Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of this thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Yunfeng Huang	Date

Gene-lifestyle Interactions in Coronary Artery Diseases

By

Yunfeng Huang
Doctor of Philosophy

Epidemiology

Yan V. Sun, PhD
Advisor

Viola Vaccarino, MD, PhD
Advisor

Arshed A. Quyyumi, MD, FACC, FRCP
Committee Member

Marta Gwinn, MD, MPH
Committee Member

Yijuan Hu, PhD
Committee Member

Accepted:

Lisa A. Tedesco, PhD
Dean of the James T. Laney School of Graduate Studies

Date

Gene-lifestyle Interactions in Coronary Artery Diseases

By

Yunfeng Huang
B.S., Fudan University, 2013
M.P.H, Emory University, 2015

Advisors:
Yan V. Sun, PhD
Viola Vaccarino, MD, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology

2019

Abstract

Gene-lifestyle Interactions in Coronary Artery Diseases

By Yunfeng Huang

Coronary artery disease (CAD) is the pre-eminent cause of death. Both genetic and lifestyle factors such as cigarette smoking and physical activity (PA) contribute to development of CAD. Over 160 loci have been linked with risk of CAD in genome-wide association studies. However, the interaction between genetic predisposition and individual lifestyle factors in CAD remains unclear. This dissertation presents research focused on exploring gene-lifestyle interactions for CAD among populations of European ancestry using data from two of the largest biobank cohorts. Multiple cardio-metabolic traits mediate the genetic effects of CAD, so this dissertation also aims to characterize the gene-lifestyle interaction driven by different mediating traits. In addition, gene-lifestyle interactions can be an important part of CAD heritability and accounting for gene-lifestyle interactions can potentially increase the power when detecting CAD-associated loci. Results of this dissertation have shown that the absolute risk elevation in CAD due to smoking is stronger among those with higher genetic susceptibility and the interaction can be driven by different mediating cardio-metabolic traits when different domains of smoking behavior is considered. Also, no evidence of interaction was identified between genetic predisposition and physical activity for CAD. Two GWASs of CAD accounting for gene-smoking interaction and gene-physical activity interaction found no novel loci, and results have shown no gain of power when a joint two degree of freedom approach was implemented. Future studies should consider exploring gene-lifestyle interactions for complex diseases such as CAD on both additive and multiplicative scale considering potential different mediating pathways. Novel methods should be developed to better incorporate gene-lifestyle interactions in genetic associations of complex diseases.

Gene-lifestyle Interactions in Coronary Artery Diseases

By

Yunfeng Huang
B.S., Fudan University, 2013
M.P.H, Emory University, 2015

Advisors:
Yan V. Sun, PhD
Viola Vaccarino, MD, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology

2019

Acknowledgements

I would like to strongly acknowledge Dr. Yan V. Sun, my mentor and chair, for all of his guidance and support through my dissertation and entire graduate program. His kindness, patience and insightful opinions have helped me grown from a student into a well-trained researcher in the field of genetic epidemiology.

I would also like to acknowledge my co-chair Dr. Viola Vaccarino, and all of my committee members: Dr. Arshed A. Quyyumi, Dr. Marta Gwinn and Dr. Yijuan Hu. Their time and guidance have been greatly appreciated.

I would like to acknowledge Dr. Qin Hui, without whom this research would not have been possible. He has provided extensive support in the data analysis of the Million Veteran Program. In addition, I would like to acknowledge all participants, collaborators, and staff of the UK Biobank and the Million Veteran Program, for their time and effort in making such great data resources available.

I would like to give special thanks to the M2M program and Dr. Nael A. McCarty, Dr. Julie A. Gazmararian, Dr. Stephanie L. Sherman, Dr. Brad Pearce, and Ms. Teresa R. Parker, for their generous guidance and support.

Finally, I would like to acknowledge the Department of Epidemiology and all faculty and staff, for their kindness, help and guidance. Also, I would give special thanks to my fellow students for their encouragement and support.

Table of Contents

Tables and Figures

Chapter 2.

Chapter 3.

Chapter 1

Introduction to gene-lifestyle interactions in coronary artery diseases

Coronary artery disease (CAD) is the most common type of cardiovascular disease, it involves narrowing or blockage of the coronary arteries (arteries that supply blood to the heart). (Parmet, Glass et al. 2004) CAD is usually caused by atherosclerosis, which is a process of buildup of plaque (deposits made up of cholesterol, other fats, and calcium) on the inner walls of the arteries. When a plaque ruptures, a blood clot will quickly form that can block blood flow in the artery and may lead to myocardial infarction which causes permanent damage to the heart muscle. Based on data from 2006 to 2016, the annual death rate attributable to CAD declined 31.8%. (Benjamin, Muntner et al. 2019) However, CAD remains the number one cause of death in the U.S. among both men and women. (CDC and NCHS) Both environmental and genetic factors contribute to the development of CAD, and lifestyle modification has played a major role in prevention of CAD in the past decades. (Lloyd-Jones, Hong et al. 2010)

CAD is a heritable condition with an estimated heritability of 50% to 60%. (Dai, Wiernek et al. 2016) A family history of cardiovascular disease has been shown as a strong predictor of incident disease. (Lloyd-Jones, Nam et al. 2004) Enormous effort as well as substantial progress has been made to understand the genetics behind CAD in the past decades and gene discovery studies in CAD have made a transition from recognition of familial patterns to discovery of the discrete genetic drivers. Similar to many other complex and common diseases, CAD has a polygenic architecture and has been treated as a good candidate for genome-wide association study (GWAS). GWAS relies on the usage of

genotype arrays to capture the majority of common inter-individual genetic variation. The very first GWASs of CAD published in 2007 reported common variants at the 9p21 locus associated with a ~30% increased risk of CAD per copy of the risk allele. (Helgadottir, Thorleifsson et al. 2007, McPherson, Pertsemlidis et al. 2007) The most recent GWAS of CAD combining data from the UK Biobank and the Coronary Artery Disease Genome-wide Replication and Meta-analysis plus the Coronary Artery Disease Genetics (CARDIoGRAMplusC4D) consortium has reported 64 novel CAD-associated loci, which accumulated the total number of CAD-associated loci to 161. (van der Harst and Verweij 2018) However, most loci have small effect size and fail to appreciably account for CAD heritability, which triggered interests in discovery of other genetic components such as rare variants and gene-environment interaction. Genetic research on CAD has also expanded progressively from pure disease gene localization to biological functions, mechanistic insights and clinical utilization.

Gene-environment (G×E) interaction is an important component of the genetic architecture of complex diseases. G×E interaction can be broadly defined as the interplay between genetic and environmental factors, and such interaction effect can be viewed as how genetic susceptibility for subpopulations modify certain environmental effects or how exposure to environmental factors modify certain genetic effects. (Gauderman, Mukherjee et al. 2017) Studying G×E interactions in complex diseases can improve understanding of disease etiology and identify susceptible or resistant sub-populations in response to environmental risk factors. (Ritz, Chatterjee et al. 2017) Several modifiable lifestyle-related factors such as smoking and physical activity are associated with cardio-metabolic conditions, (Benjamin, Muntner et al. 2019) but their interaction with genetic predisposition in the development of CAD hasn’t been fully understood. Detection of gene-lifestyle

interaction in CAD is usually hampered by availability of accurate measurements of lifestyle factors in most populations and individual studies are generally inadequately powered for exploration of gene-lifestyle interactions. Recently, consortia-based studies with large-scaled meta-analysis have been utilized to characterize gene-lifestyle interactions for complex diseases (Ahmad, Rukh et al. 2013, Nickels, Truong et al. 2013, Langenberg, Sharp et al. 2014, Usset, Raghavan et al. 2016) including CAD related traits. (Rao, Sung et al. 2017) Nevertheless, one major challenge of these consortia-based studies is the complexity of data sources that are derived from studies with heterogeneous designs and populations.

Large biobank cohorts have recently been established to achieve homogeneous measures of genetic and environmental factors while maintaining sufficient power to conduct both high-quality genomic and clinical research in complex diseases. One leading example is the UK Biobank, a large UK-based national cohort with over 500,000 participants that aims to improve prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses. It has whole-genome genomic data as well as a wide spectrum of health information including numerous environmental measurements. (Sudlow, Gallacher et al. 2015) Another example is the Million Veteran Program (MVP) which aims to build one of the world's largest medical databases by safely collecting blood samples and health information from one million veteran volunteers. (Gaziano, Concato et al. 2016) With data from survey instruments, the electronic health record, and biospecimens, the MVP is designed to facilitate scientific understanding of the potential links between genetic heterogeneity and disease. Recent genomic studies on cardio-metabolic phenotypes have benefited from the power of the UK Biobank and the MVP, (Warren, Evangelou et al. 2017, Evangelou, Warren et al. 2018, Klarin, Damrauer et al. 2018) which have also provided an

unparalleled opportunity for conduct of large-scale gene-lifestyle interaction studies for complex diseases such as CAD.

Due to the polygenic nature of complex diseases, G×E interaction studies are investigated using different analytical approaches. (Gauderman, Mukherjee et al. 2017) Candidate gene methods with a panel of SNPs have been widely used in genetic studies of CAD, including the development of a disease-specific genetic risk score (GRS). The GRS can be defined by summing the number of risk alleles for each of the disease associated SNPs weighted by their estimated effect sizes. (Ripatti, Tikkanen et al. 2010) The joint genetic effects represented by CAD-GRS have been shown to predict incident CAD. (Mega, Stitziel et al. 2015) In addition, gene-lifestyle interactions in CAD have been examined based on GRS and composite lifestyle assessments; (Khera, Emdin et al. 2016, Pazoki, Dehghan et al. 2018) however, whether effects of individual lifestyle factors on CAD risk can be modified by overall genetic susceptibility for CAD remains unclear.

Characterizing gene-lifestyle interactions for CAD can facilitate the mechanistic understanding of disease and help identify sub-populations more susceptible or resistant to CAD risk factors. For example, smoking is one of the most important modifiable CAD risk factors, but whether the smoking-related risk of CAD is modified by genomic status remains uncertain. Being physically active is protective against CAD; (Winzer, Woitek et al. 2018) however, it is not well understood whether the benefit of increased physical activity is uniform among all individuals, or whether it is modified by genomic background. Hindy et al. recently reported that the genetic predisposition to coronary heart disease (CHD) can modify the elevated CHD risk due to cigarette smoking, (Hindy, Wiberg et al. 2018) and Tikkanen et al. identified consistent effects of physical activity on CHD across low,

intermediate and high genetic risk groups in the UK Biobank. (Tikkanen, Gustafsson et al. 2018) However, current gene-lifestyle interaction studies for CAD have only focused on the multiplicative scale, ignoring potential effect modification on the additive scale. In fact, interaction on the additive scale has larger public health impact and under certain assumptions can be aligned with mechanistic interaction effects with a sufficient component cause framework. (VanderWeele 2009) Therefore, assessment of gene-lifestyle interaction effects on both scale should be performed and reported to better understand the causal mechanism and interplay between these two facets of CAD. (VanderWeele and Knol 2014)

Genetic mechanisms of CAD can be mediated through different traits including cholesterol and other lipid levels, obesity, and blood pressure (BP) level. (Webb, Erdmann et al. 2017) Previous studies predominantly focused on developing a comprehensive CAD-GRS, (Khera, Emdin et al. 2016, Hindy, Wiberg et al. 2018, Pazoki, Dehghan et al. 2018, Tikkanen, Gustafsson et al. 2018) however, whether a comprehensive CAD-GRS is specific enough to identify interaction effects with individual lifestyle factors such as smoking and physical activity remains debatable, since such interaction effects might act through different mediating traits including lipids, blood pressure or BMI. Therefore, it is important to capture the genetic predisposition to CAD mediated through different mechanisms and assess the gene-lifestyle interaction effects in parallel to provide finer evidence on how different pathways might interact with individual lifestyle factors.

In genetic studies of complex diseases, accounting for environmental exposures and G×E interactions may affect overall trait variance when investigating genetic contributions and can potentially identify novel loci, highlighting new biological processes and pathways. (Justice, Winkler et al. 2017) A two-degree-of-freedom procedure can be used to test the

combination of marginal genetic effects as well as G×E interactions. This method is proposed to be more powerful for detecting disease susceptibility loci where the true gene-environment interaction model is unknown. (Kraft, Yen et al. 2007) Lifestyle factors such as smoking and being physically inactive are important risk factors for CAD, but genetic variants that exert effects on CAD through interactions with smoking or physical activity remain undiscovered in previous CAD-GWAS due to heterogeneous main effects and stringent significance thresholds.

Focus of this dissertation

The work presented here is aimed towards the assessment of gene-lifestyle interactions for CAD with a focus on two established lifestyle-related CAD risk factors: smoking and physical activity. Specifically, genetic risk of CAD is captured by developing multiple genetic risk scores covering the overall genetic predisposition as well as genetic effects mediated through multiple clinical traits including blood pressure, lipids and BMI. Interaction effects are examined and reported on both additive and multiplicative scale in two large populations of European ancestry. In addition, CAD-GWAS accounting for gene-lifestyle interaction effects are performed for both smoking and physical activity to identify potential novel loci that haven't been discovered.

Chapter 2 presents work assessing gene-smoking interaction for incident CAD on both additive and multiplicative scale using data from the UK Biobank and the Million Veteran Program. Chapter 3 presents work assessing gene-physical activity interaction for incident CAD on both additive and multiplicative scale using data from the UK Biobank and the Million Veteran Program. Chapter 4 presents work conducting two GWASs of CAD

using a joint two degree of freedom approach accounting for gene-smoking interaction and gene-physical activity interaction with data from the UK Biobank.

Chapter 2

Assessment of gene-smoking interaction in coronary artery diseases

Introduction

CAD remains the number one of cause of death in the U. S. after the past decade when the mortality rate of CAD has been decreasing. (Ford and Capewell 2011) Lifestyle modification has played a major role in CAD prevention. Smoking is one of the most established CAD risk factor and extensive effort has been made to characterize the linkage between smoking and cardiovascular health. A meta-analysis comparing cardiovascular disease risks in 503,905 cohort participants >= 60 years of age reported an HR for cardiovascular mortality of 2.07 (95% CI, 1.82–2.36) compared with never-smokers and 1.37 (95% CI, 1.25–1.49) compared with former smokers. (Mons, Muezzinler et al. 2015) It has also been reported that female smokers have a 25% increase in risk for CAD than male smokers (RR: 1.25, 95% CI: 1.12 – 1.39). (Huxley and Woodward 2011) Despite the decreasing trend in smoking-related morbidity, it remains as the top preventable cause of death. (Leischow 2019) A recent study has shown that smoking cessation for former heavy smokers significantly reduced their risk for cardiovascular disease comparing to current smokers. (Duncan, Freiberg et al. 2019) The mechanism of smoking in CAD has been widely studied in both clinical and animal studies. Studies have shown that key processes in smoking-induced atherogenesis initiation are endothelial dysfunction and damage, increase in and oxidation of proatherogenic lipids, as well as decrease of high-density lipoprotein, induction of inflammation, and the shift toward a procoagulant state in the circulation. (Messner and Bernhard 2014) However, there is still a lack of knowledge on how smoking

behavior or cigarette smoke as a complex environmental exposure interacts with individual's genetic background to affect the development of cardiovascular disease outcomes.

CAD is a heritable condition with estimated heritability of 50% to 60%. (Dai, Wiernek et al. 2016) Substantial progress has been made to understand the genetic architecture of CAD in the past decades. Genetic studies of CAD have made a transition from recognition of familial patterns to discovery of individual genetic drivers. Recent genome-wide association studies (GWAS) of CAD have identified genetic susceptibility loci of CAD across the genome and demonstrated a polygenic architecture of CAD. Combining data from the UK Biobank and the Coronary Artery Disease Genome-wide Replication and Meta-analysis plus the Coronary Artery Disease Genetics (CARDIoGRAMplusC4D) consortium, the most recent GWAS of CAD has reported 64 novel CAD-associated loci, which accumulated the total number of CAD-associated loci to 161. (van der Harst and Verweij 2018) However, most loci have small effect size and fail to appreciably account for a large proportion of CAD heritability, which motivated the discovery of other genetic components such as gene-environment interactions. (Manolio, Collins et al. 2009) Gene-environment interaction can be broadly defined as the interplay between genetic and environmental factors, and such interaction effect can be viewed as how genetic susceptibility for subpopulations modify certain environmental effects or how exposure to environmental factors modify certain genetic effects. (Gauderman, Mukherjee et al. 2017) Understanding gene-environment interaction for CAD can expand our knowledge of potential biological mechanisms and clinical utilizations of CAD genetics. (Ritz, Chatterjee et al. 2017)

Gene-environment interaction is an important component of the genetic architecture of complex diseases such as CAD. Cigarette smoking is a complex environmental exposure

and lifestyle-related risk factor for CAD. Understanding the interaction between smoking and genetic predisposition in the development of CAD can potentially lead to better risk stratification and disease prevention. (Willett 2002, Ordovas and Tai 2008) However, very limited evidence was reported previously for gene-smoking interaction in CAD. One study has shown that smoking attenuated the increased risk for CAD associated with 9p21 risk alleles. (Hamrefors, Hedblad et al. 2014) A recent gene-smoking interaction study in CAD reported a higher magnitude of increased CAD risk by smoking among those with lower genetic risk of CAD captured by a CAD-GRS based on 50 SNPs. (Hindy, Wiberg et al. 2018) However, previous studies of gene-smoking interaction in CAD have focused and reported on multiplicative scale which by ignoring the additive scale only partially covers the potential interaction effect between genetic predisposition and smoking on CAD. In fact, interaction on the additive scale has larger public health impact and under certain assumptions can be aligned with mechanistic interaction effects with a sufficient component cause framework. (VanderWeele 2009) Therefore, assessment of gene-smoking interaction effects on both scale should be performed and reported to better understand the causal mechanism and interplay between these two risk factors. (VanderWeele and Knol 2014) In addition, genetic mechanisms of CAD can be mediated through different molecular pathways and mechanisms including cholesterol levels, obesity, and blood pressure (BP) levels, (Webb, Erdmann et al. 2017) but no study has been conducted to assess how the genetic predisposition of CAD driven by these intermediate traits can modify the increased risk due to smoking. Previous evidence of gene-smoking interaction in CAD is also limited by an incomplete list of CAD loci as well as insufficient sample size. Therefore, the purpose of this study is to assess gene-smoking interaction for CAD on both additive and multiplicative scale in populations of European ancestry using data from two of the largest biobank cohorts.

Methods

Study populations

The primary study population consists of participants with European ancestry from the UK Biobank (https://www.ukbiobank.ac.uk/) cohort. The UK Biobank is a major national and international health resource, and a registered charity in its own right, with the aim of improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Both genetic and phenotypic data for all participants in the UK Biobank were obtained for this study. The UK Biobank genetic data contains genome-wide genotypes for 488,377 participants. (Bycroft, Freeman et al. 2018) These were assayed using two very similar genotyping arrays. A subset of 49,950 participants involved in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study were genotyped at 807,411 markers using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix (now part of Thermo Fisher Scientific), and 438,427 participants were genotyped using the closely related Applied Biosystems UK Biobank Axiom Array (825,927 markers) that shares 95% of marker content with the UK BiLEVE Axiom Array. A quality control pipeline was developed and applied specifically to accommodate the large-scale dataset of ethnically diverse participants, genotyped in many batches, using two slightly different arrays, and which will be used by many researchers to tackle a wide variety of research questions. Markers that passed the quality control check were imputed using the Haplotype Reference Consortium (HRC) reference panel as well as the merged UK10K and 1000 Genomes phase 3 reference panels. Information was then

combined using the HRC data as the primary resource. For phenotype data, participants provided electronic signed consent, answered questions on socio-demographic, lifestyle and health-related factors, and completed a range of physical measures at baseline recruitment. All participants also provided consent for follow-up through linkage to their health-related records including in-patient hospital episode statistics and national death registry data.

The replication population of this study includes participants of European ancestry from the Million Veteran Program (MVP). The MVP is a national, voluntary research program funded entirely by the Department of Veterans Affairs (VA) Office of Research & Development. It is envisioned as a VA-based mega-biobank and launched to establish a national, representative, and longitudinal study of veterans for genomic and non-genomic research that combines data from survey instruments, the electronic health record and biospecimens. (Gaziano, Concato et al. 2016) The source population is defined as active users of the Veterans Health Administration (VHA), with the ability to provide informed consent as the only inclusion criterion. Recruitment is currently occurring in person at selected sites in the VHA health care system. Every Veteran is assigned a study ID number, which is used to track them throughout the entire process of recruitment, enrollment, sample collection and use. During recruitment veteran participants were informed about the MVP study via an invitation letter, explaining that participation in the study involves completing questionnaires, providing a blood sample for future research, allowing ongoing access to medical records and other health administrative data by authorized MVP staff, and agreeing to future contact by MVP staff for follow-up studies. The latest data release in 2018 contains genotype and phenotype data of over 500,000 participants among whom ~370,000 self-identified as non-Hispanic White. Genome-wide genotype data was measured using a customized Affymetrix Axiom biobank array, the MVP 1.0 Genotyping Array. With 723,305

total DNA sequence variants, the array is enriched for both common and rare variants of clinical importance in different ethnic backgrounds. Genotyped variants that were poorly called (genotype missingness > 5%) or that deviated from their expected allele frequency based on reference data from the 1000 Genomes Project were excluded. The remaining variants were used to conduct genotype imputation based on the 1000 Genomes Project phase 3, v.5 reference panel, which generated a total number of > 30 million variants. For phenotype data, participants were asked to complete two surveys: the MVP Baseline Survey and the MVP Lifestyle Survey. Conceptually, the MVP Baseline Survey was designed to collect information regarding demographics, family pedigree, health status, lifestyle habits, military experience, medical history, family history of specific illnesses, and physical features. The MVP Lifestyle Survey contains questions from validated instruments in domains selected to provide information on sleep and exercise habits, environmental exposures, dietary habits, and sense of wellbeing. Other health-related information or disease diagnosis data is collected through linkage to participants' VA electronic health record.

Outcome measurements

The disease outcome for this study is defined as primary events of incident CAD. In the UK Biobank, participants' survey data is linked to in-patient hospital episode statistics (HES) as well as national death registry data. CAD definition in the UK Biobank for this study is referenced from the most recent GWAS of CAD using the UK Biobank data. (van der Harst and Verweij 2018) A participant is defined as a CAD case if he/she has at least one occurrence of the following International Classification of Diseases, 10th edition (ICD-10) codes: I21-I25 covering ischemic heart diseases; or at least one occurrence of the following Office of Population Censuses and Surveys Classification of Interventions and Procedures,

version 4 (OPCS-4) codes: K40-K46, K49, K50 and K75 which includes replacement, transluminal balloon angioplasty, and other therapeutic transluminal operations on coronary artery and percutaneous transluminal balloon angioplasty and insertion of stent into coronary artery. Death because of CAD was defined as an occurrence of any ICD-10 codes stated above in the primary cause of death. To identify incident CAD cases, participants with CAD diagnosis before enrollment in the UK Biobank were excluded. Participants will be censored on the earliest date of CAD event/CAD death after enrollment, or the end of HES-based follow-up, or time of competing death, whichever occurs first.

In the MVP cohort, CAD definition was developed by a group of expert researchers from the MVP Cardiovascular Working Group. Disease diagnosis data was queried on two different index dates: date of enrollment and July 1st, 2017. The CAD definition has been chosen to accommodate both the number of cases for statistical power as well as accuracy in CAD diagnosis to control false positive rate. Participants were defined as a CAD case if there is occurrence of any CAD codes on two or more distinct dates on or prior to the index date, or occurrence of a revascularization procedure code on or prior to the index date. CAD codes include: International Classification of Diseases, 9th edition (ICD-9) codes 410, 411.0, 411.1, 411.81, 411.89, 412, 414.00, 414.01-414.05, 414.2-414.4, 414.8, 414.9, V45.81, V45.82; and ICD-10 codes I20.0, I21-I24, I25.1, I25.2, I25.5, I25.6, I25.70, I25.71, I25.72, I25.73, I25.79, I25.810, I25.82, I25.83, I25.84, I25.89, I25.9, Z95.1, Z98.61. Revascularization procedure codes include: International Classification of Diseases, 9th edition (ICD-9) codes 00.66, 36.0, 36.01-36.07, 36.09, 36.1, 36.11-36.17, 36.19, 36.2, 99.10; and ICD-10 codes 0210-0213, 0270-0273, 02C0, 02C1, 02C3, 02C4; and Current Procedural Terminology (CPT) codes: 33510-33514, 33516-33519, 33521-33523, 33530, 33533-33536, 33572, 92928, 92929, 92933, 92934, 92937, 92938, 92941, 92943, 92944, 92973-92975, 92977, 92980-92982,

92984, 92995, 92996, G0290, G0291, C9600-C9608. To identify incident CAD cases, participants who had CAD diagnosis on or prior to enrollment date were excluded. New CAD cases were defined as diagnosis between enrollment date and July 1st, 2017.

Smoking and covariate measurements

In the UK Biobank, smoking was self-reported in the touchscreen questionnaire on lifestyle and personal exposures. Smoking status is categorized into "current", "previous", "never" and "prefer not to answer". Age started smoking is measured among current and previous smokers, and age stopped smoking is measured among previous smokers. Number of cigarettes currently smoked daily or previously smoked daily was measured among current smokers or previous smokers, respectively. Therefore, a pack-year variable is also derived and included in the secondary analysis. In the MVP cohort, smoking was self-reported in both the MVP Baseline Survey and the MVP Lifestyle Survey. Smoking status is categorized into "current", "previous" and "never". Information on potential confounders was also collected from baseline questionnaire data of the UK Biobank and the MVP cohort including age, sex, alcohol consumption, education, history of hypertension and diabetes, usage of cholesterol medication, BMI and social economic status.

Genetic data processing and principle component analysis

Genome-wide genotyped SNP data of the UK Biobank is first examined by quality control procedures. Individuals with genetically defined non-European ancestry are excluded. Markers or individuals with a call rate less than 95 percent are also excluded. SNPs with Hardy-Weinberg Equilibrium p-value less than 10^{-6} or minor allele frequency less than

0.0001 are excluded. To remove up to the 3rd degree relatedness among the UK Biobank participants, a pairwise kinship coefficient matrix is used with kinship larger than or equal to 0.0442 as a cutoff to filter the related individual pairs. SNPs that passed the quality control procedure are then undergone a linkage disequilibrium (LD) pruning procedure with a window size of 50 kb, a step size of 5 variants, and an r^2 threshold of 0.05. LD pruned SNPs are then used in the principle component analysis. Top ten principle components are calculated and will be included in the main analysis as covariates to control for population stratification. In the MVP cohort, duplicate samples, samples with more heterozygosity than expected, an excess (>2.5%) of missing genotype calls, or discordance between genetically inferred sex and phenotypic gender are excluded. In addition, one individual from each pair of related individuals is excluded. An ethnicity-specific principle component analysis was then performed among non-Hispanic White participants who are defined as: (self-identified "non-Hispanic", "white", and > 80% genetic European ancestry).

Genetic risk score (GRS) construction

A comprehensive CAD-GRS based on 161 loci that have been reported in the most recent GWAS of CAD (van der Harst and Verweij 2018) was developed. A weighted GRS approach was implemented using the formula below:

$$GRS = \beta_1 \times SNP_1 + \beta_2 \times SNP_2 + \ldots + \beta_n \times SNP_n$$

β_i are effect sizes from GWAS or GWAS-meta-analysis; SNP_i is coded as number of risk alleles. In this study, the effect sizes for CAD-GRS construction were referenced from CAD-GWAS summary statistics of the CARDIoGRAMplusC4D consortium (Nikpay, Goel et al. 2015) to avoid sample overlap with the UK Biobank or the MVP cohort. In addition, three mediating trait-based CAD-sub-GRSs were developed based on loci that are associated with

lipids level ($GRS_{CAD\text{-}lipids}$), blood pressure ($GRS_{CAD\text{-}BP}$), or BMI ($GRS_{CAD\text{-}BMI}$). Lipids-associated loci and blood pressure-associated loci were extracted from recent GWAS publications, (Evangelou, Warren et al. 2018, Klarin, Damrauer et al. 2018) and BMI-associated loci were obtained from unpublished BMI GWAS of up to 1 million individuals of European ancestry. Loci that are associated with both CAD and only one of the three target mediating traits (Bonferroni corrected $p<0.05$) were included in the mediating trait-specific CAD-sub-GRS calculation using the same weighted approach. All GRS constructions were performed using PLINK 2.0 (https://www.cog-genomics.org/plink/2.0/) with "--score" function, and missing genotypes were imputed to the mean dosage.

Statistical analysis

Cox proportional hazards models were used to assess the association of CAD-GRSs and sub-GRSs with incident CAD as well as the interaction between GRSs and smoking in the UK Biobank. All genetic risk scores were 1) standardized and modeled as continuous variables and 2) categorized into quintiles and divided into low (lowest quintile), intermediate (quintiles 2 to 4) and high (highest quintile) genetic risk group. Smoking status was categorized as "never", "past" and "current" smokers and pack-year was divided into terciles as "low", "medium" and "high". Age, sex, alcohol consumption, education, (Davies, Dickson et al. 2018) history of hypertension, history of diabetes, cholesterol lowering medication use, BMI, Townsend deprivation index and ten principle components were included as covariates. In the CAD-sub-GRS analysis, the corresponding mediating trait was not included as a covariate to avoid over-adjustment. Therefore, history of hypertension was not adjusted for in the $GRS_{CAD\text{-}BP}$ analysis, cholesterol lowering medication was not adjusted for in the $GRS_{CAD\text{-}lipids}$ analysis, and BMI was not adjusted for in the $GRS_{CAD\text{-}BMI}$ analysis. Proportional hazards

assumption was assessed using Schoenfeld's test. When the assumption is violated, categorical variables were stratified on while interaction terms with time were added for continuous variables. Multiplicative interaction between CAD-GRSs and smoking status was assessed by including interaction terms in the model and conducting likelihood ratio tests, and additive interaction was assessed by calculating relative excess risk due to interaction (RERI) based on the hazard ratio estimates. (Li and Chambless 2007) A bootstrap resampling method was used to construct 95% confidence intervals for RERI estimates. For the replication analysis in the MVP, logistic regression models were used controlling for a similar set of covariates to assess GRS-smoking interaction on both multiplicative and additive scale. Odds ratio estimates were used to calculate RERI in the MVP replication analysis and same bootstrap procedures were done to construct confidence intervals.

Results

307,147 participants of European ancestry who were free of CAD at baseline from the UK Biobank were included in the final analysis. A detailed inclusion/exclusion and QC process was presented in Figure I. Basic characteristics of the study population were shown in Table I. 9,847 primary incident CAD events were identified from the UK Biobank. In our study sample, the mean age at baseline is 56.7 years and slightly more females (55%) than males were included. 9.5 percent of the participants were identified as current smokers and 34.9 percent as past smokers based on self-reported questionnaire data at baseline. In the replication analysis using data from the MVP, 102,283 participants with no CAD at enrollment were included. A similar inclusion/exclusion process was presented in Figure I. Using the indexing-date method, we have identified 8,016 incident CAD cases between enrollment and July 1st, 2017. Baseline characteristics of the MVP participants were

summarized in Table I. The mean age is 63.9 years old and the majority of them are male (91.7%). 17.8 percent and 48.6 percent of participants self-reported as current smokers and past smokers, respectively.

In the main analysis, a comprehensive CAD-GRS in the UK Biobank was constructed based on 161 CAD-loci that were reported in the most recent CAD-GWAS. (van der Harst and Verweij 2018) One lead SNP (rs582384) was multiallelic in the UK Biobank so a proxy SNP (rs616381, r^2=0.86 for European ancestry) was used. A detailed list of all SNPs used for score construction is presented in Table II. In the sub-score analysis, 26 SNPs were included in the $GRS_{CAD\text{-}BP}$ construction, 17 SNPs were included in the $GRS_{CAD\text{-}lipids}$ and 16 SNPs were included in the $GRS_{CAD\text{-}BMI}$. (Table III) Associations between all four CAD-GRSs with primary incident CAD in the UK Biobank were presented in Table IV-A. One standard deviation (SD) increase in the comprehensive CAD-GRS is independently associated with 33.6 percent increase in the risk of primary CAD events (HR: 1.336, 95% CI: 1.310, 1.363). Comparing to those with low genetic risk of CAD, those at intermediate genetic risk have a 43.6 percent increase in the risk or primary CAD events (HR: 1.436, 95% CI: 1.352, 1.526), and those at high genetic risk have an over two-fold increase in the risk of primary CAD events (HR: 2.196, 95% CI: 2.056, 2.345). Among the three CAD mediating trait-based sub-GRSs, the $GRS_{CAD\text{-}lipids}$ has the strongest association with incident primary CAD events. Participants with a high genetic risk captured by the $GRS_{CAD\text{-}lipids}$ have a 56.7 percent increase in CAD risk comparing to those with low genetic risk (HR: 1.567, 95% CI: 1.472, 1.668), and this genetic effect attenuated when focusing on only BP-associated CAD loci (HR: 1.422, 95% CI: 1.335, 1.514) or only BMI-associated CAD loci (HR: 1.245, 95% CI: 1.169, 1.326).

To further understand how the genetic predisposition of CAD interacts with smoking status, we assessed GRS-smoking interaction on both multiplicative and additive scale in the UK Biobank. Combined associations of CAD-GRS and smoking status with incident CAD were presented in Table V-A. Comparing to never smokers with low genetic risk, those who never smoke but possess a high genetic risk had over two-fold increase in CAD risk (HR: 2.264, 95% CI: 2.045,2.506), and those with high genetic risk who currently smoke had a four times higher risk in primary CAD events (HR: 4.077, 95% CI: 3.586,4.636). Similar patterns were also observed for each of the three CAD mediating trait-based sub-GRSs. Comparing to never smokers with low genetic risk in $GRS_{CAD\text{-}lipids}$, current smokers with high genetic risk have a much higher risk elevation (HR: 2.861, 95% CI: 2.512, 3.259) than never smokers with high genetic risk (HR: 1.658, 95% CI: 1.505, 1.827). The strongest combined effect of genetic predisposition and smoking status on CAD is observed in lipids-associated loci, and BMI-associated loci seem to have weaker effects than lipids or BP-associated CAD loci. Overall, no multiplicative interaction between CAD-GRS or CAD-sub-GRSs and smoking status was observed. In the MVP replication analysis, a similar pattern of associations between CAD-GRS and incident CAD was observed. (Table IV-B and Table V-B) The comprehensive CAD-GRS as well as all sub-GRSs were found to be associated with incident CAD except for intermediate risk category of $GRS_{CAD\text{-}BMI}$. $GRS_{CAD\text{-}lipids}$ had the strongest effect on CAD among all three intermediate traits. When combined with smoking status, no multiplicative interaction was observed between CAD-GRS and smoking status.

We then assessed the additive interaction between CAD-GRSs and smoking status by calculating RERI for each GRS. (Figure II) For the comprehensive CAD-GRS, synergistic additive interaction was observed among current smokers in both intermediate (RERI: 0.394, 95% CI: 0.097,0.729) and high (RERI: 1.051, 95% CI: 0.615,1.497) genetic risk group in the

UK Biobank (Figure II-A), meaning the absolute risk elevation due to genetic predisposition is stronger among current smokers. When comparing the three mediating trait-based CAD-sub-GRSs, the strongest additive interaction between CAD-GRS and smoking status was observed for BP-associated CAD loci. Synergistic additive interaction was observed among past smokers in intermediate genetic risk group of $GRS_{CAD\text{-}BP}$ (RERI: 0.134, 95% CI: 0.011, 0.246) and current smokers with high genetic risk in $GRS_{CAD\text{-}BP}$ (RERI: 0.650, 95% CI: 0.299, 1.032). No additive interaction was observed for $GRS_{CAD\text{-}lipids}$ or $GRS_{CAD\text{-}BMI}$. To replicative the additive interaction effect of smoking status and CAD-GRS, we also calculated RERI and corresponding bootstrap intervals in the MVP. (Figure II-B) No additive interaction effect was observed in the MVP replication analysis. An individual SNP-based interaction analysis was also conducted with current smoking in the UK Biobank, no multiplicative interaction was observed after multiple-testing correction, but one locus (rs11591147, *PCSK9*) had an additive interaction effect with current smoking. (Table VI) The effect of smoking is predominantly observed among homozygotes of the risk allele.

We also explored interaction effects between CAD-GRS and pack-year among ever smokers in the UK Biobank. Combined effects of CAD-GRS and pack-year on incident CAD among ever smokers is presented in Table VII. Comparing to the strongest effect observed when participants have high lipids genetic risk profile and were identified as current smokers in previous analysis, we observed a stronger effect combining blood pressure genetic risk and pack-year exposure. Similarly, no multiplicative interaction was observed for any CAD-GRS with pack-year. However, an additive interaction effect was observed for those who had high overall genetic risk to CAD and were in the high pack-year group (RERI: 0.519, 95% CI: 0.032, 1.005). (Figure III) After comparing three CAD-mediating traits blood pressure, lipids,

and BMI, $GRS_{CAD\text{-}lipids}$ had an additive interaction effect with pack-year when both genetic risk and pack-year exposure were high (RERI: 0.422, 95% CI: 0.072, 0.772).

Discussion

In this study, we have assessed how an overall genetic risk of CAD captured by CAD-GRS interact with smoking on both multiplicative and additive scale in European population using data from two large biobank cohorts. To further understand the role of mediating traits such as BP, lipids and BMI in such gene-smoking interaction on CAD, we also developed three separate sub-scores ($GRS_{CAD\text{-}BP}$, $GRS_{CAD\text{-}lipids}$ and $GRS_{CAD\text{-}BMI}$) focusing on genetic loci uniquely associated with one mediating trait as well as CAD risk. Our results have shown an additive and synergistic interaction effect between smoking status and CAD-GRS driven by BP-associated loci, as well as an additive and synergistic interaction effect between smoking pack-year and CAD-GRS driven by lipids-associated loci. Individual SNP analysis has pointed to one locus (rs11591147, *PCSK9*) for a positive additive interaction effect with smoking status.

Smoking has been long established as a strong risk factor for CAD. (Huxley and Woodward 2011) Mechanisms of how tobacco smoke impact cardiovascular health have been proposed but not fully understood. (Messner and Bernhard 2014) Understanding the interaction between smoking and genetic predisposition to CAD might reveal unknown disease pathways and help to stratify the populational susceptibility to CAD. However, very limited studies have been conducted to estimate the gene-smoking interaction effects on CAD for both multiplicative and additive scale. Using a weighted GRS constructed with 50 SNPs, Hindy et al. reported a higher relative increase in CAD risk due to smoking among those with

lower genetic risk. (Hindy, Wiberg et al. 2018) In our study, we didn't observe such multiplicative interaction effect, but we were able to quantify a synergistic interaction effect on the additive scale between smoking status and an overall CAD-GRS among current smokers in the UK Biobank. In addition, we observed such synergistic effect with a sub-CAD-GRS developed focusing on loci that are associated with only BP and CAD. Elevated blood pressure or hypertension is associated with CAD, and smoking can cause acute increase in blood pressure and acts synergistically with hypertension to increase the risk of CAD. (De Cesaris, Ranieri et al. 1992) Our results have supported that blood pressure related pathways might be important in the development of CAD when smoking is initiated. We also explored gene-smoking interaction using pack-year to quantify smoking intensity, which has been reported as the preeminent smoking-related risk factor for cardiovascular disease, (Lubin, Couper et al. 2016) and found an additively synergistic interaction effect between genetic risk of CAD and smoking pack-year driven by lipids associated loci. Smoking cessation among current smokers has been reported to benefit HDL cholesterol (Gepner, Piper et al. 2011) and our results have shown a larger benefit of reducing smoking intensity for those who have high genetic risk profile for lipids and CAD.

We also explored gene-smoking interaction for individual CAD risk loci, and one SNP (rs11591147) at the *PCSK9* locus was found to have synergistic interaction with current smoking on the additive scale (RERI: 0.404, 95% CI: 0.343, 0.465) but no multiplicative interaction was observed for the 161 loci tested. Our results have shown that the effect of current smoking on CAD risk is predominantly observed among homozygotes of the risk allele (G/G) for rs11591147. *PCSK9*, a gene encoding proprotein convertase subtilisin/kexin type 9 regulates cholesterol homeostasis and was found to be associated with autosomal dominant hypercholesterolemia. (Abifadel, Varret et al. 2003) Mutations reducing the

expression level of *PCSK9* were reported to be associated with lower plasma level of LDL and lower risk of CAD. (Cohen, Pertsemlidis et al. 2005, Cohen, Boerwinkle et al. 2006) Our findings have shown that *PCSK9* might also act as an effect modifier on the elevated risk of CAD due to smoking. Several other loci including *ADAMTS7*, (Saleheen, Zhao et al. 2017) *APOE* (Gustavsson, Mehlig et al. 2012, Holmes, Frikke-Schmidt et al. 2014) and *9p21* (Hamrefors, Hedblad et al. 2014) were studied previously for interaction effects with smoking on CAD, but our analysis was not able to detect interaction effects across these loci.

With data from two of the largest biobank cohorts, we assessed how genetic predisposition to CAD captured by weighted CAD-GRS interact with smoking status as well as smoking intensity (pack-year) on the risk of primary incident CAD among the European population. Our findings have shown that smoking, as one of the most established life-style related CAD risk factors, acts multiplicatively with genetic factors on increasing CAD risk. However, a synergistic interaction effect on additive scale was observed showing the absolute risk increase driven by smoking is higher among individuals with higher genetic risk. Such interaction effect was not reported in previous gene-lifestyle (Khera, Emdin et al. 2016, Said, Verweij et al. 2018) or gene-smoking interaction studies of CAD partially due to the analytical focus on multiplicative scale. However, it is often of greater public health importance to assess interaction effects on the additive scale and additive interaction effects can be better linked with mechanistic effects under a sufficient component cause framework with certain assumptions satisfied. (VanderWeele 2009) To further understand the gene-smoking interaction effect on CAD, we also developed three mediating trait-based CAD sub-GRSs and identified a BP-driven interaction effect with smoking status and a lipids-driven interaction effect with smoking intensity measured by pack-year. Our results have provided novel hypotheses on the potential different mechanisms between smoking initiation and

smoking intensity on CAD development. Individual SNP-based interaction analysis has also identified one novel additive interaction effect at the locus of *PCSK9*, which suggests the role of lipid metabolism in modifying the effect of smoking on CAD risk.

Our study also has several limitations. First, smoking is measured subjectively at baseline with questionnaire data and such measurement is susceptible to recall bias. However, to balance the statistical power required for gene-lifestyle interaction studies and measurement accuracy as well as ensuring a homogeneous study population, biobank cohorts seem to be so far the best data resource in conducting large scale gene-lifestyle interaction studies. Second, we used the MVP cohort as a replication cohort for our primary analysis conducted in the UK Biobank, but the two populations differ largely in many aspects, which limited the power and validity of the replication analysis. In addition, CAD cases were captured in slightly different ways between these two cohorts due to data availability restrictions, which also lead to slightly different definitions of some covariates in the analysis. Therefore, our results need to be interpreted with caution when comparing the primary and replication analysis.

Conclusion

Using data from the UK Biobank and the MVP cohort, we have prospectively assessed gene-smoking interaction on incident CAD on both additive and multiplicative scale. No multiplicative interaction was found between genetic predisposition and smoking status or smoking intensity, but a synergistic additive interaction driven by BP-associated loci was observed for current smoking and a synergistic additive interaction driven by lipids-associated loci was observed for smoking intensity. In addition, the *PCSK9* locus was

observed to have strong additive interaction effects with current smoking on CAD risk. Our findings have raised hypothesis with respect to different interplaying mechanisms between genetic predisposition to CAD and smoking behaviors, and highlighted the value of addressing gene-lifestyle interactions in CAD on both additive and multiplicative scale.

Table I. Characteristics of the study population

Characteristic	Mean (SD) or N (%)	
	UK Biobank (N=307,147)	MVP (N=102,283)
Age	56.7 (8.0)	63.9 (12.2)
Female	168,880 (55.0%)	8,439 (8.3%)
BMI	27.3 (4.7)	29.9 (5.7)
Smoking Status		
Current	29,281 (9.5%)	18,167 (17.8%)
Past	107,049 (34.9%)	49,670 (48.6%)
Never	170,817 (55.6%)	34,446 (33.6%)
Pack-year	22.3 (18.1)	n/a
Alcohol Consumption		
Ever	297,997 (97.0%)	61,873 (60.5%)
Never	9,150 (3.0%)	40,410 (39.5%)
Hypertension	79,678 (25.9%)	57,074 (55.8%)
Diabetes	13,059 (4.3%)	21,932 (21.4%)
Lipids Medication	45344 (14.8%)	43,593 (42.6%)
Education		
School leaving age >=15	243,327 (79.2%)	n/a
School leaving age <15	63,820 (20.8%)	n/a
Some college or higher	n/a	78,189 (76.4%)
Townsend Index	-1.6 (2.9)	n/a
Income		
$50,000 or above	n/a	37,571 (36.7%)

Table II. 161 SNPs included in the CAD-GRS construction

SNP	CHR	POS	Locus	Risk Allele	Other Allele	Weight
rs36096196	1	2252205	*MORN1*	T	C	0.043
rs2493298	1	3325912	*PRDM16*	A	C	0.057
rs61776719	1	38461319	*SF3A3*	A	C	0.052
rs11591147	1	55505647	*PCSK9*	G	T	0.257
rs56170783	1	57016131	*PLPP3*	A	C	0.127
rs7528419	1	109817192	*CELSR2*	A	G	0.115
rs11806316	1	115753482	*RP4-663N10.1*	G	A	0.032
rs11810571	1	151762308	*TDRKH*	G	C	0.056
rs6689306	1	154395946	*IL6R*	A	G	0.056
rs1892094	1	169094459	*ATP1B1*	C	T	0.018
rs6700559	1	200646073	*RP11-92G12.3*	C	T	0.028
rs2820315	1	201872264	*LMOD1*	T	C	0.047
rs60154123	1	210468999	*RP4-667H12.4*	T	C	0.045
rs67180937	1	222823743	*MIA3*	G	T	0.079
rs699	1	230845794	*AGT*	G	A	0.040
rs16986953	2	19942473	*AC019055.1*	A	G	0.085
rs585967	2	21270554	*APOB*	C	A	0.073
rs4299376	2	44072576	*ABCG8*	G	T	0.051
rs616381	2	45891708	*PRKCE*	A	G	0.033
rs7568458	2	85788175	*GGCX*	A	T	0.060
rs17678683	2	145286559	*ZEB2*	G	T	0.099
rs12999907	2	164957251	*AC092684.1*	A	G	0.047
rs840616	2	188196469	*AC007319.1*	C	T	0.046
rs114123510	2	203831212	*CARF*	A	T	0.134
rs17517928	2	216291359	*FN1*	C	T	0.057
rs2571445	2	218683154	*TNS1*	A	G	0.043
rs2972146	2	227100698	*NEU2*	T	G	0.039
rs13003675	2	233584109	*GIGYF2*	T	C	0.042
rs11677932	2	238223955	*STK25*	G	A	0.037
rs748431	3	14928077	*FGD5*	G	T	0.049
rs7633770	3	46688562	*SNORD77*	A	G	0.025
rs7617773	3	48193515	*TKT*	T	C	0.039
rs7623687	3	49448566	*RHOA*	A	C	0.070
rs17843797	3	124453022	*UMPS*	G	T	0.064
rs10512861	3	132257961	*DNAJC13*	G	T	0.041
rs667920	3	136069472	*STAG1*	T	G	0.039
rs9818870	3	138122122	*MRAS*	T	C	0.065
rs12493885	3	153839866	*ARHGEF26*	C	G	0.066
rs4266144	3	156852592	*SPTSSB*	G	C	0.032
rs12897	3	172115902	*FNDC3B*	G	A	0.050
rs16844401	4	3449652	*HGFAC*	A	G	0.072
rs72627509	4	57839051	*NOA1*	G	C	0.060
rs12500824	4	77416627	*SHROOM3*	A	G	0.029

rs10857147	4	81181072	*RP11-576N17.4*	T	A	0.055
rs11099493	4	82587050	*RASGEF1B*	A	G	0.048
rs3775058	4	96117371	*UNC5C*	A	T	0.035
rs11723436	4	120901336	*RP11-170N16.1*	G	A	0.054
rs35879803	4	146782837	*ZNF827*	C	A	0.048
rs6841581	4	148401190	*EDNRA*	A	G	0.064
rs2306556	4	156638573	*GUCY1A3*	A	G	0.067
rs7696431	4	169687725	*PALLD*	T	G	0.036
rs1508798	5	9556694	*RP11-260E18.1*	T	C	0.047
rs3936511	5	55860781	*C5orf67*	G	A	0.035
rs1800449	5	121413208	*LOX*	T	C	0.045
rs77335401	5	131759825	*C5orf56*	C	T	0.052
rs246600	5	142516897	*ARHGAP26*	T	C	0.024
rs9501744	6	1617143	*FOXC1*	C	T	0.062
rs742115	6	11327021	*NEDD9*	C	T	0.040
rs9349379	6	12903957	*PHACTR1*	G	A	0.132
rs35541991	6	22583856	*RP1-309H15.2*	C	CA	0.051
rs3130683	6	31888367	*C2*	T	C	0.090
rs4472337	6	34769765	*UHRF1BP1*	T	C	0.056
rs1321309	6	36638636	*LAP3P2*	A	G	0.029
rs56015508	6	39152041	*KCNK5*	C	A	0.063
rs6905288	6	43758873	*VEGFA*	A	G	0.040
rs9367716	6	57160572	*RNU7-66P*	G	T	0.036
rs4613862	6	82612271	*RP11-379B8.1*	A	C	0.038
rs1591805	6	126717064	*RP11-394G3.2*	A	G	0.040
rs12202017	6	134173151	*TARID*	A	G	0.067
rs17080091	6	150997401	*PLEKHG1*	C	T	0.063
rs10455872	6	161010118	*LPA*	G	A	0.319
rs10267593	7	1937261	*MAD1L1*	G	A	0.042
rs7797644	7	6486067	*DAGLB*	C	T	0.044
rs11509880	7	12261911	*TMEM106B*	A	G	0.039
rs2107595	7	19049388	*HDAC9*	A	G	0.073
rs2107732	7	45077978	*CCM2*	G	A	0.045
rs10953541	7	107244545	*BCAP29*	C	T	0.050
rs975722	7	117332914	*CFTR*	G	A	0.029
rs11556924	7	129663496	*ZC3HC1*	C	T	0.073
rs10237377	7	139757136	*PARP12*	G	T	0.032
rs3918226	7	150690176	*NOS3*	T	C	0.133
rs6997340	8	18286997	*NAT2*	T	C	0.032
rs2083636	8	19865263	*LPL*	T	G	0.051
rs6984210	8	22033615	*BMP1*	G	C	0.081
rs10093110	8	106565414	*ZFPM2*	G	A	0.026
rs2954029	8	126490972	*RP11-136O12.2*	A	T	0.044
rs2891168	9	22098619	*CDKN2B-AS1*	G	A	0.193
rs944172	9	110517794	*AL162389.1*	C	T	0.039
rs111245230	9	113169775	*SVEP1*	C	T	0.054

rs885150	9	124420173	*DAB2IP*	C	T	0.039
rs507666	9	136149399	*ABO*	A	G	0.079
rs61848342	10	12303813	*RN7SL232P*	C	T	0.029
rs1887318	10	30321598	*KIAA1462*	T	C	0.062
rs1870634	10	44480811	*LINC00841*	G	T	0.076
rs17680741	10	82251514	*TSPAN14*	T	C	0.045
rs2246942	10	91004886	*LIPA*	G	A	0.066
rs11191416	10	104604916	*PFN1P11*	T	G	0.079
rs4918072	10	105693644	*STN1*	A	G	0.040
rs4752700	10	124237612	*HTRA1*	G	A	0.029
rs11601507	11	5701074	*TRIM5*	A	C	0.073
rs10840293	11	9751196	*SWAP70*	A	G	0.055
rs1351525	11	13301548	*ARNTL*	T	A	0.048
rs7116641	11	43696917	*RP11-472I20.4*	G	T	0.030
rs12801636	11	65391317	*PCNX3*	G	A	0.050
rs590121	11	75274150	*SERPINH1*	T	G	0.032
rs7947761	11	100624599	*ARHGAP42*	G	A	0.047
rs2839812	11	103673294	*RP11-563P16.1*	T	A	0.066
rs964184	11	116648917	*ZPR1*	G	C	0.050
rs11838267	12	7175872	*C1S*	T	C	0.049
rs10841443	12	20220033	*RP11-664H17.1*	G	C	0.051
rs11170820	12	54513915	*FLJ12825*	G	C	0.091
rs2229357	12	57843711	*INHBC*	G	A	0.012
rs2681472	12	90008959	*ATP2B1*	G	A	0.074
rs7306455	12	95355541	*NDUFA12*	G	A	0.055
rs10774625	12	111910219	*ATXN2*	A	G	0.067
rs11830157	12	118265441	*KSR2*	G	T	0.035
rs2244608	12	121416988	*HNF1A*	G	A	0.048
rs11057401	12	124427306	*CCDC92*	T	A	0.027
rs1924981	13	29022645	*FLT1*	T	C	0.050
rs9591012	13	33058333	*N4BP2L2*	G	A	0.046
rs11617955	13	110818102	*COL4A1*	T	A	0.089
rs1317507	13	113631780	*MCF2L*	A	C	0.038
rs2145598	14	58794001	*ARID4A*	G	A	0.028
rs3832966	14	75614504	*TMED10*	ACCCG	A	0.052
rs112635299	14	94838142	*SERPINA1*	G	T	0.168
rs10139550	14	100145710	*HHIPL1*	G	C	0.055
rs6494488	15	65024204	*RBPMS2*	A	G	0.040
rs72743461	15	67441750	*SMAD3*	C	A	0.070
rs7164479	15	79123054	*MORF4L1*	T	C	0.077
rs2083460	15	89574484	*RP11-326A19.2*	T	C	0.080
rs2071382	15	91428197	*FES*	T	C	0.053
rs17581137	15	96146414	*RP11-61O11.1*	A	C	0.051
rs247616	16	56989590	*AC012181.1*	C	T	0.031
rs1050362	16	72130815	*DHX38*	A	C	0.029
rs3851738	16	75387533	*CFDP1*	C	G	0.045

rs7199941	16	81906423	*PLCG2*	A	G	0.036
rs7500448	16	83045790	*CDH13*	A	G	0.055
rs216172	17	2126504	*SMG6*	C	G	0.048
rs9897596	17	17593453	*RAI1*	T	C	0.041
rs13723	17	27941886	*CORO6*	G	A	0.037
rs76954792	17	30033514	*RP11-805L22.1*	T	C	0.029
rs2074158	17	40257163	*DHX58*	C	T	0.066
rs17608766	17	45013271	*GOSR2*	C	T	0.053
rs4643373	17	47123423	*IGF2BP1*	T	C	0.055
rs8068952	17	59286644	*BCAS3*	G	C	0.034
rs1867624	17	62387091	*RPL31P57*	T	C	0.040
rs9964304	18	47229717	*RP11-813F20.2*	C	A	0.035
rs663129	18	57838401	*RNU4-17P*	A	G	0.058
rs116843064	19	8429323	*ANGPTL4*	G	A	0.141
rs6511720	19	11202306	*LDLR*	G	T	0.125
rs73015714	19	17855763	*FCHO1*	G	C	0.052
rs10417115	19	33386556	*CEP89*	C	T	0.069
rs8108632	19	41854534	*TGFB1*	T	A	0.052
rs7412	19	45412079	*APOE*	C	T	0.137
rs867186	20	33764554	*PROCR*	A	G	0.061
rs6102343	20	39924279	*ZHX3*	A	G	0.045
rs3827066	20	44586023	*ZNF335*	T	C	0.052
rs260020	20	57714025	*ZNF831*	T	C	0.048
rs2832227	21	30533076	*MAP3K7CL*	G	A	0.044
rs28451064	21	35593827	*AP000318.2*	A	G	0.128
rs180803	22	24658858	*POM121L9P*	G	T	0.181

Table III. SNPs included in the sub-score construction

A. $GRS_{CAD\text{-}BP}$

SNP	CHR	POS	Locus	Risk Allele	Other Allele	Weight	P-value		
							SBP	DBP	PP
rs61776719	1	38461319	*SF3A3*	A	C	0.0518	4.249E-04	3.323E-02	4.228E-11
rs1892094	1	169094459	*ATP1B1*	C	T	0.0185	3.138E-04	8.352E-01	1.570E-05
rs699	1	230845794	*AGT*	G	A	0.0396	9.473E-10	6.681E-07	2.955E-03
rs12999907	2	164957251	*AC092684.1*	A	G	0.0467	5.496E-10	6.652E-03	1.254E-07
rs2571445	2	218683154	*TNS1*	A	G	0.0427	2.739E-05	3.214E-05	2.086E-01
rs12500824	4	77416627	*SHROOM3*	A	G	0.0293	2.018E-04	1.381E-04	2.049E-01
rs2306556	4	156638573	*GUCY1A3*	A	G	0.0673	3.738E-04	5.330E-05	4.855E-01
rs7696431	4	169687725	*PALLD*	T	G	0.0362	3.755E-04	7.901E-01	2.143E-06
rs17080091	6	150997401	*PLEKHG1*	C	T	0.0632	9.188E-07	1.613E-03	2.379E-03
rs2107595	7	19049388	*HDAC9*	A	G	0.0734	5.911E-12	1.441E-01	4.923E-24
rs11556924	7	129663496	*ZC3HC1*	C	T	0.0726	4.387E-02	1.954E-04	3.942E-01
rs3918226	7	150690176	*NOS3*	T	C	0.1333	1.312E-04	3.856E-07	7.673E-01
rs1887318	10	30321598	*KIAA1462*	T	C	0.0624	2.333E-03	6.886E-05	1.652E-14
rs10840293	11	9751196	*SWAP70*	A	G	0.0547	1.255E-05	2.922E-03	8.880E-03
rs10841443	12	20220033	*RP11-664H17.1*	G	C	0.0507	4.343E-01	2.268E-04	1.376E-05
rs2681472	12	90008959	*ATP2B1*	G	A	0.0741	4.538E-17	1.334E-09	2.131E-06
rs1317507	13	113631780	*MCF2L*	A	C	0.0383	2.553E-05	2.351E-01	2.596E-05
rs10139550	14	100145710	*HHIPL1*	G	C	0.0554	4.917E-07	9.118E-01	1.353E-10
rs2071382	15	91428197	*FES*	T	C	0.0535	1.241E-09	2.491E-04	5.027E-05
rs7500448	16	83045790	*CDH13*	A	G	0.0555	5.865E-05	1.217E-03	5.010E-16
rs17608766	17	45013271	*GOSR2*	C	T	0.0530	1.717E-08	5.656E-01	2.643E-14
rs8068952	17	59286644	*BCAS3*	G	C	0.0339	7.789E-05	2.110E-03	3.874E-02

rs8108632	19	41854534	*TGFB1*	T	A	0.0515	2.089E-01	1.790E-02	9.288E-05
rs867186	20	33764554	*PROCR*	A	G	0.0607	6.955E-03	1.613E-01	1.099E-06
rs260020	20	57714025	*ZNF831*	T	C	0.0475	2.819E-06	1.570E-07	3.267E-01
rs28451064	21	35593827	*AP000318.2*	A	G	0.1276	4.984E-02	2.391E-02	6.611E-06

B. $GRS_{CAD\text{-}lipids}$

SNP	CHR	POS	Locus	Risk Allele	Other Allele	Weight	P-value			
							HDL	LDL	TG	TC
rs11591147	1	55505647	*PCSK9*	G	T	0.2565	4.257E-04	1.709E-228	6.362E-01	6.206E-157
rs7528419	1	109817192	*CELSR2*	A	G	0.1145	1.360E-31	0.000E+00	9.278E-05	1.898E-296
rs585967	2	21270554	*APOB*	C	A	0.0731	9.090E-01	5.800E-152	9.941E-01	1.310E-111
rs4299376	2	44072576	*ABCG8*	G	T	0.0508	1.274E-01	1.698E-72	3.044E-05	1.546E-57
rs7568458	2	85788175	*GGCX*	A	T	0.0596	4.813E-02	3.437E-04	3.601E-03	1.095E-04
rs1591805	6	126717064	*RP11-394G3.2*	A	G	0.0402	1.937E-02	1.911E-01	1.030E-04	1.230E-01
rs10455872	6	161010118	*LPA*	G	A	0.3186	6.654E-04	6.167E-50	2.577E-02	7.345E-37
rs6997340	8	18286997	*NAT2*	T	C	0.0324	6.348E-01	1.383E-02	5.748E-20	6.245E-13
rs2083636	8	19865263	*LPL*	T	G	0.0514	4.002E-246	6.200E-04	2.792E-295	2.188E-04
rs2891168	9	22098619	*CDKN2B-AS1*	G	A	0.1934	5.424E-01	2.339E-10	4.376E-02	4.920E-09
rs507666	9	136149399	*ABO*	A	G	0.0788	2.257E-03	1.706E-80	6.262E-01	2.292E-72
rs2246942	10	91004886	*LIPA*	G	A	0.0662	2.049E-02	1.249E-03	2.957E-02	1.694E-04
rs11601507	11	5701074	*TRIM5*	A	C	0.0734	6.643E-04	6.652E-08	8.593E-01	4.905E-05
rs2244608	12	121416988	*HNF1A*	G	A	0.0476	1.518E-03	1.963E-23	3.819E-01	1.206E-25
rs247616	16	56989590	*AC012181.1*	C	T	0.0312	0.000E+00	3.397E-01	2.267E-15	9.586E-64
rs6511720	19	11202306	*LDLR*	G	T	0.1253	1.090E-05	1.111E-260	4.610E-01	6.767E-182
rs6102343	20	39924279	*ZHX3*	A	G	0.0451	6.733E-02	2.048E-04	1.345E-05	2.978E-08

C. $GRS_{CAD\text{-}BMI}$

SNP	CHR	POS	Locus	Risk Allele	Other Allele	Weight	BMI P-value
rs2820315	1	201872264	*LMOD1*	T	C	0.0467	3.039E-39
rs7116641	11	43696917	*AC007319.1*	G	T	0.0304	2.259E-34
rs11170820	12	54513915	*RHOA*	G	C	0.0908	5.933E-06
rs9591012	13	33058333	*STAG1*	G	A	0.0457	1.136E-15
rs13723	17	27941886	*MRAS*	G	A	0.0374	1.834E-09
rs663129	18	57838401	*ARHGEF26*	A	G	0.0582	2.730E-191
rs840616	2	188196469	*UNC5C*	C	T	0.0456	4.026E-08
rs667920	3	136069472	*KCNK5*	T	G	0.0393	2.884E-19
rs9818870	3	138122122	*MAD1L1*	T	C	0.0646	2.096E-09
rs12493885	3	153839866	*ZFPM2*	C	G	0.0661	2.808E-09
rs7623687	3	49448566	*DAB2IP*	A	C	0.0699	6.093E-07
rs3775058	4	96117371	*RP11-472I20.4*	A	T	0.0351	6.020E-07
rs56015508	6	39152041	*FLJ12825*	C	A	0.0630	1.452E-04
rs10267593	7	1937261	*N4BP2L2*	G	A	0.0418	6.131E-13
rs10093110	8	106565414	*CORO6*	G	A	0.0258	5.492E-06
rs885150	9	124420173	*RNU4-17P*	C	T	0.0389	2.903E-05

Table IV. Associations of comprehensive CAD genetic risk Score (CAD-GRS) and three CAD mediating trait-based sub-genetic risk scores (sub-GRSs) with incident CAD

A. UK Biobank (HR and 95% CI)

CAD-GRS	Comprehensive GRS	$GRS_{CAD\text{-}BP}$	$GRS_{CAD\text{-}lipids}$	$GRS_{CAD\text{-}BMI}$
per SD increase	1.336 (1.310, 1.363)	1.139 (1.117,1.161)	1.184 (1.162, 1.207)	1.076 (1.055, 1.098)
Low risk	Ref.	Ref.	Ref.	Ref.
Intermediate risk	1.436 (1.352, 1.526)	1.167 (1.105, 1.232)	1.208 (1.144, 1.276)	1.131 (1.072, 1.193)
High risk	2.196 (2.056, 2.345)	1.422 (1.335, 1.514)	1.567 (1.472, 1.668)	1.245 (1.169, 1.326)

B. MVP (OR and 95% CI)

CAD-GRS	Comprehensive GRS	$GRS_{CAD\text{-}BP}$	$GRS_{CAD\text{-}lipids}$	$GRS_{CAD\text{-}BMI}$
per SD increase	1.212 (1.184, 1.241)	1.074 (1.050,1.099)	1.134 (1.108, 1.160)	1.025 (1.003, 1.048)
Low risk	Ref.	Ref.	Ref.	Ref.
Intermediate risk	1.339 (1.254, 1.430)	1.092 (1.027, 1.161)	1.190 (1.118, 1.267)	1.036 (0.977, 1.098)
High risk	1.749 (1.622, 1.885)	1.217 (1.131, 1.309)	1.404 (1.304, 1.511)	1.074 (1.001, 1.153)

Table V. Combined associations of CAD-GRSs and smoking status with incident CAD

A. UK Biobank (HR and 95% CI)

		Smoking Status			P-value for Multiplicative Interaction
		Never	**Past**	**Current**	
CAD-GRS	Low Risk	Ref.	1.208 (1.074,1.360)	1.774 (1.503,2.094)	0.51
	Intermediate Risk	1.422 (1.295,1.562)	1.732 (1.577,1.902)	2.638 (2.365,2.942)	
	High Risk	2.264 (2.045,2.506)	2.536 (2.289,2.809)	4.077 (3.586,4.636)	
GRS_{CAD-BP}	Low Risk	Ref.	1.107 (0.998, 1.228)	1.669 (1.443, 1.931)	0.17
	Intermediate Risk	1.110 (1.021, 1.206)	1.351 (1.243, 1.468)	1.964 (1.776, 2.172)	
	High Risk	1.345 (1.221, 1.481)	1.594 (1.448, 1.755)	2.664 (2.352, 3.018)	
$GRS_{CAD-lipids}$	Low Risk	Ref.	1.237 (1.113, 1.376)	1.947 (1.682, 2.253)	0.50
	Intermediate Risk	1.222 (1.121, 1.332)	1.487 (1.364, 1.621)	2.322 (2.097, 2.572)	
	High Risk	1.658 (1.505, 1.827)	1.878 (1.702, 2.071)	2.861 (2.512, 3.259)	
$GRS_{CAD-BMI}$	Low Risk	Ref.	1.185 (1.070, 1.313)	1.719 (1.488, 1.986)	0.95
	Intermediate Risk	1.109 (1.022, 1.205)	1.351 (1.244, 1.468)	2.006 (1.818, 2.214)	
	High Risk	1.229 (1.116, 1.355)	1.467 (1.332, 1.616)	2.255 (1.978, 2.571)	

B. MVP (OR and 95% CI)

		Smoking Status			P-value for Multiplicative Interaction
		Never	**Past**	**Current**	
CAD-GRS	Low Risk	Ref.	1.130 (0.988,1.293)	1.465 (1.229,1.747)	0.26
	Intermediate Risk	1.296 (1.147,1.464)	1.560 (1.390,1.752)	1.890 (1.657,2.156)	
	High Risk	1.834 (1.598,2.104)	2.003 (1.765,2.273)	2.255 (1.920,2.649)	
GRS_{CAD-BP}	Low Risk	Ref.	1.194 (1.054, 1.352)	1.517 (1.290, 1.784)	0.63
	Intermediate Risk	1.112 (0.991, 1.247)	1.317 (1.181, 1.468)	1.561 (1.377, 1.769)	
	High Risk	1.297 (1.133, 1.485)	1.452 (1.285, 1.640)	1.656 (1.410, 1.946)	
$GRS_{CAD-lipids}$	Low Risk	Ref.	1.282 (1.126, 1.459)	1.621 (1.371, 1.916)	0.16
	Intermediate Risk	1.276 (1.132, 1.438)	1.512 (1.349, 1.694)	1.781 (1.564, 2.029)	
	High Risk	1.616 (1.409, 1.855)	1.724 (1.520, 1.955)	2.052 (1.749, 2.408)	
$GRS_{CAD-BMI}$	Low Risk	Ref.	1.195 (1.065, 1.341)	1.239 (1.056, 1.454)	0.44
	Intermediate Risk	1.014 (0.911, 1.129)	1.219 (1.102, 1.348)	1.403 (1.248, 1.578)	
	High Risk	1.114 (0.979, 1.267)	1.224 (1.091, 1.373)	1.458 (1.253, 1.697)	

Table VI. Interaction between CAD-associated SNPs and current smoking on incident CAD in the UK Biobank

SNP	CHR	POS	Locus	EA	Beta Coefficients			P-value of Interaction	RERI	P-value of RERI
					SNP	Current smoking	Interaction			
rs11591147	1	55505647	*PCSK9*	G	0.209	-0.283	0.403	0.05	0.404	1.21E-38
rs73015714	19	17855763	*FCHO1*	G	0.046	0.460	0.123	0.01	0.245	1.84E-03
rs1870634	10	44480811	*LINC00841*	G	0.027	0.391	0.091	0.04	0.158	6.71E-03
rs3918226	7	150690176	*NOS3*	T	0.031	0.483	0.163	0.02	0.317	9.03E-03
rs12897	3	172115902	*FNDC3B*	G	0.033	0.451	0.082	0.05	0.159	1.21E-02
rs4643373	17	47123423	*IGF2BP1*	T	-0.012	0.379	0.095	0.03	0.138	2.07E-02
rs2145598	14	58794001	*ARID4A*	G	0.019	0.449	0.075	0.06	0.135	2.55E-02
rs867186	20	33764554	*PROCR*	A	0.026	0.312	0.110	0.14	0.173	4.24E-02
rs12493885	3	153839866	*ARHGEF26*	C	0.078	0.402	0.065	0.27	0.149	4.48E-02
rs2891168	9	22098619	*CDKN2B-AS1*	G	0.138	0.499	0.015	0.70	0.125	4.54E-02
rs17581137	15	96146414	*RP11-61O11.1*	A	0.009	0.396	0.078	0.10	0.126	4.61E-02
rs8108632	19	41854534	*TGFB1*	T	0.048	0.468	0.054	0.19	0.123	5.28E-02
rs10455872	6	161010118	*LPA*	G	0.202	0.506	0.046	0.50	0.242	6.43E-02
rs17080091	6	150997401	*PLEKHG1*	C	0.077	0.349	0.088	0.29	0.174	6.58E-02
rs944172	9	110517794	*AL162389.1*	C	0.022	0.475	0.065	0.14	0.124	7.27E-02
rs2229357	12	57843711	*INHBC*	G	0.029	0.421	0.061	0.20	0.114	7.82E-02
rs585967	2	21270554	*APOB*	C	-0.004	0.363	0.088	0.13	0.131	8.06E-02
rs2820315	1	201872264	*LMOD1*	T	0.032	0.477	0.055	0.19	0.114	8.58E-02
rs12801636	11	65391317	*PCNX3*	G	0.013	0.409	0.068	0.16	0.113	8.66E-02
rs9367716	6	57160572	*RNU7-66P*	G	0.013	0.426	0.062	0.16	0.107	8.85E-02
rs1508798	5	9556694	*RP11-260E18.1*	T	0.038	0.415	0.061	0.25	0.118	9.46E-02
rs10267593	7	1937261	*MAD1L1*	G	0.002	0.384	0.078	0.15	0.120	9.52E-02
rs6700559	1	200646073	*RP11-92G12.3*	C	0.038	0.601	-0.082	0.04	-0.117	9.54E-02
rs216172	17	2126504	*SMG6*	C	0.037	0.571	-0.081	0.05	-0.115	1.01E-01

rs1892094	1	169094459	*ATP1B1*	C	0.013	0.461	0.055	0.17	0.099	1.03E-01
rs742115	6	11327021	*NEDD9*	C	0.003	0.452	0.059	0.15	0.097	1.13E-01
rs4918072	10	105693644	*STN1*	A	0.038	0.488	0.049	0.27	0.110	1.27E-01
rs1317507	13	113631780	*MCF2L*	A	0.044	0.557	-0.085	0.07	-0.115	1.32E-01
rs35879803	4	146782837	*ZNF827*	C	0.003	0.441	0.056	0.19	0.092	1.37E-01
rs11170820	12	54513915	*FLJ12825*	G	0.087	0.504	0.084	0.32	0.217	1.55E-01
rs2493298	1	3325912	*PRDM16*	A	0.048	0.541	-0.098	0.10	-0.133	1.59E-01
rs3775058	4	96117371	*UNC5C*	A	0.035	0.549	-0.080	0.10	-0.111	1.61E-01
rs13723	17	27941886	*CORO6*	G	0.035	0.583	-0.070	0.08	-0.096	1.63E-01
rs11509880	7	12261911	*TMEM106B*	A	0.007	0.479	0.052	0.21	0.092	1.63E-01
rs9591012	13	33058333	*N4BP2L2*	G	0.006	0.450	0.050	0.24	0.084	1.74E-01
rs2306556	4	156638573	*GUCY1A3*	A	0.074	0.679	-0.101	0.05	-0.129	1.87E-01
rs17608766	17	45013271	*GOSR2*	C	0.009	0.494	0.067	0.23	0.120	1.88E-01
rs663129	18	57838401	*RNU4-17P*	A	0.022	0.546	-0.068	0.15	-0.099	1.90E-01
rs56170783	1	57016131	*PLPP3*	A	0.159	0.823	-0.169	0.02	-0.194	1.94E-01
rs10953541	7	107244545	*BCAP29*	C	0.007	0.436	0.052	0.27	0.087	1.95E-01
rs3936511	5	55860781	*C5orf67*	G	0.038	0.544	-0.078	0.13	-0.106	1.97E-01
rs72627509	4	57839051	*NOA1*	G	0.012	0.492	0.058	0.26	0.107	1.98E-01
rs6905288	6	43758873	*VEGFA*	A	0.062	0.490	0.021	0.60	0.078	2.10E-01
rs6997340	8	18286997	*NAT2*	T	0.024	0.550	-0.064	0.16	-0.093	2.10E-01
rs7947761	11	100624599	*ARHGAP42*	G	0.031	0.553	-0.065	0.14	-0.090	2.11E-01
rs1867624	17	62387091	*RPL31P57*	T	0.033	0.593	-0.064	0.12	-0.089	2.11E-01
rs10512861	3	132257961	*DNAJC13*	G	0.034	0.669	-0.089	0.13	-0.140	2.14E-01
rs1887318	10	30321598	*KIAA1462*	T	0.025	0.482	0.035	0.38	0.076	2.26E-01
rs10774625	12	111910219	*ATXN2*	A	0.033	0.484	0.030	0.45	0.073	2.42E-01
rs60154123	1	210468999	*RP4-667H12.4*	T	0.063	0.502	0.034	0.53	0.102	2.47E-01
rs7116641	11	43696917	*RP11-472I20.4*	G	0.018	0.489	0.039	0.36	0.078	2.47E-01
rs748431	3	14928077	*FGD5*	G	0.033	0.556	-0.058	0.17	-0.077	2.71E-01

rs2571445	2	218683154	*TNS1*	A	0.010	0.484	0.038	0.35	0.069	2.75E-01
rs3851738	16	75387533	*CFDP1*	C	0.039	0.585	-0.059	0.15	-0.076	2.76E-01
rs3827066	20	44586023	*ZNF335*	T	0.029	0.535	-0.071	0.22	-0.099	2.77E-01
rs28451064	21	35593827	*AP000318.2*	A	0.112	0.510	0.016	0.78	0.110	2.80E-01
rs11556924	7	129663496	*ZC3HC1*	C	0.023	0.475	0.031	0.45	0.066	2.84E-01
rs7199941	16	81906423	*PLCG2*	A	0.036	0.558	-0.057	0.17	-0.073	2.88E-01
rs7528419	1	109817192	*CELSR2*	A	0.093	0.656	-0.090	0.06	-0.091	2.96E-01
rs6689306	1	154395946	*IL6R*	A	0.025	0.557	-0.051	0.21	-0.071	3.00E-01
rs1351525	11	13301548	*ARNTL*	T	0.006	0.460	0.039	0.38	0.066	3.10E-01
rs1591805	6	126717064	*RP11-394G3.2*	A	0.041	0.571	-0.056	0.16	-0.069	3.13E-01
rs7412	19	45412079	*APOE*	C	0.060	0.439	0.041	0.59	0.103	3.15E-01
rs11191416	10	104604916	*PFN1P11*	T	-0.003	0.397	0.064	0.38	0.096	3.24E-01
rs6841581	4	148401190	*EDNRA*	A	0.042	0.504	0.035	0.54	0.089	3.42E-01
rs11723436	4	120901336	*RP11-170N16.1*	G	0.034	0.499	0.024	0.57	0.064	3.48E-01
rs77335401	5	131759825	*C5orf56*	C	0.017	0.529	-0.062	0.32	-0.091	3.51E-01
rs11601507	11	5701074	*TRIM5*	A	0.045	0.508	0.048	0.53	0.115	3.71E-01
rs16844401	4	3449652	*HGFAC*	A	0.012	0.506	0.064	0.42	0.119	3.73E-01
rs72743461	15	67441750	*SMAD3*	C	0.032	0.475	0.025	0.60	0.063	3.77E-01
rs7623687	3	49448566	*RHOA*	A	0.033	0.624	-0.064	0.26	-0.090	3.83E-01
rs10857147	4	81181072	*RP11-576N17.4*	T	0.040	0.502	0.020	0.65	0.061	3.83E-01
rs10840293	11	9751196	*SWAP70*	A	0.038	0.495	0.017	0.67	0.054	3.90E-01
rs76954792	17	30033514	*RP11-805L22.1*	T	0.027	0.537	-0.048	0.31	-0.064	4.09E-01
rs4752700	10	124237612	*HTRA1*	G	0.040	0.556	-0.049	0.23	-0.056	4.10E-01
rs840616	2	188196469	*AC007319.1*	C	0.032	0.489	0.019	0.66	0.053	4.19E-01
rs56015508	6	39152041	*KCNK5*	C	0.012	0.465	0.031	0.53	0.058	4.25E-01
rs885150	9	124420173	*DAB2IP*	C	0.029	0.538	-0.046	0.31	-0.058	4.31E-01
rs1800449	5	121413208	*LOX*	T	0.002	0.528	-0.041	0.45	-0.067	4.33E-01
rs11617955	13	110818102	*COL4A1*	T	0.039	0.461	0.030	0.65	0.073	4.33E-01

rs8068952	17	59286644	*BCAS3*	G	0.048	0.539	-0.056	0.26	-0.063	4.35E-01
rs13003675	2	233584109	*GIGYF2*	T	0.019	0.542	-0.037	0.37	-0.050	4.58E-01
rs246600	5	142516897	*ARHGAP26*	T	0.041	0.504	0.011	0.79	0.046	4.68E-01
rs4299376	2	44072576	*ABCG8*	G	0.031	0.504	0.016	0.70	0.049	4.75E-01
rs10417115	19	33386556	*CEP89*	C	0.016	0.508	0.054	0.54	0.105	4.76E-01
rs260020	20	57714025	*ZNF831*	T	0.041	0.508	0.024	0.69	0.069	4.87E-01
rs61776719	1	38461319	*SF3A3*	A	0.021	0.554	-0.035	0.38	-0.046	4.94E-01
rs116843064	19	8429323	*ANGPTL4*	G	0.082	0.416	0.050	0.74	0.128	4.94E-01
rs17680741	10	82251514	*TSPAN14*	T	0.042	0.499	0.011	0.81	0.046	4.98E-01
rs12500824	4	77416627	*SHROOM3*	A	0.025	0.502	0.016	0.70	0.043	5.09E-01
rs7797644	7	6486067	*DAGLB*	C	0.035	0.491	0.015	0.76	0.048	5.10E-01
rs1321309	6	36638636	*LAP3P2*	A	0.030	0.502	0.012	0.77	0.041	5.19E-01
rs7500448	16	83045790	*CDH13*	A	0.029	0.490	0.016	0.73	0.046	5.19E-01
rs2954029	8	126490972	*RP11-136O12.2*	A	0.025	0.500	0.013	0.74	0.040	5.27E-01
rs35541991	6	22583856	*RP1-309H15.2*	C	0.024	0.504	0.015	0.73	0.041	5.39E-01
rs12202017	6	134173151	*TARID*	A	0.040	0.503	0.008	0.86	0.041	5.51E-01
rs61848342	10	12303813	*RN7SL232P*	C	0.056	0.514	-0.001	0.99	0.038	5.72E-01
rs2107732	7	45077978	*CCM2*	G	0.072	0.639	-0.068	0.32	-0.068	5.78E-01
rs2244608	12	121416988	*HNF1A*	G	0.062	0.545	-0.048	0.27	-0.039	5.81E-01
rs4613862	6	82612271	*RP11-379B8.1*	A	0.024	0.503	0.011	0.78	0.034	5.86E-01
rs2972146	2	227100698	*NEU2*	T	0.012	0.549	-0.027	0.52	-0.037	5.91E-01
rs11677932	2	238223955	*STK25*	G	0.043	0.509	0.004	0.92	0.036	5.93E-01
rs975722	7	117332914	*CFTR*	G	0.025	0.538	-0.031	0.45	-0.035	5.98E-01
rs11810571	1	151762308	*TDRKH*	G	0.023	0.485	0.017	0.76	0.043	6.03E-01
rs2107595	7	19049388	*HDAC9*	A	0.071	0.515	-0.001	0.98	0.047	6.05E-01
rs7617773	3	48193515	*TKT*	T	-0.007	0.538	-0.018	0.66	-0.036	6.08E-01
rs590121	11	75274150	*SERPINH1*	T	0.015	0.529	-0.027	0.55	-0.035	6.24E-01
rs2681472	12	90008959	*ATP2B1*	G	0.036	0.511	0.009	0.86	0.040	6.39E-01

rs10841443	12	20220033	*RP11-664H17.1*	G	0.036	0.508	0.004	0.92	0.032	6.39E-01
rs11838267	12	7175872	*C1S*	T	0.026	0.488	0.015	0.80	0.042	6.40E-01
rs2832227	21	30533076	*MAP3K7CL*	G	0.063	0.515	-0.002	0.97	0.040	6.54E-01
rs247616	16	56989590	*AC012181.1*	C	0.017	0.548	-0.025	0.56	-0.032	6.55E-01
rs699	1	230845794	*AGT*	G	0.037	0.513	0.002	0.97	0.028	6.63E-01
rs2839812	11	103673294	*RP11-563P16.1*	T	0.056	0.537	-0.041	0.36	-0.031	6.66E-01
rs616381	2	45891708	*PRKCE*	A	0.030	0.509	0.004	0.92	0.027	6.69E-01
rs2083460	15	89574484	*RP11-326A19.2*	T	0.051	0.500	0.008	0.92	0.048	6.75E-01
rs10139550	14	100145710	*HHIPL1*	G	0.045	0.516	-0.002	0.95	0.027	6.84E-01
rs7568458	2	85788175	*GGCX*	A	0.041	0.516	-0.001	0.97	0.025	6.93E-01
rs17843797	3	124453022	*UMPS*	G	0.057	0.515	-0.001	0.99	0.038	6.95E-01
rs6984210	8	22033615	*BMP1*	G	0.092	0.515	-0.004	0.96	0.058	6.96E-01
rs507666	9	136149399	*ABO*	A	0.008	0.509	0.016	0.76	0.032	7.06E-01
rs9501744	6	1617143	*FOXC1*	C	0.040	0.507	0.004	0.94	0.035	7.06E-01
rs6102343	20	39924279	*ZHX3*	A	0.029	0.511	0.005	0.92	0.028	7.06E-01
rs2083636	8	19865263	*LPL*	T	0.021	0.503	0.008	0.87	0.027	7.06E-01
rs4266144	3	156852592	*SPTSSB*	G	0.035	0.513	0.001	0.97	0.027	7.10E-01
rs7696431	4	169687725	*PALLD*	T	0.018	0.537	-0.022	0.59	-0.024	7.11E-01
rs9897596	17	17593453	*RAI1*	T	0.003	0.503	0.012	0.77	0.022	7.33E-01
rs16986953	2	19942473	*AC019055.1*	A	0.040	0.513	0.008	0.92	0.042	7.46E-01
rs3832966	14	75614504	*TMED10*	ACCCG	0.048	0.522	-0.007	0.85	0.020	7.52E-01
rs2074158	17	40257163	*DHX58*	C	0.029	0.513	0.002	0.97	0.024	7.80E-01
rs180803	22	24658858	*POM121L9P*	G	0.058	0.669	-0.078	0.68	-0.099	7.83E-01
rs964184	11	116648917	*ZPR1*	G	0.020	0.512	0.007	0.90	0.026	7.84E-01
rs9818870	3	138122122	*MRAS*	T	0.046	0.516	-0.005	0.93	0.023	8.00E-01
rs17678683	2	145286559	*ZEB2*	G	0.018	0.513	0.008	0.91	0.026	8.16E-01
rs7633770	3	46688562	*SNORD77*	A	0.039	0.520	-0.007	0.86	0.015	8.18E-01
rs11099493	4	82587050	*RASGEF1B*	A	0.005	0.505	0.007	0.87	0.015	8.28E-01

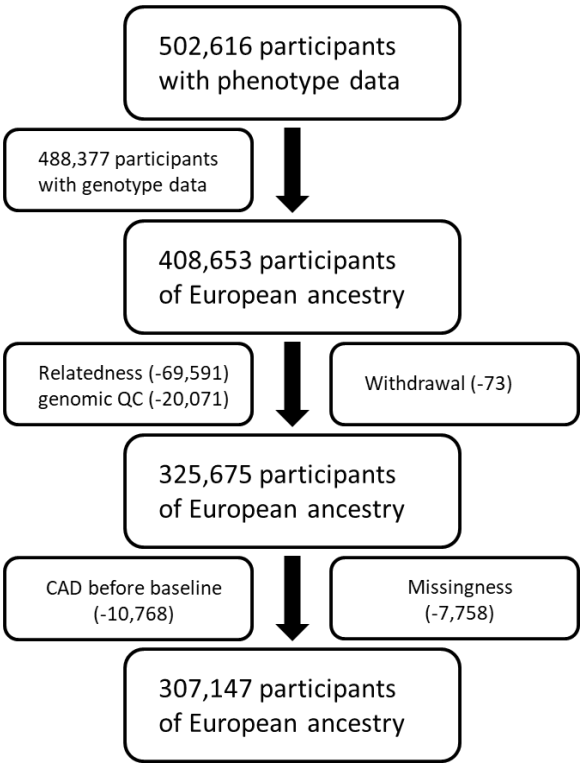
rs7164479	15	79123054	*MORF4L1*	T	0.063	0.553	-0.034	0.41	-0.014	8.39E-01
rs11057401	12	124427306	*CCDC92*	T	0.019	0.514	0.000	0.99	0.013	8.45E-01
rs4472337	6	34769765	*UHRF1BP1*	T	0.043	0.522	-0.026	0.63	-0.016	8.57E-01
rs9964304	18	47229717	*RP11-813F20.2*	C	0.072	0.527	-0.022	0.63	0.012	8.68E-01
rs3130683	6	31888367	*C2*	T	0.011	0.536	-0.013	0.82	-0.015	8.79E-01
rs7306455	12	95355541	*NDUFA12*	G	0.076	0.557	-0.024	0.72	0.014	8.95E-01
rs111245230	9	113169775	*SVEP1*	C	0.071	0.517	-0.041	0.70	-0.023	8.96E-01
rs112635299	14	94838142	*SERPINA1*	G	0.138	0.674	-0.082	0.59	-0.034	8.96E-01
rs11806316	1	115753482	*RP4-663N10.1*	G	0.034	0.537	-0.018	0.65	-0.008	9.04E-01
rs12999907	2	164957251	*AC092684.1*	A	0.069	0.554	-0.024	0.65	0.009	9.17E-01
rs667920	3	136069472	*STAG1*	T	0.023	0.536	-0.014	0.77	-0.008	9.18E-01
rs17517928	2	216291359	*FN1*	C	0.019	0.520	-0.004	0.94	0.007	9.24E-01
rs6494488	15	65024204	*RBPMS2*	A	0.031	0.545	-0.018	0.76	-0.009	9.25E-01
rs1050362	16	72130815	*DHX38*	A	0.008	0.519	-0.007	0.87	-0.006	9.25E-01
rs11830157	12	118265441	*KSR2*	G	0.011	0.515	-0.001	0.99	0.006	9.26E-01
rs2071382	15	91428197	*FES*	T	0.072	0.540	-0.026	0.52	0.006	9.26E-01
rs6511720	19	11202306	*LDLR*	G	0.075	0.578	-0.036	0.57	-0.007	9.47E-01
rs10237377	7	139757136	*PARP12*	G	0.021	0.523	-0.006	0.88	0.004	9.53E-01
rs67180937	1	222823743	*MIA3*	G	0.048	0.540	-0.017	0.71	0.004	9.58E-01
rs36096196	1	2252205	*MORN1*	T	0.064	0.522	-0.023	0.68	0.004	9.65E-01
rs1924981	13	29022645	*FLT1*	T	0.005	0.515	-0.001	0.98	0.001	9.83E-01
rs2246942	10	91004886	*LIPA*	G	0.022	0.520	-0.008	0.84	0.001	9.90E-01
rs9349379	6	12903957	*PHACTR1*	G	0.070	0.538	-0.029	0.48	-0.001	9.91E-01
rs114123510	2	203831212	*CARF*	A	0.062	0.521	-0.024	0.69	0.001	9.94E-01
rs10093110	8	106565414	*ZFPM2*	G	0.021	0.524	-0.008	0.84	0.000	9.99E-01

Table VII. Combined associations of CAD-GRSs and pack-year with incident CAD in the UK Biobank

		Smoking Pack-year			P-value for Multiplicative Interaction
		Low	**Medium**	**High**	
CAD-GRS	Low Risk	Ref.	1.292 (1.007,1.658)	1.827 (1.457,2.291)	0.32
	Intermediate Risk	1.575 (1.270,1.952)	1.963 (1.594,2.418)	2.421 (1.973,2.971)	
	High Risk	2.406 (1.907,3.035)	2.679 (2.144,3.347)	3.752 (3.033,4.641)	
GRS_{CAD-BP}	Low Risk	Ref.	1.296 (1.035, 1.622)	1.790 (1.458, 2.199)	0.15
	Intermediate Risk	1.302 (1.069, 1.586)	1.671 (1.382, 2.020)	2.039 (1.693, 2.456)	
	High Risk	1.755 (1.408, 2.189)	1.821 (1.473, 2.252)	2.618 (2.150, 3.188)	
$GRS_{CAD-lipids}$	Low Risk	Ref.	1.256 (1.015, 1.553)	1.573 (1.293, 1.913)	0.94
	Intermediate Risk	1.209 (1.003, 1.457)	1.456 (1.216, 1.743)	1.938 (1.627, 2.308)	
	High Risk	1.423 (1.145, 1.769)	1.813 (1.485, 2.213)	2.418 (2.004, 2.918)	
$GRS_{CAD-BMI}$	Low Risk	Ref.	1.222 (0.986, 1.514)	1.776 (1.461, 2.159)	0.64
	Intermediate Risk	1.197 (0.993, 1.444)	1.449 (1.210, 1.736)	1.871 (1.569, 2.231)	
	High Risk	1.270 (1.018, 1.586)	1.629 (1.330, 1.994)	2.052 (1.696, 2.482)	

Figure I. Study population QC process

A. UK Biobank



B. The MVP Cohort

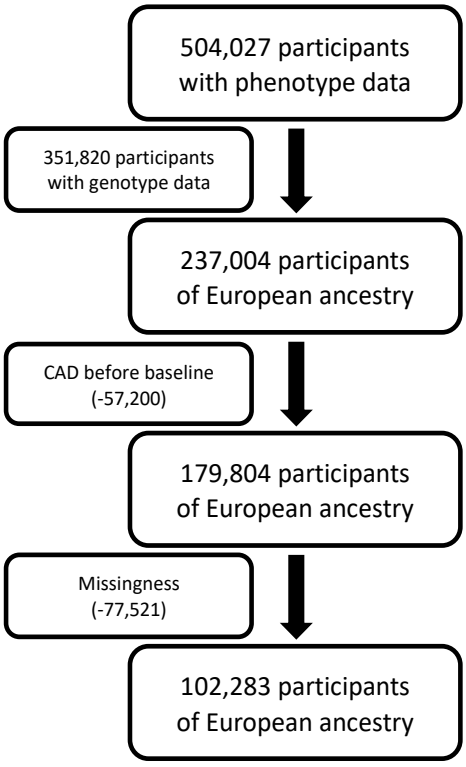
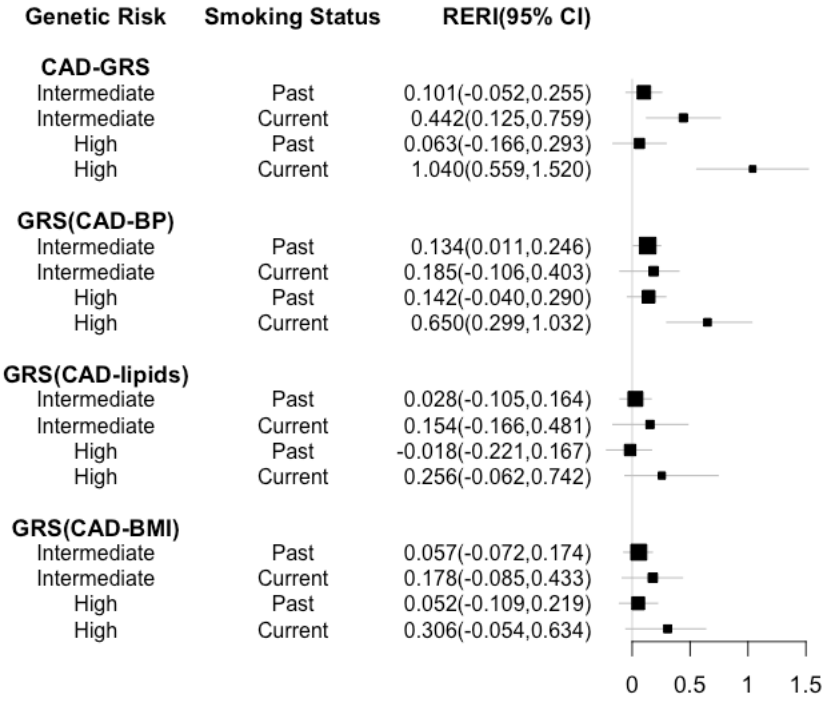


Figure II. Relative excess risk due to interaction (RERI) for CAD-GRSs and smoking status on incident CAD (Never smokers with low genetic risk were used as reference group for each score; 95% CI: 95% bootstrap confidence interval)

A. UK Biobank



B. MVP

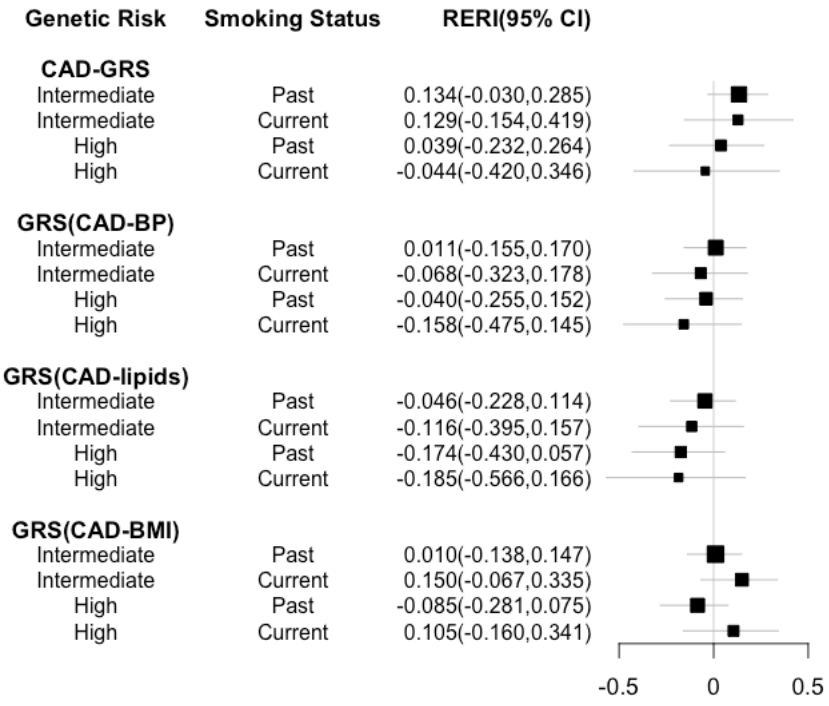
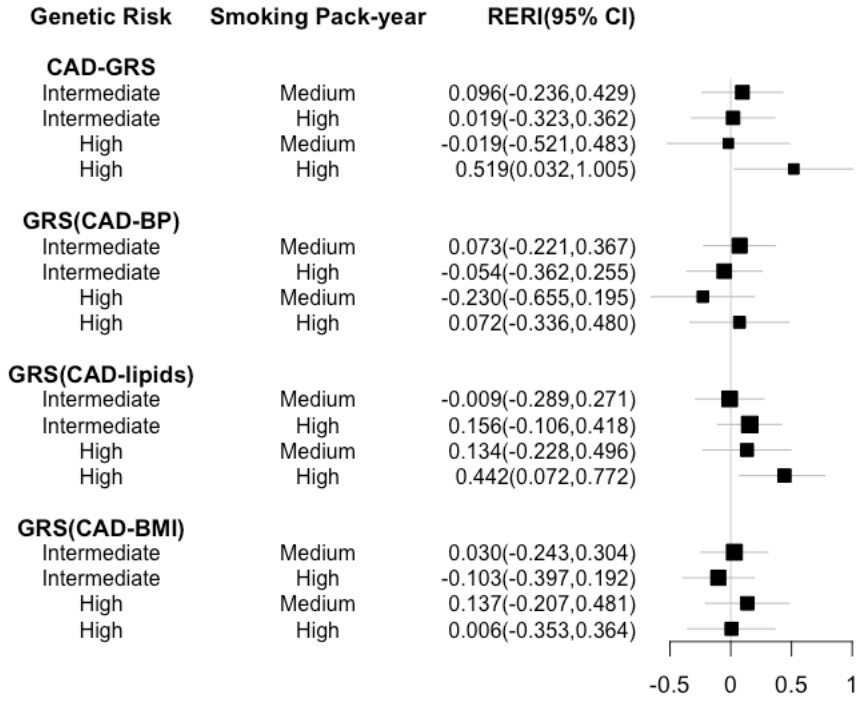

Genetic Risk
Smoking Status
RERI(95% CI)
CAD-GRS
Intermediate Past 0.134(-0.030,0.285)
Intermediate Current 0.129(-0.154,0.419)
High Past 0.039(-0.232,0.264)
High Current -0.044(-0.420,0.346)
GRS(CAD-BP)
Intermediate Past 0.011(-0.155,0.170)
Intermediate Current -0.068(-0.323,0.178)
High Past -0.040(-0.255,0.152)
High Current -0.158(-0.475,0.145)
GRS(CAD-lipids)
Intermediate Past -0.046(-0.228,0.114)
Intermediate Current -0.116(-0.395,0.157)
High Past -0.174(-0.430,0.057)
High Current -0.185(-0.566,0.166)
GRS(CAD-BMI)
Intermediate Past 0.010(-0.138,0.147)
Intermediate Current 0.150(-0.067,0.335)
High Past -0.085(-0.281,0.075)
High Current 0.105(-0.160,0.341)
-0.5
0
0.5

Figure III. Relative excess risk due to interaction (RERI) for CAD-GRSs and pack-year on incident CAD in the UK Biobank (Ever smokers with low pack-year and low genetic risk were used as reference group for each score; 95% CI: 95% bootstrap confidence interval)



Chapter 3

Assessment of gene-physical activity interaction in coronary artery diseases

Introduction

Coronary artery disease (CAD) is the pre-eminent cause of death and lifestyle modification has played a major role in CAD prevention. (Lloyd-Jones, Hong et al. 2010) Physical activity is reported to be an independent and protective risk factor associated with CAD morbidity and mortality. (Winzer, Woitek et al. 2018) A standardized case-control study of acute myocardial infarction in 52 countries has shown a 14% risk reduction due to regular physical activity and a 12% population attributable risk explained by physical activity. (Yusuf, Hawken et al. 2004) In addition, the Health Professional's Study (Tanasescu, Leitzmann et al. 2002) including 44,452 men has shown a 30% reduction in the risk of myocardial infarction, providing strong evidence for the cardiovascular benefits of exercise in primary prevention. A meta-analysis of exercise-based cardiac rehabilitation including trials conducted in the contemporary medication/intervention era, estimated a reduction in mortality of 20–32%. (Taylor, Brown et al. 2004) Physical activity can indirectly decrease CAD risk by providing a gateway through which other risk factors such as cholesterol, blood pressure and obesity can be favorably modified. However, at least 40% of the risk reduction due to exercise cannot be explained by such indirect effect through conventional CAD risk factors. Thus, a cardio-protective ‘vascular conditioning’ effect, including enhanced nitric oxide vasodilator function, improved vascular reactivity, altered vascular structure or combinations thereof, has been proposed. (Wilson, Ellison et al. 2016) In addition, exercise is reported to be a potent stimulator activating numerous downstream cascades at a molecular

and cellular level, that if sustained and intensive enough enables gross anatomical remodeling capable of enhancing functional capacity in all spectrums of the population including the casual exercisers, the sedentary individuals or those with established cardiovascular diseases.

CAD is a heritable condition with estimated heritability of 50% to 60%. (Dai, Wiernek et al. 2016) The interplay between genetic factors and environmental factors is an important part in the development of CAD. It has been proven that physical activity is beneficial for cardiovascular health, however, how the reduced risk due to physical activity interact with individual's genetic background in CAD remains unclear. Studies have taken physical activity into account when constructing a composite score for healthy lifestyle and reported a uniform benefit of such healthy lifestyle on cardiovascular health across individuals with different levels of genetic risk. (Khera, Emdin et al. 2016, Pazoki, Dehghan et al. 2018) A recent study conducted in the UK Biobank also reported similar benefit of physical activity for coronary heart disease across different genetic risk strata. (Tikkanen, Gustafsson et al. 2018) However, current evidences for gene-physical activity interaction in CAD have only focused on multiplicative scale which ignored the potential interaction effect on additive scale, even though additive scale is useful for assessing the public health importance of interventions and the public health significance of such interaction effects. (VanderWeele and Knol 2014) In addition, super-additive interaction effects if unconfounded under monotonicity assumptions can be more closely linked with mechanistic effects using a sufficient component cause model. (VanderWeele 2009) Genetic mechanisms of CAD can be mediated through different traits including cholesterol and other lipid levels, obesity, and blood pressure level (Webb, Erdmann et al. 2017), and physical activity might interact with certain mediating trait-based pathways. However, no study has been conducted to assess how the genetic predisposition of CAD driven by these mediating traits can modify the reduced

risk due to physical activity. Therefore, the purpose of this study is to assess gene-physical activity interaction for CAD on both additive and multiplicative scale in populations of European ancestry using data from two of the largest biobank cohorts.

Methods

Study populations (same as Chapter 2)

The primary study population consists of participants with European ancestry from the UK Biobank (https://www.ukbiobank.ac.uk/) cohort. The UK Biobank is a major national and international health resource, and a registered charity in its own right, with the aim of improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Both genetic and phenotypic data for all participants in the UK Biobank were obtained for this study. The UK Biobank genetic data contains genome-wide genotypes for 488,377 participants. (Bycroft, Freeman et al. 2018) These were assayed using two very similar genotyping arrays. A subset of 49,950 participants involved in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study were genotyped at 807,411 markers using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix (now part of Thermo Fisher Scientific), and 438,427 participants were genotyped using the closely related Applied Biosystems UK Biobank Axiom Array (825,927 markers) that shares 95% of marker content with the UK BiLEVE Axiom Array. A quality control pipeline was developed and applied specifically to accommodate the large-scale dataset of ethnically diverse participants, genotyped in many

batches, using two slightly different arrays, and which will be used by many researchers to tackle a wide variety of research questions. Markers that passed the quality control check were imputed using the Haplotype Reference Consortium (HRC) reference panel as well as the merged UK10K and 1000 Genomes phase 3 reference panels. Information was then combined using the HRC data as the primary resource. For phenotype data, participants provided electronic signed consent, answered questions on socio-demographic, lifestyle and health-related factors, and completed a range of physical measures at baseline recruitment. All participants also provided consent for follow-up through linkage to their health-related records including in-patient hospital episode statistics and national death registry data.

The replication population of this study includes participants of European ancestry from the Million Veteran Program (MVP). The MVP is a national, voluntary research program funded entirely by the Department of Veterans Affairs (VA) Office of Research & Development. It is envisioned as a VA-based mega-biobank and launched to establish a national, representative, and longitudinal study of veterans for genomic and non-genomic research that combines data from survey instruments, the electronic health record and biospecimens. (Gaziano, Concato et al. 2016) The source population is defined as active users of the Veterans Health Administration (VHA), with the ability to provide informed consent as the only inclusion criterion. Recruitment is currently occurring in person at selected sites in the VHA health care system. Every Veteran is assigned a study ID number, which is used to track them throughout the entire process of recruitment, enrollment, sample collection and use. During recruitment veteran participants were informed about the MVP study via an invitation letter, explaining that participation in the study involves completing questionnaires, providing a blood sample for future research, allowing ongoing access to medical records and other health administrative data by authorized MVP staff, and agreeing

to future contact by MVP staff for follow-up studies. The present study included a recent data release in 2018 containing genotype and phenotype data of over 460,000 participants among whom ~370,000 identified as non-Hispanic White. Genome-wide genotype data was measured using a customized Affymetrix Axiom biobank array, the MVP 1.0 Genotyping Array. With 723,305 total DNA sequence variants, the array is enriched for both common and rare variants of clinical importance in different ethnic backgrounds. Genotyped variants that were poorly called (genotype missingness > 5%) or that deviated from their expected allele frequency based on reference data from the 1000 Genomes Project were excluded. The remaining variants were used to conduct genotype imputation based on the 1000 Genomes Project phase 3, v.5 reference panel, which generated a total number of > 30 million variants. For phenotype data, participants were asked to complete two surveys: the MVP Baseline Survey and the MVP Lifestyle Survey. The MVP Baseline Survey was designed to collect information regarding demographics, family pedigree, health status, lifestyle habits, military experience, medical history, family history of specific illnesses, and physical features. The MVP Lifestyle Survey contains questions from validated instruments in domains selected to provide information on sleep and exercise habits, environmental exposures, dietary habits, and sense of wellbeing. Other health-related information or disease diagnosis data is collected through linkage to participants' VA electronic health record.

Outcome measurements (same as Chapter 2)

The disease outcome for this study is defined as primary events of incident CAD. In the UK Biobank, participants' survey data is linked to in-patient hospital episode statistics (HES) as well as national death registry data. CAD definition in the UK Biobank for this study is the same as the recent GWAS of CAD using the UK Biobank data. (van der Harst and Verweij

2018) A participant is defined as a CAD case if he/she has at least one occurrence of the following International Classification of Diseases, 10th edition (ICD-10) codes: I21-I25 covering ischemic heart diseases; or at least one occurrence of the following Office of Population Censuses and Surveys Classification of Interventions and Procedures, version 4 (OPCS-4) codes: K40-K46, K49, K50 and K75 which includes replacement, transluminal balloon angioplasty, and other therapeutic transluminal operations on coronary artery and percutaneous transluminal balloon angioplasty and insertion of stent into coronary artery. Death because of CAD was defined as an occurrence of any ICD-10 codes stated above in the primary cause of death. To identify incident CAD cases, participants with CAD diagnosis before enrollment in the UK Biobank were excluded. Participants will be censored on the earliest date of CAD event/CAD death after enrollment, or the end of HES-based follow-up, or time of competing death, whichever occurs first.

In the MVP cohort, CAD definition was developed by a group of expert researchers from the MVP Cardiovascular Working Group. Disease diagnosis data was queried on two different index dates: date of enrollment and July 1st, 2017. The CAD definition has been chosen to accommodate both the number of cases for statistical power as well as accuracy in CAD diagnosis to control false positive rate. Participants were defined as a CAD case if there is occurrence of any CAD codes on two or more distinct dates on or prior to the index date, or occurrence of a revascularization procedure code on or prior to the index date. CAD codes include: International Classification of Diseases, 9th edition (ICD-9) codes 410, 411.0, 411.1, 411.81, 411.89, 412, 414.00, 414.01-414.05, 414.2-414.4, 414.8, 414.9, V45.81, V45.82; and ICD-10 codes I20.0, I21-I24, I25.1, I25.2, I25.5, I25.6, I25.70, I25.71, I25.72, I25.73, I25.79, I25.810, I25.82, I25.83, I25.84, I25.89, I25.9, Z95.1, Z98.61. Revascularization procedure codes include: International Classification of Diseases, 9th edition (ICD-9) codes 00.66, 36.0,

36.01-36.07, 36.09, 36.1, 36.11-36.17, 36.19, 36.2, 99.10; and ICD-10 codes 0210-0213, 0270-0273, 02C0, 02C1, 02C3, 02C4; and Current Procedural Terminology (CPT) codes: 33510-33514, 33516-33519, 33521-33523, 33530, 33533-33536, 33572, 92928, 92929, 92933, 92934, 92937, 92938, 92941, 92943, 92944, 92973-92975, 92977, 92980-92982, 92984, 92995, 92996, G0290, G0291, C9600-C9608. To identify incident CAD cases, participants who had CAD diagnosis on or prior to enrollment date were excluded. New CAD cases were defined as diagnosis between enrollment and July 1st, 2017.

Physical activity and covariate measurements

In the UK Biobank, physical activity is measured using adapted questions from the International Physical Activity Questionnaire (IPAQ) short form. Participants were asked about duration and frequency of walks, moderate activity and vigorous activity. The Guidelines for Data Processing and Analysis of the International Physical Activity Questionnaire (IPAQ) (2005) was used to process the physical activity data in the UK Biobank. A continuous variable Metabolic Equivalent of Task (MET) min/week was calculated then categorized into high, moderate and low physical activity groups. We collapsed high and moderate physical activity groups in the analysis since no difference was found between these two groups and used this "physically active" group as reference to align the two risk factors (higher CAD risk captured by higher genetic risk score as well as higher level of physical inactivity). In the MVP cohort, due to data availability a "Life's Simple 7" criterion (Lloyd-Jones, Hong et al. 2010) was used to derive a binary physical activity variable (less active vs. more active): adults with at least 150 min/week moderate intensity activity or at least 75 min/week vigorous intensity activity or combination were defined as "more active".

Genetic data processing and principle component analysis (same as Chapter 2)

Genome-wide genotyped SNP data of the UK Biobank is first examined by quality control procedures. Markers or individuals with a call rate less than 95 percent are also excluded. SNPs with Hardy-Weinberg Equilibrium p-value less than 10^{-6} or minor allele frequency less than 0.0001 are excluded. Individuals with genetically defined non-European ancestry are excluded. To remove up to the 3rd degree relatedness among the UK Biobank participants, a pairwise kinship coefficient matrix is used with kinship larger than or equal to 0.0442 as a cutoff to filter the related individual pairs. SNPs that passed the quality control procedure are then undergone a linkage disequilibrium (LD) pruning procedure with a window size of 50 kb, a step size of 5 variants, and an r^2 threshold of 0.05. LD pruned SNPs are then used in the principle component analysis. Top ten principle components are calculated and will be included in the main analysis as covariates to control for population stratification. In the MVP cohort, duplicate samples, samples with more heterozygosity than expected, an excess (>2.5%) of missing genotype calls, or discordance between genetically inferred sex and phenotypic gender are excluded. In addition, one individual from each pair of related individuals is excluded. An ethnicity-specific principle component analysis was then performed among non-Hispanic White participants who were defined using a harmonized approach combining genetically predicted ethnicity and self-reported race/ethnicity. (Fang, Hui et al. 2019)

Genetic risk score (GRS) construction (same as Chapter 2)

A comprehensive CAD-GRS based on 161 loci that have been reported in the most recent GWAS of CAD (van der Harst and Verweij 2018) was developed. A weighted GRS approach was implemented using the formula below:

$$GRS = \beta_1 \times SNP_1 + \beta_2 \times SNP_2 + \ldots + \beta_n \times SNP_n$$

β_i are effect sizes from GWAS or GWAS-meta-analysis; SNP_i is coded as number of risk alleles. In this study, the effect sizes for CAD-GRS construction were referenced from CAD-GWAS summary statistics of the CARDIoGRAMplusC4D consortium (Nikpay, Goel et al. 2015) to avoid sample overlap with the UK Biobank or the MVP cohort. In addition, three mediating trait-based CAD-sub-GRSs were developed based on loci that are associated with lipids level ($GRS_{CAD\text{-}lipids}$), blood pressure ($GRS_{CAD\text{-}BP}$), or BMI ($GRS_{CAD\text{-}BMI}$). Lipids-associated loci and blood pressure-associated loci were extracted from recent GWAS publications, (Evangelou, Warren et al. 2018, Klarin, Damrauer et al. 2018) and BMI-associated loci were obtained from unpublished BMI GWAS of up to 1 million individuals of European ancestry. Loci that are associated with both CAD and only one of the three target mediating traits (Bonferroni corrected $p<0.05$) were included in the mediating trait-specific CAD-sub-GRS calculation using the same weighted approach. All GRS constructions were performed using PLINK 2.0 (https://www.cog-genomics.org/plink/2.0/) with "--score" function, and missing genotypes were imputed to the mean dosage.

Statistical analysis

Cox proportional hazards models were used to assess the association of CAD-GRSs and sub-GRSs with incident CAD as well as the interaction between GRSs and physical inactivity in the UK Biobank. All genetic risk scores were 1) standardized and modeled as continuous variables and 2) categorized into quintiles and divided into low (lowest quintile), intermediate

(quintiles 2 to 4) and high (highest quintile) genetic risk group. Age, sex, smoking status, alcohol consumption, education, (Davies, Dickson et al. 2018) history of hypertension, history of diabetes, cholesterol lowering medication use, BMI, Townsend deprivation index and top ten principle components of the GWAS data were included as covariates. In the CAD-sub-GRS analysis, the corresponding mediating trait was not included as a covariate to avoid over-adjustment. Thus, history of hypertension was not adjusted for in the $GRS_{CAD\text{-}BP}$ analysis, cholesterol lowering medication was not adjusted for in the $GRS_{CAD\text{-}lipids}$ analysis, and BMI was not adjusted for in the $GRS_{CAD\text{-}BMI}$ analysis. Proportional hazards assumption was assessed using Schoenfeld's test. When the assumption is violated, categorical variables were stratified on while interaction terms with time were added for continuous variables. Multiplicative interaction between CAD-GRSs and physical inactivity was assessed by including interaction terms in the model and conducting likelihood ratio tests. Additive interaction was assessed by calculating relative excess risk due to interaction (RERI) based on the hazard ratio estimates. (Li and Chambless 2007) A bootstrap method was used to construct 95% confidence intervals for RERI estimates. For the replication analysis in the MVP, logistic regression models were used controlling for a similar set of covariates to assess GRS-physical inactivity interaction on both multiplicative and additive scale. Odds ratio estimates were used to calculate RERI in the MVP replication analysis and same bootstrap procedures were done to construct confidence intervals.

Results

A detailed inclusion/exclusion and QC process was presented in Figure I. A total of 296,500 participants of European ancestry who were free of CAD at baseline from the UK Biobank were included in the final analysis. Basic characteristics of the study population

were shown in Table I. 9,434 primary incident CAD events were identified from the UK Biobank. In our study sample, the mean age at baseline is 56.7 years and slightly more females (55%) than males were included. 22.8 percent of the participants who had low physical activity level based on self-reported questionnaire data at baseline were defined as physically inactive. In the replication analysis using data from the MVP, 78,510 participants with no CAD at enrollment were included. A similar inclusion/exclusion process was presented in Figure I. Using the indexing-date method, we have identified 6,246 incident CAD cases between enrollment and July 1st, 2017. Baseline characteristics of the MVP participants were summarized in Table I. The mean age is 65.1 years old and the majority of them are male (92%). According to the "Life's Simple 7" criterion, 23.9 percent of the MVP participants were categorized as physically less active.

In the main analysis, a comprehensive CAD-GRS in the UK Biobank was constructed based on 161 CAD-loci that were reported in the most recent CAD-GWAS. (van der Harst and Verweij 2018) One lead SNP (rs582384) was multiallelic in the UK Biobank so a proxy SNP (rs616381, r^2=0.86 for European ancestry) was used. A detailed list of all SNPs used for score construction can be found in Chapter 2, Table II. In the sub-score analysis, 26 SNPs were included in the $GRS_{CAD\text{-}BP}$ construction, 17 SNPs were included in the $GRS_{CAD\text{-}lipids}$ and 16 SNPs were included in the $GRS_{CAD\text{-}BMI}$. (Chapter 2, Table III) Associations between all four CAD-GRSs with primary incident CAD in the UK Biobank were presented in Table II-A. One standard deviation (SD) increase in the comprehensive CAD-GRS is independently associated with 34.1 percent increase in the risk of primary CAD events (HR: 1.341, 95% CI: 1.314, 1.368) in the UK Biobank. Comparing to those with low genetic risk of CAD, those at intermediate genetic risk had a 44.6 percent increase in the risk or primary CAD events (HR: 1.446, 95% CI: 1.360, 1.538), and those at high genetic risk had an over two-fold increase in

the risk of primary CAD events (HR: 2.215, 95% CI: 2.070, 2.369). Among the three CAD mediating trait-based sub-GRSs, the $GRS_{CAD\text{-}lipids}$ has the strongest association with incident primary CAD events. Participants with a high genetic risk captured by the $GRS_{CAD\text{-}lipids}$ had a 56.3 percent increase in CAD risk comparing to those with low genetic risk (HR: 1.563, 95% CI: 1.466, 1.666), and this genetic effect attenuated when focusing on only BP-associated CAD loci (HR: 1.428, 95% CI: 1.340, 1.523) or only BMI-associated CAD loci (HR: 1.238, 95% CI: 1.161, 1.320). For physical activity, those who were physically inactive had a 13.1 percent increase in CAD risk comparing to those who were physically active (HR: 1.131, 95% CI: 1.079, 1.184).

To further understand how the genetic predisposition of CAD interacts with physical inactivity, we assessed GRS-physical inactivity interaction on both multiplicative and additive scale. Combined associations of CAD-GRS and physical inactivity with incident CAD in the UK Biobank were presented in Table III-A. Comparing to physically active (medium or high level of physical activity) participants with low genetic risk in the UK Biobank, those who were physically active but possess a high genetic risk had over two-fold increase in CAD risk (HR: 2.239, 95% CI: 2.070,2.422), and those with high genetic risk who were physically inactive had an additional elevation in the risk of primary CAD events (HR: 2.540, 95% CI: 2.304,2.801). Similar patterns were also observed for each of the three CAD mediating trait-based sub-GRSs. Comparing to physically active participants with low genetic risk in $GRS_{CAD\text{-}lipids}$, physically inactive participants with high genetic risk have a slightly higher risk elevation (HR: 1.794, 95% CI: 1.627, 1.977) than physically active participants with high genetic risk (HR: 1.579, 95% CI: 1.465, 1.701). The strongest combined effect of genetic predisposition and physical inactivity on CAD is observed in lipids-associated loci, and BMI-associated loci seem to have weaker effects than lipids or BP-

associated CAD loci. Overall, no multiplicative interaction between CAD-GRS or CAD-sub-GRSs and physical inactivity was observed. In the MVP replication analysis, a similar pattern of associations between CAD-GRS and incident CAD was observed for the comprehensive CAD-GRS as well as $GRS_{CAD\text{-}lipids}$ and $GRS_{CAD\text{-}BP}$. (Table II-B) $GRS_{CAD\text{-}BMI}$ was found to be not associated with incident CAD in the MVP. Lipids sub-GRS had the strongest effect on CAD among all three intermediate traits. When combined with physical inactivity, no multiplicative interaction was observed. (Table III-B)

We then assessed the additive interaction between CAD-GRSs and physical inactivity by calculating RERI for each GRS. (Figure II) No significant additive interaction with physical inactivity was observed for comprehensive CAD-GRS, $GRS_{CAD\text{-}lipids}$ or $GRS_{CAD\text{-}BMI}$. A positive but insignificant trend was found for $GRS_{CAD\text{-}BP}$ and physical inactivity in the UK Biobank. (Figure II-A) In the MVP replication analysis, negative but insignificant additive interaction effects were observed between comprehensive CAD-GRS and physical inactivity as well as between $GRS_{CAD\text{-}BP}$ and physical inactivity. (Figure II-B) The additive interaction effects between $GRS_{CAD\text{-}lipids}$ and physical inactivity or between $GRS_{CAD\text{-}BMI}$ and physical inactivity were positive. To further understand poteinal heterogeneity of gene-physical inactivity interaction effects on CAD, we also conducted an individual SNP-based analysis in the UK Biobank. No significant interaction was identified after multiple-testing correction on any SNP for either multiplicative or additive scale. (Table IV) However, among the 161 loci tested, four loci located in genes *TGFB1*, *ATXN2*, *ZPR1* and *ZEB2* had marginally significant interaction ($p<0.05$) on both multiplicative and additive scale.

Discussion

In this aim, we assessed how genetic predisposition of CAD captured by CAD-GRS interact with physical activity on both multiplicative and additive scale in European population using data from two large biobank cohorts. To further understand the role of mediating traits such as BP, lipids and BMI in such gene-physical activity interaction on CAD, we also developed three separate sub-scores ($GRS_{CAD\text{-}BP}$, $GRS_{CAD\text{-}lipids}$ and $GRS_{CAD\text{-}BMI}$) focusing on genetic loci uniquely associated with one mediating trait as well as CAD risk. Overall, we didn't detect any gene-physical activity interaction effects on either multiplicative or additive scale for incident CAD.

Physical inactivity has been established as an independent risk factor for CAD. (Winzer, Woitek et al. 2018) Mechanisms of how physical activity impact cardiovascular health have been reported for both primary prevention and secondary prevention of CAD. (Brown 2003, Linke, Erbs et al. 2008) However, very limited studies have been conducted to estimate the gene-physical activity interaction effects on CAD for both multiplicative and additive scale. One study has found no multiplicative interaction between physical activity and an overall CAD-GRS in the UK Biobank. (Tikkanen, Gustafsson et al. 2018) In our study, we confirmed this finding and observed that genetic risk and physical inactivity acted multiplicatively on increasing risk for primary incident CAD events.

In our assessment of gene-physical activity interaction on the additive scale, we observed positive but insignificant additive interaction effects between physical inactivity and $GRS_{CAD\text{-}BP}$ in the UK Biobank, which suggests a synergistic but very weak effect. However, no such interaction effect was observed for CAD-sub-GRS constructed using lipids-associated or BMI-associated loci. In the MVP replication analysis, we observed inconsistent results where the overall CAD-GRS and $GRS_{CAD\text{-}BP}$ have shown no additive

interaction but the $GRS_{CAD\text{-}lipids}$ and $GRS_{CAD\text{-}BMI}$ have shown very weak but insignificant positive additive interaction effects with physical inactivity. Nonetheless, the MVP results were still questionable since we observed null marginal effects between physical inactivity and incident CAD, which is inconsistent with physical inactivity being an independent risk factor for CAD.

We also explored gene-physical activity interaction in the UK Biobank based on individual CAD-associated loci, and no significant interaction was observed for either multiplicative or additive scale after multiple testing correction. However, four loci in genes *TGFB1*, *ATXN2*, *ZPR1* and *ZEB2* showed marginally significant interaction on both multiplicative and additive scale. TGFB1 is a member of the transforming growth factor beta (TGFB) superfamily and the role of TGFB in myocardial infarcts and ischemic heart diseases has been previously reviewed. (Bujak and Frangogiannis 2007, Gordon and Blobe 2008) *TGFB1* has also been linked with mitochondrial fuel oxidation in skeletal muscle after exercise and contribute to the level of insulin sensitivity which is relevant with prevention of type 2 diabetes. The *ATXN2* locus was also reported to be associated with expression levels of gene *MAPKAPK5*, which belongs to the same serine/threonine kinase family as type I and type II receptors of the canonical TGFB signaling cascade, but the exact role of this locus in CAD remains inconclusive. (Zeng, Dang et al. 2016) *ZPR1*(*ZNF259*) locus is associated with both lipids profile and risk of CAD (Waterworth, Ricketts et al. 2010) but the mechanism of how this locus can interact with physical inactivity remains unclear.

With data from two of the largest biobank cohorts, we assessed how genetic predisposition to CAD captured by weighted CAD-GRS interacts with physical activity on the risk of primary incident CAD among the European population. We didn't observe any

significant interaction effects on either multiplicative or additive scale, which can be partly explained by the limited power using self-reported physical activity levels. A previous study in the UK Biobank has also reported weaker effect of questionnaire-based physical activity comparing to objectively measured physical activity on cardiovascular events. (Tikkanen, Gustafsson et al. 2018) However, a positive synergistic trend was observed on additive scale between $GRS_{CAD\text{-}BP}$ and physical inactivity on CAD in the UK Biobank, meaning the mechanistic interaction between genetic predisposition and physical inactivity on CAD might predominantly act through BP-related pathways. No gene-physical activity interaction has been previously examined and reported on the additive scale despite the additive scale being of greater public heath interest as well as more closely aligned with mechanistic interaction under certain assumptions. (VanderWeele 2009) Therefore, we have made the very first effort in assessing gene-physical interaction for incident CAD on both scale and future studies with better measured physical activity levels should be conducted to further evaluate such interaction effects.

Our study also has several limitations. First, using data from two of the largest biobank cohorts we relied on self-reported measurements of physical activity level as well as other potential confounding CAD risk factors such as smoking, alcohol consumption, history of hypertension and diabetes, thus there might be residual confounding in the interaction assessment due to inaccurately measured or unmeasured confounding factors. However, to balance the statistical power required for gene-lifestyle interaction studies and measurement accuracy as well as ensuring a homogeneous study population, biobank cohorts seem to be so far the best data resource in conducting large scale gene-lifestyle interaction studies. Second, we used the MVP cohort as a replication cohort for our primary analysis conducted in the UK Biobank, but the two populations differ in many aspects, which limited the power and

validity of the replication analysis. In addition, CAD cases were captured in slightly different ways between these two cohorts due to restrictions in data availability, which also lead to different definitions of some covariates in the analysis. Therefore, our results need to be interpreted with caution when comparing the primary and replication analysis.

Conclusion

Using data from the UK Biobank and the MVP cohort, we have prospectively assessed gene-physical activity interaction on incident CAD on both additive and multiplicative scale. No multiplicative or additive interaction was found between genetic predisposition and physical activity, but our findings have raised hypothesis with respect to different interplaying mechanisms between genetic predisposition to CAD and physical activity, and highlighted the value of addressing gene-lifestyle interactions in CAD on both additive and multiplicative scale.

Table I. Characteristics of the study population

Characteristic	Mean (SD) or N (%)	
	UK Biobank (N=296,500)	MVP (N=78,510)
Age	56.7 (8.0)	65.1 (11.5)
Female	162,659 (54.9%)	6,246 (8.0%)
BMI	27.3 (4.7)	29.8 (5.6)
Physical Activity		
Low	67,726 (22.8%)	18,727 (23.9%)
Medium or High	228,774 (77.2%)	59,783 (76.1%)
Smoking Status		
Current	27,788 (9.4%)	12,265 (15.6%)
Past	103,623 (34.9%)	39,704 (50.6%)
Never	170,817 (55.7%)	26,541 (33.8%)
Alcohol Consumption		
Ever	287,870 (97.1%)	48,324 (61.6%)
Never	8,630 (2.9%)	30,186 (38.4%)
Hypertension	76,517 (25.8%)	44,468 (56.6%)
Diabetes	12,522 (4.2%)	16,809 (21.4%)
Lipids Medication	43,601 (14.7%)	34,809 (44.3%)
Education		
School leaving age >=15	236,853 (79.9%)	n/a
School leaving age <15	59,647 (20.1%)	n/a
Some college or higher	n/a	60,760 (77.4%)
Townsend Index	-1.7 (2.9)	n/a
Income		
$50,000 or above	n/a	30,243 (38.5%)

Table II. Associations of comprehensive CAD genetic risk score (CAD-GRS) and three CAD mediating trait-based sub-genetic risk scores (sub-GRSs) with incident CAD

A. UK Biobank (HR and 95% CI)

CAD-GRS	Comprehensive GRS	$GRS_{CAD\text{-}BP}$	$GRS_{CAD\text{-}lipids}$	$GRS_{CAD\text{-}BMI}$
per SD increase	1.341 (1.314, 1.368)	1.140 (1.118,1.163)	1.187 (1.164, 1.211)	1.075 (1.053, 1.097)
Low risk	Ref.	Ref.	Ref.	Ref.
Intermediate risk	1.446 (1.360, 1.538)	1.172 (1.109, 1.239)	1.202 (1.137, 1.272)	1.121 (1.061, 1.184)
High risk	2.215 (2.070, 2.369)	1.428 (1.340, 1.523)	1.563 (1.466, 1.666)	1.238 (1.161, 1.320)

B. MVP (OR and 95% CI)

CAD-GRS	Comprehensive GRS	$GRS_{CAD\text{-}BP}$	$GRS_{CAD\text{-}lipids}$	$GRS_{CAD\text{-}BMI}$
per SD increase	1.221 (1.189, 1.254)	1.069 (1.042,1.097)	1.144 (1.114, 1.174)	1.023 (0.997, 1.049)
Low risk	Ref.	Ref.	Ref.	Ref.
Intermediate risk	1.359 (1.262, 1.465)	1.092 (1.019, 1.171)	1.194 (1.112, 1.282)	1.051 (0.984, 1.122)
High risk	1.815 (1.666, 1.976)	1.211 (1.115, 1.315)	1.455 (1.339, 1.581)	1.072 (0.990, 1.161)

Table III. Combined associations of CAD-GRSs and physical inactivity with incident CAD

A. UK Biobank (HR and 95% CI)

		Physical Activity		P-value for Multiplicative Interaction
		Medium or High	**Low**	
CAD-GRS	Low Risk	Ref.	1.183 (1.043,1.341)	0.72
	Intermediate Risk	1.468 (1.366,1.578)	1.640 (1.508,1.783)	
	High Risk	2.239 (2.070,2.422)	2.540 (2.304,2.801)	
GRS_{CAD-BP}	Low Risk	Ref.	1.093 (0.977, 1.223)	0.72
	Intermediate Risk	1.157 (1.084, 1.234)	1.331 (1.233, 1.437)	
	High Risk	1.417 (1.315, 1.526)	1.598 (1.447, 1.764)	
$GRS_{CAD-lipids}$	Low Risk	Ref.	1.180 (1.056, 1.320)	0.83
	Intermediate Risk	1.215 (1.138, 1.297)	1.380 (1.277, 1.491)	
	High Risk	1.579 (1.465, 1.701)	1.794 (1.627, 1.977)	
$GRS_{CAD-BMI}$	Low Risk	Ref.	1.161 (1.041, 1.296)	0.50
	Intermediate Risk	1.110 (1.042, 1.183)	1.337 (1.240, 1.441)	
	High Risk	1.248 (1.158, 1.345)	1.408 (1.273, 1.557)	

B. MVP (OR and 95% CI)

		Physical Activity		P-value for Multiplicative Interaction
		Medium or High	**Low**	
CAD-GRS	Low Risk	Ref.	1.042 (0.892,1.217)	0.26
	Intermediate Risk	1.411 (1.213,1.643)	1.399 (1.215,1.611)	
	High Risk	1.860 (1.565,2.210)	1.875 (1.616,2.177)	
GRS_{CAD-BP}	Low Risk	Ref.	1.069 (0.927, 1.232)	0.63
	Intermediate Risk	1.146 (0.994, 1.320)	1.150 (1.009, 1.310)	
	High Risk	1.322 (1.118, 1.563)	1.258 (1.093, 1.448)	
$GRS_{CAD-lipids}$	Low Risk	Ref.	0.926 (0.802, 1.068)	0.16
	Intermediate Risk	1.135 (0.986, 1.307)	1.124 (0.988, 1.280)	
	High Risk	1.272 (1.075, 1.506)	1.406 (1.224, 1.614)	
$GRS_{CAD-BMI}$	Low Risk	Ref.	0.879 (0.770, 1.003)	0.44
	Intermediate Risk	0.986 (0.864, 1.126)	0.943 (0.836, 1.064)	
	High Risk	1.006 (0.855, 1.183)	0.962 (0.844, 1.097)	

Table IV. Interaction between CAD-associated SNPs and physical inactivity on Incident CAD in the UK Biobank

SNP	CHR	POS	Locus	EA	Coefficients			P-value (Interaction)	RERI	P-value of RERI
					SNP	Low PA	Interaction			
rs8108632	19	41854534	*TGFB1*	T	0.036	0.064	0.072	0.03	0.085	0.02
rs10774625	12	111910219	*ATXN2*	A	0.058	0.205	-0.080	0.02	-0.086	0.03
rs964184	11	116648917	*ZPR1*	G	-0.013	0.096	0.103	0.03	0.116	0.03
rs7528419	1	109817192	*CELSR2*	A	0.061	0.010	0.072	0.08	0.081	0.03
rs17678683	2	145286559	*ZEB2*	G	-0.007	0.103	0.115	0.04	0.134	0.03
rs2107595	7	19049388	*HDAC9*	A	0.046	0.097	0.086	0.05	0.108	0.04
rs975722	7	117332914	*CFTR*	G	0.009	0.075	0.064	0.06	0.072	0.04
rs76954792	17	30033514	*RP11-805L22.1*	T	-0.003	0.089	0.075	0.05	0.085	0.04
rs116843064	19	8429323	*ANGPTL4*	G	0.052	-0.227	0.179	0.17	0.154	0.05
rs2083636	8	19865263	*LPL*	T	0.013	0.027	0.065	0.09	0.070	0.05
rs11617955	13	110818102	*COL4A1*	T	0.066	0.331	-0.116	0.03	-0.136	0.06
rs13003675	2	233584109	*GIGYF2*	T	-0.007	0.080	0.061	0.07	0.067	0.06
rs4918072	10	105693644	*STN1*	A	0.029	0.090	0.062	0.09	0.075	0.06
rs1887318	10	30321598	*KIAA1462*	T	0.056	0.185	-0.067	0.04	-0.071	0.07
rs12493885	3	153839866	*ARHGEF26*	C	0.119	0.305	-0.105	0.03	-0.107	0.07
rs4643373	17	47123423	*IGF2BP1*	T	-0.011	0.043	0.058	0.11	0.061	0.09
rs9349379	6	12903957	*PHACTR1*	G	0.048	0.086	0.046	0.17	0.058	0.10
rs2107732	7	45077978	*CCM2*	G	0.096	0.324	-0.109	0.06	-0.119	0.12
rs2074158	17	40257163	*DHX58*	C	0.048	0.149	-0.071	0.10	-0.076	0.12
rs35541991	6	22583856	*RP1-309H15.2*	C	0.011	0.089	0.049	0.15	0.057	0.12
rs7623687	3	49448566	*RHOA*	A	0.007	0.014	0.064	0.18	0.068	0.13
rs7947761	11	100624599	*ARHGAP42*	G	0.044	0.159	-0.058	0.11	-0.061	0.14
rs12999907	2	164957251	*AC092684.1*	A	0.054	0.039	0.052	0.24	0.061	0.14
rs9501744	6	1617143	*FOXC1*	C	0.065	0.254	-0.075	0.13	-0.079	0.19

rs6689306	1	154395946	*IL6R*	A	0.009	0.090	0.041	0.22	0.047	0.19
rs1508798	5	9556694	*RP11-260E18.1*	T	0.038	0.049	0.046	0.29	0.053	0.21
rs2246942	10	91004886	*LIPA*	G	0.012	0.097	0.040	0.24	0.047	0.21
rs1867624	17	62387091	*RPL31P57*	T	0.034	0.182	-0.047	0.17	-0.050	0.21
rs7696431	4	169687725	*PALLD*	T	0.029	0.171	-0.044	0.18	-0.047	0.22
rs246600	5	142516897	*ARHGAP26*	T	0.055	0.169	-0.047	0.16	-0.047	0.22
rs61848342	10	12303813	*RN7SL232P*	C	0.072	0.162	-0.049	0.15	-0.047	0.23
rs1050362	16	72130815	*DHX38*	A	0.015	0.154	-0.043	0.22	-0.047	0.23
rs3936511	5	55860781	*C5orf67*	G	0.007	0.106	0.047	0.26	0.054	0.23
rs11191416	10	104604916	*PFN1P11*	T	0.031	0.265	-0.077	0.19	-0.090	0.23
rs77335401	5	131759825	*C5orf56*	C	0.027	0.140	-0.062	0.23	-0.067	0.23
rs2083460	15	89574484	*RP11-326A19.2*	T	0.081	0.292	-0.090	0.15	-0.096	0.24
rs9964304	18	47229717	*RP11-813F20.2*	C	0.063	0.106	0.034	0.36	0.048	0.24
rs3851738	16	75387533	*CFDP1*	C	0.042	0.177	-0.044	0.19	-0.045	0.25
rs6984210	8	22033615	*BMP1*	G	0.075	0.116	0.070	0.31	0.098	0.25
rs12202017	6	134173151	*TARID*	A	0.062	0.194	-0.049	0.18	-0.048	0.26
rs60154123	1	210468999	*RP4-667H12.4*	T	0.053	0.111	0.042	0.34	0.056	0.26
rs3918226	7	150690176	*NOS3*	T	0.043	0.114	0.058	0.32	0.075	0.27
rs11677932	2	238223955	*STK25*	G	0.058	0.187	-0.046	0.20	-0.045	0.28
rs17581137	15	96146414	*RP11-61O11.1*	A	0.035	0.193	-0.046	0.23	-0.048	0.28
rs1924981	13	29022645	*FLT1*	T	0.009	0.149	-0.038	0.29	-0.042	0.29
rs9367716	6	57160572	*RNU7-66P*	G	0.037	0.184	-0.042	0.25	-0.044	0.30
rs7164479	15	79123054	*MORF4L1*	T	0.052	0.094	0.027	0.43	0.037	0.30
rs11601507	11	5701074	*TRIM5*	A	0.072	0.135	-0.070	0.27	-0.073	0.31
rs7617773	3	48193515	*TKT*	T	-0.016	0.078	0.035	0.31	0.037	0.31
rs72627509	4	57839051	*NOA1*	G	0.015	0.110	0.039	0.35	0.047	0.31
rs3832966	14	75614504	*TMED10*	ACCCG	0.058	0.162	-0.039	0.24	-0.037	0.32
rs9897596	17	17593453	*RAI1*	T	0.011	0.156	-0.033	0.31	-0.037	0.32

rs67180937	1	222823743	*MIA3*	G	0.062	0.193	-0.046	0.24	-0.044	0.33
rs6700559	1	200646073	*RP11-92G12.3*	C	0.020	0.094	0.029	0.39	0.034	0.33
rs10840293	11	9751196	*SWAP70*	A	0.034	0.094	0.027	0.42	0.034	0.33
rs2820315	1	201872264	*LMOD1*	T	0.054	0.151	-0.039	0.26	-0.038	0.34
rs699	1	230845794	*AGT*	G	0.026	0.102	0.028	0.41	0.035	0.34
rs2229357	12	57843711	*INHBC*	G	0.036	0.079	0.030	0.44	0.037	0.34
rs10139550	14	100145710	*HHIPL1*	G	0.035	0.103	0.026	0.44	0.034	0.35
rs10953541	7	107244545	*BCAP29*	C	0.002	0.076	0.032	0.40	0.036	0.36
rs1800449	5	121413208	*LOX*	T	-0.005	0.111	0.039	0.37	0.043	0.36
rs6841581	4	148401190	*EDNRA*	A	0.061	0.139	-0.049	0.31	-0.049	0.37
rs10093110	8	106565414	*ZFPM2*	G	0.010	0.091	0.028	0.41	0.032	0.37
rs247616	16	56989590	*AC012181.1*	C	0.006	0.085	0.029	0.41	0.033	0.37
rs6905288	6	43758873	*VEGFA*	A	0.061	0.101	0.021	0.54	0.031	0.38
rs260020	20	57714025	*ZNF831*	T	0.032	0.115	0.038	0.44	0.049	0.38
rs16986953	2	19942473	*AC019055.1*	A	0.057	0.133	-0.062	0.35	-0.064	0.38
rs6511720	19	11202306	*LDLR*	G	0.090	0.227	-0.058	0.27	-0.053	0.39
rs216172	17	2126504	*SMG6*	C	0.034	0.148	-0.033	0.34	-0.034	0.39
rs944172	9	110517794	*AL162389.1*	C	0.018	0.108	0.028	0.44	0.034	0.39
rs11830157	12	118265441	*KSR2*	G	0.017	0.149	-0.030	0.38	-0.032	0.40
rs10237377	7	139757136	*PARP12*	G	0.030	0.168	-0.032	0.35	-0.033	0.41
rs7116641	11	43696917	*RP11-472I20.4*	G	0.018	0.108	0.026	0.46	0.032	0.41
rs6997340	8	18286997	*NAT2*	T	0.021	0.143	-0.032	0.39	-0.034	0.42
rs180803	22	24658858	*POM121L9P*	G	0.114	0.447	-0.163	0.30	-0.195	0.42
rs7412	19	45412079	*APOE*	C	0.088	0.237	-0.060	0.32	-0.057	0.43
rs35879803	4	146782837	*ZNF827*	C	0.022	0.162	-0.029	0.41	-0.030	0.45
rs3827066	20	44586023	*ZNF335*	T	0.024	0.136	-0.037	0.43	-0.039	0.45
rs73015714	19	17855763	*FCHO1*	G	0.056	0.115	0.022	0.58	0.034	0.46
rs2493298	1	3325912	*PRDM16*	A	0.049	0.136	-0.040	0.40	-0.040	0.46

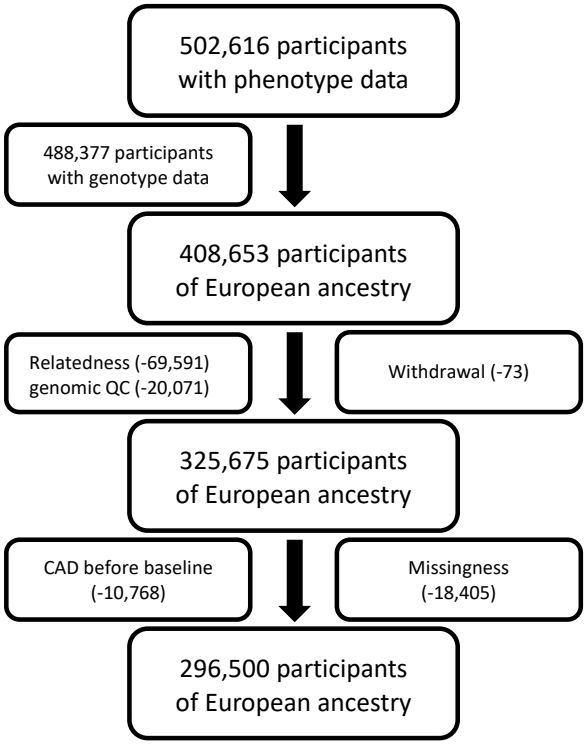
rs7306455	12	95355541	*NDUFA12*	G	0.063	0.070	0.030	0.60	0.040	0.47
rs11057401	12	124427306	*CCDC92*	T	0.027	0.164	-0.029	0.42	-0.029	0.47
rs7797644	7	6486067	*DAGLB*	C	0.028	0.089	0.023	0.55	0.029	0.47
rs616381	2	45891708	*PRKCE*	A	0.027	0.101	0.020	0.56	0.026	0.47
rs4266144	3	156852592	*SPTSSB*	G	0.052	0.143	-0.030	0.41	-0.028	0.50
rs11509880	7	12261911	*TMEM106B*	A	0.018	0.141	-0.025	0.48	-0.026	0.51
rs17517928	2	216291359	*FN1*	C	0.015	0.091	0.022	0.56	0.027	0.51
rs1321309	6	36638636	*LAP3P2*	A	0.027	0.107	0.018	0.59	0.023	0.51
rs7633770	3	46688562	*SNORD77*	A	0.034	0.110	0.017	0.62	0.023	0.52
rs2972146	2	227100698	*NEU2*	T	0.008	0.099	0.020	0.56	0.023	0.53
rs6102343	20	39924279	*ZHX3*	A	0.022	0.115	0.021	0.59	0.027	0.53
rs585967	2	21270554	*APOB*	C	0.007	0.079	0.027	0.57	0.030	0.53
rs4472337	6	34769765	*UHRF1BP1*	T	0.029	0.117	0.024	0.60	0.031	0.54
rs28451064	21	35593827	*AP000318.2*	A	0.106	0.120	0.015	0.75	0.034	0.56
rs3130683	6	31888367	*C2*	T	-0.002	0.080	0.026	0.58	0.029	0.56
rs7199941	16	81906423	*PLCG2*	A	0.030	0.114	0.014	0.67	0.020	0.59
rs1351525	11	13301548	*ARNTL*	T	0.013	0.154	-0.021	0.57	-0.022	0.60
rs56170783	1	57016131	*PLPP3*	A	0.153	0.222	-0.053	0.37	-0.034	0.60
rs840616	2	188196469	*AC007319.1*	C	0.033	0.106	0.014	0.70	0.020	0.60
rs4299376	2	44072576	*ABCG8*	G	0.029	0.115	0.014	0.69	0.020	0.60
rs10267593	7	1937261	*MAD1L1*	G	0.003	0.089	0.021	0.63	0.024	0.61
rs11838267	12	7175872	*C1S*	T	0.030	0.090	0.020	0.69	0.025	0.61
rs10455872	6	161010118	*LPA*	G	0.226	0.134	-0.051	0.36	-0.035	0.63
rs11810571	1	151762308	*TDRKH*	G	0.019	0.094	0.018	0.69	0.023	0.64
rs10841443	12	20220033	*RP11-664H17.1*	G	0.049	0.155	-0.022	0.53	-0.019	0.64
rs885150	9	124420173	*DAB2IP*	C	0.023	0.117	0.014	0.70	0.019	0.64
rs2839812	11	103673294	*RP11-563P16.1*	T	0.047	0.118	0.011	0.77	0.019	0.64
rs4752700	10	124237612	*HTRA1*	G	0.042	0.142	-0.020	0.55	-0.017	0.65

rs11723436	4	120901336	*RP11-170N16.1*	G	0.031	0.117	0.012	0.74	0.018	0.65
rs17080091	6	150997401	*PLEKHG1*	C	0.082	0.093	0.017	0.80	0.029	0.66
rs867186	20	33764554	*PROCR*	A	0.034	0.087	0.021	0.73	0.027	0.66
rs9591012	13	33058333	*N4BP2L2*	G	0.006	0.107	0.014	0.69	0.016	0.66
rs742115	6	11327021	*NEDD9*	C	0.012	0.141	-0.015	0.65	-0.016	0.67
rs6494488	15	65024204	*RBPMS2*	A	0.023	0.097	0.016	0.73	0.021	0.67
rs11099493	4	82587050	*RASGEF1B*	A	0.008	0.145	-0.015	0.66	-0.017	0.68
rs11170820	12	54513915	*FLJ12825*	G	0.103	0.129	-0.039	0.57	-0.034	0.68
rs9818870	3	138122122	*MRAS*	T	0.048	0.132	-0.023	0.60	-0.021	0.68
rs1591805	6	126717064	*RP11-394G3.2*	A	0.027	0.115	0.010	0.77	0.015	0.68
rs2145598	14	58794001	*ARID4A*	G	0.037	0.140	-0.018	0.60	-0.015	0.69
rs8068952	17	59286644	*BCAS3*	G	0.041	0.133	-0.020	0.63	-0.017	0.71
rs61776719	1	38461319	*SF3A3*	A	0.024	0.142	-0.015	0.65	-0.014	0.71
rs1892094	1	169094459	*ATP1B1*	C	0.026	0.139	-0.015	0.65	-0.014	0.71
rs56015508	6	39152041	*KCNK5*	C	0.012	0.105	0.013	0.75	0.016	0.72
rs2891168	9	22098619	*CDKN2B-AS1*	G	0.150	0.155	-0.030	0.36	-0.013	0.72
rs12500824	4	77416627	*SHROOM3*	A	0.031	0.119	0.008	0.81	0.014	0.72
rs36096196	1	2252205	*MORN1*	T	0.074	0.133	-0.024	0.60	-0.018	0.73
rs17680741	10	82251514	*TSPAN14*	T	0.053	0.152	-0.019	0.61	-0.014	0.73
rs2306556	4	156638573	*GUCY1A3*	A	0.055	0.113	0.007	0.87	0.015	0.73
rs2832227	21	30533076	*MAP3K7CL*	G	0.054	0.122	0.008	0.86	0.017	0.74
rs2244608	12	121416988	*HNF1A*	G	0.059	0.136	-0.018	0.61	-0.013	0.75
rs12801636	11	65391317	*PCNX3*	G	0.031	0.149	-0.016	0.69	-0.014	0.76
rs13723	17	27941886	*CORO6*	G	0.024	0.136	-0.012	0.72	-0.010	0.78
rs17608766	17	45013271	*GOSR2*	C	0.026	0.129	-0.014	0.76	-0.013	0.81
rs10512861	3	132257961	*DNAJC13*	G	0.025	0.150	-0.015	0.77	-0.013	0.81
rs17843797	3	124453022	*UMPS*	G	0.061	0.124	0.004	0.93	0.013	0.81
rs7500448	16	83045790	*CDH13*	A	0.035	0.144	-0.013	0.74	-0.010	0.82

rs2571445	2	218683154	*TNS1*	A	0.012	0.120	0.006	0.85	0.009	0.82
rs590121	11	75274150	*SERPINH1*	T	0.022	0.131	-0.011	0.77	-0.010	0.82
rs114123510	2	203831212	*CARF*	A	0.063	0.129	-0.018	0.71	-0.012	0.82
rs11556924	7	129663496	*ZC3HC1*	C	0.028	0.138	-0.011	0.76	-0.008	0.83
rs2681472	12	90008959	*ATP2B1*	G	0.040	0.123	0.004	0.93	0.010	0.84
rs111245230	9	113169775	*SVEP1*	C	0.069	0.126	-0.024	0.79	-0.019	0.85
rs72743461	15	67441750	*SMAD3*	C	0.039	0.142	-0.011	0.77	-0.007	0.86
rs2954029	8	126490972	*RP11-136O12.2*	A	0.030	0.134	-0.009	0.79	-0.006	0.87
rs663129	18	57838401	*RNU4-17P*	A	0.016	0.128	-0.008	0.84	-0.007	0.88
rs16844401	4	3449652	*HGFAC*	A	0.017	0.124	0.007	0.91	0.011	0.88
rs667920	3	136069472	*STAG1*	T	0.023	0.121	0.002	0.95	0.006	0.89
rs7568458	2	85788175	*GGCX*	A	0.036	0.126	-0.001	0.98	0.004	0.92
rs10857147	4	81181072	*RP11-576N17.4*	T	0.045	0.130	-0.009	0.81	-0.004	0.92
rs1317507	13	113631780	*MCF2L*	A	0.028	0.124	0.000	0.99	0.004	0.92
rs4613862	6	82612271	*RP11-379B8.1*	A	0.026	0.125	0.000	0.99	0.003	0.93
rs10417115	19	33386556	*CEP89*	C	0.026	0.125	0.002	0.98	0.006	0.95
rs2071382	15	91428197	*FES*	T	0.070	0.131	-0.006	0.85	0.002	0.95
rs3775058	4	96117371	*UNC5C*	A	0.023	0.127	-0.005	0.91	-0.002	0.96
rs11806316	1	115753482	*RP4-663N10.1*	G	0.033	0.128	-0.002	0.95	0.002	0.96
rs1870634	10	44480811	*LINC00841*	G	0.041	0.133	-0.007	0.85	-0.002	0.96
rs507666	9	136149399	*ABO*	A	0.007	0.126	-0.002	0.95	-0.002	0.97
rs748431	3	14928077	*FGD5*	G	0.025	0.128	-0.004	0.90	-0.001	0.97
rs12897	3	172115902	*FNDC3B*	G	0.051	0.130	-0.007	0.83	-0.001	0.97
rs11591147	1	55505647	*PCSK9*	G	0.274	0.214	-0.045	0.75	0.004	0.98
rs112635299	14	94838142	*SERPINA1*	G	0.122	0.163	-0.019	0.87	-0.003	0.98

Figure I. Study population QC process

A. UK Biobank



B. The MVP Cohort

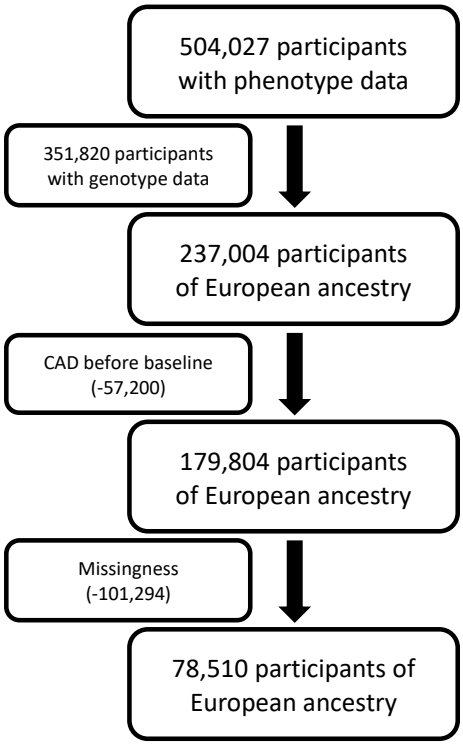
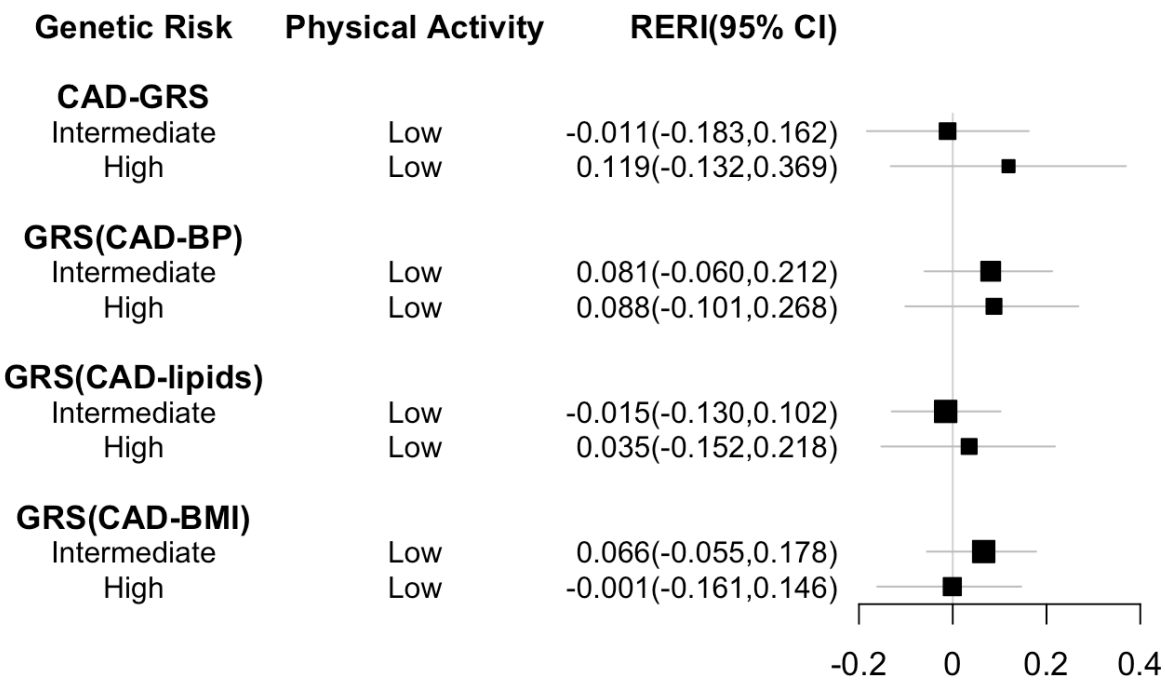
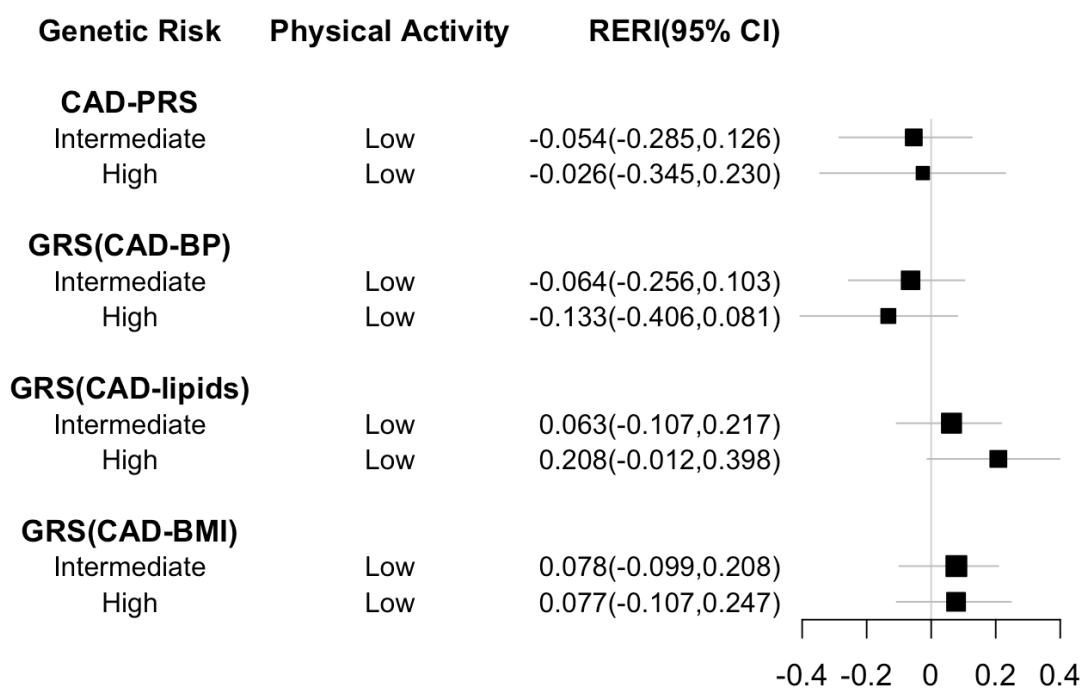


Figure II. Relative excess risk due to interaction (RERI) for CAD-GRSs and physical inactivity on incident CAD (Physically active individuals with low genetic risk were used as reference group for each score; 95% CI: 95% bootstrap confidence interval)

A. UK Biobank



B. MVP


Genetic Risk
Physical Activity
RERI(95% CI)
CAD-PRS
Intermediate
Low
-0.054(-0.285,0.126)
High
Low
-0.026(-0.345,0.230)
GRS(CAD-BP)
Intermediate
Low
-0.064(-0.256,0.103)
High
Low
-0.133(-0.406,0.081)
GRS(CAD-lipids)
Intermediate
Low
0.063(-0.107,0.217)
High
Low
0.208(-0.012,0.398)
GRS(CAD-BMI)
Intermediate
Low
0.078(-0.099,0.208)
High
Low
0.077(-0.107,0.247)
-0.4
-0.2
0
0.2
0.4

Chapter 4

Genome-wide association studies of coronary artery diseases accounting for gene-smoking interaction or gene-physical activity interaction

Introduction

Gene-environment interaction is an important component of the genetic architecture of complex diseases. Because environmental factors contribute to variation in disease development, accounting for environmental exposures and G×E interactions in genetic studies of complex diseases may affect overall trait variance when investigating genetic contributions and can potentially identify novel loci, highlighting new biological processes and pathways.(Kraft, Yen et al. 2007) Conventional GWAS of complex disease focused only on the marginal genetic effects, which may miss variants that exert effect through interactions with environmental factors. Studies have shown better power when environmental factors such as smoking and physical activity were considered for investigating novel loci for complex traits such as BMI.(Graff, Scott et al. 2017, Justice and Winkler 2017)

CAD is a complex disease with both a genetic and an environmental component and smoking as well as physical activity contributes to the variation in CAD risk, hence, accounting for smoking/physical activity or gene-smoking/gene-physical activity interaction can improve power for the discovery of CAD-associated loci. To our knowledge, no GWAS of CAD has been done incorporating such environmental factors or G×E interaction effects. A joint two degree of freedom (2-df) testing approach has been proposed to provide better power for GWAS and can be used to identify novel loci. (Kraft, Yen et al. 2007) However,

this approach hasn't been implemented in any of the large-scale GWAS of CAD. Recent efforts from the CHARGE Gene-Lifestyle Interactions Working Group have shown promising results in using the 2-df testing approach for conducting large scale GWAS of CAD risk factors such as blood pressure and lipids. (Feitosa, Kraja et al. 2018, Sung, Winkler et al. 2018, Bentley, Sung et al. 2019, de Vries, Brown et al. 2019) Therefore, the purpose of this study is to conduct two separate GWASs accounting for 1) gene-smoking interaction; and 2) gene-physical activity interaction as well as marginal genetic effect using a joint 2-df testing approach in European ancestry, respectively.

Methods

Study population (same as Chapter 2 and Chapter 3, descriptive section for the UK Biobank)

The study population of this aim consists of participants with European ancestry from the UK Biobank (https://www.ukbiobank.ac.uk/) cohort. The UK Biobank is a major national and international health resource, and a registered charity in its own right, with the aim of improving the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Both genetic and phenotypic data for all participants in the UK Biobank were obtained for this study. The UK Biobank genetic data contains genome-wide genotypes for 488,377 participants. (Bycroft, Freeman et al. 2018) These were assayed using two very similar genotyping arrays. A subset of 49,950 participants involved in the UK Biobank Lung Exome Variant Evaluation (UK

BiLEVE) study were genotyped at 807,411 markers using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix (now part of Thermo Fisher Scientific), and 438,427 participants were genotyped using the closely related Applied Biosystems UK Biobank Axiom Array (825,927 markers) that shares 95% of marker content with the UK BiLEVE Axiom Array. A quality control pipeline was developed and applied specifically to accommodate the large-scale dataset of ethnically diverse participants, genotyped in many batches, using two slightly different arrays, and which will be used by many researchers to tackle a wide variety of research questions. Markers that passed the quality control check were imputed using the Haplotype Reference Consortium (HRC) reference panel as well as the merged UK10K and 1000 Genomes phase 3 reference panels. Information was then combined using the HRC data as the primary resource. For phenotype data, participants provided electronic signed consent, answered questions on socio-demographic, lifestyle and health-related factors, and completed a range of physical measures at baseline recruitment. All participants also provided consent for follow-up through linkage to their health-related records including in-patient hospital episode statistics and national death registry data.

CAD identification

In the UK Biobank, participants' survey data is linked to in-patient hospital episode statistics (HES) as well as national death registry data. CAD definition is referenced from the recent GWAS of CAD using the UK Biobank data. (van der Harst and Verweij 2018) A participant is defined as a CAD case if he/she has at least one occurrence of the following International Classification of Diseases, 10th edition (ICD-10) codes: I21-I25 covering ischemic heart diseases; or at least one occurrence of the following Office of Population Censuses and Surveys Classification of Interventions and Procedures, version 4 (OPCS-4)

codes: K40-K46, K49, K50 and K75 which includes replacement, transluminal balloon angioplasty, and other therapeutic transluminal operations on coronary artery and percutaneous transluminal balloon angioplasty and insertion of stent into coronary artery. Death because of CAD was defined as an occurrence of any ICD-10 codes stated above in the primary cause of death. CAD cases included both CAD cases captured by HES data and deaths of CAD captured by the death registry data.

Genetic data processing and principle component analysis (same as Chapter 2 and Chapter 3)

Genome-wide genotyped SNP data of the UK Biobank was first examined by quality control procedures. Markers or individuals with a call rate less than 95 percent were also excluded. SNPs with Hardy-Weinberg Equilibrium p-value less than 10^{-6} or minor allele frequency less than 0.0001 were excluded. Individuals with genetically defined non-European ancestry were excluded. To remove up to the 3rd degree relatedness among the UK Biobank participants, a pairwise kinship coefficient matrix was used with kinship larger than or equal to 0.0442 as a cutoff to filter the related individual pairs. SNPs that passed the quality control procedure were then undergone a linkage disequilibrium (LD) pruning procedure with a window size of 50 kb, a step size of 5 variants, and an r2 threshold of 0.05. LD pruned SNPs were then used in the principle component analysis. Top ten principle components were calculated and included in the main analysis as covariates to control for population stratification.

Statistical analysis

In this aim, two GWASs of CAD accounting for 1) smoking and gene-smoking interaction and 2) physical activity and gene-physical activity interaction were conducted using a logistic regression model:

$$logit\big(P(CAD = 1)\big) = \beta_0 + \beta_1 SNP + \beta_2 E + \beta_3 SNP \times E + \beta_4 C$$

A joint 2-df testing approach was applied by testing:

$$\beta_1 = \beta_3 = 0$$

Smoking was categorized into current smokers vs. non-current smokers, and physical activity was categorized into low physical activity (inactive) vs. moderate or high physical activity (active). SNPs with minor allele frequency larger than 0.005, an imputation quality score at least 0.8, and Hardy-Weinberg Equilibrium p-value larger than 10^{-10} were kept in the analysis. A conventional Bonferroni correction $p<5\times10^{-8}$ was used to identify genome-wide significant loci. All significant loci were compared to previously reported CAD-associated loci for identification of novel loci. Age, sex and top ten principle components of the GWAS data were controlled for as covariates. Both GWASs were conducted using SUGEN (Lin, Tao et al. 2014) and robust variance estimates were used to correct for potential heteroscedasticity in GWASs accounting for environmental factors as well as gene-environment interaction effects. (Almli, Duncan et al. 2014)

Results

GWAS of CAD accounting for current smoking and gene-current smoking interaction

20,953 CAD cases and 303,547 controls of European ancestry were included in the GWAS accounting for current smoking and gene-current smoking interaction. (Figure I-A) Descriptive statistics of the study population are presented in Table I-A. 8.34 million SNPs

passed QC and were tested using the joint 2-df approach. 1,132 SNPs were genome-wide significant. (joint 2-df $p<5\times10^{-8}$) After merging significant signals with genomic distance less than 500 kb or in linkage disequilibrium (r^2 >0.1), 22 genome-wide significant loci were detected. (Figure 1) Comparing to the most recent CAD-GWAS meta-analysis conducted by van der Harst et al using data from the UK Biobank and CARDIoGRAMplusC4D consortium, (van der Harst and Verweij 2018) one locus (8q22.3) was not reported. (Figure I-B) A further examination showed that this locus was not replicated in the CARDIoGRAMplusC4D consortium and was not genome-wide significant in their meta-analysis. (van der Harst and Verweij 2018) However, the mapped gene at this locus *NCALD* has been reported to be associated with systolic blood pressure (Evangelou, Warren et al. 2018) and previous analysis in the UK Biobank has also identified this locus as associated with cardiovascular diseases. (Kichaev, Bhatia et al. 2019) Overall, the joint 2-df testing approach accounting for current smoking as well as gene-current smoking interaction effects has added little information or statistical power over the conventional CAD-GWAS of main genetic effect. The comparison between p-values of the joint 2-df approach and p-values of the conventional GWAS approach were presented in Figure II. The overall distributions were almost identical with conventional GWAS being slightly more powerful. (Figure II-A) Focusing on genome-wide significant SNPs (joint 2-df $p<5\times10^{-8}$), the two approaches had similar statistical power. (Figure II-B) When comparing suggestively associated SNPs ($p<10^{-5}$), the majority had smaller p-values as well as better power using the conventional GWAS approach. (Figure II-C)

GWAS of CAD accounting for physical activity and gene-physical activity interaction

20,154 CAD cases and 293,476 controls of European ancestry were included in the GWAS accounting for physical activity and gene-physical activity interaction. (Figure I-B) Descriptive statistics of the study population are presented in Table I-B. 8.34 million SNPs passed QC and were tested using the joint 2-df approach. 1,073 SNPs were genome-wide significant. (joint 2-df $p<5\times10^{-8}$) After merging significant signals with genomic distance less than 500 kb or in linkage disequilibrium (r^2 >0.1), 22 genome-wide significant loci were detected. (Figure III) Comparing to the most recent CAD-GWAS meta-analysis conducted by van der Harst et al using data from the UK Biobank and CARDIoGRAMplusC4D consortium, (van der Harst and Verweij 2018) all loci have been previously reported. Overall, the joint 2-df testing approach accounting for physical activity as well as gene-physical activity interaction effects has added little information or statistical power in the CAD-GWAS. The comparison between p-values of the joint 2-df approach and p-values of the conventional GWAS approach were presented in Figure IV. The overall distributions were similar with conventional GWAS being slightly more powerful. (Figure IV-A) Focusing on genome-wide significant SNPs (joint 2-df $p<5\times10^{-8}$), the 2-df approaches didn't gain any statistical power to identify CAD-associated genetic loci. (Figure IV-B) When comparing suggestively associated SNPs ($5\times10^{-8} < p < 10^{-5}$), the majority had smaller p-values as well as better power using the conventional GWAS approach (Figure IV-C). However, the variation suggested that the power was moderately improved for some SNPs when the interaction term was included in the 2-df test.

Discussion

Smoking and physical inactivity are two well-established lifestyle-related risk factors for CAD, (Benjamin, Muntner et al. 2019) however, no CAD-GWAS has been conducted

accounting for gene-smoking or gene-physical activity interaction effects. We have implemented a joint two degree of freedom approach (Kraft, Yen et al. 2007) and conducted two GWASs of CAD accounting for gene-smoking and gene-physical activity interaction in a large population of European ancestry using data from the UK Biobank. Overall, no novel locus has been identified from our analyses and p-value comparisons haven shown that comparing to the conventional GWAS method the joint 2-df testing approach added little power to the detection of novel CAD-associated loci.

CAD is a heritable condition and enormous effort has been made to identify the potential underlying genetic mechanisms of CAD. (Nikpay, Goel et al. 2015, Khera and Kathiresan 2017, van der Harst and Verweij 2018) Previously reported CAD-associated loci collectively explained 30 – 40% of CAD heritability (Khera and Kathiresan 2017) but failed to account for environmental factors or potential heterogeneity of genetic effects across different strata in environmental factors. Lifestyle-related factors such as smoking and physical inactivity are important risk factors for cardio-metabolic health, and may interact with the genetic susceptibility. Conventional GWAS is underpowered to detect susceptibility loci that act through certain strata of lifestyle factors (i.e., gene-lifestyle interaction). Alternative methods have been developed (Kraft, Yen et al. 2007) but previous studies have been limited by the lack of large, population-based cohorts with both genotypic and lifestyle-related phenotypic data. The UK Biobank is a leading resource for large-scale gene-phenotype association studies, which also enables large-scale GWAS of CAD accounting for potential gene-lifestyle interaction effects.

Our analyses have revealed that vast majority of the CAD-associated loci do not have a strong gene-lifestyle interaction component in a large population of European ancestry.

Although previous work in the CHARGE Gene-Lifestyle Interactions Working Group have shown promising results in using the 2-df testing approach for conducting large scale GWAS of CAD risk factors such as blood pressures and blood lipids, (Feitosa, Kraja et al. 2018, Sung, Winkler et al. 2018, Bentley, Sung et al. 2019, de Vries, Brown et al. 2019) our results have shown that such 2-df approach for smoking and physical activity added little power in the detection of novel CAD-associated loci comparing to the conventional GWAS approach. However, the joint 2-df approach that we have implemented only accounted for the multiplicative SNP-environmental factor interaction effects and didn't capture the potential additive interaction effect. Novel genome-wide scale statistical computing tools are awaiting to be developed for jointly testing the main genetic effect and additive gene-lifestyle interaction effects for complex diseases. Our study also has several limitations. First, CAD cases are identified based on in-patient hospital records which might lead to underdiagnosis or misclassification of disease status. Smoking and physical activity are self-reported only which might lead to inaccurate measurement of such complex lifestyle-related factors. In addition, we have dichotomized smoking into current vs. non-current and physical activity into inactive vs. active for analytical simplicity, but the actual gene-lifestyle interaction effects might act through a more complex mechanism than such binary categorization.

Conclusion

In this study, using data from the UK Biobank we have conducted two separate CAD-GWASs accounting for gene-smoking interaction and gene-physical activity interaction, respectively. Overall, no novel loci have been identified from our analyses using a joint 2-df testing approach, and p-value comparisons have shown that the majority of CAD-associated loci do not have a strong interaction effect with smoking or physical activity. However, some

aspects in the methods such as neglected potential additive interaction effects can be improved in the future for CAD-GWAS accounting for environmental factors as well as gene-environmental interactions.

Table I. Descriptive statistics of study population in the UK Biobank

A. GWAS accounting for current smoking and gene-current smoking interaction (N=324,500)

Characteristics	Mean (SD) or N (%)
Age	56.9 (8.0)
Female	174,719 (53.8%)
Current smoking	31,536 (9.7%)

B. GWAS accounting for physical activity and gene-physical activity interaction (N=313,630)

Characteristics	Mean (SD) or N (%)
Age	56.9 (8.0)
Female	168,478 (53.7%)
Physical activity	
Low	72,951 (23.3%)
Moderate or High	240,679 (76.7%)

Table II. Genome-wide significant loci of CAD using joint 2-df testing approach in the UK Biobank

A. GWAS accounting for current smoking and gene-current smoking interaction

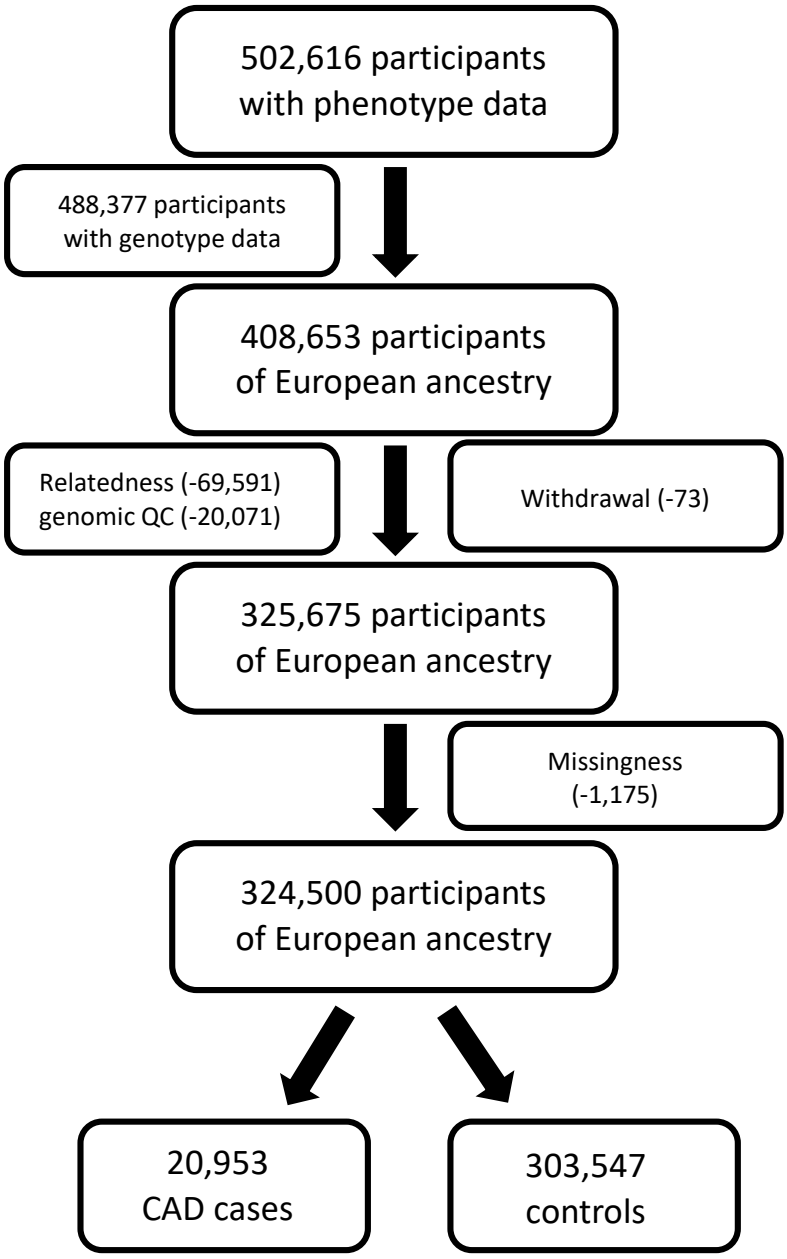
rsID	CHR	POS	REF	ALT	Gene	Joint 2-df Testing Approach					Conventional GWAS		
						Beta_G	Beta_G:Csmk	P_G	P_G:Csmk	P_2-df	Beta	SE	P
rs5896965	9	22099746	TA	T	*CDKN2B-AS1*	-0.179	-0.009	2.33E-57	7.77E-01	4.22E-65	-0.179	0.010	3.30E-66
rs55730499	6	161005610	T	C	*LPA*	-0.304	0.034	6.01E-60	5.23E-01	5.26E-65	-0.299	0.017	5.22E-66
rs12740374	1	109817590	T	G	*CELSR2*	0.129	-0.035	1.32E-20	3.59E-01	4.64E-21	0.123	0.013	8.62E-22
rs429358	19	45411941	C	T	*APOE*	-0.115	-0.020	1.21E-14	6.28E-01	3.31E-16	-0.116	0.014	9.50E-17
rs764429222	15	91428521	C	CT	*FES*	-0.095	0.069	4.88E-16	3.70E-02	3.55E-15	-0.085	0.011	6.65E-15
rs28451064	21	35593827	A	G	*AP000318.2*	-0.116	-0.025	8.48E-13	5.87E-01	3.01E-14	-0.119	0.015	3.66E-15
rs12324886	15	79054108	G	A	*ADAMTS7*	0.091	-0.015	1.38E-12	6.90E-01	1.03E-12	0.089	0.012	8.38E-14
rs11591147	1	55505647	T	G	*PCSK9*	0.295	0.245	5.33E-10	8.91E-02	1.64E-12	0.325	0.045	4.27E-13
rs56322312	1	56991890	T	G	*PPAP2B*	0.150	-0.140	3.52E-13	1.34E-02	3.20E-12	0.129	0.019	1.14E-11
rs6925904	6	12892486	G	A	*PHACTR1*	-0.081	0.011	5.64E-12	7.45E-01	3.42E-12	-0.079	0.011	6.44E-13
rs114846969	19	11191197	A	G	*LDLR*	0.119	-0.034	5.06E-11	5.03E-01	8.61E-11	0.115	0.017	1.02E-11
rs917054	17	47340153	T	C	*RP1-62O9.3*	0.067	0.014	2.11E-09	6.70E-01	4.18E-10	0.069	0.010	4.18E-11
rs145168080	2	203865822	T	TGC	*RP11-544H14.1*	-0.106	0.062	8.91E-11	2.05E-01	4.66E-10	-0.096	0.015	3.21E-10
rs2107595	7	19049388	A	G	*HDAC9*	-0.091	0.003	1.96E-09	9.41E-01	1.32E-09	-0.090	0.014	2.12E-10
rs1169288	12	121416650	C	A	*HNF1A*	-0.075	0.064	3.82E-10	5.68E-02	2.88E-09	-0.066	0.011	2.61E-09
rs10160170	10	44692843	G	A	*RP11-20J15.2*	0.101	0.036	3.47E-08	4.83E-01	4.32E-09	0.105	0.017	9.54E-10
rs1384705	11	103696851	T	C	*RP11-563P16.1*	0.071	-0.016	5.23E-09	6.47E-01	8.54E-09	0.070	0.011	9.02E-10
rs3918226	7	150690176	T	C	*NOS3*	-0.101	-0.077	5.12E-07	1.66E-01	8.98E-09	-0.113	0.019	1.10E-09
rs77215829	12	112618346	C	A	*HECTD4*	0.089	0.040	1.63E-07	3.99E-01	1.69E-08	0.093	0.016	3.72E-09
rs500546	8	102870342	A	G	*NCALD*	0.080	-0.024	1.65E-08	5.38E-01	3.89E-08	0.076	0.013	4.98E-09
rs58721068	4	148387701	G	A	*RP11-752L20.1*	-0.088	0.031	1.57E-08	4.87E-01	4.58E-08	-0.083	0.015	1.10E-08
rs35239117	12	95547732	T	A	*FGD6*	0.123	0.009	7.45E-08	8.94E-01	4.89E-08	0.123	0.021	7.58E-09

B. GWAS accounting for physical activity and gene-physical activity interaction

rsID	CHR	POS	REF	ALT	Gene	Joint 2-df Testing Approach					Conventional GWAS		
						Beta_G	Beta_G:PA	P_G	P_G:PA	P_2-df	Beta	SE	P
rs4007642	9	22093299	T	A	*CDKN2B-AS1*	-0.192	0.033	4.56E-53	1.71E-01	6.40E-65	-0.183	0.011	7.66E-66
rs55730499	6	161005610	T	C	*LPA*	-0.319	0.068	4.13E-53	9.12E-02	1.93E-63	-0.301	0.018	1.85E-64
rs12740374	1	109817590	T	G	*CELSR2*	0.122	0.018	2.46E-15	5.53E-01	7.09E-21	0.127	0.013	3.80E-22
rs769449	19	45410002	A	G	*APOE*	-0.141	0.047	2.69E-15	1.74E-01	2.06E-16	-0.127	0.015	1.57E-16
rs764429222	15	91428521	C	CT	*FES*	-0.094	0.012	7.30E-13	6.29E-01	4.65E-15	-0.090	0.011	7.06E-16
rs28451064	21	35593827	A	G	*AP000318.2*	-0.104	-0.041	9.64E-09	2.39E-01	4.09E-13	-0.115	0.015	1.14E-13
rs11591147	1	55505647	T	G	*PCSK9*	0.341	-0.023	3.97E-10	8.16E-01	2.57E-12	0.328	0.046	6.70E-13
rs12324886	15	79054108	G	A	*ADAMTS7*	0.092	-0.012	1.89E-10	6.53E-01	5.79E-12	0.087	0.012	9.75E-13
rs6925904	6	12892486	G	A	*PHACTR1*	-0.078	-0.007	2.91E-09	7.68E-01	8.64E-12	-0.080	0.011	9.49E-13
rs139853365	19	11190556	C	T	*LDLR*	0.127	0.018	1.95E-08	6.72E-01	6.67E-11	0.133	0.019	5.98E-12
rs56322312	1	56991890	T	G	*PPAP2B*	0.140	-0.028	1.20E-09	5.21E-01	1.05E-10	0.132	0.019	8.69E-12
rs2107595	7	19049388	A	G	*HDAC9*	-0.075	-0.059	9.21E-06	6.78E-02	3.61E-10	-0.092	0.014	2.15E-10
rs145168080	2	203865822	T	TGC	*RP11-544H14.1*	-0.115	0.074	2.82E-10	3.90E-02	8.93E-10	-0.095	0.016	9.84E-10
rs2011767	17	47340297	T	C	*RP1-62O9.3*	0.066	0.009	1.47E-07	7.23E-01	1.36E-09	0.069	0.011	1.33E-10
rs4846767	1	222763026	T	C	*TAF1A*	-0.069	-0.011	5.33E-07	6.72E-01	6.67E-09	-0.071	0.012	1.19E-09
rs10160170	10	44692843	G	A	*RP11-20J15.2*	0.099	0.025	1.26E-06	5.19E-01	8.33E-09	0.107	0.017	7.97E-10
rs3918226	7	150690176	T	C	*NOS3*	-0.102	-0.046	5.98E-06	2.78E-01	8.65E-09	-0.115	0.019	1.51E-09
rs1169288	12	121416650	C	A	*HNF1A*	-0.076	0.038	1.14E-08	1.35E-01	1.89E-08	-0.066	0.011	7.40E-09
rs2839812	11	103673294	A	T	*RP11-563P16.1*	0.070	0.000	3.72E-07	9.95E-01	1.90E-08	0.070	0.012	2.06E-09
rs11394930	10	12306521	CT	C	*RN7SL232P*	0.073	-0.054	6.20E-09	2.52E-02	3.01E-08	0.059	0.011	4.24E-08
rs2172725	13	110837456	A	G	*COL4A1*	-0.076	0.075	4.33E-09	2.63E-03	3.27E-08	-0.056	0.011	3.36E-07
rs7641039	3	153768638	A	C	*ARHGEF26-AS1*	0.102	-0.077	7.45E-09	2.00E-02	3.76E-08	0.080	0.015	8.65E-08

Figure I. Study population QC process

A. GWAS of CAD accounting for current smoking and gene-current smoking interaction



B. GWAS of CAD accounting for physical activity and gene-physical activity interaction

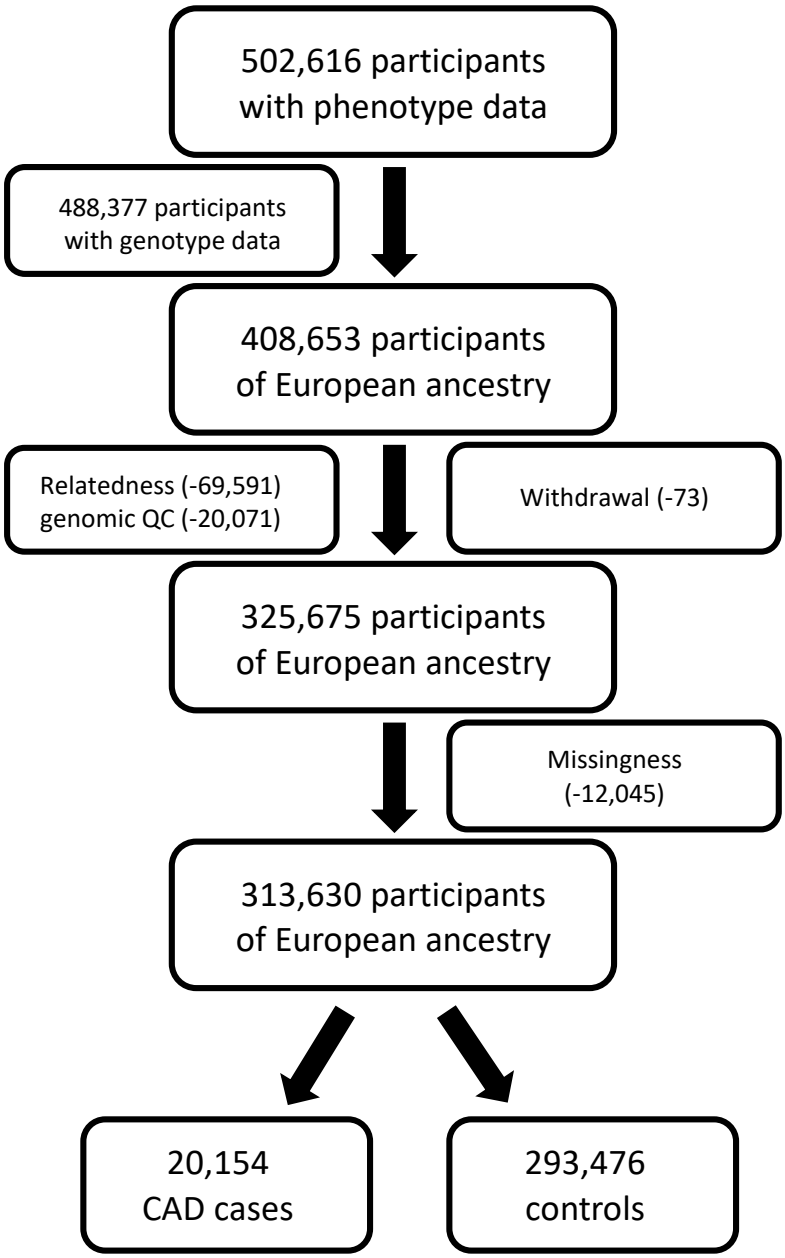
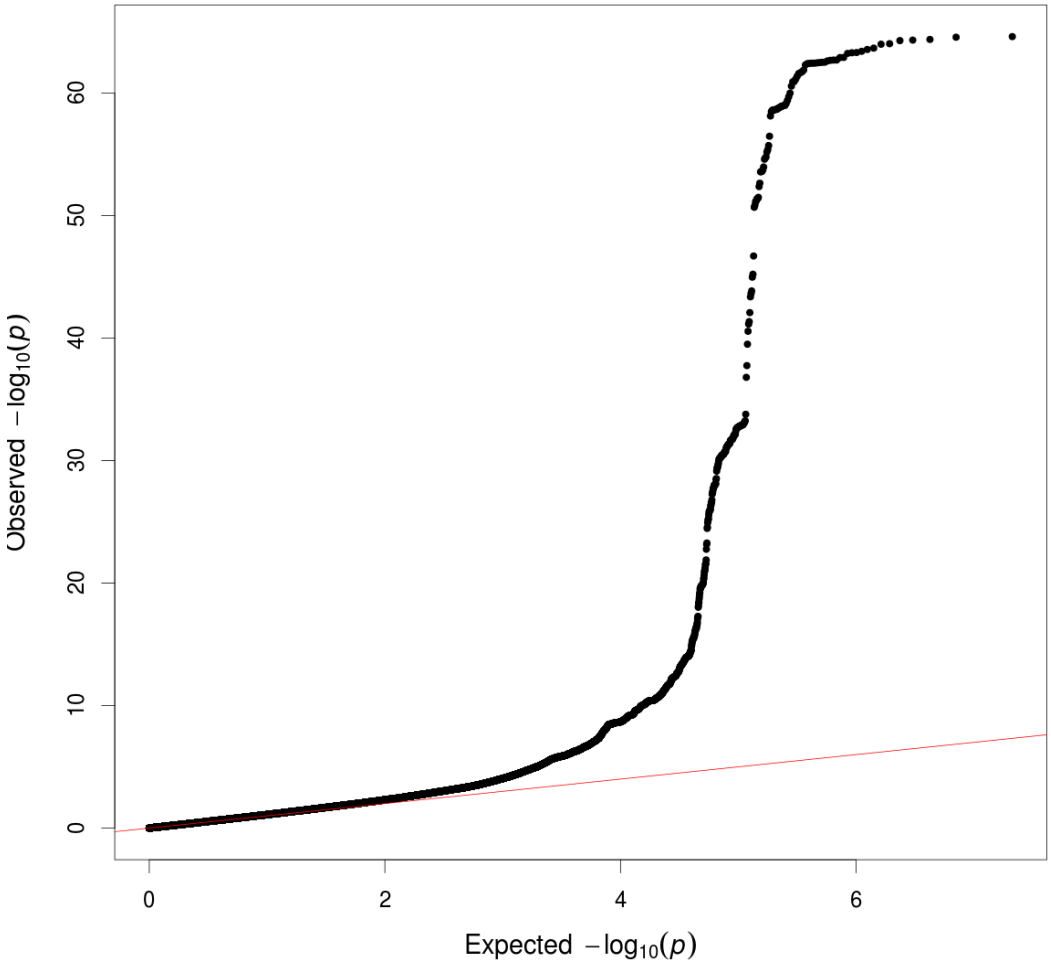


Figure II. GWAS of CAD in the UK Biobank accounting for current smoking and gene-current smoking interaction

A. Quantile-Quantile Plot (Inflation Factor = 1.13)



B. Manhattan plot (red line: genome-wide significant with $p < 5\times10^{-8}$, highlighted one locus that was not reported in the most recent CAD-GWAS by van der Harst et al.)

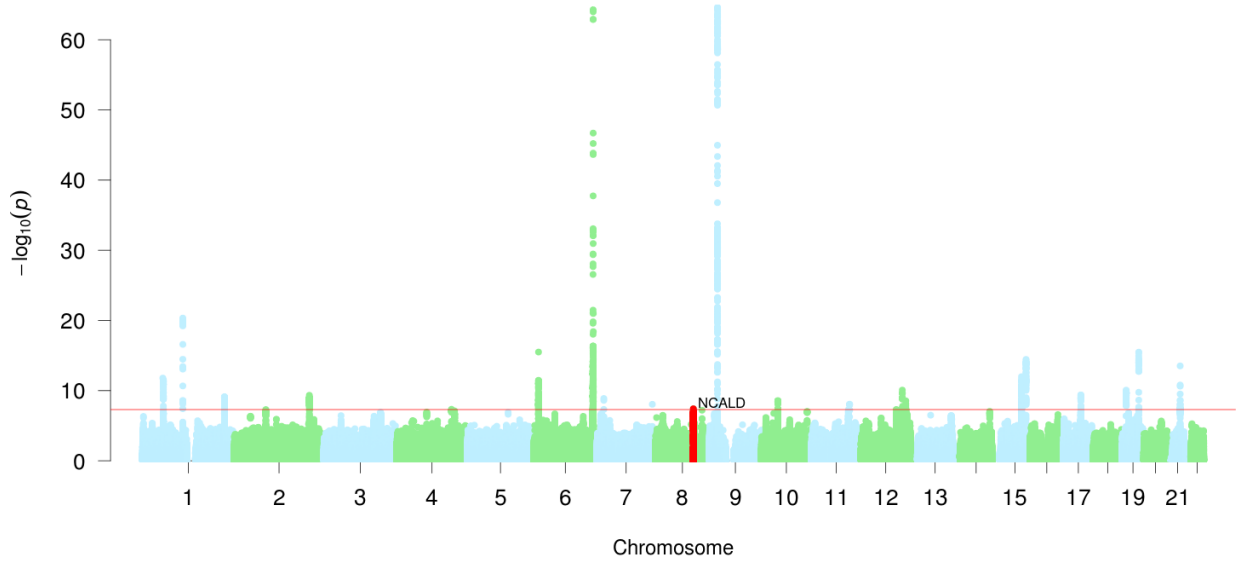
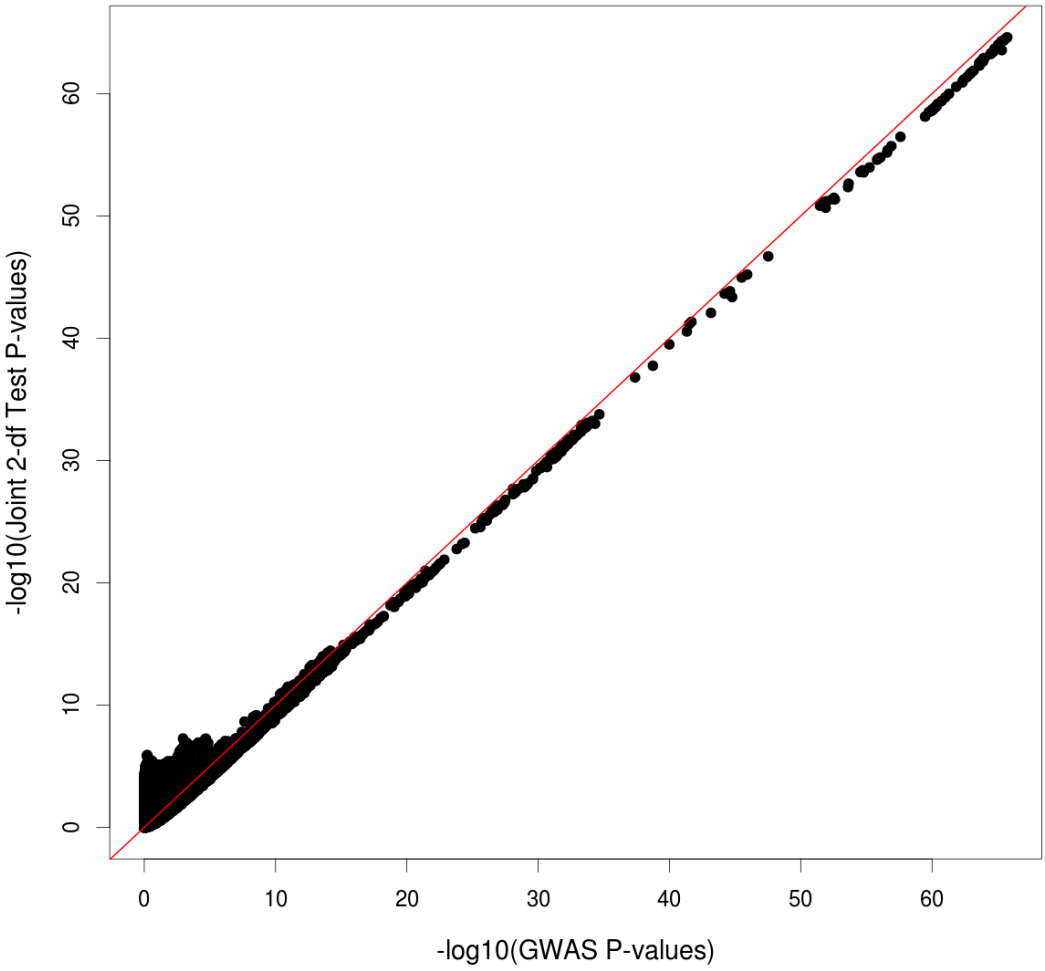
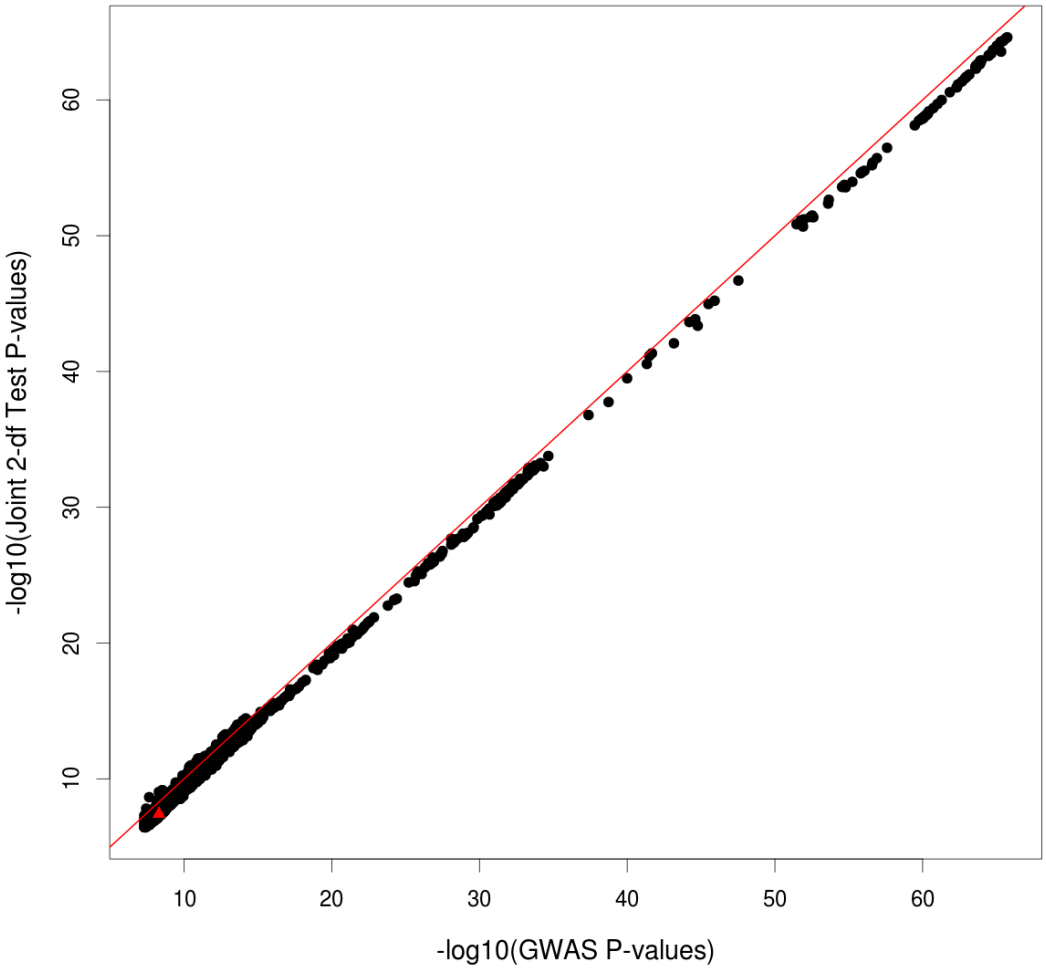


Figure III. Comparisons between p-values of joint 2-df testing approach account for current smoking and p-values using conventional GWAS approach for CAD-GWAS in the UK Biobank

A. All tested SNPs



B. Genome-wide significant SNPs (joint 2-df $p<5\times10^{-8}$) (Highlighted the locus that was not reported in the most recent CAD-GWAS by van der Harst et al.)



C. Suggestively associated SNPs (joint 2-df $5\times10^{-8}<p<10^{-5}$)

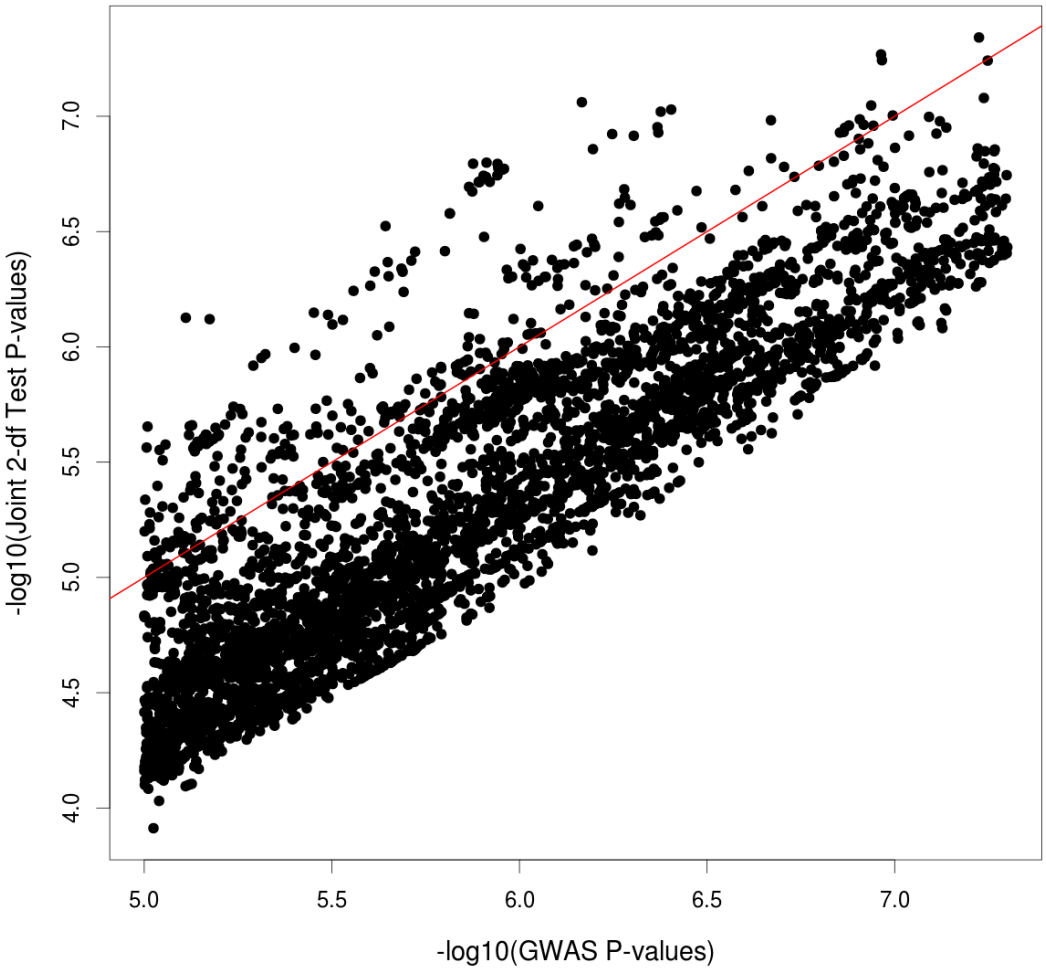
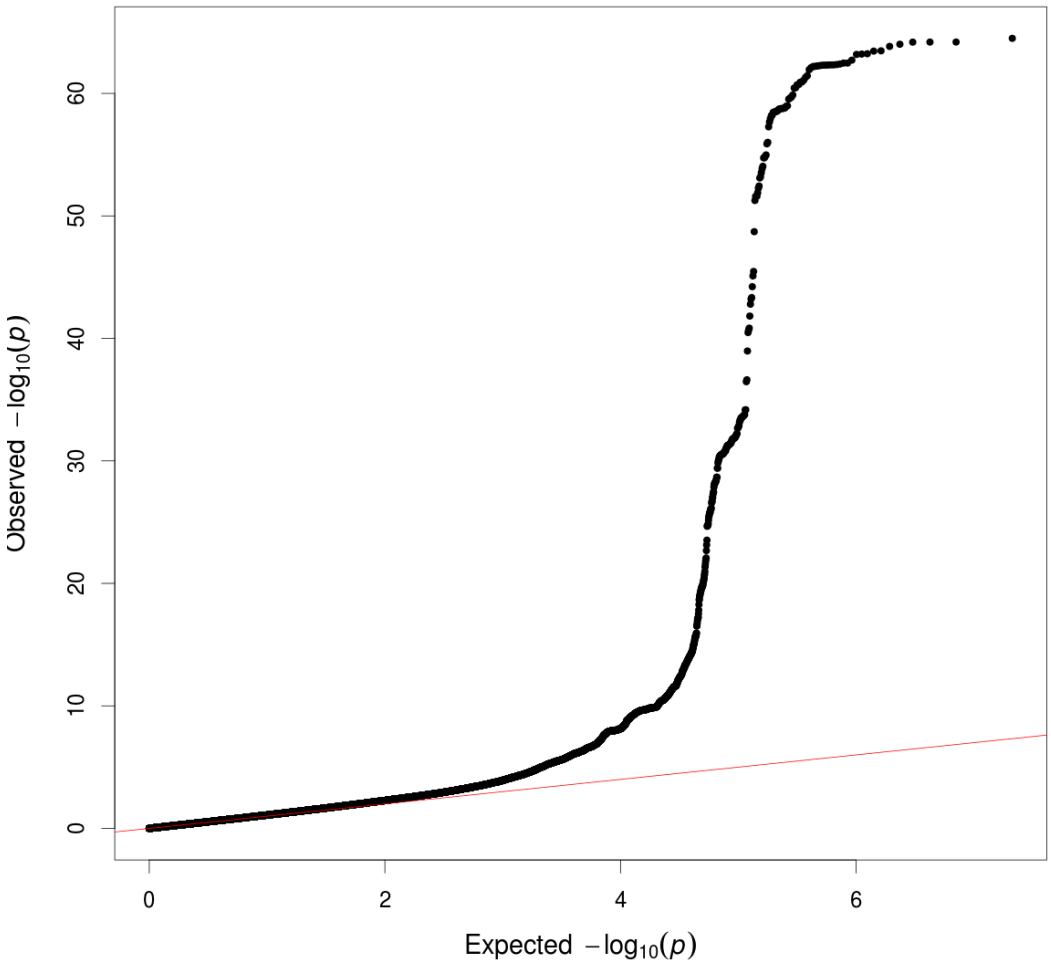


Figure IV. GWAS of CAD in the UK Biobank accounting for physical activity and gene-physical activity interaction

A. Quantile-Quantile Plot (Inflation Factor = 1.12)



B. Manhattan plot (red line: genome-wide significant with $p < 5\times10^{-8}$)

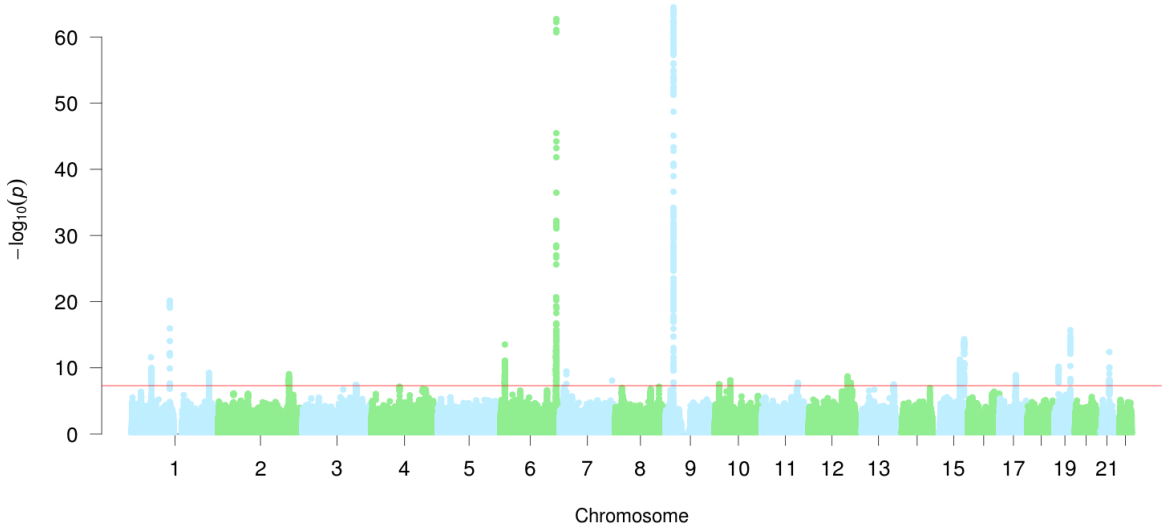
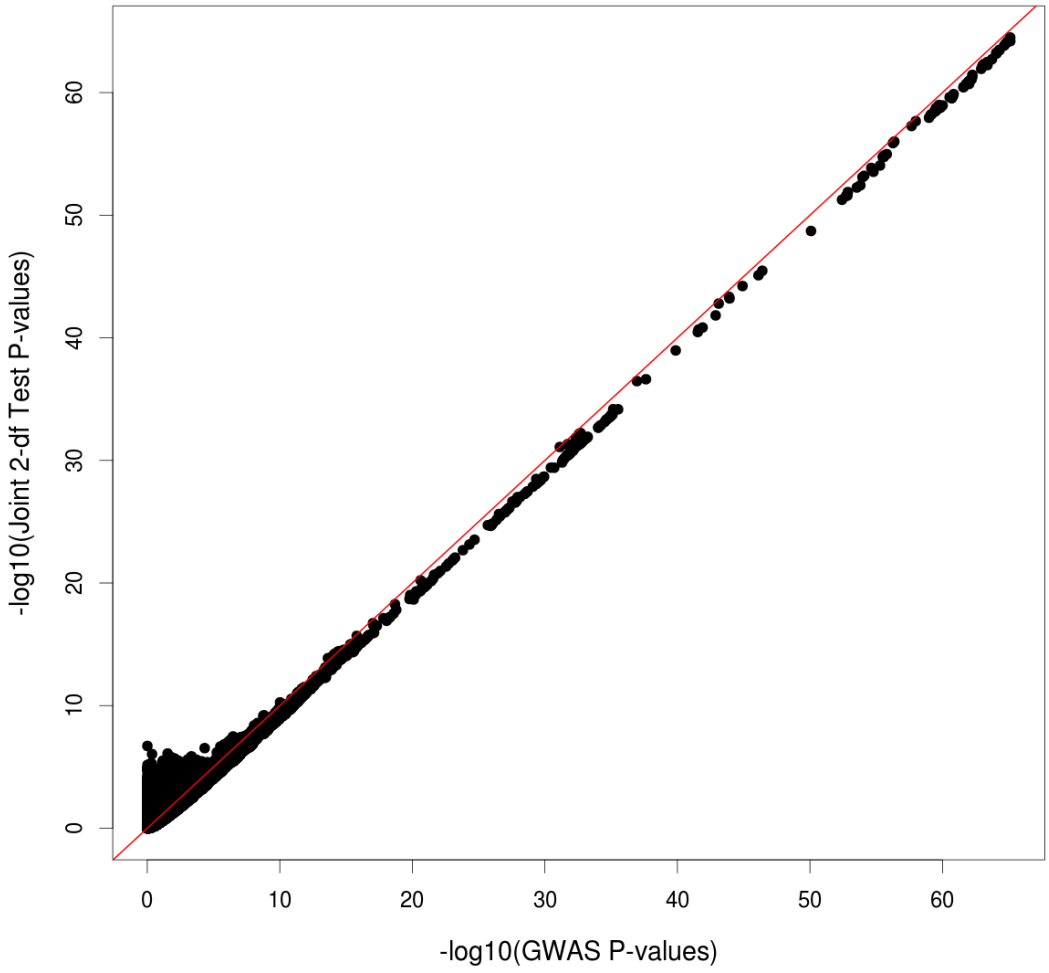
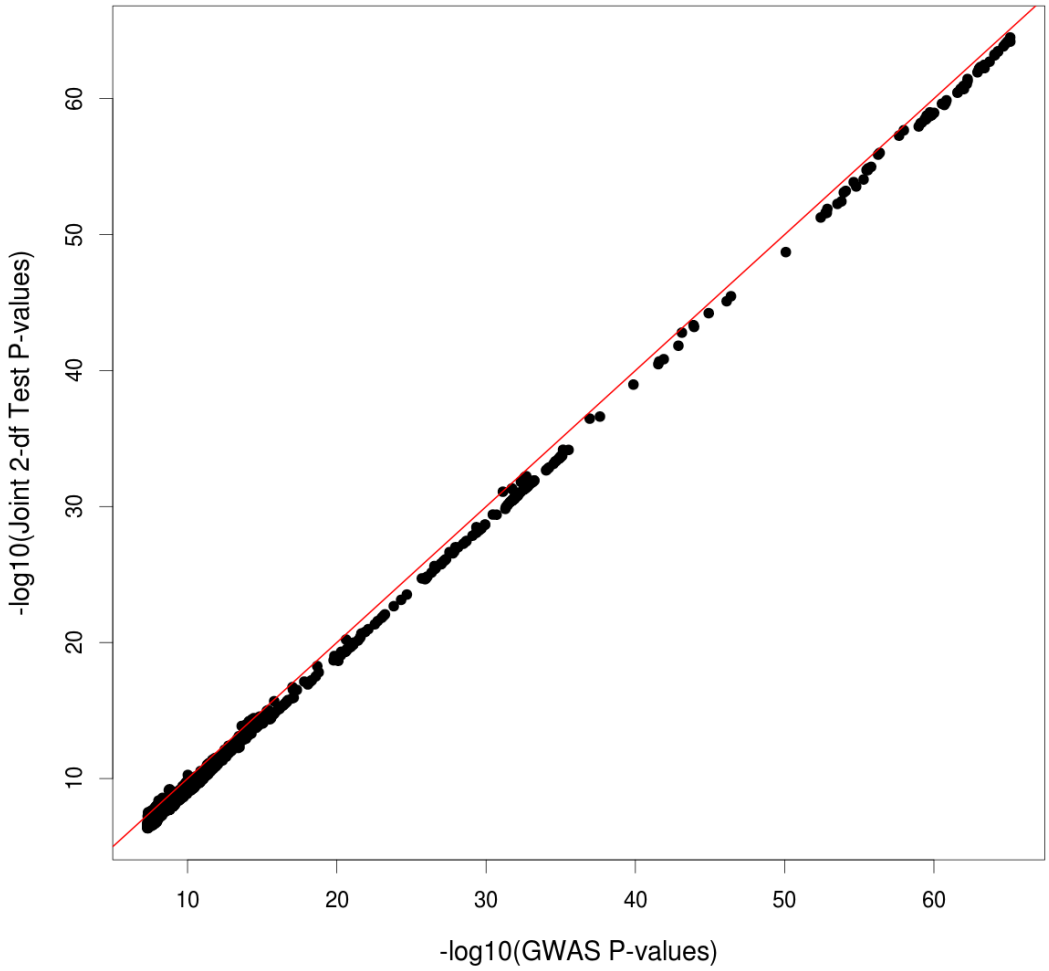


Figure V. Comparisons between p-values of joint 2-df testing approach account for physical activity and p-values using conventional GWAS approach for CAD-GWAS in the UK Biobank

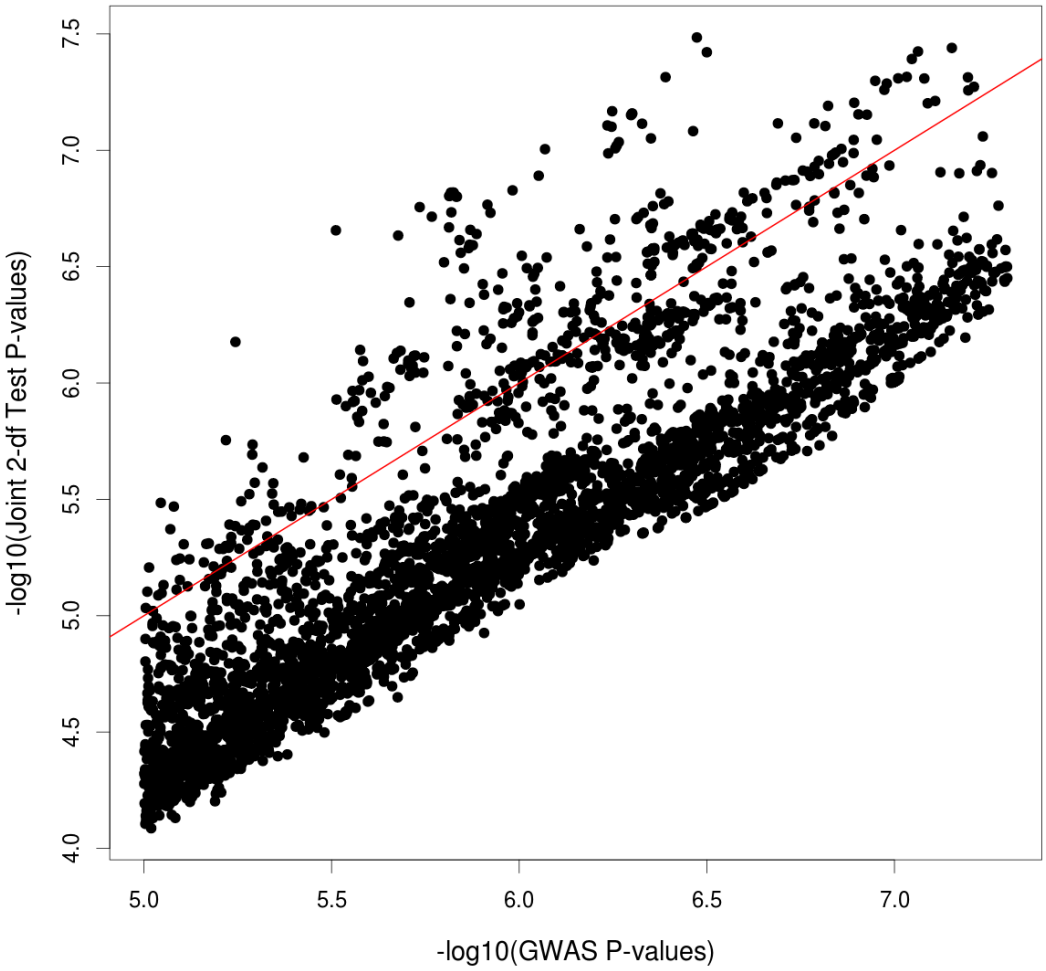
A. All tested SNPs



B. Genome-wide significant SNPs (joint 2-df $p<5\times10^{-8}$)



C. Suggestively associated SNPs (joint 2-df $5\times10^{-8}<p<1^{-5}$)



Chapter 5

Summary and Future Directions

The main goal of the work presented in this dissertation is to assess gene-lifestyle interactions in CAD with a focus on two established lifestyle-related CAD risk factors: smoking and physical inactivity, on both additive and multiplicative scale. Specifically, using data from two of the largest biobank cohorts: the UK Biobank and the Million Veteran Program, several CAD-associated genetic risk scores were constructed and their interactions with smoking and physical inactivity were assessed for incident CAD. This dissertation also aims to identify novel CAD genetic loci by conducting GWAS accounting for gene-smoking interaction and gene-physical activity interaction in a large population of European ancestry from the UK Biobank.

Absolute risk increase of CAD for smoking is more substantial among those with high genetic risk profile

We constructed genetic risk scores for CAD using a most updated list containing up to 161 CAD-associated loci, and also constructed three mediating traits-based (blood pressure, lipids and BMI) CAD genetic risk scores. Assessment of the interaction effects between these scores and smoking status has revealed a super-additive effect between genetic risk and smoking status. We also observed that such interaction effects were driven by blood pressure-associated loci when comparing current smokers with never smokers, but by lipids-associated loci when comparing those with cumulatively high smoking intensity vs. those with low cumulative smoking intensity. Prevalence of tobacco smoking has been consistently declining

but it remains a leading cause of preventable death in the U.S. and globally. (Benjamin, Muntner et al. 2019) Results from this dissertation have provided novel evidence on the interplay between genetic factors and smoking behavior in the development of CAD. Although smoking has long been realized as a risk for CAD, we conclude that the harmful impacts of smoking are higher for those with higher genetic susceptibility of developing CAD. From a public health point of view, elimination of smoking would be beneficial for CAD control, and even more CAD cases would have been prevented if those with high genetic risk hadn't started or quitted smoking. Additive interaction effects were largely omitted in the past gene-environment interaction studies, but it has been suggested that additive interaction effects provide more mechanistic insights than multiplicative interaction. We found one CAD locus at *PCSK9* with strong additive interaction with smoking, but more biological evidence is needed to characterize the role of *PCSK9* as an effect modifier for smoking-caused CAD events.

Accounting for gene-smoking or gene-physical activity interaction did not improve power in the genome-wide discovery of CAD-associated loci

Two GWASs were conducted in the work of this dissertation and we have identified no novel loci of CAD by accounting for gene-smoking or gene-physical activity interaction using a previously proposed joint 2-df testing approach. Previous work in the CHARGE Gene-Lifestyle Interactions Working Group have shown promising results in using the 2-df testing approach for conducting large scale GWAS of CAD risk factors such as blood pressure and lipids, (Feitosa, Kraja et al. 2018, Sung, Winkler et al. 2018, Bentley, Sung et al. 2019, de Vries, Brown et al. 2019) but our results have shown that this approach has added little power comparing to the conventional GWAS approach in the detection of CAD-associated loci

using data from the UK Biobank. However, the implemented joint 2-df approach only accounted for multiplicative interaction effects which might have missed CAD susceptibility loci that act through interaction with smoking and physical activity on the additive scale. Future studies should employ novel statistical methods and computational tools that can overcome such limitation and can be implemented easily on the genome-wide scale in large biobank cohorts.

Final Remarks

Previous gene-lifestyle interaction studies of cardio-metabolic health have shown equivalent benefits of a healthy lifestyle for individuals with different levels of genetic predisposition captured by genetic risk score. (Khera, Emdin et al. 2016, Pazoki, Dehghan et al. 2018, Said, Verweij et al. 2018) However, the majority of them reported on the multiplicative scale and not much attention was paid to individual lifestyle-related risk factors. Although we have concluded that the absolute risk increase due to smoking is higher for those with higher genetic risk of CAD, the clinical utilization of genetic risk score informed lifestyle intervention is still under debate. (Torkamani, Wineinger et al. 2018) It is of large interests to see future studies conducted in assessing how to best utilize genetic information to guide lifestyle interventions and improve cardio-metabolic health. Although it is relatively fast and easy to generate genetic risk scores based on candidate loci, the cost-effectiveness of such approach in improving public health remains to be seen. In addition, disparities remain in the prevalence of lifestyle related risk factors as well as incidence of CAD, but genetic evidence is lacking for many subgroups with high risk. Future gene-lifestyle interaction studies should expand to a multi-ethnic scale with more data enriched for subgroups under higher pressure of CAD development.

References

(2005). "Guidelines for data processing and analysis of the International Physical Activity Questionnaire (IPAQ)-short and long forms." IPAQ Research Committee.

Abifadel, M., M. Varret, J. P. Rabes, D. Allard, K. Ouguerram, M. Devillers, . . . C. Boileau (2003). "Mutations in PCSK9 cause autosomal dominant hypercholesterolemia." Nat Genet **34**(2): 154-156.

Ahmad, S., G. Rukh, T. V. Varga, A. Ali, A. Kurbasic, D. Shungin, . . . P. W. Franks (2013). "Gene x physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry." PLoS Genet **9**(7): e1003607.

Almli, L. M., R. Duncan, H. Feng, D. Ghosh, E. B. Binder, B. Bradley, . . . M. P. Epstein (2014). "Correcting systematic inflation in genetic association tests that consider interaction effects: application to a genome-wide association study of posttraumatic stress disorder." JAMA Psychiatry **71**(12): 1392-1399.

Benjamin, E. J., P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, . . . S. S. Virani (2019). "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association." Circulation **139**(10): e56-e528.

Bentley, A. R., Y. J. Sung, M. R. Brown, T. W. Winkler, A. T. Kraja, I. Ntalla, . . . L. A. Cupples (2019). "Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids." Nat Genet **51**(4): 636-648.

Brown, M. D. (2003). "Exercise and coronary vascular remodelling in the healthy heart." Exp Physiol **88**(5): 645-658.

Bujak, M. and N. G. Frangogiannis (2007). "The role of TGF-beta signaling in myocardial infarction and cardiac remodeling." Cardiovasc Res **74**(2): 184-195.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, . . . J. Marchini (2018). "The UK Biobank resource with deep phenotyping and genomic data." Nature **562**(7726): 203-209.

CDC and NCHS "Multiple Cause of Death 1999-2015 on CDC WONDER Online Database, released December 2016. Data are from the Multiple Cause of Death Files, 1999-2015, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/mcd-icd10.html."

Cohen, J., A. Pertsemlidis, I. K. Kotowski, R. Graham, C. K. Garcia and H. H. Hobbs (2005). "Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9." Nat Genet **37**(2): 161-165.

Cohen, J. C., E. Boerwinkle, T. H. Mosley, Jr. and H. H. Hobbs (2006). "Sequence variations in PCSK9, low LDL, and protection against coronary heart disease." N Engl J Med **354**(12): 1264-1272.

Dai, X., S. Wiernek, J. P. Evans and M. S. Runge (2016). "Genetics of coronary artery disease and myocardial infarction." World J Cardiol **8**(1): 1-23.

Davies, N. M., M. Dickson, G. Davey Smith, G. J. van den Berg and F. Windmeijer (2018). "The Causal Effects of Education on Health Outcomes in the UK Biobank." Nat Hum Behav **2**(2): 117-125.

De Cesaris, R., G. Ranieri, V. Filitti, M. V. Bonfantino and A. Andriani (1992). "Cardiovascular effects of cigarette smoking." Cardiology **81**(4-5): 233-237.

de Vries, P. S., M. R. Brown, A. R. Bentley, Y. J. Sung, T. W. Winkler, I. Ntalla, . . . A. C. Morrison (2019). "Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions." Am J Epidemiol **188**(6): 1033-1054.

Duncan, M. S., M. S. Freiberg, R. A. Greevy, Jr., S. Kundu, R. S. Vasan and H. A. Tindle (2019). "Association of Smoking Cessation With Subsequent Risk of Cardiovascular Disease." Jama **322**(7): 642-650.

Evangelou, E., H. R. Warren, D. Mosen-Ansorena, B. Mifsud, R. Pazoki, H. Gao, . . . M. J. Caulfield (2018). "Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits." Nat Genet **50**(10): 1412-1425.

Fang, H., Q. Hui, J. Lynch, J. Honerlaw, T. L. Assimes, J. Huang, . . . H. Tang (2019). "Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies." Am J Hum Genet **105**(4): 763-772.

Feitosa, M. F., A. T. Kraja, D. I. Chasman, Y. J. Sung, T. W. Winkler, I. Ntalla, . . . D. Levy (2018). "Novel genetic associations for blood pressure identified via gene-alcohol interaction in up to 570K individuals across multiple ancestries." PLoS One **13**(6): e0198166.

Ford, E. S. and S. Capewell (2011). "Proportion of the decline in cardiovascular mortality disease due to prevention versus treatment: public health versus clinical care." Annu Rev Public Health **32**: 5-22.

Gauderman, W. J., B. Mukherjee, H. Aschard, L. Hsu, J. P. Lewinger, C. J. Patel, . . . N. Chatterjee (2017). "Update on the State of the Science for Analytical Methods for Gene-Environment Interactions." Am J Epidemiol **186**(7): 762-770.

Gaziano, J. M., J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, . . . T. J. O'Leary (2016). "Million Veteran Program: A mega-biobank to study genetic influences on health and disease." J Clin Epidemiol **70**: 214-223.

Gepner, A. D., M. E. Piper, H. M. Johnson, M. C. Fiore, T. B. Baker and J. H. Stein (2011). "Effects of smoking and smoking cessation on lipids and lipoproteins: outcomes from a randomized clinical trial." Am Heart J **161**(1): 145-151.

Gordon, K. J. and G. C. Blobe (2008). "Role of transforming growth factor-beta superfamily signaling pathways in human disease." Biochim Biophys Acta **1782**(4): 197-228.

Graff, M., R. A. Scott, A. E. Justice, K. L. Young, M. F. Feitosa, L. Barata, . . . D. Hadley (2017). "Genome-wide physical activity interactions in adiposity—A meta-analysis of 200,452 adults." PLoS genetics **13**(4): e1006528.

Gustavsson, J., K. Mehlig, K. Leander, E. Strandhagen, L. Bjorck, D. S. Thelle, . . . F. Nyberg (2012). "Interaction of apolipoprotein E genotype with smoking and physical inactivity on coronary heart disease risk in men and women." Atherosclerosis **220**(2): 486-492.

Hamrefors, V., B. Hedblad, G. Hindy, J. G. Smith, P. Almgren, G. Engstrom, . . . O. Melander (2014). "Smoking modifies the associated increased risk of future cardiovascular disease by genetic variation on chromosome 9p21." PLoS One **9**(1): e85893.

Helgadottir, A., G. Thorleifsson, A. Manolescu, S. Gretarsdottir, T. Blondal, A. Jonasdottir, . . . K. Stefansson (2007). "A common variant on chromosome 9p21 affects the risk of myocardial infarction." Science **316**(5830): 1491-1493.

Hindy, G., F. Wiberg, P. Almgren, O. Melander and M. Orho-Melander (2018). "Polygenic Risk Score for Coronary Heart Disease Modifies the Elevated Risk by Cigarette Smoking for Disease Incidence." Circ Genom Precis Med **11**(1): e001856.

Holmes, M. V., R. Frikke-Schmidt, D. Melis, R. Luben, F. W. Asselbergs, J. M. Boer, . . . P. J. Talmud (2014). "A systematic review and meta-analysis of 130,000 individuals shows smoking does not modify the association of APOE genotype on risk of coronary heart disease." Atherosclerosis **237**(1): 5-12.

Huxley, R. R. and M. Woodward (2011). "Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies." Lancet **378**(9799): 1297-1305.

Justice, A. E. and T. W. Winkler (2017). "Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits." **8**: 14977.

Justice, A. E., T. W. Winkler, M. F. Feitosa, M. Graff, V. A. Fisher, K. Young, . . . L. A. Cupples (2017). "Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits." Nat Commun **8**: 14977.

Khera, A. V., C. A. Emdin, I. Drake, P. Natarajan, A. G. Bick, N. R. Cook, . . . S. Kathiresan (2016). "Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease." N Engl J Med **375**(24): 2349-2358.

Khera, A. V. and S. Kathiresan (2017). "Genetics of coronary artery disease: discovery, biology and clinical translation." Nat Rev Genet **18**(6): 331-344.

Kichaev, G., G. Bhatia, P. R. Loh, S. Gazal, K. Burch, M. K. Freund, . . . A. L. Price (2019). "Leveraging Polygenic Functional Enrichment to Improve GWAS Power." Am J Hum Genet **104**(1): 65-75.

Klarin, D., S. M. Damrauer, K. Cho, Y. V. Sun, T. M. Teslovich, J. Honerlaw, . . . T. L. Assimes (2018). "Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program." Nat Genet **50**(11): 1514-1523.

Kraft, P., Y.-C. Yen, D. O. Stram, J. Morrison and W. J. Gauderman (2007). "Exploiting gene-environment interaction to detect genetic associations." Human heredity **63**(2): 111-119.

Kraft, P., Y. C. Yen, D. O. Stram, J. Morrison and W. J. Gauderman (2007). "Exploiting gene-environment interaction to detect genetic associations." Hum Hered **63**(2): 111-119.

Langenberg, C., S. J. Sharp, P. W. Franks, R. A. Scott, P. Deloukas, N. G. Forouhi, . . . N. J. Wareham (2014). "Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study." PLoS Med **11**(5): e1001647.

Leischow, S. J. (2019). "Increasing Smoking Cessation in the United States: Expanding the Availability of Over-the-Counter Medications." Jama **321**(6): 541-542.

Li, R. and L. Chambless (2007). "Test for additive interaction in proportional hazards models." Ann Epidemiol **17**(3): 227-236.

Lin, D. Y., R. Tao, W. D. Kalsbeek, D. Zeng, F. Gonzalez, 2nd, L. Fernandez-Rhodes, . . . G. Heiss (2014). "Genetic association analysis under complex survey sampling: the Hispanic Community Health Study/Study of Latinos." Am J Hum Genet **95**(6): 675-688.

Linke, A., S. Erbs and R. Hambrecht (2008). "Effects of exercise training upon endothelial function in patients with cardiovascular disease." Front Biosci **13**: 424-432.

Lloyd-Jones, D. M., Y. Hong, D. Labarthe, D. Mozaffarian, L. J. Appel, L. Van Horn, . . . W. D. Rosamond (2010). "Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond." Circulation **121**(4): 586-613.

Lloyd-Jones, D. M., B. H. Nam, R. B. D'Agostino, Sr., D. Levy, J. M. Murabito, T. J. Wang, . . . C. J. O'Donnell (2004). "Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring." Jama **291**(18): 2204-2211.

Lubin, J. H., D. Couper, P. L. Lutsey, M. Woodward, H. Yatsuya and R. R. Huxley (2016). "Risk of Cardiovascular Disease from Cumulative Cigarette Use and the Impact of Smoking Intensity." Epidemiology **27**(3): 395-404.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, . . . P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

McPherson, R., A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts, D. R. Cox, . . . J. C. Cohen (2007). "A common allele on chromosome 9 associated with coronary heart disease." Science **316**(5830): 1488-1491.

Mega, J. L., N. O. Stitziel, J. G. Smith, D. I. Chasman, M. Caulfield, J. J. Devlin, . . . M. S. Sabatine (2015). "Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials." Lancet **385**(9984): 2264-2271.

Messner, B. and D. Bernhard (2014). "Smoking and cardiovascular disease: mechanisms of endothelial dysfunction and early atherogenesis." Arterioscler Thromb Vasc Biol **34**(3): 509-515.

Mons, U., A. Muezzinler, C. Gellert, B. Schottker, C. C. Abnet, M. Bobak, . . . H. Brenner (2015). "Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium." Bmj **350**: h1551.

Nickels, S., T. Truong, R. Hein, K. Stevens, K. Buck, S. Behrens, . . . J. Chang-Claude (2013). "Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors." PLoS Genet **9**(3): e1003284.

Nikpay, M., A. Goel, H. H. Won, L. M. Hall, C. Willenborg, S. Kanoni, . . . M. Farrall (2015). "A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease." Nat Genet **47**(10): 1121-1130.

Ordovas, J. M. and E. S. Tai (2008). "Why study gene-environment interactions?" Curr Opin Lipidol **19**(2): 158-167.

Parmet, S., T. J. Glass and R. M. Glass (2004). "JAMA patient page. Coronary artery disease." Jama **292**(20): 2540.

Pazoki, R., A. Dehghan, E. Evangelou, H. Warren, H. Gao, M. Caulfield, . . . I. Tzoulaki (2018). "Genetic Predisposition to High Blood Pressure and Lifestyle Factors: Associations With Midlife Blood Pressure Levels and Cardiovascular Events." Circulation **137**(7): 653-661.

Rao, D. C., Y. J. Sung, T. W. Winkler, K. Schwander, I. Borecki, L. A. Cupples, . . . B. M. Psaty (2017). "Multiancestry Study of Gene-Lifestyle Interactions for Cardiovascular Traits in 610 475 Individuals From 124 Cohorts: Design and Rationale." Circ Cardiovasc Genet **10**(3).

Ripatti, S., E. Tikkanen, M. Orho-Melander, A. S. Havulinna, K. Silander, A. Sharma, . . . S. Kathiresan (2010). "A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses." Lancet **376**(9750): 1393-1400.

Ritz, B. R., N. Chatterjee, M. Garcia-Closas, W. J. Gauderman, B. L. Pierce, P. Kraft, . . . K. McAllister (2017). "Lessons Learned From Past Gene-Environment Interaction Successes." Am J Epidemiol **186**(7): 778-786.

Said, M. A., N. Verweij and P. van der Harst (2018). "Associations of Combined Genetic and Lifestyle Risks With Incident Cardiovascular Disease and Diabetes in the UK Biobank Study." JAMA Cardiol **3**(8): 693-702.

Saleheen, D., W. Zhao, R. Young, C. P. Nelson, W. Ho, J. F. Ferguson, . . . M. P. Reilly (2017). "Loss of Cardioprotective Effects at the ADAMTS7 Locus as a Result of Gene-Smoking Interactions." Circulation **135**(24): 2336-2353.

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, . . . R. Collins (2015). "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." PLoS Med **12**(3): e1001779.

Sung, Y. J., T. W. Winkler, L. de Las Fuentes, A. R. Bentley, M. R. Brown, A. T. Kraja, . . . D. I. Chasman (2018). "A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure." Am J Hum Genet **102**(3): 375-400.

Tanasescu, M., M. F. Leitzmann, E. B. Rimm, W. C. Willett, M. J. Stampfer and F. B. Hu (2002). "Exercise type and intensity in relation to coronary heart disease in men." Jama **288**(16): 1994-2000.

Taylor, R. S., A. Brown, S. Ebrahim, J. Jolliffe, H. Noorani, K. Rees, . . . N. Oldridge (2004). "Exercise-based rehabilitation for patients with coronary heart disease: systematic review and meta-analysis of randomized controlled trials." Am J Med **116**(10): 682-692.

Tikkanen, E., S. Gustafsson and E. Ingelsson (2018). "Associations of Fitness, Physical Activity, Strength, and Genetic Risk With Cardiovascular Disease: Longitudinal Analyses in the UK Biobank Study." Circulation **137**(24): 2583-2591.

Torkamani, A., N. E. Wineinger and E. J. Topol (2018). "The personal and clinical utility of polygenic risk scores." Nat Rev Genet **19**(9): 581-590.

Usset, J. L., R. Raghavan, J. P. Tyrer, V. McGuire, W. Sieh, P. Webb, . . . B. L. Fridley (2016). "Assessment of Multifactor Gene-Environment Interactions and Ovarian Cancer Risk: Candidate Genes, Obesity, and Hormone-Related Risk Factors." Cancer Epidemiol Biomarkers Prev **25**(5): 780-790.

van der Harst, P. and N. Verweij (2018). "Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease." Circ Res **122**(3): 433-443.

VanderWeele, T. J. (2009). "Sufficient cause interactions and statistical interactions." Epidemiology **20**(1): 6-13.

VanderWeele, T. J. and M. J. Knol (2014). A Tutorial on Interaction. Epidemiologic Methods. **3:** 33.

Warren, H. R., E. Evangelou, C. P. Cabrera, H. Gao, M. Ren, B. Mifsud, . . . L. V. Wain (2017). "Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk." Nat Genet **49**(3): 403-415.

Waterworth, D. M., S. L. Ricketts, K. Song, L. Chen, J. H. Zhao, S. Ripatti, . . . M. S. Sandhu (2010). "Genetic variants influencing circulating lipid levels and risk of coronary artery disease." Arterioscler Thromb Vasc Biol **30**(11): 2264-2276.

Webb, T. R., J. Erdmann, K. E. Stirrups, N. O. Stitziel, N. G. Masca, H. Jansen, . . . S. Kathiresan (2017). "Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease." J Am Coll Cardiol **69**(7): 823-836.

Willett, W. C. (2002). "Balancing life-style and genomics research for disease prevention." Science **296**(5568): 695-698.

Wilson, M. G., G. M. Ellison and N. T. Cable (2016). "Basic science behind the cardiovascular benefits of exercise." Br J Sports Med **50**(2): 93-99.

Winzer, E. B., F. Woitek and A. Linke (2018). "Physical Activity in the Prevention and Treatment of Coronary Artery Disease." J Am Heart Assoc **7**(4).

Yusuf, S., S. Hawken, S. Ounpuu, T. Dans, A. Avezum, F. Lanas, . . . L. Lisheng (2004). "Effect of potentially modifiable risk factors associated with myocardial infarction in

52 countries (the INTERHEART study): case-control study." Lancet **364**(9438): 937-952.

Zeng, L., T. A. Dang and H. Schunkert (2016). "Genetics links between transforming growth factor beta pathway and coronary disease." Atherosclerosis **253**: 237-246.