influences (441). In Chapter 3 and Chapter 4 of this thesis I use GWAS summary statistics of educational attainment to instrument education in MR analyses (17, 149). To avoid sample overlap between the discovery GWAS and analyses in UK Biobank, the 2016 Okbay *et al* GWAS (17) was used in individual level MR analyses, and the Lee *et al* GWAS (149) which included UK Biobank participants in the complementary summary data MR analyses. As summary data MR analyses were carried out by co-authors, I will focus this discussion on the Okbay *et al* GWAS.

As with most PGSs, individually, each of the 74 genome-wide significant single nucleotide polymorphisms (SNPs) explains a very small amount of the variance in education. For example, the single variant explaining the greatest amount of variation, only explained 0.035% of the variance in education. However, the combined 3.2% explained by all 74 SNPs in a PGS is large enough to be meaningfully useful for social science research. Indeed, in my MR analyses, I had large F statistics and conditional F statistics, indicative of instrument strength.

Where PGSs for complex or behaviour traits are used, it is difficult to know what is being captured by the genetic variants. Of the 74 SNPs identified for educational attainment, 15 were found in genes and biological pathways involved in prenatal brain development (17). Some of the SNPs were also found to be associated with increased cognitive performance, increased risk of bipolar, decreased Alzheimer's and lower neuroticism. These pathways could either reflect vertical pleiotropic pathways, where hypothetically the genetic variants affect cognitive performance which in turn affects educational attainment, or vice versa, where the genetic variants affect educational attainment, in turn affecting cognitive performance. This form of pleiotropy does not result in bias in MR estimates (254). Conversely, these pathways could reflect horizontal pleiotropy where the SNPs independently affect cognitive performance and education. This would result in bias in MR estimates (254). To evaluate potential pleiotropy, MR-Egger was carried out, which can detect violation of the exclusion restriction criteria

(255). My MR-Egger estimates indicated the MR results were not biased by horizontal pleiotropy.

In this GWAS, educational attainment was defined in terms of the amount of formal education an individual had completed. Highest major educational qualification was converted to the International Standard Classification for Education (ISCED) definitions of years of education, allowing for comparisons between 64 heterogenous cohorts meta-analysed in the GWAS (17). This standard definition meant educational attainment in UK Biobank could be defined in the same way in analyses presented in this thesis. However, where analyses were stratified by years of education in non-genetic analyses for example in interaction analyses Chapter 5 and Chapter 6, typically non-linear associations were observed. Whilst this definition allows for a standard approach to measuring educational attainment in heterogenous studies, our results suggest this definition may not be suitable to the UK education system. In the UK context, the ISCED definitions allocate a high number of years of education (19 years) to those with a vocational qualification. However, these individuals are more likely to enter manual labour jobs. This will likely explain some of the non-linearities observed.

Non-genetic instruments are available for educational attainment, such as the policy reform, the Raising of School Leaving Age (RoSLA). Most relevantly for social science research, in 1972 in England and Wales the compulsory school leaving age was raised from 15 to 16. This means some individuals are forced to remain in education longer than they would have and increases the average education levels in the relevant cohorts (267). The validity of the RoSLA as an instrumental variable has previously been demonstrated and has widely been used in economics for causal inference (442). In the same way that genetic instrumental variables in an MR approach avoid bias by confounding and reverse causality, natural experiments cannot be biased by confounding or reverse causality and any effect must act through the change in education. Indeed, many of the MR methods have evolved from econometrics methods for instrumental variable analyses (443). However, a limitation of this approach is that the RoSLA instrument only estimates the effect of a 1-year increase in education, in those who leave school at 16 compared to 15, in a select cohort of individuals who were in education as the policy reform was enacted. Conversely, the genetic variants for educational attainment used in this thesis represent an average effect across the whole distribution of education i.e. is not restricted to a 1-year difference.

7.3.11 Triangulation of methods and data

A strength of the work presented here is the triangulation of different methods and sources of data (259). For example, in Chapter 4 I triangulated mediation results from individual level MR and summary data MR to estimate the causal role BMI, smoking and systolic blood pressure play in mediating the association between educational attainment and CVD. In Chapter 5, I used multiple different data sources to answer my research aim, including data collected at baseline UK Biobank clinics, primary care data and prescription data to estimate interactions between educational attainment and cardiovascular risk on statin use (and prescriptions).

Importantly, these different analytical designs and data sources have different sources of bias (259). For example, when comparing results from phenotypic and MR mediation methods in Chapter 4, phenotypic methods may be biased by unmeasured confounding or measurement error, whilst MR methods may be biased by pleiotropic pathways or in the presence of weak instrument bias. Similarly, when comparing estimates of interaction between QRISK scores and educational attainment on statin use in UK Biobank using data from baseline assessment centres and from primary care records in Chapter 5, the sources of bias in the data are different. For example, UK Biobank data is much more complete with little missing data, but the measurements of the data may not reflect those used in primary care and clinical practice. An example of this is cholesterol measures, where samples in UK Biobank were non-fasting, but typically would be fasting in clinical practice. Conversely, clinical data is not uniformly collected for all patients, and only clinically relevant information to the appointment is recorded.

The triangulation in this thesis could be improved and strengthened further by triangulating with data from different countries, or different cohort studies with different sociodemographic characteristics to UK Biobank.

7.4 Other potential mechanisms

Although not exhaustive, a broad scope of mechanistic pathways has been considered in this work, ranging from behavioural and lifestyle factors, to biological pathways and preventative medication. All of these mechanistic pathways were found to be involved in the aetiology of inequalities in CVD.

The risk factors considered as mediators in Chapter 3 and Chapter 4 were selected based on their known causal effects on CVD, the availability of genetic instruments and because they will be capturing a broad range of other risk factors. For example, BMI will also be capturing

related factors such as exercise and diet. Indeed, when these variables were included in a phenotypic mediation model (in the absence of genetic instruments) they explained no more of the effect of education on CVD than the three main mediators considered. However, there may be other mediators not captured, or only partially captured through these three mediators. For example, adverse mental health may be an important mediator, where lower educational attainment increases the risk of adverse mental health (444), which is suggested to be an independent risk factor for CVD (445).

In Chapter 5 I have focused on statin use, however, other preventative medications, such as antihypertensives, may be important in the development of inequalities. Additionally, adherence to statins (and other preventative medication) may be important in determining inequalities.

Interestingly in Chapter 6, the effect of education as an effect modifier for a numer of cardiovascular risk factors did not always act in the same direction. Whilst individuals with low educational attainment and higher genetic susceptibility were more likely to smoke, they were less likely to experience adverse levels of LDL-C. Understanding the mechanisms specifically involved in how educational attainment modifies genetic effects will be important. For example, these effects may be due to remaining in education leading to increased knowledge and, or because of, greater intelligence. However a number of studies have identified independent effects of education on CVD and cardiovascular risk factors after controlling for intelligence (272, 273, 398). Conversely, increased early SEP, would likely lead to higher adult SEP which may be more important for the aetiology of disease later in life.

Throughout this thesis I have only considered educational attainment as an indicator of SEP. Socioeconomic position broadly covers a number of different indicators, including individual level SEP such as education and adulthood income, occupation or employment. At the family level, early life family SEP can be captured by parental education or parental income, and in adulthood, household income can be considered (71). At the population level, deprivation indices such as the TDI can be used to estimate SEP (446). These different indicators are often, incorrectly, used interchangeably (447). Across the life course, SEP is complex, where different indicators may remain more, or less, stable. For example, income will likely change during different life stages. During active professional life, income will likely fluctuate then change again at retirement (71). Conversely, education is determined during childhood and early adulthood and will likely remain stable through adulthood and to retirement. Indeed, there is evidence that these different indicators of SEP accumulate across the life course to

affect cardiovascular risk (448). By not considering these additional indicators across the life course it is likely that the analyses presented in this thesis are not capturing the full complexity of the association between SEP and CVD. However, the focus on educational attainment means causal inference methods such as MR can be used. Although a GWAS of income has been published during the duration of this thesis work (266), genetic association estimates are not available for other indicators of SEP.

7.4.1 Individual and societal determinants of inequalities

Socioeconomic inequalities in CVD are not a new phenomenon. The Whitehall I study of civil servants demonstrated an association between occupational social class and CHD in the 1970s, where men in the lowest grade of employment had 3.6 times the CHD mortality compared with those in the highest employment grade (94). Some 50 years later, I have demonstrated similar effects of education on CHD.

In this thesis I have largely considered individual level factors in driving these inequalities. For example, by studying and demonstrating that BMI and smoking are mediators of educational inequalities, this shifts a focus and blame to individual behaviours. However, inequalities are not always an individual choice, rather the social and political structure of society dictates these inequalities exist (449). For example, the built environment around where an individual lives can result in an obesogenic environment (450). Although in the United Kingdom, we have a free at the point of use healthcare system, access to high quality healthcare is not universal (451). In analyses using primary care data in Chapter 5, I identify inequalities in statin prescriptions in individuals who attend primary care. Here, in individuals with a QRISK score recorded in primary care data, higher educated individuals are more likely to receive statin treatment compared to lower educated individuals with equivalent underlying cardiovascular risk. These results begin to elude to wider inequalities within healthcare settings which may be independent of individual health seeking behaviours.

Not considering these wider societal determinants of inequalities places potentially unfair and unjust criticism on to individuals, rather than assigning criticism to the societal structure and interventions which result in these behaviours. However, these wider determinants are difficult to quantify and often not studied in social epidemiology. As a result of the Covid-19 pandemic and the evidence of socioeconomic and racial inequalities in disease severity and mortality, these wider determinants have gained greater prominence in discussions of disease prevention (452). Whilst this pandemic may exacerbate health inequalities in the short term (453), both from Covid-19 and other conditions (454), it has been said that this could be the

turning point for inequalities (455). Any improvements implemented at a societal level to tackle Covid-19 inequalities will inevitably improve inequalities in other health outcomes, including CVD. Where the conversation shifts from blaming individuals for poor health to blaming society for poor health, we may begin to minimise health inequalities in a lasting manner.

The wider context of the mechanisms identified in this thesis should be considered in future research and in the interpretation of the results of this thesis. Whilst in this thesis I have demonstrated a number of modifiable risk factors partly explaining educational inequalities in CVD, interventions to target these risk factors should consider the societal context in which these risk factors emerge, as well as the individual behaviours.

7.5 Future work

Important future work to this thesis would include replicating analyses in different populations or cohort studies. It would be important to carry this work out in studies with different sociodemographic characteristics, in different countries and including participants from a wide range of ancestries. However, UK Biobank is not unique in in its sociodemographic characteristic, where typically cohort studies have higher recruitment and retention on wealthier, more educated, less diverse participants (168, 169). Therefore, identifying suitable studies with appropriate data and a representative population will be important. Alternatively, methods such as inverse probability weighting could be used to account for this selection bias (456, 457).

Future work would benefit from considering the role of different indicators of SEP. As more GWAS become available for SEP indicators, such as the GWAS of income (266), it may be possible to replicate analyses in this thesis. However, the more complex and difficult a phenotype is to define, the more likely it is that the exclusion-restriction criteria in MR will be violated. Similarly, future work should consider the most appropriate definition of educational attainment in a UK context. This applies mostly to future MR analyses where the GWAS define educational attainment according to ISCED definitions (17).

As more genotypic data becomes available, particularly for related individuals, this work should be replicated to account for family population structure and dynastic effects through within-family analyses.

7.5.1 Extensions to each analysis

In Chapter 3 I demonstrated how MR mediation methods, two-step MR and MVMR could be used to improve causal inference in mediation analysis. These methods have a number of clear advantages over phenotypic methods, such as not being biased by unmeasured confounding or mis-specified models resulting in reverse causality. However, there are a number of limitations to using these methods. These methods, and their usefulness, could be improved by carrying out future methodological research to be able to account for exposure-mediation interactions, such as by allowing for four-way decomposition analysis (458). Additionally, being able to account for time-varying mediators, such as childhood BMI and adulthood BMI would improve the breadth of applications possible (406, 408).

These methods developments would be beneficial for more detailed analyses of Chapter 4. In MR mediation analyses in Chapter 4 the assumption of no exposure-mediator interaction was made. However, this may not be a valid assumption. Repeating these analyses being able to account for exposure-mediator interactions may provide a more reliable causal estimate of the role of the mediators. As previously discussed, the analyses in this chapter in particular would benefit from being replicated in a more diverse population.

In Chapter 5 I demonstrated educational inequalities in statin prescribing given underlying cardiovascular risk. It would be interesting, and important for reducing disease, to identify whether these inequalities exist for other preventative medications, such as antihypertensive drugs. One challenge of this work is controlling for underlying risk. In England and Wales, QRISK3 scores are used to determine whether preventative statin treatment should be prescribed (25, 26), providing a suitable control measure. However, other preventative medications are often not prescribed on the basis of a risk score.

Inequalities in statin use were present in a number of different data sources for the exposure (QRISK) and outcome (statin), including in self-reported statin use at baseline, in statin prescriptions 3 months prior to and after baseline, and in QRISK and QRISK2 scores recorded in primary care data with statin prescriptions. In the primary care data, there was a higher prevalence of statin prescriptions than there were QRISK scores recorded, suggesting statins are readily prescribed in the absence of risk assessment. To understand the context in which these inequalities arise, it would be important to identify why or when QRISK scores are (or are not) recorded. Inequalities have been identified in attendance to NHS health checks (where QRISK scores are routinely recorded) (359-361), where health seeking behaviours may partially explain some of these differences. However, given the inequalities in primary care

data, differences must also arise at the clinical level. Engaging with clinicians and patients may help elucidate some of the decisions made when i) carrying out risk assessments and ii) when deciding whether to prescribe statins. As UK Biobank is a highly selected population, replicating these analyses of primary care data in the clinical practice research database for example will be important to understand how widespread these inequalities are. However, as educational attainment is not routinely recorded in primary care data, consideration should be given to which measures of SEP may be appropriate to explore interactions with.

Further to inequalities in statin prescriptions (or preventative medications more widely) may be inequalities in adherence. Poor adherence to medication has been shown to increase the risk of i) stroke (459) and ii) atherosclerotic cardiovascular disease (460). Currently, it is not possible to examine medication adherence in UK Biobank. However, where register data with information on repeat prescription collected is available, for example in the Finnish Drug Prescription Register it is possible to estimate adherence (459). Future work may benefit from being expanded to additionally consider the role of adherence in educational inequalities.

In Chapter 6 I identified effect modification by educational attainment on genetic susceptibility to a number of cardiovascular risk factors. These analyses can be biased by reverse causality and unmeasured confounding, unlike MR analyses. Instrumental variable analyses could be carried out to explore this effect modification in a causal framework. For example, the RoSLA could be used to instrument educational attainment (461). An avenue for future research would be to carry out MVMR to test for causal interactions between educational attainment and cardiovascular risk factors on CVD (251, 252). However, a challenge for both of these instrumental variable approaches would be having a large enough sample size to achieve adequate statistical power. In MVMR, issues of low power can often be mitigated by using summary data MR. However, MR interaction analyses currently require individual level data, and as for any interaction, require even larger sample sizes to achieve sufficient statistical power (252). As such, a continuous outcome is preferential for these analyses, but CVD is inherently a binary outcome. Therefore, identifying a suitable outcome for analyses, whilst maintaining adequate power will be important.

7.6 Implications for public health and policy

Narrowing inequalities requires large-scale interventions to address social and structural factors, including (among other factors) improved access to housing (462), improved opportunities for work and safe income (463), access to education (including higher

education) and limiting the obesogenic environment (377, 449). However, in the absence of these changes, some targeted interventions may help reduce inequalities.

Through this work I have identified a number of opportunities for interventions. Firstly, I have identified three mediators that could be intervened on to reduce CVD, these are BMI, smoking and systolic blood pressure (Chapter 4). However, despite many attempts in recent years to reduce BMI in the population, rates of obesity keep increasing (464). Understanding barriers to reducing obesity, particularly in those of lower SEP, would be important to success (465). Although it is too soon to identify any meaningful reductions in obesity attributable to the sugar sweetened beverage tax, it has been estimated that this will work to reduce obesity equitably across strata of SEP (465, 466).

The rates of smoking have successfully reduced in recent years, but in individuals with lower SEP these reductions have been smaller and are beginning to plateau (133). Some of the most successful stop smoking campaigns have involved population wide interventions, such as increasing taxation (467, 468) and banning smoking indoors in public places (469). Whilst further population wide interventions may be beneficial, such as the recently introduced plain packaging laws, there will be fewer opportunities to intervene on this scale as more policies are introduced and the effects of future interventions may be marginal. Lower educational attainment is associated both with greater uptake of smoking and lower cessation of smoking (129, 130). Therefore, targeted interventions to reduce smoking uptake in more socioeconomically deprived groups may improve health outcomes.

Reducing systolic blood pressure may also reduce CVD in lower educated individuals. Given that high BMI increases systolic blood pressure, interventions to reduce BMI would likely also result in reductions to systolic blood pressure. Unlike BMI and smoking, systolic blood pressure is not an easily observable phenotype. Opportunistic, community blood pressure programmes may increase awareness of high blood pressure (470). Targeting these community interventions to areas of greater social deprivation may result in greater reductions of inequalities. Systolic blood pressure is also the only one of these mediators which is currently a target for medication, in the form of antihypertensives. Ensuring inequalities are not present in antihypertensive medication, which can be used to prevent CVD, will be important. Should these inequalities exit, interventions should consider how they can be reduced.

In Chapter 5, I identified inequalities in self-report statin use and in statin prescribing. Although statins are the subject of considerable debate, a recent Cochrane review found statins resulted in reductions in all-cause mortality, major vascular events and

revascularisations without any excess of adverse events (139). Therefore, they represent a cost-effective, safe method of reducing CVD. Understanding more about why and how these inequalities in statin use occur will be key to developing effective interventions to improve uptake in more socioeconomically deprived individuals. Reasons for non-uptake may include differences in health seeking behaviours, particularly for primary prevention, personal decision to not take them or implicit bias by clinicians not prescribing them.

Currently, low-dose statins are available to purchase over the counter from pharmacies (471), and ongoing discussions are being had about opening this up to high-dose statins (472). This provides a useful opportunity to address some of these inequalities by removing barriers to healthcare e.g. by not needing to make an appointment for general practice. However, this also poses as an opportunity to widen inequalities. In our sample in UK Biobank, we found a high proportion of individuals who reported taking statins had no statin prescription in their primary care records. The majority of these individuals reported using Simvastatin (the only statin available over the counter) and a large proportion were under 60 (the age of free prescriptions in England and Wales); suggesting they were likely purchasing medication over the counter rather than via an NHS prescription. These medications are not available at all pharmacies; by ensuring pharmacies in more socially deprived areas are able to provide this service, and advertise this service, uptake of statins could be encouraged. However, this should not be at the detriment to primary care visits, where inequalities are not just present in cardiovascular outcomes, but a large number of health outcomes including dementia (108), mental health outcomes (473) and types of cancer (474).

Although genetics cannot be modified, the identification of education as an effect modifier of genetic susceptibility to CVD in Chapter 6 provide implications for policy. A source of considerable debate has been the value added of genetic data; with mixed conclusions (170, 195-197). The results presented in this chapter begin to suggest that although at the population level including genetic information may not add much over existing phenotypic data, at the individual level there may be some utility to considering genetically stratified risk. It will be important, before carrying out any stratified interventions, to understand more specifically what it is about educational attainment leads to these differences and why the direction of effects differs for some risk factors. This may be knowledge gained via remaining in education or later life income (and SEP) attributable to obtaining more education.

Finally, a theme which applies to all of these mechanistic pathways is understanding when interventions may be most effective during the life course. Cardiovascular risk factors,

including those studied in this thesis, have distinct life course trajectories according to SEP (154, 475, 476). Given the long latent period between educational attainment and CVD, there is a large amount of time in which to intervene. However, different periods of the life course may lead to different outcomes. It has previously been demonstrated that even heavy smokers who quit in adulthood can improve their cardiovascular risk (477). This may equally apply to BMI, systolic blood pressure and medication.

7.7 Conclusions

In this final chapter I have summarised the key findings of my thesis. I have fully explored the strengths and limitations of all analyses presented here and how this influences the causal inference to be made from the results. I have considered potential mechanisms for the aetiology of inequalities in CVD which have not been explored in this thesis, along with making recommendations for future research. Finally, I have considered the implications of this work for public health and policy.

References

1. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. J Am Coll Cardiol. 2017;70(1):1-25.
2. Holland C, Cooper Y, Shaw R, Pattison H, Cooke R. Effectiveness and uptake of screening programmes for coronary heart disease and diabetes: a realist review of design components used in interventions. BMJ Open. 2013;3(11):e003428.
3. Marmot MG, McDowall ME. Mortality decline and widening social inequalities. Lancet. 1986;2(8501):274-6.
4. Rose G, Marmot MG. Social class and coronary heart disease. Br Heart J. 1981;45(1):13-9.
5. Marmot MG, Smith GD, Stansfeld S, Patel C, North F, Head J, et al. Health inequalities among British civil servants: the Whitehall II study. Lancet. 1991;337(8754):1387-93.
6. Kaplan GA, Keil JE. Socioeconomic factors and cardiovascular disease: a review of the literature. Circulation. 1993;88(4 Pt 1):1973-98.
7. Mishra GD, Chiesa F, Goodman A, De Stavola B, Koupil I. Socio-economic position over the life course and all-cause, and circulatory diseases mortality at age 50-87 years: results from a Swedish birth cohort. Eur J Epidemiol. 2013;28(2):139-47.
8. Davies NM, Dickson M, Davey Smith G, van den Berg GJ, Windmeijer F. The causal effects of education on health outcomes in the UK Biobank. Nature Human Behaviour. 2018;2(2):117-25.
9. Tillmann T, Vaucher J, Okbay A, Pikhart H, Peasey A, Kubinova R, et al. Education and coronary heart disease: mendelian randomisation study. BMJ. 2017;358:j3542.
10. Glymour MM, Clark CR, Patton KK. Socioeconomic Determinants of Cardiovascular Disease: Recent Findings and Future Directions. Current Epidemiology Reports. 2014;1(2):89-97.
11. UNESCO IfS. International Standard Classification of Education: ISCED 2011. http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf: UNESCO; 2012.
12. Lynch JW, Kaplan GA, Cohen RD, Tuomilehto J, Salonen JT. Do cardiovascular risk factors explain the relation between socioeconomic status, risk of all-cause mortality, cardiovascular mortality, and acute myocardial infarction? American Journal of Epidemiology. 1996;144(10):934-42.
13. Nordahl H, Rod NH, Frederiksen BL, Andersen I, Lange T, Diderichsen F, et al. Education and risk of coronary heart disease: assessment of mediation by behavioral risk factors using the additive hazards model. Eur J Epidemiol. 2013;28(2):149-57.
14. Degano IR, Marrugat J, Grau M, Salvador-Gonzalez B, Ramos R, Zamora A, et al. The association between education and cardiovascular disease incidence is mediated by hypertension, diabetes, and body mass index. Sci Rep. 2017;7(1):12370.
15. Collins R. What makes UK Biobank special? Lancet. 2012;379(9822):1173-4.
16. 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;186(9):1026-34.
17. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016;533(7604):539-42.
18. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1-22.

19. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int J Epidemiol. 2012;41(1):161-76.
20. Burgess S, Daniel RM, Butterworth AS, Thompson SG, Consortium EP-I. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. Int J Epidemiol. 2015;44(2):484-95.
21. Burgess S, Freitag DF, Khan H, Gorman DN, Thompson SG. Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. PLoS One. 2014;9(10):e108891.
22. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single sample and two-sample summary data settings. bioRxiv. 2018.
23. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099.
24. NICE. Cardiovascular disease: risk assessment and reduction, including lipid modification. www.nice.org.uk/guidance/cg1812014.
25. NICE. Cardiovascular risk assessment and lipid modification. www.nice.org.uk/guidance/qs1002015.
26. NICE. Lipid modification - CVD prevention. https://cks.nice.org.uk/lipid-modification-cvd-prevention2019.
27. Collaborators GBDCoD. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1151-210.
28. Bhatnagar P, Wickramasinghe K, Wilkins E, Townsend N. Trends in the epidemiology of cardiovascular disease in the UK. Heart. 2016;102(24):1945-52.
29. Newton JN, Briggs AD, Murray CJ, Dicker D, Foreman KJ, Wang H, et al. Changes in health in England, with analysis by English regions and areas of deprivation, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;386(10010):2257-74.
30. Wilson L, Bhatnagar P, Townsend N. Comparing trends in mortality from cardiovascular disease and cancer in the United Kingdom, 1983-2013: joinpoint regression analysis. Popul Health Metr. 2017;15(1):23.
31. Smolina K, Wright FL, Rayner M, Goldacre MJ. Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study. BMJ. 2012;344:d8059.
32. Cea-Soriano L, Fowkes FGR, Johansson S, Allum AM, Garcia Rodriguez LA. Time trends in peripheral artery disease incidence, prevalence and secondary preventive therapy: a cohort study in The Health Improvement Network in the UK. BMJ Open. 2018;8(1):e018184.
33. Davies AR, Smeeth L, Grundy EM. Contribution of changes in incidence and mortality to trends in the prevalence of coronary heart disease in the UK: 1996 2005. Eur Heart J. 2007;28(17):2142-7.
34. Lampe FC, Morris RW, Whincup PH, Walker M, Ebrahim S, Shaper AG. Is the prevalence of coronary heart disease falling in British men? Heart. 2001;86(5):499-505.
35. Walley T, Folino-Gallo P, Stephens P, Van Ganse E. Trends in prescribing and utilization of statins and other lipid lowering drugs across Europe 1997-2003. Br J Clin Pharmacol. 2005;60(5):543-51.
36. O'Keeffe AG, Nazareth I, Petersen I. Time trends in the prescription of statins for the primary prevention of cardiovascular disease in the United Kingdom: a cohort study using The Health Improvement Network primary care data. Clin Epidemiol. 2016;8:123-32.

37. Vancheri F, Backlund L, Strender LE, Godman B, Wettermark B. Time trends in statin utilisation and coronary mortality in Western European countries. BMJ Open. 2016;6(3):e010500.
38. Hardoon SL, Whincup PH, Petersen I, Capewell S, Morris RW. Trends in longer-term survival following an acute myocardial infarction and prescribing of evidenced-based medications in primary care in the UK from 1991: a longitudinal population-based study. J Epidemiol Community Health. 2011;65(9):770-4.
39. Lusis AJ, Weiss JN. Cardiovascular networks: systems-based approaches to cardiovascular disease. Circulation. 2010;121(1):157-70.
40. Bell S, Daskalopoulou M, Rapsomaniki E, George J, Britton A, Bobak M, et al. Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population based cohort study using linked health records. BMJ. 2017;356:j909.
41. Millwood IY, Walters RG, Mei XW, Guo Y, Yang L, Bian Z, et al. Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China. Lancet. 2019;393(10183):1831-42.
42. Naghavi M, Abajobir AA, Abbafati C, Abbas KM, Abd-Allah F, Abera SF, et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet. 2017;390(10100):1151-210.
43. Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ. 2004;328(7455):1519.
44. Pirie K, Peto R, Reeves GK, Green J, Beral V, Million Women Study C. The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. Lancet. 2013;381(9861):133-41.
45. Asvold BO, Bjorngaard JH, Carslake D, Gabrielsen ME, Skorpen F, Smith GD, 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(5):1458-70.
46. Li J, Siegrist J. Physical activity and risk of cardiovascular disease--a meta-analysis of prospective cohort studies. Int J Environ Res Public Health. 2012;9(2):391-407.
47. Morris JN, Heady JA. Mortality in relation to the physical activity of work: a preliminary note on experience in middle age. Br J Ind Med. 1953;10(4):245-54.
48. Davey Smith G, Shipley MJ, Batty GD, Morris JN, Marmot M. Physical activity and cause-specific mortality in the Whitehall study. Public Health. 2000;114(5):308-15.
49. Abdullah SM, Defina LF, Leonard D, Barlow CE, Radford NB, Willis BL, et al. Long-Term Association of Low-Density Lipoprotein Cholesterol With Cardiovascular Mortality in Individuals at Low 10-Year Risk of Atherosclerotic Cardiovascular Disease. Circulation. 2018;138(21):2315-25.
50. Allara E, Morani G, Carter P, Gkatzionis A, Zuber V, Foley CN, et al. Genetic Determinants of Lipids and Cardiovascular Disease Outcomes: A Wide-Angled Mendelian Randomization Investigation. Circ Genom Precis Med. 2019;12(12):e002711.
51. Burgess S, Ference BA, Staley JR, Freitag DF, Mason AM, Nielsen SF, et al. Association of LPA Variants With Risk of Coronary Disease and the Implications for Lipoprotein(a)-Lowering Therapies: A Mendelian Randomization Analysis. JAMA Cardiol. 2018;3(7):619-27.
52. Kjeldsen SE. Hypertension and cardiovascular risk: General aspects. Pharmacol Res. 2018;129:95-9.
53. Stevens SL, Wood S, Koshiaris C, Law K, Glasziou P, Stevens RJ, et al. Blood pressure variability and cardiovascular disease: systematic review and meta-analysis. BMJ. 2016;354:i4098.

54. Schultz WM, Kelli HM, Lisko JC, Varghese T, Shen J, Sandesara P, et al. Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. Circulation. 2018;137(20):2166-78.
55. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. Nat Rev Cardiol. 2015;12(11):627-42.
56. Dudina A, Cooney MT, Bacquer DD, Backer GD, Ducimetiere P, Jousilahti P, et al. Relationships between body mass index, cardiovascular mortality, and risk factors: a report from the SCORE investigators. Eur J Cardiovasc Prev Rehabil. 2011;18(5):731-42.
57. Larsson SC, Back M, Rees JMB, Mason AM, Burgess S. Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study. Eur Heart J. 2020;41(2):221-6.
58. Riaz H, Khan MS, Siddiqi TJ, Usman MS, Shah N, Goyal A, et al. Association Between Obesity and Cardiovascular Outcomes: A Systematic Review and Meta-analysis of Mendelian Randomization Studies. JAMA Netw Open. 2018;1(7):e183788.
59. Holmes MV, Lange LA, Palmer T, Lanktree MB, North KE, Almoguera B, et al. Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. Am J Hum Genet. 2014;94(2):198-208.
60. Townshend T, Lake A. Obesogenic environments: current evidence of the built and food environments. Perspect Public Health. 2017;137(1):38-44.
61. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641-9.
62. Halcox JP, Deanfield JE. Childhood origins of endothelial dysfunction. Heart. 2005;91(10):1272-4.
63. McGill HC, Jr., McMahan CA, Herderick EE, Malcom GT, Tracy RE, Strong JP. Origin of atherosclerosis in childhood and adolescence. Am J Clin Nutr. 2000;72(5 Suppl):1307S-15S.
64. Baker JL, Olsen LW, Sorensen TI. Childhood body-mass index and the risk of coronary heart disease in adulthood. N Engl J Med. 2007;357(23):2329-37.
65. McCarron P, Smith GD, Okasha M, McEwen J. Blood pressure in young adulthood and mortality from cardiovascular disease. Lancet. 2000;355(9213):1430-1.
66. Krumholz HM, Larson M, Levy D. Prognosis of left ventricular geometric patterns in the Framingham Heart Study. J Am Coll Cardiol. 1995;25(4):879-84.
67. Armstrong AC, Liu K, Lewis CE, Sidney S, Colangelo LA, Kishi S, et al. Left atrial dimension and traditional cardiovascular risk factors predict 20-year clinical cardiovascular events in young healthy adults: the CARDIA study. Eur Heart J Cardiovasc Imaging. 2014;15(8):893-9.
68. Stanfield KM, Wells JC, Fewtrell MS, Frost C, Leon DA. Differences in body composition between infants of South Asian and European ancestry: the London Mother and Baby Study. Int J Epidemiol. 2012;41(5):1409-18.
69. de Swiet M, Fayers P, Shinebourne EA. Blood pressure in first 10 years of life: the Brompton study. BMJ. 1992;304(6818):23-6.
70. Howe LD, Galobardes B, Sattar N, Hingorani AD, Deanfield J, Ness AR, et al. Are there socioeconomic inequalities in cardiovascular risk factors in childhood, and are they mediated by adiposity? Findings from a prospective cohort study. Int J Obes (Lond). 2010;34(7):1149-59.
71. Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G. Indicators of socioeconomic position (part 1). J Epidemiol Community Health. 2006;60(1):7-12.
72. Krieger N. A glossary for social epidemiology. J Epidemiol Community Health. 2001;55(10):693-700.
73. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health. 1997;18:341-78.

74. Whitehead M. The concepts and principles of equity and health. Int J Health Serv. 1992;22(3):429-45.
75. Krieger N. Historical roots of social epidemiology: socioeconomic gradients in health and contextual analysis. Int J Epidemiol. 2001;30(4):899-900.
76. Bajekal M, Scholes S, O'Flaherty M, Raine R, Norman P, Capewell S. Unequal trends in coronary heart disease mortality by socioeconomic circumstances, England 1982-2006: an analytical study. PLoS One. 2013;8(3):e59608.
77. Whitehead M, Dahlgren G. What can be done about inequalities in health? Lancet. 1991;338(8774):1059-63.
78. Mackenbach JP, Stirbu I, Roskam AJ, Schaap MM, Menvielle G, Leinsalu M, et al. Socioeconomic inequalities in health in 22 European countries. N Engl J Med. 2008;358(23):2468-81.
79. Stringhini S, Sabia S, Shipley M, Brunner E, Nabi H, Kivimaki M, et al. Association of socioeconomic position with health behaviors and mortality. JAMA. 2010;303(12):1159-66.
80. Davey Smith G, Hart C, Blane D, Gillis C, Hawthorne V. Lifetime socioeconomic position and mortality: prospective observational study. BMJ. 1997;314(7080):547-52.
81. Meneton P, Kesse-Guyot E, Mejean C, Fezeu L, Galan P, Hercberg S, et al. Unemployment is associated with high cardiovascular event rate and increased all-cause mortality in middle-aged socially privileged individuals. Int Arch Occup Environ Health. 2015;88(6):707-16.
82. Rodin D, Stirbu I, Ekholm O, Dzurova D, Costa G, Mackenbach JP, et al. Educational inequalities in blood pressure and cholesterol screening in nine European countries. J Epidemiol Community Health. 2012;66(11):1050-5.
83. van Raalte AA, Kunst AE, Lundberg O, Leinsalu M, Martikainen P, Artnik B, et al. The contribution of educational inequalities to lifespan variation. Popul Health Metr. 2012;10(1):3.
84. Huisman M, Kunst AE, Bopp M, Borgan BK, Borrell C, Costa G, et al. Educational inequalities in cause-specific mortality in middle-aged and older men and women in eight western European populations. Lancet. 2005;365(9458):493-500.
85. Mosquera PA, San Sebastian M, Waenerlund AK, Ivarsson A, Weinehall L, Gustafsson PE. Income-related inequalities in cardiovascular disease from mid-life to old age in a Northern Swedish cohort: A decomposition analysis. Soc Sci Med. 2016;149:135-44.
86. Stirbu I, Looman C, Nijhof GJ, Reulings PG, Mackenbach JP. Income inequalities in case death of ischaemic heart disease in the Netherlands: a national record-linked study. J Epidemiol Community Health. 2012;66(12):1159-66.
87. Rasmussen JN, Gislason GH, Rasmussen S, Abildstrom SZ, Schramm TK, Kober L, et al. Use of statins and beta-blockers after acute myocardial infarction according to income and education. J Epidemiol Community Health. 2007;61(12):1091-7.
88. Cubbin C, Sundquist K, Ahlen H, Johansson SE, Winkleby MA, Sundquist J. Neighborhood deprivation and cardiovascular disease risk factors: protective and harmful effects. Scand J Public Health. 2006;34(3):228-37.
89. Ramsay SE, Morris RW, Whincup PH, Subramanian SV, Papacosta AO, Lennon LT, et al. The influence of neighbourhood-level socioeconomic deprivation on cardiovascular disease mortality in older age: longitudinal multilevel analyses from a cohort of older British men. J Epidemiol Community Health. 2015;69(12):1224-31.
90. Hosseinpoor AR, Bergen N, Kunst A, Harper S, Guthold R, Rekve D, et al. Socioeconomic inequalities in risk factors for non communicable diseases in low-income and middle-income countries: results from the World Health Survey. BMC Public Health. 2012;12:912.
91. Hosseinpoor AR, Bergen N, Mendis S, Harper S, Verdes E, Kunst A, et al. Socioeconomic inequality in the prevalence of noncommunicable diseases in low- and middle-income countries: results from the World Health Survey. BMC Public Health. 2012;12:474.

92. Rosengren A, Smyth A, Rangarajan S, Ramasundarahettige C, Bangdiwala SI, AlHabib KF, et al. Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: the Prospective Urban Rural Epidemiologic (PURE) study. Lancet Glob Health. 2019;7(6):e748-e60.
93. Reid DD, Brett GZ, Hamilton PJ, Jarrett RJ, Keen H, Rose G. Cardiorespiratory disease and diabetes among middle-aged male Civil Servants. A study of screening and intervention. Lancet. 1974;1(7856):469-73.
94. Marmot MG, Rose G, Shipley M, Hamilton PJ. Employment grade and coronary heart disease in British civil servants. J Epidemiol Community Health. 1978;32(4):244-9.
95. Marmot MG, Shipley MJ, Rose G. Inequalities in death--specific explanations of a general pattern? Lancet. 1984;1(8384):1003-6.
96. van Rossum CT, Shipley MJ, van de Mheen H, Grobbee DE, Marmot MG. Employment grade differences in cause specific mortality. A 25 year follow up of civil servants from the first Whitehall study. J Epidemiol Community Health. 2000;54(3):178-84.
97. Security DoHaS. Inequalities in Health: Report of a Research Working Group. 1980.
98. Gray AM. Inequalities in health. The Black Report: a summary and comment. Int J Health Serv. 1982;12(3):349-80.
99. Marmot M, Brunner E. Cohort Profile: the Whitehall II study. Int J Epidemiol. 2005;34(2):251-6.
100. Marmot M. Fair society, healthy lives: the Marmot review; strategic review of health inequalities in England post-2010. The Marmot Review. 2010.
101. Marmot M. Health equity in England: the Marmot review 10 years on. London; 2020.
102. England PH. Public Health Outcomes Framework https://fingertips.phe.org.uk/profile/public-health-outcomes-framework/data#page/7/gid/1000049/pat/6/par/E12000004/ati/102/are/E06000015/iid/90366/age/1/sex/1/cid/4/page-options/ine-vo-0_ine-ao-1_ine-yo-3:2010:-1:-1_ine-ct-9_ine-pt-0 [Available from: https://fingertips.phe.org.uk/profile/public-health-outcomes-framework/data#page/7/gid/1000049/pat/6/par/E12000004/ati/102/are/E06000015/iid/90366/age/1/sex/1/cid/4/page-options/ine-vo-0_ine-ao-1_ine-yo-3:2010:-1:-1_ine-ct-9_ine-pt-0.
103. England PH. Health profile for England: 2018. https://www.gov.uk/government/publications/health-profile-for-england-2018; 2018 11/08/2018.
104. Lomas J, Williams J. England PH, editor. https://publichealthmatters.blog.gov.uk/2019/03/04/health-matters-ambitions-to-tackle-persisting-inequalities-in-cardiovascular-disease/2019. [cited 2020 01/08/2020]. Available from: https://publichealthmatters.blog.gov.uk/2019/03/04/health-matters-ambitions-to-tackle-persisting-inequalities-in-cardiovascular-disease/.
105. Galobardes B, Davey Smith G, Jeffreys M, McCarron P. Childhood socioeconomic circumstances predict specific causes of death in adulthood: the Glasgow student cohort study. J Epidemiol Community Health. 2006;60(6):527-9.
106. Stringhini S, Zaninotto P, Kumari M, Kivimaki M, Lassale C, Batty GD. Socio-economic trajectories and cardiovascular disease mortality in older people: the English Longitudinal Study of Ageing. Int J Epidemiol. 2018;47(1):36-46.
107. Hart CL, Smith GD, Blane D. Inequalities in mortality by social class measured at 3 stages of the lifecourse. Am J Public Health. 1998;88(3):471-4.
108. Nguyen TT, Tchetgen EJT, Kawachi I, Gilman SE, Walter S, Liu SY, et al. Instrumental variable approaches to identifying the causal effect of educational attainment on dementia risk. Ann Epidemiol. 2016;26(1):71-6.
109. Courtin E, Nafilyan V, Avendano M, Meneton P, Berkman LF, Goldberg M, et al. Longer schooling but not better off? A quasi-experimental study of the effect of compulsory schooling on biomarkers in France. Soc Sci Med. 2019;220:379-86.

110. Courtin E, Nafilyan V, Glymour M, Goldberg M, Berr C, Berkman LF, et al. Long-term effects of compulsory schooling on physical, mental and cognitive ageing: a natural experiment. J Epidemiol Community Health. 2019;73(4):370-6.
111. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. Annu Rev Psychol. 2007;58:593-614.
112. Hossin MZ, Koupil I, Falkstedt D. Early life socioeconomic position and mortality from cardiovascular diseases: an application of causal mediation analysis in the Stockholm Public Health Cohort. BMJ Open. 2019;9(6):e026258.
113. Havranek EP, Mujahid MS, Barr DA, Blair IV, Cohen MS, Cruz-Flores S, et al. Social Determinants of Risk and Outcomes for Cardiovascular Disease A Scientific Statement From the American Heart Association. Circulation. 2015;132(9):873-98.
114. Kershaw KN, Droomers M, Robinson WR, Carnethon MR, Daviglus ML, Monique Verschuren WM. Quantifying the contributions of behavioral and biological risk factors to socioeconomic disparities in coronary heart disease incidence: the MORGEN study. Eur J Epidemiol. 2013;28(10):807-14.
115. Veronesi G, Ferrario MM, Kuulasmaa K, Bobak M, Chambless LE, Salomaa V, et al. Educational class inequalities in the incidence of coronary heart disease in Europe. Heart. 2016;102(12):958-65.
116. van Oort FVA, van Lenthe FJ, Mackenbach JP. Material, psychosocial, and behavioural factors in the explanation of educational inequalities in mortality in the Netherlands. J Epidemiol Commun H. 2005;59(3):214-20.
117. Mejean C, Droomers M, van der Schouw YT, Sluijs I, Czernichow S, Grobbee DE, et al. The contribution of diet and lifestyle to socioeconomic inequalities in cardiovascular morbidity and mortality. Int J Cardiol. 2013;168(6):5190-5.
118. Bockerman P, Viinikainen J, Pulkki-Raback L, Hakulinen C, Pitkanen N, Lehtimaki T, et al. Does higher education protect against obesity? Evidence using Mendelian randomization. Prev Med. 2017;101:195-8.
119. Hagenaars SP, Gale CR, Deary IJ, Harris SE. Cognitive ability and physical health: a Mendelian randomization study. Sci Rep. 2017;7(1):2651.
120. Cohen AK, Rai M, Rehkopf DH, Abrams B. Educational attainment and obesity: a systematic review. Obes Rev. 2013;14(12):989-1005.
121. Devaux M, Sassi F, Church J, Cecchini M, Borgonovi F. Exploring the Relationship Between Education and Obesity. 2011.
122. Vogel C, Lewis D, Ntani G, Cummins S, Cooper C, Moon G, et al. The relationship between dietary quality and the local food environment differs according to level of educational attainment: A cross-sectional study. PLoS One. 2017;12(8):e0183700.
123. Worsley A, Blasche R, Ball K, Crawford D. The relationship between education and food consumption in the 1995 Australian National Nutrition Survey. Public Health Nutr. 2004;7(5):649-63.
124. Shaw BA, Spokane LS. Examining the association between education level and physical activity changes during early old age. J Aging Health. 2008;20(7):767-87.
125. Liu SY, Buka SL, Linkletter CD, Kawachi I, Kubzansky L, Loucks EB. The association between blood pressure and years of schooling versus educational credentials: test of the sheepskin effect. Ann Epidemiol. 2011;21(2):128-38.
126. Di Chiara T, Scaglione A, Corrao S, Argano C, Pinto A, Scaglione R. Association between low education and higher global cardiovascular risk. J Clin Hypertens (Greenwich). 2015;17(5):332-7.
127. Stamler R, Shipley M, Elliott P, Dyer A, Sans S, Stamler J. Higher Blood-Pressure in Adults with Less Education - Some Explanations from Intersalt. Hypertension. 1992;19(3):237-41.

128. Bann D, Fluharty M, Hardy R, Scholes S. Socioeconomic inequalities in blood pressure: co-ordinated analysis of 147,775 participants from repeated birth cohort and cross-sectional datasets, 1989 to 2016. medRxiv. 2019:2019.12.19.19015313.
129. Green MJ, Leyland AH, Sweeting H, Benzeval M. Socioeconomic position and early adolescent smoking development: evidence from the British Youth Panel Survey (1994-2008). Tob Control. 2016;25(2):203-10.
130. Huisman M, Kunst AE, Mackenbach JP. Educational inequalities in smoking among men and women aged 16 years and older in 11 European countries. Tob Control. 2005;14(2):106-13.
131. Ockene IS, Miller NH. Cigarette smoking, cardiovascular disease, and stroke: a statement for healthcare professionals from the American Heart Association. American Heart Association Task Force on Risk Reduction. Circulation. 1997;96(9):3243-7.
132. Europe WHOro. European tobacco use: Trends report 2019. http://www.euro.who.int/en/health-topics/disease-prevention/tobacco/publications/2019/european-tobacco-use-trends-report-2019-20192019.
133. Cornish D BA HM, Scanlon S. . Adult smoking habits in the UK: 2018. . 2019 2 July 2019.
134. Dale CE, Fatemifar G, Palmer TM, White J, Prieto-Merino D, Zabaneh D, et al. Causal Associations of Adiposity and Body Fat Distribution With Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus: A Mendelian Randomization Analysis. Circulation. 2017;135(24):2373-88.
135. Lyall DM, Celis-Morales C, Ward J, Iliodromiti S, Anderson JJ, Gill JMR, et al. Association of Body Mass Index With Cardiometabolic Disease in the UK Biobank: A Mendelian Randomization Study. JAMA Cardiol. 2017;2(8):882-9.
136. Nazarzadeh M, Pinho-Gomes AC, Smith Byrne K, Canoy D, Raimondi F, Ayala Solares JR, et al. Systolic Blood Pressure and Risk of Valvular Heart Disease: A Mendelian Randomization Study. JAMA Cardiol. 2019.
137. Ariansen I, Mortensen LH, Graff-Iversen S, Stigum H, Kjollesdal MK, Naess O. The educational gradient in cardiovascular risk factors: impact of shared family factors in 228,346 Norwegian siblings. BMC Public Health. 2017;17(1):281.
138. Doom JR, Mason SM, Suglia SF, Clark CJ. Pathways between childhood/adolescent adversity, adolescent socioeconomic status, and long-term cardiovascular disease risk in young adulthood. Soc Sci Med. 2017;188:166-75.
139. Taylor F, Huffman MD, Macedo AF, Moore TH, Burke M, Davey Smith G, et al. Statins for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2013(1):CD004816.
140. Trusler D. Statin prescriptions in UK now total a million each week. BMJ. 2011;343:d4350.
141. Forde I, Chandola T, Raine R, Marmot MG, Kivimaki M. Socioeconomic and ethnic differences in use of lipid-lowering drugs after deregulation of simvastatin in the UK: the Whitehall II prospective cohort study. Atherosclerosis. 2011;215(1):223-8.
142. Fleetcroft R, Schofield P, Ashworth M. Variations in statin prescribing for primary cardiovascular disease prevention: cross-sectional analysis. BMC Health Serv Res. 2014;14:414.
143. Simpson CR, Hannaford PC, Williams D. Evidence for inequalities in the management of coronary heart disease in Scotland. Heart. 2005;91(5):630-4.
144. Ashworth M, Lloyd D, Smith RS, Wagner A, Rowlands G. Social deprivation and statin prescribing: a cross-sectional analysis using data from the new UK general practitioner 'Quality and Outcomes Framework'. J Public Health (Oxf). 2007;29(1):40-7.
145. Forsberg PO, Li X, Sundquist K. Neighborhood socioeconomic characteristics and statin medication in patients with myocardial infarction: a Swedish nationwide follow-up study. BMC Cardiovasc Disord. 2016;16(1):146.

146. Thomsen RW, Johnsen SP, Olesen AV, Mortensen JT, Boggild H, Olsen J, et al. Socioeconomic gradient in use of statins among Danish patients: population-based cross-sectional study. Br J Clin Pharmacol. 2005;60(5):534-42.
147. Packham C, Robinson J, Morris J, Richards C, Marks P, Gray D. Statin prescribing in Nottingham general practices: a cross-sectional study. J Public Health Med. 1999;21(1):60-4.
148. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013;340(6139):1467-71.
149. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018.
150. Ma C, Avenell A, Bolland M, Hudson J, Stewart F, Robertson C, et al. Effects of weight loss interventions for adults who are obese on mortality, cardiovascular disease, and cancer: systematic review and meta-analysis. BMJ. 2017;359:j4849.
151. Gloy VL, Briel M, Bhatt DL, Kashyap SR, Schauer PR, Mingrone G, et al. Bariatric surgery versus non-surgical treatment for obesity: a systematic review and meta-analysis of randomised controlled trials. BMJ. 2013;347:f5934.
152. Treasury H. Soft Drinks Industry Levy comes into effect https://www.gov.uk/government/news/soft-drinks-industry-levy-comes-into-effect2018 [
153. North BJ, Sinclair DA. The intersection between aging and cardiovascular disease. Circ Res. 2012;110(8):1097-108.
154. Howe LD, Tilling K, Galobardes B, Smith GD, Gunnell D, Lawlor DA. Socioeconomic differences in childhood growth trajectories: at what age do height inequalities emerge? J Epidemiol Community Health. 2012;66(2):143-8.
155. Rose G. Incubation period of coronary heart disease. Br Med J (Clin Res Ed). 1982;284(6329):1600-1.
156. Taylor AE, Richmond RC, Palviainen T, Loukola A, Wootton RE, Kaprio J, et al. The effect of body mass index on smoking behaviour and nicotine metabolism: a Mendelian randomization study. Hum Mol Genet. 2019;28(8):1322-30.
157. Mozaffarian D, Appel LJ, Van Horn L. Components of a cardioprotective diet: new insights. Circulation. 2011;123(24):2870-91.
158. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. Cell. 2012;148(6):1242-57.
159. Vinkhuyzen AA, Wray NR, Yang J, Goddard ME, Visscher PM. Estimation and partition of heritability in human populations using whole-genome analysis methods. Annu Rev Genet. 2013;47:75-95.
160. Weng LC, Choi SH, Klarin D, Smith JG, Loh PR, Chaffin M, et al. Heritability of Atrial Fibrillation. Circ Cardiovasc Genet. 2017;10(6).
161. Bluher A, Devan WJ, Holliday EG, Nalls M, Parolo S, Bione S, et al. Heritability of young- and old-onset ischaemic stroke. Eur J Neurol. 2015;22(11):1488-91.
162. Ndumele CE, Matsushita K, Lazo M, Bello N, Blumenthal RS, Gerstenblith G, et al. Obesity and Subtypes of Incident Cardiovascular Disease. J Am Heart Assoc. 2016;5(8).
163. Banks E, Joshy G, Korda RJ, Stavreski B, Soga K, Egger S, et al. Tobacco smoking and risk of 36 cardiovascular disease subtypes: fatal and non-fatal outcomes in a large prospective Australian study. BMC Med. 2019;17(1):128.
164. Hamer M, O'Donovan G, Stamatakis E. Association between physical activity and sub-types of cardiovascular disease death causes in a general population cohort. Eur J Epidemiol. 2019;34(5):483-7.
165. Dichgans M, Malik R, Konig IR, Rosand J, Clarke R, Gretarsdottir S, et al. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. Stroke. 2014;45(1):24-36.

166. Pulit SL, Weng LC, McArdle PF, Trinquart L, Choi SH, Mitchell BD, et al. Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. Neurol Genet. 2018;4(6):e293.
167. Beaty TH, Khoury MJ. Interface of genetics and epidemiology. Epidemiol Rev. 2000;22(1):120-5.
168. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2013;42(1):111-27.
169. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Smith GD, et al. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. International Journal of Epidemiology. 2013;42(1):97-110.
170. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219-24.
171. Ainsworth HF, Shin SY, Cordell HJ. A comparison of methods for inferring causal relationships between genotype and phenotype using additional biological measurements. Genet Epidemiol. 2017;41(7):577-86.
172. Pingault JB, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nat Rev Genet. 2018;19(9):566-80.
173. Maher BS. Polygenic Scores in Epidemiology: Risk Prediction, Etiology, and Clinical Utility. Curr Epidemiol Rep. 2015;2(4):239-44.
174. Dudbridge F. Polygenic Epidemiology. Genet Epidemiol. 2016;40(4):268-72.
175. Rask-Andersen M, Karlsson T, Ek WE, Johansson A. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. PLoS Genet. 2017;13(9):e1006977.
176. Reddon H, Gueant JL, Meyre D. The importance of gene-environment interactions in human obesity. Clin Sci (Lond). 2016;130(18):1571-97.
177. Brandkvist M, Bjorngaard JH, Odegard RA, Asvold BO, Sund ER, Vie GA. Quantifying the impact of genes on body mass index during the obesity epidemic: longitudinal findings from the HUNT Study. BMJ. 2019;366:l4067.
178. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science. 1989;245(4922):1066-73.
179. Visscher PM. Sizing up human height variation. Nat Genet. 2008;40(5):489-90.
180. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9(3):e1003348.
181. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.
182. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759-72.
183. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581-90.
184. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important? JAMA. 2019;321(18):1820-1.
185. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat Commun. 2019;10(1):333.
186. Brumpton B, Sanderson E, Heilbron K, Hartwig FP, Harrison S, Vie GA, et al. Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. Nat Commun. 2020;11(1):3519.

187. NICE. Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer. https://www.nice.org.uk/guidance/cg1642019.
188. NICE. Familial hypercholesterolaemia. https://www.nice.org.uk/guidance/qs412013.
189. Abul-Husn NS, Manickam K, Jones LK, Wright EA, Hartzel DN, Gonzaga-Jauregui C, et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science. 2016;354(6319).
190. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J Am Coll Cardiol. 2018;72(16):1883-93.
191. Abraham G, Malik R, Yonova-Doing E, Salim A, Wang T, Danesh J, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. Nat Commun. 2019;10(1):5819.
192. Kullo IJ, Jouni H, Austin EE, Brown SA, Kruisselbrink TM, Isseh IN, et al. Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates: Effect on Low-Density Lipoprotein Cholesterol Levels (the MI-GENES Clinical Trial). Circulation. 2016;133(12):1181-8.
193. Knowles JW, Ashley EA. Cardiovascular disease: The rise of the genetic risk score. PLoS Med. 2018;15(3):e1002546.
194. Fulda KG, Lykens K. Ethical issues in predictive genetic testing: a public health perspective. J Med Ethics. 2006;32(3):143-7.
195. Morris RW, Cooper JA, Shah T, Wong A, Drenos F, Engmann J, et al. Marginal role for 53 common genetic variants in cardiovascular disease prediction. Heart. 2016;102(20):1640-7.
196. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA. 2020;323(7):636-45.
197. Mosley JD, Gupta DK, Tan J, Yao J, Wells QS, Shaffer CM, et al. Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. JAMA. 2020;323(7):627-35.
198. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet. 2005;6(4):287-98.
199. Khoury MJ, Wagener DK. Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors. Am J Hum Genet. 1995;56(4):835-44.
200. Mackenbach JP. Genetics and health inequalities: hypotheses and controversies. J Epidemiol Community Health. 2005;59(4):268-73.
201. Rothman KJ. Synergy and antagonism in cause-effect relationships. Am J Epidemiol. 1974;99(6):385-8.
202. 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.
203. Hamrefors V, Hedblad B, Hindy G, Smith JG, Almgren P, Engstrom G, et al. Smoking modifies the associated increased risk of future cardiovascular disease by genetic variation on chromosome 9p21. PLoS One. 2014;9(1):e85893.
204. Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry. 2011;168(10):1041-9.
205. Porta M. A dictionary of epidemiology: Oxford university press; 2014.
206. Hill AB. The Environment and Disease: Association or Causation? Proc R Soc Med. 1965;58:295-300.
207. Pearl J. An introduction to causal inference. Int J Biostat. 2010;6(2):Article 7.
208. Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health. 2004;58(4):265-71.
209. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol. 2019;34(3):211-9.

210. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health. 2001;22:189-212.
211. VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat. 2013;41(1):196-220.
212. Kirkwood BR, Sterne JAC, Blackwell S. Essential medical statistics. Malden, MA: Blackwell Science; 2017.
213. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. Am J Epidemiol. 2000;151(5):531-9.
214. Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med. 1988;7(1-2):257-74.
215. Coggon D, Rose G, Barker DJP. Epidemiology for the Uninitiated: BMJ Publishing Group; 1993.
216. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med. 1998;55(10):651-6.
217. Rothman KJ, Greenland S, Lash TL. Modern epidemiology: Lippincott Williams & Wilkins; 2008.
218. Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. BMJ. 2010;340:c2289.
219. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol. 2007;166(6):646-55.
220. Kopec JA, Esdaile JM. Bias in case-control studies. A review. J Epidemiol Community Health. 1990;44(3):179-86.
221. Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol. 2017.
222. Jones DS, Podolsky SH. The history and fate of the gold standard. Lancet. 2015;385(9977):1502-3.
223. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ. 1998;316(7126):201.
224. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
225. Wright S. The method of path coefficients. Ann Math Stat. 1934;5:161-215.
226. Alwin DF, Hauser RM. Decomposition of Effects in Path Analysis. Am Sociol Rev. 1975;40(1):37-47.
227. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51(6):1173-82.
228. Vanderweele TJ. Controlled Direct and Mediated Effects: Definition, Identification and Bounds. Scand J Stat. 2011;38(3):551-63.
229. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. Int J Epidemiol. 2013;42(5):1511-9.
230. Hafeman DM. Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. Am J Epidemiol. 2011;174(6):710-7.
231. Cole SR, Hernan MA. Fallibility in estimating direct effects. Int J Epidemiol. 2002;31(1):163-5.
232. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39(2):417-20.
233. Vanderweele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology. 2014;25(2):300-6.

234. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol. 2017;46(2):756-62.
235. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550-60.
236. Pearl J. Interpretation and identification of causal mediation. Psychol Methods. 2014;19(4):459-81.
237. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992;3(2):143-55.
238. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychol Methods. 2013;18(2):137-50.
239. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. Epidemiology. 2013;24(2):224-32.
240. VanderWeele TJ. A unification of mediation and interaction: a 4-way decomposition. Epidemiology. 2014;25(5):749-61.
241. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. Annu Rev Public Health. 2016;37:17-32.
242. Naimi AI, Kaufman JS, MacLehose RF. Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects. Int J Epidemiol. 2014;43(5):1656-61.
243. Lawlor DA, Wade, K., Borges, M. C., Palmer, T., Hartwig, F. P., Hemani, G., & Bowden, J. A Mendelian Randomization dictionary: Useful definitions and descriptions for undertaking, understanding and interpreting Mendelian Randomization studies. OSF Preprints. 2019.
244. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr. 2016;103(4):965-78.
245. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27(8):1133-63.
246. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. Int J Epidemiol. 2016;45(3):908-15.
247. Pierce B, Burgess S. Efficient Design for Mendelian Randomization Studies: Subsample and Two-Sample Instrumental Variable Estimators. American Journal of Epidemiology. 2013;177:S117-S.
248. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23(R1):R89-98.
249. Ference BA, Majeed F, Penumetcha R, Flack JM, Brook RD. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 x 2 factorial Mendelian randomization study. J Am Coll Cardiol. 2015;65(15):1552-61.
250. Carter AR, Borges MC, Benn M, Tybjaerg-Hansen A, Davey Smith G, Nordestgaard BG, et al. Combined Association of Body Mass Index and Alcohol Consumption With Biomarkers for Liver Injury and Incidence of Liver Disease: A Mendelian Randomization Study. JAMA Netw Open. 2019;2(3):e190305.
251. Rees JMB, Foley CN, Burgess S. Factorial Mendelian randomization: using genetic variants to assess interactions. Int J Epidemiol. 2019.
252. North TL, Davies NM, Harrison S, Carter AR, Hemani G, Sanderson E, et al. Using Genetic Instruments to Estimate Interactions in Mendelian Randomization Studies. Epidemiology. 2019;30(6):e33-e5.

253. Burgess S, Thompson DJ, Rees JMB, Day FR, Perry JR, Ong KK. Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer. Genetics. 2017;207(2):481-7.
254. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. Hum Mol Genet. 2018;27(R2):R195-R208.
255. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512-25.
256. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016;40(4):304-14.
257. von Hinke Kessler Scholder S, Smith GD, Lawlor DA, Propper C, Windmeijer F. Mendelian randomization: the use of genes in instrumental variable analyses. Health Econ. 2011;20(8):893-6.
258. Brion MJ, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. Int J Epidemiol. 2013;42(5):1497-501.
259. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. Int J Epidemiol. 2016;45(6):1866-86.
260. Munafo MR, Davey Smith G. Robust research needs many lines of evidence. Nature. 2018;553(7689):399-401.
261. Blane D. K-IK, d'Errico A., Bartley M., Montgomery S. Social-biological transitions: how does the social become biological? Longitudinal and Life Course Studies: International Journal. 2013.
262. Kelly-Irving M, Tophoven S, Blane D. Life course research: new opportunities for establishing social and biological plausibility. Int J Public Health. 2015;60(6):629-30.
263. Krieger N. Theories for social epidemiology in the 21st century: an ecosocial perspective. International Journal of Epidemiology. 2001;30(4):668-77.
264. Fraga S, Marques-Vidal P, Vollenweider P, Waeber G, Guessous I, Paccaud F, et al. Association of socioeconomic status with inflammatory markers: a two cohort comparison. Prev Med. 2015;71:12-9.
265. Tyrrell J, Jones SE, Beaumont R, Astley CM, Lovell R, Yaghootkar H, et al. Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. BMJ. 2016;352:i582.
266. Hill WD, Davies NM, Ritchie SJ, Skene NG, Bryois J, Bell S, et al. Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. Nat Commun. 2019;10(1):5741.
267. Davies NM, Dickson M, Davey Smith G, Windmeijer F, van den Berg GJ. The Causal Effects of Education on Adult Health, Mortality and Income: Evidence from Mendelian Randomization and the Raising of the School Leaving Age. 2019.
268. Gage SH, Bowden J, Davey Smith G, Munafo MR. Investigating causality in associations between education and smoking: a two-sample Mendelian randomization study. Int J Epidemiol. 2018;47(4):1131-40.
269. Rosoff DB, Clarke TK, Adams MJ, McIntosh AM, Davey Smith G, Jung J, et al. Educational attainment impacts drinking behaviors and risk for alcohol dependence: results from a two-sample Mendelian randomization study with ~780,000 participants. Mol Psychiatry. 2019.
270. Deary IJ, Strand S, Smith P, Fernandes C. Intelligence and educational achievement. Intelligence. 2007;35(1):13-21.
271. Sanderson E, Davey Smith G, Munafo M. The effect of education and general cognitive ability on smoking: A Mendelian randomisation study. bioRxiv. 2018.

272. Davies NM, Hill WD, Anderson EL, Sanderson E, Deary IJ, Davey Smith G. Multivariable two-sample Mendelian randomization estimates of the effects of intelligence and education on health. Elife. 2019;8.
273. Gill D, Efstathiadou A, Cawood K, Tzoulaki I, Dehghan A. Education protects against coronary heart disease and stroke independently of cognitive function: evidence from Mendelian randomization. Int J Epidemiol. 2019;48(5):1468-77.
274. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods. 2002;7(1):83-104.
275. Blakely T, McKenzie S, Carter K. Misclassification of the mediator matters when estimating indirect effects. J Epidemiol Community Health. 2013;67(5):458-66.
276. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. Stat Interface. 2009;2(4):457-68.
277. Vansteelandt S, Vanderweele TJ. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. Biometrics. 2012;68(4):1019-27.
278. VanderWeele TJ, Valeri L, Ogburn EL. The role of measurement error and misclassification in mediation analysis: mediation and measurement error. Epidemiology. 2012;23(4):561-4.
279. Davey Smith G, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Med. 2007;4(12):e352.
280. Sanderson E. Multivariable Mendelian Randomization and Mediation. Cold Spring Harb Perspect Med. 2020.
281. Richmond RC, Hemani G, Tilling K, Davey Smith G, Relton CL. Challenges and novel approaches for investigating molecular mediation. Hum Mol Genet. 2016;25(R2):R149-R56.
282. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am J Epidemiol. 2015;181(4):251-60.
283. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol. 2018.
284. Varbo A, Benn M, Smith GD, Timpson NJ, Tybjaerg-Hansen A, Nordestgaard BG. Remnant cholesterol, low-density lipoprotein cholesterol, and blood pressure as mediators from obesity to ischemic heart disease. Circ Res. 2015;116(4):665-73.
285. Xu L, Borges MC, Hemani G, Lawlor DA. The role of glycaemic and lipid risk factors in mediating the effect of BMI on coronary heart disease: a two-step, two-sample Mendelian randomisation study. Diabetologia. 2017;60(11):2210-20.
286. Marouli E, Del Greco MF, Astley CM, Yang J, Ahmad S, Berndt SI, et al. Mendelian randomisation analyses find pulmonary factors mediate the effect of height on coronary artery disease. Commun Biol. 2019;2:119.
287. Carter AR, Gill D, Davies NM, Taylor AE, Tillmann T, Vaucher J, et al. Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study. BMJ. 2019;365:l1855.
288. Mitchell R, Hemani, G, Dudding, T, Paternoster, L. UK Biobank Genetic Data: MRC-IEU Quality Control, Version 1. 2017.
289. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. Genet Epidemiol. 2016;40(7):597-608.
290. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.
291. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274-83.

292. Wade KH, Carslake D, Sattar N, Davey Smith G, Timpson NJ. BMI and Mortality in UK Biobank: Revised Estimates Using Mendelian Randomization. Obesity (Silver Spring). 2018;26(11):1796-806.
293. Biobank U. Hospital inpatient data. https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/HospitalEpisodeStatistics.pdf; 2019 September 2019.
294. VanderWeele TJa. Explanation in causal inference : methods for mediation and interaction: New York, NY : Oxford University Press, [2015]; 2015.
295. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. Epidemiol Methods. 2014;2(1):95-115.
296. Sanderson E, Windmeijer F. A weak instrument [Formula: see text]-test in linear IV models with multiple endogenous variables. J Econom. 2016;190(2):212-21.
297. Rees JMB, Wood AM, Burgess S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. Stat Med. 2017;36(29):4705-18.
298. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Stat Sci. 1999;14(1):29-46.
299. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Stat Methods Med Res. 2016;25(5):1925-37.
300. Burgess S, Thompson SG. Bias in causal estimates from Mendelian randomization studies with weak instruments. Stat Med. 2011;30(11):1312-23.
301. Burgess S, Thompson SG. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. Stat Med. 2012;31(15):1582-600.
302. Sanderson E, Spiller W, Bowden J. Testing and Correcting for Weak and Pleiotropic Instruments in Two-Sample Multivariable Mendelian Randomisation. bioRxiv. 2020:2020.04.02.021980.
303. Loeys T, Moerkerke B, Vansteelandt S. A cautionary note on the power of the test for the indirect effect in mediation analysis. Front Psychol. 2014;5:1549.
304. Dyer AR, Stamler J, Shekelle RB, Schoenberger J. The relationship of education to blood pressure: findings on 40,000 employed Chicagoans. Circulation. 1976;54(6):987-92.
305. Kiely DK, Gross AL, Kim DH, Lipsitz LA. The association of educational attainment and SBP among older community-living adults: the Maintenance of Balance, Independent Living, Intellect and Zest in the Elderly (MOBILIZE) Boston Study. J Hypertens. 2012;30(8):1518-25.
306. Conen D, Glynn RJ, Ridker PM, Buring JE, Albert MA. Socioeconomic status, blood pressure progression, and incident hypertension in a prospective cohort of female health professionals. Eur Heart J. 2009;30(11):1378-84.
307. Holmes MV, 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;36(9):539-50.
308. Hamad R, Nguyen TT, Bhattacharya J, Glymour MM, Rehkopf DH. Educational attainment and cardiovascular disease in the United States: A quasi-experimental instrumental variables analysis. PLoS Med. 2019;16(6):e1002834.
309. Zuber V, Colijn JM, Klaver C, Burgess S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. Nat Commun. 2020;11(1):29.
310. Burgess S, Labrecque JA. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. Eur J Epidemiol. 2018;33(10):947-52.
311. Labrecque JA, Swanson SA. Mendelian randomization with multiple exposures: the importance of thinking about time. Int J Epidemiol. 2019.

312. Richardson TG SE, Elsworth B, Tilling K, Davey Smith G. Can the impact of childhood adiposity on disease risk be reversed? A Mendelian randomization study. BMJ. In press.
313. Zheng J, Baird D, Borges MC, Bowden J, Hemani G, Haycock P, et al. Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep. 2017;4(4):330-45.
314. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Stat Med. 2005;24(19):2911-35.
315. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet. 2018;50(10):1412-25.
316. Warren HR, Evangelou E, Cabrera CP, Gao H, Ren M, Mifsud B, et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. Nat Genet. 2017;49(3):403-15.
317. Wootton RE, Richmond RC, Stuijfzand BG, Lawn RB, Sallis HM, Taylor GMJ, et al. Causal effects of lifetime smoking on risk for depression and schizophrenia: Evidence from a Mendelian randomisation study. bioRxiv. 2018.
318. MRC IEU UK Biobank GWAS pipeline version 1, (2017).
319. Hemani G, Zheng J, Wade KH, Laurin C, Elsworth B, Burgess S, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv. 2016.
320. 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. 2019:1-9.
321. 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.
322. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015;47(10):1121-30.
323. Okbay A, Baselmans BM, De Neve JE, Turley P, Nivard MG, Fontana MA, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat Genet. 2016;48(6):624-33.
324. Palmer TM, Holmes MV, Keating BJ, Sheehan NA. Correcting the Standard Errors of 2-Stage Residual Inclusion Estimators for Mendelian Randomization Studies. Am J Epidemiol. 2017;186(9):1104-14.
325. Thompson JR, Minelli C, Del Greco MF. Mendelian Randomization using Public Data from Genetic Consortia. Int J Biostat. 2016;12(2).
326. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37(7):658-65.
327. Christopher F Baum MES, Steven Stillman. IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation. Statistical Software Components. 2002.
328. Spiller W, Davies NM, Palmer TM. Software application profile: mrrobust—a tool for performing two-sample summary Mendelian randomization analyses. International Journal of Epidemiology. 2018:dyy195-dyy.
329. Luepker RV, Rosamond WD, Murphy R, Sprafka JM, Folsom AR, McGovern PG, et al. Socioeconomic status and coronary heart disease risk factor trends. The Minnesota Heart Survey. Circulation. 1993;88(5 Pt 1):2172-9.
330. Garrison RJ, Gold RS, Wilson PW, Kannel WB. Educational attainment and coronary heart disease risk: the Framingham Offspring Study. Prev Med. 1993;22(1):54-64.
331. Mayer O, Jr., Simon J, Heidrich J, Cokkinos DV, De Bacquer D, Group EIS. Educational level and risk profile of cardiac patients in the EUROASPIRE II substudy. J Epidemiol Community Health. 2004;58(1):47-52.

332. Jacobsen BK, Thelle DS. Risk factors for coronary heart disease and level of education. The Tromso Heart Study. Am J Epidemiol. 1988;127(5):923-32.
333. Matthews KA, Kelsey SF, Meilahn EN, Kuller LH, Wing RR. Educational attainment and behavioral and biologic risk factors for coronary heart disease in middle-aged women. Am J Epidemiol. 1989;129(6):1132-44.
334. Lynch J, Smith GD, Harper S, Hillemeier M, Ross N, Kaplan GA, et al. Is income inequality a determinant of population health? Part 1. A systematic review. Milbank Q. 2004;82(1):5-99.
335. Loucks EB, Gilman SE, Howe CJ, Kawachi I, Kubzansky LD, Rudd RE, et al. Education and coronary heart disease risk: potential mechanisms such as literacy, perceived constraints, and depressive symptoms. Health Educ Behav. 2015;42(3):370-9.
336. Taylor AE, Davies NM, Ware JJ, VanderWeele T, Smith GD, Munafo MR. Mendelian randomization in health research: using appropriate genetic variants and avoiding biased estimates. Econ Hum Biol. 2014;13:99-106.
337. Pulcu E. Self-report distortions of puffing topography in daily smokers. J Health Psychol. 2016;21(8):1644-54.
338. Gkatzionis A, Newcombe PJ. Bayesian variable selection for Mendelian randomization. Genetic Epidemiology. 2018;42(7):701-.
339. Hughes RA, Davies NM, Smith GD, Tilling K. Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis. Epidemiology. 2019;30(3):350-7.
340. O'Keeffe AG, Petersen I, Nazareth I. Initiation rates of statin therapy for the primary prevention of cardiovascular disease: an assessment of differences between countries of the UK and between regions within England. BMJ Open. 2015;5(3):e007207.
341. Ramsay SE, Morris RW, Papacosta O, Lennon LT, Thomas MC, Whincup PH. Secondary prevention of coronary heart disease in older British men: extent of inequalities before and after implementation of the National Service Framework. J Public Health-Uk. 2005;27(4):338-43.
342. Mitchell AJ, Selmes T. Why don't patients attend their appointments? Maintaining engagement with psychiatric services. Advances in Psychiatric Treatment. 2018;13(6):423-34.
343. Campbell K MA, McCartney G, McCullough S (NHS Health Scotland). Who is least likely to attend? An analysis of outpatient appointment 'Did Not Attend' (DNA) data in Scotland. NHS Health Scotland; 2015.
344. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335(7611):136.
345. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475-82.
346. Farnier M, Portal JJ, Maigret P. Efficacy of atorvastatin compared with simvastatin in patients with hypercholesterolemia. J Cardiovasc Pharmacol Ther. 2000;5(1):27-32.
347. Insull W, Kafonek S, Goldner D, Zieve F. Comparison of efficacy and safety of atorvastatin (10mg) with simvastatin (10mg) at six weeks. ASSET Investigators. Am J Cardiol. 2001;87(5):554-9.
348. Karalis DG, Ross AM, Vacari RM, Zarren H, Scott R. Comparison of efficacy and safety of atorvastatin and simvastatin in patients with dyslipidemia with and without coronary heart disease. Am J Cardiol. 2002;89(6):667-71.
349. Moon JC, Bogle RG. Switching statins. BMJ. 2006;332(7554):1344-5.
350. Biobank U. Primary Care Linked Data. http://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/primary_care_data.pdf; 2019 September 2019.

351. Biobank U. Biomarker assay quality procedures: approaches used to minimise systematic and random errors (and the wider epidemiological implications). http://biobank.ctsu.ox.ac.uk/showcase/showcase/docs/biomarker_issues.pdf; 2019.
352. Royston P. Multiple Imputation of Missing Values. The Stata Journal. 2004;4(3):227-41.
353. Spratt M, Carpenter J, Sterne JA, Carlin JB, Heron J, Henderson J, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. 2010;172(4):478-87.
354. Tilling K, Williamson EJ, Spratt M, Sterne JA, Carpenter JR. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. J Clin Epidemiol. 2016;80:107-15.
355. Zhao M, Vaartjes I, Graham I, Grobbee D, Spiering W, Klipstein-Grobusch K, et al. Sex differences in risk factor management of coronary heart disease across three regions. Heart. 2017;103(20):1587-94.
356. Millett ERC, Peters SAE, Woodward M. Sex differences in risk factors for myocardial infarction: cohort study of UK Biobank participants. BMJ. 2018;363:k4247.
357. Cookson R, Propper C, Asaria M, Raine R. Socio-Economic Inequalities in Health Care in England. Fisc Stud. 2016;37(3-4):371-403.
358. England PH. Using the world leading NHS Health Check programme to prevent CVD 2018 [updated 24/01/2018. Available from: https://www.gov.uk/government/publications/using-the-nhs-health-check-programme-to-prevent-cvd/using-the-world-leading-nhs-health-check-programme-to-prevent-cvd.
359. Bunten A, Porter L, Gold N, Bogle V. A systematic review of factors influencing NHS health check uptake: invitation methods, patient characteristics, and the impact of interventions. BMC Public Health. 2020;20(1):93.
360. Dalton AR, Bottle A, Okoro C, Majeed A, Millett C. Uptake of the NHS Health Checks programme in a deprived, culturally diverse setting: cross-sectional study. J Public Health (Oxf). 2011;33(3):422-9.
361. Wilson R, Kuh D, Stafford M. Variations of health check attendance in later life: results from a British birth cohort study. BMC Public Health. 2019;19(1):1518.
362. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53.
363. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. BMJ. 2009;339:b2584.
364. Brindle PM, McConnachie A, Upton MN, Hart CL, Davey Smith G, Watt GC. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. Br J Gen Pract. 2005;55(520):838-45.
365. Tunstall-Pedoe H, Woodward M, estimation Sgor. By neglecting deprivation, cardiovascular risk scoring will exacerbate social gradients in disease. Heart. 2006;92(3):307-10.
366. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). Heart. 2007;93(2):172-6.
367. Liew SM, Blacklock C, Hislop J, Glasziou P, Mant D. Cardiovascular risk scores: qualitative study of how primary care practitioners understand and use them. Br J Gen Pract. 2013;63(611):e401-7.
368. Finnikin S, Ryan R, Marshall T. Statin initiations and QRISK2 scoring in UK general practice: a THIN database study. Br J Gen Pract. 2017;67(665):e881-e7.
369. Elliott P, Peakman TC, Biobank UK. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol. 2008;37(2):234-44.
370. VanderWeele TJ, Knol MJ. A tutorial on interaction. Epidemiologic Methods. 2014;3(1):33-72.

371. Schroder SL, Fink A, Richter M. Socioeconomic differences in experiences with treatment of coronary heart disease: a qualitative study from the perspective of elderly patients. BMJ Open. 2018;8(11):e024151.
372. Sing CF, Stengard JH, Kardia SL. Genes, environment, and cardiovascular disease. Arterioscler Thromb Vasc Biol. 2003;23(7):1190-6.
373. Schmidt B, Frolich S, Dragano N, Frank M, Eisele L, Pechlivanis S, et al. Socioeconomic Status Interacts with the Genetic Effect of a Chromosome 9p21.3 Common Variant to Influence Coronary Artery Calcification and Incident Coronary Events in the Heinz Nixdorf Recall Study (Risk Factors, Evaluation of Coronary Calcium, and Lifestyle). Circ Cardiovasc Genet. 2017;10(2).
374. Humphries SE, Talmud PJ, Hawe E, Bolla M, Day IN, Miller GJ. Apolipoprotein E4 and coronary heart disease in middle-aged men who smoke: a prospective study. Lancet. 2001;358(9276):115-9.
375. Lahoz C, Schaefer EJ, Cupples LA, Wilson PW, Levy D, Osgood D, et al. Apolipoprotein E genotype and cardiovascular disease in the Framingham Heart Study. Atherosclerosis. 2001;154(3):529-37.
376. Lindi VI, Uusitupa MI, Lindstrom J, Louheranta A, Eriksson JG, Valle TT, et al. Association of the Pro12Ala polymorphism in the PPAR-gamma2 gene with 3-year incidence of type 2 diabetes and body weight change in the Finnish Diabetes Prevention Study. Diabetes. 2002;51(8):2581-6.
377. Tyrrell J, Wood AR, Ames RM, Yaghootkar H, Beaumont RN, Jones SE, et al. Gene-obesogenic environment interactions in the UK Biobank study. Int J Epidemiol. 2017;46(2):559-75.
378. Amin V, Bockerman P, Viinikainen J, Smart MC, Bao YC, Kumari M, et al. Gene-environment interactions between education and body mass: Evidence from the UK and Finland. Soc Sci Med. 2017;195:12-6.
379. Roskam AJ, Kunst AE, Van Oyen H, Demarest S, Klumbiene J, Regidor E, et al. Comparative appraisal of educational inequalities in overweight and obesity among adults in 19 European countries. Int J Epidemiol. 2010;39(2):392-404.
380. Flowers E, Froelicher ES, Aouizerat BE. Gene-environment interactions in cardiovascular disease. Eur J Cardiovasc Nurs. 2012;11(4):472-8.
381. Ahmad OS, Morris JA, Mujammami M, Forgetta V, Leong A, Li R, et al. A Mendelian randomization study of the effect of type-2 diabetes on coronary heart disease. Nat Commun. 2015;6:7060.
382. Ference BA, Ginsberg HN, Graham I, Ray KK, Packard CJ, Bruckert E, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. Eur Heart J. 2017;38(32):2459-72.
383. Larsson SC, Mason AM, Back M, Klarin D, Damrauer SM, Million Veteran P, et al. Genetic predisposition to smoking in relation to 14 cardiovascular diseases. Eur Heart J. 2020.
384. Ettehad D, Emdin CA, Kiran A, Anderson SG, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. Lancet. 2016;387(10022):957-67.
385. Howe LJ, Lawson DJ, Davies NM, St Pourcain B, Lewis SJ, Davey Smith G, et al. Genetic evidence for assortative mating on alcohol consumption in the UK Biobank. Nat Commun. 2019;10(1):5039.
386. 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(2):237-44.
387. Harris R, Bradburn M, Deeks J, Harbord R, Altman D, Sterne J. metan: fixed- and random-effects meta-analysis. Stata Journal. 2008;8(1):3-28.

388. Ellinor PT, Lunetta KL, Albert CM, Glazer NL, Ritchie MD, Smith AV, et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. Nat Genet. 2012;44(6):670-5.
389. Scott RA, Scott LJ, Magi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66(11):2888-902.
390. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet. 2018;50(4):524-37.
391. Ruth Mitchell, Gibran Hemani, Tom Dudding, Laura Corbin, Sean Harrison, Paternoster L. UK Biobank Genetic Data: MRC-IEU Quality Control, version 2. 2019.
392. Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, Pasquale LR, et al. Sugar-sweetened beverages and genetic risk of obesity. N Engl J Med. 2012;367(15):1387-96.
393. Reddon H, Gerstein HC, Engert JC, Mohan V, Bosch J, Desai D, et al. Physical activity and genetic predisposition to obesity in a multiethnic longitudinal study. Sci Rep. 2016;6:18672.
394. Qi Q, Li Y, Chomistek AK, Kang JH, Curhan GC, Pasquale LR, et al. Television watching, leisure time physical activity, and the genetic predisposition in relation to body mass index in women and men. Circulation. 2012;126(15):1821-7.
395. Qi Q, Chu AY, Kang JH, Huang J, Rose LM, Jensen MK, et al. Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies. BMJ. 2014;348:g1610.
396. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17(10):1520-8.
397. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13(2):135-45.
398. Sanderson E, Davey Smith G, Bowden J, Munafo MR. Mendelian randomisation analysis of the effect of educational attainment and cognitive ability on smoking behaviour. Nat Commun. 2019;10(1):2949.
399. Jamieson E, Korologou-Linden R, Wootton RE, Guyatt AL, Battram T, Burrows K, et al. Smoking, DNA Methylation, and Lung Function: a Mendelian Randomization Analysis to Investigate Causal Pathways. Am J Hum Genet. 2020;106(3):315-26.
400. Howe LD, Kanayalal R, Harrison S, Beaumont RN, Davies AR, Frayling TM, et al. Effects of body mass index on relationship status, social contact and socio-economic position: Mendelian randomization and within-sibling study in UK Biobank. Int J Epidemiol. 2019.
401. Gage SH, Sallis HH, Lassi G, Wootton RE, Mokrysz C, Davey Smith G, et al. Does Smoking Cause Lower Educational Attainment and General Cognitive Ability? Triangulation of causal evidence using multiple study designs. medRxiv. 2019:19009365.
402. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol. 2017;46(4):1093-i.
403. Hingorani A, Humphries S. Nature's randomised trials. Lancet. 2005;366(9501):1906-8.
404. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601.
405. Labrecque J, Swanson SA. Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. Curr Epidemiol Rep. 2018;5(3):214-20.
406. Labrecque JA, Swanson SA. Interpretation and Potential Biases of Mendelian Randomization Estimates With Time-Varying Exposures. Am J Epidemiol. 2019;188(1):231-8.
407. Floras JS. Blood pressure variability: a novel and important risk factor. Can J Cardiol. 2013;29(5):557-63.

408. Richardson TG, Sanderson E, Elsworth B, Tilling K, Davey Smith G. Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. BMJ. 2020;369:m1203.
409. Vanderweele TJ. The sign of the bias of unmeasured confounding. Biometrics. 2008;64(3):702-6.
410. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009;20(4):488-95.
411. Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. BMJ. 2019;365:l737.
412. Hartwig FP, Davies NM, Davey Smith G. Bias in Mendelian randomization due to assortative mating. Genet Epidemiol. 2018;42(7):608-20.
413. Yengo L, Robinson MR, Keller MC, Kemper KE, Yang Y, Trzaskowski M, et al. Imprint of assortative mating on the human genome. Nat Hum Behav. 2018;2(12):948-54.
414. Kong A, Thorleifsson G, Frigge ML, Vilhjalmsson BJ, Young AI, Thorgeirsson TE, et al. The nature of nurture: Effects of parental genotypes. Science. 2018;359(6374):424-8.
415. Morris TT, Davies NM, Hemani G, Smith GD. Population phenomena inflate genetic associations of complex social traits. Sci Adv. 2020;6(16):eaay0328.
416. Krapohl E, Plomin R. Genetic link between family socioeconomic status and children's educational achievement estimated from genome-wide SNPs. Mol Psychiatry. 2016;21(3):437-43.
417. Halpern-Manners A, Helgertz J, Warren JR, Roberts E. The Effects of Education on Mortality: Evidence From Linked U.S. Census and Administrative Mortality Data. Demography. 2020.
418. Davies NM, Howe LJ, Brumpton B, Havdahl A, Evans DM, Davey Smith G. Within family Mendelian randomization studies. Hum Mol Genet. 2019;28(R2):R170-R9.
419. Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, Crawford A, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? Hum Genet. 2020;139(1):23-41.
420. Brumpton B, Sanderson E, Hartwig FP, Harrison S, Vie GÅ, Cho Y, et al. Within-family studies for Mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. bioRxiv. 2019.
421. Ekholm O. Influence of the recall period on self-reported alcohol intake. Eur J Clin Nutr. 2004;58(1):60-3.
422. Boniface S, Kneale J, Shelton N. Drinking pattern is more strongly associated with under-reporting of alcohol consumption than socio-demographic factors: evidence from a mixed-methods study. BMC Public Health. 2014;14:1297.
423. Xue A, Jiang L, Zhu Z, Wray NR, Visscher PM, Zeng J, et al. Genome-wide analyses of behavioural traits biased by misreports and longitudinal changes. medRxiv. 2020:2020.06.15.20131284.
424. Batty GD, Gale CR, Kivimaki M, Deary IJ, Bell S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. BMJ. 2020;368:m131.
425. Swanson JM. The UK Biobank and selection bias. Lancet. 2012;380(9837):110.
426. Lawlor DA, Lewcock M, Rena-Jones L, Rollings C, Yip V, Smith D, et al. The second generation of The Avon Longitudinal Study of Parents and Children (ALSPAC-G2): a cohort profile. Wellcome Open Res. 2019;4:36.
427. Strandhagen E, Berg C, Lissner L, Nunez L, Rosengren A, Toren K, et al. Selection bias in a population survey with registry linkage: potential effect on socioeconomic gradient in cardiovascular risk. Eur J Epidemiol. 2010;25(3):163-72.

428. Strandberg TE, Salomaa VV, Vanhanen HT, Naukkarinen VA, Sarna SJ, Miettinen TA. Mortality in participants and non-participants of a multifactorial prevention study of cardiovascular diseases: a 28 year follow up of the Helsinki Businessmen Study. Br Heart J. 1995;74(4):449-54.
429. Enzenbach C, Wicklein B, Wirkner K, Loeffler M. Evaluating selection bias in a population-based cohort study with low baseline participation: the LIFE-Adult-Study. BMC Med Res Methodol. 2019;19(1):135.
430. Taylor AE, Jones HJ, Sallis H, Euesden J, Stergiakouli E, Davies NM, et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2018;47(4):1207-16.
431. Howe LD, Tilling K, Galobardes B, Lawlor DA. Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities. Epidemiology. 2013;24(1):1-9.
432. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol. 2019;48(4):1294-304.
433. Tyrrell J, Zheng J, Beaumont R, Hinton K, Richardson TG, Wood AR, et al. Genetic predictors of participation in optional components of UK Biobank. bioRxiv. 2020:2020.02.10.941328.
434. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013;42(4):1012-4.
435. Wolke D, Waylen A, Samara M, Steer C, Goodman R, Ford T, et al. Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. Br J Psychiatry. 2009;195(3):249-56.
436. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ, Jr. Selection Bias Due to Loss to Follow Up in Cohort Studies. Epidemiology. 2016;27(1):91-7.
437. Nunan D, Aronson J, Bankhead C. Catalogue of bias: attrition bias. BMJ Evid Based Med. 2018;23(1):21-2.
438. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177(1):26-31.
439. Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. Nat Genet. 2020;52(3):242-3.
440. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
441. Taubman P. Earnings, education, genetics, and environment. J Hum Resour. 1976;11(4):447-61.
442. Clark D, Royer H. The Effect of Education on Adult Mortality and Health: Evidence from Britain. Am Econ Rev. 2013;103(6):2087-120.
443. von Hinke S, Davey Smith G, Lawlor DA, Propper C, Windmeijer F. Genetic markers as instrumental variables. J Health Econ. 2016;45:131-48.
444. Lorant V, Deliege D, Eaton W, Robert A, Philippot P, Ansseau M. Socioeconomic inequalities in depression: a meta-analysis. Am J Epidemiol. 2003;157(2):98-112.
445. Hemingway H, Marmot M. Evidence based cardiology: psychosocial factors in the aetiology and prognosis of coronary heart disease. Systematic review of prospective cohort studies. BMJ. 1999;318(7196):1460-7.
446. Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G. Indicators of socioeconomic position (part 2). J Epidemiol Community Health. 2006;60(2):95-101.
447. Geyer S, Hemstrom O, Peter R, Vagero D. Education, income, and occupational class cannot be used interchangeably in social epidemiology. Empirical evidence against a common practice. J Epidemiol Community Health. 2006;60(9):804-10.

448. Lawlor DA, Ebrahim S, Davey Smith G. Adverse socioeconomic position across the lifecourse increases coronary heart disease risk cumulatively: findings from the British women's heart and health study. J Epidemiol Community Health. 2005;59(9):785-93.
449. Uphoff EP, Pickett KE, Cabieses B, Small N, Wright J. A systematic review of the relationships between social capital and socioeconomic inequalities in health: a contribution to understanding the psychosocial pathway of health inequalities. Int J Equity Health. 2013;12:54.
450. Bell CN, Kerr J, Young JL. Associations between Obesity, Obesogenic Environments, and Structural Racism Vary by County-Level Racial Composition. Int J Environ Res Public Health. 2019;16(5).
451. Scobie S, Morris J. Quality and inequality: digging deeper2020 26/07/2020. Available from: https://www.nuffieldtrust.org.uk/public/files/2020-01/quality_inequality/v2/.
452. England PH. Disparities in the risk and outcomes of COVID-19. 2020.
453. Cheater S. Health inequalities – Covid-19 will widen the gap. International Journal of Health Promotion and Education. 2020;58(4):223-5.
454. Bambra C, Riordan R, Ford J, Matthews F. The COVID-19 pandemic and health inequalities. J Epidemiol Community Health. 2020.
455. Bibby J, Everest G, Abbs I. Will COVID-19 be a watershed moment for health inequalities? https://www.health.org.uk/sites/default/files/2020-05/Will%20COVID-19%20be%20a%20watershed%20moment%20for%20health%20inequalities.pdf; 2020.
456. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278-95.
457. Perkins NJ, Cole SR, Harel O, Tchetgen Tchetgen EJ, Sun B, Mitchell EM, et al. Principled Approaches to Missing Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):568-75.
458. VanderWeele TJ, Shrier I. Sufficient Cause Representation of the Four-way Decomposition for Mediation and Interaction. Epidemiology. 2016;27(5):e32-3.
459. Herttua K, Martikainen P, Batty GD, Kivimaki M. Poor Adherence to Statin and Antihypertensive Therapies as Risk Factors for Fatal Stroke. J Am Coll Cardiol. 2016;67(13):1507-15.
460. Mahtta D, Ramsey DJ, Al Rifai M, Nasir K, Samad Z, Aguilar D, et al. Evaluation of Aspirin and Statin Therapy Use and Adherence in Patients With Premature Atherosclerotic Cardiovascular Disease. JAMA Netw Open. 2020;3(8):e2011051.
461. Barcellos SH, Carvalho LS, Turley P. Education can reduce health differences related to genetic risk of obesity. Proc Natl Acad Sci U S A. 2018;115(42):E9765-E72.
462. Gibson M, Petticrew M, Bambra C, Sowden AJ, Wright KE, Whitehead M. Housing and health inequalities: a synthesis of systematic reviews of interventions aimed at different pathways linking housing and health. Health Place. 2011;17(1):175-84.
463. Marmot M. The influence of income on health: views of an epidemiologist. Health Aff (Millwood). 2002;21(2):31-46.
464. University College London DoEaPH, National Centre for Social Research (NatCen). Health Survey for England, 2017.2nd Edition.2019.
465. Peeters A, Backholer K. How to influence the obesity landscape using health policies. Int J Obes (Lond). 2017;41(6):835-9.
466. Backholer K, Sarink D, Beauchamp A, Keating C, Loh V, Ball K, et al. The impact of a tax on sugar-sweetened beverages according to socio-economic position: a systematic review of the evidence. Public Health Nutr. 2016;19(17):3070-84.
467. Sharbaugh MS, Althouse AD, Thoma FW, Lee JS, Figueredo VM, Mulukutla SR. Impact of cigarette taxes on smoking prevalence from 2001-2015: A report using the Behavioral and Risk Factor Surveillance Survey (BRFSS). PLoS One. 2018;13(9):e0204416.

468. Wilkinson AL, Scollo MM, Wakefield MA, Spittal MJ, Chaloupka FJ, Durkin SJ. Smoking prevalence following tobacco tax increases in Australia between 2001 and 2017: an interrupted time-series analysis. Lancet Public Health. 2019;4(12):e618-e27.
469. Sims M, Maxwell R, Bauld L, Gilmore A. Short term impact of smoke-free legislation in England: retrospective analysis of hospital admissions for myocardial infarction. BMJ. 2010;340:c2161.
470. Clark CE, Omboni S, McDonagh ST, McManus RJ, Sheppard JP. Effective detection and management of hypertension through community pharmacy in England. Evaluation. 2020;15:13.
471. Simvastatin over the counter. Drug and Therapeutics Bulletin. 2005;43(4):25.
472. England N. NHS to review making statins available direct from pharmacists as part of Long Term Plan to cut heart disease 2019 [updated 04/09/2019. Available from: https://www.england.nhs.uk/2019/09/nhs-to-review-making-statins-available-direct-from-pharmacists-as-part-of-long-term-plan-to-cut-heart-disease/.
473. Allen J, Balfour R, Bell R, Marmot M. Social determinants of mental health. Int Rev Psychiatry. 2014;26(4):392-407.
474. McDaniel JT, Nuhu K, Ruiz J, Alorbi G. Social determinants of cancer incidence and mortality around the world: an ecological study. Glob Health Promot. 2019;26(1):41-9.
475. Howe LD, Lawlor DA, Propper C. Trajectories of socioeconomic inequalities in health, behaviours and academic achievement across childhood and adolescence. J Epidemiol Community Health. 2013;67(4):358-64.
476. Wills AK, Lawlor DA, Matthews FE, Sayer AA, Bakra E, Ben-Shlomo Y, et al. Life course trajectories of systolic blood pressure using longitudinal data from eight UK cohorts. PLoS Med. 2011;8(6):e1000440.
477. Duncan MS, Freiberg MS, Greevy RA, Jr., Kundu S, Vasan RS, Tindle HA. Association of Smoking Cessation With Subsequent Risk of Cardiovascular Disease. JAMA. 2019;322(7):642-50.

Appendix 1: Mendelian randomisation for mediation analysis: current methods and challenges for implementation

Author affiliations

1. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
2. Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
3. Centre for Academic Mental Health, University of Bristol, Bristol, UK
4. National Institute for Health Research Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol, Bristol, UK
5. K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Norway.

Multiple mediator equations

i. The difference method to estimate the direct effect and indirect effect using phenotypic observed data mutually adjusting for all mediators

Total:

$$Y = \theta^{+}_{0} + \boldsymbol{\theta^{+}_{1}X} + \theta^{+}_{5}C$$

Direct:

$$Y = \theta_0 + \boldsymbol{\theta_1 X} + \theta_2 M1 + \theta_3 M2 + \theta_4 M3 + \theta_5 C$$

Indirect:

$$\boldsymbol{\theta^{+}_{1}X - \theta_1 X}$$

ii. The product of coefficients method to estimate the indirect effect using phenotypic observed data, considering each mediator individually

Mediator 1:

Exposure-Mediator:

$$M1 = \beta_0 + \beta_1 X + \beta_4 C$$

Direct:

$$Y = \theta_0 + \boldsymbol{\theta_1 X} + \theta_{2M_1} M1 + \theta_4 C$$

Indirect:

$$\boldsymbol{\beta_1 \theta_{2M_1}}$$

Mediator 2:

Exposure-Mediator:

$$M2 = \beta_0 + \boldsymbol{\beta_2 X} + \beta_4 C$$

Direct:

$$Y = \theta_0 + \boldsymbol{\theta_1 X} + \theta_{2M_2} M2 + \theta_4 C$$

Indirect:

$$\boldsymbol{\beta_2 \theta_{2M_2}}$$

Mediator 3:

Exposure-Mediator:

$$M3 = \beta_0 + \beta_3 X + \beta_4 C$$

Direct:

$$Y = \theta_0 + \boldsymbol{\theta_1 X} + \theta_{2M_3} M3 + \theta_4 C$$

Indirect:

$$\boldsymbol{\beta_3\theta_{2M3}}$$

Combined indirect:

$$\boldsymbol{\beta_1\theta_{2M1} + \beta_2\theta_{2M2} + \beta_3\theta_{2M3}}$$

iii. Multivariable MR to estimate the direct effect and indirect effect using a single genetic instrumental variable for each of the exposure and mediator, using two-stage least squares regression

Total:

$$X = \pi_0 + \pi_1 G_x + v_1$$
$$Y = \beta_0 + \boldsymbol{\beta_{XT}X} + \mu_1$$

Direct:

$$X = \pi_0 + \pi_{1x}G_x + \pi_{2x}G_{M1} + \pi_{3x}G_{M2} + \pi_{4x}G_{M3} + v_1$$
$$M1 = \pi_1 + \pi_{1z}G_x + \pi_{2z}G_{M1} + \pi_{3z}G_{M2} + \pi_{4z}G_{M3} + v_2$$
$$M2 = \pi_2 + \pi_{1\Omega}G_x + \pi_{2\Omega}G_{M1} + \pi_{3\Omega}G_{M2} + \pi_{4\Omega}G_{M3} + v_3$$
$$M3 = \pi_3 + \pi_{3\alpha}G_x + \pi_{2\alpha}G_{M1} + \pi_{2\alpha}G_{M2} + \pi_{4\alpha}G_{M3} + v_4$$
$$Y = \beta_0 + \boldsymbol{\beta_X X} + \beta_{M1}M1 + \beta_{M2}M2 + \beta_{M3}M3 + \mu_2$$

Indirect:

$$\boldsymbol{\beta_{XT} - \beta_X}$$

iv. Two-step MR to estimate the indirect effect using genetic instrumental variables for both the exposure and mediator, using two-stage least squares regression

Mediator 1:

Exposure-Mediator:

$$X = \pi_0 + \pi_1 G_x + v_X$$
$$M1 = \beta_0 + \boldsymbol{\beta_{XM1}X} + \mu_1$$

Direct:

$$X = \pi_0 + \pi_{1x}G_x + \pi_{2x}G_{M1} + v_{X1}$$
$$M1 = \pi_{01} + \pi_{11}G_x + \pi_{21}G_{M1} + v_{M1}$$
$$Y = \beta_0 + \boldsymbol{\beta_{X1}X} + \beta_{M1}M1 + \mu_2$$

Indirect:

$$\boldsymbol{\beta_{XM1}\beta_{M1}}$$

Mediator 2:

Exposure-Mediator:

$$X = \pi_0 + \pi_1 G_x + v_X$$
$$M2 = \beta_0 + \boldsymbol{\beta_{xM2}X} + \mu_3$$

Direct:

$$X = \pi_{02} + \pi_{12}G_x + \pi_{22}G_{M2} + v_{X2}$$
$$M2 = \pi_{0M2} + \pi_{1M2}G_x + \pi_{2M2}G_{M2} + v_{M2}$$
$$Y = \beta_1 + \boldsymbol{\beta_{X2}X} + \beta_{M2}M2 + \mu_4$$

Indirect:

$$\boldsymbol{\beta_{XM2}\beta_{M2}}$$

Mediator 3:

Exposure-Mediator:

$$X = \pi_0 + \pi_1 G_x + v_X$$

$$M3 = \beta_0 + \boldsymbol{\beta_{XM3}X} + \mu_3$$

Direct:

$$X = \pi_{03} + \pi_{13}\mathbf{G}_x + \pi_{23}G_{M3} + v_{X3}$$
$$M3 = \pi_{0M3} + \pi_{1M3}G_x + \pi_{2M3}\mathbf{G_{M2}} + v_{M3}$$
$$Y = \beta_2 + \boldsymbol{\beta_{X3}X} + \beta_{M3}M3 + \mu_6$$

Indirect:

$$\boldsymbol{\beta_{XM3}\beta_{M3}}$$

Combined indirect:

$$\boldsymbol{\beta_{XM1}\beta_{M1} + \beta_{XM2}\beta_{M2} + \beta_{XM3}\beta_{M3}}$$

Appendix 1 Table 1: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and continuous outcome (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Difference	0	0.5	1.1 (0.009)	0.60	1.20	0.833 (0.007)	0.33	0.67	0.267 (0.007)	0.27	NA	0.243 (0.006)	0.24	NA
Product									0.267 (0.007)	0.27	NA	0.243 (0.006)	0.24	0.12
Difference	-0.5	0.5	1.1 (0.009)	0.60	1.20	1.5 (0.007)	0.75	1.50	-0.4 (0.008)	-1.15	4.60	-0.364 (0.009)	0.14	-0.27
Product									-0.4 (0.008)	-1.15	4.60	-0.364 (0.009)	0.14	0.07
Difference	0.05	0	0.6 (0.009)	0.60	NA	0.333 (0.007)	0.33	NA	0.267 (0.007)	0.27	NA	0.445 (0.009)	0.39	7.90
Product									0.267 (0.007)	0.27	NA	0.445 (0.009)	0.39	0.20
Difference		0.2	0.8 (0.009)	0.80	4.00	0.507 (0.007)	0.32	1.58	0.293 (0.007)	0.28	28.33	0.367 (0.007)	0.32	6.33
Product									0.293 (0.007)	0.28	28.33	0.367 (0.007)	0.32	0.16
Difference		0.5	1.1 (0.009)	0.90	1.80	0.767 (0.007)	0.29	0.58	0.333 (0.008)	0.31	12.32	0.303 (0.006)	0.25	5.05
Product									0.333 (0.008)	0.31	12.32	0.303 (0.006)	0.25	0.13
Difference		1	1.6 (0.009)	1.10	1.10	1.2 (0.007)	0.25	0.25	0.4 (0.008)	0.35	7.00	0.25 (0.004)	0.20	4.00
Product									0.4 (0.008)	0.35	7.00	0.25 (0.004)	0.20	0.10
Difference	0.25	0	0.6 (0.008)	0.60	NA	0.334 (0.006)	0.33	NA	0.267 (0.007)	0.27	NA	0.444 (0.009)	0.19	0.78
Product									0.267 (0.007)	0.27	NA	0.444 (0.009)	0.19	0.10
Difference		0.2	0.8 (0.009)	0.80	4.00	0.4 (0.007)	0.25	1.25	0.399 (0.008)	0.35	6.99	0.499 (0.008)	0.25	1.00
Product									0.399 (0.008)	0.35	6.99	0.499 (0.008)	0.25	0.12
Difference		0.5	1.1 (0.009)	0.90	1.80	0.5 (0.01)	0.12	0.25	0.6 (0.01)	0.48	3.80	0.546 (0.008)	0.30	1.18
Product									0.6 (0.01)	0.48	3.80	0.546 (0.008)	0.30	0.15
Difference		1	1.6 (0.009)	1.10	1.10	0.667 (0.013)	-0.08	-0.08	0.933 (0.014)	0.68	2.73	0.583 (0.008)	0.33	1.33
Product									0.933 (0.014)	0.68	2.73	0.583 (0.008)	0.33	0.17
Difference	0.75	0	0.6 (0.009)	0.60	NA	0.333 (0.007)	0.33	NA	0.267 (0.007)	0.27	NA	0.445 (0.009)	-0.31	-0.41
Product									0.267 (0.007)	0.27	NA	0.445 (0.009)	-0.31	-0.15
Difference		0.2	0.8 (0.009)	0.80	4.00	0.134 (0.01)	0.08	0.42	0.666 (0.011)	0.52	3.44	0.833 (0.012)	0.08	0.11
Product									0.666 (0.011)	0.52	3.44	0.833 (0.012)	0.08	0.04
Difference		0.5	1.1 (0.008)	0.90	1.80	-0.166 (0.017)	-0.29	-0.58	1.267 (0.017)	0.89	2.38	1.151 (0.016)	0.40	0.54
Product									1.267 (0.017)	0.89	2.38	1.151 (0.016)	0.40	0.20
Difference		1	1.6 (0.009)	1.10	1.10	-0.666 (0.03)	-0.92	-0.92	2.266 (0.03)	1.52	2.02	1.416 (0.019)	0.67	0.89
Product									2.266 (0.03)	1.52	2.02	1.416 (0.019)	0.67	0.33

Difference = difference in coefficients; produce = product of coefficients; SD = standard deviation

Appendix 1 Table 2: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and continuous outcome (per unit increase in exposure), and a rare binary outcome and common binary outcome on the risk difference scale, with no residual covariance reflecting confounding (Simulated N=5000)

Outcome	Mediation method	True total effect	True proportion mediated	Total effect (SD)	Size of bias (absolute)	Direct effect (SD)	Size of bias (absolute)	Indirect effect(SD)	Size of bias (absolute)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Continuous outcome	Difference	0	0.05	0 (0.01)	0.00	0 (0.01)	0.00	0 (0.01)	0.00	0 (0.01)	0.18	3.53
	Product							0 (0.003)	0.00	0.227 (10.342)	0.18	0.09
	Difference		0.25	0 (0.01)	0.00	0 (0.01)	0.00	0 (0.01)	0.00	0 (0.01)	-0.10	-0.39
	Product							0 (0.003)	0.00	0.152 (3.448)	-0.10	-0.05
	Difference		0.75	0 (0.01)	0.00	0 (0.01)	0.00	0 (0.01)	0.00	0 (0.01)	-1.18	-1.57
	Product							0 (0.003)	0.00	-0.427 (24.509)	-1.18	-0.59
Rare binary outcome	Difference	0	0.05	0 (0.002)	0.00	0 (0.002)	0.00	0 (0.002)	0.00	0 (0.002)	-0.39	-7.83
	Product							0 (0)	0.00	-0.342 (13.323)	-0.39	-7.83
	Difference		0.25	0 (0.002)	0.00	0 (0.002)	0.00	0 (0.002)	0.00	0 (0.002)	-0.18	-0.71
	Product							0 (0)	0.00	0.073 (3.212)	-0.18	-0.71
	Difference		0.75	0 (0.002)	0.00	0 (0.002)	0.00	0 (0.002)	0.00	0 (0.002)	-0.48	-0.65
	Product							0 (0)	0.00	0.266 (5.102)	-0.48	-0.65
Common binary outcome	Difference	0	0.05	0 (0.004)	0.00	0 (0.004)	0.00	0 (0.004)	0.00	0 (0.004)	-0.01	-0.11
	Product							0 (0.001)	0.00	0.044 (3.642)	-0.01	-0.11
	Difference		0.25	0 (0.004)	0.00	0 (0.004)	0.00	0 (0.004)	0.00	0 (0.004)	-0.13	-0.50
	Product							0 (0.001)	0.00	0.125 (4.933)	-0.13	-0.50
	Difference		0.75	0 (0.004)	0.00	0 (0.004)	0.00	0 (0.004)	0.00	0 (0.004)	-0.64	-0.85
	Product							0 (0.001)	0.00	0.114 (9.155)	-0.64	-0.85

Note: Relative bias cannot be estimated for the total effect direct effect and indirect effect because there is no true total effect

Appendix 1 Table 3: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and continuous outcome using Mendelian randomisation (Simulated N=5000)

Mediation method	True prop-ortion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0	0.5	0.499 (0.017)	0.00	0.00	0.5 (0.014)	0.00	0.00	0 (0.004)	0.00	NA	0 (0.008)	0.00	NA
TSMR									0 (0.004)	0.00	NA	0.004 (0)	0.00	NA
MVMR	-0.5	0.5	0.499 (0.017)	0.00	0.00	0.75 (0.023)	0.00	0.00	-0.25 (0.018)	0.00	0.00	-0.501 (0.043)	0.00	0.00
TSMR									-0.25 (0.018)	0.00	0.00	-0.501 (0.043)	0.00	0.00
MVMR	0.05	0	0 (0.017)	0.00	NA	0 (0.014)	0.00	NA	0 (0.004)	0.00	NA	0.164 (2.176)	0.11	2.28
TSMR									0 (0.004)	0.00	NA	0.004 (0.164)	0.11	2.28
MVMR		0.2	0.2 (0.017)	0.00	0.00	0.19 (0.014)	0.00	0.00	0.01 (0.004)	0.00	0.01	0.049 (0.017)	0.00	-0.01
TSMR									0.01 (0.004)	0.00	0.01	0.004 (0.049)	0.00	-0.01
MVMR		0.5	0.5 (0.018)	0.00	0.00	0.475 (0.015)	0.00	0.00	0.025 (0.004)	0.00	-0.01	0.05 (0.008)	0.00	-0.01
TSMR									0.025 (0.004)	0.00	-0.01	0.004 (0.05)	0.00	-0.01
MVMR		1	1 (0.017)	0.00	0.00	0.95 (0.014)	0.00	0.00	0.05 (0.005)	0.00	0.00	0.05 (0.005)	0.00	0.00
TSMR									0.05 (0.005)	0.00	0.00	0.005 (0.05)	0.00	0.00
MVMR	0.25	0	0 (0.018)	0.00	NA	0 (0.014)	0.00	NA	0 (0.004)	0.00	NA	0.111 (3.554)	-0.14	-0.56
TSMR									0 (0.004)	0.00	NA	0.004 (0.111)	-0.14	-0.56
MVMR		0.2	0.199 (0.017)	0.00	-0.01	0.149 (0.014)	0.00	0.00	0.049 (0.005)	0.00	-0.01	0.249 (0.022)	0.00	0.00
TSMR									0.049 (0.005)	0.00	-0.01	0.005 (0.249)	0.00	0.00
MVMR		0.5	0.5 (0.017)	0.00	0.00	0.375 (0.017)	0.00	0.00	0.125 (0.01)	0.00	0.00	0.25 (0.019)	0.00	0.00
TSMR									0.125 (0.01)	0.00	0.00	0.01 (0.25)	0.00	0.00
MVMR		1	1.001 (0.017)	0.00	0.00	0.751 (0.023)	0.00	0.00	0.249 (0.018)	0.00	0.00	0.249 (0.018)	0.00	0.00
TSMR									0.249 (0.018)	0.00	0.00	0.018 (0.249)	0.00	0.00
MVMR	0.75	0	0 (0.017)	0.00	NA	0 (0.014)	0.00	NA	0 (0.004)	0.00	NA	-0.046 (4.949)	-0.80	-1.06
TSMR									0 (0.004)	0.00	NA	0.004 (-0.046)	-0.80	-1.06
MVMR		0.2	0.2 (0.018)	0.00	0.00	0.05 (0.018)	0.00	0.00	0.15 (0.011)	0.00	0.00	0.754 (0.074)	0.00	0.01
TSMR									0.15 (0.011)	0.00	0.00	0.011 (0.754)	0.00	0.01
MVMR		0.5	0.501 (0.017)	0.00	0.00	0.126 (0.029)	0.00	0.01	0.374 (0.026)	0.00	0.00	0.748 (0.056)	0.00	0.00
TSMR									0.374 (0.026)	0.00	0.00	0.026 (0.748)	0.00	0.00
MVMR		1	1 (0.017)	0.00	0.00	0.249 (0.057)	0.00	0.00	0.751 (0.055)	0.00	0.00	0.751 (0.056)	0.00	0.00
TSMR									0.751 (0.055)	0.00	0.00	0.055 (0.751)	0.00	0.00

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 4: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and continuous outcome (per unit increase in exposure), and a rare binary outcome and common binary outcome on the risk difference scale using Mendelian randomisation, where no residual covariance is included reflecting confounding (Simulated N=5000)

	Mediation method	True total effect	True proportion mediated	Total effect (SD)	Size of bias (absolute)	Direct effect (SD)	Size of bias (absolute)	Indirect effect (SD)	Size of bias (absolute)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Continuous outcome	MVMR	0	0.05	0 (0.003)	0.00	0.227 (10.342)	0.00	0 (0.015)	0.00	0.811 (24.209)	0.76	15.22
	TSMR							0 (0.004)	0.00	0.811 (24.209)	0.76	15.22
	MVMR		0.25	0 (0.003)	0.00	0.152 (3.448)	0.00	0 (0.015)	0.00	-0.672 (31.009)	-0.92	-3.69
	TSMR							0 (0.004)	0.00	-0.672 (31.009)	-0.92	-3.69
	MVMR		0.75	0 (0.003)	0.00	-0.427 (24.509)	0.00	0 (0.015)	0.00	-0.022 (5.39)	-0.77	-1.03
	TSMR							0 (0.004)	0.00	-0.022 (5.39)	-0.77	-1.03
Rare binary outcome	MVMR	0	0.05	0 (0)	0.00	-0.342 (13.323)	0.00	0 (0.003)	0.00	0 (0.003)	1.29	25.90
	TSMR							0 (0)	0.00	1.345 (43.052)	1.29	25.90
	MVMR		0.25	0 (0)	0.00	0.073 (3.212)	0.00	0 (0.003)	0.00	0 (0.003)	-0.21	-0.85
	TSMR							0 (0)	0.00	0.039 (1.697)	-0.21	-0.85
	MVMR		0.75	0 (0)	0.00	0.266 (5.102)	0.00	0 (0.003)	0.00	0 (0.003)	-0.81	-1.08
	TSMR							0 (0)	0.00	-0.059 (3.035)	-0.81	-1.08
Common binary outcome	MVMR	0	0.05	0 (0.001)	0.00	0.044 (3.642)	0.00	0 (0.006)	0.00	0 (0.006)	0.78	15.57
	TSMR							0 (0.001)	0.00	0.829 (30.515)	0.78	15.57
	MVMR		0.25	0 (0.001)	0.00	0.125 (4.933)	0.00	0 (0.006)	0.00	0 (0.006)	-0.24	-0.96
	TSMR							0 (0.001)	0.00	0.01 (4.135)	-0.24	-0.96
	MVMR		0.75	0 (0.001)	0.00	0.114 (9.155)	0.00	0 (0.006)	0.00	0 (0.006)	-0.19	-0.25
	TSMR							0 (0.001)	0.00	0.565 (59.327)	-0.19	-0.25

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 5: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and rare binary outcome on the risk difference scale (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Difference	0	0.025	0.064 (0.001)	0.04	1.54	0.048 (0.002)	0.02	0.92	0.015 (0.001)	0.02	NA	0.244 (0.019)	0.24	NA
Product									0.015 (0.001)	0.02	NA	0.244 (0.019)	0.24	NA
Difference	-0.5		0.064 (0.001)	0.04	1.54	0.087 (0.002)	0.05	1.31	-0.023 (0.002)	-0.06	4.86	-0.365 (0.029)	0.13	-0.27
Product									-0.023 (0.002)	-0.06	4.86	-0.365 (0.029)	0.13	-0.27
Difference	0.05	0	0.051 (0.002)	0.05	NA	0.028 (0.002)	0.03	NA	0.023 (0.001)	0.02	NA	0.447 (0.029)	0.40	7.93
Product									0.023 (0.001)	0.02	NA	0.447 (0.029)	0.40	7.93
Difference		0.1	0.058 (0.001)	-0.04	-0.42	0.037 (0.002)	-0.06	-0.61	0.021 (0.001)	0.02	3.24	0.367 (0.025)	0.32	6.35
Product									0.021 (0.001)	0.02	3.24	0.367 (0.025)	0.32	6.35
Difference		0.025	0.064 (0.001)	0.04	1.54	0.044 (0.002)	0.02	0.86	0.019 (0.001)	0.02	14.43	0.304 (0.025)	0.25	5.07
Product									0.019 (0.001)	0.02	14.43	0.304 (0.025)	0.25	5.07
Difference		0.05	0.068 (0.001)	0.02	0.36	0.051 (0.002)	0.00	0.07	0.017 (0.002)	0.01	5.83	0.251 (0.026)	0.20	4.02
Product									0.017 (0.002)	0.01	5.83	0.251 (0.026)	0.20	4.02
Difference	0.25	0	0.051 (0.002)	0.05	NA	0.028 (0.002)	0.03	NA	0.023 (0.001)	0.02	NA	0.445 (0.028)	0.19	0.78
Product									0.023 (0.001)	0.02	NA	0.445 (0.028)	0.19	0.78
Difference		0.1	0.058 (0.001)	-0.04	-0.21	0.029 (0.002)	-0.05	-0.61	0.029 (0.002)	0.00	0.16	0.5 (0.034)	0.25	1.00
Product									0.029 (0.002)	0.00	0.16	0.5 (0.034)	0.25	1.00
Difference		0.025	0.064 (0.001)	0.04	1.55	0.029 (0.003)	0.01	0.55	0.035 (0.002)	0.03	4.54	0.545 (0.043)	0.29	1.18
Product									0.035 (0.002)	0.03	4.54	0.545 (0.043)	0.29	1.18
Difference		0.05	0.068 (0.001)	0.02	0.36	0.029 (0.004)	-0.01	-0.24	0.039 (0.004)	0.03	2.16	0.581 (0.061)	0.33	1.32
Product									0.039 (0.004)	0.03	2.16	0.581 (0.061)	0.33	1.32
Difference	0.75	0	0.051 (0.002)	0.05	NA	0.028 (0.002)	0.03	NA	0.023 (0.001)	0.02	NA	0.447 (0.028)	-0.30	-0.40
Product									0.023 (0.001)	0.02	NA	0.447 (0.028)	-0.30	-0.40
Difference		0.1	0.058 (0.001)	-0.04	-0.21	0.01 (0.003)	-0.02	-0.61	0.048 (0.002)	-0.03	-0.36	0.833 (0.054)	0.08	0.11
Product									0.048 (0.002)	-0.03	-0.36	0.833 (0.054)	0.08	0.11
Difference		0.025	0.064 (0.001)	0.04	1.54	-0.01 (0.006)	-0.02	-2.55	0.073 (0.005)	0.05	2.91	1.153 (0.091)	0.40	0.54
Product									0.073 (0.005)	0.05	2.91	1.153 (0.091)	0.40	0.54
Difference		0.05	0.068 (0.001)	0.02	0.36	-0.028 (0.01)	-0.04	-3.26	0.096 (0.009)	0.06	1.57	1.416 (0.143)	0.67	0.89
Product									0.096 (0.009)	0.06	1.57	1.416 (0.143)	0.67	0.89

Difference = difference in coefficients method; produce = product of coefficient method; SD = standard deviation

Appendix 1 Table 6: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and common binary outcome on the risk difference scale (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Difference	0	0.125	0.196 (0.002)	0.07	0.57	0.148 (0.003)	0.02	0.19	0.048 (0.002)	0.05	NA	0.243 (0.011)	0.24	NA
Product									0.048 (0.002)	0.05	NA	0.243 (0.011)	0.24	NA
Difference	-0.5	0.125	0.196 (0.002)	0.07	0.57	0.267 (0.003)	0.08	0.43	-0.071 (0.003)	-0.26	4.14	-0.364 (0.018)	0.14	-0.27
Product									-0.071 (0.003)	-0.26	4.14	-0.364 (0.018)	0.14	-0.27
Difference	0.05	0	0.157 (0.003)	0.16	NA	0.087 (0.004)	0.09	NA	0.07 (0.002)	0.07	NA	0.445 (0.017)	0.40	7.91
Product									0.07 (0.002)	0.07	NA	0.445 (0.017)	0.40	7.91
Difference		0.05	0.178 (0.003)	0.13	2.56	0.113 (0.004)	0.07	1.38	0.065 (0.002)	0.06	25.09	0.366 (0.014)	0.32	6.33
Product									0.065 (0.002)	0.06	25.09	0.366 (0.014)	0.32	6.33
Difference		0.125	0.196 (0.002)	0.07	0.57	0.137 (0.004)	0.02	0.15	0.059 (0.002)	0.05	8.49	0.303 (0.014)	0.25	5.05
Product									0.059 (0.002)	0.05	8.49	0.303 (0.014)	0.25	5.05
Difference		0.25	0.21 (0.002)	-0.04	-0.16	0.157 (0.004)	-0.08	-0.34	0.052 (0.003)	0.04	3.19	0.25 (0.013)	0.20	4.00
Product									0.052 (0.003)	0.04	3.19	0.25 (0.013)	0.20	4.00
Difference	0.25	0	0.157 (0.003)	0.16	NA	0.087 (0.004)	0.09	NA	0.07 (0.002)	0.07	NA	0.444 (0.017)	0.19	0.78
Product									0.07 (0.002)	0.07	NA	0.444 (0.017)	0.19	0.78
Difference		0.05	0.178 (0.003)	0.13	2.56	0.089 (0.004)	0.05	1.37	0.089 (0.003)	0.08	6.12	0.5 (0.018)	0.25	1.00
Product									0.089 (0.003)	0.08	6.12	0.5 (0.018)	0.25	1.00
Difference		0.125	0.196 (0.002)	0.07	0.57	0.089 (0.005)	0.00	-0.05	0.107 (0.004)	0.08	2.42	0.545 (0.023)	0.29	1.18
Product									0.107 (0.004)	0.08	2.42	0.545 (0.023)	0.29	1.18
Difference		0.25	0.21 (0.002)	-0.04	-0.16	0.087 (0.007)	-0.10	-0.53	0.122 (0.006)	0.06	0.96	0.583 (0.03)	0.33	1.33
Product									0.122 (0.006)	0.06	0.96	0.583 (0.03)	0.33	1.33
Difference	0.75	0	0.157 (0.003)	0.16	NA	0.087 (0.004)	0.09	NA	0.07 (0.002)	0.07	NA	0.444 (0.016)	-0.31	-0.41
Product									0.07 (0.002)	0.07	NA	0.444 (0.016)	-0.31	-0.41
Difference		0.05	0.178 (0.003)	0.13	2.56	0.03 (0.006)	0.02	1.36	0.148 (0.004)	0.11	2.96	0.834 (0.03)	0.08	0.11
Product									0.148 (0.004)	0.11	2.96	0.834 (0.03)	0.08	0.11
Difference		0.125	0.196 (0.002)	0.07	0.57	-0.029 (0.009)	-0.06	-1.94	0.225 (0.008)	0.13	1.40	1.149 (0.047)	0.40	0.53
Product									0.225 (0.008)	0.13	1.40	1.149 (0.047)	0.40	0.53
Difference		0.25	0.21 (0.002)	-0.04	-0.16	-0.087 (0.015)	-0.15	-2.39	0.297 (0.014)	0.11	0.58	1.415 (0.073)	0.66	0.89
Product									0.297 (0.014)	0.11	0.58	1.415 (0.073)	0.66	0.89

Difference = difference in coefficients; produce = product of coefficients; SD = standard deviation

Appendix 1 Table 7: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and rare binary outcome on the risk difference scale, where simulated total effects are small (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Difference	0.05	0.0005	0.051 (0.002)	0.05	101.47	0.029 (0.002)	0.03	59.32	0.023 (0.001)	0.02	884.28	0.441 (0.029)	0.39	7.83
Product									0.023 (0.001)	0.02	884.28	0.441 (0.029)	0.39	7.83
Difference		0.0025	0.053 (0.002)	0.05	20.17	0.031 (0.002)	0.03	11.90	0.022 (0.001)	0.02	159.28	0.422 (0.027)	0.37	7.43
Product									0.022 (0.001)	0.02	159.28	0.422 (0.027)	0.37	7.43
Difference		0.005	0.055 (0.002)	0.05	9.93	0.033 (0.002)	0.03	6.80	0.022 (0.001)	0.02	85.70	0.401 (0.028)	0.35	7.02
Product									0.022 (0.001)	0.02	85.70	0.401 (0.028)	0.35	7.02
Difference	0.25	0.0005	0.051 (0.002)	0.05	101.63	0.028 (0.002)	0.03	74.49	0.023 (0.001)	0.02	181.05	0.449 (0.029)	0.20	0.80
Product									0.023 (0.001)	0.02	181.05	0.449 (0.029)	0.20	0.80
Difference		0.0025	0.053 (0.002)	0.05	20.17	0.028 (0.002)	0.03	14.19	0.024 (0.001)	0.02	36.10	0.462 (0.03)	0.21	0.85
Product									0.024 (0.001)	0.02	36.10	0.462 (0.03)	0.21	0.85
Difference		0.005	0.055 (0.001)	0.05	9.94	0.029 (0.002)	0.03	7.54	0.026 (0.001)	0.03	20.56	0.477 (0.031)	0.23	0.91
Product									0.026 (0.001)	0.03	20.56	0.477 (0.031)	0.23	0.91
Difference	0.75	0.0005	0.051 (0.002)	0.05	101.74	0.027 (0.002)	0.03	217.47	0.024 (0.001)	0.02	63.83	0.469 (0.031)	-0.28	-0.37
Product									0.024 (0.001)	0.02	63.83	0.469 (0.031)	-0.28	-0.37
Difference		0.0025	0.053 (0.002)	0.05	20.14	0.023 (0.003)	0.02	35.87	0.03 (0.002)	0.03	15.57	0.565 (0.038)	-0.19	-0.25
Product									0.03 (0.002)	0.03	15.57	0.565 (0.038)	-0.19	-0.25
Difference		0.005	0.055 (0.001)	0.05	9.93	0.018 (0.003)	0.02	14.37	0.037 (0.002)	0.04	9.71	0.67 (0.042)	-0.08	-0.11
Product									0.037 (0.002)	0.04	9.71	0.67 (0.042)	-0.08	-0.11

Difference = difference in coefficients; produce = product of coefficients; SD = standard deviation

Appendix 1 Table 8: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and common binary outcome on the risk difference scale, where true total effects are small (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Difference	0.05	0.0025	0.158 (0.003)	0.16	62.22	0.089 (0.004)	0.09	36.27	0.07 (0.002)	0.07	537.24	0.44 (0.017)	0.39	7.80
Product									0.07 (0.002)	0.07	537.24	0.44 (0.017)	0.39	7.80
Difference		0.0125	0.163 (0.003)	0.15	12.05	0.095 (0.004)	0.08	6.96	0.069 (0.002)	0.06	90.70	0.42 (0.016)	0.37	7.41
Product									0.069 (0.002)	0.06	90.70	0.42 (0.016)	0.37	7.41
Difference		0.025	0.169 (0.003)	0.14	5.74	0.101 (0.004)	0.08	3.26	0.067 (0.002)	0.04	34.97	0.4 (0.015)	0.35	7.01
Product									0.067 (0.002)	0.04	34.97	0.4 (0.015)	0.35	7.01
Difference	0.25	0.0025	0.158 (0.003)	0.16	62.22	0.087 (0.004)	0.09	45.56	0.071 (0.002)	0.07	110.19	0.448 (0.016)	0.20	0.79
Product									0.071 (0.002)	0.07	110.19	0.448 (0.016)	0.20	0.79
Difference		0.0125	0.163 (0.003)	0.15	12.06	0.088 (0.004)	0.08	8.38	0.075 (0.002)	0.07	21.08	0.461 (0.017)	0.21	0.85
Product									0.075 (0.002)	0.07	21.08	0.461 (0.017)	0.21	0.85
Difference		0.025	0.169 (0.003)	0.14	5.75	0.088 (0.004)	0.07	3.72	0.08 (0.002)	0.06	9.84	0.476 (0.017)	0.23	0.90
Product									0.08 (0.002)	0.06	9.84	0.476 (0.017)	0.23	0.90
Difference	0.75	0.0025	0.158 (0.003)	0.16	62.22	0.084 (0.004)	0.08	133.04	0.074 (0.002)	0.07	39.28	0.47 (0.017)	-0.28	-0.37
Product									0.074 (0.002)	0.07	39.28	0.47 (0.017)	-0.28	-0.37
Difference		0.0125	0.163 (0.003)	0.15	12.05	0.071 (0.004)	0.07	21.75	0.092 (0.003)	0.09	9.48	0.564 (0.02)	-0.19	-0.25
Product									0.092 (0.003)	0.09	9.48	0.564 (0.02)	-0.19	-0.25
Difference		0.025	0.169 (0.003)	0.14	5.75	0.056 (0.005)	0.05	8.01	0.112 (0.003)	0.11	5.66	0.666 (0.023)	-0.08	-0.11
Product									0.112 (0.003)	0.11	5.66	0.666 (0.023)	-0.08	-0.11

Difference = difference in coefficients; produce = product of coefficients; SD = standard deviation

Appendix 1 Table 9: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and a rare binary outcome on the risk difference scale using Mendelian randomisation (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0	0.025	0.029 (0.003)	0.00	0.15	0.029 (0.003)	0.00	0.15	0 (0)	0.00	NA	-0.001 (0.008)	0.00	NA
TSMR									0 (0)	0.00	NA	0 (-0.001)	0.00	NA
MVMR	-0.5	0.025	0.029 (0.003)	0.00	0.15	0.043 (0.004)	0.01	0.15	-0.014 (0.003)	0.00	0.16	-0.509 (0.133)	-0.01	0.02
TSMR									-0.014 (0.003)	0.00	0.16	-0.509 (0.133)	-0.01	0.02
MVMR	0.05	0	0 (0.003)	0.00	NA	0 (0.003)	0.00	NA	0 (0)	0.00	NA	-0.038 (2.344)	-0.09	-1.77
TSMR									0 (0)	0.00	NA	0 (-0.038)	-0.09	-1.77
MVMR		0.1	0.014 (0.003)	-0.09	-0.86	0.014 (0.003)	-0.08	-0.86	0.001 (0)	0.00	-0.85	0.051 (0.023)	0.00	0.02
TSMR									0.001 (0)	0.00	-0.85	0 (0.051)	0.00	0.02
MVMR		0.025	0.029 (0.003)	0.00	0.15	0.027 (0.003)	0.00	0.15	0.001 (0)	0.00	0.15	0.05 (0.015)	0.00	0.01
TSMR									0.001 (0)	0.00	0.15	0 (0.05)	0.00	0.01
MVMR		0.05	0.043 (0.003)	-0.01	-0.15	0.04 (0.003)	-0.01	-0.15	0.002 (0.001)	0.00	-0.14	0.051 (0.017)	0.00	0.02
TSMR									0.002 (0.001)	0.00	-0.14	0.001 (0.051)	0.00	0.02
MVMR	0.25	0	0 (0.003)	0.00	NA	0 (0.003)	0.00	NA	0 (0)	0.00	NA	-0.08 (3.313)	-0.33	-1.32
TSMR									0 (0)	0.00	NA	0 (-0.08)	-0.33	-1.32
MVMR		0.1	0.014 (0.003)	-0.09	-0.43	0.011 (0.003)	-0.06	-0.86	0.004 (0.001)	-0.02	-0.86	0.26 (0.078)	0.01	0.04
TSMR									0.004 (0.001)	-0.02	-0.86	0.001 (0.26)	0.01	0.04
MVMR		0.025	0.029 (0.003)	0.00	0.16	0.022 (0.003)	0.00	0.17	0.007 (0.002)	0.00	0.15	0.25 (0.066)	0.00	0.00
TSMR									0.007 (0.002)	0.00	0.15	0.002 (0.25)	0.00	0.00
MVMR		0.05	0.043 (0.003)	-0.01	-0.15	0.032 (0.004)	-0.01	-0.14	0.011 (0.004)	0.00	-0.16	0.248 (0.085)	0.00	-0.01
TSMR									0.011 (0.004)	0.00	-0.16	0.004 (0.248)	0.00	-0.01
MVMR	0.75	0	0 (0.003)	0.00	NA	0 (0.003)	0.00	NA	0 (0)	0.00		0.027 (2.198)	-0.72	-0.96
TSMR									0 (0)	0.00		0 (0.027)	-0.72	-0.96
MVMR		0.1	0.014 (0.003)	-0.09	-0.43	0.004 (0.004)	-0.02	-0.85	0.011 (0.002)	-0.06	-0.86	0.782 (0.237)	0.03	0.04
TSMR									0.011 (0.002)	-0.06	-0.86	0.002 (0.782)	0.03	0.04
MVMR		0.025	0.029 (0.003)	0.00	0.16	0.007 (0.006)	0.00	0.16	0.022 (0.005)	0.00	0.16	0.757 (0.202)	0.01	0.01
TSMR									0.022 (0.005)	0.00	0.16	0.005 (0.757)	0.01	0.01
MVMR		0.05	0.043 (0.003)	-0.01	-0.15	0.011 (0.011)	0.00	-0.15	0.032 (0.01)	-0.01	-0.15	0.753 (0.247)	0.00	0.00
TSMR									0.032 (0.01)	-0.01	-0.15	0.01 (0.753)	0.00	0.00

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 10: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and a common binary outcome on the risk difference scale using Mendelian randomisation (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0	0.125	0.089 (0.005)	-0.04	-0.29	0.089 (0.005)	-0.04	-0.29	0 (0.001)	0.00	NA	-0.001 (0.008)	0.00	NA
TSMR									0 (0.001)	0.00	NA	-0.001 (0.008)	0.00	NA
MVMR	-0.5	0.125	0.089 (0.005)	-0.04	-0.29	0.133 (0.007)	-0.05	-0.29	-0.045 (0.006)	0.02	-0.29	-0.503 (0.079)	0.00	0.01
TSMR									-0.045 (0.006)	0.02	-0.29	-0.503 (0.079)	0.00	0.01
MVMR	0.05	0	0 (0.006)	0.00	NA	0 (0.006)	0.00	NA	0 (0.001)	0.00	NA	-1.207 (38.778)	-1.26	-25.14
TSMR									0 (0.001)	0.00	NA	-1.207 (38.778)	-1.26	-25.14
MVMR		0.05	0.045 (0.006)	-0.01	-0.10	0.043 (0.005)	0.00	-0.10	0.002 (0.001)	0.00	-0.10	0.05 (0.019)	0.00	-0.01
TSMR									0.002 (0.001)	0.00	-0.10	0.05 (0.019)	0.00	-0.01
MVMR		0.125	0.089 (0.005)	-0.04	-0.29	0.085 (0.005)	-0.03	-0.29	0.004 (0.001)	0.00	-0.29	0.05 (0.01)	0.00	-0.01
TSMR									0.004 (0.001)	0.00	-0.29	0.05 (0.01)	0.00	-0.01
MVMR		0.25	0.131 (0.004)	-0.12	-0.48	0.125 (0.004)	-0.11	-0.48	0.007 (0.001)	-0.01	-0.48	0.05 (0.009)	0.00	0.00
TSMR									0.007 (0.001)	-0.01	-0.48	0.05 (0.009)	0.00	0.00
MVMR	0.25	0	0 (0.006)	0.00	NA	0 (0.006)	0.00	NA	0 (0.001)	0.00	NA	1.384 (49.314)	1.13	4.54
TSMR									0 (0.001)	0.00	NA	1.384 (49.314)	1.13	4.54
MVMR		0.05	0.044 (0.005)	-0.01	-0.12	0.033 (0.005)	0.00	-0.12	0.011 (0.002)	0.00	-0.12	0.252 (0.04)	0.00	0.01
TSMR									0.011 (0.002)	0.00	-0.12	0.252 (0.04)	0.00	0.01
MVMR		0.125	0.089 (0.005)	-0.04	-0.29	0.067 (0.006)	-0.03	-0.28	0.022 (0.003)	-0.01	-0.29	0.248 (0.037)	0.00	-0.01
TSMR									0.022 (0.003)	-0.01	-0.29	0.248 (0.037)	0.00	-0.01
MVMR		0.25	0.131 (0.004)	-0.12	-0.48	0.098 (0.007)	-0.09	-0.48	0.033 (0.006)	-0.03	-0.47	0.25 (0.044)	0.00	0.00
TSMR									0.033 (0.006)	-0.03	-0.47	0.25 (0.044)	0.00	0.00
MVMR	0.75	0	0 (0.006)	0.00	NA	0 (0.005)	0.00	NA	0 (0.001)	0.00	NA	-0.065 (9.881)	-0.81	-1.09
TSMR									0 (0.001)	0.00	NA	-0.065 (9.881)	-0.81	-1.09
MVMR		0.05	0.044 (0.006)	-0.01	-0.11	0.011 (0.006)	0.00	-0.12	0.033 (0.004)	0.00	-0.11	0.764 (0.13)	0.01	0.02
TSMR									0.033 (0.004)	0.00	-0.11	0.764 (0.13)	0.01	0.02
MVMR		0.125	0.089 (0.005)	-0.04	-0.29	0.023 (0.01)	-0.01	-0.26	0.066 (0.009)	-0.03	-0.29	0.744 (0.109)	-0.01	-0.01
TSMR									0.066 (0.009)	-0.03	-0.29	0.744 (0.109)	-0.01	-0.01
MVMR		0.25	0.131 (0.004)	-0.12	-0.48	0.033 (0.017)	-0.03	-0.47	0.098 (0.017)	-0.09	-0.48	0.748 (0.131)	0.00	0.00
TSMR									0.098 (0.017)	-0.09	-0.48	0.748 (0.131)	0.00	0.00

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 11: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and rare binary outcome on the risk difference scale using Mendelian randomisation, where simulated total effects are small (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0.05	0.0005	0.001 (0.003)	0.00	0.39	0.001 (0.003)	0.00	0.38	0 (0)	0.00	0.56	-0.021 (5.936)	-0.07	-1.42
TSMR									0 (0)	0.00	0.56	-0.021 (5.936)	-0.07	-1.42
MVMR		0.0025	0.004 (0.003)	0.00	0.64	0.004 (0.003)	0.00	0.64	0 (0)	0.00	0.71	0.05 (0.682)	0.00	-0.01
TSMR									0 (0)	0.00	0.71	0.05 (0.682)	0.00	-0.01
MVMR		0.005	0.008 (0.003)	0.01	1.45	0.007 (0.003)	0.01	1.45	0 (0)	0.00	1.45	0.05 (0.169)	0.00	-0.01
TSMR									0 (0)	0.00	1.45	0.05 (0.169)	0.00	-0.01
MVMR	0.25	0.0005	0.001 (0.003)	0.00	0.70	0.001 (0.003)	0.00	0.71	0 (0)	0.00	0.65	0.055 (1.544)	-0.20	-0.78
TSMR									0 (0)	0.00	0.65	0.055 (1.544)	-0.20	-0.78
MVMR		0.0025	0.004 (0.003)	0.00	0.61	0.003 (0.003)	0.00	0.61	0.001 (0)	0.00	0.62	1.169 (30.416)	0.92	3.68
TSMR									0.001 (0)	0.00	0.62	1.169 (30.416)	0.92	3.68
MVMR		0.005	0.008 (0.003)	0.01	1.44	0.006 (0.003)	0.01	1.43	0.002 (0)	0.00	1.45	0.275 (0.864)	0.02	0.10
TSMR									0.002 (0)	0.00	1.45	0.275 (0.864)	0.02	0.10
MVMR	0.75	0.0005	0.001 (0.003)	0.00	0.96	0 (0.003)	0.00	1.86	0.001 (0)	0.00	0.66	-0.061 (7.019)	-0.81	-1.08
TSMR									0.001 (0)	0.00	0.66	-0.061 (7.019)	-0.81	-1.08
MVMR		0.0025	0.004 (0.003)	0.00	0.64	0.001 (0.003)	0.00	0.67	0.003 (0.001)	0.00	0.63	0.492 (18.682)	-0.26	-0.34
TSMR									0.003 (0.001)	0.00	0.63	0.492 (18.682)	-0.26	-0.34
MVMR		0.005	0.008 (0.003)	0.01	1.43	0.002 (0.003)	0.00	1.33	0.006 (0.001)	0.01	1.47	1.047 (2.14)	0.30	0.40
TSMR									0.006 (0.001)	0.01	1.47	1.047 (2.14)	0.30	0.40

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 12: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and common binary outcome on the risk difference scale using Mendelian randomisation, where simulated total effects are small (Simulated N=5000):

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0.05	0.0025	0.002 (0.006)	0.00	-0.02	0.002 (0.006)	0.00	-0.02	0 (0.001)	0.00	-0.05	0.115 (2.813)	0.07	1.31
TSMR									0 (0.001)	0.00	-0.05	0.115 (2.813)	0.07	1.31
MVMR		0.0125	0.013 (0.006)	0.00	0.02	0.012 (0.006)	0.00	0.02	0.001 (0.001)	0.00	0.04	0.047 (0.533)	0.00	-0.06
TSMR									0.001 (0.001)	0.00	0.04	0.047 (0.533)	0.00	-0.06
MVMR		0.025	0.024 (0.006)	0.00	-0.04	0.023 (0.005)	0.00	-0.04	0.001 (0.001)	0.00	-0.05	0.046 (0.044)	0.00	-0.08
TSMR									0.001 (0.001)	0.00	-0.05	0.046 (0.044)	0.00	-0.08
MVMR	0.25	0.0025	0.003 (0.006)	0.00	0.04	0.002 (0.005)	0.00	0.05	0.001 (0.001)	0.00	0.01	0.243 (5.842)	-0.01	-0.03
TSMR									0.001 (0.001)	0.00	0.01	0.243 (5.842)	-0.01	-0.03
MVMR		0.0125	0.013 (0.006)	0.00	0.02	0.01 (0.006)	0.00	0.02	0.003 (0.001)	0.00	0.00	0.287 (1.014)	0.04	0.15
TSMR									0.003 (0.001)	0.00	0.00	0.287 (1.014)	0.04	0.15
MVMR		0.025	0.024 (0.006)	0.00	-0.04	0.018 (0.005)	0.00	-0.04	0.006 (0.001)	0.00	-0.05	0.26 (0.075)	0.01	0.04
TSMR									0.006 (0.001)	0.00	-0.05	0.26 (0.075)	0.01	0.04
MVMR	0.75	0.0025	0.003 (0.006)	0.00	0.01	0.001 (0.006)	0.00	-0.03	0.002 (0.001)	0.00	0.02	1.458 (28.614)	0.71	0.94
TSMR									0.002 (0.001)	0.00	0.02	1.458 (28.614)	0.71	0.94
MVMR		0.0125	0.013 (0.006)	0.00	0.02	0.003 (0.006)	0.00	0.06	0.009 (0.001)	0.00	0.01	0.877 (4.495)	0.13	0.17
TSMR									0.009 (0.001)	0.00	0.01	0.877 (4.495)	0.13	0.17
MVMR		0.025	0.024 (0.006)	0.00	-0.04	0.006 (0.006)	0.00	-0.02	0.018 (0.002)	0.00	-0.04	0.792 (0.231)	0.04	0.06
TSMR									0.018 (0.002)	0.00	-0.04	0.792 (0.231)	0.04	0.06

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 13: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and a rare binary outcome on the log odds ratio scale using Mendelian randomisation (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0	0.5	0.617 (0.063)	0.12	0.23	0.62 (0.062)	0.12	0.24	-0.003 (0.006)	0.00	NA	-0.004 (0.01)	0.00	NA
TSMR									0 (0.005)	0.00	NA	-0.001 (0.008)	0.00	NA
MVMR	0.05	0	-0.003 (0.063)	0.00	NA	-0.003 (0.062)	0.00	NA	0 (0.007)	0.00	NA	-0.049 (2.294)	-0.10	-1.99
TSMR									0 (0.007)	0.00	NA	-0.039 (2.356)	-0.09	-1.78
MVMR		0.2	0.306 (0.061)	0.11	0.53	0.292 (0.06)	0.10	0.54	0.014 (0.007)	0.00	0.39	0.046 (0.024)	0.00	-0.08
TSMR									0.016 (0.007)	0.01	0.56	0.051 (0.023)	0.00	0.03
MVMR		0.5	0.621 (0.065)	0.12	0.24	0.592 (0.064)	0.12	0.25	0.028 (0.009)	0.00	0.14	0.046 (0.014)	0.00	-0.08
TSMR									0.031 (0.01)	0.01	0.25	0.051 (0.016)	0.00	0.01
MVMR		1	0.946 (0.065)	-0.05	-0.05	0.901 (0.066)	-0.05	-0.05	0.045 (0.015)	-0.01	-0.10	0.048 (0.016)	0.00	-0.05
TSMR									0.048 (0.016)	0.00	-0.04	0.051 (0.017)	0.00	0.02
MVMR	0.25	0	0 (0.063)	0.00	NA	0 (0.061)	0.00	NA	0 (0.007)	0.00	NA	-0.048 (3.005)	-0.30	-1.19
TSMR									0 (0.007)	0.00	NA	-0.08 (3.316)	-0.33	-1.32
MVMR		0.2	0.305 (0.062)	0.10	0.52	0.23 (0.062)	0.08	0.53	0.075 (0.016)	0.02	0.50	0.256 (0.078)	0.01	0.02
TSMR									0.077 (0.017)	0.03	0.53	0.261 (0.079)	0.01	0.05
MVMR		0.5	0.625 (0.065)	0.12	0.25	0.472 (0.073)	0.10	0.26	0.153 (0.037)	0.03	0.22	0.247 (0.065)	0.00	-0.01
TSMR									0.156 (0.038)	0.03	0.25	0.251 (0.067)	0.00	0.01
MVMR		1	0.947 (0.067)	-0.05	-0.05	0.715 (0.102)	-0.04	-0.05	0.232 (0.077)	-0.02	-0.07	0.246 (0.083)	0.00	-0.02
TSMR									0.235 (0.078)	-0.02	-0.06	0.249 (0.085)	0.00	0.00
MVMR	0.75	0	-0.003 (0.065)	0.00	NA	-0.003 (0.063)	0.00	NA	0 (0.007)	0.00	NA	0.042 (2.245)	-0.71	-0.94
TSMR									0 (0.007)	0.00	NA	0.028 (2.217)	-0.72	-0.96
MVMR		0.2	0.304 (0.063)	0.10	0.52	0.077 (0.075)	0.03	0.55	0.227 (0.045)	0.08	0.51	0.781 (0.237)	0.03	0.04
TSMR									0.229 (0.046)	0.08	0.52	0.786 (0.239)	0.04	0.05
MVMR		0.5	0.622 (0.061)	0.12	0.24	0.156 (0.13)	0.03	0.25	0.465 (0.115)	0.09	0.24	0.756 (0.202)	0.01	0.01
TSMR									0.468 (0.116)	0.09	0.25	0.761 (0.204)	0.01	0.01
MVMR		1	0.945 (0.064)	-0.05	-0.05	0.234 (0.235)	-0.02	-0.06	0.711 (0.227)	-0.04	-0.05	0.756 (0.248)	0.01	0.01
TSMR									0.714 (0.229)	-0.04	-0.05	0.759 (0.25)	0.01	0.01

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 14: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and a common binary outcome on the log odds ratio scale using Mendelian randomisation (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0	0.5	0.496 (0.032)	0.00	-0.01	0.5 (0.03)	0.00	0.00	-0.004 (0.004)	0.00	NA	-0.008 (0.009)	-0.01	-0.02
TSMR									0 (0.004)	0.00	NA	-0.001 (0.008)	0.00	0.00
MVMR	0.05	0	-0.001 (0.033)	0.00	NA	-0.001 (0.03)	0.00	NA	0 (0.005)	0.00	NA	-1.244 (38.267)	-1.29	NA
TSMR									0 (0.006)	0.00	NA	-1.226 (39.475)	-1.28	NA
MVMR		0.2	0.242 (0.031)	0.04	0.21	0.232 (0.029)	0.04	0.22	0.01 (0.005)	0.00	-0.05	0.039 (0.019)	-0.01	-0.06
TSMR									0.012 (0.005)	0.00	0.23	0.05 (0.019)	0.00	0.00
MVMR		0.5	0.497 (0.03)	0.00	-0.01	0.476 (0.029)	0.00	0.00	0.021 (0.005)	0.00	-0.16	0.042 (0.009)	-0.01	-0.02
TSMR									0.025 (0.006)	0.00	0.00	0.05 (0.01)	0.00	0.00
MVMR		1	0.775 (0.032)	-0.23	-0.23	0.74 (0.032)	-0.21	-0.22	0.035 (0.006)	-0.01	-0.30	0.045 (0.008)	0.00	0.00
TSMR									0.039 (0.007)	-0.01	-0.22	0.05 (0.009)	0.00	0.00
MVMR	0.25	0	0.001 (0.034)	0.00	NA	0.001 (0.031)	0.00	NA	0 (0.006)	0.00	NA	1.345 (48.216)	1.10	NA
TSMR									0 (0.006)	0.00	NA	1.423 (50.673)	1.17	NA
MVMR		0.2	0.238 (0.03)	0.04	0.19	0.181 (0.029)	0.03	0.21	0.058 (0.008)	0.01	0.15	0.244 (0.039)	-0.01	-0.04
TSMR									0.06 (0.008)	0.01	0.20	0.255 (0.041)	0.01	0.04
MVMR		0.5	0.498 (0.031)	0.00	0.00	0.377 (0.034)	0.00	0.01	0.12 (0.017)	0.00	-0.04	0.243 (0.036)	-0.01	-0.02
TSMR									0.124 (0.018)	0.00	-0.01	0.25 (0.038)	0.00	0.00
MVMR		1	0.775 (0.032)	-0.22	-0.22	0.584 (0.044)	-0.17	-0.22	0.191 (0.033)	-0.06	-0.23	0.247 (0.043)	0.00	0.00
TSMR									0.195 (0.034)	-0.06	-0.22	0.252 (0.045)	0.00	0.00
MVMR	0.75	0	0.001 (0.032)	0.00	NA	0.001 (0.03)	0.00	NA	0 (0.005)	0.00	NA	-0.048 (9.389)	-0.80	NA
TSMR									0 (0.005)	0.00	NA	-0.066 (10.008)	-0.82	NA
MVMR		0.2	0.239 (0.031)	0.04	0.19	0.06 (0.035)	0.01	0.19	0.179 (0.021)	0.03	0.19	0.762 (0.131)	0.01	0.24
TSMR									0.182 (0.022)	0.03	0.21	0.773 (0.133)	0.02	0.46
MVMR		0.5	0.497 (0.031)	0.00	-0.01	0.129 (0.057)	0.00	0.03	0.368 (0.05)	-0.01	-0.02	0.743 (0.109)	-0.01	-0.06
TSMR									0.372 (0.051)	0.00	-0.01	0.75 (0.111)	0.00	0.00
MVMR		1	0.774 (0.031)	-0.23	-0.23	0.196 (0.103)	-0.05	-0.22	0.579 (0.099)	-0.17	-0.23	0.748 (0.131)	0.00	-0.01
TSMR									0.582 (0.101)	-0.17	-0.22	0.753 (0.133)	0.00	0.01

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 15: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and a rare binary outcome on the odds ratio scale using Mendelian randomisation (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0	1.65	1.857 (0.116)	0.21	0.13	1.862 (0.115)	0.21	0.13	-0.005 (0.011)	0.00	NA	-0.003 (0.006)	0.00	NA
TSMR									-0.001 (0.025)	0.00	NA	0.025 (-0.001)	0.00	NA
MVMR	0.05	1.00	0.999 (0.063)	0.00	0.00	0.999 (0.062)	0.05	0.05	0 (0.007)	-0.05	-1.00	0 (0.007)	-0.05	-1.00
TSMR									0 (0.029)	-0.05	-1.01	0.029 (-0.001)	-0.05	-1.02
MVMR		1.22	1.361 (0.084)	0.14	0.11	1.342 (0.081)	0.18	0.16	0.019 (0.01)	-0.04	-0.69	0.014 (0.007)	-0.04	-0.72
TSMR									0.069 (0.028)	0.01	0.13	0.028 (0.05)	0.00	0.01
MVMR		1.65	1.864 (0.121)	0.22	0.13	1.812 (0.117)	0.25	0.16	0.052 (0.017)	-0.03	-0.36	0.028 (0.009)	-0.02	-0.44
TSMR									0.161 (0.029)	0.08	0.95	0.029 (0.086)	0.04	0.72
MVMR		2.72	2.58 (0.168)	-0.14	-0.05	2.467 (0.163)	-0.12	-0.04	0.113 (0.037)	-0.02	-0.17	0.044 (0.014)	-0.01	-0.12
TSMR									0.303 (0.031)	0.17	1.23	0.031 (0.118)	0.07	1.36
MVMR	0.25	1.00	1.002 (0.063)	0.00	0.00	1.002 (0.061)	0.25	0.34	0 (0.007)	-0.25	-1.00	0 (0.007)	-0.25	-1.00
TSMR									-0.001 (0.029)	-0.25	-1.00	0.029 (-0.002)	-0.25	-1.01
MVMR		1.22	1.359 (0.084)	0.14	0.11	1.261 (0.078)	0.34	0.38	0.098 (0.022)	-0.21	-0.68	0.072 (0.015)	-0.18	-0.71
TSMR									0.339 (0.034)	0.03	0.11	0.034 (0.25)	0.00	0.00
MVMR		1.65	1.872 (0.122)	0.22	0.14	1.607 (0.117)	0.37	0.30	0.265 (0.063)	-0.15	-0.36	0.141 (0.032)	-0.11	-0.43
TSMR									0.803 (0.055)	0.39	0.95	0.055 (0.431)	0.18	0.72
MVMR		2.72	2.583 (0.173)	-0.14	-0.05	2.055 (0.209)	0.02	0.01	0.528 (0.162)	-0.15	-0.22	0.205 (0.061)	-0.05	-0.18
TSMR									1.512 (0.098)	0.83	1.23	0.098 (0.588)	0.34	1.35
MVMR	0.75	1.00	0.999 (0.065)	0.00	0.00	0.999 (0.063)	0.75	3.00	0 (0.007)	-0.75	-1.00	0 (0.007)	-0.75	-1.00
TSMR									0 (0.028)	-0.75	-1.00	0.028 (0)	-0.75	-1.00
MVMR		1.22	1.359 (0.086)	0.14	0.11	1.083 (0.081)	0.78	2.55	0.275 (0.053)	-0.64	-0.70	0.202 (0.036)	-0.55	-0.73
TSMR									1.019 (0.067)	0.10	0.11	0.067 (0.752)	0.00	0.00
MVMR		1.65	1.865 (0.113)	0.22	0.13	1.179 (0.156)	0.77	1.86	0.686 (0.143)	-0.55	-0.44	0.368 (0.073)	-0.38	-0.51
TSMR									2.412 (0.151)	1.18	0.95	0.151 (1.298)	0.55	0.73
MVMR		2.72	2.578 (0.165)	-0.14	-0.05	1.299 (0.304)	0.62	0.91	1.28 (0.308)	-0.76	-0.37	0.496 (0.115)	-0.25	-0.34
TSMR									4.545 (0.279)	2.51	1.23	0.279 (1.77)	1.02	1.36

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 16: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and a common binary outcome on the odds ratio scale using Mendelian randomisation (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0	1.65	1.643 (0.052)	-0.01	0.00	1.65 (0.05)	0.00	0.00	-0.006 (0.007)	-0.01	NA	-0.004 (0.004)	0.00	NA
TSMR									-0.001 (0.024)	0.00	NA	-0.001 (0.015)	0.00	NA
MVMR	0.05	1.00	1 (0.033)	0.00	0.00	1 (0.03)	0.05	0.05	0 (0.005)	-0.05	-1.00	0 (0.005)	-0.05	-1.00
TSMR									0 (0.027)	-0.05	-1.01	-0.001 (0.027)	-0.05	-1.02
MVMR		1.22	1.274 (0.039)	0.05	0.05	1.262 (0.036)	0.10	0.09	0.012 (0.006)	-0.05	-0.80	0.009 (0.005)	-0.04	-0.81
TSMR									0.064 (0.025)	0.00	0.05	0.05 (0.019)	0.00	0.01
MVMR		1.65	1.645 (0.049)	0.00	0.00	1.611 (0.046)	0.04	0.03	0.034 (0.008)	-0.05	-0.58	0.021 (0.005)	-0.03	-0.58
TSMR									0.152 (0.026)	0.07	0.85	0.093 (0.015)	0.04	0.85
MVMR		2.72	2.171 (0.07)	-0.55	-0.55	2.096 (0.067)	-0.49	-0.19	0.075 (0.013)	-0.06	-0.45	0.034 (0.006)	-0.02	-0.31
TSMR									0.292 (0.024)	0.16	1.15	0.134 (0.011)	0.08	1.69
MVMR	0.25	1.00	1.001 (0.034)	0.00	0.00	1.001 (0.031)	0.25	0.33	0 (0.006)	-0.25	-1.00	0 (0.006)	-0.25	-1.00
TSMR									-0.001 (0.027)	-0.25	-1.00	-0.001 (0.027)	-0.25	-1.01
MVMR		1.22	1.27 (0.038)	0.05	0.05	1.199 (0.035)	0.28	0.31	0.071 (0.01)	-0.23	-0.77	0.056 (0.007)	-0.19	-0.78
TSMR									0.317 (0.027)	0.01	0.04	0.249 (0.019)	0.00	0.00
MVMR		1.65	1.646 (0.051)	0.00	0.00	1.459 (0.05)	0.69	0.18	0.187 (0.025)	-0.07	-0.55	0.113 (0.015)	-0.14	-0.55
TSMR									0.763 (0.032)	0.51	0.85	0.464 (0.02)	0.21	0.85
MVMR		2.72	2.173 (0.069)	-0.55	-0.55	1.795 (0.08)	1.01	-0.12	0.377 (0.06)	0.11	-0.44	0.174 (0.027)	-0.08	-0.31
TSMR									1.463 (0.046)	1.20	1.15	0.674 (0.027)	0.42	1.69
MVMR	0.75	1.00	1.001 (0.032)	0.00	0.00	1.001 (0.03)	0.75	3.01	0 (0.005)	-0.75	-1.00	0 (0.005)	-0.75	-1.00
TSMR									0 (0.026)	-0.75	-1.00	0 (0.026)	-0.75	-1.00
MVMR		1.22	1.27 (0.039)	0.05	0.05	1.062 (0.038)	0.76	2.48	0.208 (0.024)	-0.71	-0.77	0.164 (0.018)	-0.59	-0.78
TSMR									0.956 (0.037)	0.04	0.04	0.753 (0.029)	0.00	0.00
MVMR		1.65	1.645 (0.051)	0.00	0.00	1.14 (0.066)	0.88	1.77	0.505 (0.059)	-0.26	-0.59	0.307 (0.035)	-0.44	-0.59
TSMR									2.288 (0.067)	1.52	0.85	1.392 (0.053)	0.64	0.86
MVMR		2.72	2.170 (0.068)	-0.55	-0.55	1.222 (0.126)	0.96	0.80	0.948 (0.126)	0.16	-0.54	0.437 (0.056)	-0.31	-0.42
TSMR									4.382 (0.12)	3.59	1.15	2.021 (0.08)	1.27	1.69

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 17: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and continuous outcome (per unit increase in exposure), and a rare binary outcome and common binary outcome on the risk difference scale, where measurement error is introduced in either the exposure or mediator (Simulated N=5000)

	Outcome	Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect(SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Measurement error in the exposure	Continuous	Difference	0.25	0.5	0.366 (0.009)	0.17	0.33	0.078 (0.004)	-0.30	-0.59	0.288 (0.008)	0.16	1.30	0.786 (0.011)	0.54	2.14
		Product									0.288 (0.008)	0.16	1.30	0.786 (0.011)	0.54	0.27
	Rare binary	Difference		0.025	0.021 (0.001)	0.00	-0.15	0.005 (0.001)	-0.01	-0.76	0.017 (0.001)	0.01	1.66	0.787 (0.049)	0.54	2.15
		Product									0.017 (0.001)	0.01	1.66	0.787 (0.049)	0.54	2.15
	Common binary	Difference		0.125	0.065 (0.002)	-0.06	-0.48	0.014 (0.002)	-0.08	-0.85	0.051 (0.001)	0.02	0.64	0.785 (0.027)	0.54	2.14
		Product									0.051 (0.001)	0.02	0.64	0.785 (0.027)	0.54	2.14
Measurement error in the mediator	Continuous	Difference	0.25	0.5	1.10 (0.009)	0.60	1.20	0.936 (0.01)	0.56	1.12	0.164 (0.006)	0.04	0.31	0.149 (0.006)	-0.10	-0.40
		Product									0.164 (0.006)	0.04	0.31	0.149 (0.006)	-0.10	-0.05
	Rare binary	Difference		0.025	0.064 (0.001)	0.04	1.54	0.054 (0.002)	0.04	1.89	0.009 (0.001)	0.00	0.51	0.148 (0.021)	-0.10	-0.41
		Product									0.009 (0.001)	0.00	0.51	0.148 (0.021)	-0.10	-0.41
	Common binary	Difference		0.125	0.196 (0.002)	0.07	0.57	0.167 (0.004)	0.07	0.78	0.029 (0.002)	0.00	-0.07	0.148 (0.012)	-0.10	-0.41
		Product									0.029 (0.002)	0.00	-0.07	0.148 (0.012)	-0.10	-0.41

Product = product in coefficients; difference = difference in coefficients; SD = standard deviation

Appendix 1 Table 18: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and continuous outcome (per unit increase in exposure), and a rare binary outcome and common binary outcome on the risk difference scale using Mendelian randomization, where measurement error is introduced in either the exposure or mediator (Simulated N=5000)

	Outcome	Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Measurement error in the exposure	Continuous	MVMR	0.25	0.5	0.499 (0.022)	0.00	0.00	0.375 (0.021)	0.00	0.00	0.124 (0.013)	0.00	-0.01	0.249 (0.024)	0.00	0.00
		TSMR									0.124 (0.013)	0.00	-0.01	0.013 (0.249)	0.00	0.00
	Rare binary	MVMR		0.025	0.029 (0.003)	0.00	0.16	0.022 (0.003)	0.00	0.16	0.007 (0.002)	0.00	0.15	0.252 (0.067)	0.00	0.01
		TSMR									0.007 (0.002)	0.00	0.15	0.002 (0.252)	0.00	0.01
	Common binary	MVMR		0.125	0.089 (0.006)	-0.04	-0.29	0.067 (0.006)	-0.03	-0.29	0.022 (0.003)	-0.01	-0.29	0.248 (0.039)	0.00	-0.01
		TSMR									0.022 (0.003)	-0.01	-0.29	0.248 (0.039)	0.00	-0.01
Measurement error in the mediator	Continuous	MVMR	0.25	0.5	0.499 (0.017)	0.00	0.00	0.374 (0.018)	0.00	0.00	0.125 (0.012)	0.00	0.00	0.251 (0.022)	0.00	0.00
		TSMR									0.125 (0.012)	0.00	0.00	0.012 (0.251)	0.00	0.00
	Rare binary	MVMR		0.025	0.029 (0.003)	0.00	0.15	0.022 (0.003)	0.00	0.15	0.007 (0.002)	0.00	0.16	0.255 (0.068)	0.00	0.02
		TSMR									0.007 (0.002)	0.00	0.16	0.002 (0.255)	0.00	0.02
	Common binary	MVMR		0.125	0.089 (0.005)	-0.04	-0.29	0.067 (0.006)	-0.03	-0.29	0.022 (0.003)	-0.01	-0.29	0.249 (0.04)	0.00	0.00
		TSMR									0.022 (0.003)	-0.01	-0.29	0.249 (0.04)	0.00	0.00

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 19: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect a continuous exposure and continuous outcome (per unit increase in exposure), and a rare binary outcome and common binary outcome on the risk difference scale using Mendelian randomisation, where simulated total effects are imprecise (Simulated N=1000)

Out-come	Mediation method	True total effect	True proportion mediated	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Contin-uous	MVMR	0.2	0.05	0.319 (0.058)	0.00	0.00	0.334 (0.066)	0.00	0.00	0.2 (0.144)	0.00	-0.07	-0.137 (14.419)	-0.19	-3.73
	TSMR									0.009 (0.015)	0.00	-0.07	-0.137 (14.419)	-0.19	-3.73
	MVMR		0.25	0.419 (0.076)	-0.01	-0.03	0.439 (0.088)	-0.01	-0.05	0.194 (0.147)	0.00	0.02	0.228 (3.527)	-0.02	-0.09
	TSMR									0.051 (0.03)	0.00	0.02	0.228 (3.527)	-0.02	-0.09
	MVMR		0.75	0.648 (0.113)	0.00	0.00	0.679 (0.131)	0.00	0.03	0.2 (0.144)	0.00	-0.01	0.849 (9.156)	0.10	0.13
	TSMR									0.148 (0.076)	0.00	-0.01	0.849 (9.156)	0.10	0.13
Rare binary	MVMR	0.01	0.05	0.01 (0.004)	0.00	-0.39	0.352 (0.154)	0.00	-0.39	0.006 (0.009)	0.00	-0.38	0.006 (0.009)	-0.13	-2.56
	TSMR									0 (0.001)	0.00	-0.38	-0.078 (2.288)	-0.13	-2.56
	MVMR		0.25	0.013 (0.005)	0.00	-0.43	0.456 (0.216)	0.00	-0.46	0.006 (0.01)	0.00	-0.36	0.004 (0.01)	-0.14	-0.55
	TSMR									0.002 (0.002)	0.00	-0.36	0.113 (2.935)	-0.14	-0.55
	MVMR		0.75	0.02 (0.008)	0.00	-0.39	0.702 (0.314)	0.00	-0.38	0.006 (0.01)	0.00	-0.39	0.002 (0.011)	-1.48	-1.97
	TSMR									0.005 (0.005)	0.00	-0.39	-0.731 (42.634)	-1.48	-1.97
Common binary	MVMR	0.05	0.05	0.03 (0.008)	-0.03	-0.62	0.334 (0.098)	-0.03	-0.62	0.019 (0.019)	0.00	-0.65	0.018 (0.019)	0.00	-0.05
	TSMR									0.001 (0.002)	0.00	-0.65	0.047 (2.153)	0.00	-0.05
	MVMR		0.25	0.04 (0.01)	-0.03	-0.63	0.44 (0.124)	-0.02	-0.64	0.018 (0.019)	-0.01	-0.62	0.013 (0.019)	-1.68	-6.72
	TSMR									0.005 (0.004)	-0.01	-0.62	-1.431 (48.68)	-1.68	-6.72
	MVMR		0.75	0.062 (0.015)	-0.03	-0.61	0.682 (0.191)	-0.01	-0.57	0.019 (0.019)	-0.02	-0.62	0.005 (0.021)	-0.02	-0.02
	TSMR									0.014 (0.01)	-0.02	-0.62	0.731 (16.786)	-0.02	-0.02

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 20: Estimated effect sizes and size of bias for simulated effect of a continuous mediator explaining the effect between a continuous exposure and continuous outcome using Mendelian randomisation, where true simulated total effects are small (Simulated N=5000)

Mediation method	True proportion mediated	True total effect	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
MVMR	0.05	0.01	0.01 (0.017)	0.00	-0.03	0.009 (0.014)	0.00	-0.02	0.00 (0.004)	0.00	-0.09	-0.448 (15.516)	-0.50	-9.97
TSMR									0.00 (0.004)	0.00	-0.09	-0.448 (15.516)	-0.50	-9.97
MVMR		0.05	0.05 (0.017)	0.00	0.01	0.048 (0.014)	0.00	0.01	0.003 (0.004)	0.00	0.04	0.027 (0.113)	-0.02	-0.46
TSMR									0.003 (0.004)	0.00	0.04	0.027 (0.113)	-0.02	-0.46
MVMR		0.1	0.1 (0.018)	0.00	0.00	0.095 (0.015)	0.00	0.00	0.005 (0.004)	0.00	-0.02	0.044 (0.039)	-0.01	-0.11
TSMR									0.005 (0.004)	0.00	-0.02	0.044 (0.039)	-0.01	-0.11
MVMR	0.25	0.01	0.01 (0.017)	0.00	-0.04	0.007 (0.014)	0.00	-0.05	0.002 (0.004)	0.00	-0.02	0.198 (5.108)	-0.05	-0.21
TSMR									0.002 (0.004)	0.00	-0.02	0.198 (5.108)	-0.05	-0.21
MVMR		0.05	0.05 (0.018)	0.00	-0.01	0.037 (0.014)	0.00	-0.01	0.012 (0.004)	0.00	-0.01	0.257 (0.124)	0.01	0.03
TSMR									0.012 (0.004)	0.00	-0.01	0.257 (0.124)	0.01	0.03
MVMR		0.1	0.099 (0.017)	0.00	-0.01	0.074 (0.014)	0.00	-0.01	0.025 (0.004)	0.00	-0.02	0.251 (0.033)	0.00	0.00
TSMR									0.025 (0.004)	0.00	-0.02	0.251 (0.033)	0.00	0.00
MVMR	0.75	0.01	0.01 (0.017)	0.00	-0.02	0.002 (0.014)	0.00	-0.04	0.007 (0.004)	0.00	-0.01	2.062 (68.799)	1.31	1.75
TSMR									0.007 (0.004)	0.00	-0.01	2.062 (68.799)	1.31	1.75
MVMR		0.05	0.051 (0.017)	0.00	0.02	0.013 (0.014)	0.00	0.06	0.038 (0.005)	0.00	0.00	0.901 (1.63)	0.15	0.20
TSMR									0.038 (0.005)	0.00	0.00	0.901 (1.63)	0.15	0.20
MVMR		0.1	0.1 (0.017)	0.00	0.00	0.025 (0.015)	0.00	0.00	0.075 (0.007)	0.00	0.00	0.767 (0.117)	0.02	0.02
TSMR									0.075 (0.007)	0.00	0.00	0.767 (0.117)	0.02	0.02

Total effect = estimated using univariate Mendelian randomisation; direct effect = estimated using multivariable Mendelian randomisation controlling for both exposure and mediator

MVMR = multivariable Mendelian randomisation; two-step = TSMR Mendelian randomisation; SD = standard deviation

Appendix 1 Table 21: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and continuous outcome (per unit increase in exposure), and a rare binary outcome and common binary outcome on the risk difference scale, where simulated total effects are imprecise (Simulated N=1000)

Outcome	Mediation method	True total effect	True proportion mediated	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Continuous	Difference	0.2	0.05	0.961 (0.076)	0.76	3.80	0.642 (0.095)	0.45	2.26	0.961 (0.076)	0.13	12.86	0.642 (0.095)	0.28	5.67
	Product									0.319 (0.058)	0.13	12.86	0.334 (0.066)	0.28	0.14
	Difference		0.25	0.962 (0.079)	0.76	3.81	0.542 (0.109)	0.39	1.96	0.962 (0.079)	0.27	5.38	0.542 (0.109)	0.19	0.76
	Product									0.419 (0.076)	0.27	5.38	0.439 (0.088)	0.19	0.09
	Difference		0.75	0.961 (0.079)	0.76	3.81	0.313 (0.137)	0.26	1.31	0.961 (0.079)	0.60	3.99	0.313 (0.137)	-0.07	-0.09
	Product									0.648 (0.113)	0.60	3.99	0.679 (0.131)	-0.07	-0.04
Rare binary	Difference	0.01	0.05	0.03 (0.005)	0.02	1.98	0.02 (0.006)	0.01	1.07	0.03 (0.005)	0.00	1.31	0.02 (0.006)	0.30	6.04
	Product									0.01 (0.004)	0.00	1.31	0.352 (0.154)	0.30	6.04
	Difference		0.25	0.03 (0.005)	0.02	1.97	0.017 (0.008)	0.01	1.22	0.03 (0.005)	0.01	2.22	0.017 (0.008)	0.21	0.82
	Product									0.013 (0.005)	0.01	2.22	0.456 (0.216)	0.21	0.82
	Difference		0.75	0.03 (0.005)	0.02	2.00	0.01 (0.01)	0.01	2.88	0.03 (0.005)	0.02	2.37	0.01 (0.01)	-0.05	-0.06
	Product									0.02 (0.008)	0.02	2.37	0.702 (0.314)	-0.05	-0.06
Common binary	Difference	0.05	0.05	0.092 (0.01)	0.04	0.84	0.062 (0.013)	0.01	0.30	0.092 (0.01)	-0.02	-6.86	0.062 (0.013)	0.28	5.69
	Product									0.03 (0.008)	-0.02	-6.86	0.334 (0.098)	0.28	5.69
	Difference		0.25	0.092 (0.009)	0.04	0.84	0.052 (0.014)	0.01	0.39	0.092 (0.009)	0.00	0.20	0.052 (0.014)	0.19	0.76
	Product									0.04 (0.01)	0.00	0.20	0.44 (0.124)	0.19	0.76
	Difference		0.75	0.092 (0.01)	0.04	0.84	0.03 (0.019)	0.02	1.41	0.092 (0.01)	0.05	1.31	0.03 (0.019)	-0.07	-0.09
	Product									0.062 (0.015)	0.05	1.31	0.682 (0.191)	-0.07	-0.09

Difference = difference in coefficients; product = product of coefficients; SD = standard deviation

Appendix 1 Table 22: Estimated effect sizes and size of bias for simulated effect of a phenotypically measured continuous mediator explaining the effect between a continuous exposure and continuous outcome, where true total effects simulated are small (Simulated N=5000)

Mediation method	True total effect	True proportion mediated	Total effect (SD)	Size of bias (absolute)	Size of bias (relative)	Direct effect (SD)	Size of bias (absolute)	Size of bias (relative)	Indirect effect (SD)	Size of bias (absolute)	Size of bias (relative)	Proportion mediated (SD)	Size of bias (absolute)	Size of bias (relative)
Difference	0.05	0.01	0.61 (0.009)	0.60	60.05	0.342 (0.007)	0.33	33.25	0.268 (0.007)	0.26	517.96	0.44 (0.009)	0.39	7.80
Product									0.268 (0.007)	0.26	517.96	0.44 (0.009)	0.39	0.19
Difference		0.05	0.65 (0.009)	0.60	12.00	0.377 (0.007)	0.33	6.58	0.273 (0.007)	0.23	90.32	0.42 (0.009)	0.37	7.41
Product									0.273 (0.007)	0.23	90.32	0.42 (0.009)	0.37	0.19
Difference		0.1	0.7 (0.009)	0.60	6.00	0.42 (0.007)	0.33	3.25	0.28 (0.008)	0.18	36.95	0.4 (0.008)	0.35	6.99
Product									0.28 (0.008)	0.18	36.95	0.4 (0.008)	0.35	0.17
Difference	0.25	0.01	0.61 (0.009)	0.60	59.95	0.336 (0.006)	0.33	32.87	0.273 (0.007)	0.27	106.31	0.448 (0.009)	0.20	0.79
Product									0.273 (0.007)	0.27	106.31	0.448 (0.009)	0.20	0.10
Difference		0.05	0.65 (0.008)	0.60	12.01	0.35 (0.007)	0.31	6.26	0.3 (0.007)	0.26	21.01	0.461 (0.009)	0.21	0.85
Product									0.3 (0.007)	0.26	21.01	0.461 (0.009)	0.21	0.11
Difference		0.1	0.7 (0.009)	0.60	6.00	0.367 (0.007)	0.29	2.92	0.333 (0.008)	0.26	10.31	0.476 (0.008)	0.23	0.90
Product									0.333 (0.008)	0.26	10.31	0.476 (0.008)	0.23	0.11
Difference	0.75	0.01	0.61 (0.009)	0.60	59.99	0.323 (0.007)	0.32	32.07	0.287 (0.008)	0.28	37.90	0.47 (0.009)	-0.28	-0.37
Product									0.287 (0.008)	0.28	37.90	0.47 (0.009)	-0.28	-0.14
Difference		0.05	0.65 (0.009)	0.60	12.00	0.284 (0.007)	0.27	5.42	0.366 (0.008)	0.35	9.44	0.564 (0.01)	-0.19	-0.25
Product									0.366 (0.008)	0.35	9.44	0.564 (0.01)	-0.19	-0.09
Difference		0.1	0.7 (0.009)	0.60	6.00	0.233 (0.008)	0.21	2.08	0.467 (0.009)	0.44	5.89	0.667 (0.01)	-0.08	-0.11
Product									0.467 (0.009)	0.44	5.89	0.667 (0.01)	-0.08	-0.04

Difference = difference in coefficients; product = product of coefficients; SD = standard deviation

Appendix 1 Table 23: Estimated indirect effect and proportion mediated by multiple continuous mediators explaining the association between a continuous exposure and continuous outcome in simulation analyses using phenotypic methods and MR methods (Simulated N = 5000)

		Total Effect (true value = 0.45)	Direct Effect (true value = 0.20)	**Mutually adjusting for all mediators (Difference in coefficients/MVMR)**							**Considering each mediator independently (Product of coefficients/TSMR)**						
				M1		M2		M3		Proportion mediated combined (true value = 0.56)	M1		M2		M3		Proportion mediated combined (true value = 0.56
				Indirect effect	Proportion mediated	Indirect effect	Proportion mediated	Indirect effect	Proportion mediated		Indirect effect	Proportion mediated	Indirect effect	Proportion mediated	Indirect effect	Proportion mediated	
Pheno -typic	Independent mediators	1.55 (0.02)	0.26 (0.01)	0.42 (0.01)	0.28 (0.01)	0.42 (0.01)	0.30 (0.01)	0.45 (0.01)	0.35 (0.01)	0.83 (0.02)	0.93 (0.02)	0.60 (0.01)	1.04 (0.02)	0.67 (0.01)	1.21 (0.02)	0.78 (0.01)	2.05
	Related mediators	1.55 (0.02)	0.26 (0.01)	0.42 (0.01)	0.28 (0.01)	0.25 (0.01)	0.16 (0.01)	0.63 (0.02)	0.48 (0.01)	0.83 (0.02)	0.94 (0.02)	0.60 (0.01)	1.04 (0.02)	0.67 (0.01)	1.37 (0.02)	0.88 (0.01)	2.15
MR	Independent mediators	0.45 (0.03)	0.20 (0.02)	0.05 (0.01)	0.11 (0.02)	0.08 (0.01)	0.18 (0.01)	0.12 (0.01)	0.27 (0.02)	0.55 (0.02)	0.05 (0.01)	0.11 (0.02)	0.12 (0.01)	0.18 (0.01)	0.12 (0.01)	0.27 (0.01)	0.56
	Related mediators	0.45 (0.03)	0.20 (0.02)	0.05 (0.01)	0.11 (0.02)	0.05 (0.01)	0.11 (0.01)	0.15 (0.01)	0.33 (0.02)	0.55 (0.02)	0.05 (0.01)	0.11 (0.02)	0.08 (0.01)	0.18 (0.01)	0.15 (0.02)	0.33 (0.04)	0.62

True indirect effect of independent mediators: M1 = 0.05; M2 = 0.08; M3 = 0.12

True indirect effect of related mediators: M1 = 0.05; M2 = 0.05; M3 = 0.12; M2 via M3; 0.03

MVMR = multivariable Mendelian randomisation; TSMR = two-step Mendelian randomisation; MR = Mendelian randomisation; SD = standard deviation

Appendix 2: Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study.

Author affiliations

[1]Medical Research Council Integrative Epidemiology Unit, University of Bristol, BS8 2BN, United Kingdom.

[2]Population Health Sciences, Bristol Medical School, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom.

[3]Department of Biostatistics and Epidemiology, School of Public Health, Imperial College London, London, UK.

[4]NIHR Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol, Bristol, UK

[5]Institute for Global Health, University College London, UK.

[6]Lausanne University Hospital, Lausanne, Switzerland

[7]Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland

[8]School of Experimental Psychology, University of Bristol, Bristol BS8 1TU, UK

[9]UK Centre for Tobacco and Alcohol Studies, School of Psychological Science, University of Bristol, BS8 1TU, United Kingdom.

[10]Institute for Stroke and Dementia Research, University Hospital of Ludwig-Maximilians University, Munich, Germany.

[11]Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases

[12]University of Texas Health Sciences Center, San Antonio, TX 78229

[13]Boston University School of Medicine

[14]The Framingham Heart Study, Framingham, MA 01702

[15]Department of Neurology and Rehabilitation Medicine, University of Cincinnati College of Medicine, Cincinnati, OH 45267-0525, USA.

[16]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[17]Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom.

[18]Medical Research Council Population Health Research Unit at the University of Oxford, Oxford, UK.

[19]Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford, UK.

[20]National Institute for Health Research, Oxford Biomedical Research Centre, Oxford University Hospital, Oxford, UK.

[21]MRC-PHE Centre for Environment, School of Public Health, Imperial College London, London, UK.

[22]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece

Appendix 2 Table 1: Genome-wide significant SNPs for SBP from split sample GWAS analysis in UK Biobank

Chromosome	RSID	Position	Beta	SE	Other Allele	P Value	Sample
1	rs5068	11905974	1.3283	0.1374	A	4.10E-22	1
1	rs448385	25395133	-0.3464	0.0630	G	3.80E-08	1
1	rs3790604	113046879	-0.8724	0.1204	C	4.30E-13	1
1	rs2765524	89417695	0.3965	0.0641	C	6.00E-10	1
2	rs953246	146335486	-0.3924	0.0685	T	1.00E-08	1
2	rs1344653	19730845	-0.3481	0.0626	A	2.70E-08	1
2	rs1009358	65276452	0.3693	0.0645	T	1.00E-08	1
2	rs268263	164954174	-0.5652	0.0735	T	1.50E-14	1
2	rs35021474	26916844	0.4639	0.0645	C	6.30E-13	1
2	rs2867114	651380	0.6149	0.1073	C	1.00E-08	1
3	rs3821843	53558012	-0.4129	0.0681	G	1.40E-09	1
3	rs1343040	169186293	-0.4317	0.0637	G	1.20E-11	1
3	rs2643826	27562988	-0.3780	0.0631	C	2.20E-09	1
3	rs263016	183502559	0.3475	0.0631	T	3.60E-08	1
4	rs10857147	81181072	-0.8157	0.0690	A	3.00E-32	1
4	rs6825268	26783453	-0.3573	0.0631	A	1.50E-08	1
4	rs13107325	103188709	0.7689	0.1190	C	1.00E-10	1
4	rs1842896	156511459	-0.4014	0.0626	G	1.40E-10	1
5	rs13436194	157803588	0.4570	0.0632	A	4.80E-13	1
5	rs12656497	32831939	-0.7161	0.0638	T	3.00E-29	1
6	rs2607015	31762843	-0.4232	0.0636	G	2.90E-11	1
6	rs2499801	96854594	0.4442	0.0806	G	3.60E-08	1
6	rs13219548	127165290	-0.4051	0.0630	C	1.30E-10	1
7	rs62481856	106412082	-0.8444	0.0789	G	9.30E-27	1
7	rs2854747	45959917	0.4198	0.0637	G	4.40E-11	1
7	rs10241964	19042114	-0.5983	0.1056	G	1.40E-08	1
7	rs10269774	92253972	0.3854	0.0669	G	8.40E-09	1
7	rs891511	150704843	0.3765	0.0681	G	3.20E-08	1
7	rs2023843	27243221	-0.8627	0.1198	C	6.00E-13	1
7	rs3823483	131010943	-0.3546	0.0634	T	2.30E-08	1
8	rs877116	10712945	0.4697	0.0636	G	1.50E-13	1
8	rs7463212	143991858	0.3776	0.0629	T	2.00E-09	1
8	rs73563812	25900405	0.4215	0.0737	G	1.10E-08	1
10	rs56137952	134376691	-0.5494	0.0985	G	2.40E-08	1
10	rs10883543	102552752	-0.6995	0.0996	G	2.20E-12	1
10	rs10995311	64564934	0.3941	0.0634	C	5.00E-10	1
10	rs11191580	104906211	1.0970	0.1180	T	1.50E-20	1
10	rs7076938	115789375	-0.4606	0.0711	C	9.30E-11	1
10	rs12258967	18727959	0.6582	0.0686	C	8.00E-22	1
10	rs7922049	63462365	0.5533	0.0869	G	1.90E-10	1
10	rs10786156	96014622	0.3767	0.0632	C	2.50E-09	1
11	rs55925664	10192809	-0.6415	0.0805	T	1.60E-15	1
11	rs7120737	47702395	0.6316	0.0891	A	1.30E-12	1
11	rs10750766	65473798	-0.3930	0.0691	C	1.30E-08	1
11	rs6331185	100593538	-0.6945	0.0698	G	2.60E-23	1
11	rs12418543	1894163	0.5668	0.0645	A	1.50E-18	1
11	rs747249	130271647	0.3747	0.0656	A	1.10E-08	1
12	rs73437338	90054619	0.7678	0.0843	T	8.80E-20	1
12	rs4766578	111904371	0.4066	0.0626	T	8.60E-11	1
12	rs35444	115552437	0.3909	0.0643	A	1.20E-09	1
12	rs73073676	20351276	0.3773	0.0670	A	1.80E-08	1
15	rs8039305	91422543	-0.5943	0.0630	T	4.00E-21	1
15	rs1717200	41368334	-0.3912	0.0628	A	4.70E-10	1
15	rs1543927	75063573	0.4218	0.0712	T	3.10E-09	1
15	rs11634851	81028965	-0.4216	0.0628	C	1.90E-11	1
16	rs77870048	69965021	-0.9117	0.1399	C	7.20E-11	1
16	rs2188717	24730230	-0.5361	0.0792	T	1.30E-11	1
17	rs9907379	59489893	-0.4257	0.0769	T	3.20E-08	1
17	rs60289499	43218677	-0.4564	0.0708	G	1.10E-10	1

17	rs34710835	45146717	0.5223	0.0643	C	4.40E-16	1
17	rs11650511	1337960	-0.4085	0.0634	C	1.20E-10	1
19	rs73046792	49605705	0.4714	0.0843	G	2.30E-08	1
19	rs12978472	7257990	0.8468	0.0941	C	2.20E-19	1
20	rs74729242	57718690	-0.5901	0.1013	T	5.80E-09	1
20	rs2423514	10693337	0.3709	0.0628	A	3.50E-09	1
1	rs55857306	11895795	0.7687	0.0850	G	1.60E-19	2
1	rs4648815	1687152	0.3807	0.0638	G	2.40E-09	2
1	rs6541328	230833262	-0.5551	0.1010	A	3.80E-08	2
1	rs778121	56620268	-0.3887	0.0659	T	3.60E-09	2
1	rs6657049	115825531	-0.3665	0.0658	G	2.50E-08	2
2	rs268263	164954174	-0.5602	0.0738	T	3.20E-14	2
2	rs4666493	19765225	-0.3641	0.0639	G	1.20E-08	2
2	rs1275988	26914364	0.4835	0.0650	C	9.90E-14	2
2	rs6724607	191466532	0.3656	0.0630	A	6.40E-09	2
3	rs2307032	27432995	0.3879	0.0664	T	5.20E-09	2
3	rs6442260	11590751	0.3595	0.0659	G	4.90E-08	2
4	rs10024506	89764197	0.4173	0.0745	G	2.10E-08	2
4	rs11099097	81167309	-0.6371	0.0696	C	5.40E-20	2
4	rs4690974	156393641	-0.3741	0.0631	T	3.10E-09	2
4	rs17010961	86723103	-0.5775	0.0913	T	2.50E-10	2
5	rs10059884	32832474	-0.5859	0.0643	C	7.80E-20	2
5	rs12652819	121244520	0.3775	0.0677	A	2.50E-08	2
5	rs17715065	158261163	-0.3662	0.0631	C	6.60E-09	2
5	rs11241959	127787964	-0.3715	0.0631	A	4.00E-09	2
5	rs2964330	157743781	-0.3645	0.0641	G	1.30E-08	2
6	rs17080069	150989698	0.7199	0.1218	A	3.40E-09	2
6	rs6923947	127098553	-0.4974	0.0635	G	5.00E-15	2
6	rs7889	31605448	-0.3792	0.0657	C	7.80E-09	2
7	rs891511	150704843	0.4141	0.0683	G	1.30E-09	2
7	rs2392929	106414069	-0.7165	0.0790	T	1.20E-19	2
7	rs42032	92237426	0.3967	0.0720	G	3.60E-08	2
7	rs57301765	19052733	-0.5059	0.0866	G	5.20E-09	2
9	rs2780072	9340831	-0.5035	0.0902	A	2.40E-08	2
10	rs76443711	75449789	-0.5442	0.0915	G	2.70E-09	2
10	rs7070797	63551773	0.6290	0.0905	G	3.60E-12	2
10	rs11187838	96038686	0.5615	0.0637	G	1.20E-18	2
10	rs732998	104897901	0.8365	0.1184	T	1.60E-12	2
10	rs12258967	18727959	0.5893	0.0691	C	1.50E-17	2
11	rs4980379	1888614	-0.5896	0.0657	C	2.70E-19	2
11	rs12807950	107057190	-0.3790	0.0631	T	1.90E-09	2
11	rs7107356	47676170	-0.4793	0.0630	A	2.90E-14	2
11	rs1216743	100573120	-0.5662	0.0705	G	9.60E-16	2
12	rs2681492	90013089	0.6765	0.0840	T	8.00E-16	2
12	rs4767328	115929396	-0.3575	0.0640	G	2.30E-08	2
12	rs35427	115556307	0.4110	0.0664	T	6.00E-10	2
12	rs4883481	50574311	0.4218	0.0652	T	1.00E-10	2
12	rs597808	111973358	0.4310	0.0634	A	1.00E-11	2
15	rs7176022	75107880	0.4365	0.0713	A	9.10E-10	2
15	rs4932373	91429287	-0.5144	0.0672	A	2.00E-14	2
15	rs117539635	69682916	1.3387	0.2009	A	2.70E-11	2
16	rs77870048	69965021	-0.9216	0.1415	C	7.40E-11	2
16	rs11646987	24832408	0.3922	0.0709	G	3.20E-08	2
17	rs1436138	75316880	0.3615	0.0661	A	4.40E-08	2
17	rs7217916	76769434	0.3944	0.0650	A	1.30E-09	2
17	rs2301597	43173273	0.5081	0.0640	T	2.00E-15	2
17	rs11874	45017193	-0.5820	0.0917	G	2.20E-10	2
17	rs4480845	1958609	0.4024	0.0662	T	1.20E-09	2
19	rs167479	11526765	0.5743	0.0630	G	8.20E-20	2
20	rs75777337	57702450	-0.6125	0.1033	T	3.10E-09	2
20	rs913220	10966476	-0.4389	0.0649	C	1.40E-11	2

Appendix 2 Table 2: Genome-wide significant SNPs for lifetime smoking from split sample GWAS analysis in UK Biobank

Chromosome	RSID	Position	Beta	SE	Other Allele	P Value	Sample
1	rs71673396	107507403	.0159243	.0029062	T	4.30e-08	1
1	rs499257	44078384	.0146849	.0024706	T	2.80e-09	1
2	rs2890772	146175106	-.0140201	.0023608	G	2.90e-09	1
3	rs4856463	83638568	.0156642	.0028286	C	3.10e-08	1
3	rs326341	107811142	.013281	.0023333	G	1.30e-08	1
4	rs6852351	28064697	.0132797	.00241	C	3.60e-08	1
5	rs17159727	106632458	.0241372	.0040898	T	3.60e-09	1
5	rs986391	166993972	.0151657	.0024028	G	2.80e-10	1
6	rs16879271	16822974	-.0325007	.0059395	A	4.50e-08	1
7	rs10226228	32315613	-.0141999	.0024083	A	3.70e-09	1
7	rs10233018	117523709	-.0129883	.0023226	A	2.20e-08	1
8	rs10093628	9393379	-.0165694	.0027221	T	1.20e-09	1
9	rs113382419	136463019	-.0242553	.0036827	C	4.50e-11	1
11	rs10750016	112837740	-.0160683	.0023918	T	1.80e-11	1
11	rs11030088	27646247	-.0157839	.002668	G	3.30e-09	1
11	rs6590701	133315869	-.0144894	.0026473	G	4.40e-08	1
12	rs4763463	10355901	.0132272	.0023863	G	3.00e-08	1
15	rs7173514	78849918	.0224245	.0027791	C	7.10e-16	1
1	rs10922907	91193049	.0134411	.0023251	A	7.40e-09	2
2	rs1863161	60139524	-.0127982	.0023263	G	3.80e-08	2
2	rs16824949	146168208	-.0145352	.0023181	G	3.60e-10	2
2	rs7559547	615627	-.0219576	.0030462	C	5.70e-13	2
2	rs263771	185921692	-.0151079	.0027443	C	3.70e-08	2
3	rs62261249	49594060	-.0158442	.0026313	T	1.70e-09	2
7	rs17657924	96625589	.0132644	.0023312	C	1.30e-08	2
9	rs12553882	128195044	-.014292	.0024004	G	2.60e-09	2
9	rs56116178	136460224	-.0306714	.0038063	A	7.70e-16	2
11	rs7948789	112839532	-.0164578	.0023768	A	4.40e-12	2
14	rs12897150	104319530	-.0136766	.0023459	A	5.50e-09	2
15	rs28669908	78910267	.0237801	.0028216	C	3.50e-17	2
15	rs34794623	47680801	-.0195197	.0028221	C	4.60e-12	2
20	rs45577732	61983934	-.0386369	.0042941	C	2.30e-19	2
20	rs159058	31108108	-.0142308	.0025321	A	1.90e-08	2

Appendix 3: Educational inequalities in statin treatment for preventing cardiovascular disease: cross-sectional analysis of UK Biobank

Author Affiliations

1) MRC Integrative Epidemiology Unit, University of Bristol Population Health Sciences, Bristol Medical School, University of Bristol
2) Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
3) Centre for Pharmacology & Therapeutics, Department of Medicine, Hammersmith Campus, Imperial College London, London, UK.
4) Novo Nordisk Research Centre Oxford, Old Road Campus, Oxford, UK
5) Clinical Pharmacology and Therapeutics Section, Institute of Medical and Biomedical Education and Institute for Infection and Immunity, St George's, University of London, London, UK
6) Clinical Pharmacology Group, Pharmacy and Medicines Directorate, St George's University Hospitals NHS Foundation Trust, London, UK
7) Centre for Academic Primary Care, University of Bristol
8) NIHR Bristol Biomedical Research Centre, University of Bristol
9) K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Norway.

Appendix 3 Table 1: ICD codes used to define incident and prevalent cases of cardiovascular disease

Cardiovascular event	ICD9	ICD10
Incident cardiovascular disease (all subtypes combined)	3900-4599	I* G45
Myocardial infarction	4100-4109, 4120-4129	I21, I22
Angina	4139	I20
Stroke	43- 4389	I6, G45
Transient ischaemic attack	4359	G45
Peripheral arterial disease	4439	I73.9
Type 1 diabetes	2500- 25011, 25013, 2504-25041, 25043, 2505-25051, 25053, 2506-25061, 25063, 2507-25071, 25073, 2509-25091, 25093	E10
Chronic kidney disease	5383, 5384, 5385	N183, N184, N185
Familial hypercholesterolaemia	2720	I78.0

Appendix 3 Table 2: Treatment codes in UK Biobank to define medications

Medication	UK Biobank treatment code
Statins	1141146234 1140888594 1140888648 1141192410 1140861958
Erectile dysfunction	1140869100 1140883010 1141168936 1141168944 1141168946 1141168948 1141187810 1141187814 1141187818 1141192248 1141192256 1141192258 1141192260
Antihypertensives	1140860332 1140860334 1140860336 1140860338 1140860340 1140860342 1140860348 1140860352 1140860356 1140860358 1140860362 1140860380 1140860382 1140860386 1140860390 1140860394 1140860396 1140860398 1140860402 1140860404 1140860406 1140860410 1140860418 1140860422 1140860426 1140860434 1140860454 1140860470 1140860478 1140860492 1140860498 1140860520 1140860532 1140860534 1140860544 1140860552 1140860558 1140860562 1140860564 1140860580 1140860590 1140860610 1140860628 1140860632 1140860638 1140860654 1140860658 1140860690 1140860696 1140860706 1140860714 1140860728 1140860736 1140860738 1140860750 1140860752 1140860758 1140860764 1140860776 1140860784 1140860790 1140860802 1140860806 1140860828 1140860830 1140860834 1140860836 1140860838 1140860840 1140860842 1140860846 1140860848 1140860862 1140860878 1140860882 1140860892 1140860904 1140860912 1140860918 1140860938 1140860942 1140860952 1140860954 1140860966 1140860972 1140860976 1140860982 1140860988 1140860994 1140861000 1140861002 1140861008 1140861010 1140861016 1140861022 1140861024 1140861034 1140861046 1140861068 1140861070 1140861088 1140861090 1140861106 1140861110 1140861114 1140861120 1140861128 1140861130 1140861136 1140861138 1140861166 1140861176 1140861190 1140861194 1140861202 1140861266 1140861268 1140861276 1140861282 1140861326 1140861384 1140864950 1140864952 1140866072 1140866074 1140866078 1140866084 1140866086 1140866090 1140866092 1140866094 1140866096 1140866102 1140866104 1140866108 1140866110 1140866116 1140866122 1140866128 1140866132 1140866136 1140866138 1140866140 1140866144 1140866146 1140866156 1140866158 1140866162 1140866164 1140866168 1140866182 1140866192 1140866194 1140866200 1140866202 1140866206 1140866210 1140866212 1140866220 1140866222 1140866226 1140866230 1140866232 1140866236 1140866244 1140866248 1140866262 1140866280 1140866282 1140866306 1140866308 1140866312 1140866318 1140866324 1140866328 1140866330 1140866332 1140866334 1140866340 1140866352 1140866354 1140866356 1140866360 1140866388 1140866390 1140866396 1140866400 1140866402 1140866404 1140866406 1140866408 1140866410 1140866412 1140866416 1140866418 1140866420 1140866422 1140866426 1140866438 1140866440 1140866442 1140866444 1140866446 1140866448 1140866450 1140866460 1140866466 1140866484 1140866506 1140866546 1140866554 1140866692 1140866704 1140866712 1140866724 1140866726 1140866738 1140866756 1140866758 1140866764 1140866766 1140866778 1140866782 1140866784 1140866798 1140866800 1140866802 1140866804 1140875808 1140879758 1140879760 1140879762 1140879778 1140879782 1140879786 1140879794 1140879798 1140879802 1140879806 1140879810 1140879818 1140879822 1140879824 1140879826 1140879830 1140879834 1140879842 1140879854 1140879866 1140888510 1140888512 1140888552 1140888556 1140888560 1140888578 1140888582 1140888586 1140888646 1140888686 1140888760 1140888762 1140909368 1140911698 1140916356 1140916362 1140917428 1140923572 1140923712 1140923718 1140926778 1140926780 1141145658 1141145660 1141145668 1141151016 1141151018 1141151382 1141152600 1141152998 1141153006 1141153026 1141153032 1141153328 1141156754 1141156808 1141156836 1141156846 1141157252 1141157254 1141164148 1141164154 1141164276 1141164280 1141165470 1141165476 1141166006 1141167822 1141167832 1141171152 1141171336 1141171344 1141172682 1141172686 1141172698 1141173888 1141180592 1141180598 1141187788 1141187790 1141190160 1141192064 1141193282 1141193346 1141194794 1141194800 1141194804 1141194808 1141194810 1141201038 1141201040

Corticosteroids	1140853854 1140854694 1140854700 1140854784 1140854788 1140854816 1140854834 1140854888 1140854916 1140854990 1140857672 1140857678 1140862572 1140868364 1140868370 1140873620 1140874790 1140874792 1140874794 1140874810 1140874814 1140874816 1140874822 1140874896 1140874930 1140874936 1140874940 1140874944 1140874950 1140874954 1140874956 1140874976 1140874978 1140875668 1140875684 1140876032 1140876036 1140876044 1140876046 1140876052 1140876058 1140876076 1140876104 1140876456 1140878562 1140879922 1140879934 1140881938 1140882152 1140882622 1140882624 1140882626 1140882630 1140882694 1140882708 1140882718 1140882722 1140882724 1140882728 1140882730 1140882732 1140882740 1140882742 1140882756 1140882758 1140882764 1140882766 1140882768 1140882774 1140882776 1140882778 1140882780 1140882782 1140882794 1140882800 1140882806 1140882808 1140882816 1140882818 1140882820 1140882822 1140882824 1140882826 1140882830 1140882832 1140882836 1140882840 1140882842 1140882844 1140882846 1140882848 1140882850 1140882852 1140882864 1140882888 1140882892 1140882894 1140882896 1140882898 1140882902 1140882904 1140882906 1140882908 1140882910 1140882914 1140882916 1140882918 1140882920 1140882926 1140882928 1140882932 1140882934 1140882938 1140883022 1140883026 1140883028 1140883030 1140883034 1140883038 1140883040 1140883044 1140883048 1140883052 1140883054 1140883056 1140883058 1140883060 1140883062 1140883064 1140884636 1140884640 1140884642 1140884646 1140884654 1140884660 1140884664 1140884672 1140884676 1140884696 1140884700 1140884704 1140884716 1140888074 1140888092 1140888098 1140888124 1140888130 1140888134 1140888142 1140888150 1140888166 1140888168 1140888172 1140888176 1140888178 1140888184 1140888194 1140909786 1140909894 1140910424 1140910634 1141151424 1141157294 1141157402 1141157418 1141162532 1141164086 1141167174 1141169844 1141173346 1141174512 1141174520 1141174548 1141174552 1141179072 1141179982 1141180342 1141181062 1141181554 1141181610 1141189464 1141191748 1141194840 1141195232 1141195280
Second generation atypical Psychotics	1140867420 1140867432 1140867444 1140927956 1140927970 1140928916 1141152848 1141152860 1141153490 1141167976 1141177762 1141195974 1141202024

Appendix 4: Interactions between educational attainment and polygenic scores for cardiovascular risk factors: cross-sectional and prospective analysis of UK Biobank

Author affiliations

10) MRC Integrative Epidemiology Unit, University of Bristol
11) Population Health Sciences, Bristol Medical School, University of Bristol
12) Clinical Pharmacology and Therapeutics Section, Institute of Medical and Biomedical Education and Institute for Infection and Immunity, St George's, University of London, London, United Kingdom
13) Clinical Pharmacology Group, Pharmacy and Medicines Directorate, St George's University Hospitals NHS Foundation Trust, London, United Kingdom
14) Novo Nordisk Research Centre Oxford, Old Road Campus, Oxford, United Kingdom
15) Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom
16) Centre for Academic Primary Care, University of Bristol
17) NIHR Bristol Biomedical Research Centre, University of Bristol
18) K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Norway.

Tables: List of single nucleotide polymorphisms (SNPs) included in polygenic scores

Appendix 4 Table 1: List of SNPs used in polygenic score for alcohol consumption, measured as drinks per week, at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs11940694	G	A	0.0229	0.00308	4.82E-14
2	rs1229978	C	T	0.0208	0.00301	3.04E-12
3	rs1229984	C	T	0.186	0.0105	1.12E-65
4	rs145441283	G	A	-0.197	0.0196	5.07E-24
5	rs151180	T	G	-0.0229	0.00415	3.84E-08
6	rs1789889	A	G	0.022	0.0038	6.33E-09
7	rs181163639	C	A	-0.0708	0.0115	8.33E-10
8	rs1919208	T	C	0.137	0.0212	4.30E-08
9	rs3114045	C	T	0.0336	0.00461	2.08E-13
10	rs4699680	A	G	0.0428	0.00762	1.97E-08
11	rs55872084	T	G	0.0203	0.00359	2.20E-08
12	rs6676388	C	T	0.0183	0.00307	2.31E-09
13	rs71612659	A	G	-0.0398	0.00719	1.26E-08
14	rs7187575	T	C	0.02	0.00338	5.23E-09

Appendix 4 Table 2: List of SNPs used in polygenic score for body mass index at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs1000940	G	A	0.0192	0.0034	1.28E-08
2	rs10132280	A	C	-0.023	0.0034	1.14E-11
3	rs1016287	C	T	-0.0229	0.0034	2.25E-11
4	rs10176391	G	C	-0.0793	0.0133	2.64E-09
5	rs10182181	G	A	0.0307	0.0031	8.78E-24
6	rs10484664	A	G	0.0185	0.0031	3.76E-09
7	rs10493499	C	T	-0.0263	0.0037	1.48E-12
8	rs10733682	G	A	-0.0174	0.0031	1.83E-08
9	rs10798580	A	G	-0.0177	0.0031	1.16E-08
10	rs10938397	G	A	0.0402	0.0031	3.21E-38
11	rs10968576	G	A	0.0249	0.0033	6.61E-14
12	rs11030066	T	C	0.0282	0.0047	2.14E-09
13	rs11030104	G	A	-0.0414	0.0038	5.56E-28
14	rs11057405	A	G	-0.0307	0.0055	2.02E-08
15	rs11074422	T	A	0.0188	0.0034	3.34E-08
16	rs11074446	C	T	-0.0256	0.0045	1.31E-08
17	rs11075986	G	C	-0.0423	0.006	1.23E-12
18	rs11126666	A	G	0.0207	0.0034	1.33E-09
19	rs11165643	T	C	0.0218	0.0031	2.07E-12
20	rs11191560	C	T	0.0308	0.0053	8.45E-09
21	rs11583200	T	C	-0.0177	0.0031	1.48E-08
22	rs1167827	G	A	0.0202	0.0033	6.33E-10
23	rs11688816	A	G	-0.0172	0.0031	1.89E-08
24	rs11727676	C	T	-0.0358	0.0064	2.55E-08
25	rs11847697	T	C	0.0492	0.0084	3.99E-09
26	rs12286929	G	A	0.0217	0.0031	1.31E-12
27	rs12354124	T	A	-0.0224	0.0031	2.18E-13
28	rs12361415	G	T	-0.0234	0.0034	1.09E-11
29	rs12401738	A	G	0.0211	0.0033	1.15E-10
30	rs12429545	A	G	0.0334	0.0047	1.09E-12
31	rs12446632	A	G	-0.0403	0.0046	1.48E-18
32	rs12566985	A	G	-0.0242	0.0031	3.28E-15
33	rs12607795	C	T	-0.0287	0.0037	1.47E-14
34	rs12885454	A	C	-0.0207	0.0033	1.94E-10
35	rs1292637	G	C	-0.0196	0.0033	2.88E-09
36	rs12940622	A	G	-0.0182	0.0031	2.49E-09
37	rs12996547	T	C	0.0246	0.0033	3.67E-14
38	rs12999373	A	G	-0.0199	0.0036	4.10E-08
39	rs13021737	G	A	0.0601	0.004	1.11E-50
40	rs13078960	G	T	0.0297	0.0039	1.74E-14
41	rs13107325	T	C	0.0477	0.0068	1.83E-12
42	rs1317006	C	T	0.0214	0.0034	4.97E-10
43	rs13191362	G	A	-0.0277	0.0048	7.34E-09
44	rs13411762	T	C	-0.0325	0.0057	1.08E-08
45	rs1344840	A	G	-0.0206	0.0035	2.88E-09
46	rs1477199	G	A	0.0242	0.0044	4.56E-08
47	rs1516725	C	T	0.0451	0.0046	1.89E-22
48	rs1519480	T	C	-0.0306	0.0033	9.58E-21
49	rs1528435	T	C	0.0178	0.0031	1.20E-08
50	rs1558902	A	T	0.0818	0.0031	7.51E-153
51	rs1620977	G	A	-0.0241	0.0035	5.28E-12
52	rs16851483	T	G	0.0483	0.0077	3.55E-10
53	rs16951275	C	T	-0.0311	0.0037	1.91E-17
54	rs17001654	G	C	0.0306	0.0053	7.76E-09

55	rs17024393	C	T	0.0658	0.0088	7.03E-14
56	rs17066842	A	G	-0.0626	0.0083	6.40E-14
57	rs17094222	C	T	0.0249	0.0038	5.94E-11
58	rs17115529	C	A	0.0233	0.0042	2.90E-08
59	rs17391694	T	C	0.0299	0.0052	1.03E-08
60	rs17405819	C	T	-0.0224	0.0033	2.07E-11
61	rs17724992	G	A	-0.0194	0.0035	3.42E-08
62	rs1808579	T	C	-0.0167	0.0031	4.17E-08
63	rs1927850	A	C	-0.0187	0.0031	1.69E-09
64	rs1928295	C	T	-0.0188	0.0031	7.91E-10
65	rs1943229	T	G	0.0274	0.0039	2.80E-12
66	rs2033529	G	A	0.019	0.0033	1.39E-08
67	rs2033732	C	T	0.0192	0.0035	4.89E-08
68	rs205262	G	A	0.0221	0.0035	1.75E-10
69	rs2058908	C	T	0.0637	0.0039	5.87E-60
70	rs2075650	G	A	-0.0258	0.0045	1.25E-08
71	rs2112347	G	T	-0.0261	0.0031	6.19E-17
72	rs2121279	T	C	0.0245	0.0044	2.31E-08
73	rs2176598	C	T	-0.0198	0.0036	2.97E-08
74	rs2207139	G	A	0.0447	0.004	4.13E-29
75	rs2245368	T	C	-0.0317	0.0057	3.19E-08
76	rs2287019	T	C	-0.036	0.0042	4.59E-18
77	rs2365389	T	C	-0.02	0.0031	1.63E-10
78	rs2650492	A	G	0.0207	0.0035	1.92E-09
79	rs2744489	A	G	-0.0171	0.0031	1.94E-08
80	rs2817419	A	G	0.0275	0.0035	3.66E-15
81	rs2820292	C	A	0.0195	0.0031	1.83E-10
82	rs2821236	C	T	-0.0282	0.0038	6.30E-14
83	rs29941	G	A	0.0182	0.0033	2.41E-08
84	rs3101336	C	T	0.0334	0.0031	2.66E-26
85	rs3736485	G	A	-0.0176	0.0031	7.41E-09
86	rs3810291	A	G	0.0283	0.0036	4.81E-15
87	rs3817334	T	C	0.0262	0.0031	5.15E-17
88	rs3849570	A	C	0.0188	0.0034	2.60E-08
89	rs3888190	A	C	0.0309	0.0031	3.14E-23
90	rs423934	C	T	0.0226	0.0035	5.59E-11
91	rs4256980	G	C	0.0209	0.0031	2.90E-11
92	rs4280233	T	G	-0.0374	0.0067	3.00E-08
93	rs4514364	T	C	-0.0195	0.0033	4.46E-09
94	rs4611674	G	A	-0.0218	0.0031	4.54E-12
95	rs4671328	G	T	-0.0215	0.0037	6.22E-09
96	rs4740619	C	T	-0.0179	0.0031	4.56E-09
97	rs4788115	A	T	-0.0285	0.0048	3.28E-09
98	rs4940929	G	C	0.0274	0.0042	6.85E-11
99	rs543874	G	A	0.0482	0.0039	2.62E-35
100	rs561634	T	A	0.0195	0.0031	2.32E-10
101	rs6477694	T	C	-0.0174	0.0031	2.67E-08
102	rs6499653	C	T	-0.0269	0.0037	2.32E-13
103	rs6567160	C	T	0.0556	0.0036	3.93E-53
104	rs657452	G	A	-0.0227	0.0031	5.48E-13
105	rs6749646	T	A	0.027	0.0047	9.21E-09
106	rs6804842	G	A	0.0185	0.0031	2.48E-09
107	rs6845132	T	C	0.0182	0.0031	2.50E-09
108	rs7138803	A	G	0.0315	0.0031	8.15E-24
109	rs7141420	T	C	0.0235	0.0031	1.23E-14
110	rs7144011	T	G	0.0289	0.0038	1.38E-14
111	rs7186521	G	A	0.0279	0.0031	1.50E-19
112	rs7203521	A	G	0.0326	0.0032	3.46E-24
113	rs7243357	G	T	-0.0217	0.004	3.86E-08

114	rs758747	T	C	0.0225	0.0037	7.47E-10
115	rs7599312	A	G	-0.022	0.0034	1.17E-10
116	rs7629375	A	C	-0.021	0.0031	1.85E-11
117	rs7899106	G	A	0.0395	0.0071	2.96E-08
118	rs7903146	T	C	-0.0234	0.0034	1.11E-11
119	rs8097783	A	G	-0.0398	0.006	4.20E-11
120	rs879620	T	C	0.0244	0.004	1.06E-09
121	rs9400239	C	T	0.0188	0.0033	1.61E-08
122	rs9579083	C	G	0.0295	0.0047	3.46E-10
123	rs9829032	G	A	0.0194	0.0032	1.86E-09
124	rs9925964	G	A	-0.0192	0.0031	8.11E-10
125	rs9945063	T	C	0.0217	0.0038	1.35E-08
126	rs9947301	T	C	-0.0377	0.0057	3.70E-11
127	rs9956279	T	C	0.0348	0.0033	2.62E-25

Appendix 4 Table 3:List of SNPs used in polygenic score for Low-density lipoprotein cholesterol at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs10101167	G	C	0.0208	0.0038	4.41E-08
2	rs10102164	A	G	0.0301	0.0043	2.56E-12
3	rs10102352	G	A	0.0399	0.005	1.46E-15
4	rs10178381	T	A	0.0557	0.0064	3.23E-18
5	rs10184004	T	C	-0.0205	0.0037	3.02E-08
6	rs10208987	G	T	-0.0439	0.0065	1.44E-11
7	rs10209020	T	C	0.0295	0.0039	3.91E-14
8	rs1025447	C	T	0.0344	0.0046	7.53E-14
9	rs103294	T	C	0.0314	0.0045	3.00E-12
10	rs10401969	C	T	-0.1369	0.007	3.59E-85
11	rs10402271	G	T	0.0702	0.0037	2.85E-80
12	rs10403668	A	G	-0.0392	0.005	4.51E-15
13	rs10455872	G	A	0.1238	0.014	9.33E-19
14	rs10468017	T	C	0.0617	0.004	1.11E-53
15	rs1048699	T	C	0.0315	0.0056	1.86E-08
16	rs10489488	A	G	-0.0906	0.0165	4.00E-08
17	rs10490626	A	G	-0.0415	0.0066	3.22E-10
18	rs10493329	G	A	-0.0452	0.0048	4.66E-21
19	rs10515214	G	A	0.0456	0.0048	2.10E-21
20	rs10757056	T	C	0.0265	0.0047	1.72E-08
21	rs10773003	A	G	0.0369	0.0058	1.99E-10
22	rs1077514	T	C	0.0301	0.0052	7.10E-09
23	rs1077834	C	T	0.0652	0.0043	6.24E-52
24	rs10832962	T	C	0.0315	0.0039	6.64E-16
25	rs10838738	G	A	-0.0208	0.0038	4.41E-08
26	rs10903129	G	A	0.029	0.0035	1.17E-16
27	rs10904908	G	A	0.025	0.0036	3.80E-12
28	rs10910490	A	G	0.0398	0.0049	4.57E-16
29	rs1107851	C	T	-0.0243	0.0035	3.84E-12
30	rs11096689	T	C	-0.0667	0.0039	1.42E-65
31	rs11102964	C	T	-0.0405	0.0047	6.87E-18
32	rs1111826230	G	A	0.0615	0.0094	6.05E-11
33	rs11206510	C	T	-0.069	0.0048	7.43E-47
34	rs11206514	A	C	0.0355	0.0039	8.82E-20
35	rs11208004	A	G	-0.0758	0.0037	2.84E-93
36	rs11216137	A	G	-0.0742	0.0077	5.61E-22
37	rs11220462	A	G	0.0474	0.0058	3.02E-16
38	rs11230815	C	G	-0.0361	0.0061	3.26E-09
39	rs1129555	G	A	-0.0317	0.0039	4.36E-16
40	rs11563251	T	C	0.0368	0.0059	4.45E-10
41	rs11591147	T	G	-0.3341	0.0173	4.25E-83
42	rs11603023	C	T	-0.0216	0.0036	1.97E-09
43	rs11659960	C	G	0.0267	0.0041	7.41E-11
44	rs11668536	T	C	-0.0388	0.0044	1.16E-18
45	rs11672862	T	C	-0.0537	0.0078	5.79E-12
46	rs11679386	C	T	0.0393	0.0059	2.72E-11
47	rs11685356	T	C	0.0474	0.0041	6.50E-31
48	rs11694172	G	A	0.0277	0.0041	1.42E-11
49	rs11699690	A	G	-0.0362	0.0062	5.26E-09
50	rs11709504	C	T	-0.0322	0.0045	8.33E-13
51	rs11742194	T	C	0.056	0.0058	4.67E-22
52	rs1174604	C	T	0.021	0.0038	3.27E-08
53	rs11753995	A	G	0.0489	0.0048	2.25E-24
54	rs117733303	G	A	0.1303	0.0213	9.51E-10

55	rs11789603	T	C	0.0427	0.0062	5.69E-12
56	rs11820504	C	T	0.0266	0.0045	3.40E-09
57	rs11858279	C	T	0.0431	0.0044	1.18E-22
58	rs11875600	G	A	0.0533	0.0076	2.33E-12
59	rs11881156	T	C	-0.0689	0.0048	1.00E-46
60	rs12052201	T	G	-0.0613	0.0042	3.01E-48
61	rs12122434	G	A	-0.0666	0.0098	1.08E-11
62	rs12123703	G	A	-0.0567	0.01	1.43E-08
63	rs12208357	T	C	0.0576	0.0098	4.16E-09
64	rs12270837	C	A	-0.056	0.0074	3.80E-14
65	rs12285095	G	T	0.1024	0.0072	6.67E-46
66	rs12309	T	C	0.0258	0.0047	4.03E-08
67	rs12321904	T	G	0.0222	0.0036	6.97E-10
68	rs12448528	G	A	0.0461	0.005	2.97E-20
69	rs1264344	T	C	0.0218	0.0038	9.65E-09
70	rs12660382	T	C	0.0266	0.0046	7.36E-09
71	rs12670798	C	T	0.0364	0.0041	6.80E-19
72	rs12691202	T	C	-0.1031	0.0109	3.12E-21
73	rs12708454	C	A	0.0366	0.0057	1.35E-10
74	rs12710745	G	A	-0.0405	0.0037	6.95E-28
75	rs12720796	C	A	0.0821	0.0133	6.70E-10
76	rs12720842	C	T	0.0892	0.0111	9.28E-16
77	rs12721109	A	G	-0.3234	0.0179	5.79E-73
78	rs12749263	C	T	0.0357	0.0041	3.11E-18
79	rs12916	C	T	0.0684	0.0036	1.71E-80
80	rs12920974	T	G	-0.029	0.0053	4.46E-08
81	rs12924285	A	G	0.0371	0.006	6.28E-10
82	rs12931964	G	T	0.0356	0.0049	3.72E-13
83	rs12983316	G	A	0.0402	0.005	8.98E-16
85	rs13277646	G	A	-0.0217	0.0039	2.64E-08
86	rs13292582	G	A	-0.0572	0.0052	3.82E-28
87	rs13315871	A	G	-0.0355	0.0061	5.90E-09
88	rs13344893	T	C	-0.0352	0.0045	5.19E-15
89	rs13375691	T	C	-0.06	0.0062	3.76E-22
90	rs13396400	G	A	0.0294	0.0036	3.17E-16
91	rs13465	G	A	0.084	0.0082	1.26E-24
92	rs1367117	A	G	0.0995	0.0038	4.02E-151
93	rs138764	C	T	-0.0214	0.0037	7.30E-09
94	rs1475701	C	T	0.0652	0.0089	2.37E-13
95	rs1475961	G	A	0.0206	0.0037	2.58E-08
96	rs1494369	G	A	-0.0292	0.0053	3.60E-08
97	rs1501909	T	G	0.0347	0.0038	6.75E-20
98	rs1525764	A	T	0.0237	0.004	3.12E-09
99	rs1529711	T	C	0.0315	0.0049	1.29E-10
100	rs1534842	G	A	-0.0416	0.0076	4.41E-08
101	rs1535	G	A	-0.0497	0.0037	3.90E-41
102	rs157580	A	G	0.0969	0.0043	1.89E-112
103	rs1594895	C	T	-0.0259	0.0045	8.64E-09
104	rs1604144	T	C	0.0215	0.0039	3.53E-08
105	rs16831243	T	C	0.0378	0.0053	9.89E-13
106	rs16872670	A	G	0.0525	0.0081	9.08E-11
107	rs16941759	A	G	0.0264	0.0048	3.80E-08
108	rs16942887	A	G	0.031	0.0052	2.50E-09
109	rs16970670	T	A	0.0529	0.0092	8.92E-09
110	rs16979372	G	T	-0.1057	0.0094	2.46E-29
111	rs16979595	A	G	0.0296	0.005	3.22E-09
112	rs17035630	A	G	0.0402	0.0058	4.18E-12
113	rs17035665	T	C	-0.0594	0.0057	1.99E-25
114	rs17035949	G	T	-0.0944	0.0133	1.27E-12

115	rs17248748	T	C	-0.1062	0.0176	1.60E-09
116	rs17301746	T	C	0.0798	0.0146	4.61E-08
117	rs17398765	G	A	0.0734	0.0073	8.75E-24
118	rs17405319	T	C	0.0429	0.0049	2.04E-18
119	rs17424122	A	T	0.0653	0.0087	6.11E-14
120	rs174468	A	G	0.0232	0.0038	1.03E-09
121	rs174532	A	G	0.0331	0.0041	6.85E-16
122	rs174602	C	T	-0.0334	0.0056	2.46E-09
123	rs174634	C	G	-0.0267	0.0042	2.06E-10
124	rs17584208	A	G	-0.0803	0.0064	4.14E-36
125	rs17630235	A	G	-0.0298	0.0036	1.25E-16
126	rs17649913	C	T	-0.0243	0.0044	3.34E-08
127	rs17651629	T	C	-0.0324	0.0059	3.98E-08
128	rs17661330	G	T	-0.0261	0.0042	5.16E-10
129	rs17800819	T	C	-0.033	0.0053	4.77E-10
130	rs17819328	G	T	0.0277	0.0037	7.07E-14
131	rs17821316	C	A	-0.0535	0.0067	1.40E-15
132	rs1787328	C	T	0.0268	0.0038	1.76E-12
133	rs1800562	A	G	-0.0565	0.0077	2.17E-13
134	rs1800961	T	C	-0.1062	0.0101	7.38E-26
135	rs1801701	T	C	0.0497	0.0062	1.09E-15
136	rs180326	T	G	-0.0443	0.0039	6.69E-30
137	rs181360	G	T	-0.0278	0.0043	1.01E-10
138	rs1825955	A	C	0.0438	0.0058	4.30E-14
139	rs1874776	C	T	0.0339	0.0042	6.95E-16
140	rs1883025	T	C	-0.0671	0.0042	1.87E-57
141	rs1943681	T	A	-0.0295	0.0037	1.55E-15
142	rs1943979	A	G	0.0298	0.0036	1.25E-16
143	rs1997243	G	A	0.0332	0.005	3.14E-11
144	rs2000813	T	C	0.0226	0.004	1.60E-08
145	rs2000999	A	G	0.0617	0.0044	1.13E-44
146	rs2006760	G	C	0.0534	0.0074	5.35E-13
147	rs2023472	G	A	0.0214	0.0039	4.08E-08
148	rs2035191	C	T	0.0556	0.0045	4.55E-35
149	rs207145	T	C	0.042	0.0055	2.23E-14
150	rs2071593	A	G	0.0389	0.0063	6.63E-10
151	rs2073048	A	G	0.0298	0.0048	5.35E-10
152	rs2073547	G	A	0.0456	0.0047	2.95E-22
153	rs2075650	G	A	0.1432	0.0052	6.08E-167
154	rs2148489	C	T	-0.028	0.0042	2.62E-11
155	rs2149116	A	G	0.0407	0.0052	5.00E-15
156	rs2155216	T	C	0.0895	0.0146	8.78E-10
157	rs2156499	A	G	-0.0262	0.0039	1.84E-11
158	rs2156552	T	A	-0.057	0.0047	7.54E-34
159	rs217181	T	C	-0.0572	0.0047	4.48E-34
160	rs217386	A	G	-0.0338	0.0036	6.06E-21
161	rs217420	C	A	0.023	0.0042	4.35E-08
162	rs2178198	T	C	0.0393	0.0053	1.22E-13
163	rs2194562	A	G	0.0419	0.0059	1.23E-12
164	rs2199403	T	C	-0.0221	0.0036	8.31E-10
165	rs2230365	T	C	0.0307	0.0049	3.72E-10
166	rs2230808	C	T	0.031	0.0042	1.57E-13
167	rs2235367	G	A	0.0357	0.0035	1.98E-24
168	rs2244608	G	A	0.0313	0.0037	2.69E-17
169	rs2247056	C	T	0.0391	0.0041	1.48E-21
170	rs2248372	A	G	0.0257	0.0037	3.76E-12
171	rs2266788	A	G	-0.1138	0.0071	8.12E-58
173	rs2287019	T	C	-0.0292	0.0046	2.18E-10
174	rs2287623	A	G	-0.0273	0.0036	3.37E-14

175	rs2288904	G	A	0.0409	0.0045	1.00E-19
176	rs2294261	C	A	-0.0245	0.0036	1.01E-11
177	rs2297374	T	C	-0.0311	0.0036	5.68E-18
178	rs2305929	G	A	0.0277	0.0046	1.73E-09
179	rs2326077	T	C	-0.0388	0.0036	4.38E-27
180	rs2336438	C	T	0.0613	0.0111	3.34E-08
181	rs2390536	A	G	0.0221	0.0037	2.33E-09
182	rs2394427	A	G	0.0388	0.005	8.49E-15
183	rs2395471	A	G	0.0334	0.0039	1.09E-17
184	rs2479394	A	G	-0.0359	0.0039	3.41E-20
185	rs2479409	A	G	-0.054	0.004	1.56E-41
186	rs2495477	G	A	-0.0452	0.0052	3.55E-18
187	rs2516440	A	G	0.028	0.0041	8.53E-12
188	rs2517546	T	C	-0.0407	0.0056	3.65E-13
189	rs2521567	A	G	-0.0205	0.0035	4.71E-09
190	rs2523864	T	C	-0.0236	0.004	3.64E-09
191	rs2596501	T	C	0.0233	0.0036	9.66E-11
192	rs2621321	G	A	0.0242	0.004	1.45E-09
193	rs2642438	G	A	0.037	0.004	2.24E-20
194	rs2737252	A	G	-0.0331	0.0039	2.12E-17
195	rs2738464	C	G	-0.0364	0.0059	6.85E-10
196	rs2758886	A	G	0.0232	0.0039	2.70E-09
197	rs2770	A	G	0.0327	0.0055	2.76E-09
199	rs2814982	T	C	-0.0441	0.0057	1.02E-14
200	rs283813	A	T	0.1104	0.009	1.37E-34
201	rs2845573	G	A	-0.051	0.0059	5.42E-18
202	rs2857595	A	G	-0.0369	0.0048	1.50E-14
203	rs2858331	G	A	0.0279	0.0037	4.68E-14
204	rs28718232	G	A	0.0383	0.0062	6.52E-10
205	rs2886232	C	T	-0.0358	0.0062	7.73E-09
206	rs2894254	G	T	-0.0466	0.0059	2.83E-15
207	rs289741	A	G	-0.0296	0.004	1.36E-13
208	rs2899624	G	A	-0.0382	0.0051	6.88E-14
209	rs2902940	G	A	-0.0241	0.0039	6.43E-10
210	rs2920500	A	G	0.0243	0.0036	1.48E-11
212	rs2960420	G	C	0.024	0.0039	7.56E-10
213	rs2965101	C	T	-0.0499	0.0038	2.17E-39
214	rs2965156	C	G	-0.0353	0.0057	5.90E-10
215	rs2965157	C	T	-0.1222	0.0107	3.30E-30
216	rs2965185	C	T	-0.042	0.0039	4.81E-27
217	rs2972564	G	A	0.0474	0.0065	3.05E-13
218	rs2980885	A	G	-0.0337	0.0044	1.87E-14
219	rs3312046	T	C	0.0406	0.0038	1.21E-26
220	rs3124785	A	G	0.0374	0.0057	5.33E-11
221	rs3125055	A	T	0.0467	0.0053	1.24E-18
222	rs3132454	G	A	0.0234	0.0037	2.54E-10
223	rs314253	C	T	-0.0233	0.0037	3.03E-10
224	rs3184504	C	T	0.0318	0.0037	8.36E-18
225	rs3208856	T	C	-0.2034	0.0188	2.79E-27
226	rs3213422	C	A	-0.0277	0.0037	7.07E-14
227	rs3745157	C	T	-0.0252	0.0038	3.32E-11
228	rs3757354	T	C	-0.0348	0.0042	1.17E-16
229	rs3764261	A	C	0.0503	0.004	2.90E-36
230	rs3780181	G	A	-0.0442	0.0071	4.80E-10
232	rs3786721	C	T	-0.0366	0.0037	4.51E-23
233	rs379309	T	C	-0.0266	0.0038	2.56E-12
234	rs3798180	G	A	-0.0254	0.0036	1.72E-12
235	rs3798221	T	G	-0.0317	0.0043	1.68E-13
236	rs3800406	G	A	-0.0437	0.006	3.26E-13

237	rs3810444	A	T	0.0648	0.0081	1.24E-15
238	rs3817588	C	T	-0.0438	0.0044	2.41E-23
239	rs3823151	C	A	0.0554	0.01	3.02E-08
240	rs3873380	T	C	0.0243	0.0038	1.61E-10
241	rs387976	C	A	-0.0697	0.0056	1.46E-35
242	rs3891175	T	C	-0.0302	0.0047	1.31E-10
243	rs389883	T	G	0.0346	0.0043	8.52E-16
244	rs3935470	G	A	0.0389	0.0038	1.36E-24
245	rs4148177	A	G	-0.0364	0.0049	1.10E-13
246	rs4148218	A	G	-0.0385	0.0045	1.17E-17
247	rs4149311	T	C	0.0456	0.0055	1.12E-16
248	rs4245791	T	C	-0.078	0.004	1.10E-84
249	rs4253772	T	C	0.0322	0.0058	2.83E-08
250	rs4360309	T	C	0.0387	0.0036	5.93E-27
251	rs4382144	A	G	0.0285	0.0035	3.86E-16
252	rs4530754	A	G	0.0228	0.0035	7.30E-11
253	rs461473	A	G	-0.0384	0.0062	5.88E-10
254	rs4622454	T	C	0.0234	0.0036	8.03E-11
255	rs4635554	G	T	0.0691	0.004	7.26E-67
256	rs4666366	C	T	-0.0219	0.004	4.38E-08
257	rs4704810	A	G	0.0219	0.0036	1.18E-09
258	rs4711268	T	C	0.0335	0.0041	3.07E-16
259	rs4722551	C	T	0.029	0.0047	6.82E-10
260	rs4752805	G	A	0.0251	0.0041	9.24E-10
261	rs4783962	C	T	0.0242	0.0043	1.82E-08
262	rs4788589	A	T	-0.0296	0.0043	5.83E-12
263	rs4803750	G	A	-0.1485	0.0075	2.98E-87
264	rs4803760	C	T	0.0837	0.0048	4.28E-68
265	rs4803767	T	C	0.0417	0.0058	6.50E-13
266	rs4803770	G	C	0.0431	0.0042	1.05E-24
267	rs4804158	C	T	0.0243	0.0038	1.61E-10
268	rs4808802	C	G	0.0251	0.0044	1.17E-08
269	rs486394	C	A	0.0307	0.004	1.65E-14
270	rs488191	G	A	0.0358	0.0062	7.73E-09
271	rs4921914	T	C	-0.0332	0.0042	2.68E-15
272	rs4926670	T	C	-0.0636	0.0058	5.60E-28
273	rs4938303	T	C	-0.0414	0.0039	2.53E-26
274	rs4953023	A	G	-0.1249	0.0072	2.07E-67
275	rs4968255	T	C	0.0275	0.0048	1.01E-08
276	rs4988235	A	G	-0.0308	0.004	1.36E-14
277	rs505000	T	C	0.0253	0.0045	1.89E-08
278	rs5110	A	C	0.0857	0.0138	5.29E-10
279	rs511676	G	T	-0.0523	0.0054	3.49E-22
280	rs516246	T	C	0.0315	0.0037	1.69E-17
282	rs533556	C	A	-0.0433	0.0037	1.23E-31
283	rs533617	C	T	-0.1111	0.0095	1.36E-31
284	rs541041	A	G	0.1245	0.0046	2.53E-161
285	rs548638	G	T	-0.0253	0.0045	1.89E-08
286	rs558971	G	A	0.0398	0.0036	2.06E-28
287	rs570877	G	T	0.0864	0.0063	8.34E-43
288	rs572512	T	C	0.0361	0.0045	1.04E-15
289	rs5742911	G	A	-0.0468	0.0055	1.75E-17
290	rs5763662	T	C	0.0692	0.0117	3.33E-09
291	rs579459	C	T	0.062	0.0044	4.32E-45
292	rs581080	C	G	-0.0377	0.0047	1.05E-15
293	rs584626	T	C	0.0437	0.0047	1.43E-20
294	rs585362	T	C	0.0703	0.0053	3.74E-40
295	rs5880	C	G	0.0622	0.0092	1.37E-11
296	rs599839	A	G	0.1281	0.0042	2.61E-204

297	rs6016381	C	T	-0.0328	0.0036	8.15E-20
298	rs6124309	G	A	0.0286	0.0046	5.05E-10
299	rs629001	T	C	0.0847	0.0076	7.60E-29
300	rs630014	G	A	0.0295	0.0036	2.52E-16
301	rs633862	T	C	0.0244	0.0036	1.22E-11
302	rs6413458	A	G	-0.0802	0.0134	2.16E-09
303	rs6435161	G	T	-0.0243	0.0039	4.64E-10
304	rs648673	G	C	-0.0411	0.0055	7.85E-14
305	rs6504872	T	C	0.025	0.0035	9.14E-13
306	rs6511720	T	G	-0.1851	0.0059	4.74E-216
307	rs655246	G	A	0.0439	0.0036	3.33E-34
308	rs6587970	A	G	-0.0286	0.0046	5.05E-10
309	rs6603981	T	C	0.0351	0.0043	3.27E-16
310	rs6662286	C	T	0.0691	0.007	5.54E-23
311	rs6664692	T	C	0.0289	0.0044	5.09E-11
312	rs6689614	A	G	0.0507	0.0036	4.81E-45
313	rs6725189	T	G	-0.0548	0.0043	3.36E-37
314	rs6728178	A	G	-0.0499	0.004	1.02E-35
315	rs6729410	G	A	-0.036	0.0038	2.70E-21
316	rs6730157	G	A	0.0298	0.0039	2.15E-14
317	rs6739502	G	A	-0.0315	0.0035	2.26E-19
318	rs6756743	T	C	0.0551	0.0089	5.98E-10
319	rs676385	G	A	0.0254	0.0042	1.47E-09
320	rs6818397	G	T	-0.0254	0.0039	7.38E-11
321	rs6831256	G	A	0.023	0.0037	5.09E-10
322	rs6859	G	A	-0.0636	0.0037	3.20E-66
323	rs6873053	G	A	0.0396	0.0063	3.26E-10
324	rs688	T	C	0.0416	0.0036	6.92E-31
325	rs6882076	C	T	0.0508	0.0037	6.74E-43
326	rs6917747	A	G	0.0324	0.0052	4.64E-10
327	rs6935921	T	C	0.0254	0.0041	5.82E-10
328	rs709167	T	G	-0.0227	0.0037	8.51E-10
329	rs7117842	C	T	0.0294	0.0036	3.17E-16
330	rs714948	A	C	0.0364	0.006	1.31E-09
331	rs7164909	T	C	-0.0397	0.0068	5.28E-09
332	rs7188	C	A	0.0426	0.0042	3.57E-24
333	rs7193549	C	T	0.0332	0.0048	4.62E-12
334	rs7229377	T	C	0.0362	0.0046	3.56E-15
335	rs7235005	A	G	0.0244	0.0036	1.22E-11
336	rs7241596	T	C	0.0274	0.0037	1.31E-13
337	rs7255743	A	G	-0.1066	0.015	1.19E-12
338	rs7259004	C	G	0.1247	0.0088	1.40E-45
339	rs7264396	T	C	-0.0313	0.0043	3.36E-13
340	rs72703204	A	G	-0.1193	0.0136	1.75E-18
341	rs73015030	A	G	-0.1319	0.0142	1.56E-20
342	rs732839	A	G	0.0269	0.0043	3.95E-10
343	rs7349418	T	C	-0.0201	0.0036	2.36E-08
344	rs739468	G	T	-0.0378	0.0061	5.77E-10
345	rs73959582	C	T	0.0437	0.0076	8.92E-09
346	rs74019428	T	C	-0.0805	0.0135	2.48E-09
347	rs7412	T	C	-0.3736	0.0096	0
348	rs742748	C	T	-0.022	0.0037	2.75E-09
349	rs7499892	T	C	-0.0513	0.005	1.07E-24
350	rs7512480	T	C	0.0339	0.0036	4.66E-21
351	rs7515901	T	C	-0.0407	0.005	3.95E-16
352	rs7544735	A	G	0.0273	0.0039	2.56E-12
353	rs7550711	T	C	-0.0566	0.0101	2.10E-08
354	rs7551981	T	G	0.0358	0.0037	3.83E-22
355	rs7567653	A	G	-0.1009	0.0107	4.10E-21

356	rs7578637	A	G	-0.1124	0.0181	5.30E-10
357	rs7616006	G	A	-0.0315	0.0036	2.13E-18
358	rs7640978	T	C	-0.0376	0.0066	1.22E-08
359	rs7715806	T	C	0.0355	0.0037	8.42E-22
360	rs7742144	C	T	-0.0238	0.004	2.68E-09
361	rs7770628	T	C	-0.0245	0.0036	1.01E-11
362	rs7774197	C	A	0.0507	0.007	4.39E-13
363	rs780093	C	T	-0.0515	0.0036	2.02E-46
364	rs7832643	T	G	0.0289	0.0037	5.68E-15
365	rs8017377	A	G	0.0251	0.0037	1.17E-11
366	rs8044335	C	A	0.0288	0.0035	1.89E-16
367	rs8044476	G	A	0.0317	0.0051	5.11E-10
368	rs8060878	G	A	0.0287	0.0035	2.40E-16
369	rs8069974	C	G	0.0244	0.0041	2.66E-09
370	rs8103315	A	C	0.0422	0.0055	1.68E-14
371	rs8176720	C	T	-0.0257	0.0037	3.76E-12
372	rs8180991	G	C	0.0483	0.0043	2.82E-29
373	rs865774	T	C	-0.0328	0.0053	6.07E-10
374	rs868943	A	G	-0.0292	0.0036	5.02E-16
375	rs873870	A	G	-0.0216	0.0038	1.31E-08
376	rs887829	T	C	-0.0228	0.0037	7.18E-10
377	rs888246	T	C	0.0617	0.0061	4.75E-24
378	rs889545	A	G	-0.0581	0.0091	1.72E-10
379	rs892114	G	A	-0.0294	0.0045	6.43E-11
380	rs904743	G	A	0.0611	0.0052	7.06E-32
381	rs914547	T	C	-0.0269	0.0046	4.98E-09
382	rs926054	G	T	-0.0334	0.0057	4.64E-09
383	rs9273363	A	C	-0.0232	0.004	6.63E-09
384	rs9275406	T	G	0.0328	0.0043	2.39E-14
385	rs9282575	A	G	-0.1076	0.0128	4.23E-17
386	rs9302635	C	T	-0.0366	0.0047	6.85E-15
387	rs9334287	G	A	0.0278	0.0046	1.51E-09
388	rs936960	G	T	-0.0376	0.0066	1.22E-08
389	rs9376090	C	T	-0.0254	0.004	2.15E-10
390	rs9378212	T	C	0.0355	0.0052	8.68E-12
391	rs9391858	G	A	0.0495	0.005	4.16E-23
392	rs940434	T	C	-0.021	0.0038	3.27E-08
393	rs9457843	T	C	-0.0297	0.0054	3.80E-08
394	rs9501587	A	G	0.0292	0.0046	2.18E-10
395	rs970548	C	A	0.025	0.004	4.10E-10
396	rs9951669	G	A	0.027	0.0043	3.41E-10
397	rs9972882	C	A	0.0243	0.004	1.24E-09
398	rs9987289	G	A	0.0842	0.0063	9.67E-41

Appendix 4 Table 4: List of SNPs used in polygenic score for lifetime smoking behaviour in the sample 1 GWAS at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs10093628	C	T	0.0165694	0.00272207	1.20E-09
2	rs10187072	T	C	-0.0130905	0.00235359	2.70E-08
3	rs10226228	G	A	0.0141999	0.0024083	3.70E-09
4	rs10233018	G	A	0.0129883	0.00232256	2.20E-08
5	rs10750016	A	T	0.0160683	0.00239178	1.80E-11
6	rs11030088	A	G	0.0157839	0.00266804	3.30E-09
7	rs112151537	T	C	0.0289527	0.00500391	7.20E-09
8	rs113382419	A	C	0.0242553	0.00368267	4.50E-11
9	rs16879271	C	A	0.0325007	0.00593948	4.50E-08
10	rs17159727	C	T	-0.0241372	0.0040898	3.60E-09
11	rs2890772	T	G	0.0140201	0.00236079	2.90E-09
12	rs326341	A	G	-0.013281	0.00233329	1.30E-08
13	rs4763463	A	G	-0.0132272	0.00238627	3.00E-08
14	rs4841235	G	A	-0.0130052	0.00234148	2.80E-08
15	rs4856463	T	C	-0.0156642	0.00282864	3.10E-08
16	rs4957528	C	A	0.0160705	0.00287694	2.30E-08
17	rs499257	C	T	-0.0146849	0.0024706	2.80E-09
18	rs6590701	T	G	0.0144894	0.00264731	4.40E-08
19	rs6852351	T	C	-0.0132797	0.00240999	3.60E-08
20	rs71673396	C	T	-0.0159243	0.00290616	4.30E-08
21	rs7173514	T	C	-0.0224245	0.00277906	7.10E-16
22	rs8042849	T	C	-0.0175789	0.00244843	7.00E-13
23	rs986391	A	G	-0.0151657	0.00240275	2.80E-10

Appendix 4 Table 5: List of SNPs used in polygenic score for lifetime smoking behaviour in the sample 2 GWAS at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs10922907	T	A	-0.0134411	0.00232508	7.40E-09
2	rs112151537	T	C	0.0294478	0.00497115	3.10E-09
3	rs12897150	T	A	0.0136766	0.00234592	5.50E-09
4	rs12900091	G	A	-0.0131482	0.00231911	1.40E-08
5	rs12914385	T	C	0.0189068	0.00237702	1.80E-15
6	rs13292239	A	G	0.0152012	0.00248573	9.60E-10
7	rs159058	C	A	0.0142308	0.0025321	1.90E-08
8	rs16824949	T	G	0.0145352	0.00231811	3.60E-10
9	rs17657924	A	C	-0.0132644	0.00233124	1.30E-08
10	rs1863161	A	G	0.0127982	0.00232631	3.80E-08
11	rs263771	A	C	0.0151079	0.00274433	3.70E-08
12	rs28669908	A	C	-0.0237801	0.00282163	3.50E-17
13	rs3025354	T	C	0.0150973	0.00237188	2.00E-10
14	rs34794623	A	C	0.0195197	0.00282212	4.60E-12
15	rs45568238	G	C	0.0225856	0.00308616	2.50E-13
16	rs45577732	G	C	0.0386369	0.00429408	2.30E-19
17	rs56116178	G	A	0.0306714	0.00380628	7.70E-16
18	rs62261249	C	T	0.0158442	0.00263127	1.70E-09
19	rs7559547	T	C	0.0219576	0.00304622	5.70E-13
20	rs7948789	G	A	0.0164578	0.00237675	4.40E-12
21	rs8043105	T	C	-0.0287382	0.0052151	3.60E-08

Appendix 4 Table 6: List of SNPs used in polygenic score for systolic blood pressure in the sample 1 GWAS at the genome-wide significance level with a clumping threshold of 500kb and an R2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs1009358	C	T	-0.369257	0.0644871	1.00E-08
2	rs10171080	C	G	0.462857	0.0647265	8.60E-13
3	rs10269774	A	G	-0.385427	0.0669084	8.40E-09
4	rs1032777	C	T	-0.475747	0.0649262	2.30E-13
5	rs1053924	C	T	0.388981	0.0675514	8.50E-09
6	rs10750766	A	C	0.393005	0.0691166	1.30E-08
7	rs10764329	C	G	0.3665	0.0636834	8.70E-09
8	rs10769602	T	A	-0.556702	0.0890515	4.10E-10
9	rs1077394	T	C	0.373884	0.0668653	2.20E-08
10	rs10786156	G	C	-0.376662	0.0631706	2.50E-09
11	rs10838873	C	T	-0.562113	0.0859116	6.00E-11
12	rs10857147	T	A	0.815677	0.0690015	3.00E-32
13	rs10883543	T	G	0.699532	0.0996261	2.20E-12
14	rs10995311	G	C	-0.394142	0.0633664	5.00E-10
15	rs11105429	C	T	0.487352	0.0710168	6.80E-12
16	rs11187837	C	T	-0.556212	0.100308	2.90E-08
17	rs11191580	C	T	-1.09699	0.118001	1.50E-20
18	rs1121450	T	C	-0.575418	0.0903921	1.90E-10
19	rs11246486	T	C	-0.5443	0.0920554	3.40E-09
20	rs11634851	G	C	0.421599	0.062824	1.90E-11
21	rs11646677	C	T	0.360837	0.063622	1.40E-08
22	rs11650511	T	C	0.408482	0.0633763	1.20E-10
23	rs1173690	G	A	-0.417164	0.0645566	1.00E-10
24	rs1175651	T	C	0.420797	0.0769911	4.60E-08
25	rs11853441	G	T	0.357549	0.0626497	1.10E-08
26	rs12136566	G	A	0.410475	0.0670484	9.20E-10
27	rs12221645	C	T	-0.551699	0.0895822	7.30E-10
28	rs12258967	G	C	-0.658175	0.0685584	8.00E-22
29	rs12418543	G	A	-0.566817	0.0644978	1.50E-18
30	rs12656497	C	T	0.716061	0.0637776	3.00E-29
31	rs12677146	G	C	0.42607	0.065947	1.00E-10
32	rs12978472	G	C	-0.84682	0.0940538	2.20E-19
33	rs13107325	T	C	-0.768905	0.119024	1.00E-10
34	rs1320340	T	G	0.824129	0.148358	2.80E-08
35	rs13219548	T	C	0.405146	0.0630438	1.30E-10
36	rs13328893	T	C	-0.558092	0.088806	3.30E-10
37	rs1343040	A	G	0.431735	0.063705	1.20E-11
38	rs13436194	G	A	-0.457014	0.0632129	4.80E-13
39	rs1344653	G	A	0.348057	0.0625861	2.70E-08
40	rs147045545	G	A	-0.544935	0.0932389	5.10E-09
41	rs1543927	C	T	-0.4218	0.0711512	3.10E-09
42	rs1548391	G	A	-0.417502	0.0660732	2.60E-10
43	rs167479	T	G	-0.470301	0.0628261	7.10E-14
44	rs16952009	T	C	-0.365616	0.0658472	2.80E-08
45	rs1703982	T	A	-0.447104	0.0629233	1.20E-12
46	rs17117200	G	A	0.391185	0.0627883	4.70E-10
47	rs17713163	G	C	-0.957472	0.172387	2.80E-08
48	rs1801131	G	T	-0.555138	0.0674219	1.80E-16
49	rs1842896	T	G	0.401431	0.062567	1.40E-10
50	rs1864587	A	G	0.39065	0.0658407	3.00E-09
51	rs1939309	T	C	-0.36633	0.0637749	9.20E-09
52	rs1939310	A	G	0.379368	0.0628021	1.50E-09
53	rs1945211	T	A	0.545985	0.0955267	1.10E-08
54	rs1989803	G	C	0.399848	0.0667533	2.10E-09

55	rs2023843	T	C	0.862691	0.119811	6.00E-13
56	rs204883	A	G	0.360713	0.0644452	2.20E-08
57	rs2082450	G	A	0.582839	0.105292	3.10E-08
58	rs2107595	A	G	0.536691	0.0876715	9.30E-10
59	rs2188717	C	T	0.536128	0.0791637	1.30E-11
60	rs2423514	G	A	-0.370939	0.0628149	3.50E-09
61	rs2499801	A	G	-0.44416	0.0806357	3.60E-08
62	rs2524099	A	G	0.36444	0.0640551	1.30E-08
63	rs2607015	C	G	0.423217	0.0636473	2.90E-11
64	rs2610989	C	T	0.393514	0.0712999	3.40E-08
65	rs263016	C	T	-0.34751	0.0630696	3.60E-08
66	rs2643826	T	C	0.377984	0.0631442	2.20E-09
67	rs268263	A	T	0.565244	0.0734925	1.50E-14
68	rs2765524	T	C	-0.396517	0.0640623	6.00E-10
69	rs2854747	A	G	-0.419782	0.0636961	4.40E-11
70	rs2867114	T	C	-0.614906	0.10733	1.00E-08
71	rs3131007	T	A	0.367603	0.0634629	6.90E-09
72	rs34071855	G	C	0.370798	0.0663072	2.20E-08
73	rs34406901	G	A	0.74674	0.135852	3.90E-08
74	rs34710835	T	C	-0.52234	0.0642725	4.40E-16
75	rs34742161	T	C	0.453475	0.0823403	3.60E-08
76	rs35021474	G	C	-0.463939	0.0644971	6.30E-13
77	rs35444	G	A	-0.390865	0.0642817	1.20E-09
78	rs35726503	T	A	-0.449535	0.0634517	1.40E-12
79	rs360153	C	T	0.389126	0.0632899	7.80E-10
80	rs3790604	A	C	0.872402	0.120391	4.30E-13
81	rs3821843	A	G	0.412853	0.0681401	1.40E-09
82	rs3823483	C	T	0.354643	0.0634247	2.30E-08
83	rs3828591	C	G	-0.521162	0.0642105	4.80E-16
84	rs4480845	C	T	-0.406496	0.0656129	5.80E-10
85	rs448385	A	G	0.34644	0.0629656	3.80E-08
86	rs4753981	C	G	0.365393	0.0636624	9.50E-09
87	rs4766578	A	T	-0.406594	0.0626484	8.60E-11
88	rs5068	G	A	-1.32827	0.137364	4.10E-22
89	rs55840650	T	C	0.422052	0.0671703	3.30E-10
90	rs55925664	A	T	0.641451	0.0805291	1.60E-15
91	rs56137952	A	G	0.549446	0.0984794	2.40E-08
92	rs60289499	A	G	0.456364	0.070768	1.10E-10
93	rs620315	A	G	0.478634	0.0651849	2.10E-13
94	rs62481856	A	G	0.844443	0.078857	9.30E-27
95	rs633185	C	G	0.694507	0.0698276	2.60E-23
96	rs6825268	G	A	0.357322	0.0631452	1.50E-08
97	rs6982308	G	C	-0.422162	0.0627861	1.80E-11
98	rs7076938	T	C	0.460615	0.0711061	9.30E-11
99	rs7107356	G	A	0.386324	0.0625051	6.40E-10
100	rs7120737	G	A	-0.63164	0.0890932	1.30E-12
101	rs7136259	C	T	0.523423	0.0634848	1.70E-16
102	rs71373532	T	C	0.707799	0.11879	2.50E-09
103	rs72843938	A	G	-0.511856	0.0766282	2.40E-11
104	rs73007683	T	A	-0.448092	0.0800131	2.10E-08
105	rs73046792	A	G	-0.471445	0.0843393	2.30E-08
106	rs73073676	T	A	-0.377265	0.0670153	1.80E-08
107	rs73437338	C	T	-0.767786	0.0843419	8.80E-20
108	rs73563812	T	G	-0.421515	0.0737441	1.10E-08
109	rs7463212	A	T	-0.377575	0.0629451	2.00E-09
110	rs747249	G	A	-0.374676	0.065604	1.10E-08
111	rs74729242	C	T	0.59008	0.101342	5.80E-09
112	rs75230966	A	G	0.565429	0.0950299	2.70E-09
113	rs753012	C	T	-0.408732	0.0659309	5.70E-10

114	rs77870048	T	C	0.911699	0.139917	7.20E-11
115	rs7822500	C	T	-0.367521	0.0654572	2.00E-08
116	rs7835002	C	G	0.40281	0.0652194	6.60E-10
117	rs7909027	C	T	0.422909	0.0658512	1.30E-10
118	rs7922049	A	G	-0.553349	0.0869331	1.90E-10
119	rs7930107	G	A	-0.523191	0.0935296	2.20E-08
120	rs79780963	T	C	-1.06814	0.118415	1.90E-19
121	rs8039305	C	T	0.594252	0.0630046	4.00E-21
122	rs877116	T	G	-0.46967	0.0636099	1.50E-13
123	rs890431	C	T	0.616178	0.108373	1.30E-08
124	rs891511	A	G	-0.376501	0.0680582	3.20E-08
125	rs953246	A	T	0.392405	0.0685268	1.00E-08
126	rs9907379	C	T	0.425715	0.0769493	3.20E-08

Appendix 4 Table 7: List of SNPs used in polygenic score for systolic blood pressure in the sample 2 GWAS at the genome-wide significance level with a clumping threshold of 500kb and an R2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs10059884	A	C	0.585862	0.0642636	7.80E-20
2	rs1051006	A	G	-0.505227	0.0826819	9.90E-10
3	rs10783339	G	A	-0.389187	0.0714149	5.00E-08
4	rs10839472	C	T	-0.405148	0.0737475	3.90E-08
5	rs10849937	G	A	-0.529839	0.0781559	1.20E-11
6	rs10882412	C	T	-0.387459	0.0654848	3.30E-09
7	rs11014012	T	G	-0.387608	0.0640686	1.40E-09
8	rs11099097	T	C	0.637052	0.0695747	5.40E-20
9	rs11105358	G	C	0.442412	0.0644037	6.40E-12
10	rs11187793	A	G	0.429409	0.063869	1.80E-11
11	rs11187838	A	G	-0.561505	0.0637082	1.20E-18
12	rs11188220	T	C	0.473499	0.0863154	4.10E-08
13	rs11224417	C	A	0.365772	0.0633882	7.90E-09
14	rs11241959	G	A	0.371451	0.0631332	4.00E-09
15	rs11246667	A	G	0.482296	0.087972	4.20E-08
16	rs112873218	T	C	-0.636788	0.105091	1.40E-09
17	rs11646987	T	G	-0.392246	0.0709123	3.20E-08
18	rs11671314	C	G	0.581767	0.0973419	2.30E-09
19	rs117539635	G	A	-1.33873	0.200873	2.70E-11
20	rs117574138	G	C	0.58339	0.106127	3.90E-08
21	rs117754181	A	G	0.848858	0.13851	8.90E-10
22	rs11894064	T	A	-0.358367	0.0633502	1.50E-08
23	rs12046278	C	T	0.438565	0.0660808	3.20E-11
24	rs12053529	A	G	-0.47616	0.085436	2.50E-08
25	rs12130314	T	G	-0.414228	0.070268	3.70E-09
26	rs1216743	A	G	0.566196	0.0704896	9.60E-16
27	rs12229946	T	G	0.409993	0.0706661	6.60E-09
28	rs12258967	G	C	-0.589339	0.0690826	1.50E-17
29	rs12575654	A	G	0.473968	0.0729086	8.00E-11
30	rs12652819	G	A	-0.377534	0.0677462	2.50E-08
31	rs1275988	T	C	-0.483465	0.0649665	9.90E-14
32	rs12788272	A	C	0.930086	0.163013	1.20E-08
33	rs12946454	T	A	0.459599	0.0712814	1.10E-10
34	rs12951057	G	C	0.530415	0.0855947	5.80E-10
35	rs12978472	G	C	-0.70494	0.0947152	9.90E-14
36	rs142289341	A	T	0.737241	0.121777	1.40E-09
37	rs1436138	G	A	-0.361529	0.0660546	4.40E-08
38	rs167479	T	G	-0.574313	0.0630398	8.20E-20
39	rs17010961	A	T	0.577535	0.0912832	2.50E-10
40	rs17080069	G	A	-0.719939	0.121839	3.40E-09
41	rs17715065	T	C	0.366183	0.0631246	6.60E-09
42	rs2031323	T	C	0.418446	0.0652904	1.50E-10
43	rs2301597	C	T	-0.508119	0.0639747	2.00E-15
44	rs2307032	C	T	-0.387941	0.0664115	5.20E-09
45	rs2392929	G	T	0.716545	0.0789963	1.20E-19
46	rs2472299	G	A	-0.458136	0.0709387	1.10E-10
47	rs2681492	C	T	-0.676486	0.0839838	8.00E-16
48	rs268263	A	T	0.560215	0.0738018	3.20E-14
49	rs2760748	A	T	0.639762	0.105431	1.30E-09
50	rs2780072	T	A	0.503507	0.0902206	2.40E-08
51	rs2964330	T	G	0.364522	0.0641049	1.30E-08
52	rs3020644	G	A	0.359631	0.0655733	4.10E-08
53	rs34040136	A	G	1.08492	0.17975	1.60E-09
54	rs34477350	T	G	1.03609	0.172755	2.00E-09

55	rs35312823	T	C	1.13407	0.205392	3.40E-08
56	rs35427	G	T	-0.411027	0.0663981	6.00E-10
57	rs35838379	G	A	0.464904	0.0831492	2.30E-08
58	rs3740360	C	A	-0.700725	0.100635	3.30E-12
59	rs3792765	G	A	-0.553424	0.064696	1.20E-17
60	rs42032	A	G	-0.396654	0.0719921	3.60E-08
61	rs4425336	G	A	-0.453477	0.0770397	3.90E-09
62	rs4468343	T	C	0.589876	0.094465	4.30E-10
63	rs4480845	C	T	-0.402416	0.066164	1.20E-09
64	rs448798	G	A	0.371658	0.0636445	5.20E-09
65	rs4648815	A	G	-0.380689	0.0637656	2.40E-09
66	rs4690974	C	T	0.374146	0.0631476	3.10E-09
67	rs4767328	A	G	0.357508	0.0639895	2.30E-08
68	rs4883481	C	T	-0.421841	0.0652497	1.00E-10
69	rs4932373	C	A	0.514421	0.0672123	2.00E-14
70	rs4980379	T	C	0.58963	0.0656634	2.70E-19
71	rs5066	A	C	-0.707267	0.128245	3.50E-08
72	rs55714388	C	A	-0.676772	0.110871	1.00E-09
73	rs55857306	A	G	-0.76866	0.0850282	1.60E-19
74	rs57301765	A	G	0.505862	0.0865903	5.20E-09
75	rs59652089	T	C	-0.496269	0.0910536	5.00E-08
76	rs597808	G	A	-0.430999	0.0633734	1.00E-11
77	rs61572747	G	A	-0.403188	0.0720379	2.20E-08
78	rs61867141	A	G	0.553681	0.0807567	7.10E-12
79	rs61868776	T	C	-0.488816	0.0710773	6.10E-12
80	rs62049035	A	G	0.736701	0.128326	9.40E-09
81	rs6442260	A	G	-0.359497	0.0658933	4.90E-08
82	rs6504591	T	G	-0.352197	0.0641313	4.00E-08
83	rs6541328	G	A	0.555125	0.100955	3.80E-08
84	rs6657049	A	G	0.366463	0.0657613	2.50E-08
85	rs6668659	G	T	-0.49247	0.066692	1.50E-13
86	rs6724607	G	A	-0.365573	0.0629604	6.40E-09
87	rs6923947	A	G	0.49736	0.0635489	5.00E-15
88	rs7070797	A	G	-0.628952	0.0904607	3.60E-12
89	rs7102374	A	G	-0.372813	0.0674884	3.30E-08
90	rs7107356	G	A	0.479333	0.0630371	2.90E-14
91	rs7129056	A	G	0.403445	0.0631572	1.70E-10
92	rs7217916	G	A	-0.39437	0.0650295	1.30E-09
93	rs73143584	A	G	-0.561522	0.100825	2.60E-08
94	rs732998	C	T	-0.836528	0.118393	1.60E-12
95	rs74679637	A	G	0.929741	0.15789	3.90E-09
96	rs75615848	A	G	-0.757813	0.130765	6.80E-09
97	rs75777337	A	T	0.612514	0.103339	3.10E-09
98	rs7588932	T	C	-0.471728	0.0741433	2.00E-10
99	rs7600124	T	A	0.418679	0.0644378	8.20E-11
100	rs76443711	C	G	0.544241	0.0915247	2.70E-09
101	rs778121	C	T	0.388712	0.0658549	3.60E-09
102	rs77870048	T	C	0.921644	0.141505	7.40E-11
103	rs7889	G	C	0.379228	0.065705	7.80E-09
104	rs7908334	T	C	0.540497	0.0779396	4.10E-12
105	rs7918142	G	T	-0.359412	0.0639049	1.90E-08
106	rs7953257	T	A	-0.370313	0.0653828	1.50E-08
107	rs79553110	G	A	0.893733	0.155023	8.20E-09
108	rs79780963	T	C	-0.831243	0.118335	2.10E-12
109	rs80335285	G	A	0.598781	0.091777	6.80E-11
110	rs891511	A	G	-0.414076	0.0682559	1.30E-09
111	rs913220	G	C	0.43889	0.064947	1.40E-11
112	rs9898793	T	C	0.414299	0.0719183	8.40E-09

Appendix 4 Table 8: List of SNPs used in polygenic score for atrial fibrillation at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs10005432	A	G	0.1432	0.0111	4.76E-38
2	rs10109521	A	G	-0.0471	0.0071	3.35E-11
3	rs1011441	G	A	0.0939	0.0082	4.66E-30
4	rs1015864	C	T	0.0763	0.0122	4.19E-10
5	rs10165883	T	C	-0.0642	0.0072	5.83E-19
6	rs10213171	G	C	0.1041	0.0139	6.09E-14
7	rs10222783	T	C	-0.1406	0.0131	4.99E-27
8	rs1044258	C	T	-0.0463	0.0076	1.07E-09
9	rs10466138	C	T	-0.0534	0.0073	1.97E-13
10	rs10516564	G	A	0.0548	0.0079	4.74E-12
11	rs10520260	G	A	-0.0539	0.0079	8.98E-12
12	rs10753933	G	T	-0.0743	0.0072	5.83E-25
13	rs10760361	T	G	-0.0434	0.0075	7.03E-09
14	rs10786662	C	G	-0.0403	0.0072	1.99E-08
15	rs10800507	G	C	-0.0814	0.0072	7.01E-30
16	rs10800898	G	A	-0.0404	0.0073	2.67E-08
17	rs10822152	G	A	0.0565	0.0098	8.90E-09
18	rs10842383	T	C	-0.1088	0.0104	1.02E-25
19	rs10873299	G	A	-0.0483	0.0075	9.62E-11
20	rs10883913	T	C	-0.0626	0.0073	7.27E-18
21	rs10919364	C	T	-0.0587	0.0102	7.70E-09
22	rs11001667	G	A	0.0619	0.0091	1.06E-11
23	rs11075959	G	A	0.1397	0.0216	1.03E-10
24	rs111233078	A	G	-0.0426	0.0074	8.12E-09
25	rs111621680	T	C	0.0754	0.011	6.15E-12
26	rs11180703	A	G	-0.0457	0.0073	3.58E-10
27	rs11191801	C	A	0.0533	0.0077	4.89E-12
28	rs112156684	C	T	0.2125	0.0241	1.07E-18
29	rs112453500	A	G	0.228	0.0384	2.82E-09
30	rs112515238	T	C	0.1065	0.0172	5.77E-10
31	rs112599895	G	A	0.643	0.0361	6.57E-71
32	rs11264280	T	C	0.127	0.0078	4.60E-59
33	rs113378881	A	G	0.1314	0.0135	2.75E-22
34	rs11353611	A	T	0.1512	0.0274	3.49E-08
35	rs113640213	A	G	-0.2895	0.0467	5.72E-10
36	rs113654447	T	C	-0.0564	0.0078	6.40E-13
37	rs113819537	G	C	-0.049	0.0082	2.23E-09
38	rs113832645	A	G	-0.2785	0.02	3.58E-44
39	rs114014056	C	T	-0.1957	0.0303	1.02E-10
40	rs114691030	C	G	0.2167	0.0227	1.39E-21
41	rs11598047	G	A	0.1533	0.0095	4.83E-58
42	rs116202356	A	G	0.1971	0.0323	1.01E-09
43	rs11641227	A	G	0.0449	0.0074	1.16E-09
44	rs116455344	A	G	-0.1153	0.0187	6.59E-10
45	rs11717092	G	A	0.0465	0.0073	2.33E-10
46	rs11768850	T	C	0.0392	0.0072	4.96E-08
47	rs11773845	A	C	0.1162	0.0072	4.61E-58
48	rs117984853	T	G	0.1132	0.0136	8.38E-17
49	rs11180286	C	T	-0.0682	0.008	2.29E-17
50	rs11814244	T	G	0.0432	0.0072	2.57E-09
51	rs11835327	G	A	0.0681	0.0122	2.14E-08
52	rs11848040	T	G	0.047	0.0086	4.76E-08
53	rs12044963	T	G	0.0795	0.0113	1.61E-12
54	rs12046897	G	A	-0.0676	0.0102	3.32E-11

55	rs12121494	A	G	0.2056	0.0367	2.11E-08
56	rs12122060	A	T	0.1373	0.0112	2.73E-34
57	rs12131638	G	A	0.1082	0.0178	1.27E-09
58	rs12142379	C	T	0.1058	0.0156	1.32E-11
59	rs1218574	G	A	0.0718	0.0109	5.18E-11
60	rs1218577	C	T	0.0522	0.0075	3.60E-12
61	rs1218578	G	A	-0.0406	0.0072	1.83E-08
62	rs1218598	A	G	-0.0591	0.0088	1.78E-11
63	rs12189392	T	A	0.0581	0.0095	9.54E-10
64	rs12208899	A	G	0.049	0.0087	1.95E-08
65	rs1229741	A	G	0.0459	0.0076	1.71E-09
66	rs12298484	T	C	-0.0455	0.0076	2.05E-09
67	rs12325558	C	A	0.0583	0.0071	2.90E-16
68	rs12360357	T	C	-0.1093	0.0088	1.04E-35
69	rs12589834	G	A	0.0639	0.0089	7.66E-13
70	rs12591736	A	G	-0.0606	0.0102	2.47E-09
71	rs12647973	T	G	-0.0732	0.0093	3.62E-15
72	rs12649917	A	G	0.0926	0.0119	6.11E-15
73	rs12730906	T	C	0.0743	0.0104	1.06E-12
74	rs12809354	C	T	0.081	0.01	5.48E-16
75	rs12810346	T	C	0.0658	0.011	2.34E-09
76	rs12812948	G	A	-0.0509	0.0084	1.20E-09
77	rs12908004	G	A	0.0753	0.0098	1.95E-14
78	rs12908437	C	T	-0.0468	0.0073	1.25E-10
79	rs12992412	T	A	0.0406	0.0073	2.30E-08
80	rs13061421	A	G	0.0476	0.0081	4.48E-09
81	rs1307274	G	T	-0.0741	0.0135	3.85E-08
82	rs13105878	A	C	-0.1787	0.0136	1.47E-39
83	rs13121747	A	G	-0.0626	0.0087	5.61E-13
84	rs13126426	C	T	0.1875	0.0186	6.56E-24
85	rs13191450	C	A	-0.0704	0.0075	8.92E-21
86	rs13242816	T	C	-0.1213	0.0126	8.53E-22
87	rs13334473	C	A	0.0561	0.0089	2.83E-10
88	rs139811148	A	G	-0.0627	0.0108	6.01E-09
89	rs140185678	A	G	0.1813	0.0241	5.57E-14
90	rs141221125	A	G	0.2181	0.0289	4.55E-14
91	rs141752220	A	G	0.4067	0.0291	1.86E-44
92	rs142822330	C	T	0.12	0.0215	2.29E-08
93	rs1443926	G	A	-0.0482	0.008	1.94E-09
94	rs1448813	C	T	-0.0471	0.0076	4.63E-10
95	rs145538762	C	T	0.1041	0.0179	5.84E-09
96	rs146269981	A	G	0.3064	0.0398	1.44E-14
97	rs146518726	A	G	0.1617	0.0254	2.05E-10
98	rs1470618	T	C	0.1357	0.0094	7.12E-47
99	rs147352248	C	T	-0.3097	0.0429	5.53E-13
100	rs151107921	A	G	-0.083	0.0129	1.32E-10
101	rs1538575	T	A	-0.0473	0.0071	2.86E-11
102	rs1562641	G	A	-0.0604	0.0094	1.17E-10
103	rs168367	C	T	-0.1108	0.018	7.38E-10
104	rs17041835	G	A	-0.1209	0.0212	1.29E-08
105	rs17042059	A	G	0.4252	0.0097	0
106	rs17079881	G	A	0.0851	0.0105	4.23E-16
107	rs17341992	C	T	-0.0502	0.0083	1.51E-09
108	rs174048	C	T	0.0665	0.0098	1.05E-11
109	rs17490701	A	G	-0.07	0.0107	5.43E-11
110	rs17507821	C	T	0.0459	0.0078	4.67E-09
111	rs17513625	A	G	0.464	0.0249	9.55E-78
112	rs17513772	A	T	-0.1568	0.0171	4.96E-20
113	rs17513814	T	C	0.1905	0.0242	3.55E-15

114	rs17552555	C	T	0.18	0.0218	1.45E-16
115	rs17662087	G	A	-0.1039	0.0132	2.90E-15
116	rs17746631	G	A	-0.1718	0.0183	4.98E-21
117	rs1822273	A	G	-0.0683	0.0082	8.99E-17
118	rs1866961	C	T	0.0572	0.0074	9.80E-15
119	rs187311	G	A	-0.1388	0.0191	3.35E-13
120	rs1896002	C	A	0.0582	0.0071	2.69E-16
121	rs192667187	C	T	0.3073	0.0466	4.08E-11
122	rs1963560	T	C	0.0599	0.0106	1.36E-08
123	rs2047036	T	C	-0.0549	0.0075	2.80E-13
124	rs2072412	G	C	0.0515	0.0084	7.43E-10
125	rs2073341	A	G	-0.0421	0.0073	9.60E-09
126	rs2145274	C	A	-0.1015	0.0141	6.97E-13
127	rs2145587	A	G	0.0754	0.0079	2.32E-21
128	rs214575	T	C	-0.0539	0.0073	2.23E-13
129	rs2216553	T	C	-0.0489	0.0076	1.14E-10
130	rs2240331	A	C	0.0499	0.0072	4.87E-12
131	rs2286466	G	A	0.0718	0.0095	3.53E-14
132	rs2291437	G	T	0.0786	0.0105	7.53E-14
133	rs2306272	C	T	0.0512	0.0078	4.54E-11
134	rs2359171	A	T	0.1884	0.0089	2.94E-100
135	rs242557	A	G	-0.0439	0.0075	4.35E-09
136	rs2500549	T	C	0.0498	0.0081	7.50E-10
137	rs2540949	T	A	-0.0752	0.0073	8.17E-25
138	rs2595110	G	A	-0.0725	0.0083	3.77E-18
139	rs2604195	T	C	-0.0581	0.0082	1.39E-12
140	rs2660824	T	C	0.0403	0.0071	1.47E-08
141	rs2723307	T	A	-0.1577	0.0075	3.33E-99
142	rs2738413	G	A	-0.0807	0.0072	1.81E-29
143	rs2739200	C	G	-0.1711	0.0076	4.24E-112
144	rs2810915	T	G	-0.0536	0.0078	7.33E-12
145	rs2834618	G	T	-0.1096	0.0126	2.93E-18
146	rs28436726	A	G	0.0977	0.0157	4.48E-10
147	rs28488916	A	G	-0.0488	0.0089	4.23E-08
148	rs28587043	A	G	-0.0929	0.0076	8.77E-35
149	rs28601812	C	A	-0.0901	0.0083	1.24E-27
150	rs28631169	T	C	0.07	0.0093	3.80E-14
151	rs2894040	C	A	0.0452	0.0074	8.62E-10
152	rs295114	T	C	-0.0676	0.0073	1.76E-20
153	rs2984131	T	C	-0.1023	0.012	1.59E-17
154	rs2986036	C	T	0.0419	0.0072	5.36E-09
155	rs3014204	A	C	0.0571	0.01	1.33E-08
156	rs3112133	G	A	-0.0442	0.0077	8.52E-09
157	rs3176326	A	G	-0.0599	0.0092	7.95E-11
158	rs34195153	G	C	0.2303	0.0393	4.62E-09
159	rs34515871	T	C	0.1253	0.008	1.19E-55
160	rs34750263	T	C	0.0873	0.0076	2.89E-30
161	rs34969716	A	G	0.0875	0.0084	2.91E-25
162	rs35006907	A	C	0.0454	0.0076	2.76E-09
163	rs35056927	A	G	-0.0826	0.0131	3.18E-10
164	rs35176054	A	T	0.1458	0.0109	8.47E-41
165	rs35349325	C	T	-0.0524	0.0073	9.04E-13
166	rs35504893	T	C	0.09	0.0087	6.89E-25
167	rs3761834	A	G	-0.047	0.0075	3.49E-10
168	rs3769081	C	T	0.0414	0.0075	4.16E-08
169	rs3731640	A	G	-0.0667	0.0115	6.32E-09
170	rs374582	A	G	-0.0938	0.0076	9.66E-35
171	rs3781339	T	C	-0.0834	0.009	2.56E-20
172	rs3784193	A	T	0.0678	0.0082	1.64E-16

173	rs3796097	C	T	-0.0404	0.0073	3.01E-08
174	rs3796903	C	T	0.0911	0.0132	4.79E-12
175	rs3822259	T	G	0.0463	0.0077	1.93E-09
176	rs3849045	C	T	-0.0393	0.0072	4.34E-08
177	rs3853444	C	T	-0.0706	0.0086	1.99E-16
178	rs3855819	G	C	-0.141	0.0098	3.07E-47
179	rs3922844	C	T	-0.0478	0.0078	6.84E-10
180	rs3925798	T	C	-0.0461	0.0073	2.80E-10
181	rs3951016	A	T	0.0543	0.0072	4.62E-14
182	rs396024	G	C	0.0662	0.0118	1.85E-08
183	rs3960788	C	T	0.0507	0.0072	2.09E-12
184	rs3968564	A	G	-0.05	0.0091	3.58E-08
185	rs4115273	A	C	0.1766	0.0082	1.42E-103
186	rs4124174	T	G	0.1048	0.0081	6.01E-38
187	rs41264253	A	G	0.1138	0.0125	8.17E-20
188	rs412768	G	A	0.0467	0.0078	2.56E-09
189	rs41298968	T	C	-0.0603	0.0096	3.18E-10
190	rs41312411	G	C	-0.0605	0.0105	8.32E-09
191	rs4146379	C	T	0.0505	0.0072	2.39E-12
192	rs42874	C	T	-0.0485	0.0074	4.81E-11
193	rs438258	T	A	-0.053	0.0081	5.88E-11
194	rs4385527	A	G	0.092	0.0073	2.26E-36
195	rs4401702	A	G	-0.0522	0.0091	1.14E-08
196	rs4414093	A	C	-0.0466	0.0074	3.05E-10
197	rs4484922	C	G	-0.063	0.0078	4.57E-16
198	rs449333	G	C	-0.0716	0.0077	1.38E-20
199	rs4607376	G	A	0.0397	0.0073	4.64E-08
200	rs4656215	T	C	0.0656	0.0084	4.45E-15
201	rs4656754	A	G	0.1603	0.028	1.04E-08
202	rs4656794	A	G	-0.0795	0.0075	2.20E-26
203	rs4673891	G	C	0.0543	0.0079	7.45E-12
204	rs4743034	A	G	0.049	0.0083	3.98E-09
205	rs4744374	A	G	-0.0539	0.008	1.70E-11
206	rs478454	C	T	-0.0633	0.0072	1.96E-18
207	rs4788489	G	T	0.0413	0.0071	6.50E-09
208	rs4788490	C	G	0.1002	0.0079	4.13E-37
209	rs4788697	A	G	0.0785	0.0078	3.96E-24
210	rs480667	A	C	-0.051	0.0081	2.51E-10
211	rs4845703	T	C	-0.0447	0.008	2.02E-08
212	rs4855075	T	C	0.0604	0.0103	4.00E-09
213	rs4871397	C	G	-0.1018	0.0145	1.95E-12
214	rs4951261	C	A	0.0441	0.0072	1.17E-09
215	rs4977397	G	A	-0.0432	0.0075	8.60E-09
216	rs4981979	T	C	0.0459	0.0078	3.16E-09
217	rs4986938	T	C	-0.0516	0.0075	6.25E-12
218	rs4999127	A	G	0.0891	0.011	4.77E-16
219	rs514739	G	A	0.1231	0.0217	1.35E-08
220	rs524788	C	T	-0.1592	0.0193	1.59E-16
221	rs532748	A	T	-0.0589	0.0094	4.11E-10
222	rs55734480	A	G	0.0504	0.0082	7.34E-10
223	rs55754224	T	C	0.0477	0.0083	9.25E-09
224	rs55947985	T	C	0.0728	0.0107	9.76E-12
225	rs55985730	G	T	0.0957	0.017	1.81E-08
226	rs56103503	T	C	-0.0823	0.0078	6.31E-26
227	rs56181519	T	C	-0.0778	0.0086	1.52E-19
228	rs56305400	T	C	0.0539	0.0083	8.37E-11
229	rs56308529	G	C	0.1198	0.018	3.01E-11
230	rs58847541	A	G	0.054	0.0099	4.63E-08
231	rs591715	A	G	-0.076	0.0089	1.26E-17

232	rs60029182	T	G	-0.0671	0.0104	1.07E-10
233	rs60050852	A	G	-0.0543	0.0094	6.66E-09
234	rs60212594	C	G	-0.1097	0.0102	6.48E-27
235	rs608930	T	G	-0.0968	0.0071	1.94E-42
236	rs61150523	G	A	0.072	0.0117	6.90E-10
237	rs61826205	G	T	-0.081	0.0141	8.70E-09
238	rs62011291	G	A	0.0519	0.0089	6.14E-09
239	rs62055086	T	C	-0.0657	0.0088	6.92E-14
240	rs62059797	G	A	-0.0529	0.0086	8.72E-10
241	rs62337205	G	A	-0.2505	0.0211	2.13E-32
242	rs62337249	G	A	-0.1193	0.0107	8.25E-29
243	rs62380877	A	G	-0.0565	0.0096	4.14E-09
244	rs62483627	A	G	0.0489	0.0084	5.17E-09
245	rs62521286	G	A	0.1224	0.0148	1.24E-16
246	rs634851	T	C	0.0636	0.0096	3.28E-11
247	rs6427245	T	C	0.0531	0.0072	2.20E-13
248	rs6462078	A	C	0.058	0.0086	1.35E-11
249	rs6499606	C	T	0.0781	0.0074	4.13E-26
250	rs6546620	C	T	0.0708	0.0093	2.96E-14
251	rs6553712	A	T	-0.14	0.0225	5.25E-10
252	rs6661079	C	T	-0.053	0.0071	9.75E-14
253	rs6680785	T	C	0.0505	0.0073	4.51E-12
254	rs6701640	C	A	-0.0733	0.0095	9.36E-15
255	rs6708345	C	A	0.0456	0.0072	2.20E-10
256	rs6742276	A	G	0.0485	0.0073	2.42E-11
257	rs6790396	G	C	0.0636	0.0073	4.13E-18
258	rs6793245	A	G	-0.0413	0.0076	4.57E-08
259	rs6810325	C	T	0.0747	0.0076	5.24E-23
260	rs6823804	A	G	-0.0817	0.0099	1.48E-16
261	rs6838973	T	C	-0.1842	0.0072	1.35E-142
262	rs6882776	A	G	-0.06	0.0079	3.18E-14
263	rs6907805	T	G	-0.0405	0.0071	1.10E-08
264	rs6907980	G	A	0.0402	0.0072	2.41E-08
265	rs6931433	G	C	0.0523	0.009	7.53E-09
266	rs6993266	A	G	0.0443	0.0072	9.73E-10
267	rs700607	C	T	-0.0631	0.0088	7.70E-13
268	rs700613	C	A	-0.0462	0.0082	1.65E-08
269	rs7067666	T	C	0.0955	0.0071	2.57E-41
270	rs710768	A	T	-0.043	0.0079	4.71E-08
271	rs71419908	A	C	-0.0469	0.0072	7.22E-11
272	rs71424150	C	T	-0.0792	0.0137	6.73E-09
273	rs71628635	C	A	0.1615	0.015	7.27E-27
274	rs716845	A	G	0.0594	0.008	1.16E-13
275	rs7219869	G	C	0.046	0.0072	1.49E-10
276	rs721994	G	A	0.0724	0.008	9.65E-20
277	rs723363	C	T	-0.2414	0.0075	2.89E-229
278	rs723840	T	C	0.0444	0.0073	1.23E-09
279	rs72666200	C	T	0.0854	0.0104	2.47E-16
280	rs72667931	T	C	0.1546	0.0147	5.61E-26
281	rs72674110	T	G	0.088	0.0105	5.03E-17
282	rs72690464	G	T	0.1375	0.0226	1.12E-09
283	rs7269123	T	C	-0.0443	0.0076	5.59E-09
284	rs72700114	C	G	0.2026	0.0139	7.32E-48
285	rs72702041	T	C	0.0608	0.0109	2.30E-08
286	rs72712048	A	G	0.1288	0.0209	7.52E-10
287	rs72715944	A	G	0.1463	0.0193	3.91E-14
288	rs72802815	G	A	0.0495	0.0085	6.48E-09
289	rs72811294	C	G	-0.0667	0.0115	6.87E-09
290	rs728713	G	A	-0.0969	0.0137	1.42E-12

291	rs72926475	A	G	-0.0708	0.0113	3.49E-10
292	rs73032363	G	A	-0.0432	0.0078	3.59E-08
293	rs73241997	T	C	0.072	0.0097	1.10E-13
294	rs73366713	A	G	-0.1052	0.0112	5.80E-21
295	rs7349311	A	G	-0.0758	0.0097	4.80E-15
296	rs73666807	T	C	0.1083	0.019	1.29E-08
297	rs7373065	C	T	-0.2151	0.0287	6.50E-14
298	rs74022964	T	C	0.1059	0.0097	1.27E-27
299	rs7460121	A	G	0.0708	0.0125	1.65E-08
300	rs74910854	G	A	0.0942	0.0159	3.36E-09
301	rs7502669	G	A	-0.0417	0.0075	3.21E-08
302	rs7508	A	G	0.072	0.008	2.22E-19
303	rs7514023	T	C	0.1654	0.0293	1.74E-08
304	rs7526113	A	G	-0.0718	0.0116	7.15E-10
305	rs7549338	G	C	-0.0454	0.0071	1.71E-10
306	rs75577686	G	T	0.1239	0.0203	9.52E-10
307	rs76097649	A	G	0.1264	0.0137	2.19E-20
308	rs76306191	G	C	0.058	0.0094	6.19E-10
309	rs7632427	C	T	-0.0425	0.0074	1.10E-08
310	rs76774446	A	C	0.0633	0.0111	1.13E-08
311	rs76872986	T	C	-0.1651	0.027	9.94E-10
312	rs77316573	T	C	0.0528	0.0093	1.48E-08
313	rs7755375	T	C	0.0406	0.0071	1.03E-08
314	rs77668866	T	C	-0.1314	0.0146	2.82E-19
315	rs7789146	A	G	-0.0571	0.0092	6.51E-10
316	rs77953709	T	C	0.1728	0.0201	9.49E-18
317	rs77955149	G	C	0.0822	0.0093	1.05E-18
318	rs78053786	A	G	0.1085	0.0148	2.44E-13
319	rs7835298	A	G	0.0392	0.0072	4.49E-08
320	rs7846485	A	C	-0.0872	0.0111	3.71E-15
321	rs78710246	T	A	-0.0875	0.0143	1.07E-09
322	rs79187193	A	G	-0.1116	0.0182	8.07E-10
323	rs7919685	T	G	-0.0579	0.0071	5.00E-16
324	rs7953024	G	C	0.0626	0.0091	6.12E-12
325	rs7966951	G	A	0.044	0.008	3.31E-08
326	rs7978685	C	T	-0.0547	0.0079	5.99E-12
327	rs8005490	C	T	-0.0465	0.0073	1.94E-10
328	rs80141833	G	A	0.2211	0.013	1.71E-64
329	rs8073937	A	G	-0.0504	0.0074	1.02E-11
330	rs876727	G	T	-0.0905	0.0089	4.69E-24
331	rs880315	C	T	0.0437	0.0075	5.04E-09
332	rs883079	T	C	0.1196	0.0079	1.26E-51
333	rs926198	T	C	0.0533	0.0079	1.20E-11
334	rs9414802	C	T	-0.0534	0.009	3.18E-09
335	rs9428207	G	A	0.0474	0.0078	1.01E-09
336	rs9481825	A	G	0.0639	0.0092	4.54E-12
337	rs949078	T	C	-0.0534	0.0081	4.77E-11
338	rs9580438	C	T	0.0568	0.0076	1.01E-13
339	rs9669457	G	A	0.0472	0.0085	2.52E-08
340	rs9920	C	T	-0.1075	0.0128	3.99E-17
341	rs9953366	C	T	0.0504	0.0078	9.03E-11

Appendix 4 Table 9: List of SNPs used in polygenic score for coronary heart disease at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs10139550	G	C	0.05538	0.0097569	1.38E-08
2	rs10174652	G	A	0.079086	0.0144747	4.66E-08
3	rs10840293	A	G	0.054714	0.009619	1.28E-08
4	rs11065979	T	C	0.068556	0.0107672	1.93E-10
5	rs11066188	A	G	0.063162	0.0108943	6.72E-09
6	rs11191416	G	T	-0.079249	0.0135252	4.65E-09
7	rs11206510	C	T	-0.074519	0.0133438	2.34E-08
8	rs11556924	T	C	-0.072569	0.0110605	5.34E-11
9	rs115654617	A	C	0.137846	0.0158314	3.12E-18
10	rs11617955	A	T	-0.088766	0.0161041	3.55E-08
11	rs11790231	A	G	0.118907	0.0163887	4.00E-13
12	rs11838776	A	G	0.068566	0.0107552	1.83E-10
13	rs12202017	G	A	-0.066813	0.0099612	1.98E-11
14	rs12202891	T	C	0.076537	0.0133417	9.65E-09
15	rs13209002	T	C	0.105577	0.0161942	7.06E-11
16	rs1333050	T	C	0.140772	0.010438	1.88E-41
17	rs1412444	T	C	0.066812	0.0096809	5.15E-12
18	rs16986953	A	G	0.08516	0.0150265	1.45E-08
19	rs17087335	T	G	0.060764	0.0111159	4.59E-08
20	rs1746050	A	C	-0.092397	0.0128431	6.28E-13
21	rs17678683	G	T	0.098786	0.0166548	3.00E-09
22	rs180803	T	G	-0.180923	0.0283062	1.64E-10
23	rs1833024	A	G	0.08283	0.0149787	3.20E-08
24	rs1855185	G	T	0.13824	0.0248682	2.71E-08
25	rs186696265	T	C	0.550351	0.0481949	3.35E-30
26	rs1870634	G	T	0.075878	0.0097113	5.55E-15
27	rs2107595	A	G	0.073415	0.0112951	8.05E-11
28	rs2128739	C	A	-0.065565	0.0100568	7.05E-11
29	rs2487928	A	G	0.062633	0.0095049	4.41E-11
30	rs2519093	T	C	0.079704	0.0117524	1.19E-11
31	rs2681472	G	A	0.074114	0.0113331	6.17E-11
32	rs28451064	A	G	0.127571	0.015952	1.33E-15
33	rs2891168	G	A	0.193401	0.0091877	2.29E-98
34	rs3120147	T	C	0.077862	0.0136132	1.07E-08
35	rs36049381	A	G	-0.083549	0.0140684	2.87E-09
36	rs3731249	T	C	0.171038	0.0301757	1.44E-08
37	rs3743058	T	C	0.06925	0.0101317	8.20E-12
38	rs3918226	T	C	0.133315	0.0221275	1.69E-09
39	rs4420638	G	A	0.091906	0.0140977	7.07E-11
40	rs4468572	C	T	0.077234	0.0095277	4.44E-16
41	rs4593108	G	C	-0.07083	0.0115558	8.82E-10
42	rs4773141	G	C	0.069732	0.0116482	2.14E-09
43	rs515135	C	T	0.067499	0.0121924	3.09E-08
44	rs55730499	T	C	0.316641	0.0242403	5.39E-39
45	rs56031815	A	G	0.075204	0.0127681	3.86E-09
46	rs56062135	T	C	-0.069743	0.0118937	4.52E-09
47	rs56289821	A	G	-0.13361	0.0170415	4.44E-15
48	rs56336142	C	T	-0.066813	0.0118763	1.85E-08
49	rs61271866	A	T	-0.112191	0.0107857	2.43E-25
50	rs624249	A	C	-0.061265	0.0105717	6.82E-09
51	rs6511721	A	G	-0.061674	0.0111409	3.10E-08
52	rs663129	A	G	0.058163	0.0105173	3.20E-08
53	rs66478960	A	G	-0.124511	0.0129191	5.54E-22
54	rs6689306	G	A	-0.056012	0.0094061	2.60E-09

55	rs67180937	G	T	0.078807	0.0110551	1.01E-12
56	rs6905073	T	G	-0.058185	0.0096896	1.91E-09
57	rs7041637	A	C	0.099171	0.0103044	6.33E-22
58	rs7212798	C	T	0.079961	0.0142216	1.88E-08
59	rs72652411	T	G	0.131271	0.0239959	4.49E-08
60	rs72652478	G	C	0.202956	0.0354724	1.06E-08
61	rs72689147	T	G	-0.068558	0.0117905	6.07E-09
62	rs73013166	C	T	-0.131384	0.0234973	2.25E-08
63	rs73015007	A	G	-0.08328	0.0116884	1.04E-12
64	rs7412	T	C	-0.137045	0.0210923	8.17E-11
65	rs74923585	G	A	0.167726	0.0277262	1.45E-09
66	rs7528419	G	A	-0.11453	0.011482	1.97E-23
67	rs7568458	A	T	0.059618	0.0095093	3.62E-10
68	rs762158	G	C	0.078	0.0114591	9.98E-12
69	rs8042271	A	G	-0.096711	0.0175662	3.68E-08
70	rs9349379	G	A	0.131836	0.0096527	1.81E-42
71	rs9457861	T	C	0.097078	0.0177867	4.82E-08
72	rs9457995	G	A	0.068647	0.0101947	1.66E-11
73	rs9515203	C	T	-0.071146	0.0116243	9.33E-10
74	rs9804352	G	A	0.058132	0.0092898	3.91E-10
75	rs9970807	T	C	-0.12575	0.016695	5.00E-14

Appendix 4 Table 10: List of SNPs used in polygenic score for Type 2 diabetes at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs10954284	A	T	-0.0953102	0.01616572	1.20E-08
2	rs11196175	C	T	0.26236426	0.02543351	4.80E-24
3	rs11196212	C	T	0.10436002	0.01602064	1.90E-10
4	rs11709077	A	G	-0.1570037	0.02618728	1.10E-09
5	rs12110493	G	A	0.23111172	0.04025689	2.20E-08
6	rs12266632	G	C	0.25464222	0.03932415	8.50E-11
7	rs1801214	T	C	0.11332869	0.01822933	1.30E-08
8	rs2383208	G	A	-0.1655144	0.02369738	6.00E-14
9	rs3802177	A	G	-0.14842	0.02410481	2.10E-11
10	rs3915932	C	G	-0.0953102	0.01856106	4.70E-08
11	rs4506565	T	A	0.31481074	0.01862886	4.90E-68
12	rs5015480	T	C	-0.1397619	0.01775339	2.20E-16
13	rs7651090	G	A	0.12221763	0.01806787	2.00E-11
14	rs7901275	C	A	0.14842001	0.01760022	3.90E-18
15	rs7933855	A	G	0.11332869	0.02041906	1.30E-09
16	rs864745	C	T	-0.1133287	0.01602064	7.30E-11
17	rs9368222	A	C	0.19062036	0.01906213	4.80E-23
18	rs9936385	C	T	0.12221763	0.02023896	4.70E-11

Appendix 4 Table 11: List of SNPs used in polygenic score for stroke at the genome-wide significance level with a clumping threshold of 500kb and an R^2 threshold of 0.25

	SNP (RSID)	Effect allele	Other allele	Beta	Standard error	P value
1	rs1052053	G	A	-0.0675	0.0096	2.25E-12
2	rs10774624	A	G	-0.0654	0.0094	4.04E-12
3	rs11066283	G	A	0.0692	0.0104	2.36E-11
4	rs11242678	T	C	0.0643	0.0105	8.71E-10
5	rs1537375	C	T	0.0519	0.0091	1.24E-08
6	rs2107595	A	G	0.0803	0.0121	3.59E-11
7	rs2634074	A	T	-0.084	0.0112	6.56E-14
8	rs475937	C	A	-0.0757	0.0137	2.92E-08
9	rs4942561	T	G	0.064	0.0107	2.05E-09
10	rs76110445	C	T	0.0814	0.0147	2.94E-08
11	rs847892	A	G	-0.054	0.0098	3.28E-08