

POLYGENIC RISK SCORES FOR THE PREDICTION OF CARDIOVASCULAR  
DISEASES

NOVEL STATISTICAL METHODS FOR POLYGENIC  
RISK SCORE GENERATION IN CARDIOVASCULAR  
DISEASES

By Ann LE,

*A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of  
the Requirements for the Degree Doctor of Philosophy in Medical Sciences*

McMaster University © Copyright by Ann LE December 29, 2024

McMaster University

Doctor of Philosophy (2024)

Hamilton, Ontario (Department of Medical Sciences)

TITLE: NOVEL STATISTICAL METHODS FOR POLYGENIC RISK SCORE GENERATION IN CARDIOVASCULAR DISEASES

AUTHOR: Ann LE (McMaster University)

SUPERVISOR: Dr. Guillaume PARÉ

NUMBER OF PAGES: xiv, 258

## Lay Abstract

Many common diseases, like coronary artery disease (CAD) and diabetes, are influenced not only by lifestyle and environmental factors, but also by genetics. Therefore, incorporating genetic information into disease risk prediction for patients in clinical settings would be logical, especially since genetic data can be obtained early in life. One tool for quantifying risk based on genetics is the polygenic risk score (PRS). PRS assigns a numerical value based on an individual's genetic profile, calculated by summing up risk variants in their DNA. The risk level corresponds to the variant's association with the trait, as determined by genome-wide association studies (GWAS). PRS have become increasingly popular for guiding disease treatment and personalized medicine. However, there's still work to be done to make PRS suitable for clinical use. Many methods have attempted to enhance the predictive ability of PRS, but there's still room for improvement. This thesis introduces various applications for PRS, along with a novel prediction method that potentially addresses some limitations and explores the applications of PRS in common diseases.

# Abstract

Polygenic risk scores (PRS) are relatively novel tools for risk prediction, serving as a quantitative singular value which depicts a patient's genetic disposition for a certain disease. Given that many current clinical risk predictors do not address heritability within their calculations, PRS are likely to improve prediction, especially in the case of complex diseases which are influenced by a combination of genetic, environmental and lifestyle factors. Altogether, PRS studies have been pursued for their abilities in trait detection, therapeutic intervention and disease protection, with much potential in personalized/precision medicine where each interpretation is unique and based on a patient's genotype. However, despite the numerous advances over years, PRS have yet to reach the level where they can be implemented into standard clinical practices as originally intended. The goal is to develop PRS which are applicable to global populations, which has yet to be achieved due to the inconsistency and general skepticism regarding the method. Furthermore, PRS have yet to reach the upper threshold for risk prediction, as indicated by the heritability that remains unaccounted for with PRS calculations. Thus, this thesis addresses how PRS can inform and guide clinical decision-making for complex decisions with strong, genetic dispositions. It also presents novel approach to PRS aimed at mitigating some of its current limitations.

## Acknowledgements

First and foremost, I would like to express my deepest gratitude to Dr. Guillaume Paré. His mentorship and guidance in the Genetics and Molecular Epidemiology Lab (GMEL) have been fundamental in my academic journey, and his insights and expertise were essential for these complex and rewarding projects. I am eternally grateful for the opportunities to utilize my background in mathematics and statistics for applications in genetics and epidemiology, allowing me to practice the interdisciplinary sciences as I'd always intended. It has truly been a privilege to have had his mentorship over the years.

I am also sincerely thankful to my supervisory committee members, Dr. Angelo Canty and Dr. Darryl Leong. Your guidance, expertise, and thoughtful feedback have enriched my work, and every discussion we shared has profoundly enhanced my understanding of my research. I deeply appreciate your unwavering support throughout this journey.

I would also like to acknowledge the members of the GMEL lab. As colleagues and friends, each of you contributed unique perspectives and support that I will carry with me. I'm grateful for the conversations and camaraderie that propelled me forward. Finally, I extend my heartfelt thanks to all those who have supported and encouraged me through the ups and downs of this PhD journey, particularly during the most challenging moments. I couldn't have done this without you all.

# Table of Contents

<b>1</b>	<b>General Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.1.1	Genetic Architecture in Complex Diseases . . . . .	1
1.1.2	Polygenic Risk Scores . . . . .	11
1.1.3	Study Populations . . . . .	32
1.1.4	Standards for Reporting PRS & Quality Control . . . . .	35
	References . . . . .	39
<b>2</b>	<b>Hypothesis</b>	<b>66</b>
2.1	General Hypothesis, Objective & Approach . . . . .	66
2.1.1	General Hypothesis . . . . .	66
2.1.2	General Objectives . . . . .	66
2.1.3	Rationale and Approach . . . . .	67
<b>3</b>	<b>What Causes Premature Coronary Artery Disease?</b>	<b>68</b>
3.1	Forward . . . . .	69
3.2	Abstract . . . . .	71
3.3	Condensed Abstract . . . . .	72
3.4	Introduction . . . . .	72
3.5	Genetics of pCAD . . . . .	73

3.5.1	Monogenic Causes of pCAD . . . . .	75
3.6	Polygenic Causes of pCAD . . . . .	81
3.6.1	Common Genetic Variant Studies . . . . .	81
3.6.2	Rare Genetic Variant Studies . . . . .	83
3.7	Clonal Hematopoiesis of Indeterminate Potential (CHIP) . . . . .	84
3.8	Non-Genetic Risk Factors of pCAD . . . . .	86
3.8.1	Clinical Risk Factors & Clinical Risk Scores . . . . .	86
3.9	Smoking & Other Drugs of Abuse . . . . .	87
3.9.1	Smoking . . . . .	87
3.9.2	Opioid Usage . . . . .	88
3.9.3	Alcohol . . . . .	90
3.9.4	Amphetamines . . . . .	90
3.9.5	Stress and Exercise . . . . .	91
3.10	Spontaneous Coronary Artery Dissection (SCAD) . . . . .	92
3.11	Conclusion . . . . .	92
	References . . . . .	94
<b>4</b>	<b>Polygenic risk scores in Myocardial Injury after Non-cardiac Surgery:</b>	
	<b>a VISION substudy</b>	<b>125</b>
4.1	Forward . . . . .	126
4.2	Abstract . . . . .	128
4.3	Condensed Abstract . . . . .	130
4.4	Introduction . . . . .	130
4.5	Methods . . . . .	132
4.5.1	Study Population & Definition . . . . .	132
4.6	Calculation and Derivation of Polygenic Risk Scores . . . . .	135

4.7	Statistical Analyses . . . . .	136
4.8	Results . . . . .	137
4.9	Discussion . . . . .	143
	References . . . . .	147
4.10	Supplementary Materials . . . . .	153
<b>5</b>	<b>Performance of polygenic risk score methodologies in the absence of external GWAS summary statistics</b>	<b>161</b>
5.1	Forward . . . . .	163
5.2	Abstract . . . . .	165
5.3	Condensed Abstract . . . . .	166
5.4	Introduction . . . . .	167
5.5	Methods . . . . .	169
5.5.1	Study Populations . . . . .	169
5.5.2	Genome-wide association study (GWAS) data . . . . .	170
5.5.3	Polygenic Risk Score Methodologies . . . . .	171
5.5.4	Internal UKB Genotype Association . . . . .	172
5.5.5	EX-TERR Pipeline . . . . .	172
5.5.6	Principal Component Analysis (PCA) Technique for Dimension Re- duction . . . . .	173
5.5.7	Multiple Train-Test Split: Participant & Genotype Levels . . . . .	174
5.5.8	Discrimination & Calibration Tests . . . . .	175
5.6	Results . . . . .	176
5.7	Discussion . . . . .	190
5.8	Conclusion . . . . .	193
	References . . . . .	195

<b>6 Conclusion</b>	<b>243</b>
6.1 General Overview . . . . .	243
6.2 Chapter 3 Overview . . . . .	244
6.3 Chapter 4 Overview . . . . .	245
6.4 Chapter 5 Overview . . . . .	246
6.5 Clinical & Research Implications . . . . .	247
6.5.1 Genetics Risk Prediction of Complex Traits in Clinical Settings . . .	247
6.5.2 Comparison of Different Leading PRS Methods . . . . .	248
6.5.3 Availability of External GWAS for PRS . . . . .	249
6.5.4 Novel Insights for Genetic and Biological Pathways . . . . .	250
6.5.5 Extension to Other Diseases & Traits . . . . .	251
6.6 Limitations & Considerations . . . . .	251
6.7 Conclusion . . . . .	255
References . . . . .	256

# List of Figures

- 1.1 Typical workflow for utilizing machine learning for creating multi-PRS. . . . 18
- 3.1 This review focuses on the genetic and non-genetic causes of premature coronary artery disease (pCAD). The genetic risk factors include monogenic and polygenic causes, including both common and rare variants. Environmental risk factors include smoking, drug usage and lifestyle choices (stress/exercise). 74
- 4.1 **Overview of Experimental Design for VISION MINS PRS study.**  
This figure shows the flow of participant selection, along with the analyses that were conducted on the sample to determine the association between PRS and MINS. The VISION Biobank sample originally consists of 4,428 patients, from which 600 patients were randomly selected for case-control matching. Amongst the 600 patients, those without matches, misrepresented ancestry and lack of RCRI availability were filtered out, resulting in a final sample of 253 cases (patients with MINS) and 253 controls (patients without MINS) matched for age and sex. Conditional logistic regression and discrimination capacity analyses were performed on this sample. . . . . 134

4.2 **MINS Odds Ratio according to quintile of HbA1c PRS.** The figure displays the odds ratio of association per quintile of HbA1c PRS, with the 1<sup>st</sup> quintile as a reference. The HbA1c PRS can stratify MINS risk without RCRI. Confidence interval bars for logsitic regression estimates are also shown. 138

4.3 **Association between PRS and MINS.** The following forest plots display the association of all traits for which PRS were created by order of descending odds ratio with a 95% confidence interval, as determined through conditional logistic regression. Figure 3a. shows association of each traits PRS without adjustment for RCRI. Figure 3b. shows association of each traits PRS with adjustment for RCRI. . . . . 142

5.1 **Description of PRS methodologies with masking process.** LDpred2 is used to create single-trait PRS applied to each of these methods. From left to right, the PRS techniques being compared are 1) baseline: the best-performing single-trait PRS, 2) PRS<sub>multi</sub>: a multi-trait PRS based in elastic net regression and 3) EX-TERR: a multi-trait PRS based in Multi Adaptive Regression Splines (MARS). Masking (exclusion of GWAS directly matching to outcome) is performed across all methods, to simulate the context of no external GWAS corresponding to the target outcome. The GWAS used in the analysis were required to meet the following criteria: inclusion of at least 5,000,000 SNPs, exclusion of UKB data, and being the most up-to-date summary statistics available. . . . . 180

**5.2 Overview of EX-TERR pipeline.** A total of 408,182 related participants from the UK Biobank (UKB) were used to create EX-TERR PRS. Regression coefficients are obtained from external GWAS, and linear or logistic regression between UKB genotypes and outcomes. These coefficients are converted into polygenic risk score (PRS) weights using LDpred2. Genotypes are divided into approximately 5,000 single-nucleotide polymorphism (SNP) blocks, and within these blocks, rotations are performed using rotational matrix ( $\mathbb{V}_n$ ) derived from the UKB training set genotypes (coefficient matrix  $\times \mathbb{V}_n$ ). The predictor for the MARS regression is the rotated external GWAS LDpred2 weights, and the outcome are rotated UKB LDpred2 weights. A secondary train/test split is conducted through 5-fold cross-validation in the predictors. This process allows the MARS weights to be applied to both training and/or validation sets for PRS generation. . . . . 181

**5.3 Single-trait PRS performance with (unmasked) and without (masked) its corresponding external GWAS. a) Continuous outcomes.** Comparative performance of single-trait, baseline PRS with and without masking of the matching GWAS corresponding to the outcome. A list of outcomes with matching GWAS traits is presented in Supplementary Table S5.5. . . . 182

**5.4 Predictive performance of three PRS methodologies for 69 outcomes under the masked condition.** Forest plot illustrating predictive performances of the three PRS methodologies: baseline, PRS<sub>multi</sub> and EX-TERR though logistic regression. Error bars are for 95% confidence intervals (CI). **a) Continuous outcomes.** Forest plot of adjusted  $r^2$  with Cohens 95% confidence intervals (CI) for continuous traits. . . . . 185

5.5	<b>Association of masked baseline PRS, masked PRS<sub>multi</sub>, masked EX-TERR PRS, and unmasked baseline PRS with diabetes mellitus.</b> a) Prevalence of diabetes mellitus according to PRS percentile. b) Odds ratio of diabetes per quintile of PRS. The first quintile acts as the reference, and standard error bars are for 95% confidence intervals. . . . .	222
5.6	b) Odds ratio of diabetes per quintile of PRS. The first quintile acts as the reference, and standard error bars are for 95% confidence intervals. . . . .	223
5.7	<b>Traits of highest importance for diabetes mellitus (DM) as determined by EX-TERR.</b> Variable importance plots (VIPs) for a single fold instance of EX-TERR for diabetes mellitus. Outcome structure is denoted as binary or continuous. Results are presented as residual sum of squares (RSS). . . . .	224
S5.1	<b>Pipeline for assessing correlation of phenotypes for masking.</b> Process of determining correlation between outcomes to determine which traits should be masked. This was originally done for 71 outcomes (two were removed due to insufficient data.). The cor() function in R is used to determine pairwise correlation. Results are shown in Supplementary Table S3 and final masked GWAS are reported in Supplementary Table S4. . . . .	239

S5.2	EX-TERR earth visualization of genetic effects on target outcome. Partial dependence plots (PDP) for the top six significant traits for a single cross-validation fold. For the continuous traits (TG HDLs), the peaks are center around (0,0) and reflect the general direction of the genetic effect of these predictors on the DM outcome. The TG partial dependence plot (PDP) shows positive DM values, with a negative slope at lower values and a positive slope at higher values. This suggests that as TG coefficients reach extreme values, the effect on DM increases, indicating a deleterious effect. Conversely, HDL coefficients show a decreasing effect or negative DM estimates at extreme values, implying a protective effect. For dichotomous traits such as peripheral artery disease (PAD), heart failure, and attention deficit hyperactivity disorder (ADHD), there is a positive correlation with DM estimators, indicating that there is an estimated increase of disease risk as the trait coefficients increase. . . . .	240
S5.3	Simulation results for MARS algorithm as performed by EX-TERR in one fold for diabetes mellitus outcome (threshold = 0.6). a. Term selection b. Cumulative distribution of residuals c. Fitted vs. residuals d. QQ plot . . .	241
S5.4	<b>Goodness of fit calibration results for EX-TERR PRS with diabetes mellitus (DM) outcome.</b> A Hosmer-Lemeshow plot depicting the calibration test between observed and expected values for the DM outcome and its corresponding EX-TERR PRS. The accompanying table reports results for DM outcome across the three methodologies. In this example, PRS <sub>multi</sub> is not significantly calibrated, while both baseline and EX-TERR PRS are significantly calibrated. . . . .	242

# List of Tables

3.1	Monogenic diseases associated with pCAD pathogenesis. . . . .	80
4.1	Baseline characteristics of VISION Biobank case-control participants of European ancestry. . . . .	139
4.2	Association between CAD PRS and preoperative CAD and T2D PRS and HbA1c PRS with preoperative T2D. . . . .	139
4.3	Polygenic risk score and revised cardiac risk index (RCRI) in patients with and without MINS. . . . .	140
4.4	Logistic regression models studying association between revised cardiac risk index (RCRI) score, polygenic risk scores (PRS) and myocardial injury after non-cardiac surgery (MINS). All models are adjusted for 10 PCs accounting for genetic ancestry as a confounder. . . . .	141
S4.1	List of genome-wide summary statistics consortium summary statistics. . .	157
S4.2	Discriminative capacity using c-statistic (with 95% confidence intervals) in conditional logistic regressions for MINS within 30 days after surgery among participants in the T2D PRS. Figure S2a. shows discriminative capacity of CAD PRS, S2b. for T2D PRS and S2c. for HbA1c PRS. . . . .	158
S4.3	DeLong analyses to discriminative capacity significance for PRS within each subset. . . . .	159

S4.4 Discriminative capacity using Net Reclassification Improvement (NRI) in logistic regressions for MINS within 30 days after surgery among participants for RCRI with the addition of PRS. . . . .	160
5.1 <b>Proportion of best performing PRS methodology.</b> Proportion of best performing methodology compared between baseline, PRS <sub>multi</sub> , and EX-TERR, categorized by continuous or dichotomous traits. Percentages represent the proportion of instances where each score performs best within each category, with the corresponding count shown in brackets. . . . .	177
5.2 <b>PRS performance with and without masking of GWAS information matching to the outcome.</b> Performance for masked and unmasked PRS is reported across the three methodologies: baseline, PRS <sub>multi</sub> , and EX-TERR. PRS performance for continuous outcomes is reported in adjusted $r^2$ values, while performance for dichotomous outcomes is reported as odd ratios. A list of GWAS considered matching can be referred to in Supplementary Table S5.5. HbA1c = Glycated haemoglobin, HDL = High density lipoprotein, LDL = Low density lipoprotein, CAD = Coronary artery disease, DM = Diabetes mellitus, AFF = Atrial Fibrillation and flutter, MI = Myocardial infarction, T2D = Type II diabetes . . . . .	179
S5.1 Baseline characteristics for UK Biobank British related participants (n = 408 160). Standard deviations shown for continuous traits and percentage of total participants shown for dichotomous traits. . . . .	203
S5.2 Complete list of 61 external genome-wide association study (GWAS) summary statistics utilized in the analysis. . . . .	207

S5.3	Pairs of phenotypes with high ( $R^2 > 0.7$ ) correlation. Used to aid in decision of trait masking i.e. determine which traits were closely related enough for masking. . . . .	210
S5.4	Complete list of outcome list with their definitions and their corresponding masked traits. . . . .	211
S5.5	Matching GWAS for each outcome where available for PRS generation. The data consortium is specified to clarify situations in which duplicate GWAS summary statistics exist. Supplementary Table S5.2 provides full details regarding GWAS information. . . . .	225
S5.6	<b>Significance of each methodology’s PRS to each outcome.</b> $p$ -values for multivariate regression models between the three PRS for baseline, PRS <sub>multi</sub> , and EX-TERR and their corresponding outcomes. Logistic regression was performed for dichotomous outcomes and linear regression for continuous outcomes. All models adjusted age, sex, and 10 principal components (PCs). Significance denoted with (*). . . . .	226
S5.7	Calibration for all outcomes, comparison between PRS <sub>multi</sub> and EX-TERR. Fig a. Residual Mean Squared Error (RMSE) for continuous traits. The RMSE difference is calculated as PRS <sub>multi</sub> - EX-TERR. A lower RMSE statistic indicates a better fit. . . . .	229
S5.9	Discriminative capacities of PRS <sub>multi</sub> and EX-TERR PRS for each available outcome, with Age + Sex as the base model. . . . .	231
S5.10a.	<b>Continuous traits (regression coefficient, adjusted <math>R^2</math>.)</b> PRS results in validation set for baseline PRS, multi LDpred2 PRS and EX-TERR PRS (earth degree = 1). All results are highly significant (Rotation SD = 0.6, $p$ -value $\ll 0.01$ ) . . . . .	232

S5.12 Best-performing PRS after masking for each of the 69 outcomes in the validation set. . . . .	237
--	-----

McMASTER UNIVERSITY

DEPARTMENT OF

MEDICAL SCIENCES

The undersigned hereby certify that they have read the thesis entitled "**Novel statistical methods using Multi Adaptive Regression Splines to optimize predictiveness of Polygenic Risk Scores**" by **Ann Le** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dated: \_\_\_\_\_

Supervisor: \_\_\_\_\_

Dr. Guillaume Paré

Second Reader: \_\_\_\_\_

Dr. Angelo Canty / Dr. Darryl Leong

# McMASTER UNIVERSITY

SEPTEMBER 2024

AUTHOR: ANN LE  
TITLE: NOVEL STATISTICAL METHODS USING MULTI ADAPTIVE  
REGRESSION SPLINES TO OPTIMIZE PREDICTIVENESS OF  
POLYGENIC RISK SCORES FOR COMPLEX DISEASES  
DEPARTMENT: MEDICAL SCIENCES  
FACULTY: HEALTH SCIENCES  
CONVOCATION: SEPTEMBER 2024

I, Ann LE, declare that this thesis title, NOVEL STATISTICAL METHODS FOR POLYGENIC RISK SCORE GENERATION IN CARDIOVASCULAR DISEASES and the work presented in it are my own and has not been presented for any other degree, published or submitted for examination in this or any other university. Permission is herewith granted to McMaster University to circulate and to have copied for non-commercial purposes, at its direction, the above title upon request of individuals or institutions.

---

Ann Le

# Chapter 1

## General Introduction

### 1.1 Background

#### 1.1.1 Genetic Architecture in Complex Diseases

The field of genetic epidemiology has been rapidly expanding over the past few years. Genetic epidemiology focuses explicitly on genetics, usually within a general population context, with the ultimate goal of addressing both new and existing challenges related to diseases [1]. The substantial growth in this field can be attributed to the drastic advancements in technological innovations over the last few decades. This includes improvements in genomic technologies, increased capacity for data storage enabling larger population-based studies, novel statistical methodologies, and overall increased efficiencies in computational processes. The focus on genetics is especially notable for its potential for precision medicine and disease risk prediction. This is especially relevant for diseases of early-onset, due to the accessibility of genotypic information at younger ages. The field of genetic epidemiol-

ogy combines elements of biological, statistical and computing approaches to gain insights into disease etiologies and their global impact. Advancements in technology have significantly transformed this field over the years, facilitating improved processing and analysis of genomic information. The focus on genetics previously demonstrated its utility in medical and clinical settings, particularly in cases of monogenic diseases where single genes or chromosome numbers are affected [2, 3]. Typically genetic predisposition is measured by heritability, which is defined as the proportion of variance in a trait that can be ascribed to genetic factors [4, 5]. Many prevalent diseases exhibit high heritability. These include coronary artery disease (35%-57%) [4, 6], type II diabetes (25-80%) [7, 8, 9] and Alzheimer's disease ( $\approx 70\%$ )[10, 11].

A Mendelian, or monogenic, disorder is typically determined by the mutation of a single gene [3, 12]. While the risk of most diseases are commonly associated with rare monogenic variants, virtually all conditions are inherently complex, involving a combination of effects from the environment or other genes [13]. In actuality, most traits exhibit polygenicity, in which they are influenced by several different genetic loci to varying degrees [12]. Polygenic risk differs from monogenic risk in that, rather than a small number of rare, high-effect mutations contributing to the overall risk of disease, polygenic risk reflects the cumulative effect of many common variants, each with lower individual effects [1, 14].

Detection of trait polygenicity has been greatly expedited through the recent developments of genome-wide association studies (GWAS), which identify associations between common genetic variants and phenotypic traits [15, 16]. The advent of these studies has allowed for discoveries, such as the 9p21 locus most strongly associated with coronary artery disease (CAD) and myocardial infarction (MI) in 2007 when the first CAD GWAS was performed [17, 18, 19]. GWAS tend to adhere to the common disease/common variant

(CDCV) hypothesis which posits that if a heritable disease is common within a population, then the variants which contribute to said disease will also be common to the population [1, 20]. Contrarily, the aggregation of rare alleles (common disease/rare variant [CDRV]) has also gained relevance due to the improvement in technological ability to detect variants of rarer frequencies. Regardless, aggregation of multiple variants has proven to be effective in risk prediction. Numerous methodologies have developed to analyze polygenic risk, as will be mentioned proceedingly.

#### **1.1.1.1 Cardiovascular Disease & Coronary Artery Disease**

With approximately 375,000 deaths reported annually in 2021, coronary artery disease (CAD) affects 1 in 20 adults globally at any given time and remains one of the leading causes of death globally and in North America [21]. Generally, CAD refers to the atherosclerotic (obstructive) form, in which plaque build-up occurs within the coronary arteries of the heart, resulting in narrowing or blockage of the arteries [22, 23]. Disruption in the vascular endothelium of the coronary arteries leads to plaque accumulation, which can eventually lead to thrombosis or other conditions [24]. Clinical manifestations of CAD include myocardial infarctions (MI), angina, acute coronary syndrome and sudden cardiac death [25]. CAD accounts for 42% of all cardiovascular cases, and was declared to be the global leading cause of death by the World Health Organization (WHO) in 2020 [26, 25]. CAD is considered a complex chronic inflammatory disease, manifesting from the combination of genetic, environmental and lifestyle factors, the etiology of CAD and the interaction among these risk factors remain elusive. The risk factors of CAD can be considered as modifiable (smoking, exercise, obesity, metabolism) or non-modifiable (genetics [family history], age, gender, ethnicity) [24, 27, 28]. Additionally, many conditions act as comorbidities to CAD, due to the similarities in risk factors such as other cardiovascular diseases (e.g.

stroke), hypertension, type II diabetes, dyslipidemias, chronic kidney disease (CKD), level of physical activity, alcohol consumption obesity and other metabolic disorders. Previously, preventative efforts towards CAD tended to focus on favourably adjusting modifiable risk factors, such as monitoring blood pressure levels, encouraging healthy habits and exercise or smoking cessation. Despite a significant decrease in mortality rates across North America in the recent decades [29, 30], certain populations, particularly those of younger age or lower socioeconomic status, have reported stagnant mortality rates [31, 32]. Despite the stagnancy, approximately 1 in 5 deaths from CAD (20%) occur in adults less than 65 years old [33]. This underscores the necessity for alternative intervention methods, which may detect CAD susceptibilities earlier. Recent research has highlighted the importance of incorporating genetics into interventions and preventative measures for CAD.

In order for genetic epidemiology and genetic risk predictors to be pertinent to a disease of interest, it is essential that the disease itself has a genetic component, and the magnitude of this influence should be understood. Typically, this is addressed by considering the heritability of a phenotype. As previously alluded, there is much evidence that CAD is a disease with a strong hereditary component. Notable longitudinal studies to uncover CAD heritability include the Swedish Twins Study[6] and the Framingham Heart Study[34]. The Swedish Twins Study reports CAD heritability to be 57% (95% CI: 45% - 69%) amongst male twins and 38% (95% CI: 26% - 50%) amongst female twins. Furthermore, the Framingham Heart Study reports higher incidence of CAD amongst participants with a first-degree relative who also had CAD [35, 36, 37].

Initially, the genetic architecture of CAD was first unearthed through family aggregation studies. A disease predominantly associated with CAD is familial hypercholesterolemia (FH), a Mendelian autosomal dominant disorder characterized by elevated plasma levels

of low-density lipoprotein cholesterol (LDLc) concentrations ( $\geq 190$  mg/dl in adults) [38]. FH is most frequently linked to the *LDLR* gene, which encodes the low-density lipoprotein receptor [38]. *LDLR* was first demonstrated to be responsible for CAD risk in 1985, when a deletion in the gene was discovered in a patient with FH and his mother [35, 39]. Additional genes determined to be causative of CAD were similarly initially identified through family-based studies, namely *APOB* (apolipoprotein B) and *PCSK9* (proprotein convertase subtilisin/kexin type 9) [40, 41, 35]. The mechanism through which these genes cause CAD is thought to stem from the impairment of low-density lipoprotein receptor function, resulting in decreased hepatic clearance of LDLc particles and increased cholesterol deposition in the inner layer of arteries [42, 39]. However, there are limitations to these family studies, as certain findings of genotypes relating to the extreme CAD phenotype are so rare (e.g. *MEF2A*, *DYRK1B*) that they have proven difficult to apply to general populations [43, 44].

Fortunately, the gradual emergence of genome-wide association studies (GWAS) significantly mitigated the potential statistical power limitations inherent in family aggregation studies. Significant insights into the genetic architecture of CAD were elucidated from the first GWAS for CAD conducted in 2007 [17, 18, 19]. The GWAS successfully identified 97 genetic loci associated with CAD at genome-wide significance, including the 9p21 locus, which remains the locus most strongly linked to CAD and MI. Although the precise mechanisms linking 9p21 to CAD are still not fully understood, the locus contains enhancers for regulation of cell growth, which could potentially heighten cell proliferation within arterial walls, resulting in plaque enlargement and atherosclerosis [45, 14, 46]. Over the years, continuous efforts have been made to uncover more regarding the genetic architecture of CAD. To date, GWAS have successfully identified 279 genetic loci that exhibit significant associations with CAD [47].

While GWAS have uncovered much of the heritability for CAD throughout the years, the problem of “missing heritability” remains, which states that the numerous associations discovered by GWAS remain insufficient to completely explain 100% of the heritability for complex traits [1, 48]. Due to the drastically lowered costs of genome sequencing over the years, it has become much more feasible to investigate rare variants which may also contribute to heritability. Rare variants may have stronger effects on risk, but are much rarer in occurrence (minor allele frequency [MAF] < 0.5%) and are prone to being overlooked in regular GWAS. Typically, GWAS have been noted to have poorer coverage of variants in the 0.5-5% frequency range [49]. Usually, rare variants can be identified through exome sequencing [50]. As an alternative to whole-genome sequencing (WGS), whole-exome sequencing (WES) focusses solely on the protein coding subset of the genome, accounting for 2% of the genome including splice sites and micro RNA [1]. The sole focus on the exome increases cost-effectiveness and overall efficiency of genetic sequencing, while also targeting variants more likely to exert larger effects and functional consequences. Consequentially, exome sequencing may miss non-coding variants which contribute to risk. Rare variants occur with such infrequency that it is insufficient to establish associations based solely on individual variants. Therefore, burden tests (aggregation) is necessary to conclude rare variants’ risk association with disease. Thus, rare variant association studies (RVAS) are designed to capture these variants that are typically not found on genotyping chips designed to capture common variants. Exome sequencing has confirmed loss-of-function mutations in *LDLR*, *LPL* and *APOA5* leads to increased risk of CAD by increasing LDLc or triglyceride levels [51, 35, 52, 50]. Additionally, inactivating rare mutations in *PCSK9*, *ASGR1*, *APOC3* and *ANGPTL4* are seen to decrease risk of CAD, by method of reducing LDLc and triglyceride levels [53, 54, 55, 56, 57, 58]. A more thorough list of genes associated with CAD risk can be found in Chapter 3. Rare variant genetic risk scores (RVGRS) can

also be created, as opposed to a typical PRS which uses common variants (common variant genetic risk score [CVGRS]).

In general, the evidence supporting a genetic basis for CAD is significant, underscoring the necessity for the use of genetic methodologies to predict risk for the disease. Conventionally, associations of CAD are to monogenic causes such as FH. However, since the proliferation of GWAS, it has been observed that genetic polygenicity also contributes to the development of CAD [59]. This further highlights the importance of genetics, along with the employment of polygenic risk scores (PRS) to enhance prevention and prediction of CAD. Typically, the performance of PRS in context of CAD is relative to clinical risk factors, which are usually more conventional for the disease. The American College of Cardiology and American Heart Association suggests 10-year cardiovascular-risk calculator (ACC/AHA ASCVD cardiovascular-risk calculator), which are pooled cohort equations (PCE) utilized to predict risk in adults aged 40-79 [4, 60]. This risk calculator was first published in 2013 and intended to be an extension of the Framingham Risk Score (FRS) derived from the Framingham Heart Study. The current edition of the ACC/AHA ASCVD risk calculator considers clinical risk factors age, sex, race, blood pressure, cholesterol levels, and past history of diabetes, smoking or hypertension. This tool is intended for patients with low-density lipoprotein cholesterol (LDLc) levels under 190 mg/dL and not on LDLc lowering treatments. There exists alternative forms of clinical risk scores including the Framingham Risk Score (FRS) [61, 62], the European System Coronary Risk Evaluation (SCORE) [63, 64], the Reynolds risk score [65], the INTERHEART risk score [66], the Assign risk score [67], the QRISK3 score [68], the PROCAM risk score [69] and the CUORE risk score [70]. These address similar risk factors as the conventional ACC/AHA risk calculator. While these are deemed the standard in clinical settings, their efficacy can be considerably constrained due to the exclusive focus on non-genetic risk factors, overlooking

the substantial heritability component in cardiovascular diseases. In fact, newer iterations of these clinical scores have attempted to incorporate a component of family history into the prediction of CAD and related cardiovascular conditions, in order to address the complex nature of their etiologies [71, 72, 73]. Furthermore, these clinical scores are afflicted by a fatal flaw in that persons of younger age are automatically assigned a lower risk of CAD, which cannot apply to cohorts who suffer from premature onset. While prevalence of CAD greatly increases in older individuals, mortality rates for patients younger than 65 years for women and 55 years in men have not decreased over the years [30]. This is of particular concern due to the potential societal burden of this younger cohort. Moreover, consulting genetics would be of particular benefit to these younger patients, as CAD has a strong genetic predisposition, as is common among many diseases of early onset [74, 75].

#### **1.1.1.2 Myocardial Infarction after Non-cardiac Surgery**

Myocardial injury after noncardiac surgery (MINS) is a cardiovascular complication occurring after surgery, and is defined to be myocardial injury occurring during or within 30 days after surgery that is not associated with underlying cardiac ischemia [76]. Although it is presently the most common cardiovascular complication after surgery, affecting as many as 1 in 6 patients undergoing surgery, its underlying causes are not fully comprehended [76, 77]. Annually, over 300 million non-cardiac surgeries are conducted globally, with approximately up to 1 in 30 adults undergoing such procedures every year [78, 77]. The VISION study (Vascular Events in Noncardiac Surgery Patients Cohort Evaluation) revealed the estimated incidence of MINS to be 18% [76, 77]. The overall global prevalence of MINS is estimated at 8%, as established by a conventional fourth-generation TnT assay [79]. As non-cardiac surgery is corrective procedure that can significantly enhance a patient's quality of life, there is strong motivation to deepen our understanding of MINS,

its most common potential outcome. One such approach is to investigate the potential genetic influences of MINS.

Myocardial injury is defined as the presence of at least one cardiac troponin (cTn) value above the 99<sup>th</sup> percentile upper reference limit (URL) which may or may not be associated with ischemic symptoms in the absence of [80]. Generally, two troponin biomarkers are used to diagnose cardiac injury: cardiac troponin I (cTnI) and cardiac troponin T (cTnT) [81]. Note that myocardial injury is not equivalent to myocardial infarction (MI). MI is a form of myocardial injury, but requires evidence of acute myocardial ischemia [82, 80]. The origin of MINS is ischemic, caused by a supply-demand mismatch or atherothrombosis. The elevation of cTn levels can be attributed to a variety of factors, often stemming from surgical trauma [83]. Spikes in catecholamines, cortisol, inflammatory cytokines, platelet activation and fluid shifts can result in a mismatch of oxygen supply demand. Moreover, occurrences of hypertension, hypotension, tachycardia, bradycardia, hypoxemia and anemia may also arise, inducing heightened stress in the coronary arteries. This may potentially culminate in plaque disruptions which could even precipitate CAD. Myocardial injury can also be non-ischemic. Other co-morbidities for MINS include sepsis, renal failure, volume overload, valvular heart disease and pulmonary embolism.

Past evidence has shown that MINS is associated with CAD, due to the stress it induces on the cardiovascular system. The consequences of MINS may lead to MI, either type 1 or type 2. By definition, Type 1 MI is the result of atherosclerotic CAD, usually resulting from plaque erosion. Alternatively, Type 2 MI is caused by a mismatch in oxygen supply and demand [80]. Both scenarios are potential direct causes of MINS. In the case of Type I MI, atherosclerotic CAD was found to be present in 72% to 94% of patients with perioperative MI through angiographic studies [84, 83, 85, 86]. Evidence of atherosclerotic CAD was

also found in an autopsy series where 46% of patients who had fatal MI after noncardiac surgery [87]. For type 2 MI, ischemic conditions can arise due to stress occurring from anesthetic and surgical procedures, such as bleeding, hypotension and pain triggering stress response. Hypoxemia, anemia, hypotension, and bradyarrhythmia can all lead to type 2 MI, while increased myocardial oxygen demand can be attributable to tachyarrhythmia or severe hypertension[83].

There are existing methods to predict risk of MINS, the most common of which is the Revised Cardiac Risk Index (RCRI). As a conventional clinical approach for prediction of perioperative cardiovascular complications, the RCRI score considers six factors: history of CAD, history of congestive heart failure (CHF), history of cerebrovascular disease, diabetes on insulin, creatinine levels ( $> 177$  mmol/L) and high risk surgery [78, 88]. However, there is criticism of current methods for prediction of perioperative cardiovascular complications, with claims that they oversimplify or do not fully consider the physiology of complications [88]. In advent of increased feasibility of the processing genetic information, there is motivation to include genetics into risk calculation for MINS. Notably, many of the risk factors considered, such as CAD and diabetes, have been proven to have strong genetic dispositions [89, 90, 35]. A study by Douville et al. in 2021 proposed using a CAD PRS in order to improve risk prediction for MINS [91]. The study population was acquired from the Michigan Genomics Initiative (MGI), and consisted of 429 cases of MINS with 89 624 controls. The CAD PRS was independently associated with MINS (OR = 1.12, 95% CI = 1.02 - 1.24,  $p = 0.023$ ), and patients who developed MINS after surgery had a higher standard CAD PRS than those who did not. Additionally, utilizing clinical risk scores with the CAD PRS was seen to further improve predictive value, as assessed by discriminative capacity measures (EHR c-statistic =  $0.921 \pm 0.006$  vs.  $0.720 \pm 0.011$ ,  $p < 0.001$ , RCRI c-statistic =  $0.921 \pm 0.006$  vs.  $0.786 \pm 0.013$ ,  $p < 0.001$ ). Genetic influence was also seen

to have a greater impact in higher RCRI risk classes. Overall, this strongly encourages the exploration of integrating genetics into prediction of MINS.

### **1.1.2 Polygenic Risk Scores**

The purpose of this study is to promote and optimize the application of genetic or polygenic risk scores (PRS), typically defined as a weighted sum of risk alleles across a large number of genetic variants associated with a specific polygenic trait. PRS are practical as quantitative measures of genetic susceptibility, and can be readily derived as long as a patient’s genotype is available. They hold particular relevance for complex diseases with high heritability, such as cardiovascular disease, diabetes or obesity [92, 93, 94].

#### **1.1.2.1 Genome-wide association studies**

Genome-wide association studies (GWAS), also known as common-variant association studies, seek to detect associations between common genetic variants and phenotypic traits [15]. The advent of GWAS had enabled novel discoveries about genes and pathways for complex diseases and enhanced clinical applications based on its discoveries [1, 95]. GWAS serve as the initial foundation for PRS by assigning a value to each variant based on the variant’s strength of association to a given phenotype. They were formulated as an application of the “common disease/common variant” (CDCV) hypothesis first proposed in 1990s, which states that relatively common variants (minor allele frequency [MAF] = 0.05) of smaller effect sizes (odd ratio [OR] = 1.1 to 1.5) may collectively aggregate to a significant effect for a given trait [1, 49]. This is in direct contrast to the definition of a Mendelian disease, which are caused by extremely rare variants with high effect sizes generally detected through exome sequencing [96]. Before the advent of GWAS, most genes were identified through family-based linkage analyses, which can be greatly lacking in statistical power.

Unlike candidate gene studies, a GWAS requires no prior knowledge regarding the associative strength of variants nor known etiologic mechanisms. A GWAS observes the entire genome by means of high-throughput technologies (SNP arrays) combined with statistical imputation techniques in order to assess millions of SNPs at once [1, 15]. As such, GWAS tend to require large samples from population-based databases. Associations from GWAS can usually be interpreted through a Manhattan plot (variant location sorted by chromosome plotted against  $-\log_{10}$  p-value association significance) or a Quantile-Quantile (QQ) plot ( $-\log_{10}$  p-value under null hypothesis against observed  $-\log_{10}$  p-value) [1].

Typically, genotyping chips for GWAS detect common variants ( $\text{MAF} \geq 0.5$ ), in alignment with the CDCV hypothesis. As such, the PRS based on GWAS in this study are common-variant genetic risk scores (CVGRS). Rare-variant genetic risk scores (RVGRS) also exist, however they rely on rare-variant association study (RVAS) which can detect variants with below the 0.5% MAF threshold. Rare variants are usually detected through exome sequencing, and require burden tests in order to analyze due to the lower power of rare variants. However, several studies claim that common variants captured by genotyping chips are sufficient to capture significant, clinically relevant associations in prevalent diseases such as diabetes, breast cancer and Alzheimer’s disease [97, 98, 99, 100, 101, 102].

### **1.1.2.2 Calculation and Development of PRS**

A polygenic risk score (PRS) is typically a weighted sum of a large number of risk alleles associated with a given trait within a genotype. PRS provides a quantitative overview of an individual’s genetic predisposition towards a phenotype, adjusting the count of each allele based on a weight reflecting the variant’s association with the phenotype, obtained from GWAS. PRS development was stemmed from the hypothesis that while common variants with smaller effects may not contribute to risk prediction on their own, their accumulation

may result in detectable, significant associations [103]. The definition for PRS can be specified as followed:

$$r_i = \sum_{j=1}^m \beta_j x_{ij} \quad (1.1.1)$$

where  $r_i$  represents the risk,  $i$  represents the individual number,  $j$  is the SNP number,  $\beta_j$  is the weight for each SNP derived from GWAS summary statistics, and  $x_{ij}$  corresponds to the allele count for the  $j^{\text{th}}$  SNP of the  $i^{\text{th}}$  individual (where the full genotype is equivalent to  $[x_1 x_2 \dots x_j]^T$ ). Once a PRS is standardized, the relative genetic risk can be determined for the assessed individual. For instance, individuals with a score above a specific high-risk threshold or percentile would likely benefit from early intervention and prevention, due to their perceived higher genetic predisposition to a specific condition.

With the widespread adoption of GWAS and availability of large databases in recent years, there has been a notable surge of interest in using PRS alongside clinical methods for predicting risk [102]. Since genotyping information can be obtained early in life and is specific to each individual, PRS enables early intervention and prevention, and also holds potential for precision medicine. Incorporating genetic information into a patient's treatment plan may be greatly improve their overall well-being. A PRS may be used to influence early intervention, clinical decisions, and treatment choices [104]. However, the goal of normalizing implementation of PRS in a clinical setting remains to be achieved, whether by optimizing predictiveness of PRS or to improve discrimination when used in tandem with clinical risk predictors. Ideally, a PRS should be universally applicable across diverse populations, and reliably stratify patients irrespective of ethnicity.

PRS calculation is limited in that PRS are entirely dependent on the GWAS from which

they are derived. Bias exists within all GWAS depending on the population from which they were derived. Wang et al. gives an example regarding the UK Biobank, in which all participants are volunteers who are more likely to be healthier, wealthier and higher educated than the average population [105, 106]. While it has been consistently demonstrated that common variants are sufficient to capture significant levels of heritability, the problem of missing heritability remains. Thus, even in their most optimized form, CV-GRS are technically unable to capture all the phenotypic variance attributed to genetics. Ultimately, there exists an upper limit of  $\approx 30\%$  for SNP-based heritability [104]. Based on twin studies, the heritability for most behavioural traits is at most 50%, thus a PRS can never perfectly predict a complex trait [107, 108]. Most PRS also do not account for gene-environment interactions, where the effect of a genotype can differ depending on the environment [109, 110]. The quality of GWAS summary statistics can also vary depending on methodology and imputation. New GWAS are also consistently being published, as such PRS must be constantly updated. As well, current GWAS are mostly performed with individuals of European ancestry. In other words, PRS created from European GWAS are the most accurate for populations with European ancestry [111, 106]. More specifically, it has been observed that PRS created from European ancestry GWAS are only approximately 20-40% accurate when applied to African populations [112, 106]. Furthermore, linkage disequilibrium (LD, the tendency of non-random association of alleles. See Section 1.1.2.3) and minor allele frequencies can differ significantly across ethnicities, further impacting PRS calculations [106]. This limits the ultimate goal of creating PRS which are applicable to global populations.

### 1.1.2.3 PRS Approaches & Methodologies

Within the recent years, novel statistical methodologies have been adopted with the intent to maximize the utility of PRS. In particular, the typical aim is to maximize predictiveness by training the method on a sample subset and validating the PRS on independent test sample. The accuracy or predictiveness of the PRS is usually assessed through a regression (linear or logistic) on the test sample. Alternatively, the discriminative capacity of PRS can also be determined. This can be done through determining the area under the ROC curve (AUC ROC), net reclassification index (NRI) or integrated discrimination index (IDI) of the regression model. The  $R^2$  and  $p$ -value can also be noted, as well as the relative risk between various thresholds within the stratified sample.

#### *LD adjustment*

First coined in the 1960s by population geneticists, the term linkage disequilibrium (LD) refers to non-random association between alleles at different loci, and the tendency for certain alleles to be inherited together [1, 113]. According to Mendel’s 2nd law, or the Law of Equal Segregation: “the alleles of different genes will be inherited independently of one another” [114]. Currently, it’s understood that this notion is not universally accurate, and several factors may influence genes being inherited today, with LD being one such example.

As initially described by Lewontin, the coefficient of linkage disequilibrium, denoted as  $D$ , is the difference between the frequency of a haplotype comprising two alleles and the product of the frequencies of the two alleles (probability of both alleles independently occurring together) [113]. When  $D \neq 0$ , there exists LD between the two alleles.  $D$  can be standardized by dividing it by the maximum difference between the observed and expected haplotypes, such that  $D' = D/D_{\max}$ . With this definition,  $D$  depends on the

allele frequencies and can be positive or negative. Alternatively, one can quantify LD by calculating a correlation coefficient, or  $r^2$ .

As such, some modern reiterations of Mendel’s 2nd Law states that equal segregation applies to genes which are on different chromosomes [1]. Prior to 2015, several PRS methods did not account for LD adjustments, but it has since become indispensable in modern PRS approaches. LD can result in inaccuracy for GWAS as it is dependent on MAF, which can greatly differ between ethnicities [106]. Moreover, neglecting to consider LD could lead to an overestimation of the genetic impact of variants, because it effectively counts them twice when, in reality, the alleles are inherited together.

### *Machine Learning for Multi-PRS*

Amongst the statistical methods used to optimize PRS, machine learning approaches are more frequently utilized in order to increase prediction accuracy for PR. As a wide class of statistical analysis methods, machine learning (ML) models are automated by artificial technology, allowing for pattern recognition and computation beyond human abilities [115]. These methods are able to fit models based on given data, making them suitable for building PRS suited for a particular dataset. In the context of PRS, ML methods can automatically select for variants or risk factors that are considered significant to the phenotype of interest without human intervention. Examples of ML techniques that have been applied to PRS in the past include random forests, gradient boosted regression trees, elastic net regression and neural networks. Advancements in computing technologies have greatly facilitated the ability for these methods to handle large datasets comprising of full genotype information and hundreds of thousands of patients.

A common usage of ML for PRS involves creating “multi-PRS”. The process usually

involves inputting multiple PRS into an ML method of choice, which selects the most significant PRS and combines them into a singular PRS. Each individual PRS is derived from selected GWAS representing various traits, and is usually re-weighted by a regression coefficient from a training set. A typical pipeline is shown in Figure 1.1. As such, most ML methods used are supervised (requires training set) rather than unsupervised (no training set required, method recognizes its own patterns). This method is impartial in its selection of traits, and decides which input PRS are significant purely based on the data it is given. By combining multiple PRS into a singular predictor, a multi-PRS incorporates information from numerous GWAS representing different traits. This can also account for genetic pleiotropy, as the final regression model will account for multiple traits that could be related.

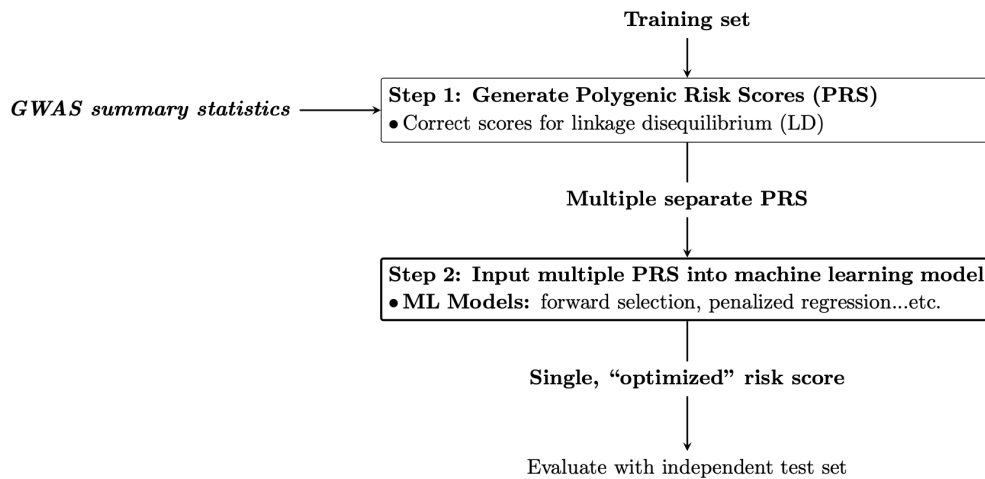


Figure 1.1: Typical workflow for utilizing machine learning for creating multi-PRS.

### *Regional Genomic Correlation*

While combining multiple PRS can improve overall prediction accuracy and information captured, it does not address the limitations of relying solely on GWAS. As previously mentioned, PRS will generally account for additive genetic effect which are below genome-wide significance. Regular GWAS assumes that all regions allow for detection of heterogeneity, when there is evidence that polygenic genetic effects may concentrate at certain regions within the genome [116, 117, 118]. Currently, GWAS and consequently CVGRS assume an “infinitesimal-model”, which presumes that all SNPs contribute to the trait in a normal distribution around the parents’ average [119]. This assumption has proven to be untrue at many different risk loci [118]. For instance, in the case of regional genomic correlation between BMI and type II diabetes (T2D), certain regions had positive correlation between BMI and T2D, other regions had negative correlation, while others were observed to be neutral [120]. This variance in directionality can especially become an issue when multiple polygenic PRS are combined to create a singular PRS, as region directionality is not accounted and all segments of the genome are assumed to contribute equally. In summary, incorporating regional genetic correlations may enhance the predictive accuracy of PRS by addressing several key factors: i) the common disease/common variant (CDCV) hypothesis which applies to GWAS, suggesting that additive genetic effects often fall below genome-wide significant but can exert a significant collective impact, ii) the tendency for polygenic inheritance to concentrate in specific genomic regions, and iii) the previously observed genome-wide correlations between pairs of complex traits. Moreover, investigating regional genetic correlations may offer insights into novel biological pathways by exploring shared heritability.

As such, there is a motivation to create PRS methods which consider polygenic heritabil-

ity within smaller regions. In 2018, Paré et al. developed a method to determine regional genetic correlation between polygenic traits [117]. As its name suggests, the “weighted maximum likelihood-regional polygenic correlation” (WML-RPC) method adopts a weighted maximum likelihood model to estimate the genetic correlation within predefined regional blocks ( $\sim 1$  Mb) between two complex traits. The method considers LD and only requires GWAS summary statistics. A likelihood ratio test is performed with maximum likelihood estimates for a genetic covariance matrix to determine the regional genetic correlation between the two traits. The method was able to identify seven correlated regions between HDLc and CAD, with one region having positive correlation and six regions having negative correlation. The six negatively correlated regions were all directed related to triglyceride metabolism (*LPL*, *TRIB1*, *MC4R*), while the positively correlated region contained the gene encoding hepatic lipase (*LIPC*), a lipolytic enzyme which regulates triglyceride levels [121]. The results suggested that high HDLc levels caused by *LIPC* increases the risk of CAD, which aligns with previous evidence that *LIPC* can lead to both increased HDLc and triglyceride levels [122]. Additionally, the additional loci of *TRIB1* and *MC4R* were identified. This singular example demonstrates the potential for regional genetic correlation to identify established biological pathways and gain novel insights. Thus, applying the concept of genomic regions could improve predictiveness of PRS.

### *Clumping & Thresholding*

One of the most common methods for constructing PRS is the clumping and thresholding (C+T), or pruning and thresholding (P+T) method [123, 124]. The PRS calculation is as seen in Equation 1.1.1, with additional filtering steps which seek to correct for LD and reduce noise. The “clumping” step involves correlating variants within a specified genetic distance, and removes all SNPs above a pre-defined  $r^2$  correlation coefficient. Subsequently,

the “thresholding” step refers to removing all variants above a certain  $p$ -value threshold (i.e. significance level) according to the GWAS summary statistics. While this method is appealing due to its simplicity, complete removal of variants from the calculation results in a loss of statistical power, and may inadvertently overlook variants which can make meaningful contributions. Regardless, it remains the most commonly used method of PRS calculation [123]. Despite its limitations, this method is still frequently regarded as the “standard” for PRS calculation and can often generate PRS that significantly associated, particularly when used as baseline for comparison to other methods.

### *LDadj*

Proposed the “GraBLD” paper by Paré et al. in 2017, LDadj offers an alternative to the C+T method by incorporating LD without removing any variants [125]. More specifically, instead of completely eliminating variants above a certain threshold, LDadj calculates the pairwise correlation between variants within a pre-specified window (distance upstream and downstream of target variant) and adjusts the original weighting obtained from GWAS summary statistics based on pairwise correlation. More specifically, the LD adjustment is as defined in Equation 1.1.2:

$$\eta_j = \sum_{k=j-l}^{j+l} r_{j,k}^2 \tag{1.1.2}$$

where the LD adjustment ( $\eta_j$ ) is the total summation of pairwise linkage disequilibrium ( $r_{j,k}^2$ ) between the  $j^{\text{th}}$  and  $k^{\text{th}}$  SNP within a pre-defined distance ( $l$ ) for a window upstream and downstream of the target SNP. Thus, when adjusting for LD, the modified  $\beta$  takes the form  $\tilde{\beta} = \beta_j/\eta_j$  and can be substituted into Equation 1.1.1 to obtain the LDadj PRS:

$$r_i = \sum_{j=1}^m \frac{\beta_j}{\eta_j} x_{i,j} \tag{1.1.3}$$

*PRSice-2*

The original PRSice method, introduced in 2015 by Eusden et al., was distinctive in its capability to compute PRS at any number of p-value thresholds in a singular execution [126]. By autonomously identifying the p-value threshold which returns the optimal results, PRSice relieves the need for external tuning. The method requires GWAS summary statistics and phenotypic information in order to determine the optimal p-value threshold. It also allows for pruning of SNPs based on LD. PRSice-2 is an improvement of the previous method, with expedited code and optimized efficiency, and relies on the same approach [127]. PRSice-2 also features automatic strand flipping, LD clumping and adjustments for overfitting.

PRSice-2 relies on an empirical p-value to optimize predictive accuracy and avoid overfitting [127, 126]. Ideally, an independent validation sample is used to evaluate performance, but the calculation of the empirical *p*-value can determine association while controlling for Type I error. The empirical *p*-value is calculated through *k* permutations of the sample, and the “best-fit” PRS is obtained for each permutation across the range of *p*-value thresholds tested. The empirical *p*-value is calculated as:

$$\text{empirical } p = \frac{\sum_{n=1}^N I(P_n < P_0) + 1}{N + 1}$$

where *N* denotes the number of permutations and *I* is an indicator function which maps

values to 1 if the  $p$ -value of the best-fit PRS of the  $n^{\text{th}}$  ( $P_n$ ) permutation is less (more significant) than the observed  $p$ -value ( $P_0$ ) and 0 otherwise. In the situation that an independent validation is not available, the authors recommend using cross-validation to prevent overfitting.

It was demonstrated that PRSice-2’s predictive performance is comparable to other leading PRS methodologies LDpred and LASSOSUM [127]. In 2021, PRSice-2 was utilized to create a T2D PRS which showed potential for identifying individuals who were at high risk for T2D (AUC = 0.795) [128]. The method was also used to create PRS for stroke and various stroke subtypes from MEGASTROKE to identify etiological pathways and analyze heritability. The PRS minorly, but significantly improved the prediction of ischemic stroke (AUC = 0.596 vs. AUC = 0.554 in a base model) [129].

### *LASSOSUM2*

This method was first proposed in 2017 as ‘lassosum’, which takes advantage of a novel penalized regression methodology in order to optimize single trait PRS [130]. lassoSum is robust to LD corrections by utilizing an external LD reference panel. The revamped methodology “lassosum2” claims to better handle differences in GWAS per-variant sample sizes [131]. This misspecification could otherwise lead to potential violation of model assumptions such as estimates for the marginal GWAS effects, resulting in inaccurate prediction. The updated approach also incorporates supplementary quality checks (WC), such as accommodations for diverse GWAS imputations and new adjustable parameters for further customization according to user requirements.

lassosum was initially created in response to the issue of the optimal  $p$ -value for thresholding generally being unknown [130]. While this can be determined by training in an

independent sample with the phenotype of interest, there can be circumstances where this data can be unavailable. The basis of lassosum2 lies in an elastic net penalized regression framework, representing a deterministic model with convex optimization. Penalized regression refers to a class of regression analyses which assign a penalty to the least squares criterion which results in shrinkage of the regression coefficients towards 0 [132]. Ridge regression, elastic net regression and LASSO regression are all penalized regression techniques, each with slightly differing sum of square penalty equations based on input parameters. For the purposes of defining the method, a general elastic net sum of squares regression term can be defined as:

$$SS = \sum_{i=1}^n (y_i - \sum_j x_{ij} \beta_j)^2 + s\lambda \sum_{j=1}^p |\beta_j| + (1-s)\lambda \sum_{j=1}^p |\beta_j|^2 \quad (1.1.4)$$

where  $y_i$  represents the  $i^{\text{th}}$  observation of the independent response variable,  $x_{ij}$  is the  $i^{\text{th}}$  observation of the  $j^{\text{th}}$  dependent explanatory variable,  $\beta_j$  is the (weighted) regression coefficient for the  $j^{\text{th}}$  response variable,  $\lambda$  is the shrinkage parameter (controls strength of penalty) and  $s$  determines whether the model weighs more towards L1 (Least Absolute Shrinkage and Selection Operator [LASSO]) regularization or L2 (Ridge) regularization. The  $\lambda$  directly affects the regression coefficient  $\beta_j$  and is one of the important modifiable parameters for elastic net regression. It controls the strengths of the penalty: if  $\lambda = 0$ , this is a regular least squares model and all dependent variables are retained. As  $\lambda$  increases, there will usually be a large enough value such that all dependent variables will be shrunk to 0. The second tuning parameter is  $s$ , which determines which of the two penalty term takes precedence. If  $s = 0$ , this becomes a Ridge regression problem, where variables can shrink towards 0, but never become exactly 0. If  $s = 1$ , this becomes a LASSO regression problem, where some variables can be entirely be removed from the equation as their value

can be reduced completely to 0. In the case where  $0 < s < 1 = 0.5$ , the penalty is balanced between each penalty term and becomes an elastic net problem. Note that if  $s = 0.5$ , the penalty is evenly balanced between the two terms and the two penalty terms are treated equally.

At the time of their introductions, both `lassosum` and `lassosum2` have demonstrated superior efficiency and enhanced predictive accuracy compared to alternative methods [130, 131]. This has been further verified through external publications, and `lassosum` was seen to be highly accurate with complex diseases such as CAD and T2D [133, 134, 135].

### *LDpred2*

The `LDpred2` method operates alongside `lassosum2`, utilizing identical inputs and capable of simultaneous execution within a single instance [131]. `LDpred2` adopts an alternative approach of utilizing Bayesian statistics, which has been recently gaining interest for PRS development [136, 137, 138].

Bayesian statistics are based off Bayes' theorem, which calculates the probability of an event happening given that another event is true, based on prior knowledge of conditions related to the event [139]. Mathematically, Bayes' theorem is defined as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1.5}$$

where  $P(A|B)$  is the conditional probability of event  $A$  occurring given event  $B$  is true, or the posterior probability;  $P(B|A)$  is the conditional probability of event  $B$  occurring given event  $A$  is true, or the likelihood;  $P(A)$  is the probability of observing  $A$ , or the

prior probability; and  $P(B)$  is the probability of observing  $B$ , or the marginal probability. The purpose of the Bayes theorem is to determine the posterior distribution from the prior distribution, which is a quantification of the beliefs regarding the parameter before the data is seen. The prior is usually based on previous independent work and can be inferred through previously well-established methods which require more advanced computation.

In the context of LDpred2, the posterior probability is defined as  $P(Y|\bar{\beta}, \hat{D})$ , where  $Y$  is a phenotype vector,  $\bar{\beta}$  is a vector of marginally estimated least-squares estimates from GWAS summary statistics. When using squared error loss, the mean of the posterior distribution can be used as Bayes estimator, or an estimator which minimizes the posterior expected loss (e.g. error). Thus, the posterior mean phenotype given GWAS summary statistics and LD under a linear model assumption is:

$$E(Y|\bar{\beta}, \hat{D}) = \sum_{i=1}^M X_i' E(\beta_i|\bar{\beta}, \hat{D}). \quad (1.1.6)$$

Note that  $X_i$  is notation for the  $i^{\text{th}}$  genetic and  $\beta_i$  is the true genetic effect. With the incorporation of LD adjustments, the prior becomes non-infinitesimal, thus is approximated through a Markov chain Monte Carlo (MCMC) Gibbs sampler [140]. The MCMC methods are algorithms used to create samples from a continuous random variable to evaluate the expected value or variance of the given variable. The Gibbs sampler is one such example of an MCMC algorithm used when the conditional distribution is easier to sample from relative to the joint distribution. Ultimately, LDpred2 is able to calculate a weighting for each variant based on the probability of the phenotype given the GWAS effects and LD corrections. The posterior mean effect sizes obtained from the MCMC Gibbs sampling can subsequently be applied to validation data.

Since its release, LDpred2 has been considered to be a widely referenced method for PRS creation [141]. The updated method involves improvements in QC, accounting for variations in imputations, input consistency and optional adjustable parameters to fit the user data [131]. A CAD PRS created using the LDpred2 method was able to stratify CAD risk when used in combination with lipoprotein[a] levels, where individuals with high CAD PRS and high Lp[a] levels were seen to have a 5-fold increased risk in CAD relative to those with low CAD-PRS and low Lp[a] levels [142]. LDpred2 has also seen successful applications with other complex diseases, such as T2D and psychiatric disorders [143, 144].

#### **1.1.2.4 Clinical Utility & Applications of PRS**

The impetus for the development of PRS arises from its potential utility in precision or personalized medicine. Precision medicine, an innovative approach that has garnered recent attention, integrates genetic, environmental, and lifestyle factors of individual patients to determine optimal treatment approaches for diseases [145, 102]. PRS are particularly attractive for precision medicine, as genotypes can be obtained shortly after birth and are unique to each patient. Hence, they can prove highly valuable for the early detection, intervention and prevention of diseases. Once a PRS is acquired, it serves as a singular quantitative measure encapsulating the genetic risk for an individual patient. In this form, PRS are adaptable, and can seamlessly integrate into existing approaches such as clinical risk predictors which often lack a genetic component. The cost of genotyping has been reduced dramatically over the years, with commercial prices for WGS now less than \$1000 [146]. A single genetic test for an individual is approximated to be around \$47 (\$35 USD), comprising of a genome-wide array with automated bioinformatics [147]. A single test alone would be sufficient to create multiple PRS allowing for diagnosis for multiple diseases. A 2022 cost-utility analysis on PRS for cardiovascular diseases indicated that

using PRS alongside existing clinical methods could be cost-effective, with an incremental cost-effectiveness ratio of \$172 906 (\$143 685 USD) per quality-adjusted life-year (QALY) at a base-case analysis genotyping cost of \$70 [148].

### *Disease Detection with PRS*

PRS holds utility in enhancing disease detection, specifically early detection. As inherited genetics are established at conception, PRS can be applied early in life, independent of other clinical factors such as age, environmental conditions and lifestyle choices. Moreover, clinical risk factors may take time to emerge, enhancing the utility of PRS in identifying diseases of early onset. A PRS is also inclined to estimate lifetime risk trajectories, rather than focusing on shorter-term risk typical of clinical risk scores, which usually span 5 to 10 years [147].

A common suggestion for PRS utility involves integrating PRS alongside clinical risk scores in order to enhance risk prediction. Clinical risk scores often heavily consider age as a risk factor, rendering them inadequate for assessing early-onset cases. For instance, the AHA/ACC ASCVD risk calculator recommended by the American College of Cardiology and American Heart Association estimates cardiovascular risk over a 10-year period and predominantly considers age as a risk factor[4]. Additionally, family history or genetic disposition is not considered in these scores, despite evidence that has demonstrated incorporation of family history can significantly improve risk assessment CAD [149, 73]. Consequently, such calculators are ill-suited for predicting early-onset CAD, where younger individuals are deemed at lower risk solely based on age despite the potentially significant contribution of genetic factors, which are particularly influential in early-onset cases. Previous studies have shown that incorporating a CAD PRS with clinical risk scores lead to

improved risk prediction for cardiovascular events, independent of clinical risk factors, thus affirming the efficacy of PRS in this regard [134, 150, 151, 152].

There is also potential for utilizing PRS independently of clinical risk factors. The heritable component alone for certain phenotypes can be as effective at detecting onset, especially in certain circumstances. As previously stated, this is especially relevant in cases of early onset, where genetic disposition tends to have a higher impact. A 2021 study demonstrated that for certain common diseases (e.g. hypertension, atherosclerotic CAD, hypothyroidism, unspecified malignant neoplasm of the skin, calculus of gallbladder without cholecystitis), the genetic relative risk decreased with age when compared to the risk associated with environmental or non-genetic factors[153]. PRS have also demonstrated the ability to significantly improve risk prediction in carriers of high-impact causal disease variants, such as in the case of FH or breast cancer [147, 154, 151]. Thus, a highly predictive PRS could account for genetic risk that a monogenic mutation may not account for due to the effect of polygenicity. The risk conferred by a PRS alone has also been comparable to risk conferred by a monogenic mutation. In particular, the top 8% of persons in a CAD PRS distribution were seen to have a risk comparable to those with a monogenic FH mutation [103]. Thus, there may also be benefits in using PRS alone as genetic risk predictor.

PRS may also be used to aid in the interpretation of disease screening. This is of particular relevance to cancers, where various risk factors are considered as to when screening should be initiated[102]. PRS have seen particular utility in detecting false-positive rates in screening tests. The age for screening for breast cancer and colorectal cancer is recommended based on the average age risk of cancer and the risk occurring due to false-positive mammography results. A breast cancer PRS was able to determine 16% of the popula-

tion who should start at 40 years of age as their risk was greater than that of an average 50-year-old [155]. Similarly, a PRS for colorectal cancer demonstrated that individuals in the highest PRS decile should initiate colonoscopy screening at 42 years old rather than 52 years old for individuals in the lowest decile [156]. Finally, PRS have also aided in the interpretation of screening tests with high false-positive rates, particularly in the case of prostate cancer. Despite the high prevalence (second highest cause of cancer death in men), the screening for prostate cancer is often not recommended as the risk of detecting a false positive result has been seen to outweigh the benefits [157]. Thus, PRS were able to differentiate between men who may have higher benefit from screening due to their elevated risk. The positive predictive value for detection of aggressive prostate cancer by screening was approximately 25% in individuals in the top 5% of the PRS relatively to about 12.5% in the general population [158]. This can also be extended to diagnostic refinement in instances where certain conditions can be difficult to differentiate [147]. For example, type I and type II diabetes can be difficult to differentiate clinically, due to the similarity in symptoms and lab results. A PRS was able to reasonably distinguish between T1D and T2D, despite not being clinically applicable [159].

#### *Disease Intervention with PRS*

PRS also have potential in therapeutic intervention, particularly to determine whether the risk of disease onset surpasses the unwanted consequences of particular interventions [102]. This aligns with the principles of precision medicine, as it allows for personalized care plans and treatment methods based on the individual. A prevalent example is with statin therapy, which is known for its use in lowering blood cholesterol levels for the primary (first event) and secondary prevention of cardiovascular events [160]. Numerous studies over the years have demonstrated the benefits of statin therapy and its efficiency in reducing the

risk of coronary events, however it has also been known to create adverse effects. There is a chance that statins may induce diabetes over a 5-year period, with 1 in 100 individuals developing the disease. Contrarily, less than 2 out of 100 individuals taking statins avoid a heart attack or a stroke within a 5-year period [160, 161, 162]. Current recommendations for statin therapy initiation rely on clinical risk factors, which have been shown to overestimate the absolute risk of coronary events [163, 164, 102]. Thus, studies have shown that CAD PRS have dispersed some uncertainty regarding which individuals have greater benefit from statin therapy, independent of family history [102]. More specifically, it has been demonstrated that patients within the highest quintile of a CAD PRS are at an approximately 30% increased risk of an adverse coronary events, and that these individuals are able to reduce the 10-year risk of heart attack or CAD-related event by approximately 45% upon initiation of statin therapy [165, 166, 167]. Similarly, this intervention may apply to other therapies, such as the prescription of insulin for diabetes.

#### *Disease Prevention with PRS*

Lastly, the genetic insights provided by PRS may facilitate lifestyle adjustments that could potentially obviate the necessity for disease detection or intervention altogether. Alternatively, the awareness of an increased genetic predisposition might motivate individuals to adopt healthier behaviours or be more vigilant about screening and taking medications for relevant diseases. It has been demonstrated that individuals were more likely to continue with cancer screenings after genetic test results were disclosed [168]. Additionally, disclosure of CAD PRS to patients have shown improved risk-reduction behaviours [169]. Patients were more likely to seek information and adopt favourable lifestyle choices regarding health. Of course, this is not exclusive to the disclosure of genetic information, and other factors could motivate these preventative measures [170].

### 1.1.2.5 Current Limitations of PRS

Despite much evidence to continue the pursuit of optimizing and utilizing PRS as genetic risk predictors, there exist limitations that currently prevent their effectiveness in a clinical setting. The goal remains to create a PRS that is accurate to the global population and is generalizable to all patients. This can be difficult to achieve as polygenic variants are not as directly correlated to outcomes as monogenic, high-risk variants are [102]. PRS are susceptible to noise, and, as previously stated, are entirely reliant on the GWAS from which they originate. This is discussed in detail in Sections 1.1.2.1 & 1.1.2. Briefly, the variations imputation quality of a GWAS can affect the predictive accuracy of a PRS. Furthermore, GWAS tend to be formed from European populations, which can lead to inaccuracy when projecting to global populations.

The general perception of PRS is also important in its clinical applications. The sole dependence on genetics alone has stemmed criticism, especially from proponents of using traditional, clinical risk factors for risk prediction. Firstly, ethical and privacy concerns surrounding the acquisition and storage of genotyping data may cause patients to hesitate in providing their consent [35]. The lack of physician and public knowledge could further reinforce these concerns. Additionally, there is no standard guideline for the implementation of PRS under a clinical setting [160, 102]. This can be vital especially when a physician may have to delay interventions to consult a PRS. Of course, an inaccurate PRS could also be detrimental to the patient, with false positives causing unneeded and possibly invasive interventions and false negatives instilling false senses of security. In conclusion, there is still considerable progress needed in the development of PRS before it can be deemed clinically safe and established as a standard practice. Further general limitations of PRS will be discussed in the Conclusion chapter of this thesis.

### 1.1.3 Study Populations

#### 1.1.3.1 UK Biobank

The UK Biobank (UKB) is a large epidemiological study consistent of over 500 000 individuals aged 40 to 69 from across the United Kingdom [171]. The resource contains extensive data regarding participants' characteristics and measurements including demographics, health diagnoses, physical measurements, environmental and lifestyle factors. Baseline assessments were conducted in 2006 with the intention of follow-up, and was completed in 2010. Methods of baseline data collection included a self-completed questionnaire, computer-assisted interview, physical measures, function measures and samples of blood, urine and saliva. As such, the UK Biobank contains a plethora of genetic and phenotypic information, including WGS available for all 500m000 patients and WES for 470,000 patients. Overall, the UK Biobank offers a wealth of resources with the aim of advancing research in public health and epidemiology.

The primary objective of the main study was to optimize and train PRS models for a range of common diseases, requiring access to comprehensive genotypic and phenotypic datasets. The UK Biobank was selected for its ability to accommodate specific phenotype definitions through its field IDs, and it also provides support for International Classification of Diseases, 9<sup>th</sup> and 10<sup>th</sup> Revisions (ICD-9, ICD-10) and Office of Population Censuses Surveys Classification of Intervention and Procedures, version 4 (OPCS-4) codes[172, 173]. For instance, the CAD phenotype was defined based on myocardial infarctions (MI) and coronary revascularization. Various types of MI were defined by ICD-10 codes I21 (acute MI), I22 (subsequent MI), I23 (certain complications following acute myocardial infarction), I24.1 (Dressler syndrome) and I25.2 (old myocardial infarction; occurring more than 4 weeks prior to cardiac rehabilitation services), while coronary revascularization was based

on OPCS-4 codes for coronary bypass grafting or coronary angioplasty with or without stenting: K40.1-40.4, K41.1-41.4, K45.1-45.5, K49.1-49.2, K49.8-49.9, K50.2, K75.1-75.4 & K75.8-75.9. The stroke phenotype also encompasses multiple fields (algorithmically defined outcomes for all stroke, ischemic stroke and hemorrhagic stroke). Thus, the UK Biobank population of British related patients serves as one of the main study populations for the construction of our novel PRS method.

### **1.1.3.2 VISION study**

The Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) study was a large, international prospective cohort study consisting of over 40,000 patients aged 45 and older who underwent inpatient noncardiac surgery [174, 76]. The main focus of VISION is to investigate adverse major vascular events occurring during non-cardiac surgeries. The study is motivated by changes in surgery over the years, with uncertainty regarding the invasiveness of certain surgical interventions despite the increase in elderly patients over the years. This international study spanned 28 academic hospital centres and 14 countries across the Americas, Asia, Europe, Africa and Australia, with patient recruitment occurring from August 2007 to November 2013. All patients had received general or regional anesthesia and remained in the hospital for at least one night after surgery. It was observed among VISION participants that MINS was one of the three most important perioperative complications associated with perioperative mortality, with the others being major bleeding and sepsis [174]. Within the observed cohort, MINS ranked as the second most prevalent complication following major bleeding, affecting 13.0% of patients (5 191 cases). Additionally, MINS showed an independent association with 30-day mortality (314 deaths, HR: 2.2 [95% CI: 1.9 - 2.6]). Thus, VISION is an ideal cohort for evaluating MINS-related PRS.

For the substudy, VISION participants were investigated for ischemic symptoms according to non-ischemic definition of MINS. Troponin elevation levels surpassing the 99<sup>th</sup> percentile were adjudicated to have an ischemic origin. Clinical data is currently stored at the Population Health Research Institute (PHRI) in Hamilton, Ontario. Blood samples were collected, processed, frozen and stored for genotyping. The Precision Medicine Research Array (PMRA) was used for genotyping. Quality control was performed using PLINK software, with further imputations performed. Participants were restricted to those with European genetic ancestry.

### **1.1.3.3 External Genome-wide Association Study Summary Statistics**

This thesis is based on PRS, and thus relies numerous consortia conducting GWAS for analyses. Examples of commonly referenced consortia and their GWAS include C4D/Cardiogram for CAD and MI, DIAGRAM for T2D, or MEGASTROKE for stroke phenotypes. Tables S4.1 and S5.2 displays the full list of GWAS referenced for PRS creation. All PRS utilized within these studies underwent filtering criteria specific to the outcome of interest. For the MINS study in Chapter 4, GWAS was selected according to related MINS risk factors (e.g. cardiovascular conditions and related co-morbidities). For the PRS methodology study in Chapter 6, GWAS were selected from an pre-downloaded internal database, which is constantly updated for prevalent traits. GWAS were required to no contain UKB data (due to outcomes being based in UKB), be the most recently updated version of the summary statistics and contain at least 5,000,000 variants.

### **1.1.4 Standards for Reporting PRS & Quality Control**

The influx of novel PRS publication and methodologies within the recent decade warrants a standardized method of developing and reporting these studies. Within the last decade,

over 900 publications have mentioned PRS, involving evaluation, development and utility of PRS [175]. Several guides have been developed to aid in this matter, such as Choi et al.’s “A guide to performing Polygenic Risk Score analyses” [123] and Wand et al.’s Polygenic Risk Score Reporting Standards (PRS-RS) [175], both published in 2020. These guidelines aim to establish a standard for PRS by providing robust and sensitive data management practices, as well as clear reporting standards.

#### 1.1.4.1 Quality Control

The PRS is inherently reliant on the base data (GWAS) and the target data (cohort study). Several important quality control (QC) conditions should be met for the GWAS itself, in order to ensure robustness in the PRS generated from them [176]. These considerations including missingness of variants and individuals, sex discrepancy, minor allele frequency (MAF), Hardy-Weinbury equilibrium (HWE), heterozygosity, relatedness & population stratification.

The GWAS base data should undergo a few QC steps. First, it is recommended to do a heritability check on the GWAS to ensure it will sufficiently power the PRS. A chip-heritability estimate (heritability captured by genotyping chip) threshold of  $h_{\text{snp}}^2 > 0.05$  for GWAS data is recommended[123]. Additionally, it is crucial to identify the effect allele within the GWAS to ensure the correct effect direction of the PRS. More specific thresholds recommended to ensure a high quality GWAS include but are not limited to: genotyping rate  $> 0.99$ , sample missingness  $< 0.02$ , heterozygosity  $P > 10 \times 10^{-6}$ , minor allele frequency (MAF)  $> 1\%$ , imputation ‘info score’  $> 0.8$  [123]. In terms of the target (outcome) data, there is a sample size recommendation of at least 100 individuals, or effective sample size (minimum sample size which achieves adequate statistical power) of over 100 for case/control studies [177]. When large sample sizes are available for both the

GWAS and outcome data, certain thresholds may be slightly adjusted (e.g. SNPs with  $MAF < 1\%$  may be included).

Further checks should be ensured in both GWAS and outcome data. The correct genome build version should be ensured. The bioinformatics tool LiftOver is suggested for standardizing genomic builds across different datasets [178]. Ambiguous SNPs, or variants with complementary alleles, should also be considered. These are usually removed from the analysis, as it can be difficult to match these alleles if the GWAS and outcome data were generated using different genotyping chips and chromosome strand is unknown for either. Another option is to infer the allele match based on allele frequencies, but this has been shown to result in bias[179]. Strand mismatching may also occur (non-ambiguous mismatch), though most PRS programs will flip this automatically. Duplicate SNPs should also be removed. Additionally, sex chromosomes can be complex to handle, as males are hemizygous for the X chromosome while females are homozygous. If they are included, scaling and/or dosage compensation must be performed to mimic the effects of X-inactivation in females. Various models exist to adjust for sex chromosomes [180]. However, if it suffices to analyze chromosomal genetics, sex chromosomes can be typically removed from analysis. It is also recommended to remove overlapping samples/samples generated from identical sources from the GWAS and outcome data, as this may result in overinflation of results. Note this issue can also be addressed through a training/test split of the original sample. Relatedness should also be considered, as a high degree of relatedness may also result in inflation of results. It may be optimal to remove any data containing highly related individuals (1<sup>st</sup> and 2<sup>nd</sup> degree relatives).

After the quality control is performed on the data, considerations should be made during the actual process of PRS calculation[123]. Firstly, the influence of each SNP may not be

equivalent, thus shrinkage of effect sizes may be considered. Some novel PRS methods which utilize LASSO regression or Bayesian approaches automatically consider shrinkage of variants [136, 181, 131, 137]. A  $p$ -value threshold can also be applied, such that only variants under a certain significance threshold are retained for PRS calculations. The optimal  $p$ -value threshold for a given dataset can be obtained through tuning and simulations, as a formal method to remove weakly associated variants from the calculation. Finally, for the current literature, adjustments for linkage disequilibrium (LD) are crucial in PRS calculations to manage the non-random associations of certain alleles.

#### **1.1.4.2 PRS-RS: Polygenic Risk Score Reporting Standards**

The Polygenic Risk Score Reporting Standards (PRS-RS) serve as a minimal criterion to ensure clarity and reproducibility in reports on PRS development, with the ultimate goal of clinical implementation [175].

First, the study should be outlined with appropriate outcomes, to describe the intent and relevancy for using the population of choice. The risk being measured should be clearly stated along with purpose and clinical relevancy. Dataset characteristics should also be reported (e.g. demographics). Study population should also be described, with a focus on ancestry and definitions regarding outcomes and genotyping assays. Next, the method should be clearly outlined with descriptions regarding variation selection, statistical model of choice and any other non-genetic variables included (e.g. correction for environmental factor or clinical risk score).

It is also important to establish a standard for reporting PRS findings. First, the PRS distribution should be described. Next, the predictive accuracy of the PRS should be relayed, generally through performance measures (e.g.  $R^2$ , proportion of variance) or

risk estimates (e.g. odds ratio [OR] / hazards ratio [HR]). Discrimination and calibration analyses should also be included, to ensure the fitted model will improve prediction or is accurate to the actual observed data. Finally, demographics and clinical characteristics for any subgroup analyses should also be reported.

There is also guidance for the discussion of PRS studies. A summary should be provided regarding the interpretation of the risk model, including overall performance of the PRS. This is usually relative to other PRS or clinical risk models. Limitations should be outlined, along with generalizability to the target or other populations. Additionally, the intended uses and future directions of the method should be stated. Finally, there should be sufficient data transparency and availability to be able to replicate the results.

## References

- [1] M. A. Austin. *Genetic Epidemiology: Methods and Applications*. Cabi, Wallingford, Oxfordshire, UK, illustrated edition edition, 2013. ISBN 978-1-78064-181-2.
- [2] I. Condó. Rare Monogenic Diseases: Molecular Pathophysiology and Novel Therapies. *Int J Mol Sci*, 23(12), 2022. doi: 10.3390/ijms23126525.
- [3] P. Duggal, C. Ladd-Acosta, D. Ray, and T. H. Beaty. The Evolving Field of Genetic Epidemiology: From Familial Aggregation to Genomic Sequencing. *Am J Epidemiol*, 188(12):2069–2077, 2019. doi: 10.1093/aje/kwz193.
- [4] D. K. Arnett, R. S. Blumenthal, M. A. Albert, A. B. Buroker, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, 140(11):e596e646, 2019. doi: 10.1161/CIR.0000000000000678.
- [5] P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era concepts and misconceptions. *Nature Reviews Genetics*, 9(44):255266, 2008. ISSN 1471-0064. doi: 10.1038/nrg2322.
- [6] S. Zdravkovic, A. Wienke, N. L. Pedersen, M. E. Marenberg, et al. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *Journal of Internal Medicine*, 252(3):247254, 2002. ISSN 1365-2796. doi: 10.1046/j.1365-2796.2002.01029.x.
- [7] J. Kaprio, J. Tuomilehto, M. Koskenvuo, K. Romanov, et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in

- a population-based cohort of twins in Finland. *Diabetologia*, 35(11):10601067, 1992. ISSN 0012-186X. doi: 10.1007/BF02221682.
- [8] M. McCarthy and S. Menzel. The genetics of type 2 diabetes. *British Journal of Clinical Pharmacology*, 51(3):195199, 2001. ISSN 0306-5251. doi: 10.1046/j.1365-2125.2001.00346.x.
- [9] R. B. Prasad and L. Groop. Genetics of type 2 diabetes pitfalls and possibilities. *Genes*, 6(1):87123, 2015. ISSN 2073-4425. doi: 10.3390/genes6010087.
- [10] S. J. Andrews, A. E. Renton, B. Fulton-Howard, A. Podlesny-Drabiniok, et al. The complex genetic architecture of Alzheimers disease: novel insights and future directions. *eBioMedicine*, 90, April 2023. ISSN 2352-3964. doi: 10.1016/j.ebiom.2023.104511.
- [11] I. K. Karlsson, V. Escott-Price, M. Gatz, J. Hardy, et al. Measuring heritable contributions to Alzheimers disease: polygenic risk score analysis with twins. *Brain Communications*, 4(1):fcab308, 2022. ISSN 2632-1297. doi: 10.1093/braincomms/fcab308.
- [12] D. Lvovs, O.O. Favorova, and A.V. Favorov. A Polygenic Approach to the Study of Polygenic Diseases. *Acta Naturae*, 4(3):5971, 2012. ISSN 2075-8251.
- [13] L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nature Reviews Genetics*, 2(22):9199, 2001. ISSN 1471-0064. doi: 10.1038/35052543.
- [14] E. D. Muse, S. Chen, and A. Torkamani. Monogenic and Polygenic Models of Coronary Artery Disease. *Current cardiology reports*, 23(8):107, 2021. ISSN 1523-3782. doi: 10.1007/s11886-021-01540-0.
- [15] V. Tam, N. Patel, M. Turcotte, Y. Bosse, et al. Benefits and limitations of genome-

- wide association studies. *Nature Reviews Genetics*, 20(8):467485, 2019. ISSN 14710056. doi: 10.1038/s41576-019-0127-1.
- [16] E. Uffelmann, Q. Huang, Nchangwi S. Munung, de Vries, et al. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(11):121, 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00056-9.
- [17] A. Helgadottir, G. Thorleifsson, A. Manolescu, S. Gretarsdottir, et al. A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science*, 316(5830):14911493, 2007. doi: 10.1126/science.1142842.
- [18] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, et al. A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science (New York, N.Y.)*, 316(5830):14881491, 2007. ISSN 0036-8075. doi: 10.1126/science.1142447.
- [19] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, et al. Genomewide association analysis of coronary artery disease. *The New England Journal of Medicine*, 357(5):443453, 2007. ISSN 1533-4406. doi: 10.1056/NEJMoa072366.
- [20] N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol. Common vs. Rare Allele Hypotheses for Complex Diseases. *Current opinion in genetics & development*, 19(3):212219, 2009. ISSN 0959-437X. doi: 10.1016/j.gde.2009.04.010.
- [21] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, Cheryl A.M. Anderson, et al. Heart Disease and Stroke Statistics2023 Update: A Report From the American Heart Association. *Circulation*, 147(8):e93e621, 2023. doi: 10.1161/CIR.0000000000001123.
- [22] "Heart and Stroke Foundation of Canada". Coronary artery disease, 2023. URL <https://www.heartandstroke.ca/en/heart-disease/conditions/coronary-artery-disease/>.

- [23] A. V. Khera, C. A. Emdin, I. Drake, P. Natarajan, and other. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *New England Journal of Medicine*, 375(24):23492358, 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1605086.
- [24] P. Libby and P. Theroux. Pathophysiology of Coronary Artery Disease. *Circulation*, 111(25):34813488, 2005. doi: 10.1161/CIRCULATIONAHA.105.537878.
- [25] S. Sayols-Baixeras, C. Lluís-Ganella, G. Lucas, and R. Elosua. Pathogenesis of coronary artery disease: focus on genetic risk factors and identification of genetic variants. *The Application of Clinical Genetics*, 7:1532, 2014. ISSN 1178-704X. doi: 10.2147/TACG.S35301.
- [26] P. Premsagar, C. Aldous, and T. Esterhuizen. Cardiac scoring systems, coronary artery disease and major adverse cardiovascular events: A scoping review. *South African Family Practice*, 65(1):5683, 2023. ISSN 2078-6190. doi: 10.4102/safp.v65i1.5683.
- [27] Public Health Agency of Canada. Heart Disease in Canada, 2017. URL <https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html>. Last Modified: 2022-07-28.
- [28] M. Regmi and M. A. Siccardi. *Coronary Artery Disease Prevention*. StatPearls Publishing, Treasure Island (FL), 2023. URL <http://www.ncbi.nlm.nih.gov/books/NBK547760/>.
- [29] E. S. Ford, U. A. Ajani, Janet B. Croft, J. A. Critchley, et al. Explaining the Decrease in U.S. Deaths from Coronary Disease, 1980-2000. *New England Journal of Medicine*, 356(23):23882398, 2007. ISSN 0028-4793. doi: 10.1056/NEJMsa053935.

- [30] K. A. Wilmot, M. O'Flaherty, S. Capewell, E. S. Ford, and V. Vaccarino. Coronary Heart Disease Mortality Declines in the United States From 1979 Through 2011. *Circulation*, 132(11):9971002, 2015. doi: 10.1161/CIRCULATIONAHA.115.015293.
- [31] G. A. Roth, . Johnson, A. Abajobir, F. Abd-Allah, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *Journal of the American College of Cardiology*, 70(1):125, 2017. ISSN 1558-3597. doi: 10.1016/j.jacc.2017.04.052.
- [32] B. Vogel, M. Acevedo, Y. Appelman, Bairey .l, et al. The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030. *The Lancet*, 397(10292):23852438, 2021. ISSN 01406736. doi: 10.1016/S0140-6736(21)00684-X.
- [33] National Center for Health Statistics. Multiple Cause of Death Data on CDC WONDER, 2023. URL <https://wonder.cdc.gov/mcd.html>.
- [34] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang. The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective. *Lancet*, 383(9921):9991008, 2014. ISSN 0140-6736. doi: 10.1016/S0140-6736(13)61752-3.
- [35] A. V. Khera and S. Kathiresan. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature Reviews Genetics*, 18(66):331344, 2017. ISSN 1471-0064. doi: 10.1038/nrg.2016.160.
- [36] D. M. Lloyd-Jones, B. Nam, R. B. D'Agostino, D. Levy, et al. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: A prospective study of parents and offspring. *JAMA*, 291(18):22042211, 2004. ISSN 0098-7484. doi: 10.1001/jama.291.18.2204.

- [37] J. M. Murabito, M. J. Pencina, B. Nam, Ralph B. D'Agostino, et al. Sibling Cardiovascular Disease as a Risk Factor for Cardiovascular Disease in Middle-aged Adults. *JAMA*, 294(24):31173123, 2005. ISSN 0098-7484. doi: 10.1001/jama.294.24.3117.
- [38] R. N. Pejic. Familial Hypercholesterolemia. *The Ochsner Journal*, 14(4):669672, 2014. ISSN 1524-5012.
- [39] M. A. Lehrman, W. J. Schneider, T. C. Südhof, M. S. Brown, J. L. Goldstein, and D. W. Russell. Mutation in LDL Receptor: Alu-Alu Recombination Deletes Exons Encoding Transmembrane and Cytoplasmic Domains. *Science (New York, N.Y.)*, 227(4683):140146, 1985. ISSN 0036-8075.
- [40] M. Abifadel, M. Varret, J. P. Rabès, D. Allard, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature Genetics*, 34(22):154156, 2003. ISSN 1546-1718. doi: 10.1038/ng1161.
- [41] L. F. Soria, E. H. Ludwig, H. R. Clarke, G. L. Vega, et al. Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. *Proceedings of the National Academy of Sciences of the United States of America*, 86(2):587, 1989. doi: 10.1073/pnas.86.2.587.
- [42] G. Go and A. Mani. Low-Density Lipoprotein Receptor (LDLR) Family Orchestrates Cholesterol Homeostasis. *The Yale Journal of Biology and Medicine*, 85(1):1928, 2012. ISSN 0044-0086.
- [43] A. R. Keramati, M. Fathzadeh, G. Go, R. Singh, et al. A form of the metabolic syndrome associated with mutations in DYRK1B. *The New England Journal of Medicine*, 370(20):19091919, 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1301824.

- [44] W. Lieb, B. Mayer, I. R. König, I. Borwitzky, et al. Lack of Association Between the MEF2A Gene and Myocardial Infarction. *Circulation*, 117(2):185191, 2008. doi: 10.1161/CIRCULATIONAHA.107.728485.
- [45] N. A. Almontashiri. The 9p21.3 risk locus for coronary artery disease: A 10-year search for its mechanism. *Journal of Taibah University Medical Sciences*, 12(3):199204, 2017. ISSN 16583612. doi: 10.1016/j.jtumed.2017.03.001.
- [46] J. Zhuang, W. Peng, H. Li, W. Wang, et al. Methylation of p15INK4b and Expression of ANRIL on Chromosome 9p21 Are Associated with Coronary Artery Disease. *PLoS ONE*, 7(10):e47193, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0047193.
- [47] K. G. Aragam, T. Jiang, A. Goel, S. Kanoni, et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nature Genetics*, 54(1212):18031815, 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01233-6.
- [48] S. Sandoval-Motta, M. Aldana, E. Martínez-Romero, and A. Frank. The Human Microbiome and the Missing Heritability Problem. *Frontiers in Genetics*, 8:80, 2017. ISSN 1664-8021. doi: 10.3389/fgene.2017.00080.
- [49] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):15161518, 1996. ISSN 00368075.
- [50] A. Warr, C. Robert, D. Hume, A. Archibald, et al. Exome Sequencing: Current and Future Perspectives. *G3: Genes/Genomes/Genetics*, 5(8):15431550, 2015. ISSN 2160-1836. doi: 10.1534/g3.115.018564.
- [51] J. Hartiala, W. S. Schwartzman, J. Gabbay, A. Ghazalpour, et al. The Genetic Architecture of Coronary Artery Disease: Current Knowledge and Future Oppor-

- tunities. *Current atherosclerosis reports*, 19(2):6, 2017. ISSN 1523-3804. doi: 10.1007/s11883-017-0641-6.
- [52] A. V. Khera, H. Won, G. M. Peloso, C. ODushlaine, et al. Association of Rare and Common Variation in the Lipoprotein Lipase Gene With Coronary Artery Disease. *JAMA*, 317(9):937946, 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.0972.
- [53] J. C. Cohen, E. Boerwinkle, T. H. Mosley, and H. H. Hobbs. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *New England Journal of Medicine*, 354(12):12641272, 2006. ISSN 0028-4793. doi: 10.1056/NEJMoa054013.
- [54] F. E. Dewey, Vi. Gusarova, C. ODushlaine, O. Gottesman, et al. Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *New England Journal of Medicine*, 374(12):11231133, 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1510926.
- [55] Anders Berg Jørgensen, Ruth Frikke-Schmidt, Børge G. Nordestgaard, and Anne Tybjærg-Hansen. Loss-of-function mutations in APOC3 and risk of ischemic vascular disease. *The New England Journal of Medicine*, 371(1):3241, 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1308027.
- [56] "Myocardial Infarction Genetics, CARDIoGRAM Exome Consortia Investigators", et al. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *The New England Journal of Medicine*, 374(12):11341144, 2016. ISSN 1533-4406. doi: 10.1056/NEJMoa1507652.
- [57] P. Nioi, A. Sigurdsson, G. Thorleifsson, H. Helgason, et al. Variant ASGR1 Associated with a Reduced Risk of Coronary Artery Disease. *New England Journal of Medicine*, 374(22):21312141, 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1508419.

- [58] "TG, Lung HDL Working Group of the Exome Sequencing Project, National Heart, Blood Institute, Crosby", et al. Loss-of-function mutations in apoc3, triglycerides, and coronary disease. *The New England Journal of Medicine*, 371(1):2231, 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1307095.
- [59] S. Theriault, R. Lali, M. Chong, J. L. Velianou, M. K. Natarajan, and G. Paré. Polygenic Contribution in Individuals With Early-Onset Coronary Artery Disease. *Circ Genom Precis Med*, 11(1):e001849, 2018. doi: 10.1161/CIRCGEN.117.001849.
- [60] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation*, 129(25):S49S73, 2014. doi: 10.1161/01.cir.0000437741.48606.98.
- [61] R. B. DAgostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6):743753, 2008. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.107.699579.
- [62] C. Koz, O. Baysan, A. Hasimi, M. Cihan, et al. Conventional and non-conventional coronary risk factors in male premature coronary artery disease patients already having a low Framingham risk score. *Acta Cardiologica*, 63(5):623628, 2008. ISSN 0001-5385. doi: 10.2143/AC.63.5.2033231.
- [63] D. H. Becker and L. B. Gardner. *Prevention in Clinical Practice*. Springer Science Business Media, 2012. ISBN 978-1-4684-5356-0. Google-Books-ID: EVRD-BAAAQBAJ.
- [64] I. M. Graham, A. E. Di, F. Visseren, B. D. De, et al. Systematic Coronary Risk Evaluation (SCORE). *Journal of the American College of Cardiology*, 77(24):30463057, 2021. doi: 10.1016/j.jacc.2021.04.052.

- [65] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA*, 297(6):611619, 2007. ISSN 1538-3598. doi: 10.1001/jama.297.6.611.
- [66] S. Yusuf, S. Rangarajan, K. Teo, S. Islam, et al. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. *The New England Journal of Medicine*, 371(9):818827, 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1311890.
- [67] M. Woodward, P. Brindle, H. Tunstall-Pedoe, and SIGN group on risk estimation. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart (British Cardiac Society)*, 93(2):172176, 2007. ISSN 1468-201X. doi: 10.1136/hrt.2006.108167.
- [68] J. Hippisley-Cox, C. Coupland, and P. Brindle. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*, 357:j2099, 2017. ISSN 1756-1833. doi: 10.1136/bmj.j2099.
- [69] G. Assmann, P. Cullen, and H. Schulte. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. *Circulation*, 105(3):310315, 2002. ISSN 1524-4539. doi: 10.1161/hc0302.102575.
- [70] L. Palmieri, R. Rielli, L. Demattè, C. Donfrancesco, et al. CUORE project: implementation of the 10-year risk score. *European Journal of Cardiovascular Prevention Rehabilitation*, 18(4):642649, 2011. ISSN 1741-8267. doi: 10.1177/1741826710389925.

- [71] T. J. Anderson, J. Grégoire, R. A. Hegele, P. Couture, G.B. John Mancini, et al. 2012 update of the canadian cardiovascular society guidelines for the diagnosis and treatment of dyslipidemia for the prevention of cardiovascular disease in the adult. *Canadian Journal of Cardiology*, 29(2):151167, 2013. ISSN 0828282X. doi: 10.1016/j.cjca.2012.11.032.
- [72] J. Patel, M. Al Rifai, M. T. Scheuner, S. Shea, et al. Basic vs More Complex Definitions of Family History in the Prediction of Coronary Heart Disease: The Multi-Ethnic Study of Atherosclerosis. *Mayo Clinic Proceedings*, 93(9):12131223, 2018. ISSN 1942-5546. doi: 10.1016/j.mayocp.2018.01.014.
- [73] S. Sivapalaratnam, S.M. Boekholdt, M.D. Trip, M.S. Sandhu, et al. Family history of premature coronary heart disease and risk prediction in the EPIC-Norfolk prospective population study. *Heart (British Cardiac Society)*, 96(24):19851989, 2010. ISSN 1355-6037. doi: 10.1136/hrt.2010.210740.
- [74] D. Klarin and P. Natarajan. Clinical utility of polygenic risk scores for coronary artery disease. *Nature reviews. Cardiology*, 19(5):291301, 2022. ISSN 1759-5002. doi: 10.1038/s41569-021-00638-w.
- [75] R. Lali, E. Cui, A. Ansarikaleibari, M. Pigeyre, and G. Paré. Genetics of early-onset coronary artery disease: from discovery to clinical translation. *Current Opinion in Cardiology*, 34(6):706713, 2019. ISSN 1531-7080. doi: 10.1097/HCO.0000000000000676.
- [76] "Writing Committee for the VISION Study Investigators" et al. Association of Post-operative High-Sensitivity Troponin Levels With Myocardial Injury and 30-Day Mortality Among Patients Undergoing Noncardiac Surgery. *JAMA*, 317(16):16421651, 2017. ISSN 1538-3598. doi: 10.1001/jama.2017.4360.

- [77] N. R. Smilowitz, G. Redel-Traub, A. Hausvater, A. Armanious, et al. Myocardial Injury After Noncardiac Surgery: A Systematic Review and Meta-Analysis. *Cardiology in Review*, 27(6):267273, 2019. ISSN 1538-4683. doi: 10.1097/CRD.0000000000000254.
- [78] E. Duceppe, J. Parlow, P. MacDonald, K. Lyons, et al. Canadian Cardiovascular Society Guidelines on Perioperative Cardiac Risk Assessment and Management for Patients Who Undergo Noncardiac Surgery. *The Canadian Journal of Cardiology*, 33(1):1732, 2017. ISSN 1916-7075. doi: 10.1016/j.cjca.2016.09.008.
- [79] F. Botto, P. Alonso-Coello, M. T. V. Chan, J. C. Villar, et al. Myocardial injury after noncardiac surgery: a large, international, prospective cohort study establishing diagnostic criteria, characteristics, predictors, and 30-day outcomes. *Anesthesiology*, 120(3):564578, 2014. ISSN 1528-1175. doi: 10.1097/ALN.0000000000000113.
- [80] K. Thygesen, J. S. Alpert, A. S. Jaffe, B. R. Chaitman, et al. Fourth Universal Definition of Myocardial Infarction (2018). *Journal of the American College of Cardiology*, 72(18):22312264, 2018. ISSN 0735-1097. doi: 10.1016/j.jacc.2018.08.1038.
- [81] L. Babuin and A. S. Jaffe. Troponin: the biomarker of choice for the detection of cardiac injury. *CMAJ: Canadian Medical Association Journal*, 173(10):11911202, 2005. ISSN 0820-3946. doi: 10.1503/cmaj.050141.
- [82] Cian McCarthy, Sean Murphy, Joshua A. Cohen, Saad Rehman, Maeve Jones-OConnor, David S. Olshan, A. Singh, M. Vaduganathan, J. L. Januzzi, and J. H. Wasfy. Misclassification of myocardial injury as myocardial infarction. *JAMA Cardiology*, 4(5):460464, 2019. ISSN 2380-6583. doi: 10.1001/jamacardio.2019.0716.
- [83] K. Ruetzler, N. R. Smilowitz, J. S. Berger, P. J. Devereaux, et al. Diagnosis and Man-

- agement of Patients With Myocardial Injury After Noncardiac Surgery: A Scientific Statement From the American Heart Association. *Circulation*, 144(19):e287e305, 2021. doi: 10.1161/CIR.0000000000001024.
- [84] D. M. Gualandro, C. A. Campos, D. Calderaro, P. C. Yu, et al. Coronary plaque rupture in patients with myocardial infarction after noncardiac surgery: Frequent and dangerous. *Atherosclerosis*, 222(1):191195, 2012. ISSN 00219150. doi: 10.1016/j.atherosclerosis.2012.02.021.
- [85] T. Sheth, M. Chan, C. Butler, B. Chow, et al. Prognostic capabilities of coronary computed tomographic angiography before non-cardiac surgery: prospective cohort study. *BMJ (Clinical research ed.)*, 350:h1907, 2015. ISSN 1756-1833. doi: 10.1136/bmj.h1907.
- [86] F. Ujueta, J. S. Berger, and N. Smilowitz. Coronary Angiography in Patients With Perioperative Myocardial Injury After Non-Cardiac Surgery. *The Journal of Invasive Cardiology*, 30(9):E90E92, 2018. ISSN 1557-2501.
- [87] M. C. Cohen and T. H. Aretz. Histological analysis of coronary artery lesions in fatal postoperative myocardial infarction. *Cardiovascular Pathology*, 8(3):133139, 1999. ISSN 1054-8807. doi: 10.1016/S1054-8807(98)00032-5.
- [88] T. H. Lee, E. R. Marcantonio, C. M. Mangione, E. J. Thomas, C. A. Polanczyk, et al. Derivation and Prospective Validation of a Simple Index for Prediction of Cardiac Risk of Major Noncardiac Surgery. *Circulation*, 100(10):10431049, 1999. doi: 10.1161/01.CIR.100.10.1043.
- [89] O. Ali. Genetics of type 2 diabetes. *World Journal of Diabetes*, 4(4):114123, 2013. ISSN 1948-9358. doi: 10.4239/wjd.v4.i4.114.

- [90] J. B. Cole and J. C. Florez. Genetics of diabetes and diabetes complications. *Nature reviews. Nephrology*, 16(7):377390, 2020. ISSN 1759-5061. doi: 10.1038/s41581-020-0278-5.
- [91] N. J. Douville, I. Surakka, A. Leis, C. B. Douville, et al. Use of a Polygenic Risk Score Improves Prediction of Myocardial Injury after Non-cardiac Surgery. *Circulation. Genomic and precision medicine*, 13(4):e002817, 2020. ISSN 2574-8300. doi: 10.1161/CIRCGEN.119.002817.
- [92] M. Fischer, U. Broeckel, S. Holmer, A. Baessler, et al. Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. *Circulation*, 111(7):855862, 2005. ISSN 1524-4539. doi: 10.1161/01.CIR.0000155611.41961.BB.
- [93] V. Hyttinen, J. Kaprio, L. Kinnunen, M. Koskenvuo, and J. Tuomilehto. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes*, 52(4):10521055, 2003. ISSN 0012-1797. doi: 10.2337/diabetes.52.4.1052.
- [94] A. J. Walley, A. I. F. Blakemore, and P. Froguel. Genetics of obesity and the prediction of risk for health. *Human Molecular Genetics*, 15:124–130, 2006. ISSN 0964-6906. doi: 10.1093/hmg/ddl215.
- [95] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five Years of GWAS Discovery. *American Journal of Human Genetics*, 90(1):724, 2012. ISSN 0002-9297. doi: 10.1016/j.ajhg.2011.11.029.
- [96] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, et al. Exome sequencing as a

- tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745756, 2011. ISSN 14710056. doi: 10.1038/nrg3031.
- [97] D. F. Easton, Paul D.P. Pharoah, A. C. Antoniou, M. Tischkowitz, et al. Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *The New England journal of medicine*, 372(23):22432257, 2015. ISSN 0028-4793. doi: 10.1056/NEJMSr1501341.
- [98] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):4147, 2016. ISSN 1476-4687. doi: 10.1038/nature18642.
- [99] M. Gatz, C. A. Reynolds, L. Fratiglioni, B. Johansson, et al. Role of Genes and Environments for Explaining Alzheimer Disease. *Archives of General Psychiatry*, 63(2):168174, 2006. ISSN 0003-990X. doi: 10.1001/archpsyc.63.2.168.
- [100] K. Michailidou, S. Lindström, J. Dennis, J. Beesley, S. Hui, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):9294, 2017. ISSN 0028-0836. doi: 10.1038/nature24284.
- [101] R. Sims, S. J. Lee, A. C. Naj, C. Bellenguez, N. Badarinarayan, et al. Rare coding variants in *PLCG2*, *ABI3* and *TREM2* implicate microglial-mediated innate immunity in Alzheimers disease. *Nature genetics*, 49(9):1373, 2017. doi: 10.1038/ng.3916.
- [102] A. Torkamani, N. E. Wineinger, and E. J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581581, 2018. ISSN 14710056. doi: 10.1038/s41576-018-0018-x.
- [103] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic

- mutations. *Nature Genetics*, 50(99):12191224, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0183-z.
- [104] N. R. Wray, T. Lin, J. Austin, J. J. McGrath, et al. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA psychiatry*, 78(1):101109, 2021. ISSN 2168-6238. doi: 10.1001/jamapsychiatry.2020.3049.
- [105] A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, et al. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9):10261034, 2017. ISSN 0002-9262. doi: 10.1093/aje/kwx246.
- [106] Y. Wang, K. Tsuo, M. Kanai, B. M. Neale, and A. R. Martin. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annual Review of Biomedical Data Science*, 5:293320, 2022. doi: 10.1146/annurev-biodatasci-111721-074830.
- [107] R. Plomin and S. von Stumm. Polygenic scores: prediction versus explanation. *Molecular Psychiatry*, 27(11):4952, 2022. ISSN 1476-5578. doi: 10.1038/s41380-021-01348-y.
- [108] T. J. C. Polderman, B. Benyamin, Ch. A. de Leeuw, P. F. Sullivan, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(77):702709, 2015. ISSN 1546-1718. doi: 10.1038/ng.3285.
- [109] G. H. Freeman. Statistical methods for the analysis of genotype-environment interactions. *Heredity*, 31(33):339354, 1973. ISSN 1365-2540. doi: 10.1038/hdy.1973.90.
- [110] A. I. Young, S. Benonisdottir, M. Przeworski, and A. Kong. Deconstructing the

- sources of genotype-phenotype associations in humans. *Science (New York, N.Y.)*, 365(6460):13961400, 2019. ISSN 0036-8075. doi: 10.1126/science.aax3710.
- [111] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100(4):635649, 2017. ISSN 1537-6605. doi: 10.1016/j.ajhg.2017.03.004.
- [112] L. Duncan, H. Shen, B. Gelaye, J. Meijisen, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10(11):3328, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11112-0.
- [113] R.C. Lewontin. The Interaction of Selection and Linkage. I. General considerations, heterotic models. *Genetics*, 1964.
- [114] W. F. R. Weldon. MENDELS LAWS OF ALTERNATIVE INHERITANCE IN PEAS. *Biometrika*, 1(2):228233, 1902. ISSN 0006-3444. doi: 10.1093/biomet/1.2.228.
- [115] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones. A guide to machine learning for biologists. *Nature Reviews. Molecular Cell Biology*, 23(1):4055, 2022. ISSN 14710072. doi: 10.1038/s41580-021-00407-0.
- [116] P. Loh, G. Bhatia, A. Gusev, H. K. Finucane, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):13851396, 2015. ISSN 10614036. doi: 10.1038/ng.3431.
- [117] G. Paré, S. Mao, and W. Q. Deng. A robust method to estimate regional polygenic correlation under misspecified linkage disequilibrium structure. *Genetic Epidemiology*, (7):636647, 2018. ISSN 1098-2272. doi: 10.1002/gepi.22149.

- [118] H. Shi, G. Kichaev, and B. Pasaniuc. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics*, 99(1):139153, 2016. ISSN 00029297. doi: 10.1016/j.ajhg.2016.05.013.
- [119] N. H. Barton, A. M. Etheridge, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:5073, 2017. ISSN 0040-5809. doi: 10.1016/j.tpb.2017.06.001.
- [120] M. Pigeyre, J. Sjaarda, S. Mao, M. Chong, et al. Identification of Novel Causal Blood Biomarkers Linking Metabolically Favorable Adiposity With Type 2 Diabetes Risk. *Diabetes Care*, 42(9):18001808, 2019. ISSN 0149-5992. doi: 10.2337/dc18-2444.
- [121] C. Chatterjee and D. L. Sparks. Hepatic Lipase, High Density Lipoproteins, and Hypertriglyceridemia. *The American Journal of Pathology*, 178(4):14291433, 2011. ISSN 0002-9440. doi: 10.1016/j.ajpath.2010.12.050.
- [122] I.L. Ruel, P. Couture, J.S. Cohn, and B. Lamarche. Plasma metabolism of apob-containing lipoproteins in patients with hepatic lipase deficiency. *Atherosclerosis*, 180(2):355366, 2005. ISSN 0021-9150. doi: 10.1016/j.atherosclerosis.2004.12.014.
- [123] S. W. Choi, T. S. H. Mak, and O'reilly P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, 15(9):27592773, September 2020. ISSN 17542189. doi: 10.1038/s41596-020-0353-1.
- [124] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G.B. Blum. Making the most of clumping and thresholding for polygenic scores. *American Journal of Human Genetics*, 105(6):12131221, 2019. ISSN 0002-9297. doi: 10.1016/j.ajhg.2019.11.001.
- [125] G. Paré, Sh. Mao, and W. Q. Deng. A machine-learning heuristic to improve gene

- score prediction of polygenic traits. *Scientific Reports*, 7(11):12665, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-13056-1.
- [126] J. Euesden, C. M. Lewis, and P. F. O'Reilly. PRSice: Polygenic Risk Score software. *Bioinformatics (Oxford, England)*, 31(9):14661468, 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu848.
- [127] S. W. Choi and P. F. O'Reilly. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, 8(7), 2019. doi: 10.1093/gigascience/giz082. URL [https://journals.scholarsportal.info/details/2047217x/v08i0007/nfp\\_pprssfbd.xml](https://journals.scholarsportal.info/details/2047217x/v08i0007/nfp_pprssfbd.xml).
- [128] W. Liu, Z. Zhuang, W. Wang, T. Huang, and Z. Liu. An Improved Genome-Wide Polygenic Score Model for Predicting the Risk of Type 2 Diabetes. *Frontiers in Genetics*, 12:632385, 2021. ISSN 1664-8021. doi: 10.3389/fgene.2021.632385.
- [129] J. Li, D. P. Chaudhary, A. Khan, C. Griessenauer, et al. Polygenic Risk Scores Augment Stroke Subtyping. *Neurology: Genetics*, 7(2):e560, 2021. ISSN 2376-7839. doi: 10.1212/NXG.0000000000000560.
- [130] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469480, 2017. ISSN 1098-2272. doi: 10.1002/gepi.22050.
- [131] F. Privé, J. Arbel, H. Aschard, and B. J. Vilhjálmsson. Identifying and correcting for misspecifications in gwas summary statistics and polygenic scores. *Human Genetics and Genomics Advances*, 3(4):100136, 2022. ISSN 2666-2477. doi: 10.1016/j.xhgg.2022.100136.

- [132] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267288, 1996. ISSN 0035-9246.
- [133] M. Elgart, G. Lyons, S. Romero-Brufau, N. Kurniansyah, et al. Non-linear machine learning models incorporating snps and prs improve polygenic prediction in diverse human populations. *Communications Biology*, 5, 2022. doi: 10.1038/s42003-022-03812-z. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9395509/>.
- [134] J. Elliott, B. Bodinier, T. A. Bond, M. Chadeau-Hyam, et al. Predictive Accuracy of a Polygenic Risk Score Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA*, 323(7):636645, 2020. ISSN 0098-7484. doi: 10.1001/jama.2019.22241.
- [135] I. Shim, H. Kuwahara, N. Chen, M. O. Hashem, et al. Clinical utility of polygenic scores for cardiometabolic disease in Arabs. *Nature Communications*, 14(11):6535, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-41985-1.
- [136] T. Ge, C. Chen, Y. Ni, Y. A. Feng, and J. W. Smoller. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(11):1776, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09718-5.
- [137] S. Song, L. Hou, and J. S. Liu. A data-adaptive Bayesian regression approach for polygenic risk prediction. *Bioinformatics*, 38(7):19381946, 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac024.
- [138] S. Zabad, S. Gravel, and Y. Li. Fast and accurate Bayesian polygenic risk modeling with variational inference. *The American Journal of Human Genetics*, 110(5):741761, 2023. ISSN 0002-9297. doi: 10.1016/j.ajhg.2023.03.009.

- [139] J. Joyce. *Bayes Theorem*. Metaphysics Research Lab, Stanford University, fall 2021 edition, 2021. URL <https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/>.
- [140] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 1st ed. 2006. corr. 2nd printing 2011 edition edition, 2006. ISBN 978-0-387-31073-2.
- [141] F. Privé, C. Albiñana, J. Arbel, B. Pasaniuc, and B. J. Vilhjálmsson. Inferring disease architecture and predictive ability with LDpred2-auto. *The American Journal of Human Genetics*, 110(12):20422055, 2023. ISSN 00029297. doi: 10.1016/j.ajhg.2023.10.010.
- [142] H. D. Manikpurage, A. Paulin, A. Girard, A. Eslami, et al. Contribution of Lipoprotein(a) to Polygenic Risk Prediction of Coronary Artery Disease: A Prospective UK Biobank Analysis. *Circulation: Genomic and Precision Medicine*, 16(5):470477, 2023. doi: 10.1161/CIRCGEN.123.004137.
- [143] G. Ni, J. Zeng, J. A. Revez, Y. Wang, et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. page 2020.09.10.20192310, 2021. doi: 10.1101/2020.09.10.20192310. URL <https://www.medrxiv.org/content/10.1101/2020.09.10.20192310v2>.
- [144] S. K. Pemmasani, S. Atmakuri, and A. Acharya. Genome-wide polygenic risk score for type 2 diabetes in indian population. *Scientific Reports*, 13(11):11568, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-38768-5.
- [145] "National Research Council", "Division on Earth, Life Studies", "Board on Life Sciences", and "Committee on a Framework for Developing a New Taxonomy of Disease"s. *Toward Precision Medicine: Building a Knowledge Network for Biomedical*

- Research and a New Taxonomy of Disease*. National Academies Press, Washington, illustrated edition edition, 2012. ISBN 978-0-309-22222-8.
- [146] K. D. Christensen, D. Dukhovny, U. Siebert, and R. C. Green. Assessing the Costs and Cost-Effectiveness of Genomic Sequencing. *Journal of Personalized Medicine*, 5(4):470486, 2015. ISSN 2075-4426. doi: 10.3390/jpm5040470.
- [147] A. Adeyemo, M. K. Balaconis, D. R. Darnes, S Fatumo, et al. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nature Medicine*, 27(1111):18761884, 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01549-6.
- [148] M. Kiflen, A. Le, S. Mao, R. Lali, et al. Cost-Effectiveness of Polygenic Risk Scores to Guide Statin Therapy for Cardiovascular Disease Prevention. *Circulation: Genomic and Precision Medicine*, 15(5):e003423, 2022. doi: 10.1161/CIRCGEN.121.003423.
- [149] M. T. Scheuner, W. C. Whitworth, H. McGruder, P. W. Yoon, and M. J. Khoury. Familial risk assessment for early-onset coronary heart disease. *Genetics in Medicine*, 8(8):525531, 2006. ISSN 1098-3600. doi: 10.1097/01.gim.0000232480.00293.00.
- [150] M. Inouye, G. Abraham, C. P. Nelson, A. M. Wood, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *Journal of the American College of Cardiology*, 72(16):18831893, 2018. ISSN 0735-1097. doi: 10.1016/j.jacc.2018.07.079.
- [151] N. Mars, J. T. Koskela, P. Ripatti, T. T. J. Kiiskinen, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine*, 26(44):549557, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0800-0.

- [152] L. Weng, S. Khurshid, S. Gunn, L. Trinquart, et al. Clinical and Genetic Atrial Fibrillation Risk and Discrimination of Cardioembolic From Noncardioembolic Stroke. *Stroke*, 54(7):17771785, 2023. doi: 10.1161/STROKEAHA.122.041533.
- [153] X. Jiang, C. Holmes, and G. McVean. The impact of age on genetic risk for common diseases. *PLoS Genetics*, 17(8):e1009723e1009723, 2021. ISSN 15537390. doi: 10.1371/journal.pgen.1009723.
- [154] A. C. Fahed, M. Wang, J. R. Homburger, A. P. Patel, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications*, 11(11):3635, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17374-3.
- [155] N. Mavaddat, P. D. P. Pharoah, K. Michailidou, J. Tyrer, M. N. Brook, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *Journal of the National Cancer Institute*, 107(5):djv036, 2015. ISSN 1460-2105. doi: 10.1093/jnci/djv036.
- [156] L. Hsu, J. Jeon, H. Brenner, S. B. Gruber, et al. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology*, 148(7):1330–1339.e14, 2015. ISSN 1528-0012. doi: 10.1053/j.gastro.2015.02.010.
- [157] K. Bibbins-Domingo, D. C. Grossman, and S. J. Curry. The US Preventive Services Task Force 2017 Draft Recommendation Statement on Screening for Prostate Cancer: An Invitation to Review and Comment. *JAMA*, 317(19):19491950, 2017. ISSN 1538-3598. doi: 10.1001/jama.2017.4413.
- [158] N. Pashayan, S. W. Duffy, D. E. Neal, F. C. Hamdy, et al. Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genetics in Medicine*, 17(10):789, 2015. doi: 10.1038/gim.2014.192.

- [159] R. A. Oram, K. Patel, A. Hill, B. Shields, et al. A Type 1 Diabetes Genetic Risk Score Can Aid Discrimination Between Type 1 and Type 2 Diabetes in Young Adults. *Diabetes Care*, 39(3):337344, 2015. ISSN 0149-5992. doi: 10.2337/dc15-1111.
- [160] B. B. Adhyaru and T. A. Jacobson. Safety and efficacy of statin therapy. *Nature Reviews Cardiology*, 15(12):757760, 2018. ISSN 17595002. doi: 10.1038/s41569-018-0098-5.
- [161] A. F. Macedo, F. C. Taylor, J. P. Casas, A. Adler, et al. Unintended effects of statins from observational studies in the general population: systematic review and meta-analysis. *BMC medicine*, 12:51, 2014. ISSN 1741-7015. doi: 10.1186/1741-7015-12-51.
- [162] N. Sattar, D. Preiss, H. M. Murray, P. Welsh, et al. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. *The Lancet*, 375(9716):735742, 2010. ISSN 01406736. doi: 10.1016/S0140-6736(09)61965-6.
- [163] N. R. Cook and P. M. Ridker. Calibration of the Pooled Cohort Equations for Atherosclerotic Cardiovascular Disease: An Update. *Annals of Internal Medicine*, 165(11):786794, 2016. ISSN 0003-4819. doi: 10.7326/M16-1739.
- [164] J. S. Rana, G. H. Tabada, M. D. Solomon, J. C. Lo, et al. Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic real-world population. *Journal of the American College of Cardiology*, 67(18):21182130, 2016. ISSN 0735-1097. doi: 10.1016/j.jacc.2016.02.055.
- [165] G. Abraham, A. S. Havulinna, O. G. Bhalala, S. G. Byars, et al. Genomic prediction of coronary heart disease. *European Heart Journal*, 37(43):32673278, 2016. ISSN 0195-668X. doi: 10.1093/eurheartj/ehw450.

- [166] J. L. Mega, N. O. Stitzel, J. G. Smith, D. I. Chasman, et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet*, 385(9984):22642271, 2015. ISSN 01406736. doi: 10.1016/S0140-6736(14)61730-X.
- [167] P. Natarajan, R. Young, N. O. Stitzel, S. Padmanabhan, et al. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation*, 135(22):20912101, 2017. doi: 10.1161/CIRCULATIONAHA.116.024436.
- [168] S. Saya, J. G. McIntosh, I. M. Winship, M. Clendenning, et al. A Genomic Test for Colorectal Cancer Risk: Is This Acceptable and Feasible in Primary Care? *Public Health Genomics*, 23(3/4):110121, 2020. ISSN 1662-4246.
- [169] E. Widén, N. Junna, S. Ruotsalainen, I. Surakka, et al. How Communicating Polygenic and Clinical Risk for Atherosclerotic Cardiovascular Disease Impacts Health Behavior: an Observational Follow-up Study. *Circulation: Genomic and Precision Medicine*, 15(2):e003459, 2022. doi: 10.1161/CIRCGEN.121.003459.
- [170] B. P. Turnwald, J. P. Goyer, D. Z. Boles, A. Silder, et al. Learning ones genetic risk changes physiology independent of actual genetic risk. *Nature Human Behaviour*, 3(11):4856, 2019. ISSN 2397-3374. doi: 10.1038/s41562-018-0483-4.
- [171] C. Sudlow, J. Gallacher, N. Allen, V. Beral, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):e1001779, 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779.

- [172] "National Health Service (NHS)". *Office of Population Censuses and Surveys Classification of Interventions and Procedures, Version 4*. 1989. Available online: URL.
- [173] "World Health Organization (WHO)". *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. World Health Organization, Genève, Switzerland, 1993. ISBN 9789241544559.
- [174] "The Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Study Investigators", J. Spence, Y. LeManach, M. T. V. Chan, et al. Association between complications and death within 30 days after noncardiac surgery. *CMAJ*, 191(30):E830E837, 2019. ISSN 0820-3946, 1488-2329. doi: 10.1503/cmaj.190221.
- [175] H. Wand, S. A. Lambert, C. Tamburro, M. A. Iacocca, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, 591(78497849): 211219, March 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03243-6.
- [176] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, et al. A tutorial on conducting genomewide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608, February 2018. ISSN 1049-8931. doi: 10.1002/mpr.1608.
- [177] E. P. Hong and J. W. Park. Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, 10(2):117122, June 2012. ISSN 1598-866X. doi: 10.5808/GI.2012.10.2.117.
- [178] L. R. Nassar, G. P. Barber, A. Benet-Pagès, J. Casper, et al. The ucsc genome browser database: 2023 update. *Nucleic Acids Research*, 51(D1):D1188D1195, January 2023. ISSN 1362-4962. doi: 10.1093/nar/gkac1072.

- [179] L. M. Chen, N. Yao, E. Garg, Y. Zhu, et al. Prs-on-spark (prsos): a novel, efficient and flexible approach for generating polygenic risk scores. *BMC bioinformatics*, 19(1):295, August 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2289-9.
- [180] I. R. König, C. Loley, J. Erdmann, and A. Ziegler. How to include chromosome x in your genomewide association study. *Genetic Epidemiology*, 38(2):97103, 2014. ISSN 0741-0395. doi: 10.1002/gepi.21782.
- [181] F. Privé, B. J. Vilhjálmsson, and T. S. H. Mak. lassosum2: an updated version complementing ldpred2. page 2021.03.29.437510, March 2021. doi: 10.1101/2021.03.29.437510. URL <https://www.biorxiv.org/content/10.1101/2021.03.29.437510v1>.

## Chapter 2

# Hypothesis

### 2.1 General Hypothesis, Objective & Approach

#### 2.1.1 General Hypothesis

Our hypothesis suggests that incorporating innovative statistical techniques to refine polygenic risk scores will enhance their predictive accuracy for complex diseases. Furthermore, it is anticipated that polygenic risk scores will offer valuable insight into the etiology and pathophysiology of cardiovascular diseases.

#### 2.1.2 General Objectives

This PhD thesis is centrally focused on the optimization of polygenic risk scores for disease risk prediction, and exploring their potential clinical applications.

### 2.1.3 Rationale and Approach

Previous research has shown the substantial influence of genetics on prevalent disease. PRS have emerged as a promising tool for improving risk prediction. In the recent years, researchers worldwide have explored various strategies to develop a PRS suitable for clinical use, with broad application to a global scale. Despite significant achievements, there is still room for improvement, as the anticipated maximum potential for risk prediction (derived from heritability studies) has not yet been reached. By employing less conventional statistical approaches and meticulously selecting GWAS and phenotype data, our methodology aims to enhance upon methods. Additionally, there is substantial evidence indicating that integrating multiple PRS and regional genetic correlation information could enhance predictive accuracy, both of which are encompassed by our methods. Furthermore, the application of PRS to diseases should unveil novel insights regarding etiologies and pathophysiologies of certain conditions. This is particularly pertinent in the context of cardiovascular diseases, given the considerable evidence of the underlying polygenic architecture. We review the genetic and environmental causes of premature coronary artery disease (pCAD), a condition characterized by genetic influences which affects younger cohorts (Chapter 3). To highlight potential applications of PRS, we analyzed various MINS-related PRS to ascertain their effectiveness in predicting MINS, both independently and in conjunction with clinical risk scores (Chapter 4). Finally, we developed a PRS method incorporating Multi Adaptive Regression Splines (MARS) to combine multiple regional genotype principal components to attain optimal predictiveness (Chapter 5).

## Chapter 3

# What Causes Premature Coronary Artery Disease?

Ann Le<sup>1,2</sup>, Helen Peng<sup>1,8</sup>, Danielle Golinsky<sup>1,10</sup>, Matteo Di Scipio<sup>1,2,7</sup>, Ricky Lali<sup>1,6</sup> and Guillaume Paré<sup>1,2,3,4,5,6</sup>

1. Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada.
2. Department of Medical Sciences, McMaster University, Faculty of Health Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.
3. Department of Biochemistry and Biomedical Sciences, McMaster University, Faculty of Health Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.
4. Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada
5. Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada

6. Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street West, Hamilton ON L8L 4K1, Canada.

7. Department of Medicine, McMaster University, 1280 Main Street West, Hamilton ON L8L 4K1, Canada.

### 3.1 Forward

Coronary artery disease (CAD) is a complex disease which remains a leading cause of death worldwide, and while prevention and treatment of CAD itself has significantly improved and progressed steadily throughout the years, prognosis and overall understanding for early-onset or premature cases remains poor. Premature coronary artery disease (pCAD) refers to the condition of CAD occurring in patients younger than 65 years for women and 55 years for men. Mortality rates for pCAD have not declined over the years, despite the high societal burden of affected demographic being younger individuals. This manuscript provides a review for pCAD, seeking to address potential theories regarding the complex pathophysiology and etiologies of the condition. In particular, the genetic and environmental factors of pCAD are examined due to the complex nature of the trait. Additionally, we also explored the common clinical risk predictors for pCAD, along with the related conditions of spontaneous coronary artery dissection (SCAD) and clonal hematopoiesis of indeterminate potential (CHIP).

Due to pCAD being a condition of early onset, much of its etiology can be attributed to genetic causes. The monogenic causes are mostly centred around dyslipidemias, mainly familial hypercholesterolemia (elevated low-density lipoprotein [LDL] levels). pCAD also exhibits underlying polygenic influences, as polygenic risk scores have shown to be predictive of pCAD. Similarly, rare variant polygenic risk scores have demonstrated decent potential for pCAD prediction as well. Regarding clinical risk factors, hypertension and

type 2 diabetes mellitus were covered, along with traditional risk scores used for CAD predictions (e.g. Framingham risk score, AHA/ACC ASCVD risk calculator). As for environmental risk factors, smoking, opioid usage, alcohol, amphetamines, stress and exercise may have potential association with pCAD. In conclusion, the current nature of research limits the full understanding of pCAD, and many cases remain without a clear and identifiable cause. There is overall need to gain a better understanding of the etiological and pathophysiological causes of pCAD, in order to provide satisfactory approaches in pCAD prevention and treatment.

This manuscript was recently accepted to *Current Atherosclerosis Reports* in April 2024. Guillaume Paré was invited to compose a review on the etiology of pCAD and structured the thematic content accordingly. Ann Le spearheaded the writing of the manuscript and wrote a majority of the paper regarding genetics and clinical risk scores. All authors contributed to the review and critical reading and revision of the manuscript.

## 3.2 Abstract

**PURPOSE OF REVIEW:** This review provides an overview of genetic and non-genetic causes of premature coronary artery disease (pCAD).

**RECENT FINDINGS:** pCAD refers to coronary artery disease (CAD) occurring before the age of 65 years in women and 55 years in men. Both genetic and non-genetic risk factors may contribute to the onset of pCAD. Recent advances in the genetic epidemiology of pCAD have revealed the importance of both monogenic and polygenic contributions to pCAD. Familial hypercholesterolemia (FH) is the most common monogenic disorder associated with atherosclerotic pCAD. However, clinical overreliance on monogenic genes can result in overlooked genetic causes of pCAD, especially polygenic contributions. Non-genetic factors, notably smoking and drug use, are also important contributors to pCAD. Cigarette smoking has been observed in 25.5% of pCAD patients relative to 12.2% of non-pCAD patients. Finally, myocardial infarction (MI) associated with spontaneous coronary artery dissection (SCAD) may result in similar clinical presentations as atherosclerotic pCAD.

**SUMMARY:** Recognizing the genetic and non-genetic causes underlying premature coronary artery disease (pCAD) is important for appropriate prevention and treatment. Despite recent progress, pCAD remains incompletely understood, highlighting the need for both awareness and research.

### 3.3 Condensed Abstract

As a leading cause of global death, coronary artery disease (CAD) has been well investigated throughout the years and clinical approaches for its prevention and treatment has drastically improved over the years. Despite the improvements in clinical treatment of CAD, the early-onset or premature cases of CAD are not as well investigated. Premature coronary artery disease (pCAD) is defined as CAD occurring before the age of 65 years in women and 55 years in men. This manuscript provides a review for pCAD, covering genetic influences (monogenic and polygenic causes), environmental influences (drugs, stress/exercise), clinical risk factors and the related conditions of spontaneous coronary dissection (SCAD) and clonal hematopoiesis of indetermined potential (CHIP).

**KEYWORDS:** Premature coronary artery disease; Risk factors of premature coronary artery disease; Genetics of coronary artery disease; Polygenic risk scores; Lifestyle factors of coronary artery disease; Spontaneous coronary artery dissection

### 3.4 Introduction

Coronary artery disease (CAD) is the leading cause of death worldwide, with a reported 1 in 30 patients with CAD experiencing death each year [1]. In this review, the term CAD is used generally to refer to atherosclerotic CAD, which is pathologically characterized by the narrowing or blockage of the coronary arteries due to the buildup of cholesterol-based plaques [2, 3]. CAD is influenced by both genetic and environmental factors. The most common complication of CAD is myocardial infarction (MI), defined as the death of cardiac cells due to blockage of the coronary arteries, which can lead to sudden cardiac death (SCD) [4]. Early-onset or premature coronary artery disease (pCAD) is defined as

CAD occurring in patients younger than 65 years for women and 55 years for men [3]. Despite progress in prevention and treatment of CAD, mortality rates for pCAD have not declined and overall prognosis is poor, which is particularly sobering given the high societal burden of CAD in young individuals [5]. In individuals aged between 35 and 44 years of age, 62% of SCD cases have been ascribed to CAD [6].

As a disease of early onset, pCAD is often attributable to genetic risk factors. The most common Mendelian disease associated with pCAD is familial hypercholesterolemia (FH), characterized by lifelong, severe elevations of plasma low-density lipoprotein cholesterol (LDLc) concentrations [7, 8]. However, detection of FH has been reported to be low, especially in young adults, with only 2% of Europeans being diagnosed before the age of 18, and fewer than half of adult cases diagnosed before 40 years of age [9]. Furthermore, only a minority of patients with pCAD are found to carry a FH-causing mutation, and thus alternative methods of inheritance have been suggested, including polygenic contributions from both common and rare variants. Environmental and non-genetic risk factors also contribute to pCAD. These factors include smoking, opioid consumption, alcohol intake, amphetamine usage, stress, and obesity. Spontaneous Coronary Artery Dissection (SCAD), a condition characterized by incidental tearing within the coronary artery wall leading to the formation of a false lumen, has similar manifestations to atherosclerotic pCAD and can be easily overlooked [10]. This review explores the genetic and non-genetic risk determinants of pCAD, with a focus on pathophysiological perspectives (see Figure 3.1).

### **3.5 Genetics of pCAD**

Early onset diseases are frequently associated with stronger genetic predisposition [11, 12]. Heritability describes the proportion of variance in a trait that can be attributed to genetics

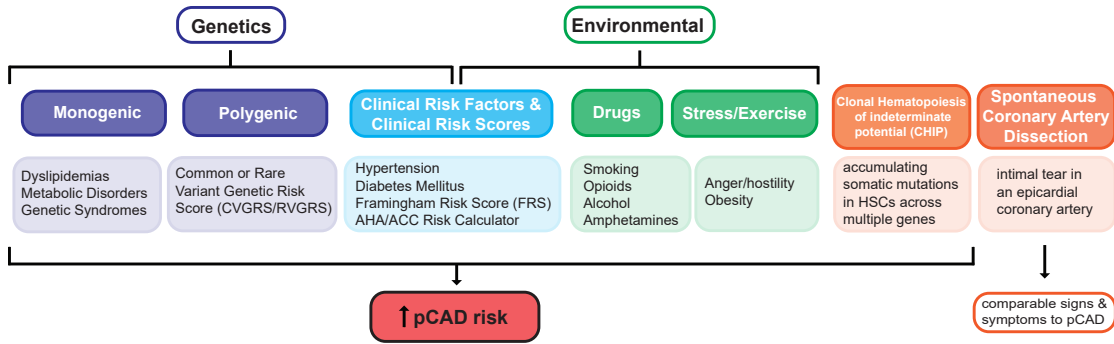


Figure 3.1: This review focuses on the genetic and non-genetic causes of premature coronary artery disease (pCAD). The genetic risk factors include monogenic and polygenic causes, including both common and rare variants. Environmental risk factors include smoking, drug usage and lifestyle choices (stress/exercise).

[13, 14]. Amongst other late-onset diseases, the heritability of CAD is relatively strong, ranging from 37% to as high as 57%, as estimated from twin and family studies [13, 15, 14]. Heritability for specific atherosclerotic phenotypes, including acute myocardial infarction, coronary artery calcification and carotid artery atherosclerosis, are also relatively high and generally ranges from 40 to 55% [16, 17]. Heritability tends to be higher for pCAD than for general CAD. Having a monozygotic twin who died from pCAD increases the relative hazard of pCAD substantially, relative to not having a twin who died from pCAD (male monozygotic HR = 8.1, male dizygotic HR = 10.5) [18]. More generally, individuals with a first-degree relative (parent or sibling) diagnosed with pCAD exhibit an approximately two-fold increase in risk of pCAD [19]. While late-onset CAD risk in first-degree relatives is also significantly associated with pCAD, the strength of the association is reduced by half [20].

The genetic risk underlying pCAD can be stratified into three categories: 1) monogenic, 2) common variant polygenic and 3) rare variant polygenic. While each mechanism can

independently accrue risk, they may also act in combination.

### 3.5.1 Monogenic Causes of pCAD

Monogenic or Mendelian disorders are single-gene diseases, in which affected individuals carry one or two copies of the effect allele, inherited in a specific p pattern (autosomal dominant, autosomal recessive, X-linked dominant and X-linked recessive) [21]. A monogenic predisposition to pCAD typically arises from rare, high impact mutations in genes responsible for encoding proteins which play pivotal roles in the metabolism of atherogenic lipoproteins (see Table 3.1).

The predominant monogenic disease associated with pCAD is familial hypercholesterolemia (FH), defined as a Mendelian autosomal dominant disorder characterized by lifelong levels of severely elevated plasma low-density lipoprotein cholesterol (LDLc) concentrations ( $\geq 190$  mg/dl in adults) [7, 8]. According to a recent review, it is estimated to affect approximately 1 in 300 individuals globally [22]. However, it should be noted that FH prevalence can vary significantly depending on ancestry [23, 24]. FH can clinically manifest as tuberous or tendinous xanthoma, which are localized cutaneous lipid deposits [25, 26, 27]. The three genes most frequently linked to FH are *LDLR* (low-density lipoprotein receptor), *APOB* (apolipoprotein B) and *PCSK9* (proprotein convertase subtilisin/kexin 9) [16, 28, 29]. Autosomal recessive forms of FH also exist and involve the gene encoding for the LDLR adaptor protein 1 (*LDRAP1*) [30, 31]. Irrespective of the gene involved, the condition leads to the impaired low-density lipoprotein receptor function, decreased hepatic clearance of LDLc particles, increased deposition of cholesterol in the inner layer of arteries, and ultimately atherosclerotic cardiovascular disease (ASCVD). FH mutations are associated with an average 50 mg/dl increase in LDLc [2]. Notably, even when LDLc levels are below the risk range ( $< 130$  mg/dl), individuals with FH mutations

are still associated with a 2.6-fold increase in general CAD risk. In pCAD populations, the same mutations confer an even higher association with a 3.7-fold increase in pCAD risk [32].

While elevated LDLc levels are often associated with the risk of CAD, there are monogenic mutations affecting other lipoproteins which also contribute. An ongoing debate revolves around elevated serum triglyceride (TG) levels and their role as an independent risk factor for CAD and atherogenesis [33]. Familial hyperchylomicronemia syndrome (FCS) is most commonly attributed to autosomal recessive mutation in the *LPL* gene coding for the lipoprotein lipase (LPL) enzyme [34, 35, 36]. The LPL enzyme is responsible for the hydrolysis of triglycerides found in chylomicrons and very low-density lipoproteins (VLDL). It is characterized by severely elevated triglycerides, including the pathological presence of chylomicrons during fasting states [36]. While genetic studies have demonstrated that Mendelian forms of FCS are causally linked to increased CAD and pCAD risk, the association between moderately elevated TG levels and CAD remains inconclusive due to confounding risk factors. Many risk factors for hypertriglyceridemia, such as body mass index (BMI), type II diabetes (T2D) and alcohol usage, are also risk factors for CAD [37]. Furthermore, results of clinical trials testing the effects of TG lowering agents on cardiovascular events have been mostly null [38, 39, 40, 41, 42], suggesting that only certain pathways leading to high TG levels are associated with CAD, while others are not. Through the use of Mendelian Randomization (MR) studies, which examine the causal effects of risk factors on diseases using genetic variants as instruments [43], LPL has been associated with an independent effect of TGs on CAD [44, 45]. Additionally, it has been observed that heterozygous carriers of the deleterious LPL mutation had 1.84-time greater odds of pCAD than non-carriers [16]. FCS may also arise from mutations that indirectly impair function of LPL. This includes biallelic loss-of-function mutations in genes related

to LPL function, such as *APOC2*, *APOA5*, *GP1HBP1* and *LMF1* [46, 47, 48, 49] Carriers of the *APOA5* mutation were observed to have a 2.2-fold increased risk for early onset MI and identified to have higher plasma TG levels relative to non-carriers [46]. Alternatively, mutations in *APOC3* have been shown to reduce TG levels and risk of CAD [50, 51, 52]. Apolipoprotein C-III (*APOC3*) inhibits LPL, leading to increased TG levels. Randomized control trials have shown it to be a promising target for treatment of elevated TG levels.

Intermediate density lipoproteins can contribute to elevated total cholesterol and triglyceride levels, and are considered to also increase the risk of pCAD [53]. Familial dysbetalipoproteinemia (FD), also known as type III hyperlipoproteinemia, is an underdiagnosed autosomal recessive disorder characterized by the accumulation of triglyceride-rich remnant lipoproteins. It occurs in individuals homozygote for the epsilon 2 ( $\epsilon/\epsilon2$ ) isoform of apolipoprotein E [54]. Other mutations can also lead to FD, and approximately 10% of FD patients exhibit dominant or codominant inheritance patterns [55, 56]. ApoE regulates lipoprotein levels via receptors of the LDL receptor gene family and cell-surface sulfate proteoglycans [57, 58]. The defining characteristic of FD is Palmar striated xanthomas affecting approximately 10% of FD patients [59]. It is characterized by yellow to brownish colouration of palmar and finger creases. A cross-section study of 305 European FD patients suggests peripheral artery disease (PAD) and CAD are the most common cardiovascular manifestations of FD (CAD prevalence = 19%, PAD prevalence = 4%) [60]. The prevalence of FD can vary under different study definitions. When defined using a total cholesterol level greater than the 90th percentile and the  $\epsilon2/\epsilon2$  genotype, the total global prevalence was estimated at 1 in 825 [61]. The apoE2 allele has a relatively high global frequency of 0.08 relative to other rare diseases, and only 15% of homozygous carriers develop dysbetalipoproteinemia [54, 62, 55]. Additionally, individuals who develop the disease can have regular lipid levels for decades, due to the absence of additional inherited or acquired

risk factors [62].

Sitosterolemia is a rare monogenic condition caused by recessive autosomal mutations in ATP-binding cassette subfamily G genes (*ABCG5* and *ABCG8*) and is associated with CAD [63, 64]. It is characterized by increased plasma concentrations of plant sterols (sitosterol, campesterol, stigmasterol), resulting from increased absorption of plant sterols and decreased secretion from the liver. It has similar manifestations to FH, such as tendinous, tuberous xanthomas and premature coronary atherosclerosis [63, 64, 65].

Elevated plasma homocysteine can present with asymptomatic or symptomatic homocystinuria, which has been established to be an independent risk factor for atherosclerosis [66, 67]. Typically inherited in an autosomal recessive pattern, it affects genes which code for key enzymes in homocysteine metabolism: cystathionine-beta-synthase (CBS) and methylenetetrahydrofolate reductase (*MTHFR*) [68]. Carriers of these mutations develop severe cardiovascular disease before the age of thirty, which is hypothesized to be the result of alterations in arterial structure and function as a result of damage to the vascular endothelium and smooth muscle cells [69, 70]. However, the role of homocystinuria on arterial function is not fully established, and the evidence supporting moderate increases in blood homocysteine levels as a CAD risk factor is inconclusive [68]. Recent randomized clinical trials assessing the effect of folic acid and vitamin B<sub>12</sub> (vital precursors to homocysteine metabolism) on homocysteine levels did not show evidence supporting their use in primary and secondary prevention of CAD [71, 72, 73, 74, 75].

There exist rarer monogenic conditions which lead to complications which may evolve into pCAD. Although their association to pCAD or CAD is considered controversial, some studies have shown their connections to cardiovascular events. These conditions include familial high-density lipoprotein deficiency I (Tangier Disease), autosomal dominant coro-

nary disease II, Williams Syndrome, Hutchinson-Gilford Progeria syndrome and pseudoxanthoma elasticum.

Disease	Global prevalence	Typical inheritance pattern	Genes affected	Pathogenic characteristics
<b>Dyslipidemias</b>				
Familial hypercholesterolemia	1/311 [22]	Autosomal dominant / autosomal recessive	<i>LDLR, APOB, PCSK9</i>	Severe elevated plasma LDLc levels for a lifetime leading to atherosclerotic plaques
Familial hyperchylomicronemia syndrome	1/500 [76]	Autosomal recessive	<i>LPL, APOC2, APOA5, GPIIIBP1, LMF1</i>	Elevated fasting triglyceride serum levels [77]
Familial dysbetalipoproteinemia	1/825[61]	Autosomal dominant / autosomal recessive	<i>APOE</i>	Presence of intermediate density lipoproteins leading to increased total cholesterol and triglycerides
Sitosterolemia	< 1/1000000 [63]	Autosomal recessive	<i>ABCG5/ABCG8</i>	Accumulation of plant sterols in the blood
Tangier disease (familial high-density lipoprotein deficiency I)	1/1000000 [78]	Autosomal recessive	<i>ABCA1</i>	Low HDL and tissue accumulation of cholesteryl esters [78, 79]
<b>Metabolic disorders</b>				
Homocystinuria	1/100 000[80]	Autosomal recessive	<i>CBS, MTHFR</i>	Elevated plasma and urinary homocysteine levels, causes damage to vascular endothelial cells, resulting CAD
Autosomal dominant coronary artery disease 2	Unknown	Autosomal dominant	<i>LRP6</i>	Loss-of-function of gene coding for Wnt/B-catenin signaling co-receptor regulating cell proliferation and tissue homeostasis [81]. Cardiovascular events (MI, stroke, sudden cardiac death) with a feature of metabolic syndrome (hypertension, hyperlipidemia and diabetes) [82, 83, 84, 85]
<b>Atherosclerotic complications of genetic syndromes</b>				
Williams Syndrome	< 1/75000[86]	Autosomal recessive	Deletion at 7q11.23; including <i>ELN</i>	Disrupted formation of elastic fibers in various tissues resulting in neurodevelopmental intellectual ability. [86]
Hutchinson-Gilford progeria syndrome	1/2 000 000[88]	Autosomal dominant	<i>LMNA</i>	Increased risk of occult CAD and progression of multi-site arterial stenosis. [87]
Pseudoxanthoma elasticum (Gronblad-Strandberg syndrome)	1/25 000 to 1/100 000	Autosomal recessive	<i>ABCC6</i>	Incompletely processed variant of the nuclear fibrillar protein lamin A resulting in premature, rapid aging [89, 90? ]
				Ectopic mineralization appearing in elastic tissues of the skin, eyes and the vascular system. Increases risk of MI, cerebral and peripheral artery disease due to structural and function changes in the arterial wall [91, 92]

Table 3.1: Monogenic diseases associated with pCAD pathogenesis.

## 3.6 Polygenic Causes of pCAD

### 3.6.1 Common Genetic Variant Studies

CAD is a complex disease, and while many monogenic forms have been described, genetic polygenicity also plays a role in its development[93]. Polygenicity refers to when a phenotypic trait is influenced by several genetic loci to varying degrees. Genome-wide association studies (GWAS) seek to detect associations between common genetic variants and a phenotypic trait[94]. Genotyping is usually performed using SNP arrays combined with statistical imputation of unobserved genotypes from reference panels. Genotyping chips typically cover common variants which occur at sufficient frequencies, usually defined as having an allele frequency of  $\geq 0.5\%$  (one carrier per 100 individuals)[95]. The influx of data from GWAS has allowed for identification of numerous CAD loci [96, 97]. A recent GWAS study by Aragam et al. revealed 279 genome-wide significant associations for CAD [96]. The conditionally independent causal loci reaching genome-wide significance accounted for 15.5% of CAD heritability. All suggestively significant associations ( $p < 2.52 \times 10^{-5}$ ) approximating a 1% false discovery rate cumulatively accounted for 36.1% of CAD heritability.

In 2007, the first CAD GWAS discovered the 9p21 locus, which remains the genetic locus most strongly associated with CAD and MI [98, 99, 100, 101]. Homozygous carriers of the 9p21 variant have an approximate 2-fold greater risk of disease within pCAD populations [99]. The locus contains enhancers which affect expression of tumor suppressor genes for cyclin-dependent kinases 2A and 2B (CDKN21A and CDKN2B), which regulate cell growth and proliferation [102]. Carriers of these variants have a population attributable risk of 21% for MI, and 31% in early onset cases [99]. While the exact mechanism of the 9p21 locus

remains elusive, it has been associated with atherosclerosis and plaque enlargement due to excessive cell proliferation within the arterial walls [103, 104, 42]. Additionally, studies have observed a decrease in expression of tumour suppressors p15 and p16 within aortic smooth muscle cells in the presence of the 9p21.3 locus [102, 105, 106]. This is associated with higher proliferation of aortic smooth muscle cells and lack of cell senescence, leading to further build-up of atherosclerotic plaques.

Polygenic risk scores (PRS) are often used as quantitative measures of an individual's genetic susceptibility towards a given phenotype. PRS are typically computed as a weighted sum of risk alleles across a large number of genetic variants associated with a given trait. More recent methodologies incorporate advanced statistical methods and machine learning to construct PRS [107, 108]. Since genotypes are determined at conception and PRS can be calculated at early ages, PRS hold much potential for prediction of early onset diseases. For various diseases, including CAD, it has been demonstrated that genetic relative risk tends to decrease with age when compared to the risk associated with environmental factors or other non-genetic factors [107]. The odds ratio (OR) for the 90th percentile PRS for CAD in the youngest age group is 3.63 (age < 45 years), dropping to 1.77 for the 90<sup>th</sup> percentile PRS for CAD in the oldest age group (age > 75 years). This suggests that while genetic risk might decline with age, it remains highly relevant in younger cohorts. It has been noted that risk detected by CAD PRS is comparable to risk detected by monogenic variants for FH, with individuals in the highest 5% of PRS having a comparable 3-fold increase in CAD risk, despite the much lower prevalence of FH (0.3%) [11, 109, 110]. Additionally, it has been observed that polygenic background can influence penetrance of monogenic mutations in CAD, and that risk captured by a CAD PRS may act largely independently from LDLc pathways [111].

The addition of CAD PRS to the clinically conventional AHA/ACC ASCVD risk calculator has been demonstrated to improve prediction for cardiovascular events, independent of clinical risk factors [112, 113]. Individuals with a higher genetic risk as determined by CAD PRS derive a greater relative and absolute benefit from both statin therapy and PCSK9 inhibitors to prevent a first CAD event [114, 115]. Thus, the use of PRS may prompt earlier intervention in younger adults for earlier statin initiation or lifestyle adjustments to prevent CAD onset.

There exist limitations to the use of PRS. Design of PRS is dependent on GWAS summary statistics, which may vary in quality and are regularly being published. Hence, PRS are constantly being updated. Currently, the majority of GWAS are performed in individuals of European ancestry, limiting global applications because differences in genetic ancestries lead to lower PRS predictiveness in non-European patients. The acquisition and storage of genetic information for PRS has also raised ethical and privacy concerns [16]. Finally, while it is recommended to use PRS in conjunction with clinical risk scores, there are currently no established standards or guidelines for the clinical application of PRS.

### **3.6.2 Rare Genetic Variant Studies**

The recent decline in cost for sequencing has allowed for rare variant association studies (RVAS), which can include rare variants that are not typically included on genotyping chips. GWAS have been noted to have poorer coverage of variants in the 0.5-5% allele frequency range [95]. This was first proposed due to the missing heritability problem, in which variants of lower ( $0.5\% \leq \text{MAF} \leq 5\%$ ) and rare ( $\text{MAF} < 0.5\%$ ) frequency could potentially account for genetic effects that can be missed in regular GWAS studies. Rare variants do not occur with sufficient frequency to associate tests based on individual variants, and thus require aggregation in order to assess association with disease [114]. Their role in

association with diseases could be significant due to their lower frequencies, which are likely due to purifying selection [95].

A recent study developed a method coined “RV-EXCALIBER” which utilizes summary statistics from an exome sequencing database to calculate rare variant burden over many genes to determine genetic association [115]. The method creates a weighted rare-variant genetic risk score (RVGRS) for CAD cases and controls in UK Biobank participants. The UK Biobank is a large epidemiology study consisting of over 500, 000 individuals aged 40-69 from across the UK [116]. The resource contains variables for patient characteristics and measurements including demographics, health diagnoses, physical measurements and lifestyle factors. RV-EXCALIBER was found a strong association between RVGRS and CAD in European and South Asian populations. Even after adjustment for known Mendelian CAD genes, clinical risk factors (Framingham Risk Score) and common-variant genetic scores (CVGRS), the RVGRS could identify 1.5% of the population with greater than a 2-fold risk of pCAD, despite the lower power of rare variants relative to common variants.

### **3.7 Clonal Hematopoiesis of Indeterminate Potential (CHIP)**

Clonal hematopoiesis of indeterminate potential (CHIP) has been demonstrated to be associated with premature MI and CAD. CHIP is an aging related condition occurring when hematopoietic stem cells (HSCs) acquire somatic mutations, leading to the development of blood cells with leukemic properties [117, 118, 119]. Over time, somatic mutations gradually accumulate within HSCs, with certain mutations eventually gaining an advantage over others [120]. Consequently, the prevalence of CHIP can vary significantly across age groups, ranging from 5% in individuals under 60 years old to 30% in those over 80 years

old [118, 121].

Given the absence of discernible signs and symptoms, CHIP can only be identified through genetic analysis. The two genes most commonly associated with CHIP are epigenetic regulators *DNMT3A* and *TET2*, which play a role in modification of DNA methylation [122, 123]. Mutations in *ASXL*, a gene involved in chromatin remodeling, are also associated with CHIP. The presence of CHIP has been linked to a two-fold increase in the occurrence of CAD [118]. Moreover, in patients between 45 and 50 years old, the presence of CHIP is associated with a four-fold increase in occurrence of premature MI. CHIP has been shown to contribute to accelerated atherogenesis, likely to due to its effect on inflammatory properties [118]. Specifically, this may result from alterations in transcriptional regulation of macrophages, which play a role in mediating inflammatory responses and are found within atherosclerotic plaques. Recently, CHIP has been found to be associated with an adverse outcome in patients with atherosclerotic cardiovascular disease in the presence of *TET2* or spliceosome mutations (*SFSB1/SRSF2/U2AF1*) [124].

However, it is important to acknowledge that the influence of CHIP on cardiovascular diseases remains a matter of debate, and some reports suggest a lack of association. For instance, a study involving 200,453 participants found no significant association between clonal hematopoiesis and certain ischemic cardiovascular diseases including CAD and stroke, with age as a strong confounding factor [125]. This suggests that previous associations may simply be a result of aging. Furthermore, causal relationships were not found between variants associated with clonal hematopoiesis and outcomes such as CAD, ischemic stroke and heart failure through Mendelian randomization analysis [125, 126]. Thus, while investigating the relationship between CHIP and cardiovascular diseases hold promise, further refinement is needed, considering that identifying the presence of CHIP

can be challenging and results may heavily rely on the quality of genetic variant calls.

## **3.8 Non-Genetic Risk Factors of pCAD**

While CAD is highly heritable, non-genetic risk including environmental and lifestyle factors also contribute to pCAD.

### **3.8.1 Clinical Risk Factors & Clinical Risk Scores**

While clinical risk factors and clinical risk scores are often viewed as separate from genetic factors, they entail elements of both environmental and genetic factors. Hypertension and type 2 diabetes mellitus (T2D) are two important risk factors for CAD [127, 128, 129, 130]. Hypertension has been shown to increase risk of pCAD by as much as 60% [131]. In the case of T2D, even when pCAD patients are not diagnosed with T2D, they are often seen to have mild disruption in glucose metabolism and compromised glucose tolerance [132]. A family history of diabetes is also associated with ASCVD and pCAD [133].

The Framingham risk score (FRS) was originally developed in 2008 and is one of the most used clinical risk predictors for CAD [134, 135]. It considers clinical risk factors: age, total cholesterol, high-density lipoprotein cholesterol (HDL-C), systolic blood pressure, blood pressure, smoking and diabetes to predict 10-year risk of incident cardiovascular events. Currently, the American College of Cardiology and American Heart Association recommends 10-year cardiovascular risk calculations using the AHA/ACC ASCVD risk calculator for adults aged 40-75, which references similar clinical risk factors as the FRS [5, 147]. Other variations of clinical risk scores include the European Systematic Coronary Risk Evaluation (SCORE) [136], the Reynolds risk score [137], the INTERHEART risk score [137, 138], the Assign risk score [139], the QRISK3 score [140], the PROCAM risk

score [141] and the CUORE risk score [142, 143, 144]. It has been demonstrated that individuals with pCAD have many modifiable risk factors (i.e. controllable risk factors such as diet, exercise, smoking) [145]. Some notable risk factors of the pCAD population include smoking and body mass index (BMI).

All these clinical scores were designed to predict later onset CAD and consider age as a strong risk factor. Hence, younger individuals are almost invariably considered at low risk and these scores have limited utility to predict pCAD [143]. Additionally, these scores do not consider family history and thus do not account for the genetic contributions to pCAD. It has been shown that incorporation of family history can significantly improve risk assessment, especially in pCAD populations [20, 146]. Furthermore, these scores tend to only be well calibrated to the population in which they are developed. For example, FRS, developed from a primarily white male cohort from 2008, has been shown to overestimate cardiovascular risk in women [144, 147, 148].

## **3.9 Smoking & Other Drugs of Abuse**

### **3.9.1 Smoking**

Smoking is a well-known non-genetic, modifiable risk factor for cardiovascular diseases [149]. The Framingham Heart study identified a 1.92-fold and 1.70-fold higher risk in pCAD with heavy smokers aged 35-44 relative to nonsmokers in men and women respectively [134]. In a study comparing gender and age-stratified pCAD cases and non-pCAD controls, cigarette smoking was observed in 25.5% of pCAD patients relative to 12.2% in non-pCAD patients [150].

Tobacco smoke contains a variety of harmful chemicals and substances which can con-

tribute to pCAD, including reactive oxygen species (ROS). The oxidative stress resulting from ROS is thought to be a contributor to inflammation and carcinogenesis. Specifically, ROS can cause endothelial activation, resulting in proinflammatory proliferation of endothelial cells and impaired vasodilation [151, 152]. Additionally, a simultaneous increase in ROS and decrease in antioxidant capacities can impact the oxidative potential of LDL. This allows macrophage uptake via scavenger receptors A and CD36, ultimately leading to foam cell creation and atherosclerotic deposits on arterial walls [153].

While cigarette smoking has been steadily decreasing over the last two decades, nicotine consumption remains on the rise due to increased consumption of e-cigarettes [154, 155, 156]. The 2020 National Youth Tobacco Survey (NYTS) reports 19.6% of American high school students and 4.7% of American middle school students as regular e-cigarette users [157, 158]. It has also been reported that e-cigarettes can act as a gateway to traditional cigarettes [159]. Though the association between e-cigarettes and CAD remains controversial, studies have suggested that certain toxic components within e-cigarettes may induce harmful cardiovascular effects. Regardless of history of traditional cigarette usage, adults who have normal flow-mediated vasodilation exhibit pronounced vasoconstriction upon exposure to e-cigarettes [160, 161, 162]. The increased stress from vasoconstriction may lead to an increase in atherosclerotic deposits [163]. Additionally, acrolein is a carcinogenic substance found within e-cigarettes, and has been shown to induce inflammation, oxidative dysfunction, and endothelial dysfunction [164].

### **3.9.2 Opioid Usage**

Opiates are non-synthetic narcotics derived from the opium poppy plant, belonging to the opioid class of drugs (which includes synthetic derivatives) [165]. Opioid use disorder (OUD) and opioid addiction is considered an epidemic in the United States and worldwide,

with over 2.1 million people affected in the United States and 16 million people worldwide [165, 166]. Over 120,000 deaths are attributed to opioids annually. Opiate usage has been increasing in North America over the past two decades, as indicated by the 53% rise in hospitalizations and deaths attributed to their usage [167].

Complications of chronic use and overdose of opioids include acute MI, arrhythmias and heart failure [167, 168]. The use of opium has been reported to be directly associated with cardiovascular complications such as hyperlipidemia (changes in LDL, HDL, and TG levels), oxidative stress, decreased physical activity, insulin resistance, and increased levels of homocysteine, fibrinogen, and PAI-1 [169, 170, 171]. Increased inflammation may also occur, due to an increase of antagonistic activity on inflammatory mediators such as interleukin-17, interleukin-1 and C-reactive protein (CRP) receptors, all of which have been associated with atherosclerosis. It should also be noted that opioids are commonly associated with respiratory depression, due to the impact on the receptors mediating respiratory neurons [184]. This may also manifest as an alternative mechanism in which opioids may cause cardiovascular complications through the induction of oxidative stress. Heritability of OUD is estimated to be approximately 50% [172]. Opioid overdose is the highest among individuals between the ages of 40 and 50, while the most common age for treatment of OUD is between 20 to 35 years old [173]. Notably, heroin overdoses occur mostly between the ages of 20 and 30. Opium is most commonly used in Southeast and Central Asia, with Iran accounting for 42% of global consumption [174]. The Milano-Iranian Study (MIran) observed 1011 young patients (males < 45 years, females < 55 years) with severe CAD who underwent diagnostic coronary angiography (CAG) at the Tehran Heart Center against 2002 controls [175]. Active opiate consumers had a 3.8-fold increase in pCAD risk relative to non-consumers. However, the effect of opioids on pCAD remains controversial, and some studies have found a lower use of opioids in pCAD cases

as compared to controls [176].

### 3.9.3 Alcohol

Young adults (aged 18 to 30 years) are considered the group that consumes the most alcohol and binge-drinks the most [177, 178]. The effect of alcohol on cardiovascular diseases remains contentious, due to the observations of both protective and deleterious effects depending on consumption levels [177, 179, 180]. Alcohol consumption has also been linked to an increase in CAD risk factors, such as high systolic blood pressure and LDLc levels. Abusive drinkers (consumption of  $\geq 45$  g of alcohol per day) were observed to have a 1.38 times higher chance of developing CAD [181, 182]. Recent MR have demonstrated a causal association between alcohol consumption and CAD. A linear increase in CAD risk with increasing levels of alcohol consumption was reported, suggesting alcohol increases CAD risk even at low consumption levels [183]. The CARDIA study found a significant association between binge-drinking and atherosclerosis in young drinkers (aged 33 to 45 years) [184]. Notably, the largest percentage of binge-drinkers (43.5%) was found among ages 18 to 49 [185].

### 3.9.4 Amphetamines

Amphetamines belong to a drug class of central nervous system stimulants, acting by increasing the amount of dopamine and serotonin in the synaptic space [186, 187]. Additionally, amphetamine has been found to reduce complex IV and cytochrome oxidase transient activity in mitochondria, leading to vascular damage and an increase in ROS levels due to a decrease in antioxidant capacity [188]. Increased dopamine levels can lead to the formation of reactive quinones, which may also increase ROS levels [189]. The use of amphetamine is observed to be independently associated with a 2.74-fold increase in risk

of pCAD [190]. In a group of participants under 30 years old, a 12-fold increase of risk in premature cardiovascular disease was seen in amphetamine users relative to non-users [191].

### **3.9.5 Stress and Exercise**

pCAD has been found to be associated with elevated levels of stress and hostility-related factors. Characterized as any physiological response arising from a physical stressor or heightened emotional reactivity, stress can lead to extended periods of activation of the sympathetic nervous system and depletion of both nervous and immune systems [187, 192, 193, 194, 195, 196, 197, 198, 199]. In a 2002 prospective study observing anger and stress levels in young men, individuals categorized in the highest level of anger and stress were associated with increased pCAD and premature MI [200]. Low physical activity and the increase of childhood obesity have also been linked to pCAD [201]. Generalized obesity is defined as a body mass index (BMI)  $> 30 \text{ kg/ m}^2$  [202]. The worldwide prevalence of childhood obesity has risen 60% from 1990 to 2010 [203]. Childhood obesity can be attributed to a variety of environmental, social or genetic risk factors [204]. Individuals identified with a higher genetic risk of obesity were associated with a 50% greater risk of developing CAD [201, 205, 206, 207, 208, 209]. Additionally, the same cohort was identified with increased levels of artery calcification [210]. Within a group of younger male participants (age 15 to 34 years), an increased amount of atherosclerotic fatty streaks within the right coronary artery was observed in those who were considered clinically obese or had a BMI greater than 30 [211].

### **3.10 Spontaneous Coronary Artery Dissection (SCAD)**

Spontaneous coronary artery dissection (SCAD) is often confused with atherosclerotic pCAD due to its similar manifestations and the predominantly younger age of the affected population [204]. SCAD is characterized by an incidental tear within the coronary artery wall, leading to the formation of a false lumen that compresses the true lumen which drastically increases MI risk [212, 213, 214]. It is predominantly present in younger female populations, with 80% of cases occurring in females between the ages of 43 and 53 years [215]. Up to 35% of MI cases in women under the age of 50 are caused by SCAD [215]. It has been suggested that drastic shifts in hormonal levels can lead to the weakening of arterial walls. Estrogen stimulates the release of nitric oxide (vasodilator) while also increasing HDL levels, consequently reducing plaque formation [216, 217, 218, 219, 220]. During menstruation and pregnancy, high levels of estrogen are present. However, the steep drop after this period has been suggested to induce SCAD risk. Individuals within the typical age range of pregnancy (3336 years) diagnosed with SCAD are associated with a higher risk of MI and ventricular fibrillation, due to the sudden drop in estrogen-induced vasodilatory and hypotensive effects on vascular smooth muscle [216, 220, 221, 222].

### **3.11 Conclusion**

Despite recent progress, there are limitations to our current understanding of pCAD. Up to 20% of pCAD cases are not attributed to conventional atherosclerotic CAD, and remain without a clear identifiable cause [223, 224, 225, 226]. This creates challenges both for personalized therapy and prevention of pCAD through identification of at-risk individuals. Furthermore, genetic studies remain predominantly conducted in European populations, hindering generalizability to non- European populations. Finally, recognition of pCAD

as a unique clinical entity requiring tailored medical history, laboratory investigations and treatments is necessary. Particularly, a detailed multi-generation pedigree is valuable (with accurate phenotyping) in identifying Mendelian inheritance patterns and disease type. Inquiry with focus on age of symptoms onset, drug and alcohol use, social history, and diet/exercise profiles can also aid in risk stratification of pCAD. However, there is a need to gain a better understanding of biological pathways and genetic variants linked to pCAD, otherwise it will be challenging to offer satisfactory preventative approaches.

## References

- [1] GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. *Lancet*, 385(9963):117–171, 2015.
- [2] A. V. Khera, C. A. Emdin, I. Drake, P. Natarajan, and other. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *New England Journal of Medicine*, 375(24):2349-2358, 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1605086.
- [3] D. Malakar, A. K. and Choudhury, B. Halder, P. Paul, et al. A review on coronary artery disease, its risk factors, and therapeutics. *J. Cell. Physiol.*, 234(10):16812–16823, 2019.
- [4] K. Thygesen, J. S. Alpert, A. S. Jaffe, B. R. Chaitman, et al. Fourth Universal Definition of Myocardial Infarction (2018). *Journal of the American College of Cardiology*, 72(18):2231-2264, 2018. ISSN 0735-1097. doi: 10.1016/j.jacc.2018.08.1038.
- [5] K. A. Wilmot, M. O'Flaherty, S. Capewell, E. S. Ford, and V. Vaccarino. Coronary Heart Disease Mortality Declines in the United States From 1979 Through 2011. *Circulation*, 132(11):997-1002, 2015. doi: 10.1161/CIRCULATIONAHA.115.015293.
- [6] Z J Zheng, J B Croft, W H Giles, and G A Mensah. Sudden cardiac death in the united states, 1989 to 1998. *Circulation*, 104(18):2158–2163, 2001.
- [7] H E Ison, S L Clarke, and J W Knowles. Familial hypercholesterolemia. In Margaret P Adam, Ghayda M Mirzaa, Roberta A Pagon, Stephanie E Wallace, Lora

- J H Bean, Karen W Gripp, and Anne Amemiya, editors, *GeneReviews*. University of Washington, Seattle, Seattle (WA), 2014.
- [8] R. N. Pejic. Familial Hypercholesterolemia. *The Ochsner Journal*, 14(4):669672, 2014. ISSN 1524-5012.
- [9] A J Vallejo-Vaz, C A T Stevens, Alexander R M Lyons, Kanika I Dharmayat, et al. Global perspective of familial hypercholesterolaemia: a cross-sectional study from the EAS familial hypercholesterolaemia studies collaboration (FHSC). *Lancet*, 398(10312):1713–1725, 2021.
- [10] R. Albiero and G. Seresini. Atherosclerotic spontaneous coronary artery dissection (A-SCAD) in a patient with COVID-19: case report and possible mechanisms. *Eur Heart J Case Rep*, 4(FI1):1–6, 2020.
- [11] D. Klarin and P. Natarajan. Clinical utility of polygenic risk scores for coronary artery disease. *Nature reviews. Cardiology*, 19(5):291301, 2022. ISSN 1759-5002. doi: 10.1038/s41569-021-00638-w.
- [12] R. Lali, E. Cui, A. Ansarikaleibari, M. Pigeyre, and G. Paré. Genetics of early-onset coronary artery disease: from discovery to clinical translation. *Current Opinion in Cardiology*, 34(6):706713, 2019. ISSN 1531-7080. doi: 10.1097/HCO.0000000000000676.
- [13] D. K. Arnett, R. S. Blumenthal, M. A. Albert, A. B. Buroker, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, 140(11):e596e646, 2019. doi: 10.1161/CIR.0000000000000678.

- [14] S. Zdravkovic, A. Wienke, N. L. Pedersen, M. E. Marenberg, et al. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *Journal of Internal Medicine*, 252(3):247254, 2002. ISSN 1365-2796. doi: 10.1046/j.1365-2796.2002.01029.x.
- [15] P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era concepts and misconceptions. *Nature Reviews Genetics*, 9(44):255266, 2008. ISSN 1471-0064. doi: 10.1038/nrg2322.
- [16] A. V. Khera and S. Kathiresan. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature Reviews Genetics*, 18(66):331344, 2017. ISSN 1471-0064. doi: 10.1038/nrg.2016.160.
- [17] S. Sayols-Baixeras, C. Lluís-Ganella, G. Lucas, and R. Elosua. Pathogenesis of coronary artery disease: focus on genetic risk factors and identification of genetic variants. *The Application of Clinical Genetics*, 7:1532, 2014. ISSN 1178-704X. doi: 10.2147/TACG.S35301.
- [18] M. E. Marenberg, N. Risch, L. F. Berkman, B. Floderus, and U. de Faire. Genetic susceptibility to death from coronary heart disease in a study of twins. *The New England Journal of Medicine*, 330(15):10411046, 1994. ISSN 0028-4793. doi: 10.1056/NEJM199404143301503.
- [19] M. T. Scheuner, W. C. Whitworth, H. McGruder, P. W. Yoon, and M. J. Khoury. Familial risk assessment for early-onset coronary heart disease. *Genetics in Medicine*, 8(8):525531, 2006. ISSN 1098-3600. doi: 10.1097/01.gim.0000232480.00293.00.
- [20] M. T. Scheuner, W. C. Whitworth, H. McGruder, P. W. Yoon, et al. Expanding the

- definition of a positive family history for early-onset coronary heart disease. *Genetics in Medicine*, 8(8). ISSN 1098-3600. doi: 10.1097/01.gim.0000232582.91028.03.
- [21] Genetic Alliance and New York-Mid-Atlantic Consortium for Genetic and Newborn Screening Services. *Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals*. Lulu.com, 2009.
- [22] P. Hu, K. I. Dharmayat, C A T Stevens, M T A Sharabiani, et al. Prevalence of familial hypercholesterolemia among the general population and patients with atherosclerotic cardiovascular disease. *Circulation*, 2020.
- [23] J. Gratton, S. E. Humphries, and M. Futema. Prevalence of FH-Causing variants and impact on LDL-C concentration in european, south asian, and african ancestry groups of the UK Biobank-Brief report. *Arterioscler. Thromb. Vasc. Biol.*, 43(9): 1737–1742, 2023.
- [24] F. Toft-Nielsen, F. Emanuelsson, and M. Benn. Familial hypercholesterolemia prevalence among Ethnicities-Systematic review and Meta-Analysis. *Front. Genet.*, 13: 840797, 2022.
- [25] F. Alnouri, F. A Al-Allaf, M. Athar, Z. Abduljaleel, et al. Xanthomas can be misdiagnosed and mistreated in homozygous familial hypercholesterolemia patients: A call for increased awareness among dermatologists and health care practitioners. *Glob. Heart*, 15(1), 2020.
- [26] F Civeira, S Castillo, R Alonso, et al. Tendon xanthomas in familial hypercholesterolemia are associated with cardiovascular risk independently of the Low-Density lipoprotein receptor gene mutation. *Arterioscler. Thromb. Vasc. Biol.*, 2005.

- [27] A. Bell and A. P. Shreenath. Xanthoma. In *StatPearls [Internet]*. StatPearls Publishing, 2022.
- [28] M. Abifadel, M. Varret, J. P. Rabès, D. Allard, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature Genetics*, 34(22):154156, 2003. ISSN 1546-1718. doi: 10.1038/ng1161.
- [29] M Sharifi, M Futema, D Nair, and S E Humphries. Genetic architecture of familial hypercholesterolaemia. *Curr. Cardiol. Rep.*, 19(5):44, 2017.
- [30] C. K. Garcia, K. Wilund, M. Arca, G. Zuliani, et al. Autosomal recessive hypercholesterolemia caused by mutations in a putative ldl receptor adaptor protein. *Science (New York, N.Y.)*, 292(5520):13941398, May 2001. ISSN 0036-8075. doi: 10.1126/science.1060458.
- [31] K E Berge, H Tian, G A Graf, L Yu, N V Grishin, J Schultz, P Kwiterovich, B Shan, R Barnes, and H H Hobbs. Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. *Science*, 290(5497):1771–1775, 2000.
- [32] N S Abul-Husn, K Manickam, L K Jones, E A Wright, et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science*, 2016.
- [33] Željko Reiner. Hypertriglyceridaemia and risk of coronary artery disease. *Nat. Rev. Cardiol.*, 14(7):401–411, 2017.
- [34] L. Berglund, J. D. Brunzell, A. C. Goldberg, I. J. Goldberg, et al. Evaluation and treatment of hypertriglyceridemia: An endocrine society clinical practice guideline. *J. Clin. Endocrinol. Metab.*, 97(9):2969–2989, 2012.

- [35] S B Hulley, R H Rosenman, R D Bawol, and R J Brand. Epidemiology as a guide to clinical decisions. the association between triglyceride and coronary heart disease. *N. Engl. J. Med.*, 302(25):1383–1389, 1980.
- [36] Silv Santamarina-Fojo. The familial hyperchylomicronemia syndrome. *JAMA*, 265(7):904, 1991.
- [37] A M Gotto, Jr. Triglyceride as a risk factor for coronary artery disease. *Am. J. Cardiol.*, 82(9A):22Q–25Q, 1998.
- [38] A. D. Pradhan, R. J. Glynn, J. Fruchart, et al. Triglyceride lowering with pemafibrate to reduce cardiovascular risk. *N. Engl. J. Med.*, 2022.
- [39] The AIM-HIGH Investigators. Niacin in patients with low HDL cholesterol levels receiving intensive statin therapy. 2011.
- [40] H N Ginsberg, M B Elam, L C Lovato, J R Crouse, et al. Effects of combination lipid therapy in type 2 diabetes mellitus. *N. Engl. J. Med.*, 362(17), 2010.
- [41] The HPS2-THRIVE Collaborative Group. Effects of Extended-Release Niacin with Laropiprant in High-Risk Patients. 2014.
- [42] Kiran Musunuru. Enduring Mystery of the Chromosome 9p21.3 Locus. *Circulation: Cardiovascular Genetics*, 6(2):224–225, 2013. doi: 10.1161/CIRCGENETICS.113.000132.
- [43] S. Burgess and S. G. Thompson. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. CRC Press, 2015.
- [44] M Thomsen, A Varbo, A Tybjaerg-Hansen, and B G Nordestgaard. Low nonfasting

- triglycerides and reduced all-cause mortality: a mendelian randomization study. *Clin. Chem.*, 60(5):737–746, 2014.
- [45] M V Holmes, F W Asselbergs, T M Palmer, F Drenos, et al. Mendelian randomization of blood lipids for coronary heart disease. *Eur. Heart J.*, 36(9):539–550, 2015.
- [46] R. Do, N. O. Stitzel, H. Won, A. B. Jørgensen, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*, 518(7537):102–106, 2015.
- [47] Anders Berg Jørgensen, Ruth Frikke-Schmidt, Børge G. Nordestgaard, and Anne Tybjærg-Hansen. Loss-of-function mutations in APOC3 and risk of ischemic vascular disease. *The New England Journal of Medicine*, 371(1):3241, 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1308027.
- [48] F. E. Dewey, Vi. Gusarova, C. ODushlaine, O. Gottesman, et al. Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *New England Journal of Medicine*, 374(12):1123–1133, 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1510926.
- [49] J R Burnett, A J Hooper, and R A Hegele. Familial lipoprotein lipase deficiency. In *GeneReviews® [Internet]*. University of Washington, Seattle, 2017.
- [50] D. Gaudet, V. J. Alexander, B. F. Baker, D. Brisson, et al. Antisense inhibition of apolipoprotein C-III in patients with hypertriglyceridemia. *N. Engl. J. Med.*, 373(5):438–447, 2015.
- [51] R A Hegele. APOC3 interference for familial chylomicronaemia syndrome. *touchREV Endocrinol*, 18(2):82–83, 2022.
- [52] R. Pechlaner, S. Tsimikas, X. Yin, P. Willeit, et al. Very-Low-Density Lipoprotein-

- Associated apolipoproteins predict cardiovascular events and are lowered by inhibition of APOC-III. *J. Am. Coll. Cardiol.*, 69(7):789–800, 2017.
- [53] P N Hopkins, L L Wu, S C Hunt, and E A Brinton. Plasma triglycerides and type III hyperlipidemia are independently associated with premature familial coronary artery disease. *J. Am. Coll. Cardiol.*, 45(7):1003–1012, 2005.
- [54] A. M. Bennet, E. Di Angelantonio, Z. Ye, F. Wensley, et al. Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA*, 298(11):1300–1311, 2007.
- [55] C. Koopal, A David M., J. Westerink, and F. L. J. Visseren. Autosomal dominant familial dysbetalipoproteinemia: A pathophysiological framework and practical approach to diagnosis and therapy. *J. Clin. Lipidol.*, 11(1):12–23.e1, 2017.
- [56] Zannis V. Genetic polymorphism in human apolipoprotein E. In *Methods in Enzymology*, volume 128, pages 823–851. Academic Press, 1986.
- [57] R W Mahley. Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. *Science*, 240(4852):622–630, 1988.
- [58] Mahley RW. Pathogenesis of type III hyperlipoproteinemia (dysbetalipoproteinemia): questions, quandaries, and paradoxes. *J. Lipid Res.*, 40(11):1933–1949, 1999.
- [59] N. Roy, D. Gaudet, and D. Brisson. Palmar Striated Xanthomas in Clinical Practice. *J Endocr Soc*, 6(8):bvac103, 2022.
- [60] C Koopal, K Retterstøl, B Sjouke, G K Hovingh, E Ros, et al. Vascular risk factors, vascular disease, lipids and lipid targets in patients with familial dysbetalipoproteinemia: a european cross-sectional study. *Atherosclerosis*, 240(1), 2015.

- [61] F. de Beer, A. F. H. Stalenhoef, N. Hoogerbrugge, J. J. P. Kastelein, et al. Expression of type III hyperlipoproteinemia in apolipoprotein E2 (arg158  $\rightarrow$  cys) homozygotes is associated with hyperinsulinemia. *Arterioscler. Thromb. Vasc. Biol.*, 22(2):294–299, 2002.
- [62] S Villeneuve, D Brisson, N L Marchant, and D Gaudet. The potential applications of apolipoprotein E in personalized medicine. *Front. Aging Neurosci.*, 6:154, 2014.
- [63] S. B. Myrie, R. D. Steiner, and D. Mymin. *Sitosterolemia*. University of Washington, Seattle, 2020.
- [64] Hayato Tada, Akihiro Nomura, Masatsune Ogura, Katsunori Ikewaki, Yasushi Ishigaki, Kyoko Inagaki, Kazuhisa Tsukamoto, Kazushige Dobashi, Kimitoshi Nakamura, Mika Hori, Kota Matsuki, Shizuya Yamashita, Shinji Yokoyama, Masa-Aki Kawashiri, and Mariko Harada-Shiba. Diagnosis and management of sitosterolemia 2021. *J. Atheroscler. Thromb.*, 28(8):791–801, 2021.
- [65] J. M. Bastida, M. L. Girós, R. Benito, et al. Sitosterolemia: Diagnosis, metabolic and hematological abnormalities, cardiovascular disease and management. *Curr. Med. Chem.*, 26(37):6766–6775, 2019.
- [66] X Pang, J Liu, J Zhao, J Mao, et al. Homocysteine induces the expression of c-reactive protein via NMDAR-ROS-MAPK-NF- $\kappa$ B signal pathway in rat vascular smooth muscle cells. *Atherosclerosis*, 236(1):73–81, 2014.
- [67] V. Shenoy, V. Mehendale, K. Prabhu, R. Shetty, and P. Rao. Correlation of serum homocysteine levels with the severity of coronary artery disease. *Indian J. Clin. Biochem.*, 29(3):339–344, 2014.

- [68] P Ganguly and Sreyoshi F Alam. Role of homocysteine in the development of cardiovascular disease. *Nutr. J.*, 14:6, 2015.
- [69] S. Zhang, Y. Bai, L. Luo, W. Xiao, et al. Association between serum homocysteine and arterial stiffness in elderly: a community-based study. *J. Geriatr. Cardiol.*, 11(1):32–38, 2014.
- [70] S. Moll and E. A Varga. Homocysteine and MTHFR mutations. *Circulation*, 132(1):e6–9, 2015.
- [71] M Ebbing, O Bleie, P M Ueland, J E Nordrehaug, et al. Mortality and cardiovascular events in patients treated with homocysteine-lowering B vitamins after coronary angiography: a randomized controlled trial. *JAMA*, 300(7):795–804, 2008.
- [72] C. Antoniades, A. S. Antonopoulos, D. Tousoulis, K. Marinou, and C. Stefanadis. Homocysteine and coronary atherosclerosis: from folate fortification to the recent clinical trials. *Eur. Heart J.*, 30(1):6–15, 2009.
- [73] K. H. Bønaa, I. Njølstad, P. M. Ueland, et al. Homocysteine lowering and cardiovascular events after acute myocardial infarction. *N. Engl. J. Med.*, 354(15):1578–1588, 2006.
- [74] E Lonn, S Yusuf, M J Arnold, P Sheridan, et al. Homocysteine lowering with folic acid and B vitamins in vascular disease. *N. Engl. J. Med.*, 354(15):1567–1577, 2006.
- [75] J F Toole, M R Malinow, L E Chambless, et al. Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the vitamin intervention for stroke prevention (VISP) randomized controlled trial. *JAMA*, 291(5):565–575, 2004.

- [76] A S Shah and D P Wilson. *Genetic Disorders Causing Hypertriglyceridemia in Children and Adolescents*. MDText.com, Inc., 2023.
- [77] A Goyal, A S Cusick, and E Reilly. *Familial Hypertriglyceridemia*. StatPearls Publishing, 2023.
- [78] A. Alshaikhli and S. Vaqar. *Tangier Disease*. StatPearls Publishing, 2023.
- [79] J R Burnett, A J Hooper, S P A McCormick, and R A Hegele. *Tangier Disease*. University of Washington, Seattle, 2019.
- [80] D. Faeh, A. Chiolero, and F. Paccaud. Homocysteine as a risk factor for cardiovascular disease: should we (still) worry about? *Swiss Med. Wkly*, 136(47-48):745–756, 2006.
- [81] A. F. Alrefaei and M. Abu-Elmagd. LRP6 receptor plays essential functions in development and human diseases. *Genes*, 13(1):120, 2022.
- [82] R Singh, E Smith, M Fathzadeh, W Liu, et al. Rare nonconservative LRP6 mutations are associated with metabolic syndrome. *Hum. Mutat.*, 34(9):1221–1225, 2013.
- [83] A Mani, J Radhakrishnan, H Wang, A Mani, et al. LRP6 mutation in a family with early coronary disease and metabolic risk factors. *Science*, 315(5816):1278–1282, 2007.
- [84] Y Xu, W Gong, J Peng, H Wang, et al. Functional analysis LRP6 novel mutations in patients with coronary artery disease. *PLoS One*, 9(1):e84345, 2014.
- [85] R Srivastava, J Zhang, G Go, A Narayanan, et al. Impaired LRP6-TCF7L2 activity enhances smooth muscle cell plasticity and causes coronary artery disease. *Cell Rep.*, 13(4):746–759, 2015.

- [86] Colleen A Morris. Williams syndrome. In Margaret P Adam, Ghayda M Mirzaa, Roberta A Pagon, Stephanie E Wallace, Lora J H Bean, Karen W Gripp, and Anne Amemiya, editors, *GeneReviews®*. University of Washington, Seattle, Seattle (WA), 1999.
- [87] G H Dadlani, C Mercado, V Roberts, H Blackwelder, et al. Cardiovascular screening in williams syndrome. *Prog. Pediatr. Cardiol.*, 58:101267, 2020.
- [88] C. K. Sickles and G. P. Gross. *Progeria*. StatPearls Publishing, 2022.
- [89] L B Gordon, W Ted Brown, and F S Collins. *Hutchinson-Gilford Progeria Syndrome*. University of Washington, Seattle, 2019.
- [90] M. Olive, I. Harten, R. Mitchell, J. K. Beers, et al. Cardiovascular pathology in Hutchinson-Gilford progeria: correlation with the vascular pathology of aging. *Arterioscler. Thromb. Vasc. Biol.*, 30(11):2301–2309, 2010.
- [91] D. P. Germain. Pseudoxanthoma elasticum. *Orphanet J. Rare Dis.*, 12(1):1–13, 2017.
- [92] Diana N. Vikulova, Mark Trinder, G.B. John Mancini, Simon N. Pimstone, and Liam R. Brunham. Familial Hypercholesterolemia, Familial Combined Hyperlipidemia, and Elevated Lipoprotein(a) in Patients with Premature Coronary Artery Disease. *Canadian Journal of Cardiology*, 37(11):17331742, 2021. ISSN 0828282X. doi: 10.1016/j.cjca.2021.08.012.
- [93] S. Theriault, R. Lali, M. Chong, J. L. Velianou, M. K. Natarajan, and G. Paré. Polygenic Contribution in Individuals With Early-Onset Coronary Artery Disease. *Circ Genom Precis Med*, 11(1):e001849, 2018. doi: 10.1161/CIRCGEN.117.001849.
- [94] V. Tam, N. Patel, M. Turcotte, Y. Bosse, et al. Benefits and limitations of genome-

- wide association studies. *Nature Reviews Genetics*, 20(8):467485, 2019. ISSN 14710056. doi: 10.1038/s41576-019-0127-1.
- [95] O. Zuk, S. F. Schaffner, K. Samocha, et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.*, 111(4):E455–64, 2014.
- [96] K. G. Aragam, T. Jiang, A. Goel, S. Kanoni, et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nature Genetics*, 54(1212):18031815, 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01233-6.
- [97] H Matsunaga, K Ito, M Akiyama, A Takahashi, et al. Transethnic Meta-Analysis of Genome-Wide association studies identifies three new loci and characterizes Population-Specific differences for coronary artery disease. *Circ Genom Precis Med*, 13(3):e002670, 2020.
- [98] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, et al. Genomewide association analysis of coronary artery disease. *The New England Journal of Medicine*, 357(5):443453, 2007. ISSN 1533-4406. doi: 10.1056/NEJMoa072366.
- [99] A. Helgadottir, G. Thorleifsson, A. Manolescu, S. Gretarsdottir, et al. A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science*, 316(5830):14911493, 2007. doi: 10.1126/science.1142842.
- [100] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, et al. A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science (New York, N. Y.)*, 316(5830):14881491, 2007. ISSN 0036-8075. doi: 10.1126/science.1142447.

- [101] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.
- [102] N. A. Almontashiri. The 9p21.3 risk locus for coronary artery disease: A 10-year search for its mechanism. *Journal of Taibah University Medical Sciences*, 12(3):199204, 2017. ISSN 16583612. doi: 10.1016/j.jtumed.2017.03.001.
- [103] R Ross. Atherosclerosis—an inflammatory disease. *N. Engl. J. Med.*, 340(2):115–126, 1999.
- [104] J. Zhuang, W. Peng, H. Li, W. Wang, et al. Methylation of p15INK4b and Expression of ANRIL on Chromosome 9p21 Are Associated with Coronary Artery Disease. *PLoS ONE*, 7(10):e47193, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0047193.
- [105] O Jarinova, A F Stewart, R Roberts, G Wells, P Lau, T Naing, C Buerki, B W McLean, et al. Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. *Arterioscler. Thromb. Vasc. Biol.*, 29(10), 2009.
- [106] A Motterle, X Pu, H Wood, Q Xiao, S Gor, et al. Functional analyses of coronary artery disease associated variation on chromosome 9p21 in vascular smooth muscle cells. *Hum. Mol. Genet.*, 21(18), 2012.
- [107] F. Privé, J. Arbel, H. Aschard, and B. J. Vilhjálmsson. Identifying and correcting for misspecifications in gwas summary statistics and polygenic scores. *Human Genetics and Genomics Advances*, 3(4):100136, 2022. ISSN 2666-2477. doi: 10.1016/j.xhgg.2022.100136.
- [108] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. LDpred2: better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431, 2020.

- [109] A. P. Patel, M. Wang, Y. Ruan, S. Koyama, et al. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.*, 29(7):1793–1803, 2023.
- [110] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(99):12191224, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0183-z.
- [111] A. C. Fahed, M. Wang, J. R. Homburger, A. P. Patel, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications*, 11(11):3635, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17374-3.
- [112] N. Mars, J. T. Koskela, P. Ripatti, T. T. J. Kiiskinen, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine*, 26(44):549557, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0800-0.
- [113] M. Inouye, G. Abraham, C. P. Nelson, A. M. Wood, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *Journal of the American College of Cardiology*, 72(16):18831893, 2018. ISSN 0735-1097. doi: 10.1016/j.jacc.2018.07.079.
- [114] P. Natarajan, R. Young, N. O. Stitzel, S. Padmanabhan, et al. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation*, 135(22):20912101, 2017. doi: 10.1161/CIRCULATIONAHA.116.024436.
- [115] J. L. Mega, N. O. Stitzel, J. G. Smith, D. I. Chasman, et al. Genetic risk, coronary

- heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet*, 385(9984):22642271, 2015. ISSN 01406736. doi: 10.1016/S0140-6736(14)61730-X.
- [116] C. Sudlow, J. Gallacher, N. Allen, V. Beral, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):e1001779, 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779.
- [117] M. D. Mesbah Uddin, N. Q. H. Nguyen, B. Yu, J. A. Brody, et al. Clonal hematopoiesis of indeterminate potential, dna methylation, and risk for coronary artery disease. *Nature Communications*, 13(1):5350, September 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33093-3.
- [118] S. Jaiswal, P. Natarajan, A. J. Silver, J. Gibson, C, et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *The New England Journal of Medicine*, 377(2):111121, July 2017. ISSN 1533-4406. doi: 10.1056/NEJMoa1701719.
- [119] A. Stein, K. Metzeler, A. S. Kubasch, K. P. Rommel, et al. Clonal hematopoiesis and cardiovascular disease: deciphering interconnections. *Basic Research in Cardiology*, 117(1):55, 2022. ISSN 0300-8428. doi: 10.1007/s00395-022-00969-w.
- [120] G. A. Challen and M. A. Goodell. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. *Blood*, 136(14):15901598, October 2020. ISSN 1528-0020. doi: 10.1182/blood.2020006510.
- [121] E. Mayerhofer, C. Strecker, H. Becker, M. K. Georgakis, et al. Prevalence and therapeutic implications of clonal hematopoiesis of indeterminate potential in young

- patients with stroke. *Stroke*, 54(4):938946, April 2023. ISSN 1524-4628. doi: 10.1161/STROKEAHA.122.041416.
- [122] I. Cobo, T. Tanaka, C. K. Glass, and C. Yeang. Clonal hematopoiesis driven by *dnmt3a* and *tet2* mutations: role in monocyte and macrophage biology and atherosclerotic cardiovascular disease. *Current Opinion in Hematology*, 29(1):17, January 2022. ISSN 1531-7048. doi: 10.1097/MOH.0000000000000688.
- [123] C. S. Marnell, A. Bick, and P. Natarajan. Clonal hematopoiesis of indeterminate potential (chip): Linking somatic mutations, hematopoiesis, chronic inflammation and cardiovascular disease. *Journal of Molecular and Cellular Cardiology*, 161:98105, December 2021. ISSN 1095-8584. doi: 10.1016/j.yjmcc.2021.07.004.
- [124] E. D. Gumuser, A. Schuermans, J. Cho, Z. A. Sporn, et al. Clonal hematopoiesis of indeterminate potential predicts adverse outcomes in patients with atherosclerotic cardiovascular disease. *Journal of the American College of Cardiology*, 81(20):19962009, May 2023. ISSN 1558-3597. doi: 10.1016/j.jacc.2023.03.401.
- [125] S. P. Kar, P. M. Quiros, M. Gu, T. Jiang, et al. Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nature Genetics*, 54(8):11551166, August 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01121-z.
- [126] M. Nikpay, A. Goel, H. Won, L. M. Hall, et al. A comprehensive 1000 genomesbased genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):11211130, October 2015. ISSN 1546-1718. doi: 10.1038/ng.3396.
- [127] J H Cole, J I Miller, 3rd, L S Sperling, and W S Weintraub. Long-term follow-up

- of coronary artery disease presenting in young adults. *J. Am. Coll. Cardiol.*, 41(4):521–528, 2003.
- [128] E L Navas-Nacher, L Colangelo, C Beam, and P Greenland. Risk factors for coronary heart disease in men 18 to 39 years of age. *Ann. Intern. Med.*, 134(6):433–439, 2001.
- [129] J R Petrie, T J Guzik, and R M Touyz. Diabetes, hypertension, and cardiovascular disease: Clinical insights and vascular mechanisms. *Can. J. Cardiol.*, 34(5):575–584, 2018.
- [130] E. Jorge-Galarza, F. D. Martínez-Sánchez, C. I. Javier-Montiel, A. X. Medina-Urrutia, et al. Control of blood pressure levels in patients with premature coronary artery disease: Results from the genetics of atherosclerotic disease study. *J. Clin. Hypertens.*, 22(7):1253–1262, 2020.
- [131] H. Poorzand, K. Tsarouhas, S. A. Hozhabrossadati, N. Khorrampazhouh, et al. Risk factors of premature coronary artery disease in iran: A systematic review and meta-analysis. *Eur. J. Clin. Invest.*, 49(7):e13124, 2019.
- [132] K Malmberg, P Båvenholm, and A Hamsten. Clinical and biochemical factors associated with prognosis after myocardial infarction at a young age. *J. Am. Coll. Cardiol.*, 24(3):592–599, 1994.
- [133] D. M Maahs, S. R Daniels, S. D de Ferranti, H. L. Dichek, et al. Cardiovascular disease risk factors in youth with diabetes mellitus: a scientific statement from the american heart association. *Circulation*, 130(17):1532–1558, 2014.
- [134] R. B. DAgostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6):743–753, 2008. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.107.699579.

- [135] C. Koz, O. Baysan, A. Hasimi, M. Cihan, et al. Conventional and non-conventional coronary risk factors in male premature coronary artery disease patients already having a low Framingham risk score. *Acta Cardiologica*, 63(5):623628, 2008. ISSN 0001-5385. doi: 10.2143/AC.63.5.2033231.
- [136] D. H. Becker and L. B. Gardner. *Prevention in Clinical Practice*. Springer Science Business Media, 2012. ISBN 978-1-4684-5356-0. Google-Books-ID: EVRD-BAAAQBAJ.
- [137] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA*, 297(6):611619, 2007. ISSN 1538-3598. doi: 10.1001/jama.297.6.611.
- [138] S. Yusuf, S. Rangarajan, K. Teo, S. Islam, et al. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. *The New England Journal of Medicine*, 371(9):818827, 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1311890.
- [139] M. Woodward, P. Brindle, H. Tunstall-Pedoe, and SIGN group on risk estimation. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart (British Cardiac Society)*, 93(2):172176, 2007. ISSN 1468-201X. doi: 10.1136/hrt.2006.108167.
- [140] J. Hippisley-Cox, C. Coupland, and P. Brindle. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*, 357:j2099, 2017. ISSN 1756-1833. doi: 10.1136/bmj.j2099.

- [141] G. Assmann, P. Cullen, and H. Schulte. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. *Circulation*, 105(3):310315, 2002. ISSN 1524-4539. doi: 10.1161/hc0302.102575.
- [142] L. Palmieri, R. Rielli, L. Demattè, C. Donfrancesco, et al. CUORE project: implementation of the 10-year risk score. *European Journal of Cardiovascular Prevention Rehabilitation*, 18(4):642649, 2011. ISSN 1741-8267. doi: 10.1177/1741826710389925.
- [143] A. Sofogianni, N. Stalikas, C. Antza, and K. Tziomalos. Cardiovascular risk prediction models and scores in the era of personalized medicine. *J Pers Med*, 12(7), 2022.
- [144] A P DeFilippis, R Young, C J Carrubba, et al. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Ann. Intern. Med.*, 162(4):266–275, 2015.
- [145] M Zeitouni, R M Clare, K Chiswell, J Abdulrahim, et al. Risk factor burden and LongTerm prognosis of patients with premature coronary artery disease. *J. Am. Heart Assoc.*, 2020.
- [146] S. Sivapalaratnam, S.M. Boekholdt, M.D. Trip, M.S. Sandhu, et al. Family history of premature coronary heart disease and risk prediction in the EPIC-Norfolk prospective population study. *Heart (British Cardiac Society)*, 96(24):19851989, 2010. ISSN 1355-6037. doi: 10.1136/hrt.2010.210740.
- [147] S. Bastuji-Garin, A. Deverly, D. Moyse, A. Castaigne, et al. The framingham prediction rule is not valid in a european population of treated hypertensive patients. *J. Hypertens.*, 20(10):1973–1980, 2002.

- [148] P. Brindle, J. Emberson, F. Lampe, M. Walker, et al. Predictive accuracy of the framingham coronary risk score in british men: prospective cohort study. *BMJ*, 327 (7426):1267, 2003.
- [149] R J Khan, C P Stewart, S K Davis, D J Harvey, et al. The risk and burden of smoking related heart disease mortality among young people in the united states. *Tob. Induc. Dis.*, 13(1):16, 2015.
- [150] S. Sadeghian, P. Graili, M. Salarifar, A. A. Karimi, et al. Opium consumption in men and diabetes mellitus in women are the most important risk factors of premature coronary artery disease in iran. *Int. J. Cardiol.*, 141(1):116–118, 2010.
- [151] A W Caliri, S Tommasi, and A Besaratinia. Relationships among smoking, oxidative stress, inflammation, macromolecular damage, and cancer. *Mutat. Res. - Rev. Mut. Res.*, 787:108365, 2021.
- [152] M A Incalza, R D’Oria, A Natalicchio, S Perrini, et al. Oxidative stress and reactive oxygen species in endothelial dysfunction associated with cardiovascular and metabolic diseases. *Vascul. Pharmacol.*, 100:1–19, 2018.
- [153] K. M. Patel, A. Strong, J. Tohyama, X. Jin, et al. Macrophage sortilin promotes LDL uptake, foam cell formation, and atherosclerosis. *Circ. Res.*, 116(5):789–796, 2015.
- [154] R. Sun, D. Mendez, and K. E. Warner. Trends in nicotine product use among US adolescents, 1999-2020. *JAMA Netw Open*, 4(8):e2118788, 2021.
- [155] L M Dutra and S A Glantz. E-cigarettes and national adolescent cigarette use: 2004–2014. *Pediatrics*, 139(2), 2017.

- [156] M B Harrell, S R Weaver, A Loukas, M Creamer, et al. Flavored e-cigarette use: Characterizing youth, young adult, and adult users. *Prev Med Rep*, 5:33–40, 2017.
- [157] T. W. Wang, L. J. Neff, E. Park-Lee, C. Ren, et al. E-cigarette use among middle and high school students — united states, 2020. *MMWR Surveill. Summ.*, 69(37):1310, 2020.
- [158] A M Leventhal, D R Strong, Matthew G Kirkpatrick, J B Unger, et al. Association of electronic cigarette use with initiation of combustible tobacco product smoking in early adolescence. *JAMA*, 314(7):700, 2015.
- [159] T Martinelli, M J J M Candel, H de Vries, R Talhout, et al. Exploring the gateway hypothesis of e-cigarettes and tobacco: a prospective replication study among adolescents in the netherlands and flanders. *Tob. Control*, 32(2):170–178, 2023.
- [160] A. Matsuzawa, Y. and Lerman. Endothelial dysfunction and coronary artery disease: assessment, prognosis, and treatment. *Coron. Artery Dis.*, 25(8):713–724, 2014.
- [161] M. Kuntic, M. Oelze, S. Steven, S. Kröller-Schön, et al. Short-term e-cigarette vapour exposure causes vascular oxidative stress and dysfunction: evidence for a close connection to brain damage and a key role of the phagocytic nadph oxidase (nox-2). *European Heart Journal*, 41(26):24722483, July 2020. ISSN 1522-9645. doi: 10.1093/eurheartj/ehz772.
- [162] P. von Hundelshausen, K. S. C. Weber, Y. Huo, et al. Rantes deposition by platelets triggers monocyte arrest on inflamed and atherosclerotic endothelium. *Circulation*, 103(13):17721777, April 2001. doi: 10.1161/01.CIR.103.13.1772.
- [163] L Badimon, J J Badimon, W Penny, M W Webster, et al. Endothelium and atherosclerosis. *J. Hypertens. Suppl.*, 10(2):S43–50, 1992.

- [164] L E Wold, R Tarran, Crotty A, et al. Cardiopulmonary consequences of vaping in adolescents: A scientific statement from the american heart association. *Circ. Res.*, 131(3):e70–e82, 2022.
- [165] A. M. Dydyk, N. K. Jain, and M. Gupta. Opioid use disorder. In *StatPearls [Internet]*. StatPearls Publishing, 2023.
- [166] H Y Chang, H Kharrazi, D Bodycombe, J P Weiner, and G C Alexander. Healthcare costs and utilization associated with high-risk prescription opioid use: a retrospective cohort study. *BMC Med.*, 16(1), 2018.
- [167] R. Doshi, M. Majmundar, T. Kansara, R. Desai, et al. Frequency of cardiovascular events and in-hospital mortality with opioid overdose hospitalizations. *Am. J. Cardiol.*, 124(10):1528–1533, 2019.
- [168] M J Krantz, R B Palmer, and M C P Haigney. Cardiovascular complications of opioid use: JACC state-of-the-art review. *J. Am. Coll. Cardiol.*, 77(2):205–223, 2021.
- [169] S. Nakhaee, S. Ghasemi, K. Karimzadeh, N. Zamani, et al. The effects of opium on the cardiovascular system: a review of side effects, uses, and potential mechanisms. *Substance Abuse Treatment, Prevention, and Policy*, 15:30, April 2020. ISSN 1747-597X. doi: 10.1186/s13011-020-00272-8.
- [170] S. Asgary, N. Sarrafzadegan, G. Naderi, and R. Rozbehani. Effect of opium addiction on new and traditional cardiovascular risk factors: do duration of addiction and route of administration matter? *Lipids in Health and Disease*, 7:42, November 2008. ISSN 1476-511X. doi: 10.1186/1476-511X-7-42.
- [171] M. Ziaee, R. Hajizadeh, A. Khorammi, N. Sephervand, S. Momtaz, et al. Cardio-

- vascular complications of chronic opium consumption: A narrative review article. *Iranian Journal of Public Health*, 48(12):21542164, December 2019. ISSN 2251-6085.
- [172] D M Dick and A Agrawal. The genetics of alcohol and other drug dependence. *Alcohol Res. Health*, 31(2), 2008.
- [173] J M Rosenbloom, S M Burns, E Kim, D A August, V E Ortiz, and T T Houle. Race/Ethnicity and sex and opioid administration in the emergency room. *Anesth. Analg.*, 128(5), 2019.
- [174] S. Momtazi and R. Rawson. Substance abuse among Iranian high school students. *Curr. Opin. Psychiatry*, 23(3):221–226, 2010.
- [175] A. Maino, S. Sadeghian, I. Mancini, S. H. Abbasi, et al. Opium as a risk factor for early-onset coronary artery disease: Results from the Milano-Iran (MIran) study. *PLoS One*, 18(4):e0283707, 2023.
- [176] M. Marmor, A. Penn, K. Widmer, R. I Levin, and R. Maslansky. Coronary artery disease and opioid use. *Am. J. Cardiol.*, 93(10):1295–1297, 2004.
- [177] M R Piano. Alcohol’s effects on the cardiovascular system. *Alcohol Res.*, 38(2): 219–241, 2017.
- [178] L. A. Quigley and G. A. Marlatt. Drinking among young adults: Prevalence, patterns, and consequences. *Alcohol Health and Research World*, 20(3):185191, 1996. ISSN 0090-838X.
- [179] W B Kannel and R C Ellison. Alcohol and coronary heart disease: the evidence for a protective effect. *Clin. Chim. Acta*, 246(1-2):59–76, 1996.

- [180] R. D. Langer, M. H. Criqui, and D. M. Reed. Lipoproteins and blood pressure as biological pathways for effect of moderate alcohol consumption on coronary heart disease. *Circulation*, 85(3):910915, March 1992. ISSN 0009-7322. doi: 10.1161/01.cir.85.3.910.
- [181] K J Biddinger, C A Emdin, M E Haas, M Wang, et al. Association of habitual alcohol intake with risk of cardiovascular disease. *JAMA Netw Open*, 5(3):e223849, 2022.
- [182] S C Larsson, S Burgess, A M Mason, and K Michaëlsson. Alcohol consumption and cardiovascular disease: A mendelian randomization study. *Circ Genom Precis Med*, 13(3):e002814, 2020.
- [183] C Hu, C Huang, J Li, F Liu, et al. Causal associations of alcohol consumption with cardiovascular diseases and all-cause mortality among chinese males. *Am. J. Clin. Nutr.*, 116(3):771–779, 2022.
- [184] M. J. Pletcher, P. Varosy, C.I. Kiefe, C. E. Lewis, et al. Alcohol consumption, binge drinking, and early coronary calcification: findings from the coronary artery risk development in young adults (CARDIA) study. *Am. J. Epidemiol.*, 161(5):423–433, 2005.
- [185] M. M. Englund, B. Egeland, E. M. Oliva, and W. A. Collins. Childhood and adolescent predictors of heavy drinking and alcohol use disorders in early adulthood: a longitudinal developmental analysis. *Addiction (Abingdon, England)*, 103(Suppl 1): 23, May 2008. ISSN 0965-2140. doi: 10.1111/j.1360-0443.2008.02174.x.
- [186] N. S. Shoar, R. Marwaha, and M. Molla. Dextroamphetamine-Amphetamine. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

- [187] A. Schrantee, L. Václav, D. F. R. Heijtel, M. W. A. Caan, et al. Dopaminergic system dysfunction in recreational dexamphetamine users. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 40(5):11721180, March 2015. ISSN 1740-634X. doi: 10.1038/npp.2014.301.
- [188] R. Potula, B. J. Hawkins, J. M. Cenna, S. Fan, et al. Methamphetamine causes mitochondrial oxidative damage in human T lymphocytes leading to functional impairment. *J. Immunol.*, 185(5):2867–2876, 2010.
- [189] D G Graham, S M Tiffany, W R Bell, Jr, and W F Gutknecht. Autoxidation versus covalent binding of quinones as the mechanism of toxicity of dopamine, 6-hydroxydopamine, and related compounds toward C1300 neuroblastoma cells in vitro. *Mol. Pharmacol.*, 14(4):644–653, 1978.
- [190] D Mahtta, D Ramsey, C Krittanawong, Al R, et al. Recreational substance use among patients with premature atherosclerotic cardiovascular disease. *Heart*, 107(8):650–656, 2021.
- [191] V Batra, K. S. Murnane, B. Knox, A. N. Edinoff, et al. Early onset cardiovascular disease related to methamphetamine use is most striking in individuals under 30: A retrospective chart review. *Addict Behav Rep*, 15:100435, 2022.
- [192] N Schneiderman, G Ironson, and S D Siegel. Stress and health: psychological, behavioral, and biological determinants. *Annu. Rev. Clin. Psychol.*, 1(1):607–628, 2005.
- [193] M. Y. Henein, S. Vancheri, G. Longo, and F. Vancheri. The impact of mental stress on cardiovascular Health-Part II. *J. Clin. Med. Res.*, 11(15), 2022.
- [194] B. Sadeghi, H. Mashalchi, S. Eghbali, M. Jamshidi, et al. The relationship between

- hostility and anger with coronary heart disease in patients. *J. Educ. Health Promot.*, 9:223, 2020.
- [195] F. Satyjeet, S. Naz, V. Kumar, N. H. Aung, et al. Psychological stress as a risk factor for cardiovascular disease: A Case-Control study. *Cureus*, 12(10):e10757, 2020.
- [196] B. Yao, L. Meng, M. Hao, Y. Zhang, et al. Chronic stress: a critical risk factor for atherosclerosis. *J. Int. Med. Res.*, 47(4):1429–1440, 2019.
- [197] M. F. Di Carli, M. C. Tobes, T. Mangner, A. B. Levine, O. Muzik, P. Chakroborty, and T. B. Levine. Effects of cardiac sympathetic innervation on coronary blood flow. *The New England Journal of Medicine*, 336(17):12081215, April 1997. ISSN 0028-4793. doi: 10.1056/NEJM199704243361703.
- [198] M. Liu, J. Liu, L. Zhang, W. Xu, et al. An evidence of brain-heart disorder: mental stress-induced myocardial ischemia regulated by inflammatory cytokines. *Neurological Research*, 42(8):670675, August 2020. ISSN 1743-1328. doi: 10.1080/01616412.2020.1783879.
- [199] K. N. Kershaw, A. D. Lane-Cordova, M. R. Carnethon, H. A. Tindle, and K. Liu. Chronic stress and endothelial dysfunction: The multi-ethnic study of atherosclerosis (mesa). *American Journal of Hypertension*, 30(1):7580, January 2017. ISSN 1941-7225. doi: 10.1093/ajh/hpw103.
- [200] P. P Chang, D. E. Ford, L. A. Meoni, N. Wang, and M. J. Klag. Anger in young men and subsequent premature cardiovascular disease: the precursors study. *Arch. Intern. Med.*, 162(8):901–906, 2002.
- [201] R J Henning. Obesity and obesity-induced inflammatory disease contribute to

- atherosclerosis: a review of the pathophysiology and treatment of obesity. *Am. J. Cardiovasc. Dis.*, 11(4):504–529, 2021.
- [202] M. Piché, A. Tchernof, and J. P. Després. Obesity phenotypes, diabetes, and cardiovascular diseases. *Circulation Research*, 126(11):14771500, May 2020. ISSN 1524-4571. doi: 10.1161/CIRCRESAHA.120.316101.
- [203] M de Onis, M Blössner, and E Borghi. Global prevalence and trends of overweight and obesity among preschool children. *Am. J. Clin. Nutr.*, 92(5):1257–1264, 2010.
- [204] S M Grundy. Multifactorial causation of obesity: implications for prevention. *Am. J. Clin. Nutr.*, 67(3 Suppl):563S–72S, 1998.
- [205] R. Din-Dzietham, Y. Liu, M. V. Bielo, and F. Shamsa. High blood pressure trends in children and adolescents in national surveys, 1963 to 2002. *Circulation*, 116(13):14881496, September 2007. doi: 10.1161/CIRCULATIONAHA.106.683243.
- [206] Christopher B Cole, Majid Nikpay, Alexandre F R Stewart, and Ruth McPherson. Increased genetic risk for obesity in premature coronary artery disease. *Eur. J. Hum. Genet.*, 24(4):587–591, 2016.
- [207] M. Raj. Obesity and cardiovascular risk in children and adolescents. *Indian J. Endocrinol. Metab.*, 16(1):13–19, 2012.
- [208] M. Mihuta, C. Paul, A. Ciulpan, F. Dacca, et al. Subclinical atherosclerosis progression in obese children with relevant cardiometabolic risk factors can be assessed through carotid intima media thickness. *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, 11(22):10721, 2021.
- [209] D. S. Freedman, L. K. Khan, W. H. Dietz, S. R. Srinivasan, and G. S. Berenson. Relationship of childhood obesity to coronary heart disease risk factors in adulthood:

- the bogalusa heart study. *Pediatrics*, 108(3):712718, September 2001. ISSN 1098-4275. doi: 10.1542/peds.108.3.712.
- [210] V. Beauloye, F. Zech, H. Tran, P. Clapuyt, et al. Determinants of early atherosclerosis in obese children and adolescents. *J. Clin. Endocrinol. Metab.*, 92(8):3025–3032, 2007.
- [211] H C McGill, Jr, A C McMahan, E E Herderick, A W Zieske, et al. Obesity accelerates the progression of coronary atherosclerosis in young men. *Circulation*, 105(23):2712–2718, 2002.
- [212] A. Yip and J. Saw. Spontaneous coronary artery dissection-a review. *Cardiovasc Diagn Ther*, 5(1):37–48, 2015.
- [213] M. Garcia-Guimarães, T. Bastante, P. Antuña, C. Jimenez, et al. Spontaneous coronary artery dissection: Mechanisms, diagnosis and management. *Eur Cardiol*, 15:1–8, 2020.
- [214] T. Nakashima, S. Yasuda, T. Noguchi, S. Haruta, et al. Abstract 13596: The prognostic impact of spontaneous coronary artery dissection in younger female patients with acute myocardial infarction; report from angina pectoris-myocardial infarction multicenter investigations in japan. *Circulation*, 130(suppl\_2):A13596A13596, November 2014. doi: 10.1161/circ.130.suppl\\_2.13596.
- [215] T Nishiguchi, A Tanaka, Y Ozaki, A Taruya, et al. Prevalence of spontaneous coronary artery dissection in patients with acute coronary syndrome. *Eur Heart J Acute Cardiovasc Care*, 5(3):263–270, 2016.
- [216] R E White. Estrogen and vascular function. *Vascul. Pharmacol.*, 38(2):73–80, 2002.
- [217] B. J. Rensing, M. Kofflard, M. J.B.M. van den Brand, and D. P. Foley. Spontaneous dissections of all three coronary arteries in a 33-week-pregnant woman. *Catheteri-*

- zation and Cardiovascular Interventions*, 48(2):207210, 1999. ISSN 1522-726X. doi: 10.1002/(SICI)1522-726X(199910)48:2<207::AID-CCD19>3.0.CO;2-2.
- [218] T. S. Mikkola and T. B. Clarkson. Estrogen replacement therapy, atherosclerosis, and vascular function. *Cardiovasc. Res.*, 53(3):605–619, 2002.
- [219] M J Tikkanen, E A Nikkila, and E Vartiainen. Natural oestrogen as an effective treatment for type-II hyperlipoproteinaemia in postmenopausal women. *Lancet*, 2(8088):490–491, 1978.
- [220] F. Grodstein, J. E. Manson, G. A. Colditz, W. C. Willett, et al. A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Annals of Internal Medicine*, 133(12):933941, December 2000. ISSN 0003-4819. doi: 10.7326/0003-4819-133-12-200012190-00008.
- [221] S N Hayes, M S Tweet, D Adlam, Esther S H Kim, et al. Spontaneous coronary artery dissection: JACC State-of-the-Art review. *J. Am. Coll. Cardiol.*, 76(8):961–984, 2020.
- [222] L. Marcoff and E. Rahman. Menstruation-associated spontaneous coronary artery dissection. *The Journal of Invasive Cardiology*, 22(10):E183–185, October 2010. ISSN 1557-2501.
- [223] Amitesh Aggarwal, Saurabh Srivastava, and M Velmurugan. Newer perspectives of coronary artery disease in young. *World J. Cardiol.*, 8(12):728, 2016.
- [224] L W Klein. Acute coronary syndromes in young patients with angiographically normal coronary arteries. *Am. Heart J.*, 152(4), 2006.
- [225] M. Tanaka, K. Tomiyasu, M. Fukui, S. Akabame, et al. Evaluation of characteristics

and degree of remodeling in coronary atherosclerotic lesions by 64-detector multislice computed tomography (MSCT). *Atherosclerosis*, 203(2), 2009.

- [226] I J Kullo, W D Edwards, and R S Schwartz. Vulnerable plaque: pathobiology and clinical implications. *Ann. Intern. Med.*, 129(12), 1998.

## Chapter 4

# Polygenic risk scores in Myocardial Injury after Non-cardiac Surgery: a VISION substudy

Ann Le<sup>1,2</sup>, Guillaume Paré<sup>1,2,3,4,5,6</sup>, PJ Devereaux, Ibrahim Quazi, Shihong Mao<sup>1</sup>, Michael Chong<sup>1,2,3</sup>, Diane Heels-Ansdell, Michael Wang, Sandra N. Ofori<sup>1,6,7</sup>, David Conen, Emmanuelle Duceppe, Jessica Spence, Emilie Belley-Cote, Caleb Beck, William McIntyre, Richard Whitlock, Jeff Healey, Shirley Pettit, and Flavia K. Borges<sup>1,6,7</sup>. On behalf of VISION Investigators.

1. Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada.
2. Department of Medical Sciences, McMaster University, Faculty of Health Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.
3. Department of Biochemistry and Biomedical Sciences, McMaster University, Faculty of Health Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.

4. Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada
5. Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroot School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada
6. Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street West, Hamilton ON L8L 4K1, Canada.
7. Department of Medicine, McMaster University, 1280 Main Street West, Hamilton ON L8L 4K1, Canada.

## 4.1 Forward

Perioperative complications have garnered attention recently, especially with the growing population of elderly individuals requiring surgeries and the ongoing gaps of understanding certain aspects in their etiologies or pathophysiologies. Myocardial injury after non-cardiac surgery (MINS) is one such example. Despite being the most prevalent cardiovascular complication post-surgery, its root causes remain incompletely elucidated, prompting investigation into genetic causes. Notably, evidence has previously demonstrated that many clinical risk factors related to MINS, such as coronary artery disease (CAD) or diabetes, can have strong genetic dispositions. Currently, the most common risk predictor for MINS is the Revised Cardiac Risk Index (RCRI), which considers several components that exhibit significant genetic predisposition. Thus, polygenic risk scores (PRS) related to MINS were considered as a potential quantitative measure of risk association with MINS. PRS calculations were performed for various risk factors associated with MINS, and their associations with MINS were examined to gain insights into its underlying cause.

Our application of PRS to MINS unveiled the association between type II diabetes (T2D) PRS and MINS, and a hemoglobin A1c (HbA1c) PRS and MINS. The case-control study was based within the VISION cohort, an international cohort of individuals aged 45

and older unergoing inpatient noncardiac surgery across 28 centres and 14 countries. MINS was determined by daily troponin levels on days 1, 2, and 3 after surgery, after which PRS were computed using publicly available genome-wide association study (GWAS) summary statistics. RCRI was also computed for each participant for comparison purposes. Conditional logistic regression was used to evaluate PRS association with MINS, and C-statistics from the receiver-operator curve (ROC area under the curve [AUC]) were calculated to determine discriminative capacity of PRS. Apart from the T2D and HbA1c PRSs, no other PRS were associated with MINS, including the PRS for CAD.

In summary, the findings revealed that when the T2D and HbA1c PRSs are combined with RCRI, they are associated with MINS and can potentially improve prediction. The association is encouraging and unveils potential insights into MINS pathophysiology and treatment. In particular, the observed association between T2D and HbA1c PRS may suggest a potential role of microvascular disease in MINS onset as opposed to being driven by atherosclerotic CAD as formerly hypothesized. However, it should be noted that the study is potentially limited by the lower sample size, and previous studies have shown associations between CAD PRS and MINS. The findings suggest that the underlying pathophysiology of MINS is potentially multifaceted, and it may be beneficial to consult clinical trials which consider diabetes and glucose management.

This manuscript is currently undergoing minor revisions for *JAAC: Advances*. Flavia K. Borges and Guillaume Paré conceptualized and designed the study. Ibrahim Quazi conducted initial statistical analyses. Ann Le followed up with the analysis plan, performed further statistical analyses (rerun of logistic regressions, and reclassification analyses), and wrote the manuscript. All authors contributed to the interpretation of the findings and to the critical reading and revision of the manuscript.

## 4.2 Abstract

**BACKGROUND:** Myocardial injury after noncardiac surgery (MINS) is the most prevalent vascular complication following surgical procedures. Despite the widespread use of the Revised Cardiac Risk Index (RCRI) score for prediction of postoperative cardiovascular complications, predictive accuracy remains suboptimal. Considering genetic influences may improve risk prediction, through the use of polygenic risk scores (PRS). We propose integration of PRS with the RCRI score to enhance prediction of MINS, while also seeking to identify PRS associated with MINS to gain insights into its pathophysiology.

**METHODS:** This is a case-control study nested within the VISION cohort, a large-scale international prospective representative cohort comprising 40,004 individuals aged 45 and older undergoing inpatient noncardiac surgery across 28 centres and 14 countries. Daily troponin levels were measured preoperatively and on days 1, 2 and 3 after surgery. 3,264 blood samples were processed, frozen, and stored for future genotyping. PRS were computed for MINS risk factors using publicly available summary statistics to compare predictive and discrimination performances of PRS and RCRI. Logistic regression techniques were employed to evaluate the association between PRS and MINS, adjusting for the RCRI score and genetic principal components (PCs). PRS discrimination was determined both independently using c-statistics and in combination with the RCRI score.

**RESULTS:** Among participants from the VISION Biobank, 253 MINS cases were matched with an equal number of controls, adjusting for age, sex, and limited to individuals with European genetic ancestry ( $n_{\text{total}} = 506$ ). In the conditional logistic regression model adjusting for matched pairs, the Type II Diabetes (T2D) PRS (adjusted OR = 1.26, 95% CI: 1.00–1.58, p-value = 0.047), and the HbA1c PRS (adjusted OR = 1.26, 95% CI: 1.03

1.54, p-value = 0.026) were associated with MINS. No other PRS was associated with MINS, including PRS for coronary artery disease (CAD), stroke and lipid biomarkers. C-statistics reported no significant improvement in discrimination capacity for PRS scores in addition to the RCRI score.

**CONCLUSION:** The T2D PRS and the HbA1c PRS was associated with an increased risk of MINS. However, no other PRS were associated with MINS. As previous studies have indicated associations between cardiovascular conditions (CAD) and MINS, these findings may reflect the multifactorial pathophysiology of MINS and the need for larger, better powered, genetic studies. Given prior research indicating a connection between perioperative glucose levels and MINS, trials evaluating interventions effective in managing diabetes and glucose during the perioperative period warrant consideration.

### 4.3 Condensed Abstract

Despite being the most common vascular complication following surgical procedures, the etiology and predictive accuracy for myocardial injury after noncardiac surgery (MINS) remains suboptimal. Currently, the Revised Cardiac Risk Index (RCRI) is used to predict MINS. We propose integration of genetic risk into MINS prediction, through the use of polygenic risk scores (PRS). PRS were created for MINS risk factor with the aim of improving risk prediction for MINS. A significant association was found between T2D PRS and MINS, and HbA1c PRS and MINS. No other significant associations were found between any other PRS and MINS, including the PRS for coronary artery disease (CAD). These findings suggest a potential underlying multifactorial pathophysiology for MINS, and it may be beneficial to consult glucose biomarkers to improve risk prediction for MINS.

**KEYWORDS:** Myocardial injury after noncardiac surgery, Risk prediction of myocardial injury after noncardiac surgery, Polygenic risk scores

### 4.4 Introduction

Worldwide, 1 in every 30 to 40 adults will undergo a noncardiac surgery [1]. While noncardiac surgeries can greatly improve a patients quality of life, perioperative complications may occur due to the patients circumstances regarding the underlying clinical condition requiring surgery, the anesthetic, and the surgical procedure. The most common cardiovascular complication after surgery is myocardial injury after noncardiac surgery (MINS), defined as myocardial injury presumed to be due to underlying cardiac underlying cardiac ischemia occurring during or within 30 days after surgery [2]. It occurs in up to 1 in 6 patients undergoing noncardiac surgery. In a recent systematic review with over 530

000 participants, the estimated incidence of MINS was 18% [2, 3]. MINS is an important perioperative event as it is associated with increased risk of further cardiovascular events and death in the coming 30 days after surgery [2, 3]. The VISION (Vascular Events in Noncardiac Surgery Participants Cohort Evaluation) study has demonstrated that among 40,004 patients undergoing inpatient noncardiac surgery, MINS was one of the three most important perioperative complications associated with perioperative mortality [4]. Despite its high incidence, the etiology of MINS remains incompletely understood, although some studies have demonstrated that a high proportion of patients suffering MINS have underlying coronary artery disease (CAD). A better understanding of MINS predictors and underlying pathophysiology may improve its prevention, early detection and management [5].

Currently, the Revised Cardiac Risk Index (RCRI) score is the most common method recommended for clinical prediction of perioperative cardiovascular complications. The RCRI score consists of six variables: history of CAD, history of congestive heart failure (CHF), history of cerebrovascular disease, diabetes on insulin, creatinine levels greater than 177 mmol/L and high risk surgery[5, 6]. Previous evidence indicates that certain traits considered in the RCRI score (such as CAD, and diabetes) can exhibit significant genetic predisposition, suggesting that employing genetic tools for prediction of MINS could be successful [7, 8].

The advent of genome-wide association studies (GWAS) has facilitated the identification of numerous genetic loci linked to various traits [9]. By leveraging known genetic associations from GWAS, polygenic risk scores (PRS) can be computed for individuals, providing quantitative measures of their genetic predisposition to specific phenotypes. PRS have recently gained relevance for their potential in early disease intervention and prevention,

notably in conditions such as premature CAD or Alzheimers disease [10, 11]. For instance, GWAS have pinpointed multiple genetic loci, including the 9p21 locus, known for its robust association with CAD and myocardial infarction (MI) [12, 13, 14]. Previous studies have indicated that PRS outperform individual clinical risk factors in CAD prediction [15, 12]. Additionally, PRS have demonstrated utility in diabetes prediction and differentiation between type I diabetes (T1D) and type II diabetes (T2D) [16, 17]. Recent findings further highlight the significance of leveraging genetic predictions, with both PRS and family risk scores independently linked to T2D risk [18]. Considering PRS alongside clinical risk scores can be advantageous, as genetic factors can often serve as the earliest indicators for many diseases or traits with heritability. PRS can also provide insight into a patient’s lifelong predisposition to disease, capturing additional risk information.

In this study, we computed PRS for traits corresponding to potential risk factors for myocardial injury after noncardiac surgery (MINS) in participants with and without MINS using patient genotyping data from the VISION study [4, 19]. We hypothesized that individuals with a history of MINS would exhibit a higher standardized PRS for cardio-metabolic risk factors such as CAD, T2D and lipid biomarkers. We additionally aimed to assess the additional clinical value of PRS in identifying high-risk MINS patients beyond conventional perioperative risk scores.

## **4.5 Methods**

### **4.5.1 Study Population & Definition**

This is a case-control study nested within the VISION cohort. The study population is derived from the VISION Study, an international initiative comprising 28 centers across 14 countries with over 40,000 representative participants aged 45 and above undergoing

inpatient noncardiac surgery (Figure 4.1) [4, 2]. This cohort was prospectively evaluated for major complications occurring within 30 days after surgery. MINS outcome was determined by measuring daily troponin levels pre- and post-operatively. Blood samples were obtained preoperatively and on days 1, 2 and 3 post-surgery. 4,428 patients were included in the VISION Biobank, and 3,264 had a buffy coat sample available.

Among Biobank participants, we randomly selected 300 cases of participants who experienced MINS, followed by the selection of 300 controls who did not exhibit troponin elevation within the initial 30 days post-surgery (i.e. without MINS), matched by age ( $\pm 5$  years) and sex. We restricted the Biobank samples to patients of European ancestry (representing 85% of VISION Biobank samples). Additionally, 83 participants were excluded for lack of case-control match, misrepresented ancestry and lack of RCRI availability. The final sample comprised 506 participants, evenly divided between those with MINS (253 participants) and those without MINS (253 participants).

MINS diagnosis was established based on a new troponin elevation judged to be due to acute ischemic etiology within 30 days post-surgery, excluding alternative causes such as chronic troponin elevation, pulmonary embolism, sepsis or rapid atrial fibrillation. All patients with a troponin elevation in the main VISION Study were assessed for ischemic symptoms and electrocardiographic ischemic findings. Continuous monitoring was conducted by research personnel during hospitalization, along with a 30-day follow-up telephone call. Data collection on variables and outcomes involved interviews with patients or their next of kin, along with chart reviews. Source documents pertaining to outcome events were obtained from primary-care physician or hospital records. Case report forms and supporting documentation were securely stored at the coordination centre (Population Health Research Institute [PHRI], Hamilton, Ontario, Canada) in an online data management

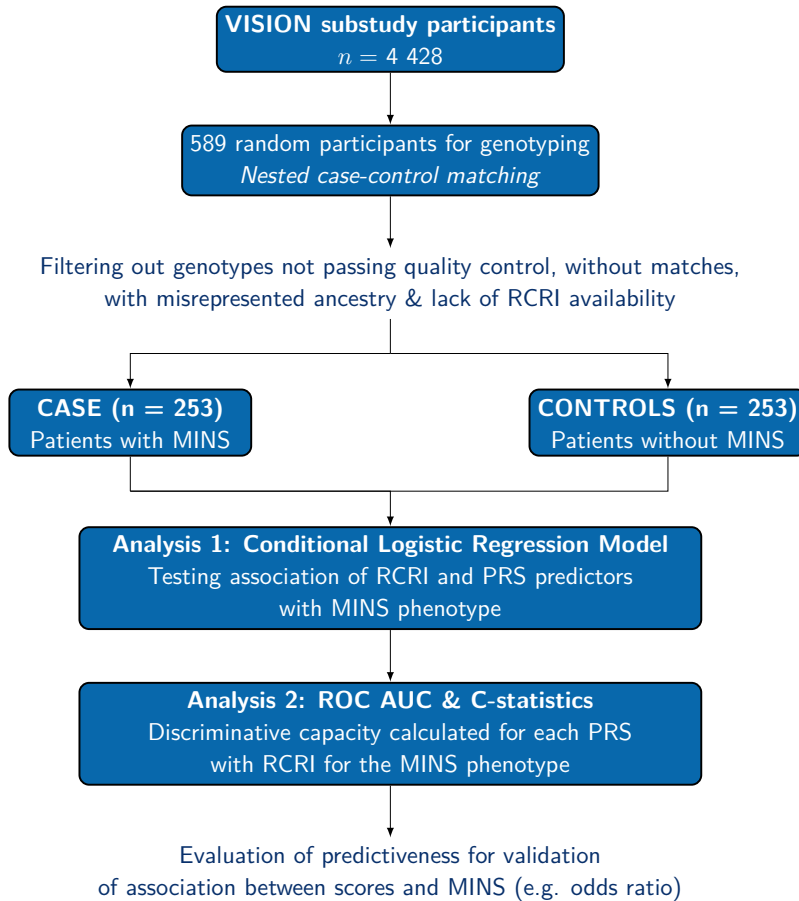


Figure 4.1: **Overview of Experimental Design for VISION MINS PRS study.**

This figure shows the flow of participant selection, along with the analyses that were conducted on the sample to determine the association between PRS and MINS. The VISION Biobank sample originally consists of 4,428 patients, from which 600 patients were randomly selected for case-control matching. Amongst the 600 patients, those without matches, misrepresented ancestry and lack of RCRI availability were filtered out, resulting in a final sample of 253 cases (patients with MINS) and 253 controls (patients without MINS) matched for age and sex. Conditional logistic regression and discrimination capacity analyses were performed on this sample.

RCRI = revised cardiac risk index, PRS = polygenic risk score, ROC = receiver-operator curve

system (iDataFax).

Blood samples from the VISION Biobank were collected, processed, frozen, and stored for genotyping purposes. Genotyping was carried out using the Axiom Precision Medicine Research Array (PMRA) release 3, with quality control conducted using PLINK software. Among the 589 genotyped samples, 583 passed quality control. The genotyping chip assays up to 841,064 variants, of which approximately 578,220 variants (69%) were detected within the VISION cohort; 265,804 variants (31%) were excluded because the second allele was not observed within study samples due to the relatively moderate sample size ( $N < 1000$ ). Of these 578,220 variants detectable within the VISION-CS participants, 573,428 (99%) passed quality control. Genetic imputation was then performed on the directly genotyped variants against the TOPMED release 2 reference panel using the TOPMED imputation server (EAGLE2 for phasing and Positional Burrows Wheeler for imputation). Finally, we excluded variants with a minor allele frequency (MAF)  $< 0.1\%$  and imputation  $r^2 < 0.3$ .

## 4.6 Calculation and Derivation of Polygenic Risk Scores

Genome-wide association study (GWAS) summary statistics were obtained from large, external genetic meta-analysis consortia, corresponding to CAD and other traits relevant to MINS risk factors S4.1). The chosen traits were cardiovascular risk factors, which are hypothesized to be related to MINS. LASSOSUM2, a method for computing PRS using penalized regression, allows for adjustments of tuning parameters through an embedded reference panel [19]. The computation efficiency and predictive accuracy of LASSOSUM2 have been evidenced to surpass those of comparable methods [20, 21]. PRS for each trait were computed using LASSOSUM2 for the 506 participants, with variants filtered based on a significance threshold of  $p < 0.01$ .

## 4.7 Statistical Analyses

We calculated that, to detect an odds ratio of 1.5 for MINS per 1 standard deviation (SD) increase in CAD PRS with 90% power at significance level of 0.05 (two-sided), a sample of 128 matched pairs would be necessary, with each pair comprising 1 case and 1 control.

Conditional logistic regression was utilized to explore the association between MINS risk factor PRS and MINS, with and without RCRI adjustment, while adjusting for 10 principal components (PCs) to address genetic ancestry. The `clogit()` function from the R package `survival` was used to construct all conditional logistic regression models. Genetic PCs were derived using GCTA (Genome-wide Complex Trait Analysis), which estimates genetic relatedness among SNPs through variance calculations. PCs were generated using the 1000 Genomes database as a reference, retaining clusters within three SDs of the reference data as the 1<sup>st</sup> and 2<sup>nd</sup> PCs.

The discriminative performance of the PRS was assessed using the area under the receiver operating characteristic (ROC) curve. C-statistics were computed to determine whether PRS improves the discrimination of predictive model for MINS. The R package `pROC` was used to determine C-statistics through the `roc()` function. Initially, discrimination was evaluated for PRS alone, and then combined with the clinical risk prediction RCRI score. Furthermore, Net Reclassification Improvement (NRI) was also calculated, using the `improveProb()` function from the `Hmisc` R package. R version 4.2.0 was used for all statistical analyses.

## 4.8 Results

The baseline characteristics of participants according to case (participants with MINS) and control (participants without MINS) status are displayed in Table 4.1. The cohort had an average age of 71.7 years, with 282 male and 224 female participants. The most prevalent comorbidities included hypertension (67.7%), CAD (22.8%), and diabetes (21.3%). The distributions of CAD PRS, T2D PRS, Hemoglobin A1c (HbA1c) PRS and RCRI scores among patients with and without MINS is shown in Table 4.3.

The CAD PRS exhibited a significant association with the CAD baseline condition (OR 2.63, 95% CI 1.54 - 4.50,  $p < 0.001$ ) (4.2). Similarly, the HbA1c PRS was associated with T2D (OR 1.67, 95% CI 1.12 - 2.48,  $p = 0.011$ ). However, there was no significant association between the T2D PRS and the occurrence of T2D (OR 1.21, 95% CI 0.82 - 1.81,  $p = 0.34$ ). Other phenotypes apart were not examined due to insufficient data.

Among all PRS traits analyzed, the HbA1c PRS exhibited a significant association with MINS, with an odds ratio (OR) of 1.30 per SD (95% CI 1.07 - 1.57,  $p = 0.0081$ ). Moreover, the HbA1c PRS could stratify patients into risk categories without RCRI (Figure 4.2). No other PRS demonstrated a significant association with MINS.

We subsequently assessed the association of each PRS further adjusting for RCRI (Table 4.4). With the inclusion of RCRI, the T2D PRS exhibited a significant association with MINS, with an OR of 1.26 per SD (95% CI 1.00 - 1.58,  $p = 0.047$ ). The HbA1c PRS also showed significance, with an OR of 1.26 per SD (95% CI 1.03 - 1.54,  $p = 0.026$ ). There were no significant associations with any other PRS, nor did the inclusion of any other traits PRS enhance prediction when combined with RCRI (Figure 4.2).

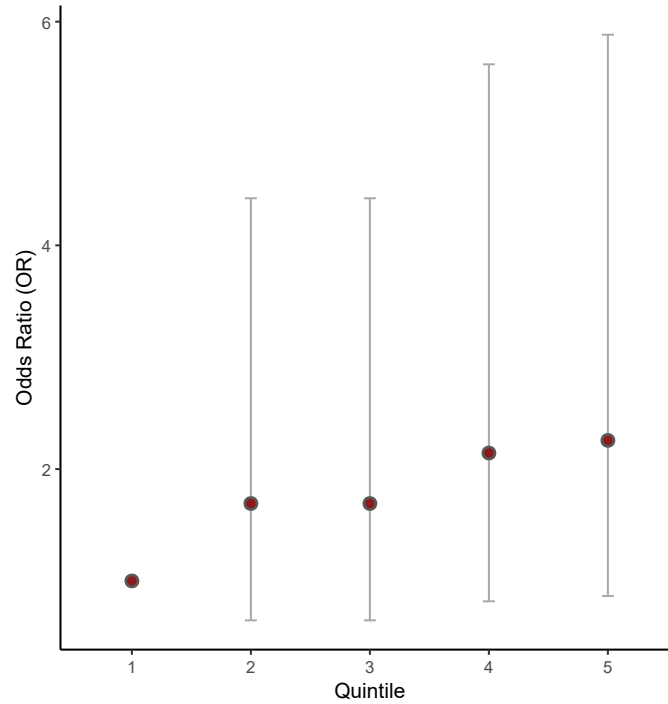


Figure 4.2: **MINS Odds Ratio according to quintile of HbA1c PRS.** The figure displays the odds ratio of association per quintile of HbA1c PRS, with the 1<sup>st</sup> quintile as a reference. The HbA1c PRS can stratify MINS risk without RCRI. Confidence interval bars for logsitic regression estimates are also shown.

MINS = myocardial injury after non-cardiac surgery, HbA1C = Hemoglobin A1c, PRS = polygenic risk score, RCRI = revised cardiac risk index

Characteristic	All	MINS group	No MINS group	p-value
N (%)	506	253 (50.0)	253 (50.0)	.
Age (years), mean $\pm$ SD	71.7 $\pm$ 9.1	71.8 $\pm$ 9.0	71.7 $\pm$ 9.1	.
Age - n (%)				
45 - 64	113 (22.3)	57 (22.5)	56 (22.1)	.
65 - 74	185 (36.6)	92 (36.4)	93 (36.8)	.
$\geq$ 75	208 (41.1)	104 (41.1)	104 (41.1)	.
Male n (%)	282 (55.7)	141 (55.7)	141 (55.7)	.
Current tobacco use - n (%)	58 (11.5)	34 (13.4)	24 (9.5)	0.157
History of atherosclerotic disease* - n (%)	161 (31.8)	98 (38.7)	63 (24.9)	< 0.001
History of coronary artery disease - n (%)	120 (22.8)	68 (25.9)	52 (19.8)	0.077
History of diabetes - n (%)	108 (21.3)	65 (25.7)	43 (17.0)	0.018
History of congestive heart failure - n (%)	18 (3.6)	12 (4.7)	6 (2.4)	0.157
Hypertension - n (%)	342 (67.6)	182 (71.9)	160 (63.2)	0.026
Type of surgery - n (%)				
Vascular	35 (6.9)	19 (7.5)	16 (6.3)	0.590
Thoracic	20 (4.0)	12 (4.7)	8 (3.2)	0.317
Major Urology/gynecology	88 (17.4)	43 (17.0)	45 (17.8)	0.811
General	78 (15.4)	45 (17.8)	33 (13.0)	0.140
Orthopedic	212 (41.9)	108 (42.7)	104 (41.1)	0.715
Neuro	19 (3.8)	9 (3.6)	10 (4.0)	0.819
Low risk	85 (16.8)	33 (13.0)	52 (20.6)	0.017
Urgent/Emergent surgery - n (%)	11 (2.2)	4 (1.6)	7 (2.8)	0.366
Preoperative eGFR (MDRD, mL/min/1.76 m <sup>2</sup> ) - n (%)				
< 30 or on dialysis	16 (3.2)	14 (5.5)	2 (0.8)	0.003
30 to <45	45 (8.9)	31 (12.3)	14 (5.5)	0.005
45 to <60	93 (18.4)	51 (20.2)	42 (16.6)	0.272
$\geq$ 60	352 (69.6)	157 (62.1)	195 (77.1)	< 0.001
RCRI score - n (%)				
0	262 (51.8)	106 (41.9)	156 (61.7)	< 0.001
1	161 (31.8)	92 (36.4)	69 (27.3)	0.021
2	64 (12.6)	41 (16.2)	23 (9.1)	0.014
$\geq$ 3	19 (3.8)	14 (5.5)	5 (2.0)	0.039

\*Coronary artery disease/peripheral vascular disease/cerebrovascular event

eGFR = estimated glomerular filtration rate, MDRD = modification of diet in renal disease, RCRI = revised cardiac risk index

Table 4.1: Baseline characteristics of VISION Biobank case-control participants of European ancestry.

Regression Model	Odds Ratio (95% CI)	p-value	Nagelkerke Pseudo $R^2$	c-statistic (95% CI)
<b>CAD PRS with CAD baseline condition</b>	2.63 (1.54 - 4.50)	< 0.001	0.27	0.83 (0.77 - 0.90)
<b>T2D PRS with T2D baseline condition</b>	1.21 (0.82 - 1.81)	0.34	0.18	0.72 (0.69 - 0.85)
<b>HbA1c PRS with T2D baseline condition</b>	1.67 (1.12 - 2.48)	0.011	0.25	0.73 (0.60 - 0.86)

Table 4.2: Association between CAD PRS and preoperative CAD and T2D PRS and HbA1c PRS with preoperative T2D.

	<i>N</i>	Mean $\pm$ SD	Median (IQR: p25, p75)	Minimum	Maximum	<i>p</i> -value
<b>Preoperative CAD PRS</b>						
MINS	253	-0.01 $\pm$ 0.94	0.01 (-0.66, 0.55)	-2.37	2.59	0.51 <sup>¶</sup>
No MINS	253	0.05 $\pm$ 0.96	0.06 (-0.60, 0.68)	-2.52	2.63	
<b>Preoperative T2D PRS</b>						
MINS	253	-0.028 $\pm$ 0.85	0.28 (-0.33, 0.78)	-2.24	2.08	0.072 <sup>¶</sup>
No MINS	253	0.21 $\pm$ 0.90	-0.10 (-0.61, 0.52)	-1.76	2.10	
<b>Preoperative HbA1c PRS</b>						
MINS	253	0.11 $\pm$ 1.0	0.11 (-0.42, 0.68)	-3.16	2.97	0.0066 <sup>¶</sup>
No MINS	253	-0.12 $\pm$ 0.98	0.065 (-0.81, 0.48)	-3.06	2.50	
<b>Revised cardiac risk index (RCRI)</b>						
MINS	253	0.88 $\pm$ 0.96	1.00 (0.00, 1.00)	0.00	5.00	< 0.001 <sup>§</sup>
No MINS	253	0.51 $\pm$ 0.74	0.00 (0.00, 1.00)	0.00	3.00	

<sup>¶</sup> Paired t-test comparing means between the groups

<sup>§</sup> *p*-value on Wilcoxon signed rank comparing medians between the groups and assuming non-normally distributed data

CAD = coronary artery disease, T2D = type 2 diabetes, HbA1c = Hemoglobin A1c

Table 4.3: Polygenic risk score and revised cardiac risk index (RCRI) in patients with and without MINS.

Outcome: MINS within 30 days after surgery	Odds Ratio (95% CI)	<i>p</i> -value	Nagelkerke Pseudo $R^2$	c-statistic (95% CI)
<b>Model 1:</b> CAD PRS	0.92 (0.75 - 1.12)	0.16	0.057	0.63 (0.58 - 0.68)
<b>Model 2:</b> RCRI	1.78 (1.38 - 2.30)	< 0.001***	0.14	0.70 (0.66 - 0.75)
<b>Model 3:</b> CAD PRS + RCRI				0.72 (0.67 - 0.76)
CAD PRS	0.86 (0.69 - 1.06)	0.16	0.15	
RCRI	1.70 (1.36 - 2.14)	< 0.001***		
<b>Model 4:</b> T2D PRS	1.20 (0.97 - 1.48)	0.10	0.065	0.64 (0.60 - 0.69)
<b>Model 5:</b> T2D PRS + RCRI				0.72 (0.67 - 0.76)
T2D PRS	1.26 (1.00 - 1.58)	0.047*	0.16	
RCRI	1.82 (1.40 - 2.36)	< 0.001***		
<b>Model 6:</b> LDL PRS	1.04 (0.86 - 1.27)	0.67	0.056	0.63 (0.58 - 0.68)
<b>Model 7:</b> LDL PRS + RCRI				0.71 (0.66 - 0.75)
LDL PRS	1.04 (0.85 - 1.28)	0.70	0.14	
RCRI	1.78 (1.38 - 2.30)	< 0.001***		
<b>Model 8:</b> HDL PRS	0.93 (0.75 - 1.15)	0.48	0.48	0.63 (0.58 - 0.68)
<b>Model 9:</b> HDL PRS + RCRI				0.70 (0.66 - 0.75)
HDL PRS	0.99 (0.80 - 1.24)	0.96	0.14	
RCRI	1.78 (1.37 - 2.30)	< 0.001***		
<b>Model 10:</b> TG PRS	0.99 (0.83 - 1.18)	0.90	0.054	0.63 (0.58 - 0.68)
<b>Model 11:</b> TG PRS + RCRI				0.71 (0.66 - 0.75)
TG PRS	0.99 (0.80 - 1.16)	0.99	0.043	
RCRI	1.78 (1.38 - 2.31)	< 0.001***		
<b>Model 12:</b> HbA1c PRS	1.30 (1.07 - 1.57)	0.0081**	0.082	0.66 (0.61 - 0.70)
<b>Model 13:</b> HbA1c PRS + RCRI				0.73 (0.68 - 0.77)
HbA1c PRS	1.26 (1.03 - 1.54)	0.026*	0.16	
RCRI	1.75 (1.35 - 2.27)	< 0.001***		

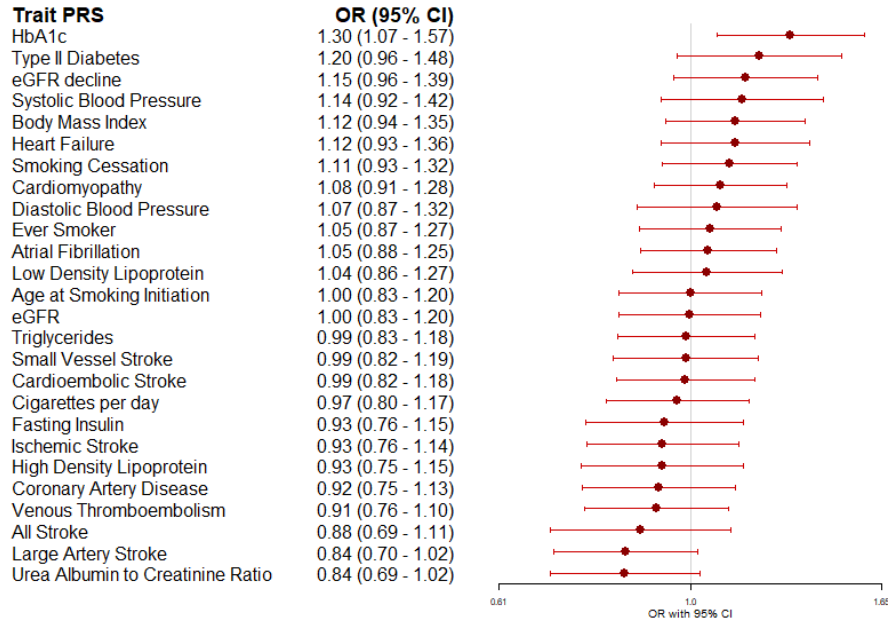
MINS = myocardial injury after noncardiac surgery (MINS), CI = confidence interval, CAD = coronary artery disease, RCRI = revised cardiac risk index, T2D = type 2 diabetes (diabetes mellitus), LDL = low density lipoprotein, HDL = high density lipoprotein, TG = triglycerides, HbA1c = Hemoglobin A1c

‡ Adjusted for 10 Principal Components (PCs) as covariables considering genetic ancestry as a confounder.

(\*) denotes significance

Table 4.4: Logistic regression models studying association between revised cardiac risk index (RCRI) score, polygenic risk scores (PRS) and myocardial injury after non-cardiac surgery (MINS). All models are adjusted for 10 PCs accounting for genetic ancestry as a confounder.

a.



b.

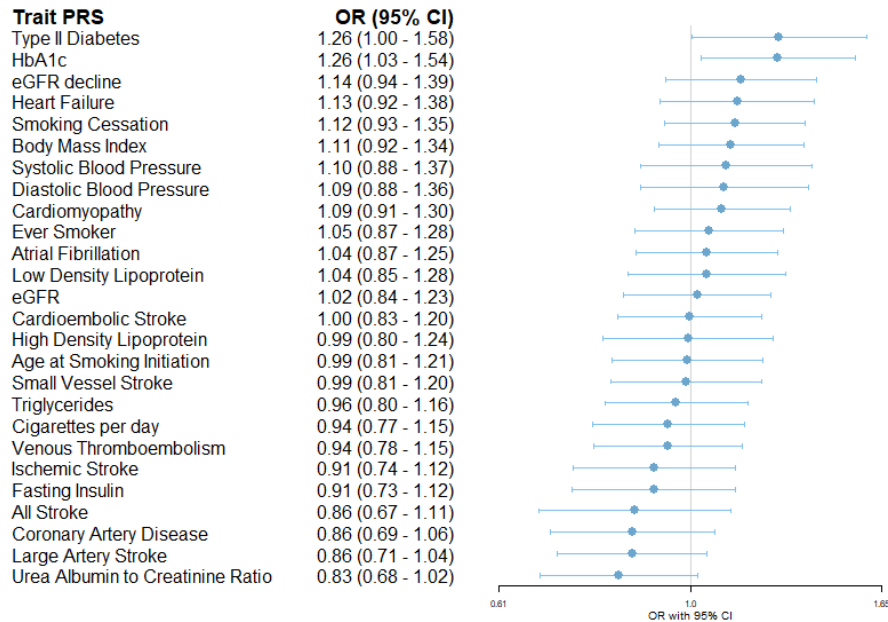


Figure 4.3: **Association between PRS and MINS.** The following forest plots display the association of all traits for which PRS were created by order of descending odds ratio with a 95% confidence interval, as determined through conditional logistic regression. Figure 3a. shows association of each traits PRS without adjustment for RCRI. Figure 3b. shows association of each traits PRS with adjustment for RCRI.

MINS = myocardial injury after non-cardiac surgery, HbA1c = Hemoglobin A1c, PRS = polygenic risk score, RCRI = revised cardiac risk index

The discriminative capacity of the CAD, T2D and HbA1c PRS for MINS was determined across different age and sex categories (Supplementary Table S4.2). Throughout, the RCRI score maintained a higher discrimination threshold, and the inclusion of these PRS alongside RCRI score did not significantly enhance discrimination compared to RCRI score alone. This finding was confirmed through a Delong test comparing ROC AUCs for the CAD, T2D, and HbA1c PRS across all subgroups, where no significant differences were observed. To corroborate these findings, NRI analysis was conducted (Supplementary Table S4.4). T2D PRS and HbA1c PRS displayed significant improvements in discrimination, while CAD PRS did not.

## 4.9 Discussion

Our study represents the first prospective case-control cohort investigation employing standardized perioperative troponin monitoring to explore associations between PRS and MINS within a perioperative context. Our findings reveal that the combination of T2D PRS and HbA1c PRS with RCRI improves MINS prediction. No other PRS models exhibited significant improvements or associations with MINS, including the CAD PRS which was highly predictive of CAD at baseline. The association between the T2D PRS and HbA1c PRS with MINS is encouraging and may offer further insight into MINS pathophysiology and treatment.

The definition of MINS encompasses patients exhibiting asymptomatic troponin elevation attributed to underlying ischemia, as well as those meeting the universal definition of MI [22]. Approximately one-fifth of MINS cases align with the criteria for perioperative MI [2]. Type I MI is caused by atherothrombotic CAD, often triggered by erosion of atherosclerotic plaque, while Type 2 MI arises from an oxygen supply-demand mismatch

due to pathophysiology mechanisms other than coronary atherothrombosis. Although most MINS patients have underlying CAD, only a minority experience underlying ischemia due to plaque rupture during the perioperative period [5, 23, 22]. Indeed, Type 2 MI can develop in surgical contexts due to factors such as hypoxemia, anemia, hypotension, bradyarrhythmia, or increased myocardial oxygen demand induced by tachyarrhythmia or severe hypertension. Hence, in the perioperative scenario, MINS typically reflects a positive stress test, wherein patients with underlying CAD suffer insults inherent to anesthetic and surgical procedures. These include bleeding, hypotension and pain triggering stress response, which can lead to type 2 demand ischemia. Therefore, the observation that MINS exhibited weak association with the CAD PRS implies a multifaceted underlying physiology for MINS. Given that MINS is closely related to Type 2 ischemia, an isolated genetic predictor would be unlikely to stand out in a statistical model. However, its important to note that other studies have reported an association between CAD and MINS [24, 25]. Obstructive CAD was detected in 72% of patients who underwent coronary computed (CT) angiography before noncardiac surgery [5]. Similarly, alternative angiographic investigations have suggested a link between MINS and pre-existing obstructive CAD or unstable coronary plaques [26, 27, 28, 29]. Autopsies revealed evidence of coronary artery plaque rupture in 46% of patients who succumbed to post-operative MI [30]. Additionally, Douville et al. identified an independent association between CAD PRS and MINS (OR 1.12, 95% CI 1.02 - 1.24,  $p = 0.023$ ) in a cohort comprising 429 MINS cases and 89,624 controls without MINS) [24]. The absence of an association between the CAD PRS and MINS in the current study could be attributed to a smaller number of cases, resulting in insufficient statistical power.

Diabetes is strongly linked to microvascular complications such as retinopathy, nephropathy and neuropathy [31]. Microvascular disease encompasses vascular alterations in small

vessels, such as capillaries, ultimately leading to organ dysfunction [31, 32]. Microvascular disease is majorly induced by chronic hyperglycemia, which is progressive in nature. Previous research has demonstrated that T2D is associated with coronary microvascular dysfunction, resulting in consequences such as decreased nitric oxide, impaired vasodilation, and alterations in cardiomyocyte contractility and stiffness. The observed association between T2D PRS and MINS in our study may suggest a potential role of microvascular disease in MINS onset. Additionally, prior studies have indicated an association between intraoperative hyperglycemia and increased MINS incidence. Notably, patients with intraoperative peak glucose levels  $\geq 180$  mg/dL exhibited a significantly higher MINS occurrence relative to those with lower glucose levels (24.2% vs. 17.2%, OR = 1.26 [CI: 1.14 - 1.40,  $p < 0.001$ ]) [33].

This study has several limitations. Specifically, the association between T2D PRS and clinically diagnosed T2D status was not statistically significant, likely due to limited statistical power resulting from a small sample size. While the current PRS methodologies demonstrate limited strength in association and discrimination, current results should be considered as hypothesis-generating, emphasizing the need for replication with larger sample sizes. PRS investigations typically involve extensive cohorts, often comprising hundreds of thousands of participants, with tens of thousands available cases for highly relevant conditions such as CAD and T2D [15, 20, 34]. For instance, the UK Biobank, housing over 500,000 British participants of various ancestries, is a common resource for PRS research[35]. Improved accuracy is anticipated with larger-scale genetic studies focusing on MINS. Furthermore, the absence of MINS-specific GWAS and PRS poses a challenge. MINS may possess a unique genetic architecture driven by MINS-specific genetic pathways distinct from known MINS risk factors. In such scenarios, accurate MINS prediction using PRS may necessitate appropriately powered MINS GWAS. Notably, MINS in the

perioperative setting is associated with various baseline clinical conditions (e.g., CAD, diabetes, chronic kidney disease) and perioperative complications (e.g. bleeding, hypoxia, hypotension)[25]. The multifactorial etiology of MINS often involves multiple baseline risk factors and postoperative complications. Currently, no heritability estimates are available for MINS, which means that genetics could play a minor role in MINS susceptibility, potentially limiting the predictive utility of PRS. Moreover, the scarcity of GWAS data for other MINS risk factors, such as perioperative complications, may similarly constrain our analysis. Also, it should be acknowledged that the discrimination improvements for current PRS were negligible, and is unlikely to make significance changes clinically. Further investigation is warranted to conclude that PRS can enhance risk prediction in a clinical setting.

In conclusion, our study reveals that T2D PRS and HbA1c PRS are associated with elevated MINS risk. These results likely underscore the multifaceted pathophysiology of MINS, emphasizing the necessity for extensive genetic studies involving larger cohorts to delve deeper into these relationships. Given the association between T2D PRS and MINS, coupled with prior findings indicating a correlation between perioperative glucose levels and MINS, clinical trials assessing interventions effective in diabetes and glucose management warrant consideration.

## References

- [1] E. Duceppe, J. Parlow, P. MacDonald, K. Lyons, et al. Canadian Cardiovascular Society Guidelines on Perioperative Cardiac Risk Assessment and Management for Patients Who Undergo Noncardiac Surgery. *The Canadian Journal of Cardiology*, 33(1):1732, 2017. ISSN 1916-7075. doi: 10.1016/j.cjca.2016.09.008.
- [2] "Writing Committee for the VISION Study Investigators" et al. Association of Postoperative High-Sensitivity Troponin Levels With Myocardial Injury and 30-Day Mortality Among Patients Undergoing Noncardiac Surgery. *JAMA*, 317(16):16421651, 2017. ISSN 1538-3598. doi: 10.1001/jama.2017.4360.
- [3] N. R. Smilowitz, G. Redel-Traub, A. Hausvater, A. Armanious, et al. Myocardial Injury After Noncardiac Surgery: A Systematic Review and Meta-Analysis. *Cardiology in Review*, 27(6):267273, 2019. ISSN 1538-4683. doi: 10.1097/CRD.0000000000000254.
- [4] "The Vascular Events in Noncardiac Surgery Patients Cohort Evaluation (VISION) Study Investigators", J. Spence, Y. LeManach, M. T. V. Chan, et al. Association between complications and death within 30 days after noncardiac surgery. *CMAJ*, 191(30):E830E837, 2019. ISSN 0820-3946, 1488-2329. doi: 10.1503/cmaj.190221.
- [5] T. Sheth, M. Chan, C. Butler, B. Chow, et al. Prognostic capabilities of coronary computed tomographic angiography before non-cardiac surgery: prospective cohort study. *BMJ (Clinical research ed.)*, 350:h1907, 2015. ISSN 1756-1833. doi: 10.1136/bmj.h1907.
- [6] T. H. Lee, E. R. Marcantonio, C. M. Mangione, E. J. Thomas, C. A. Polanczyk, et al. Derivation and Prospective Validation of a Simple Index for Prediction of

- Cardiac Risk of Major Noncardiac Surgery. *Circulation*, 100(10):10431049, 1999. doi: 10.1161/01.CIR.100.10.1043.
- [7] O. Ali. Genetics of type 2 diabetes. *World Journal of Diabetes*, 4(4):114123, 2013. ISSN 1948-9358. doi: 10.4239/wjd.v4.i4.114.
- [8] A. V. Khera and S. Kathiresan. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature Reviews Genetics*, 18(66):331344, 2017. ISSN 1471-0064. doi: 10.1038/nrg.2016.160.
- [9] I. Ntalla, S. Kanoni, L. Zeng, O. Giannakopoulou, et al. Genetic risk score for coronary disease identifies predispositions to cardiovascular and noncardiovascular diseases. *Journal of the American College of Cardiology*, 73(23):29322942, 2019. ISSN 0735-1097. doi: 10.1016/j.jacc.2019.03.512.
- [10] Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimers disease. 62. ISSN 0197-4580.
- [11] S. Theriault, R. Lali, M. Chong, J. L. Velianou, M. K. Natarajan, and G. Paré. Polygenic Contribution in Individuals With Early-Onset Coronary Artery Disease. *Circ Genom Precis Med*, 11(1):e001849, 2018. doi: 10.1161/CIRCGEN.117.001849.
- [12] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, et al. A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science (New York, N.Y.)*, 316(5830):14881491, 2007. ISSN 0036-8075. doi: 10.1126/science.1142447.
- [13] A. Helgadottir, G. Thorleifsson, A. Manolescu, S. Gretarsdottir, et al. A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science*, 316(5830):14911493, 2007. doi: 10.1126/science.1142842.

- [14] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, et al. Genomewide association analysis of coronary artery disease. *The New England Journal of Medicine*, 357(5): 443453, 2007. ISSN 1533-4406. doi: 10.1056/NEJMoa072366.
- [15] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(99):12191224, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0183-z.
- [16] S. J. Hahn, S. Kim, Y. S. Choi, J. Lee, and J. Kang. Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study. *EBioMedicine*, 86: 104383, December 2022. ISSN 2352-3964. doi: 10.1016/j.ebiom.2022.104383.
- [17] F. Padilla-Martínez, F. Collin, M. Kwasniewski, and A. Kretowski. Systematic review of polygenic risk scores for type 1 and type 2 diabetes. *International Journal of Molecular Sciences*, 21(5):1703, March 2020. ISSN 1422-0067. doi: 10.3390/ijms21051703.
- [18] E. Duschek, L. Forer, S. Schönherr, C. Gieger, et al. A polygenic and family risk score are both independently associated with risk of type 2 diabetes in a population-based study. *Scientific Reports*, 13(1):4805, March 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-31496-w.
- [19] F. Privé, J. Arbel, H. Aschard, and B. J. Vilhjálmsón. Identifying and correcting for misspecifications in gwas summary statistics and polygenic scores. *Human Genetics and Genomics Advances*, 3(4):100136, 2022. ISSN 2666-2477. doi: 10.1016/j.xhgg.2022.100136.
- [20] M. Elgart, G. Lyons, S. Romero-Brufau, N. Kurniansyah, et al. Non-linear machine

learning models incorporating snps and prs improve polygenic prediction in diverse human populations. *Communications Biology*, 5, 2022. doi: 10.1038/s42003-022-03812-z. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9395509/>.

- [21] J. Elliott, B. Bodinier, T. A. Bond, M. Chadeau-Hyam, et al. Predictive Accuracy of a Polygenic Risk Score Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA*, 323(7):636645, 2020. ISSN 0098-7484. doi: 10.1001/jama.2019.22241.
- [22] K. Thygesen, J. S. Alpert, A. S. Jaffe, B. R. Chaitman, et al. Fourth Universal Definition of Myocardial Infarction (2018). *Journal of the American College of Cardiology*, 72(18):22312264, 2018. ISSN 0735-1097. doi: 10.1016/j.jacc.2018.08.1038.
- [23] T. Sheth, M.K. Natarajan, V. Hsieh, N. Valettas, et al. Incidence of thrombosis in perioperative and non-operative myocardial infarction. *British Journal of Anaesthesia*, 120(4):725733, 2018. ISSN 00070912. doi: 10.1016/j.bja.2017.11.063.
- [24] N. J. Douville, I. Surakka, A. Leis, C. B. Douville, et al. Use of a Polygenic Risk Score Improves Prediction of Myocardial Injury after Non-cardiac Surgery. *Circulation. Genomic and precision medicine*, 13(4):e002817, 2020. ISSN 2574-8300. doi: 10.1161/CIRCGEN.119.002817.
- [25] K. Ruetzler, N. R. Smilowitz, J. S. Berger, P. J. Devereaux, et al. Diagnosis and Management of Patients With Myocardial Injury After Noncardiac Surgery: A Scientific Statement From the American Heart Association. *Circulation*, 144(19):e287e305, 2021. doi: 10.1161/CIR.0000000000001024.
- [26] D. M. Gualandro, C. A. Campos, D. Calderaro, P. C. Yu, et al. Coronary plaque rupture in patients with myocardial infarction after noncardiac surgery: Frequent

- and dangerous. *Atherosclerosis*, 222(1):191195, 2012. ISSN 00219150. doi: 10.1016/j.atherosclerosis.2012.02.021.
- [27] W. L. Duvall, B. Sealove, C. Pungoti, D. Katz, et al. Angiographic investigation of the pathophysiology of perioperative myocardial infarction. *Catheterization and Cardiovascular Interventions*, 80(5):768776, 2012. ISSN 1522-1946. doi: 10.1002/ccd.23446.
- [28] G. Landesberg, W. S. Beattie, M. Mosseri, A. S. Jaffe, and J. S. Alpert. Perioperative Myocardial Infarction. *Circulation*, 119(22):29362944, 2009. doi: 10.1161/CIRCULATIONAHA.108.828228.
- [29] I. Hanson, J. Kahn, S. Dixon, and J. Goldstein. Angiographic and clinical characteristics of type 1 versus type 2 perioperative myocardial infarction. *Catheterization and Cardiovascular Interventions*, 82(4):622628, 2013. ISSN 1522-1946. doi: 10.1002/ccd.24626.
- [30] M. C. Cohen and T. H. Aretz. Histological analysis of coronary artery lesions in fatal postoperative myocardial infarction. *Cardiovascular Pathology*, 8(3):133139, 1999. ISSN 1054-8807. doi: 10.1016/S1054-8807(98)00032-5.
- [31] T. W. Cade. Diabetes-Related Microvascular and Macrovascular Diseases in the Physical Therapy Setting. *Physical Therapy*, 88(11):13221335, 2008. ISSN 0031-9023. doi: 10.2522/ptj.20080008.
- [32] Beckman, J. A. and Creager, M. A. Vascular Complications of Diabetes. *Circulation Research*, 118(11):17711785, 2016. doi: 10.1161/CIRCRESAHA.115.306884.
- [33] S. Kim, J. Park, H. Kim, K. Yang, et al. Intraoperative Hyperglycemia May Be Associated with an Increased Risk of Myocardial Injury after Non-Cardiac Surgery in

Diabetic Patients. *Journal of Clinical Medicine*, 10(2222):5219, 2021. ISSN 2077-0383. doi: 10.3390/jcm10225219.

[34] A. P. Patel, M. Wang, Y. Ruan, S. Koyama, et al. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.*, 29(7):1793–1803, 2023.

[35] C. Sudlow, J. Gallacher, N. Allen, V. Beral, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):e1001779, 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779.

## 4.10 Supplementary Materials

### VISION Funding Sources

Canada

1. Canadian Institutes of Health Research - 7 grants
2. Heart and Stroke Foundation of Ontario - 2 grants
3. Academic Health Science Centres Alternative Funding Plan Innovation Fund Grant  
Ontario
4. Population Health Research Institute Grant
5. CLARITY Research Group Grant
6. McMaster University, Department of Surgery, Surgical Associates Research Grant
7. Hamilton Health Science New Investigator Fund Grant
8. Hamilton Health Sciences Grant
9. Ontario Ministry of Resource and Innovation Grant
10. Stryker Canada
11. McMaster University, Department of Anesthesiology 2 grants
12. Saint Josephs Healthcare, Department of Medicine 2 grants
13. Father Sean OSullivan Research Centre 2 grants

14. McMaster University, Department of Medicine 2 grants
15. Roche Diagnostics Global Office 5 grants
16. Hamilton Health Sciences Summer Studentships 6 grants
17. McMaster University, Department of Clinical Epidemiology and Biostatistics Grant
18. McMaster University, Division of Cardiology Grant
19. Canadian Network and Centre for Trials Internationally Grant
20. Winnipeg Health Sciences Foundation Operating Grant
21. University of Manitoba, Department of Surgery Research Grant 2 grants
22. Diagnostic Services of Manitoba Research Grant
23. Manitoba Medical Services Foundation Grant
24. Manitoba Health Research Council Grant
25. University of Manitoba, Faculty of Dentistry Operational Fund
26. University of Manitoba, Department of Anesthesia Grant
27. University Medical Group, Department of Surgery, University of Manitoba, start-up Fund

#### Australia

1. National Health and Medical Research Council Program Grant

2. Australian and New Zealand College of Anaesthetists Grant (13/008), Melbourne, Australia Colombia

#### Brazil

1. Projeto Hospitais de Excelência a Serviço do SUS (PROADI-SUS) grant from the Brazilian Ministry of Health in Partnership with Hcor (Cardiac Hospital Sao Paulo-SP)
2. National Council for Scientific and Technological Development (CNPq), grant from the Brazilian Ministry of Science and Technology

#### China

1. Public Policy Research Fund (CUHK-4002-PPR-3), Research Grant Council, Hong Kong SAR
2. General Research Fund (461412), Research Grant Council, Hong Kong SAR

#### Colombia

1. School of Nursing, Universidad Industrial de Santander
2. Grupo de Cardiología Preventiva, Universidad Autónoma de Bucaramanga
3. Fundación Cardioinfantil - Instituto de Cardiología
4. Alianza Diagnóstica S.A.

#### France

1. Université Pierre et Marie Curie, Département d'anesthésie Réanimation, Pitié-Salpêtrière,

Assistance Publique-Hôpitaux de Paris Grant India

2. St. John's Medical College and Research Institute Grant, Division of Clinical Research and Training Grant

Malaysia

1. University of Malaya Research Grant (RG302-14AFR)
2. University of Malaya, Penyelidikan Jangka Pendek Grant (PJP)

Poland

1. Polish Ministry of Science and Higher Education (NN402083939) Grant

South Africa

1. University of KwaZulu-Natal Grant

Spain

1. Instituto de Salud Carlos III
2. Fundació La Marató de TV3

United States

1. American Heart Association Grant
2. Covidien Grant United Kingdom
3. National Institute for Health Research (NIHR)

Centre		<i>n</i> participants		
		Overall	Control	Cases
2	Juravinski Hospital, Hamilton, Ontario	291	136	155
3	Hamilton General Hospital, Hamilton, Ontario	91	50	41
4	St Joseph's, Hamilton, Ontario	113	64	49
11	Victoria Hospital, London, Ontario	11	3	8
Total		506	253	253

Consortium	Trait	Sample Size (n)
Coronary artery disease (CAD)	Cardiogram/C4D + UK Biobank	332 477
Type II diabetes (T2D)	MVP + UK Biobank	1 114 458
Stroke + stroke-related phenotypes	GIGASTROKE	~1 500 000
HbA1c	MAGIC + 1000Genomes	123 665
Atrial fibrillation (AF)	AFGen + deCODE + UK Biobank	~1 000 000
Body mass index (BMI)	GIANT + UK Biobank	681 725
High density lipoproteins (HDL)	MVP	210 967
Low density lipoproteins (LDL)	MVP	215 196
Triglycerides (TG)	MVP	211 491
Systolic blood pressure (SBP)	MVP	220 501
Diastolic blood pressure (DBP)	MVP	220 501
Cardiomyopathy	GBMI	745 451
Heart failure	GBMI	821 198
Venous thromboembolism (VTE)	GBMI	747 540

Table S4.1: List of genome-wide summary statistics consortium summary statistics.

## a. CAD PRS

Sex group	Age group	n (MINS/no MINS)	Model 1: CAD PRS	Model 2: RCRI	Model 3: CAD PRS + RCRI
All	All	253	0.63 (0.58 - 0.68)	0.70 (0.66 - 0.75)	0.72 (0.67 - 0.76)
	45-64	57	0.65 (0.54 - 0.76)	0.82 (0.74 - 0.90)	0.83 (0.76 - 0.91)
	65-74	92	0.60 (0.52 - 0.68)	0.65 (0.57 - 0.73)	0.67 (0.63 - 0.76)
	≥ 75	104	0.65 (0.57 - 0.72)	0.69 (0.62 - 0.76)	0.70 (0.48 - 0.75)
Men	All	141	0.63 (0.57 - 0.70)	0.72 (0.66 - 0.78)	0.72 (0.67 - 0.78)
	45-64	33	0.64 (0.58 - 0.71)	0.84 (0.74 - 0.94)	0.85 (0.76 - 0.94)
0.84 (0.74 - 0.94)	65-74	52	0.65 (0.55 - 0.76)	0.69 (0.59 - 0.79)	0.70 (0.60 - 0.80)
	≥ 75	56	0.63 (0.52 - 0.73)	0.68 (0.58 - 0.77)	0.69 (0.59 - 0.78)
	All	141	0.63 (0.56 - 0.70)	0.68 (0.61 - 0.75)	0.71 (0.64 - 0.78)
Women	All	141	0.63 (0.56 - 0.70)	0.68 (0.61 - 0.75)	0.71 (0.64 - 0.78)
	45-64	33	0.69 (0.52 - 0.85)	0.82 (0.69 - 0.94)	0.86 (0.76 - 0.97)
	65-74	52	0.55 (0.41 - 0.68)	0.60 (0.47 - 0.74)	0.64 (0.51 - 0.77)
	≥ 75	56	0.67 (0.56 - 0.77)	0.70 (0.60 - 0.80)	0.70 (0.60 - 0.80)

## b. T2D PRS

Sex group	Age group	n (MINS/no MINS)	Model 1: T2D PRS	Model 2: RCRI	Model 3: T2D PRS + RCRI
All	All	253	0.64 (0.60 - 0.69)	0.70 (0.66 - 0.75)	0.72 (0.67 - 0.76)
	45-64	57	0.69 (0.58 - 0.79)	0.82 (0.74 - 0.90)	0.84 (0.76 - 0.91)
	65-74	92	0.62 (0.54 - 0.70)	0.65 (0.57 - 0.73)	0.68 (0.60 - 0.75)
	≥ 75	104	0.64 (0.57 - 0.72)	0.69 (0.62 - 0.76)	0.69 (0.62 - 0.76)
Men	All	141	0.64 (0.58 - 0.71)	0.72 (0.66 - 0.78)	0.73 (0.67 - 0.79)
	45-64	33	0.65 (0.51 - 0.79)	0.84 (0.74 - 0.94)	0.85 (0.76 - 0.94)
	65-74	52	0.64 (0.53 - 0.75)	0.69 (0.59 - 0.79)	0.69 (0.59 - 0.79)
	≥ 75	56	0.64 (0.53 - 0.74)	0.68 (0.58 - 0.77)	0.68 (0.59 - 0.78)
Women	All	141	0.65 (0.58 - 0.72)	0.68 (0.61 - 0.75)	0.71 (0.64 - 0.77)
	45-64	33	0.73 (0.58 - 0.88)	0.82 (0.69 - 0.94)	0.82 (0.70 - 0.94)
	65-74	52	0.59 (0.46 - 0.72)	0.60 (0.47 - 0.74)	0.65 (0.53 - 0.78)
	≥ 75	56	0.65 (0.55 - 0.76)	0.70 (0.60 - 0.80)	0.69 (0.59 - 0.79)

## c. HbA1c PRS

Sex group	Age group	n (MINS/no MINS)	Model 1: HbA1c PRS	Model 2: RCRI	Model 3: HbA1c PRS + RCRI
All	All	253	0.66 (0.61 - 0.70)	0.70 (0.66 - 0.75)	0.73 (0.68 - 0.77)
	45-64	57	0.69 (0.59 - 0.79)	0.82 (0.74 - 0.90)	0.85 (0.78 - 0.92)
	65-74	92	0.61 (0.53 - 0.69)	0.65 (0.57 - 0.73)	0.66 (0.59 - 0.74)
	≥ 75	104	0.68 (0.61 - 0.75)	0.69 (0.62 - 0.76)	0.72 (0.66 - 0.79)
Men	All	141	0.66 (0.60 - 0.72)	0.72 (0.66 - 0.78)	0.74 (0.68 - 0.80)
	45-64	33	0.67 (0.57 - 0.77)	0.84 (0.74 - 0.94)	0.88 (0.80 - 0.96)
	65-74	52	0.64 (0.53 - 0.74)	0.69 (0.59 - 0.79)	0.69 (0.59 - 0.79)
	≥ 75	56	0.67 (0.57 - 0.77)	0.68 (0.58 - 0.77)	0.72 (0.63 - 0.81)
Women	All	141	0.66 (0.59 - 0.73)	0.68 (0.61 - 0.75)	0.72 (0.65 - 0.79)
	45-64	33	0.71 (0.56 - 0.87)	0.82 (0.69 - 0.94)	0.82 (0.70 - 0.94)
	65-74	52	0.58 (0.44 - 0.71)	0.60 (0.47 - 0.74)	0.65 (0.52 - 0.78)
	≥ 75	56	0.69 (0.59 - 0.79)	0.70 (0.60 - 0.80)	0.73 (0.62 - 0.83)

Table S4.2: Discriminative capacity using c-statistic (with 95% confidence intervals) in conditional logistic regressions for MINS within 30 days after surgery among participants in the T2D PRS. Figure S2a. shows discriminative capacity of CAD PRS, S2b. for T2D PRS and S2c. for HbA1c PRS.

Trait	Age Group	Sex Group	p-value
CAD	45-64	All	0.44
CAD	45-64	Male	0.052
CAD	45-64	Female	0.0057
CAD	65-74	All	0.10
CAD	65-74	Male	0.71
CAD	65-74	Female	0.58
CAD	75+	All	0.65
CAD	75+	Male	0.56
CAD	75+	Female	0.76
CAD	All	Male	0.66
CAD	All	Female	0.022
HbA1c	45-64	All	0.31
HbA1c	45-64	Male	0.0022
HbA1c	45-64	Female	0.064
HbA1c	65-74	All	0.41
HbA1c	65-74	Male	0.60
HbA1c	65-74	Female	0.63
HbA1c	75+	All	0.030
HbA1c	75+	Male	0.94
HbA1c	75+	Female	0.42
HbA1c	All	Male	0.13
HbA1c	All	Female	0.038
T2D	45-64	All	0.44
T2D	45-64	Male	0.024
T2D	45-64	Female	0.049
T2D	65-74	All	0.12
T2D	65-74	Male	0.66
T2D	65-74	Female	0.69
T2D	75+	All	0.84
T2D	75+	Male	0.54
T2D	75+	Female	0.92
T2D	All	Male	0.48
T2D	All	Female	0.061

Table S4.3: DeLong analyses to discriminative capacity significance for PRS within each subset.

a. CAD PRS

	NRI	p-value	95% CI
NRI	0.087	0.33	-0.09 - 0.26
NRI for MINS	0.051	0.41	-0.072 - 0.17
NRI for no MINS	0.036	0.57	-0.062 - 0.16

b. T2D PRS

	NRI	p-value	95% CI
NRI	0.20	0.025*	0.024 - 0.37
NRI for MINS	0.059	0.041	-0.072 - 0.17
NRI for no MINS	0.14	0.019*	0.024 - 0.27

c. HbA1c PRS

	NRI	p-value	95% CI
NRI	0.18	0.040*	0.0083 - 0.36
NRI for MINS	0.11	0.067*	-0.0078 - 0.24
NRI for no MINS	0.067	0.28	-0.056 - 0.19

Table S4.4: Discriminative capacity using Net Reclassification Improvement (NRI) in logistic regressions for MINS within 30 days after surgery among participants for RCRI with the addition of PRS.

## Chapter 5

# Performance of polygenic risk score methodologies in the absence of external GWAS summary statistics

Ann Le<sup>1,2</sup>, Shihong Mao<sup>1</sup>, Angelo Canty<sup>3</sup>, Yanran Li<sup>1,6</sup>, Alice Man<sup>1,2,4</sup>, Keona Pang<sup>1,2</sup> & Guillaume Paré<sup>1,2,3,5,6,7,8</sup>

1. Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada.
2. Department of Medical Sciences, McMaster University, Faculty of Health Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.

3. Department of Mathematics and Statistics, McMaster University, Faculty of Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.
4. Michael G. DeGroot School of Medicine, McMaster University, Faculty of Health Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.
5. Department of Biochemistry and Biomedical Sciences, McMaster University, Faculty of Health Sciences, 1280 Main Street West, Hamilton ON L8S 4K1, Canada.
6. Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, 237 Barton Street East, Hamilton, ON L8L 2X2, Canada
7. Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroot School of Medicine, 1280 Main Street West, Hamilton ON L8S 4K1, Canada
8. Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street West, Hamilton ON L8L 4K1, Canada.

## 5.1 Forward

Polygenic risk scores (PRS) are a relatively new method of risk prediction, as a follow-up to genome-wide association studies (GWAS) which gained relevance in the recent decades. PRS exist as quantitative measures of individual’s genetic susceptibility towards a given phenotype. They have great potential for improving current clinical predictor tools and precision medicine. There has been great progress with PRS throughout the years, incorporating numerous techniques involving advanced statistical and machine learning methods. While there have been great improvements in PRS development over the years, current methodologies rely on the existence of an external GWAS that corresponds to the desired outcome. Moreover, the issue of a completely unavailable GWAS has not been thoroughly investigated. Thus, this study was proposed to assess the performance of various PRS methods without an available external trait-specific GWAS. Additionally, we developed an innovative approach employing Multi Adaptive Regression Splines (MARS) to calculate PRS.

The analyses focussed on three distinct PRS methodologies with the third-party method LDpred2 acting as the baseline, single-trait PRS. Namely, the three methods are a baseline, single-trait PRS, PRS<sub>multi</sub>: a multi-trait PRS trained used elastic net regression, and EX-TERR: a multi-trait PRS trained using Multivariate Adaptive Regression Splines (MARS). The PRS were trained and tested on 408,160 British related participants from the UK Biobank. This method uses principal component analysis (PCA) techniques on 5 Mb genotype blocks to generate separate rotated matrices. This is done to reduce dimensionality of the dataset based on a variance threshold. Due to a built-in cross-validation step for rotated genotype blocks, EX-TERR does not require an initial train-test split for the sample. This could be of further benefit when no trait-specific GWAS are available,

as a robust PRS can be dependent on GWAS powered by larger samples. We simulated the situation of a lack of corresponding GWAS by masking (omitted) specific GWAS with traits which are directly or closely related to target outcome.

A drastic decrease in performance was observed when the corresponding external GWAS was masked. All outcomes with directly matched external GWAS showed a decrease in performance. Masked PRS for continuous traits showed decreased performance reflecting lower adjusted  $R^2$  of up to 98.7%, while masked PRS for dichotomous traits showed a decrease in OR of up to 41.8%. The independent performance of each PRS methodology was also observed. While no method universally outperformed the rest, multi-PRS methods (PRS<sub>multi</sub> and EX-TERR) generally improved performance in most outcomes, with PRS<sub>multi</sub> and EX-TERR improving predictive accuracy in 71.0% and 62.3% of outcomes respectively. The novel EX-TERR method performed comparably to the pre-existing PRS<sub>multi</sub> methodology. Under the masked condition, the outcomes most strongly associated with PRS methods were urate (PRS<sub>multi</sub> adj  $r^2 = 0.039$ , EX-TERR adj.  $r^2 = 0.057$ ) and type II diabetes (T2D) (PRS<sub>multi</sub> OR per SD = 1.40, EX-TERR OR per SD = 1.49). In conclusion, we presented an overview of the leading PRS methods and introduced our novel approach to address the issue of calculating PRS in the absence of an external GWAS.

This manuscript is in progress for submission. Guillaume Paré conceptualized and designed the study. Ann Le designed analysis plan, ran the EX-TERR pipeline, conducted statistical analyses, and wrote the manuscript. Shihong Mao initiated the EX-TERR pipeline and also conducted statistical analyses. Angelo Canty provided instruction and detailed insights into underlying statistical concepts used within the model. All authors contributed to the interpretation of findings and to the critical reading and revision of the manuscript.

## 5.2 Abstract

There is strong interest in polygenic risk scores (PRS), due to their capabilities to predict risk by solely using genotype information. PRS have been shown in many cases to have a predictiveness similar to clinical risk factors. However, a limitation of PRS is their reliance on genome-wide association study (GWAS) data, which might not always be available for the target trait. This study aims to determine the most effective PRS methodology to utilize in this context. We propose a new methodology called EX-TERR (EXternal Technique with Earth Regional Regression) which utilizes Multi-Adaptive Regression Splines (MARS) and single-trait PRS as training data to develop target outcome PRS. We also adapt established approaches, baseline PRS, which selects the most predictive existing single-trait PRS, and PRS<sub>multi</sub>, an elastic net regression method integrating multiple single-trait PRS, to approximate the target outcome PRS within UK Biobank participants of British ancestry ( $n = 408,160$ ). Target outcome PRS were generated using single-trait LDpred2 PRS for 61 traits with external (non-UKB) GWAS summary statistics and 69 UKB outcomes. External GWAS corresponding to each target outcome were masked and output PRS were subsequently compared for discrimination and calibration. 93.7% of the 207 PRS validated were significantly associated with the 69 outcomes ( $p < 0.05$ ). Predictive accuracy was shown to be drastically decreased when the GWAS matching the outcome was masked. Relative to the best performing single-trait baseline PRS, PRS<sub>multi</sub> and EX-TERR was able to improve predictive accuracy in 71.0% and 62.3% of outcomes when target trait GWAS was masked. The reduction in performance was substantially greater for continuous traits compared to dichotomous traits. The average relative decrease in performance from unmasked to masked traits was 86.4% in continuous traits and 19.1% in dichotomous traits across the three PRS methodologies tested. When an external GWAS for a specific

outcome is not available, the performance of PRS was significantly diminished across all tested methods. While no method was universally best, methods integrating multiple external GWAS had the best predictive performance for most outcomes. Overall, significant reduction in predictive accuracy is to be expected in the absence of an external GWAS for the outcome and current methods only partially address this challenge.

### 5.3 Condensed Abstract

While development of polygenic risk scores (PRS) have improved through the years, there are a lack of studies which addresses the best approach to take when a corresponding external genome-wide association study (GWAS) is not available for PRS generation. We tested several existing leading PRS methodologies, as well as developed a new methodology coined “EX-TERR” (EXternal Technique with Earth Regional Regression) to simulate PRS predictiveness in situations where the external GWAS corresponding to the target trait does not exist. EX-TERR is based on utilizing adaptive regression splines (MARS) to train genetic variant data, and uniquely does not require an initial train-test split for the original participant sample. Overall, the lack of availability in external GWAS appears to impede PRS performance. Additionally, analyses further revealed no PRS method consistently outperformed the rest, suggesting there is no single methodology that is generalizable across all outcomes.

**KEYWORDS:** polygenic risk scores; genetic risk prediction; genome-wide association studies; genetics; genetic epidemiology

## 5.4 Introduction

Polygenic risk scores (PRS) are defined as quantitative measures of an individual's genetic predisposition to a specified disease or trait, derived from the cumulative effect of multiple genetic variants across the genome. PRS have potential for the early detection, intervention, and prevention of diseases, with evidence that genetic influences on common diseases are at least as significant as environmental influences for certain traits[1, 2, 3, 4, 5]. Since genotypes can be obtained early in life, PRS are of particular interest for predicting disease of high heritability. Coronary artery disease (CAD), for example, has a relatively strong heritability, estimated to be as high as 57%, and is often associated with early onset cases due to a strong genetic predisposition[6, 7, 8, 9, 10, 11]. The development of PRS is motivated by the concept that, while individual single nucleotide polymorphisms (SNPs) may have minor effects alone, their accumulated effect may notably contribute to a given trait[11, 12, 13]. Typically, PRS are computed as weighted sums of alleles across a large number of SNPs associated with a trait. Over the years, various novel approaches incorporating advanced statistical methods and machine learning have been employed to enhance the predictive accuracy of PRS (e.g. LDpred2[14], LASSOSUM2[15], PRSice-2[16], PRS-CS[17]).

While novel PRS methodologies have achieved noteworthy improvements in risk prediction relative to traditional allelic sum method, they all rely on the existence of an external genome-wide association study (GWAS) summary statistic corresponding to the desired outcome. In these methods, GWAS identify associations between SNPs and specific outcomes, which are then used to assign weights to individual variants. For instance, creating a PRS for CAD would require a corresponding CAD GWAS summary statistic, such as those from the CARDIOGRAM/C4D consortium[18]. In many instances, external GWAS

summary statistics directly corresponding to the particular trait or disease might not be available. Previously, studies have demonstrated that PRS performance can vary greatly depending on the quality, ancestry representation, power and other aspects of GWAS. However, the issue of a completely unavailable GWAS has not been extensively investigated. For example, more than 85% of GWAS are conducted within European populations, and evidence has shown that PRS generated for a specific ancestry are not applicable to other populations[19, 20]. More notably, the issue of a completely unavailable GWAS has not been extensively investigated. While this can be potentially mitigated by using external GWAS summary statistics for risk factors related to the trait of interest, availability of the trait-specific GWAS can be crucial for investigation of distinct genetic architectures driven by trait-specific genetic pathways that do not overlap with known risk factors. This can be of particular importance when heritability estimates are unavailable, as it is entirely possible that genetics may only play a minor role in an individual's trait susceptibility, rendering the PRS futile.

This indicates the need for a PRS methodology that does not depend on the availability of an external GWAS corresponding trait. In this study, we propose adapting several previously established methods and implementing a novel approach to address the absence of external GWAS, aiming to identify the most effective strategy under these circumstances. Within a cohort of 408,160 British related participants from the UK Biobank (UKB), we employed three different PRS methodologies to create and compare three types of PRS in the absence of external PRS data: 1) a baseline, single-trait PRS, 2) PRS<sub>multi</sub>: a multi-trait PRS trained using elastic net regression, and 3) EX-TERR: a multi-trait PRS trained using Multivariate Adaptive Regression Splines (MARS). A multi-trait PRS refers to in which multiple traits are inputted into a machine learning algorithm to reassess the strength of association for each trait with the outcome based on the chosen regression

model. EX-TERR is a novel concept for PRS involving the integration of MARS to combine multiple, separate rotated genotype matrices generated from principal component analysis (PCA) techniques within blocks of 5,000 variants. PCA is used to reduce dimensionality by minimizing the number of closely related variant in the MARS model, based on a variance threshold. The scenario of lacking a corresponding GWAS is simulated by masking (omitting) specific traits that directly correspond to or are closely related to the target outcome. The overall objective is to evaluate and compare the performance of both new and old PRS methodologies in situations where the outcome trait of interest does not have a corresponding external GWAS summary statistic.

## 5.5 Methods

### 5.5.1 Study Populations

The study cohort is a subset of the UK Biobank (UKB), a large epidemiological study consisting of over 500,000 individuals aged 40 to 69 from across the United Kingdom[21]. The resource contains extensive data regarding participants characteristics and measurements including demographics, health diagnoses, physical measurements and lifestyle factors. The UKB also contains phenotypic information, including patient information linked to the 9th and 10th editions of the International Classification of Diseases (ICD-9, ICD-10) and Classification of Interventions and Procedures, version 4 of the Office of Population, Censuses and Surveys (OPCS-4). Variants included those found in the Haplotype Reference Consortium and 1000 Genomes panels from release version 3 of the UKB data. Variants had no deviation from Hardy-Weinberg equilibrium ( $p > 1 \times 10^{-10}$ ). Further SNP exclusion criteria included minor allele frequency (MAF)  $< 0.001$ , SNPs with low imputation quality (INFO score  $< 0.30$ ), and ambiguous or duplicated SNPs. For our analyses, we used the subset of

related British participants ( $n = 408,160$ ). Participants were additionally excluded based on substantial genotype missingness ( $>5\%$  missing genotype), elevated ancestry-specific heterozygosity, incongruent genetic ancestry, sex chromosome aneuploidy, and inconsistencies between reported and genetic sex. In total, 69 outcomes were derived from the UKB. Both continuous and dichotomous outcomes were included, such as blood biomarkers and disease status (Supplementary Table S5.1). Phenotypes were defined through various identifiers including UKB phecodes, ICD-10 codes, OPCS-4 codes, self-reported data or UKB fields with algorithmically-defined outcomes. Overall, phenotypes were categorized into thirteen distinct groups for further clarification. These included seven continuous trait categories (anthropometrics, lipids & lipoproteins, liver function, endocrine function, renal function, inflammatory biomarkers and electrolytes) and six dichotomous trait categories (cardiovascular conditions, metabolic conditions, respiratory conditions, cancers, lifestyle factors and others). Both incident and prevalent cases were considered as positive outcomes.

### **5.5.2 Genome-wide association study (GWAS) data**

A comprehensive internal database containing pre-downloaded GWAS summary statistics across a wide range of traits and ancestries was available for analysis. 61 external GWAS summary statistics from various consortia were utilized (Supplementary Table S5.2). The GWAS data had to contain no UKB participant to prevent circularity in the analysis. The GWAS selection criteria required that summary statistics be either the most recently updated from the consortium and/or contain the largest number of variants. Furthermore, the summary statistics had to include at least 5,000,000 single nucleotide polymorphisms (SNPs). A subset of variants common to the UKB genotype and all external summary statistics was created. Ambiguous SNPs were removed. To simulate the context of the lack

of a corresponding external GWAS, a GWAS would be removed (masked) from the PRS construction if they directly matched the corresponding outcome or had clinical evidence of being directly correlated. For example, given the outcome of CAD, the corresponding GWAS for CAD from CARDIOGRAM/C4D was masked from the analysis. Additionally, the GWAS for heart failure from GBMI was also masked for CAD due to the large proportion of heart failure cases directly caused by CAD resulting in high correlation between outcomes. To further aid masking decisions, correlation was observed between all phenotypic outcomes. Outcomes with correlation  $R^2$  higher than 0.7 also had their corresponding GWAS masked (Supplementary Table S5.3). Each of the 69 outcomes thus had a specific set of masked GWAS that were excluded during PRS calculations (Supplementary Table S5.4).

### 5.5.3 Polygenic Risk Score Methodologies

External summary statistics for 61 different traits and 69 UKB outcomes were used to train and derive PRS across the 408,160 participants. Of the total 69 outcomes, there were 35 continuous outcomes and 34 dichotomous outcomes. Continuous outcomes were further divided into seven categories: anthropometrics (5), lipids lipoproteins (7), liver function (9), endocrine function (8), renal function (2), inflammatory biomarkers (2) and electrolytes (2). Dichotomous outcomes were divided into six categories: cardiovascular conditions (14), metabolic conditions (5), respiratory conditions (3), cancers (3), lifestyle factors (2) and others (7). To ensure equivalent comparison across all three PRS methodologies, LDpred2 was used to generate the baseline single-trait PRS for all methods. LDpred2 is a widely used PRS method, incorporating Bayesian statistics with an external panel for linkage disequilibrium (LD) reference[17, 22, 23]. The three PRS methodologies used for comparison in this study are: 1) a baseline PRS: the most predictive single trait LDpred2

score, 2)  $\text{PRS}_{\text{multi}}$ : a multi-trait score with multiple LDpred2 inputs trained using elastic net regression and 3) EX-TERR: a multi-trait score with multiple LDpred2 inputs trained using MARS (Figure 5.1). To create single-trait PRS, variant associations are derived from external GWAS, reweighted using LDpred2, and then applied to each variant before being summed. For  $\text{PRS}_{\text{multi}}$ , multiple baseline PRS are combined as input into an elastic net regression model using the “glmnet” function in R ( $\alpha = 0.5$ , with  $\lambda$  determined via built-in cross-validation) [24]. The elastic net model determined the importance of each separate single-trait PRS on the outcome through penalized regression, essentially reweighing each PRS before combining them into a single final score.

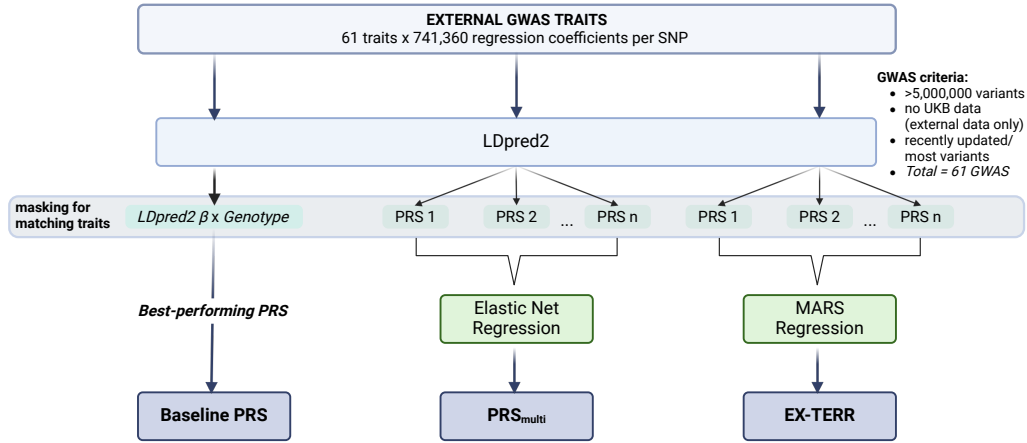


Figure 5.1: **Description of PRS methodologies with masking process.** LDpred2 is used to create single-trait PRS applied to each of these methods. From left to right, the PRS techniques being compared are 1) baseline: the best-performing single-trait PRS, 2)  $\text{PRS}_{\text{multi}}$ : a multi-trait PRS based in elastic net regression and 3) EX-TERR: a multi-trait PRS based in Multi Adaptive Adaptive Regression Splines (MARS). Masking (exclusion of GWAS directly matching to outcome) is performed across all methods, to simulate the context of no external GWAS corresponding to the target outcome. The GWAS used in the analysis were required to meet the following criteria: inclusion of at least 5,000,000 SNPs, exclusion of UKB data, and being the most up-to-date summary statistics available.

#### 5.5.4 Internal UKB Genotype Association

In this study, we define external regression coefficients (genetic associations) as those obtained from GWAS summary statistics. Alternatively, internal UKB regression coefficients refer to those derived from regression analyses between UKB genotype data and the 69 UKB-derived phenotypes. REGENIE[25] was used to compute association summary statistics for the UK Biobank (UKB) dataset. The linear and logistic functions were applied with default settings and a batch size of 1000. Adjustments were made for age, sex and the first 10 principal components (PCs) to account for population structure, with the PCs being pre-calculated. Overall, internal UKB association correlation between outcomes were conducted for 69 UKB genotype predictors and phenotype outcomes, employing linear regression for continuous traits and logistic regression for dichotomous traits.

#### 5.5.5 EX-TERR Pipeline

The EX-TERR pipeline is a supervised PRS methodology using Multivariate Adaptive Regression Splines models (MARS; Figure 5.2). MARS is a non-parametric regression technique which allows for the construction of a non-linear regression model and fitting data without assuming a predetermined form[26]. The method is a greedy algorithm, operating in a forward pass followed by a backward pass to optimize model fit while avoiding overfitting. The general MARS model is based on the form:

$$\hat{f}(x) = \sum_{i=1}^k c_i B_i(x) \quad (5.5.1)$$

where  $c_i$  is a constant and  $B_i(x)$  is a basis function. Basis functions consist of hinge functions, which takes on the form  $\max(0, x - c)$ . These hinge functions create knots or kinks,

which are characteristic features in the visualization of a MARS model. A basis function can also be a product of two or more hinge functions, indicating interactions between variables. Notably, MARS penalizes excessive addition of knots to prevent overfitting.

The earth package version 5.3.3 in R version 3.6.0 and 4.4.0 was used to implement MARS. Default earth parameters were used. Variable importance is assessed using residual sum-of-squares (RSS). MARS outputs a set of regression coefficients trained on rotated LDpred2 UKB weights as outcomes, which are then applied to the rotated genotype matrices from the participant validation sample to generate the final EX-TERR PRS.

### **5.5.6 Principal Component Analysis (PCA) Technique for Dimension Reduction**

Techniques from principal component analyses (PCA) are adapted to project genotype and coefficient matrices through a rotation matrix intended for principal component (PCs) [27]. A PCA rotation matrix was used to transform blocks containing genotypic information into its rotated form. Rotation matrix columns can be subsequently excluded based on standard deviation (SD) to reduce dimensionality and potential noise. First, all regression coefficients derived from the associations between external GWAS and UKB genotypes with the outcome are converted into LDpred2 weights. Next, LDpred2 weight matrices are standardized such that each matrix column has a mean of 0 and SD of 1. Then, the LDpred2 weights are divided into matching contiguous non-overlapping blocks approximately 5,000 SNPs in length. Finally, each block is rotated using a 5,000 x 5,000 rotation matrix derived from the training set UKB genotype matrix ( $V$ ). Since the original genotype matrix is standardized, a rotated matrix column with an SD of 1 is considered to capture information equivalent to that of a single genetic variant. The SD threshold for pruning was empirically determined using pairwise correlation between outcomes as reference (Supplementary Table

S5.3). Ultimately, an SD of 1.0 ( $\sigma^2 = 1.0$ ) was decided as the filtering threshold. The resulting rotated matrices conserve genetic information provided by variants.

### 5.5.7 Multiple Train-Test Split: Participant & Genotype Levels

Originally, initial participant sample is divided into an 20/80 train-test split. However, EX-TERR also implements a 5-fold cross validation on genotypic information, thus does not require this initial train-test split. Rotated UKB LDpred2 weight groups remaining after SD threshold filtering are divided into approximately five independent groups, after which each group alternates to become the test set. This ensures training on independent groups of variants, regardless of whether the initial participant train-test split was implemented. Thus, this feature negates the need for the initial participant train-test split when using EX-TERR, potentially allowing for better training and increased power. Thus, while baseline PRS and PRS<sub>multi</sub> were performed on the initial 20/80 train-test split, EX-TERR is performed on all available participants and trained on genotypes.

### 5.5.8 Discrimination & Calibration Tests

For discrimination, Net Reclassification Improvement (NRI) indices were computed, using a base model that included age, sex, and 10 genetically derived principal components. Calibration tests were also performed to ensure goodness-of-fit, as a comparison of how closely predicted and actual observations align in the risk model[28]. In dichotomous outcomes, the Hosmer-Lemeshow (HL) test was used to compare the observed and expected outcome frequencies within PRS decile bins using a chi-square test[29]. The default bin number of 10 (8 degrees of freedom) was used for HL tests. For continuous outcomes, the root mean squared error (RMSE) was calculated. RMSE assesses the mean difference between observed and predicted observations for continuous outcomes[30].

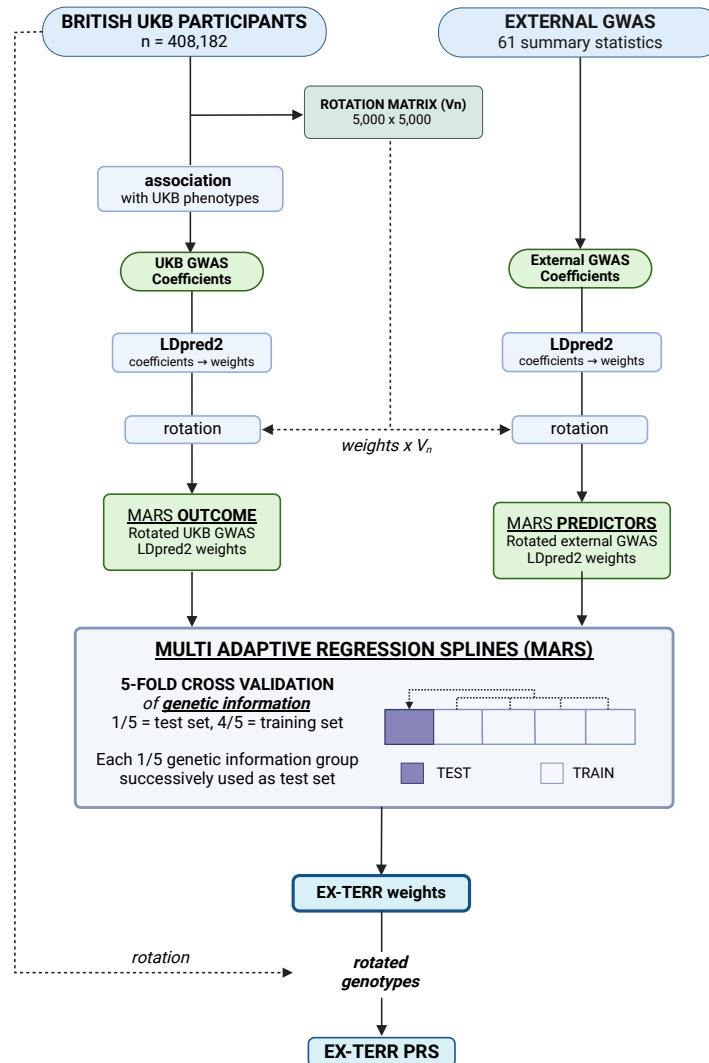


Figure 5.2: Overview of EX-TERR pipeline. A total of 408,182 related participants from the UK Biobank (UKB) were used to create EX-TERR PRS. Regression coefficients are obtained from external GWAS, and linear or logistic regression between UKB genotypes and outcomes. These coefficients are converted into polygenic risk score (PRS) weights using LDpred2. Genotypes are divided into approximately 5,000 single-nucleotide polymorphism (SNP) blocks, and within these blocks, rotations are performed using rotational matrix ( $V_n$ ) derived from the UKB training set genotypes (coefficient matrix  $\times V_n$ ). The predictor for the MARS regression is the rotated external GWAS LDpred2 weights, and the outcome are rotated UKB LDpred2 weights. A secondary train/test split is conducted through 5-fold cross-validation in the predictors. This process allows the MARS weights to be applied to both training and/or validation sets for PRS generation.

## 5.6 Results

Characteristics of the 408,182 participants included are shown in Supplementary Table S5.1. The average age of participants was 71.7 years, with 220,467 (54.1%) females and 187,513 (45.9%) males. The most prevalent outcomes were hypertension (32.26%) and suspected or other cancer (15.16%). First, single-trait PRS performance was compared with and without masking for all outcomes considered to have a directly matching external GWAS. For example, the T2D DIAGRAM consortium GWAS is excluded (masked) from all PRS calculations for diabetes outcomes (refer to Supplementary Table S5.5 for the complete list of outcomes considered to have matching GWAS). When the directly matched external GWAS is excluded from the analysis, predictive accuracy significantly declined for all 20 tested outcomes (Figure 5.3). A clear reduction in adjusted  $R^2$  in continuous outcomes and odds ratio (OR) in dichotomous outcomes was observed across all traits, though the extent of decrease varied. Among continuous traits, high-density lipoprotein (HDL) had the greatest decrease in PRS performance (98.7% decrease in adj.  $R^2$ ), while cancer (with breast cancer) had the greatest decrease for dichotomous traits (41.8% decrease in OR).

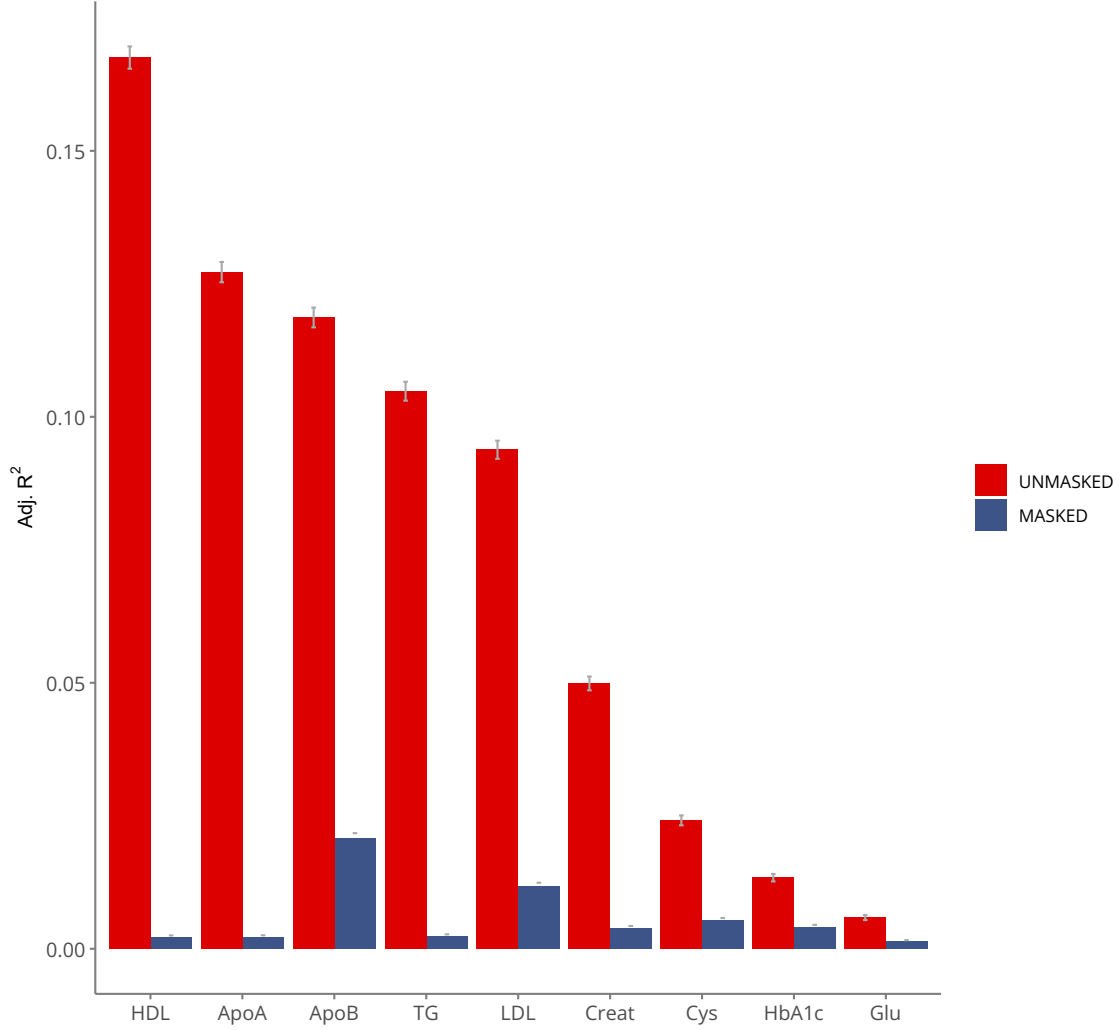
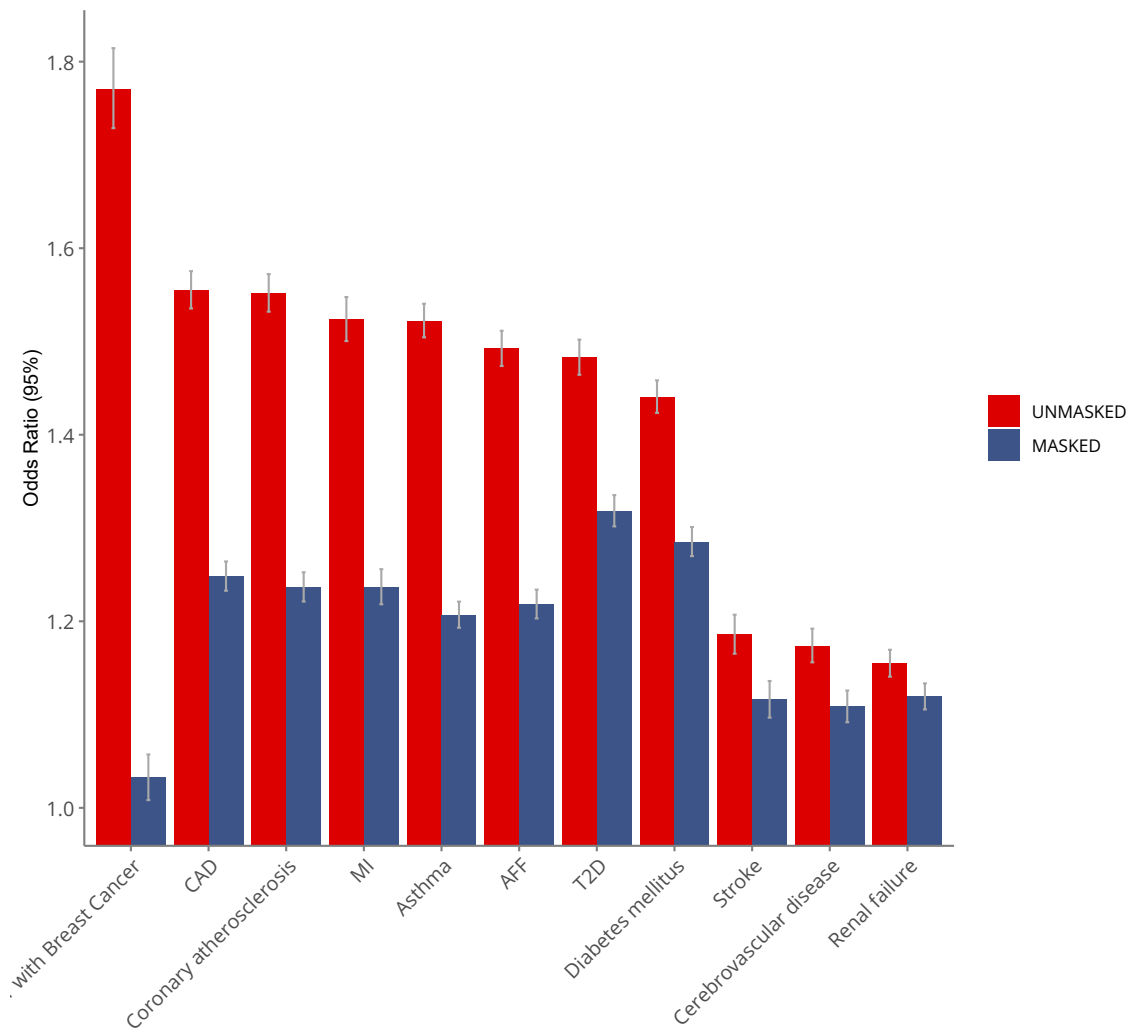


Figure 5.3: **Single-trait PRS performance with (unmasked) and without (masked) its corresponding external GWAS. a) Continuous outcomes.** Comparative performance of single-trait, baseline PRS with and without masking of the matching GWAS corresponding to the outcome. A list of outcomes with matching GWAS traits is presented in Supplementary Table S5.5.



**b) Dichotomous outcomes.** Comparative performance of single-trait, baseline PRS with and without masking of the matching GWAS corresponding to the outcome. A list of outcomes with matching GWAS traits is presented in Supplementary Table S5.5.

Next, all three PRS methods were observed under the masking condition (Figure 5.1). Three approaches best-performing single-trait baseline PRS, PRS<sub>multi</sub>, and EX-TERR PRS were applied to 69 UKB-derived outcomes (35 continuous and 34 dichotomous), generating a total of 207 PRS (69 outcomes  $\times$  3 methods) for validation. Of the PRS tested, all were PRS significantly associated ( $p < 0.05$ ) with their corresponding outcome with the exception of 6 PRS<sub>multi</sub> scores (8.7%) and 7 EX-TERR PRS scores (10.1%). PRS<sub>multi</sub> and EX-TERR had a greater predictive accuracy in 71.0% and 62.3% of outcome traits, respectively, relative to the best performing single-trait baseline PRS. The relative performance of each method was evaluated based on the outcome type, represented by the proportion of outcomes where the methodology yielded the highest adjusted  $R^2$  for continuous outcomes or the highest odds ratio (OR) for dichotomous outcomes (Table 5.1). Cumulatively, the multi-PRS methods (PRS<sub>multi</sub> and EX-TERR) outperformed the single-trait baseline PRS for 60% of continuous outcomes, 82.4% of dichotomous outcomes, and 71.0% of outcomes in total.

	Continuous	Dichotomous	TOTAL
Baseline	40.00% (14)	17.64% (6)	28.99% (20)
PRS <sub>multi</sub>	40.00% (14)	52.94% (18)	46.38% (32)
EX-TERR	20.00% (7)	29.41% (10)	24.64% (17)
Total #	35	34	69

Table 5.1: **Proportion of best performing PRS methodology.** Proportion of best performing methodology compared between baseline, PRS<sub>multi</sub>, and EX-TERR, categorized by continuous or dichotomous traits. Percentages represent the proportion of instances where each score performs best within each category, with the corresponding count shown in brackets.

Masked PRS performance can also be compared within the assigned outcome categories (Figure 5.4). Overall, PRS<sub>multi</sub> performed best overall, obtaining the highest results in 32 out of 69 outcomes (46.4% overall; 40.0% continuous, 52.9% dichotomous). Baseline PRS

performed best in 20 outcomes (29.0% overall; 40.0% continuous, 17.6% dichotomous). Finally, EX-TERR outperformed the two other methods in 17 outcomes (24.6% overall; 20.0% continuous, 29.4% dichotomous). In 36.2% of total outcomes, EX-TERR outperformed PRS<sub>multi</sub> (42.8% continuous, 29.4% dichotomous). Although EX-TERR did not outperform all other methods in any specific category, it showed the best performance in the continuous “anthropometrics” category, returning the highest predictive accuracy in 2 out of 3 (66%) traits. The continuous outcome with the highest predictive accuracy was urate (baseline = 0.036, PRS<sub>multi</sub> adj.  $r^2$  = 0.039, EX-TERR adj.  $r^2$  = 0.057) and the most accurately predicted dichotomous trait was type II diabetes (T2D) (baseline odds ratio (OR) per SD = 1.32, PRS<sub>multi</sub> OR per SD = 1.40, EX-TERR OR per SD = 1.49). In the diabetes mellitus (DM) example, while all three masked PRS methods are independently significantly and highly predictive, the EX-TERR is the best performing overall followed closely by the unmasked single-trait PRS (Figure 5.5). Additionally, EX-TERR identified predictor traits that were most significantly associated with the DM outcome (Figure 5.7).

Masked PRS performance was drastically reduced relative to unmasked PRS performance, consistent across all three PRS methodologies tested (Table 5.2). This decrease is consistently more pronounced in continuous traits rather than dichotomous traits. Predictive accuracy is substantially lower in continuous traits relative to dichotomous traits. In baseline PRS, the average reduction in performance from masked to unmasked PRS is 88.6% for continuous traits and 16.5% for dichotomous traits. Similarly, the decrease for PRS<sub>multi</sub> is 84.8% for continuous traits and 16.4% for dichotomous traits. Finally, for EX-TERR PRS, it is 85.9% for continuous traits and 24.4% for dichotomous traits.

	Baseline		PRS <sub>multi</sub>		EX-TERR	
	MASKED	UNMASKED	MASKED	UNMASKED	MASKED	UNMASKED
<b>Continuous Traits</b>						
Apolipoprotein A	0.0022	0.13	0.0018	0.13	0.0020	0.12
Apolipoprotein B	0.021	0.12	0.0038	0.12	0.0017	0.12
Creatinine	0.0039	0.050	0.0037	0.055	0.0037	0.050
Cystatin C	0.0053	0.024	0.0095	0.031	0.0099	0.031
Glucose	0.0014	0.0059	0.0024	0.0074	0.0025	0.0071
HbA1c	0.0041	0.013	0.0069	0.018	0.0072	0.017
HDL	0.0022	0.17	0.033	0.17	0.0063	0.16
LDL	0.012	0.094	0.00025	0.095	0.00029	0.086
Triglycerides	0.0019	0.11	0.0046	0.11	0.0026	0.098
<b>Dichotomous Traits</b>						
Asthma	1.21	1.52	1.26	1.54	1.04	1.50
CAD	1.25	1.56	1.42	1.66	1.21	1.64
Cancer with breast cancer	1.03	1.77	1.00	1.77	0.99	1.62
DM	1.29	1.44	1.42	1.62	1.29	1.63
AFF	1.22	1.49	1.27	1.54	1.13	1.47
MI	1.24	1.52	1.38	1.61	1.13	1.57
Stroke	1.12	1.19	1.19	1.27	1.06	1.24
T2D	1.32	1.48	1.45	1.68	1.31	1.68
Renal failure	1.12	1.15	1.24	1.34	1.12	1.32

Table 5.2: **PRS performance with and without masking of GWAS information matching to the outcome.** Performance for masked and unmasked PRS is reported across the three methodologies: baseline, PRS<sub>multi</sub>, and EX-TERR. PRS performance for continuous outcomes is reported in adjusted  $r^2$  values, while performance for dichotomous outcomes is reported as odd ratios. A list of GWAS considered matching can be referred to in Supplementary Table S5.5. HbA1c = Glycated haemoglobin, HDL = High density lipoprotein, LDL = Low density lipoprotein, CAD = Coronary artery disease, DM = Diabetes mellitus, AFF = Atrial Fibrillation and flutter, MI = Myocardial infarction, T2D = Type II diabetes

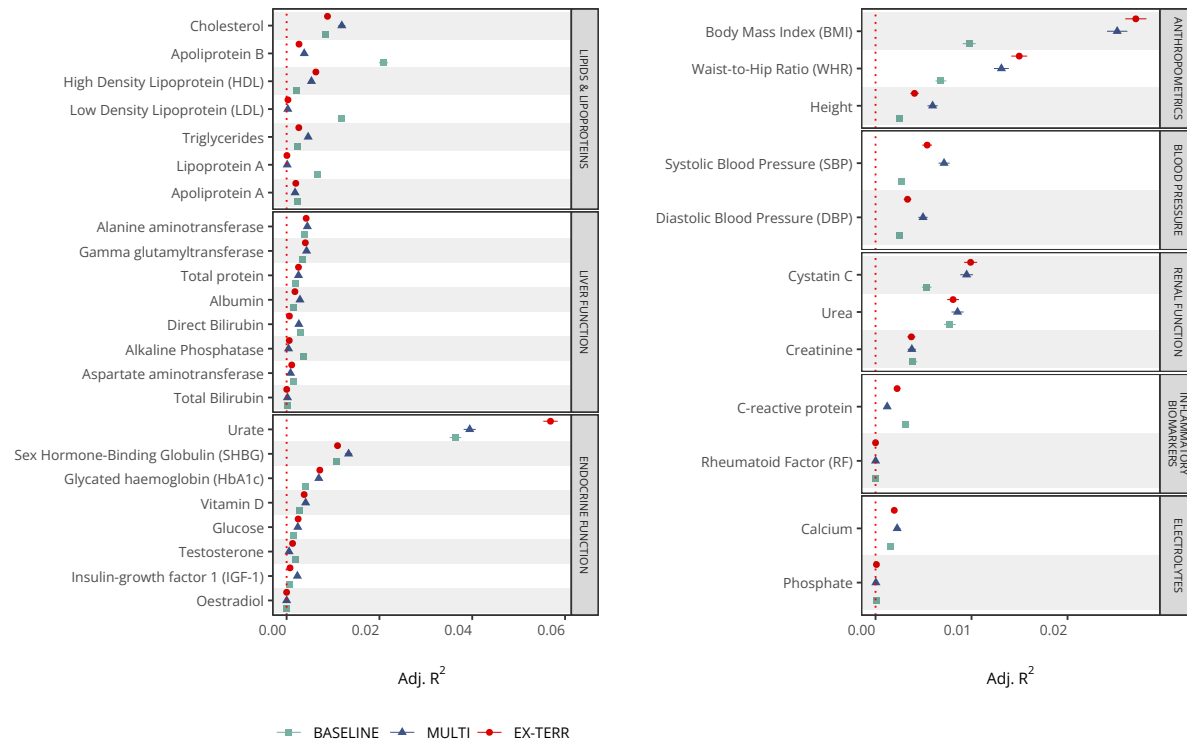
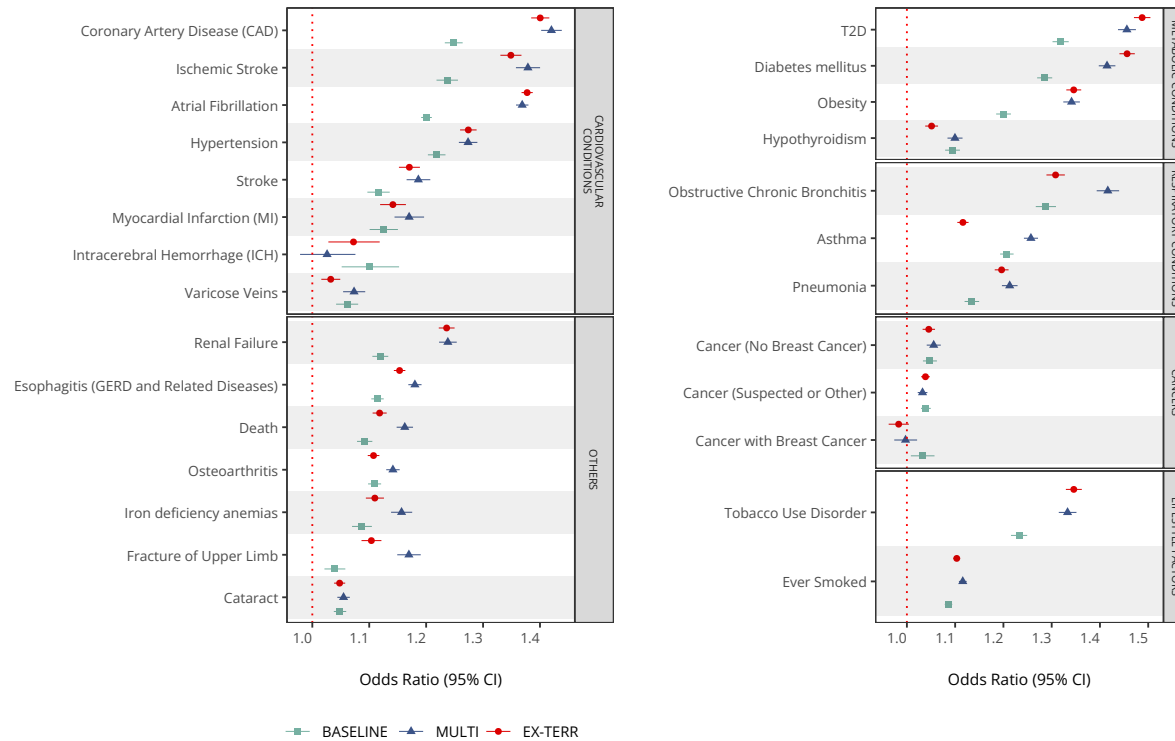


Figure 5.4: **Predictive performance of three PRS methodologies for 69 outcomes under the masked condition.** Forest plot illustrating predictive performances of the three PRS methodologies: baseline, PRS<sub>multi</sub> and EX-TERR through logistic regression. Error bars are for 95% confidence intervals (CI). **a) Continuous outcomes.** Forest plot of adjusted  $r^2$  with Cohens 95% confidence intervals (CI) for continuous traits.



**b) Dichotomous outcomes.** Forest plot illustrating predictive performances of the three PRS methodologies: baseline, PRS<sub>multi</sub> and EX-TERR through logistic regression. Performance presented as odds ratios (ORs) and 95% confidence interval (CI) error bars.

a.

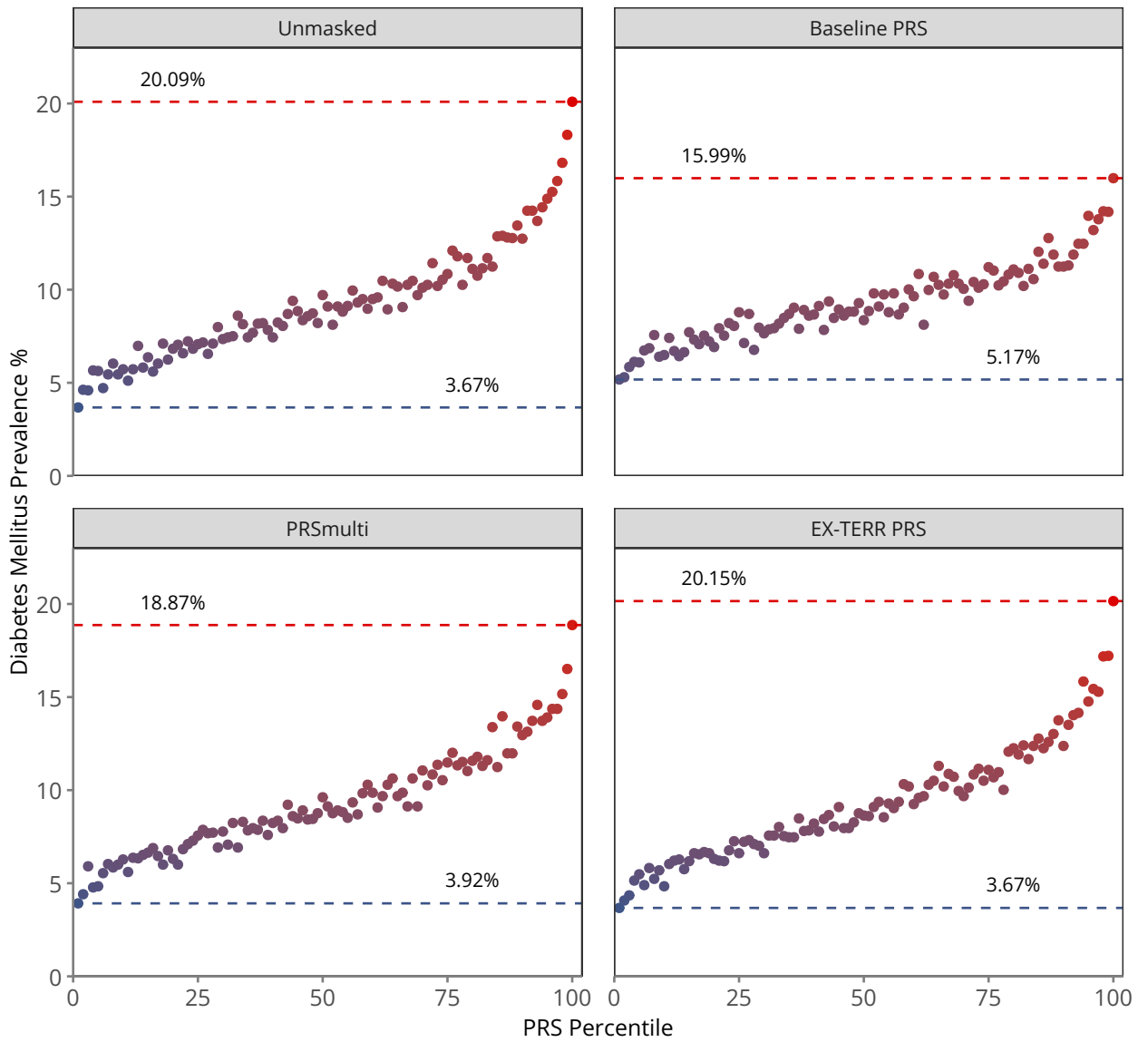


Figure 5.5: **Association of masked baseline PRS, masked PRS<sub>multi</sub>, masked EX-TERR PRS, and unmasked baseline PRS with diabetes mellitus.** a) Prevalence of diabetes mellitus according to PRS percentile. b) Odds ratio of diabetes per quintile of PRS. The first quintile acts as the reference, and standard error bars are for 95% confidence intervals.

b.

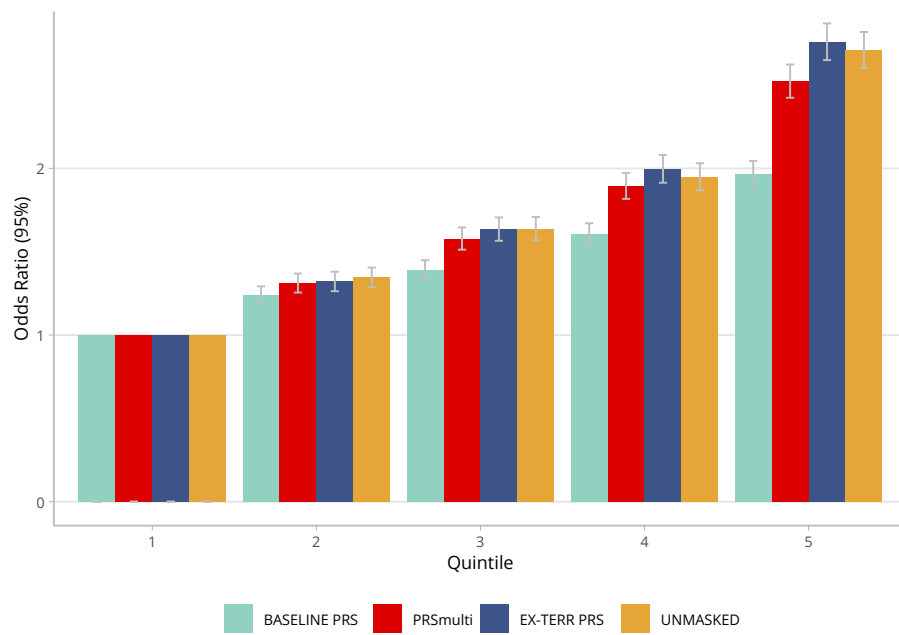


Figure 5.6: b) Odds ratio of diabetes per quintile of PRS. The first quintile acts as the reference, and standard error bars are for 95% confidence intervals.

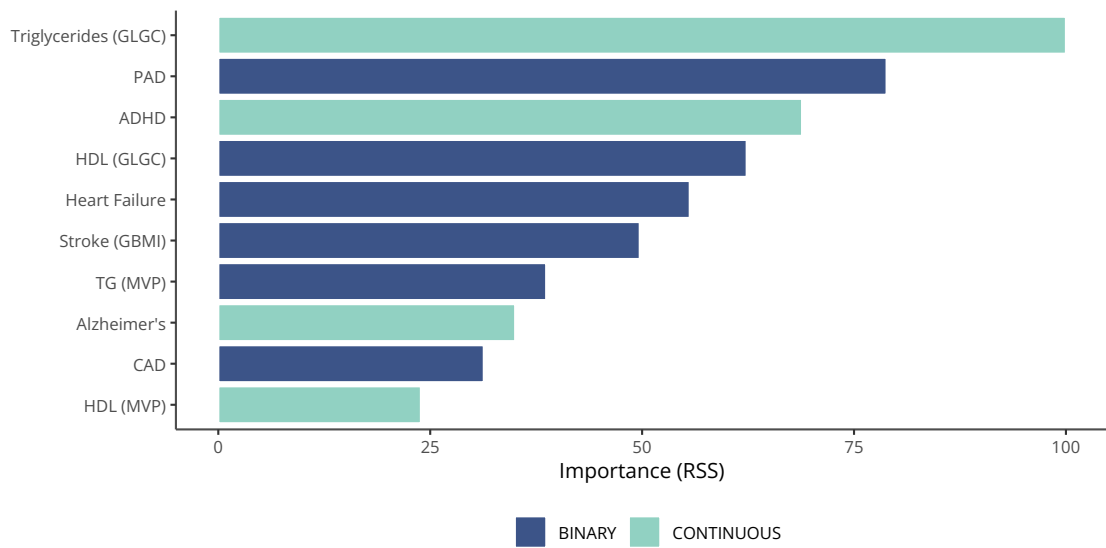


Figure 5.7: **Traits of highest importance for diabetes mellitus (DM) as determined by EX-TERR.** Variable importance plots (VIPs) for a single fold instance of EX-TERR for diabetes mellitus. Outcome structure is denoted as binary or continuous. Results are presented as residual sum of squares (RSS).

To test whether the different PRS are complementary or not, we repeated analyses while including all three PRS as independent variables for prediction of their corresponding outcome, adjusting for age, sex and 10 genetic PCs. Apart from iron deficiency anemias, there was no evidence PRS methods provided complementary information (Supplementary Table S5.6).

Discrimination and calibration of PRS<sub>multi</sub> and EX-TERR PRS were comparable. Across the 34 dichotomous outcomes, 23.5% of PRS<sub>multi</sub> and 20.6% of EX-TERR had evidence of miscalibration (p-value > 0.05) according to Hosmer-Lemeshow (HL) tests (Supplementary Table S5.8). When comparing HL or RMSE values, EX-TERR was better calibrated in 40.0% (14/35) of continuous outcomes and 41.2% (14/34) of dichotomous outcomes. Furthermore, 76.4% of HL tests for EX-TERR were significant while 73.5% of PRS<sub>multi</sub> HL tests were significant. Overall, this suggests that PRS<sub>multi</sub> is slightly better calibrated than EX-TERR overall. NRI was used to assess discriminative capacity of PRS<sub>multi</sub> and EX-TERR PRS (Supplementary Table S5.9). PRS<sub>multi</sub> and EX-TERR PRS each significantly discriminated 91.2% (31/34) of 36 dichotomous outcomes.

## 5.7 Discussion

Despite significant advancements in PRS methodologies within recent years, there remains uncertainty regarding implications when an external GWAS for the outcome trait is unavailable. To address this, we investigate the scenario by excluding external GWAS from PRS analyses and evaluating three different methods. We also introduce the novel methodology, EX-TERR, which combines MARS regression with 5-fold cross validation within groups containing genetic information equivalent to variants. This approach negates the need for an initial participant train-test split typical of most PRS methods incorporating super-

vised learning techniques, thereby addressing potential power reductions when an external GWAS is unavailable. EX-TERR utilizes UK Biobank (UKB) associations as outcomes for the MARS regression, with external GWAS associations for multiple traits serving as predictor. All association coefficients are adjusted using LDpred2, followed by dimensionality reduction through PCA. The MARS-derived regression coefficients are then applied as weights to construct the final EX-TERR PRS.

The absence of an external GWAS corresponding to a target trait greatly limited the performance of PRS. Under this scenario, PRS<sub>multi</sub> was the best performing method overall, achieving the highest prediction accuracy for 46.4% of all outcomes tested, followed by EX-TERR at 29.0% and baseline at 24.6%. The decrease in performance can be attributed to various underlying factors, including genetic architecture. For a PRS to be effective, the target outcome trait falls under several assumptions: 1) it is highly heritable 2) it has polygenic characteristics and 3) the data from summary statistics is applicable to the cohort for which the PRS is being generated (e.g. matching ethnicities). If a trait does not satisfy these conditions, PRS performance will be greatly diminished regardless of methodology. For example, associations are weak for all three methods for lipoprotein(a). Lp(a) is considered an independent risk factor for cardiovascular diseases with a unique genetic architecture. Lp(a) is highly heritable, with estimates as high as 95% [22]. However, this genetic influence mainly stems from the LPA locus, making it unlikely that any unrelated GWAS will substantially contribute to PRS predictiveness.

The magnitude of decrease varied across different traits. This may be attributed to a mismatch in phenotypic definition between the external GWAS and UKB outcome. For example, the CKDGen GWAS for Chronic Kidney Disease (CKD) was used to generate the “renal failure” outcome[23]. While kidney disease is defined as the gradual loss of kidney

function due to the presence of kidney damage, this can manifest in stages and may not match directly to our definition of “renal failure” in UKB. This could lead to a decrease in predictiveness accuracy in the unmasked score due to a mismatch in GWAS and outcome definition, translating to a smaller gap in performance between the unmasked and masked scores. The larger reduction observed in continuous outcomes compared to dichotomous outcomes can be explained by the higher precision of empirical measurements in continuous outcomes, while dichotomous outcomes often involve multiple comorbidities and risk factors. For most outcomes,  $\text{PRS}_{\text{multi}}$  and EX-TERR outperformed the best-performing baseline single-trait PRS. However, the predictiveness of the baseline PRS should not be entirely dismissed, as it outperformed all other methods in approximately 25% of outcomes. Overall, it is evident that no single method universally outperforms others. This may also be attributable to underlying genetic architectures as discussed above. Additionally, different outcomes will be unique in their co-morbidities, heritability, and environmental influence, which can be difficult to capture within a single PRS methodology.

Our novel methodology of EX-TERR offers several advantages for PRS calculation. The use of the underlying MARS regression model offers a flexible, non-parametric approach that adapts well to various datasets and offers valuable insights into the biological interactions among predictor traits. It requires no assumption of the input data distribution and can work well with both continuous and dichotomous outcomes. Our results demonstrate that, under certain circumstances, EX-TERR is comparable to other penalized regression multi-PRS methods which are commonly utilized in current PRS studies [32, 33, 34, 35]. However,  $\text{PRS}_{\text{multi}}$  remains the best performing technique overall. The internal cross-validation of genetic information (rotated blocks) removes the necessity of the initial sample train/test split, allowing for potential mitigation of issues arising from smaller sample sizes. This additional cross validation step could potentially mitigate the

issue of low-powered or unavailable external GWAS matching to the outcome. Finally, MARS basis functions are relatively easier to interpret than alternative regression models (e.g. neural networks, random foresting) and allow for determination of the relative genetic contribution of various predictor traits to a given outcome.

The study is not without limitations. While the multi-PRS methods address the situation of a missing GWAS, it still requires GWAS data from similar or matching ancestry. This is one of the greatest challenges for the clinical use of PRS on a global basis, as non-European ancestries remain underrepresented in GWAS studies. If information is only available from European ancestries, then  $\text{PRS}_{\text{multi}}$  and EX-TERR will not mitigate this. Also, all methods utilized assume polygenicity and will underperform if the genetic architecture is different (e.g.  $L_p(a)$ ), or there is little heritability. We also acknowledge some limitations of the novel EX-TERR method. EX-TERR requires supervised learning at the variant level, which may necessitate a large sample size to achieve robust results. While this is manageable in larger databases like the UKB, smaller clinical trials for rarer diseases may not have sufficient participants. Additionally, MARS, as a highly flexible non-parametric technique, is prone to overfitting, particularly with small sample sizes or high-dimensional datasets. Its computational intensity and sensitivity to outliers, combined with the need for meticulous tuning of numerous parameters (e.g. maximum number of interactions, penalty factor per knot), can lead to high variance and fluctuating performance.

## 5.8 Conclusion

Despite the significant dependence on external GWAS data corresponding to the target trait for PRS generation, little research has explored potential workarounds when such data is unavailable. Although no single PRS method universally outperforms all others,

PRS<sub>multi</sub> were able to improve predictive accuracy in most outcomes. Nonetheless, there are significant gaps in performance when external GWAS are not available, irrespective of methods. As more GWAS are available and with further methodological advances, the predictive accuracy of PRS when GWAS are not available can be expected to improve.

## References

- [1] E. Duschek, L. Forer, S. Schönherr, C. Gieger, et al. A polygenic and family risk score are both independently associated with risk of type 2 diabetes in a population-based study. *Scientific Reports*, 13(1):4805, March 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-31496-w.
- [2] M. Gatz, C. A. Reynolds, L. Fratiglioni, B. Johansson, et al. Role of Genes and Environments for Explaining Alzheimer Disease. *Archives of General Psychiatry*, 63(2):168174, 2006. ISSN 0003-990X. doi: 10.1001/archpsyc.63.2.168.
- [3] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(99):12191224, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0183-z.
- [4] A. V. Khera, C. A. Emdin, I. Drake, P. Natarajan, and other. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *New England Journal of Medicine*, 375(24):23492358, 2016. ISSN 0028-4793. doi: 10.1056/NEJMoa1605086.
- [5] P. Klimek, S. Aichberger, and S. Thurner. Disentangling genetic and environmental risk factors for individual diseases from multiplex comorbidity networks. *Scientific Reports*, 6(1):39658, December 2016. ISSN 2045-2322. doi: 10.1038/srep39658.
- [6] Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimers disease. 62. ISSN 0197-4580.
- [7] R. Lali, E. Cui, A. Ansarikaleibari, M. Pigeyre, and G. Paré. Genetics of early-onset coronary artery disease: from discovery to clinical translation. *Current*

*Opinion in Cardiology*, 34(6):706713, 2019. ISSN 1531-7080. doi: 10.1097/HCO.0000000000000676.

- [8] A. Le, H. Peng, D. Golinsky, M. Di Scipio, et al. What causes premature coronary artery disease? *Current Atherosclerosis Reports*, 26(6):189203, June 2024. ISSN 1534-6242. doi: 10.1007/s11883-024-01200-y.
- [9] O’Sullivan J. W., Raghavan S., Marquez-Luna C., Luzum J. L., et al. Polygenic risk scores for cardiovascular disease: A scientific statement from the american heart association. *Circulation*, 146(8):e93–e118, 2022.
- [10] S. Theriault, R. Lali, M. Chong, J. L. Velianou, M. K. Natarajan, and G. Paré. Polygenic Contribution in Individuals With Early-Onset Coronary Artery Disease. *Circ Genom Precis Med*, 11(1):e001849, 2018. doi: 10.1161/CIRCGEN.117.001849.
- [11] A. Torkamani, N. E. Wineinger, and E. J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581581, 2018. ISSN 14710056. doi: 10.1038/s41576-018-0018-x.
- [12] D. Klarin and P. Natarajan. Clinical utility of polygenic risk scores for coronary artery disease. *Nature reviews. Cardiology*, 19(5):291301, 2022. ISSN 1759-5002. doi: 10.1038/s41569-021-00638-w.
- [13] Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature genetics*, 49(3):403415, 2017. ISSN 1061-4036. doi: 10.1038/ng.3768.
- [14] F. Privé, J. Arbel, and B. J. Vilhjálmsson. Ldpred2: better, faster, stronger. *Bioinformatics*, 36(2223):54245431, April 2021. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btaa1029.

- [15] F. Privé, B. J. Vilhjálmsón, and T. S. H. Mak. lassosum2: an updated version complementing ldpred2. page 2021.03.29.437510, March 2021. doi: 10.1101/2021.03.29.437510. URL <https://www.biorxiv.org/content/10.1101/2021.03.29.437510v1>.
- [16] S. W. Choi and P. F. O'Reilly. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, 8(7), 2019. doi: 10.1093/gigascience/giz082. URL [https://journals.scholarsportal.info/details/2047217x/v08i0007/nfp\\_pprssfbd.xml](https://journals.scholarsportal.info/details/2047217x/v08i0007/nfp_pprssfbd.xml).
- [17] T. Ge, C. Chen, Y. Ni, Y. A. Feng, and J. W. Smoller. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(11):1776, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09718-5.
- [18] M. Nikpay, A. Goel, H. Won, L. M. Hall, et al. A comprehensive 1000 genomesbased genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):11211130, October 2015. ISSN 1546-1718. doi: 10.1038/ng.3396.
- [19] L. Kachuri, N. Chatterjee, J. Hirbo, D. J. Schaid, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews. Genetics*, 25(1):825, January 2024. ISSN 1471-0064. doi: 10.1038/s41576-023-00637-2.
- [20] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(44):584591, 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0379-x.
- [21] C. Sudlow, J. Gallacher, N. Allen, V. Beral, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):e1001779, 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779.

- [22] F. Kronenberg. Prediction of cardiovascular risk by lp(a) concentrations or genetic variants within the lpa gene region. *Clinical Research in Cardiology Supplements*, 14 (Suppl 1):512, April 2019. ISSN 1861-0714. doi: 10.1007/s11789-019-00093-5.
- [23] T. K. Chen, D. H. Knicely, and M. E. Grams. Chronic kidney disease diagnosis and management: A review. *JAMA*, 322(13):12941304, October 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.14745.
- [24] Friedman J., Hastie T., Tibshirani R., Balasubramanian N., et al. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2021. URL <https://cran.r-project.org/web/packages/glmnet/index.html>. R package version 4.1-8.
- [25] J. Mbatchou, L. Barnard, J. Backman, A. Marcketta, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7):10971103, July 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00870-7.
- [26] Friedman J. H. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1991.
- [27] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):3752, 1987. ISSN 0169-7439. doi: 10.1016/0169-7439(87)80084-9.
- [28] N. R. Cook, O. V. Demler, and N. P. Paynter. Clinical risk reclassification at 10 years. *Statistics in medicine*, 36(28):44984502, 2017. ISSN 0277-6715. doi: 10.1002/sim.7340.
- [29] D. W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):10431069, 1980. ISSN 0361-0926. doi: 10.1080/03610928008827941.

- [30] T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):12471250, 2014. ISSN 1991-959X. doi: 10.5194/gmd-7-1247-2014.
- [31] M. Hoekstra, H. Y. Chen, J. Rong, L. Dufresne, et al. Genome-wide association study highlights apoh as a novel locus for lipoprotein(a) levels. *Arteriosclerosis, thrombosis, and vascular biology*, 41(1):458464, January 2021. ISSN 1079-5642. doi: 10.1161/ATVBAHA.120.314965.
- [32] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469480, 2017. ISSN 1098-2272. doi: 10.1002/gepi.22050.
- [33] J. Pattee and W. Pan. Penalized regression and model selection methods for polygenic scores on summary statistics. *PLoS Computational Biology*, 16(10):e1008271, October 2020. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1008271.
- [34] F. Privé, B. J. Vilhjálmsón, H. Aschard, and M. G.B. Blum. Making the most of clumping and thresholding for polygenic scores. *American Journal of Human Genetics*, 105(6):12131221, 2019. ISSN 0002-9297. doi: 10.1016/j.ajhg.2019.11.001.
- [35] Q. F. Xu, X. H. Ding, C. X. Jiang, K. M. Yu, and L. Shi. An elastic-net penalized expectile regression with applications. *Journal of Applied Statistics*, 48(12):22052230, 2021. ISSN 0266-4763. doi: 10.1080/02664763.2020.1787355.
- [36] J. Elliott, B. Bodinier, T. A. Bond, M. Chadeau-Hyam, et al. Predictive Accuracy of a Polygenic Risk Score Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA*, 323(7):636645, 2020. ISSN 0098-7484. doi: 10.1001/jama.2019.22241.

- [37] K. M. de Lange, L. Moutsianas, J. C. Lee, C. A. Lamb, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256261, February 2017. ISSN 1061-4036. doi: 10.1038/ng.3760.
- [38] J. Z Liu, S. van Sommeren, H. Huang, S. C Ng, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979986, September 2015. ISSN 1061-4036. doi: 10.1038/ng.3359.
- [39] L. Duncan, Z. Yilmaz, R. Walters, J. Goldstein, et al. Genome-wide association study reveals first locus for anorexia nervosa and metabolic correlations. *The American journal of psychiatry*, 174(9):850858, 2017. ISSN 0002-953X. doi: 10.1176/appi.ajp.2017.16121402.
- [40] W. Zhou, M. Kanai, K. H. Wu, H. Rasheed, et al. Global biobank meta-analysis initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2(10):100192, October 2022. ISSN 2666-979X. doi: 10.1016/j.xgen.2022.100192.
- [41] G. Ni, J. Zeng, J. A. Revez, Y. Wang, et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. page 2020.09.10.20192310, 2021. doi: 10.1101/2020.09.10.20192310. URL <https://www.medrxiv.org/content/10.1101/2020.09.10.20192310v2>.
- [42] I. E. Christophersen, M. Rienstra, C. Roselli, X. Yin, et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nature Genetics*, 49(6):946952, June 2017. ISSN 1546-1718. doi: 10.1038/ng.3843.
- [43] C. M. Middeldorp, J. F. Felix, A. Mahajan, T. S. Ahluwalia, et al. The early growth

genetics (egg) and early genetics and lifecourse epidemiology (eagle) consortia: design, results and future prospects. *European Journal of Epidemiology*, 34(3):279300, March 2019. ISSN 1573-7284. doi: 10.1007/s10654-019-00502-9.

- [44] K. Estrada, E. Styrkarsdottir, U. and Evangelou, Y. Hsu, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature Genetics*, 44(5):491501, 2012. ISSN 1546-1718. doi: 10.1038/ng.2249.
- [45] H. Zhang, T. U. Ahearn, J. Lecarpentier, D. Barnes, J. Beesley, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature genetics*, 52(6):572581, June 2020. ISSN 1061-4036. doi: 10.1038/s41588-020-0609-2.
- [46] Anna Köttgen and Cristian Pattaro. The ckdgen consortium: ten years of insights into the genetic basis of kidney function. *Kidney International*, 97(2):236242, February 2020. ISSN 1523-1755. doi: 10.1016/j.kint.2019.10.027.
- [47] R. Malik, G. Chauhan, M. Traylor, M. Sargurupremraj, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature Genetics*, 50(4):524537, April 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0058-3.
- [48] J. F. Meschia, D. K. Arnett, H. Ay, R. D. Brown, et al. Stroke genetics network (sign) study: Design and rationale for a genome-wide association study of ischemic stroke subtypes. *Stroke; a journal of cerebral circulation*, 44(10):26942702, October 2013. ISSN 0039-2499. doi: 10.1161/STROKEAHA.113.001857.
- [49] A. Villaplana-Velasco, M. Pigeire, J. Engelmann, K. Rawlik, et al. Fine-mapping of

retinal vascular complexity loci identifies notch regulation as a shared mechanism with myocardial infarction outcomes. *Communications Biology*, 6(1):113, May 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04836-9.

[50] M. E. K. Niemi, J. Karjalainen, R. G. Liao, B. M. Neale, et al. Mapping the human genetic architecture of covid-19. *Nature*, 600(7889):472477, December 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03767-x.

[51] D. L. McCartney, J. L. Min, R. C. Richmond, A. T. Lu, et al. Genome-wide association studies identify 137 genetic loci for dna methylation biomarkers of aging. *Genome Biology*, 22(1):194, June 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02398-9.

Table S5.1: Baseline characteristics for UK Biobank British related participants (n = 408 160). Standard deviations shown for continuous traits and percentage of total participants shown for dichotomous traits.

Baseline characteristic	Female	Male	Total
No. of participants	220 647	187 513	408 160
Age of recruitment, years	56.7 (7.9)	57.1 (8.1)	56.9 (8.0)
<i>Continuous Traits</i>			
<b>Anthropometrics</b>			
Body Mass Index (BMI)	27.04 (5.09)	27.80 (4.19)	27.39 (4.71)
Diastolic Blood Pressure (DBP)	80.61 (9.59)	83.78 (9.56)	82.07 (9.71)
Height	162.55 (6.22)	175.82 (6.75)	168.67 (9.25)
Systolic Blood Pressure (SBP)	136.41 (18.62)	141.40 (16.76)	138.71 (17.96)
Waist-Hip Ratio (WHR)	103.35 (10.24)	103.46 (7.50)	103.40 (9.08)
<b>Lipids &amp; Lipoproteins</b>			
Apolipoprotein A	1.63 (0.25)	1.45 (0.22)	1.54 (0.25)
Apolipoprotein B	1.04 (0.23)	1.03 (0.23)	1.03 (0.23)
Cholesterol	5.90 (1.11)	5.50 (1.09)	5.71 (1.12)
High Density Lipoprotein (HDL)	1.58 (0.36)	1.30 (0.29)	1.45 (0.36)
Low Density Lipoprotein (LDL)	3.64 (0.85)	3.48 (0.84)	3.57 (0.85)
Lipoprotein A	44.34 (43.32)	43.72 (43.26)	44.06 (43.29)
Triglycerides	1.57 (0.83)	1.97 (1.12)	1.76 (0.99)
<b>Liver Function</b>			
Alanine aminotransferase (ALT)	20.44 (11.87)	27.10 (14.56)	23.51 (13.59)
Albumin	45.03 (2.42)	45.51 (2.45)	45.25 (2.45)
Alkaline phosphatase (ALP)	85.06 (26.81)	82.25 (25.21)	83.77 (26.12)
Aspartate aminotransferase	24.63 (9.28)	28.06 (11.36)	26.21 (10.43)
Bilirubin (direct)	1.69 (0.60)	1.99 (0.84)	1.83 (0.74)

Continued on next page

Table S5.1 – continued from previous page

Baseline characteristic	Female	Male	Total
Bilirubin (total)	8.17 (3.58)	10.212 (4.67)	9.11 (4.24)
Gamma glutamyltransferase	30.57 (32.76)	44.92 (46.63)	37.19 (40.40)
Total protein	72.31 (3.78)	72.48 (3.79)	72.39 (3.78)
Urea	5.28 (1.29)	5.63 (1.42)	5.44 (1.36)
<b>Endocrine</b>			
Glucose	5.07 (0.94)	5.17 (1.25)	5.11 (1.10)
Glycated haemoglobin (HbA1c)	35.68 (5.41)	36.25 (7.28)	35.94 (6.35)
Insulin-growth factor 1 (IGF-1)	20.93 (5.56)	21.90 (5.37)	21.38 (5.50)
Oestradiol	474.16 (210.44)	438.20 (70.42)	457.57 (162.67)
Sex Hormone-Binding Globulin (SHBG)	60.70 (28.84)	41.54 (16.22)	51.86 (25.70)
Testosterone	2.21 (2.26)	11.69 (3.79)	6.58 (5.63)
Urate	272.58 (64.65)	352.34 (71.03)	309.38 (78.48)
Vitamin D	49.86 (19.82)	50.06 (20.43)	49.95 (20.10)
<b>Renal Function</b>			
Creatinine	64.82 (13.93)	81.38 (17.73)	72.46 (17.83)
Cystatin C	0.88 (0.16)	0.94 (0.17)	0.91 (0.17)
<b>Inflammatory Biomarkers</b>			
C-reactive protein	2.68 (4.19)	2.46 (4.17)	2.58 (4.18)
Rheumatoid Factor (RF)	24.46 (6.12)	24.38 (5.53)	24.42 (5.86)
<b>Electrolytes</b>			
Calcium	2.38 (0.091)	2.37 (0.084)	2.38 (0.0877)
Phosphate	1.190 (0.14)	1.12 (0.153)	1.16 (0.15)
<i>Dichotomous Traits</i>			
<b>Cardiovascular Conditions</b>			
Atrial Fibrillation	2536 (5.77)	4426 (11.75)	6962 (8.53)
Cardiac dysrhythmias	4533 (10.31)	6330 (16.81)	10863 (13.31)
Cerebrovascular disease	1822 (4.14)	2492 (6.62)	4314 (5.28)

Continued on next page

Table S5.1 – continued from previous page

Baseline characteristic	Female	Male	Total
Cerebrovascular disease (Ever smoked)	695 (1.85)	688 (1.56)	1383 (1.69)
Cerebrovascular disease (Never smoked)	336 (0.89)	519 (1.18)	855 (1.05)
Coronary Artery Disease (CAD)	2207 (5.02)	5118 (13.59)	7325 (8.97)
Coronary atherosclerosis	2045 (4.65)	5035 (13.37)	7080 (8.67)
Hemorrhagic stroke	198 (0.45)	171 (0.45)	369 (0.45)
Hypertension	12413 (28.23)	13920 (36.96)	26333 (32.26)
Intracerebral hemorrhage (ICH)	206 (0.47)	237 (0.63)	443 (0.54)
Ischemic heart disease	3788 (8.61)	6984 (18.54)	10772 (13.20)
Ischemic stroke	736 (1.67)	1177 (3.13)	1913 (2.34)
Myocardial infarction	1321 (3.00)	3330 (8.84)	4651 (5.70)
Stroke	1341 (3.05)	1897 (5.04)	3238 (3.97)
Subarachnoid hemorrhage (SAH)	215 (0.49)	140 (0.37)	355 (0.43)
Varicose veins	1216 (3.23)	1834 (4.17)	3050 (3.74)
<b>Metabolic conditions</b>			
Diabetes mellitus	3057 (6.95)	4429 (11.76)	7486 (9.17)
Hypothyroidism	4182 (9.51)	1102 (2.93)	5284 (6.47)
Hypothyroidism (NOS)	3979 (9.05)	1034 (2.75)	5013 (6.14)
Obesity	3417 (7.77)	3062 (8.13)	6479 (7.93)
Type II Diabetes (T2D)	2683 (6.10)	4078 (10.83)	6761 (8.28)
<b>Respiratory conditions</b>			
Asthma	4651 (10.58)	3324 (8.83)	7975 (9.77)
Pneumonia	2480 (5.64)	3398 (9.02)	5878 (7.20)
Obstructive chronic bron- chitis	1808 (4.11)	2152 (5.71)	3960 (4.85)
<b>Cancers</b>			
Cancer (No breast cancer)	2259 (5.14)	3433 (9.12)	5692 (6.97)
Cancer (suspected or other)	6428 (14.62)	5949 (15.80)	12377 (15.16)

Continued on next page

Table S5.1 – continued from previous page

Baseline characteristic	Female	Male	Total
Cancer (with breast cancer)	19 (0.050)	1693 (3.85)	1712 (2.10)
<b>Lifestyle factors</b>			
Tobacco use disorder	2423 (5.51)	3128 (8.31)	5551 (6.80)
Ever smoked	24593 (55.92)	24684 (65.54)	49277 (60.36)
<b>Others</b>			
Cataract	6632 (15.08)	4765 (12.65)	11397 (13.96)
Death	2932 (6.67)	4275 (11.35)	7207 (8.83)
Esophagitis (GERD or related diseases)	6301 (14.33)	5472 (14.53)	11773 (14.42)
Iron deficiency anemias	2250 (5.12)	1750 (4.65)	4000 (4.90)
Fracture of upper limb	2123 (4.83)	1073 (2.85)	3196 (3.91)
Osteoarthritis	6407 (14.57)	4417 (11.73)	10824 (13.26)
Renal failure	3110 (7.07)	3946 (10.48)	7056 (8.64)

Table S5.2: Complete list of 61 external genome-wide association study (GWAS) summary statistics utilized in the analysis.

<b>Trait</b>	<b>Consortium</b>	<b>No. of participants</b>	<b>No. of variants</b>
Abdominal aortic aneurysm	GBMI	1 048 616	26 386 613
Acute appendicitis	GBMI	688 774	22 168 360
Attention Deficit Hyperactivity Disorder (ADHD)	PGC	55 374	8 007 024
Atrial Fibrillation (AF)	AFGen	109 934	10 506 607
All stroke	MEGASTROKE2	446 290	7 886 680
Alzheimers Disease	IGAP	54 162	7 046 850
Autism spectrum disorder (ASD)	PGC	46 350	9 063 082
Asthma	GBMI	1 399 372	36 673 123
Atopic dermatitis	EAGLE	116 863	9 980 328
Bone mass density (femoral - neck)	GEFOS	32 961	10 564 712
Bone mass density (forearm)	GEFOS	53 236	9 937 231
Bone mass density (lumbar - spine)	GEFOS	31 800	10 561 152
Breast cancer	BCAC	214 675	10 442 405
Blood urea nitrogen (BUN)	CKDGen	243 031	8 348 409
Coronary artery disease (CAD)	CARDIOGRAMplusC4D	184 305	8 607 408
Cardioembolic stroke	MEGASTROKE2	322 150	7 932 693
Cardiomyopathy	GBMI	776 580	22 565 896
Cryptogenic stroke (cardioembolism minor)	SiGN	25 841	10 524 062
Chronic Kidney Disease (CKD)	CKDGen	480 698	9 152 454
Chronic obstructive pulmonary disease (COPD)	GBMI	992 289	33 638 441
COVID-19	COVID-19 HGI	1 299 010	11 310 908
COVID-19 - transethnic	COVID-19 HGI	1 388 512	12 402 708
COVID-19 (hospitalized)	COVID-19 HGI	908 494	9 667 688
COVID-19 (hospitalized) - transethnic	COVID-19 HGI	10 908	14 625 029
COVID-19 respiratory failure	Ellinghaus et al[36]	3 790	8 313 919
COVID-19 severe respiratory failure	COVID-19 HGI	387 440	11 571 054
Chrons disease	de Lange et al[37]	40 266	9 560 360

Continued on next page

**Table S5.2 – continued from previous page**

<b>Trait</b>	<b>Consortium</b>	<b>No. of participants</b>	<b>No. of variants</b>
Eating disorder		14 477	10 583 669
Estimated Glomerular-Filtration Rate (eGFR)	CKDGen	567 460	8 875 632
Embolic Strokes of Undetermined Source (ESUS) - Causative Classification System causative (CCSc)	SiGN	389 324	8 525 922
Embolic Strokes of Undetermined Source (ESUS) - TOAST classification	SiGN	389 324	8 393 468
Glaucoma	GBMI	837 753	29 582 517
Gout	GBMI	836 867	36 628 642
Granulocyte % - epigenetic prediction	GoDMC	34 710	7 559 111
High density lipoproteins (HDL)	GLGC	888 227	36 448 786
High density lipoproteins (HDL)	MVP	210 967	16 534 376
Heart failure	GBMI	821 198	33 810 461
Idiopathic pulmonary fibrosis	GBMI	848 798	28 936 945
Inflammatory bowel disease	Liu et al[38]	59 957	9 724 768
Ischemic stroke	MEGASTROKE2	434 418	7 964 850
Lacunar stroke	MEGASTROKE2	254 959	6 931 872
Large artery stroke	MEGASTROKE2	204 911	8 071 617
Low density lipoproteins (LDL)	GLGC	842 660	35 745 393
Low density lipoproteins (LDL)	MVP	215 196	16 551 672
Myocardial infarction (MI)	CARDIOGRAMplusC4D	163 654	8 453 993
Obsession-Compulsive Disorder (OCD)	PGC	9 725	8 394 162
Peripheral artery disease (PAD)	MVP	174 902	16 164 110
Retinal fractal dimension	CLSA	20 025	10 205 093
Schizophrenia	PGC	150 064	8 059 459
Sclerosing cholangitis	Duncan et al[39]	14 890	7 650 626
Small vessel disease	MEGASTROKE2	255 458	7 940 270
Stroke	GBMI	961 310	33 559 955
Triglycerides (TG)	GLGC	864 240	35 884 163
Triglycerides (TG)	MVP	211 491	16 511 606
Type II diabetes (T2D)	DIAGRAM	159 208	10 775 984

Continued on next page

**Table S5.2 – continued from previous page**

<b>Trait</b>	<b>Consortium</b>	<b>No. of participants</b>	<b>No. of variants</b>
Thyroid cancer	GBMI	1 201 302	30 468 380
Ulcerative colitis	de Lange et al.	45 975	9 577 406
Uterine cancer	GBMI	491 827	23 547 992
Venous thromboembolism	GBMI	628 300	28 586 040
White Matter Hyperintensities (WMH)	Traylor et al.	11 226	7 590 417

GBMI = Global Biobank Meta-analysis Initiative[40], PGC = Psychiatric Genomics Consortium[41], AFGen = Atrial Fibrillation Consortium[42], EAGLE = Early Genetics and Lifecourse Epidemiology[43], GEFOS = Genetic Factors for Osteoporosis Consortium[44], BCAC = Breast Cancer Association Consortium[45], CKDGen = CKDGen (Chronic Kidney Disease Genetics) Consortium[46], CARDIOGRAMplusC4D = Coronary Artery Disease Genome-wide Replication and Meta-analysis plus The Coronary Artery Disease (C4D) Genetics Consortium[18], MEGASTROKE2 = MEGASTROKE Consortium[47], SiGN = NINDS (U.S. National Institute of Neurologic Disorders and Stroke) Stroke Genetics Network[48], CLSA = Canadian Longitudinal Study on Aging[49], COVID-19 HGI = COVID-19 Host Genetics Initiative[50] GoDMC = Genetics of DNA Methylation Consortium[51]

<b>Trait 1</b>	<b>Trait 2</b>	$R^2$
T2D	Diabetes mellitus	0.898
BMI	WHR	0.891
Coronary atherosclerosis	Ischemic heart disease	0.891
Coronary atherosclerosis	Myocardial infarction	0.886
Ischemic heart disease	Myocardial infarction	0.831
Apolipoprotein B	Cholesterol	0.769
Glucose	Glycated haemoglobin (HbA1c)	0.764
BMI	Obesity	0.738
SBP	Hypertension	0.736
Cerebrovascular Disease (Ever Smoked)	SAH	0.734
SBP	DBP	0.716
Osteoarthritis	Death	0.716
Creatinine	Cystatin C	0.715
Diabetes mellitus	Hypertension	0.714
HDL cholesterol	Triglycerides	0.710

Table S5.3: Pairs of phenotypes with high ( $R^2 > 0.7$ ) correlation. Used to aid in decision of trait masking i.e. determine which traits were closely related enough for masking.

Table S5.4: Complete list of outcome list with their definitions and their corresponding masked traits.

Phenotype	Definition (UKB Field IDs)	UKB Category	Masked GWAS Traits
Continuous Traits			
Alanine aminotransferase	30620	Blood biochemistry	
Albumin	30600	Blood biochemistry	
Alkaline phosphatase	30610	Blood biochemistry	
Apolipoprotein A	30630	Blood biochemistry	HDL (GLGC), HDL (MVP), TG (GLGC), TG (MVP)
Apolipoprotein B	30640	Blood biochemistry	LDL (GLGC), LDL (MVP)
Aspartate aminotransferase	30650	Blood biochemistry	
Bilirubin (direct)	30660	Blood biochemistry	
Bilirubin (total)	30840	Blood biochemistry	
Body Mass Index (BMI)	21001	Body size measures	
Calcium	30680	Blood biochemistry	
Cholesterol	30690	Blood biochemistry	
C-reactive protein	30710	Blood biochemistry	
Creatinine	30700	Blood biochemistry	CKD (CKDGen), eGFR (CKDGen)
Cystatin C	30720	Blood biochemistry	CKD (CKDGen), eGFR (CKDGen)
Diastolic Blood Pressure (DBP)	4079		
Blood pressure			
Gamma glutamyltransferase	30730	Blood biochemistry	
Glucose	30740	Blood biochemistry	T2D (DIAGRAM)
Glycated haemoglobin (HbA1c)	30750	Blood biochemistry	T2D (DIAGRAM)

Continued on next page

Table S5.4 – continued from previous page

Phenotype	Definition (UKB Field IDs)	UKB Category	Masked GWAS Traits
High Density Lipoprotein (HDL) Cholesterol	30760	Blood biochemistry	HDL (GLGC), HDL (MVP)
Height	12144	Body size measures	
Insulin-like Growth Factor 1 (IGF-1)	30770	Blood biochemistry	
Low Density Lipoprotein (LDL) Cholesterol	30780	Blood biochemistry	
Lipoprotein A	30790	Blood biochemistry	
Oestradiol	30800	Blood biochemistry	
Phosphate	30810	Blood biochemistry	
Protein (total)	30860	Blood biochemistry	
Rheumatoid Factor	30820	Blood biochemistry	
Systolic Blood Pressure (SBP)	4080	Blood pressure	
Sex Hormone Binding Globulin (SHBG)	30830	Blood biochemistry	
Testosterone	30850	Blood biochemistry	
Triglycerides	30870	Blood biochemistry	HDL (GLGC), HDL (MVP), TG (GLGC), TG (MVP)
Urate	30880	Blood biochemistry	
Urea	30670	Blood biochemistry	BUN (CKDGen)
Vitamin D	30890	Blood biochemistry	
Waist-Hip Ratio	48 / 49	Body size measures	

Continued on next page

**Table S5.4 – continued from previous page**

Phenotype	Definition (UKB Field IDs)	UKB Category	Masked GWAS Traits
Dichotomous Traits			
Atrial Fibrillation (AF)	427.2 / I48	Circulatory system	CAD (CARDIOGRAMplusC4D), AF (AFGen)
Asthma	495 / J45, J450, J451, J458, J459, J46	Respiratory system	Asthma (GBMI)
Cancer (suspected or other)	195 / D050, D09, D093, D099, Z031, B217, C068, C45, C457, C459, C76, C762, C763, C764, C765, C767, C768, C80, C97, M907, Z85, Z854, Z859, Z860	Neoplasms	Breast cancer (BCAC), Thyroid cancer (GBMI), Uterine cancer (GBMI)
Cancer (no breast cancer)	Artificial trait - cancer without breast cancer	Neoplasms	
Cancer (with breast cancer)	Artificial trait - cancer with breast cancer	Neoplasms	Breast cancer (BCAC)
Cardiac dysrhythmias	427 / I47, I479, I471, I492, I470, I472, I48, I494, I498, R001, I499, I493, I491, R000, I495, R002, I490, I46, I460, I469	Circulatory system	AF (AFGen)
Cataract	366 / H26, H262, H263, H264, H268, H269, H280, H281, H282, Z961, H260, H25, H250, H251, H252, H258, H259, H261	Sense organs	

Continued on next page
------------------------

Table S5.4 – continued from previous page

Phenotype		Definition (UKB Field IDs)	UKB Category	Masked GWAS Traits
Cerebrovascular (CeD)	disease	433 / G454, G463, G464, G465, G466, G467, G468, I600, I677, I679, I68, I680, I682, I688, I65, I650, I651, I652, I653, I658, I659, I631, I632, I672, I63, I632, I635, I639, I66, I660, I661, I662, I663, I664, I668, I669, I676, I630, I633, I634, I635, I636, I638, I64, I67, I678, G450, G451, G452, G458, G459, G460, G461, G462, I675, I671, I69, I690, I691, I692, I693, I694, I698	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)
Cerebrovascular (ever smoked)	disease	Artificial trait - CeD cases with ever smoked status	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)

Continued on next page

Table S5.4 – continued from previous page

Phenotype	Definition (UKB Field IDs)	UKB Category	Masked GWAS Traits
Cerebrovascular disease (never smoked)	Artificial trait - CeD cases without smokers	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)
Coronary Artery Disease (CAD)	Definition from Khera et al[3]: I21, I22, I23, I241, I252 + OPSC-4 codes: K40.1 40.4, K41.1-41.4, K45.1-45.5, K49.1-49.2, K49.8-49.9, K50.2, K75.1-75.6, K75.8-75.9	Circulatory system	CAD (CARDIOGRAMplusC4D), HF (GBMI)
Coronary atherosclerosis	411.4 / I240, I241, Z951, Z955, I253, I254, I341	Circulatory system	CAD (CARDIOGRAMplusC4D), HF (GBMI)
Death	All-cause mortality in death register (ICD10 Fields: 40001/40002)	Death register	

Continued on next page

**Table S5.4 – continued from previous page**

<b>Phenotype</b>	<b>Definition (UKB Field IDs)</b>	<b>UKB Category</b>	<b>Masked GWAS Traits</b>
Diabetes mellitus	250 / E10, E100, E106, E107, E108, E109, E101, E102, E103, E104, E11, E110, E116, E117, E118, E119, E13, E135, E136, E137, E138, E139, E149, E111, E131, E112, E103, E113, E123, E104, E114, E134, G590, Z964, R81, R824, G632, E103, E133, H360, R730, R739	Metabolic traits	T2D (DIAGRAM)
Esophagitis (GERD and related diseases)	530.1 / K20, K21, K219, K221, K210	Digestive traits	
Ever smoked	20610	Lifestyle factors	
Fracture of upper limb	803 / S427, S527, S528, T022, T024, T10, T922, S422, S423, S424, S429, S52, S520, S521, S522, S523, S524, S525, S529, S224, S420, S421	Injuries & poisonings	

Continued on next page

**Table S5.4 – continued from previous page**

<b>Phenotype</b>	<b>Definition (UKB Field IDs)</b>	<b>UKB Category</b>	<b>Masked GWAS Traits</b>
Hemorrhagic stroke	Algorithmically-defined outcomes: 42010 or 42011 + 42012 or 42013	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)
Hypertension	401 / I10, I13, I130, I131, I132, I139, I11, I110, I119, I12, I120, I129, I15, I150, I151, I152, I158, I159, I674	Circulatory system	
Hypothyroidism	244 / E032, E890, E018, E02, E033, E038, E039, E00, E000, E001, E002, E009, E030, E031	Endocrine system	
Hypothyroidism (NOS)	244.4 / E039	Endocrine system	

Continued on next page

Table S5.4 – continued from previous page

Phenotype	Definition (UKB Field IDs)	UKB Category	Masked GWAS Traits
Intracerebral Hemorrhage (ICH)	Algorithmically-defined outcomes: 42010 or 42011	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)
Iron deficiency anemias	280 / D50, D501, D508, D509, D500	Hematopoietic system	
Ischemic heart disease	411 / I200, I21, I210, I211, I212, I213, I214, I219, I22, I220, I221, I228, I229, I23, I230, I231, I232, I233, I236, I238, I241, I252, I510, I513, I20, I201, I208, I209, I240, I251, Z951, Z955, I253, I254, I341, I25, I255, I256, I258, I259, I24, I248, I249	Circulatory system	CAD (CARDIOGRAMplusC4D), HF (GBMI)

Continued on next page

**Table S5.4 – continued from previous page**

<b>Phenotype</b>	<b>Definition (UKB Field IDs)</b>	<b>UKB Category</b>	<b>Masked GWAS Traits</b>
Ischemic stroke	Algorithmically-defined outcomes: 42008, 42009	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)
Myocardial infarction	411.2 / I21, I210, I211, I212, I213, I214, I219, I22, I220, I221, I228, I229, I23, I230, I231, I232, I233, I236, I238, I241, I252, I510, I513	Circulatory system	CAD (CARDIOGRAMplusC4D), HF (GBMI)
Obesity	278.1 / E66, E660, E661, E668, E669	Metabolic traits	
Obstructive chronic bronchitis	496.21 / J44, J440, J441	Respiratory system	COPD (GBMI)
Osteoarthritis	740 / M16, M163, M169, M189, M160, M161, M171, M180, M181, M166, M167, M174, M175, M185, M192, M150, M151, M152, M139, M153, M154, M190	Musculoskeletal system	BMD: femoral neck, lumbar spine, forearm (GEFOS)

Continued on next page

**Table S5.4 – continued from previous page**

<b>Phenotype</b>	<b>Definition (UKB Field IDs)</b>	<b>UKB Category</b>	<b>Masked GWAS Traits</b>
Pneumonia	480 / B583, J100, J110, J168, J172, J18, J188, J189, A202, A212, A221, A310, A420, A430, A481, A78, B960, B961, J14, J150, J152, J153, J154, J155, J156, J157, J158, J159, J160, J13, J181, J151, B012, B052, B250, J12, J120, J121, J122, J128, J129, B206, B371, B380, B381, B382, B440, B59, J17, J180, J85, J850, J851, J852, J853	Respiratory system	
Renal failure	585 / N17, N170, N171, N172, N178, N179, N19, N18, N180, N189, Y602, Y612, Y620, Y841, Z491, Z492, Z992	Genitourinary system	CKD (CKDGen), eGFR (CKDGen)
Subarachnoid hemorrhage (SAH)	Algorithmically-defined outcomes: 42011 or 42012	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)

Continued on next page

Table S5.4 – continued from previous page

Phenotype	Definition (UKB Field IDs)	UKB Category	Masked GWAS Traits
Stroke	Algorithmically-defined outcomes: 42006 or 42007	Circulatory system	All stroke (MEGASTROKE2), Cardioembolic stroke (GBMI), CCSc (SiGN), ESUS CCSc (SiGN), ESUS TOAST (SiGN), Ischemic stroke (MEGASTROKE2), Lacunar stroke (MEGASTROKE2), Large artery stroke (MEGASTROKE2), Small vessel disease (MEGASTROKE2), STROKE (GBMI)
Type II Diabetes (T2D)	E11	Metabolic traits	T2D (DIAGRAM)
Tobacco use disorder	318 / F170, F171, F172, F173, F174, F179, Z720	Mental disorders	
Varicose veins	454 / I86, I860, I861, I862, I863, I864, I868, I83, I839, I830, I831, I832	Circulatory system	

<sup>1</sup>Phecodes based off of ICD10 UKB fields: 41270 (ICD10 diagnoses), 40001 (ICD10 underlying primary cause of death) 40002 (ICD10 contributory secondary cause of death)

<sup>2</sup>Algorithmically-defined outcomes: for all-cause stroke are from UKB field number 42006.

<b>Outcome</b>	<b>Matched GWAS Trait</b>	<b>Consortium</b>
<b>Continuous Outcomes</b>		
Apolipoprotein A	HDL	MVP
Apolipoprotein B	LDL	MVP
Creatinine	eGFR	CKDGen
Cystatin C	eGFR	CKDGen
Glucose	T2D	DIAGRAM
HbA1c	T2D	DIAGRAM
HDL	HDL	MVP
LDL	LDL	MVP
<b>Dichotomous Outcomes</b>		
AFF	AF	AFGen
Asthma	Asthma	GBMI
CAD	CAD	CARDIOGRAMplusC4D
Cancer (with Breast Cancer)	Breast cancer	BCAC
Cardiac dysrhythmias	AF	AFGen
Cerebrovascular disease	Stroke	GBMI
Coronary atherosclerosis	CAD	CARDIOGRAMplusC4D
Diabetes mellitus	T2D	DIAGRAM
Myocardial infarction	MI	CARDIOGRAMplusC4D
Renal failure	CKD	CKDGen
Stroke	Stroke	GBMI
T2D	T2D	DIAGRAM

Table S5.5: Matching GWAS for each outcome where available for PRS generation. The data consortium is specified to clarify situations in which duplicate GWAS summary statistics exist. Supplementary Table S5.2 provides full details regarding GWAS information.

Table S5.6: **Significance of each methodology’s PRS to each outcome.**  $p$ -values for multivariate regression models between the three PRS for baseline, PRS<sub>multi</sub>, and EX-TERR and their corresponding outcomes. Logistic regression was performed for dichotomous outcomes and linear regression for continuous outcomes. All models adjusted age, sex, and 10 principal components (PCs). Significance denoted with (\*).

Trait	Baseline PRS p-value	PRS <sub>multi</sub> p-value	EX-TERR PRS p-values
Atrial fibrillation and flutter (AFF)	0.25	0.17	0.83
Alanine aminotransferase	0.078	0.020*	0.59
Albumin	0.15	0.69	0.41
Alkaline phosphatase	0.55	0.027*	0.99
Apolipoprotein A	0.70	0.81	0.82
Apolipoprotein B	0.80	0.75	0.50
Aspartate aminotransferase	0.31	0.008*	0.34
Asthma	0.02	0.85	0.059
Body mass index (BMI)	0.63	0.68	0.66
Bilirubin (total)	0.75	0.38	0.79
Coronary artery disease (CAD)	0.26	0.81	0.84
Calcium	0.61	0.30	0.32
Cancer (no breast cancer)	0.73	0.12	0.052
Cancer (suspected or other)	0.79	0.039*	0.08
Cancer (with breast cancer)	0.78	0.76	0.57
Cardiac dysrhythmias	0.22	0.045*	0.25
Cataract	0.47	0.76	0.39
Cerebrovascular disease	0.21	0.64	0.30
Cerebrovascular disease (without smoking)	0.41	0.61	0.39
Cerebrovascular disease (with smoking)	0.38	0.21	0.51
Cholesterol	0.64	0.41	0.78
Coronary atherosclerosis	0.63	0.56	0.60
C-reactive protein	0.25	0.18	0.36
Creatinine	0.25	0.46	0.76
Cystatin C	0.090	0.27	0.85

Continued on next page

Table S5.6 – continued from previous page

Trait	Baseline PRS p-value	PRS <sub>multi</sub> p-value	EX-TERR PRS p-values
DBP	0.74	0.66	0.79
Death	0.31	0.41	0.031*
Diabetes mellitus	0.065	0.031*	0.29
Bilirubin (direct)	0.38	0.34	0.96
Esophagitis (GERD and related dis- eases)	0.29	0.99	0.57
Ever smoker	0.54	0.22	0.62
Fracture of upper limb	0.53	0.63	0.20
Gamma glutamyltransferase	0.36	0.34	0.19
Glucose	0.92	0.52	0.58
Glycated haemoglobin (HbA1c)	0.026	0.10	0.60
High-density lipoprotein (HDL) cholesterol	0.86	0.72	0.21
Height	0.062	0.16	0.12
Hemorrhagic stroke	0.59	0.20	0.24
Hypertension	0.73	0.86	0.89
Hypothyroidism	0.38	0.75	0.58
Hypothyroidism (NOS)	0.42	0.83	0.89
Intracerebral Hemorrhage (ICH)	0.19	0.73	0.87
Insulin-like growth factor 1 (IGF-1)	0.98	0.14	0.39
Iron deficiency anemias	0.95	0.012*	0.018*
Ischemic heart disease	0.95	0.38	0.47
Ischemic stroke	0.021*	0.89	0.67
Lipoprotein A	0.017*	0.10	0.91
Low-density lipoprotein (LDL) choles- terol	0.54	0.34	0.31
Myocardial infarction (MI)	0.95	0.81	0.73
Obesity	0.32	0.65	0.87
Obstructive chronic bronchitis	0.56	0.56	0.83
Oestradiol	0.21	0.91	0.24
Osteoarthritis	0.38	0.11	0.48
Phosphate	0.46	0.83	0.55

Continued on next page

Table S5.6 – continued from previous page

Trait	Baseline PRS p-value	PRS <sub>multi</sub> p-value	EX-TERR PRS p-values
Pneumonia	0.71	0.44	0.63
Protein (total)	0.95	0.43	0.68
Renal failure	0.16	0.35	0.28
Rheumatoid factor	0.90	0.95	0.44
SAH	0.39	0.79	0.18
SBP	0.15	0.67	0.29
Sex hormone binding globulin (SHBG)	0.037	0.06	0.83
Stroke	$2.4 \times 10^{-3**}$	0.43	0.85
T2D	0.088	0.016*	0.25
Testosterone	0.67	0.46	0.41
Tobacco use disorder	0.42	0.23	0.03
Triglycerides	0.25	$6.2 \times 10^{-5***}$	0.60
Urate	0.75	0.47	0.19
Urea	0.14	0.14	0.59
Varicose veins	0.45	0.28	0.31
Vitamin D	0.90	0.026*	0.80
Waist-to-Hip Ratio (WHR)	0.66	0.45	0.98

a. Continuous traits

Trait	PRS <sub>multi</sub> RMSE	EX-TERR RMSE	RMSE Difference
Alanine aminotransferase	13.227	13.211	0.0166
Albumin	2.394	2.398	-0.00336
Alkaline phosphatase	25.339	25.424	-0.0858
Apolipoprotein A	0.234	0.234	0.0000797
Apolipoprotein B	0.231	0.231	-0.000150
Aspartate aminotransferase	10.026	10.073	-0.0477
BMI	4.671	4.659	0.0124
Calcium	0.087	0.087	-0.0000698
Cholesterol	1.091	1.094	-0.00208
C-reactive protein	4.232	4.217	0.0150
Creatinine	15.449	15.505	-0.0564
Cystatin C	0.157	0.158	-0.000681
DBP	9.550	9.557	-0.00714
Direct bilirubin	0.753	0.747	0.00558
Gamma glutamyltransferase	39.991	39.923	0.0684
Glucose	1.127	1.119	0.00762
Glycated haemoglobin (HbA1c)	6.134	6.172	-0.0373
HDL cholesterol	0.329	0.329	0.000371
Height	6.288	6.296	-0.00840
IGF-1	5.319	5.319	0.0000474
LDL cho	0.845	0.845	-0.0000312
Lipoprotein <sub>A</sub>	43.203	43.220	-0.0174
Oestradiol	167.686	166.356	1.33
Phosphate	0.145	0.146	-0.000142
Rheumatoid factor	5.769	5.787	-0.0179
SBP	16.783	16.807	-0.0241
SHBG	23.664	23.699	-0.0350
Testosterone	3.061	3.059	0.00165
Bilirubin (total)	4.182	4.169	0.0124
Protein (total)	3.748	3.753	-0.00453
Triglycerides	0.974	0.974	-0.000327
Urate	65.730	65.169	0.561
Urea	1.299	1.300	-0.00104
Vitamin D	19.870	19.896	-0.0252
WHR	9.075	9.055	0.0204

Table S5.7: Calibration for all outcomes, comparison between PRS<sub>multi</sub> and EX-TERR. Fig a. Residual Mean Squared Error (RMSE) for continuous traits. The RMSE difference is calculated as PRS<sub>multi</sub> - EX-TERR. A lower RMSE statistic indicates a better fit.

**b. Dichotomous traits**

Fig b. Hosmer-Lemeshow (HL) test for dichotomous outcomes, with p-values shown in brackets. The HL difference is calculated as  $PRS_{\text{multi}} - \text{EX-TERR}$ . A lower HL statistic indicates a better fit.

Trait	$PRS_{\text{multi}}$ HL/ $\lambda^2$	$PRS_{\text{multi}}$ HL/ $\lambda^2$	EX-TERR HL/ $\lambda^2$	EX-TERR p-value
AFF	33.140	$5.81 \times 10^5$	31.087	0.000136
Asthma	6.358	0.61	7.157	0.520
CAD	94.885	$\lll 0.001$	97.531	$\lll 0.001$
Cancer, no breast cancer	147.841	$\lll 0.001$	135.520	$\lll 0.001$
Cancer, with breast cancer	17.270	0.0274	15.754	0.0460
Cancer, suspected or other	7.075	0.529	4.263	0.833
Cardiac dysrhythmias	134.644	$\lll 0.001$	139.925	$\lll 0.001$
Cataract	6.322	0.6115	3.984	0.859
Cerebrovascular disease	33.330	$5.37 \times 10^{-5}$	38.874	$5.19 \times 10^{-6}$
Coronary atherosclerosis	99.298	$\lll 0.001$	104.708	$\lll 0.001$
Death	96.009	$\lll 0.001$	104.593	$\lll 0.001$
Diabetes mellitus	11.772	0.162	17.316	0.0270
Esophagitis (GERD+)	57.133	$1.70 \times 10^{-9}$	56.787	$1.98 \times 10^{-9}$
Ever smoked	238.015	$\lll 0.001$	280.171	$\lll 0.001$
Fracture of upper limb	242.194	$\lll 0.001$	286.529	$\lll 0.001$
Hemorrhagic stroke	14.230	0.0760	11.402	0.180
Hypertension	17.555	0.0248	17.554	0.0248
Hypothyroidism	22.261	0.00445	23.458	0.00282
Hypothyroidism (NOS)	27.189	0.000656	24.574	0.00183
ICH	5.892	0.659	7.950	0.438
Iron deficiency anemias	25.062	0.00152	37.901	$7.85 \times 10^{-6}$
Ischemic heart disease	51.414	$2.18 \times 10^{-8}$	61.692	$2.17 \times 10^{-10}$
Ischemic stroke	11.057	0.199	12.098	0.147
Myocardial infarction	66.640	$2.29 \times 10^{-11}$	69.532	$6.09 \times 10^{-12}$
Obesity	25.564	0.00125	29.882	0.000222
Obs. chronic bronchitis	46.573	$1.85 \times 10^{-7}$	37.658	$8.71 \times 10^{-6}$
Osteoarthritis	25.248	0.00141	24.598	0.00182
Pneumonia	97.807	$\lll 0.001$	104.138	$\lll 0.001$
Renal failure	35.480	$2.19 \times 10^{-5}$	34.093	$3.91 \times 10^{-5}$
SAH	2.345	0.969	3.907	0.865
Stroke	18.622	0.0170	17.910	0.0219
T2D	38.012	$7.49 \times 10^{-6}$	57.372	$1.52 \times 10^{-9}$
Tobacco use disorder	14.934	0.0604	14.448	0.0708
Varicose veins	15.867	0.0443	16.364	0.0375

Trait	PRS <sub>multi</sub> NRI	EX-TERR PRS NRI	PRS <sub>multi</sub> p-value	EX-TERR PRS p-value
CAD	0.283	0.272	<<<0.001	<<<0.001
T2D	0.297	0.326	<<<0.001	<<<0.001
Stroke	0.129	0.129	<<<0.001	<<<0.001
AFF	0.189	0.197	<<<0.001	<<<0.001
Asthma	0.180	0.089	<<<0.001	<<<0.001
Cancer, no breast cancer	0.045	0.033	<<<0.001	<<<0.001
Cancer, suspected or other	0.111	0.030	<<<0.001	<<<0.001
Cancer, with breast cancer	0.016	0.002	0.172	0.879
Cardiac dysrhythmias	0.151	0.161	<<<0.001	<<<0.001
Cataract	0.044	0.041	<<<0.001	<<<0.001
Cerebrovascular disease	0.142	0.139	<<<0.001	<<<0.001
Coronary atherosclerosis	0.268	0.260	<<<0.001	<<<0.001
Diabetes mellitus	0.277	0.309	<<<0.001	<<<0.001
Esophagitis (GERD+)	0.139	0.122	<<<0.001	<<<0.001
Ever smoked	0.088	0.076	<<<0.001	<<<0.001
Fracture of upper limb	0.136	0.080	<<<0.001	<<<0.001
Hemorrhagic stroke	0.032	0.019	0.238	0.467
Hypertension	0.249	0.254	<<<0.001	<<<0.001
Hypothyroidism	0.074	0.047	<<<0.001	<<<0.001
Hypothyroidism (NOS)	0.075	0.061	<<<0.001	<<<0.001
ICH	-0.153	0.052	<<<0.001	0.031
Iron deficiency anemias	0.114	0.081	<<<0.001	<<<0.001
Ischemic Heart Disease	0.240	0.237	<<<0.001	<<<0.001
Ischemic stroke	0.135	0.108	<<<0.001	<<<0.001
Myocardial infarction	0.263	0.237	<<<0.001	<<<0.001
Obesity	0.237	0.240	<<<0.001	<<<0.001
Obstructive chronic bronchitis	0.278	0.216	<<<0.001	<<<0.001
Osteoarthritis	0.106	0.088	<<<0.001	<<<0.001
Pneumonia	0.148	0.137	<<<0.001	<<<0.001
Renal failure	0.168	0.162	<<<0.001	<<<0.001
SAH	0.013	0.035	0.641	0.207
Tobacco use disorder	0.224	0.231	<<<0.001	<<<0.001
Varicose veins	0.051	0.035	<<<0.001	<<<0.001
Death	0.122	0.089	<<<0.001	<<<0.001

Table S5.9: Discriminative capacities of PRS<sub>multi</sub> and EX-TERR PRS for each available outcome, with Age + Sex as the base model.

Table S5.10: **a. Continuous traits (regression coefficient, adjusted  $R^2$ .)** PRS results in validation set for baseline PRS, multi LDpred2 PRS and EX-TERR PRS (earth degree = 1). All results are highly significant (Rotation SD = 0.6, p-value  $\ll$  0.01)

Trait	Baseline		PRS <sub>multi</sub>		EX-TERR	
	Beta (SE)	Adj $R^2$	Beta (SE)	Adj $R^2$	Beta (SE)	Adj $R^2$
Alanine aminotransferase	0.0616 (0.00175)	0.00380	0.0669 (0.00175)	0.00447	0.0649 (0.00156)	0.00421
Albumin	0.0394 (0.00175)	0.00155	0.0532 (0.00175)	0.00283	0.0424 (0.00156)	0.00179
Alkaline phosphatase	-0.0604 (0.00175)	0.00365	0.0201 (0.00175)	0.000403	0.0238 (0.00156)	0.000563
Apolipoprotein A	-0.0474 (0.00175)	0.00224	0.0426 (0.00175)	0.00181	0.0445 (0.00156)	0.00198
Apolipoprotein B	0.144 (0.00173)	0.0209	0.0616 (0.00175)	0.00380	0.0517 (0.00156)	0.00267
Aspartate aminotransferase	0.0373 (0.00175)	0.00139	0.0291 (0.00175)	0.000843	0.0332 (0.00156)	0.00110
Bilirubin (direct)	-0.058 (0.00175)	0.00289	0.0513 (0.00175)	0.00263	0.0246 (0.00156)	0.000604
Bilirubin (total)	-0.0172 (0.00175)	0.000294	0.0135 (0.00175)	0.000180	0.00499 (0.00157)	0.0000225
Body Mass Index (BMI)	0.0988 (0.00174)	0.00975	0.158 (0.00173)	0.0251	0.165 (0.00154)	0.02711
Calcium	0.0389 (0.00175)	0.00151	0.0467 (0.00175)	0.00218	0.0474 (0.00156)	0.00225
Cholesterol	0.0910 (0.00174)	0.00829	0.307 (0.00167)	0.0945	0.0442 (0.00156)	0.00195
C-reactive protein	-0.0559 (0.00175)	0.00312	0.0359 (0.00175)	0.00129	0.0939 (0.00156)	0.00881
Creatinine	0.0625 (0.00175)	0.00390	0.0610 (0.00175)	0.00372	0.0611 (0.00156)	0.00373

Continued on next page

Table S5.10 – continued from previous page

Trait	Baseline		PRS <sub>multi</sub>		EX-TERR	
	Beta (SE)	Adj $R^2$	Beta (SE)	Adj $R^2$	Beta (SE)	Adj $R^2$
Cystatin C	0.0730 (0.00175)	0.00533	0.0973 (0.00174)	0.00946	0.0995 (0.00156)	0.00990
Diastolic Blood Pressure (DBP)	0.0501 (0.00175)	0.00250	0.0699 (0.00175)	0.00488	0.0578 (0.00156)	0.00333
Gamma glutamyltransferase	0.0580 (0.00175)	0.00336	0.0657 (0.00175)	0.00431	0.0635 (0.00156)	0.00403
Glucose	-0.0373 (0.00175)	0.00138	0.0487 (0.00175)	0.00237	0.0499 (0.00156)	0.00248
Glycated haemoglobin (HbA1c)	0.0637 (0.00175)	0.00406	0.0833 (0.00174)	0.00694	0.0846 (0.00156)	0.00715
High Density Lipoprotein (HDL)	-0.0471 (0.00175)	0.00221	0.182 (0.00172)	0.03316	0.0794 (0.00156)	0.00630
Height	-0.0499 (0.00175)	0.00248	0.0770 (0.00174)	0.00593	0.0637 (0.00156)	0.00406
Insulin-like Growth Factor 1 (IGF-1)	-0.0236 (0.00175)	0.000555	0.0481 (0.00175)	0.00231	0.0272 (0.00156)	0.000736
Low Density Lipoprotein (LDL) Cholesterol	0.108 (0.00174)	0.01175	0.0160 (0.00175)	0.000253	0.0171 (0.00157)	0.000290
Lipoprotein A	0.0811 (0.00174)	0.00657	0.0113 (0.00175)	0.000124	0.00808 (0.00156)	6.28 x 10 <sup>-5</sup>
Oestradiol	-0.00622 (0.00175)	0.0000357	0.00336 (0.00175)	8.20 x 10 <sup>-6</sup>	-0.00141 (0.00156)	-4.62 x 10 <sup>-7</sup>
Phosphate	0.0118 (0.00175)	0.000136	0.00710 (0.00175)	4.73 x 10 <sup>-5</sup>	0.00985 (0.00156)	9.46 x 10 <sup>-5</sup>
Protein (total)	0.0428 (0.00175)	0.00183	0.0498 (0.00175)	0.00248	0.0506 (0.00156)	0.00256
Rheumatoid Factor	0.00409 (0.00175)	0.0000136	-0.00357 (0.00175)	9.65 x 10 <sup>-6</sup>	0.00172 (0.00157)	5.22 x 10 <sup>-7</sup>
Systolic Blood Pressure (SBP)	0.0521 (0.00175)	0.00271	0.0845 (0.00174)			

Continued on next page

Table S5.10 – continued from previous page

Trait	Baseline		PRS <sub>multi</sub>		EX-TERR	
	Beta (SE)	Adj $R^2$	Beta (SE)	Adj $R^2$	Beta (SE)	Adj $R^2$
0.00713	0.0733 (0.00156)	0.00538				
Sex Hormone Binding Globulin (SHBG)	-0.104 (0.00174)	0.0108	0.115 (0.00174)	0.0133	0.105 (0.00156)	0.0110
Testosterone	-0.0431 (0.00175)	0.00185	0.0217 (0.00175)	0.000468	0.0359 (0.00156)	0.00128
Triglycerides	0.0427 (0.00175)	0.00312	0.0680 (0.00175)	0.00129	0.0513 (0.00156)	0.00263
Urate	0.195 (0.00172)	0.0364	0.199 (0.00172)	0.0395	0.239 (0.00152)	0.0569
Urea	-0.0880 (0.00174)	0.00775	0.0920 (0.00174)	0.00846	0.0899 (0.00156)	0.00808
Vitamin D	-0.0523 (0.00175)	0.00275	0.0644 (0.00175)	0.00414	0.0615 (0.00156)	0.00378
Waist-Hip Ratio (WHR)	0.0825 (0.00174)	0.00680	0.115 (0.00174)	0.0132	0.122 (0.00155)	0.0150

**b. Dichotomous traits (OR).** PRS results in validation sets for baseline PRS, multi LDpred2 PRS and EX-TERR PRS (earth degree = 1). All results are highly significant (Rotation SD = 1.0, p-value  $\lll 0.01$ )

Trait	Baseline OR	SD	PRS <sub>multi</sub> OR	SD	EX- TERR OR	SD
Atrial Fibrillation & Flutter (AFF)	1.218	0.00647	1.265	0.00650	1.274	0.00581
Asthma	1.207	0.00590	1.259	0.00592	1.116	0.00532
Cancer (no breast cancer)	1.048	0.00703	1.056	0.00704	1.045	0.00627
Cancer (suspected or other)	1.039	0.00496	1.031	0.00495	1.038	0.00444
Cancer (with breast cancer)	1.032	0.0121	1.000	0.0121	0.983	0.0109
Cardiac dysrhythmias	1.165	0.00530	1.219	0.00533	1.220	0.00477
Cataract	1.049	0.00530	1.055	0.00529	1.048	0.00474
Cerebrovascular disease	1.218	0.00782	1.193	0.00784	1.185	0.0153
Coronary Artery Disease (CAD)	1.248	0.00641	1.419	0.00650	1.400	0.0125
Coronary atherosclerosis	1.237	0.00646	1.390	0.00655	1.384	0.00703
Death	1.237	0.00628	1.162	0.00631	1.118	0.00580
Diabetes mellitus	1.285	0.00619	1.415	0.00621	1.456	0.00585
Esophagitis (GERD and re- lated diseases)	1.114	0.00499	1.180	0.00501	1.153	0.00564
Ever smoked	1.086	0.00362	1.116	0.00365	1.103	0.00558
Fracture of upper limb	1.040	0.00900	1.158	0.00900	1.104	0.00447
Hemorrhagic stroke	1.079	0.02678	1.036	0.0266	1.021	0.00325
Hypertension	1.201	0.00399	1.369	0.00407	1.377	0.00808
Hypothyroidism (NOS)	1.094	0.00715	1.103	0.00722	1.051	0.0238
Hypothyroidism	1.096	0.00731	1.102	0.00740	1.071	0.00365
Intracerebral Hemorrhage (ICH)	1.101	0.0234	1.023	0.02418	1.072	0.00645
Iron deficiency anemias	1.087	0.00817	1.156	0.00818	1.110	0.00659
Ischemic heart disease	1.211	0.00542	1.339	0.00549	1.349	0.0214
Ischemic stroke	1.125	0.0113	1.156	0.0113	1.142	0.00730
Myocardial infarction	1.237	0.00776	1.384	0.00784	1.349	0.00490

Continued on next page

Table S5.11 – continued from previous page

Trait	Baseline OR	SD	PRS <sub>multi</sub> OR	SD	EX- TERR OR	SD
Obesity	1.200	0.00650	1.342	0.00654	1.346	0.0101
Obstructive chronic bronchitis	1.288	0.00829	1.416	0.00837	1.308	0.00700
Osteoarthritis	1.109	0.00527	1.144	0.00528	1.108	0.00585
Pneumonia	1.135	0.00679	1.214	0.00681	1.196	0.00743
Renal failure	1.119	0.00636	1.239	0.00639	1.236	0.00471
Stroke	1.116	0.00897	1.189	0.00897	1.170	0.00610
Subarachnoid hemorrhage (SAH)	1.077	0.0279	0.996	0.0279	1.043	0.00573
Tobacco use disorder	1.232	0.00694	1.334	0.00701	1.346	0.00803
Type II diabetes (T2D)	1.318	0.00649	1.453	0.00651	1.487	0.0248
Varicose veins	1.061	0.00931	1.076	0.00922	1.032	0.00626

Table S5.12: Best-performing PRS after masking for each of the 69 outcomes in the validation set.

Trait	Best Performing PRS	Trait	Best Performing PRS
Continuous Traits		Dichotomous Traits	
Alanine aminotransferase	TG (GLGC)	Atrial Fibrillation (AF)	Heart failure
Albumin	TG (GLGC)	Asthma	COPD
Alkaline phosphatase	Venous thromboembolism	Cancer (suspected or other)	ADHD
Apolipoprotein A	Alzheimers disease	Cancer (no breast cancer)	COPD
Apolipoprotein B	Alzheimers disease	Cancer (with breast cancer)	ADHD
Aspartate aminotransferase	TG (GLGC)	Cardiac dysrhythmias	Heart failure
Bilirubin (direct)	LDL (GLGC)	Cataract	Stroke
Bilirubin (total)	LDL (MVP)	Cerebrovascular disease (CeD)	PAD
Body Mass Index (BMI)	Heart failure	Cerebrovascular disease (ever smoked)	Thyroid cancer
Calcium	TG (GLGC)	Cerebrovascular disease (never smoked)	COVID-19 (transethnic)
Cholesterol	LDL (GLGC)	Coronary Artery Disease (CAD)	PAD
C-reactive protein	Alzheimers disease	Coronary atherosclerosis	PAD
Creatinine	BUN	Death	COPD
Cystatin C	BUN	Diabetes mellitus	TG (GLGC)
Diastolic Blood Pressure (DBP)	Heart failure	Esophagitis (GERD and related diseases)	ADHD
Gamma glutamyltransferase	TG (GLGC)	Ever smoked	ADHD
Glucose	HDL (GLGC)	Fracture of upper limb	HDL
Glycated haemoglobin (HbA1c)	TG (GLGC)	Hemorrhagic stroke	PAD
High Density Lipoprotein (HDL) Cholesterol	PAD	Hypertension	Heart failure
Height	TG (GLGC)	Hypothyroidism	ADHD
Insulin-like Growth Factor 1 (IGF-1)	eGFR	Intracerebral Hemorrhage (ICH)	Alzheimer's disease
Low Density Lipoprotein (LDL) Cholesterol	Alzheimer's disease	Iron deficiency anemias	COPD

Continued on next page

Table S5.12 – continued from previous page

Trait	Best Performing PRS	Trait	Best Performing PRS
Continuous Traits		Dichotomous Traits	
Lipoprotein A	MI	Ischemic heart disease	PAD
Oestradiol	T2D	Ischemic stroke	Heart failure
Phosphate	Schizophrenia	Myocardial infarction	PAD
Protein (total)	TG (GLGC)	Hypothyroidism (NOS)	ADHD
Rheumatoid Factor	Crohns' Disase	Obesity	ADHD
Systolic Blood Pressure (SBP)	Heart failure	Obstructive chronic bronchitis	Asthma
Sex Hormone Binding Globulin (SHBG)	TG (GLGC)	Osteoarthrosis	ADHD
Testosterone	TG (GLGC)	Pneumonia	COPD
Triglycerides	T2D	Renal failure	Heart failure
Urate	Gout	Subarachnoid hemorrhage (SAH)	Abdominal aortic aneurysm
Urea	eGFR	Stroke	PAD
Vitamin D	TG (GLGC)	Type II Diabetes (T2D)	TG (GLGC)
Waist-Hip Ratio	Heart failure	Tobacco use disorder	ADHD
		Varicose veins	Heart failure

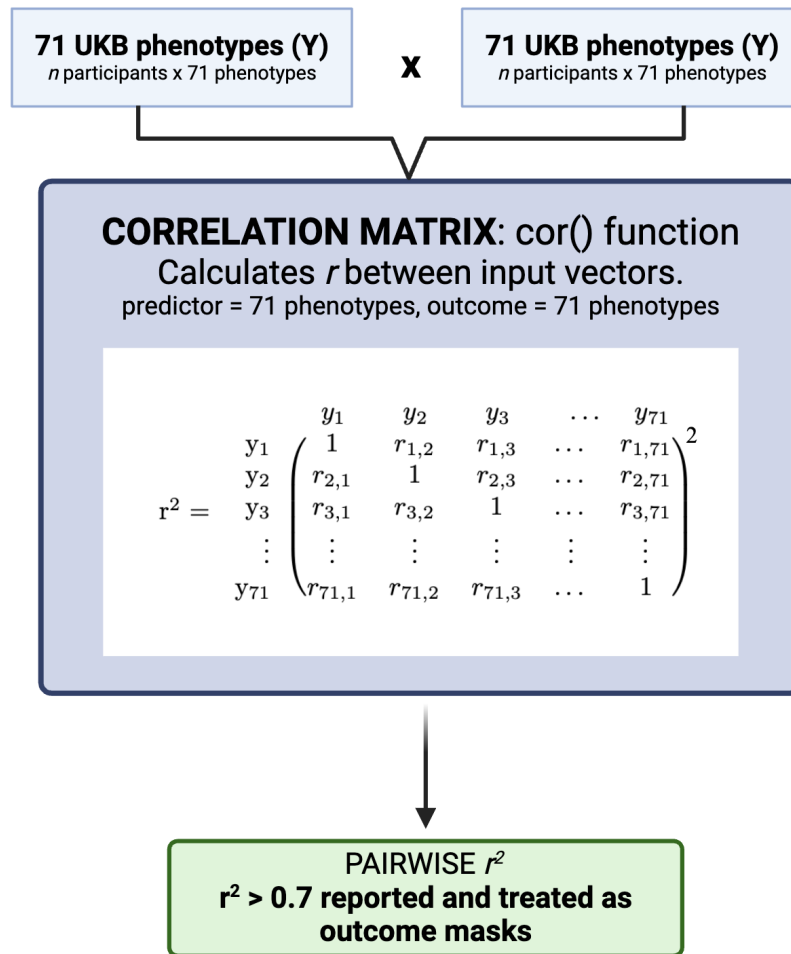


Figure S5.1: **Pipeline for assessing correlation of phenotypes for masking.** Process of determining correlation between outcomes to determine which traits should be masked. This was originally done for 71 outcomes (two were removed due to insufficient data.). The cor() function in R is used to determine pairwise correlation. Results are shown in Supplementary Table S3 and final masked GWAS are reported in Supplementary Table S4.

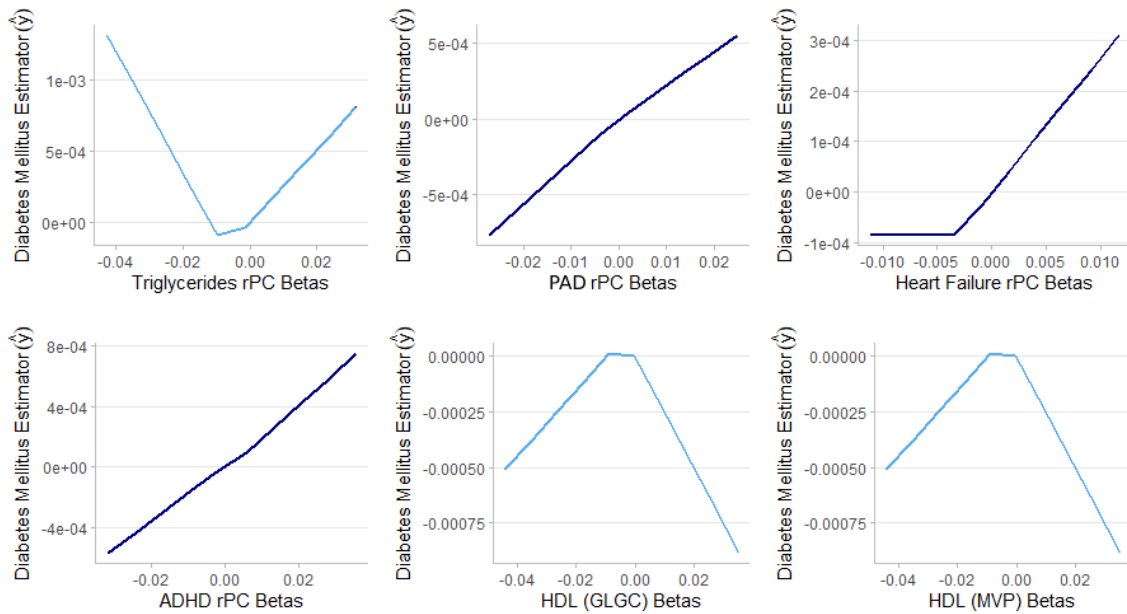


Figure S5.2: EX-TERR earth visualization of genetic effects on target outcome. Partial dependence plots (PDP) for the top six significant traits for a single cross-validation fold. For the continuous traits (TG HDLs), the peaks are center around (0,0) and reflect the general direction of the genetic effect of these predictors on the DM outcome. The TG partial dependence plot (PDP) shows positive DM values, with a negative slope at lower values and a positive slope at higher values. This suggests that as TG coefficients reach extreme values, the effect on DM increases, indicating a deleterious effect. Conversely, HDL coefficients show a decreasing effect or negative DM estimates at extreme values, implying a protective effect. For dichotomous traits such as peripheral artery disease (PAD), heart failure, and attention deficit hyperactivity disorder (ADHD), there is a positive correlation with DM estimators, indicating that there is an estimated increase of disease risk as the trait coefficients increase.

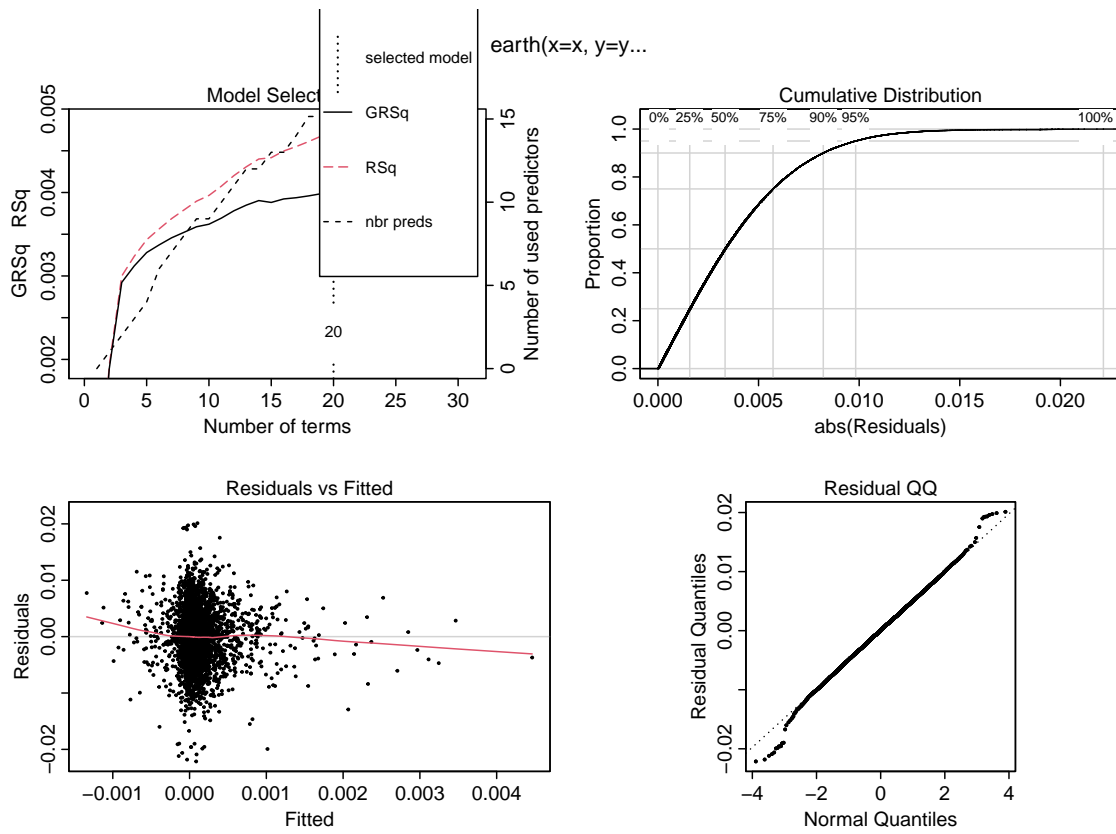
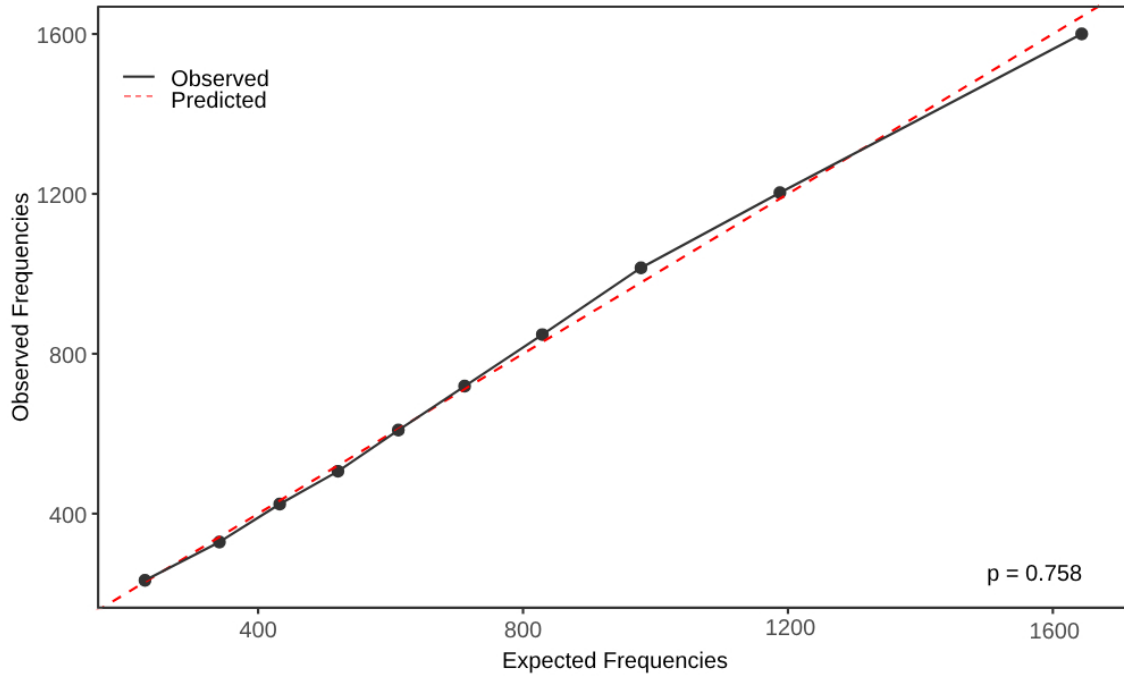


Figure S5.3: Simulation results for MARS algorithm as performed by EX-TERR in one fold for diabetes mellitus outcome (threshold = 0.6). a. Term selection b. Cumulative distribution of residuals c. Fitted vs. residuals d. QQ plot



	$\lambda^2/\text{HL}$ statistic	p-value
<b>Baseline PRS</b>	8.39	0.40**
<b>PRS<sub>multi</sub></b>	16.83	0.032
<b>EX-TERR</b>	4.99	0.76*

Figure S5.4: **Goodness of fit calibration results for EX-TERR PRS with diabetes mellitus (DM) outcome.** A Hosmer-Lemeshow plot depicting the calibration test between observed and expected values for the DM outcome and its corresponding EX-TERR PRS. The accompanying table reports results for DM outcome across the three methodologies. In this example, PRS<sub>multi</sub> is not significantly calibrated, while both baseline and EX-TERR PRS are significantly calibrated.

# Chapter 6

## Conclusion

### 6.1 General Overview

Cardiovascular diseases continue to be a significant global concern, and utilizing genetic data for risk prediction can play a crucial role in the prevention, detection, and treatment of these conditions. Polygenic risk scores (PRS) hold great clinical potential, as they rely solely on a patient's genotype, which can be obtained from an early age. Thus, PRS provide a non-invasive and cost-effective tool which can be used in tandem with current clinical risk predictors or alone. PRS are especially relevant to high prevalent cardiovascular diseases such coronary artery disease (CAD) or stroke, as these are highly heritable and there has been much evidence of genetic associations throughout the years. Additionally, because PRS only require a patient's genotypic information, they can be developed for numerous traits and phenotypes, facilitated by the recent advent of large-scale genome-wide association studies (GWAS). Throughout the years, PRS have demonstrated prediction performance at least equivalent to comparable clinical risk predictors, and continue to be

improved through methodologies incorporating machine learning and advanced statistical techniques. However, limitations remain with PRS which prevent them from being fully adapted into clinical practice. The performance of PRS greatly relies on the quality and scale of the GWAS available. Furthermore, GWAS generated within a given ancestry have proven to be less applicable to a different ancestry. This further poses a problem, as a large majority of GWAS are conducted within European ancestries. Additionally, there exists many traits for which a corresponding GWAS does not exist. This can be a great hindrance to PRS quality. While this issue could be mitigated through using GWAS pertaining to relevant risk factors to the target trait, this strategy would not address genetic architectures where trait-specific genetic pathways exist. This is especially prevalent when heritability estimates are unavailable, furthering the difficulty to discern between genetic and environmental influences within these complex traits. Therefore, the lack of GWAS can pose a major challenge in PRS optimization. Priorly, the issue of unavailable external trait-specific GWAS has not been thoroughly investigated, and there are no methodologies which specifically address it. Overall, this thesis addresses the use of PRS for risk prediction of cardiovascular traits. By exploring established genetic influences and pathways related to cardiovascular conditions, further advancements in PRS could yield new insights into biological and genetic pathways. Additionally, the aim is to overcome existing limitations and develop optimized, standardized PRS suitable for clinical application. This section provides a concise overview of research findings, clinical implications, current challenges, and future directions in PRS research for cardiovascular diseases.

## **6.2 Chapter 3 Overview**

Current research and literature pertaining to premature coronary artery disease (pCAD) was reviewed. pCAD refers to the early-onset condition of coronary artery disease (CAD)

occurring in patients younger than 65 years for women and 55 years for men. Both genetic and environmental influences on pCAD were investigated, along with the related conditions of clonal hematopoiesis of indeterminate potential (CHIP) and spontaneous coronary artery dissection (SCAD). Since pCAD is a condition of early-onset, it is highly attributable to genetic influences. pCAD is reported to be highly heritable as reported from twin and family studies. Monogenic causes are mainly attributed to familial hypercholesterolemia (elevated LDLc levels) and other dyslipidemias. Polygenic causes cited the first CAD GWAS in 2007 and the discovery of the 9p21 locus which is strongly associated with CAD. Rare variant studies have also shown promising results for pCAD. The advantages and disadvantages of PRS and clinical risk scores were also mentioned, as clinical risk scores are not particularly useful for pCAD due to their emphasis on age. Environmental risk factors include smoking, opioid usage, alcohol, amphetamines, stress and exercise. Finally, we reviewed conditions of similar manifestations to pCAD: clonal hematopoiesis and indeterminate potential (CHIP) and spontaneous coronary artery dissection (SCAD). Cumulatively, there is a need for a better understanding of pCAD, due to high societal burden of CAD in younger individuals.

### **6.3 Chapter 4 Overview**

Myocardial infarction after noncardiac surgery (MINS), the most prevalent vascular complication preceding surgical procedures, was the target outcome for applying polygenic risk score (PRS). The involvement of genetic influences was expected to improve predictive accuracy beyond the commonly used clinically risk predictor Revised Cardiac Risk Index (RCRI). The case-control study consisted of 506 matched patients nested within the international prospective representative Vascular Events in Noncardiac Surgery Participants Cohort Evaluation (VISION) cohort. Conditional logistic regression revealed significant

associations between MINS and the type 2 diabetes (T2D) PRS and the HbA1c PRS. No other PRS was associated with with MINS. Discrimination capacity was also observed for significant PRS. These results imply a multifactorial etiology of MINS, suggesting the potential influence of microvascular disease in MINS. However, this study was constrained by its smaller sample size, possibly resulting in a lower association due to limited statistical power. Additionally, no GWAS is currently available for MINS, which may overlook unique genetic architectures driven by MINS-specific genetic pathways distinct from known MINS risk factors. This warrants the need for better powered genetic studies, especially pertaining to MINS and MINS risk factors. Overall, the PRS results suggest potential links between perioperative glucose levels and MINS.

## 6.4 Chapter 5 Overview

Despite the growing popularity of polygenic risk scores (PRS) in recent years, there has been no clear guidance on the appropriate approach to take when a genome-wide association study (GWAS) for the target outcome is unavailable. We develop a novel method EX-TERR to address this issue, and compare with leading methods such as LDpred2 and elastic net regression multi PRS. Overall, we created PRS from 61 genome-wide association studies for 71 outcomes. Three PRS methods were observed: 1) a baseline, single-trait PRS, 2) a multi-trait PRS based in elastic net regression and 3) EX-TERR: a multi-trait PRS based in multi-adaptive regression splines (MARS). EX-TERR is based on MARS, a flexible, non-parametric and non-linear regression technique. EX-TERR is unique in that it does not require a train/test split for the original patient sample. This is due to a training and cross-validation step of components equivalent to genetic variants, mitigating the need for original participant train/test set. To simulate the condition of an unavailable GWAS, GWAS which had traits matching to a specific outcome were removed or “masked” from

the analysis. Within the 71 outcomes, multi-trait PRS outperformed single-trait PRS for 77% of total outcomes. However, the results also demonstrated that there is not a general optimal PRS methodology across all outcomes. Furthermore, significant decreases were indeed observed when the matching trait was masked relative to when they were left unmasked. Finally, it was noted that EX-TERR's performance was comparable to the current leading PRS methods.

## **6.5 Clinical & Research Implications**

### **6.5.1 Genetics Risk Prediction of Complex Traits in Clinical Settings**

Our pCAD review summarizes both genetic and non-genetic risk factors for CAD, including key monogenic and polygenic loci associated with the disease. Most prevalent disease traits are complex, arising from a combination of environmental factors and multiple genetic loci. Current clinical risk scores for cardiovascular risk typically overlook genetic risk factors and family history, focusing instead on clinical risk factors such as age and relevant biomarkers [1, 2]. Consequently, these scores are particularly limited for complex conditions with high heritability, as they may not be applicable to early-onset cases (e.g. younger individuals are automatically categorized as lower risk) and fail to account for genetic contributions. This highlights the importance of integrating PRS into clinical practices to account for the genetic component. Including family history with clinical risk scores has been shown to enhance risk assessment[3, 4, 5]. Similarly, the use of PRS in tandem with clinical risk predictors has shown to significantly improve prediction in CAD and related cardiovascular events.[6, 7]. Furthermore, PRS have also independently performed comparably to other genetic and non-genetic risk predictors[8, 9]. From a clinical perspective, PRS can be very beneficial as they only require the initial genotype information, which can be obtained

at an early age, and can be applicable to numerous traits. This can be convenient for disease prevention, particularly in early-onset cases, and may help circumvent challenges posed by long-term follow-up or the acquisition of clinical risk factors that can change over a lifetime (e.g. modifiable risk factors, biomarker levels, co-morbidities) [10]. The use of PRS has also been demonstrated to be cost-effective for CVD, which can be attributed to lowered genotyping costs over the years[11, 12]. While there can be several advantages to implementing PRS, a standard must first be established. There is further interest for PRS in precision medicine and stratification for drug treatments[3]. Each PRS is specifically catered to the patient, and can inform insights into treatment decisions. For examples, PRS have been conducted to assess major adverse cardiovascular events (MACE) and determine whether or not they should be treated with drugs for PCSK9 inhibition [13, 14]. It was concluded that patients concurring higher genetic risk may have greater benefit in drug treatment rather than those with low genetic risk.

### **6.5.2 Comparison of Different Leading PRS Methods**

Over the years, numerous PRS methods have been developed, each with their unique set of advantages. Our PRS analyses utilize the current leading techniques in tandem, allowing for a direct comparison of the preferred methodologies. During the development of the PRS project, simulations were conducted using the most frequently cited PRS methods from the past decade within the UK Biobank. The simulations encompassed both single-trait approaches and multi-trait techniques that employed advanced statistical methods and machine learning. Although not exhaustive, the methods tested included clumping and thresholding (C + T), LD adjustment within a window, forward selection, elastic net regression, LASSO regression, regionally correlated techniques, random forests, and Bayesian approaches. Essentially, a review of the currently relevant PRS methods was

conducted. Although multi-trait PRS generally outperformed single-trait PRS, no single method consistently emerged as the best across all outcomes. This insight should be taken into account in future PRS development, as it indicates that more complex methods do not necessarily yield better results. In some instances, the simplest methods are more than adequate and offer the advantages of greater efficiency and interpretability. Overall, the advantages and disadvantages of many PRS methods were explored, and the direct predictive accuracy was observed across many outcomes. It is important to recognize that there may not be a universally optimized PRS method, as different outcomes with distinct genetic architectures may require varying risk prediction approaches.

### **6.5.3 Availability of External GWAS for PRS**

The gap in predictive accuracy of PRS significantly differs when a matching external GWAS is available relative to when the matching GWAS is not available. While previous studies have consistently shown that PRS performance depends on the base GWAS, there is a lack of information on the best approach when an external matching GWAS is entirely unavailable. Myocardial infarction after non-cardiac surgery (MINS) is one such example. While GWAS for risk factors related to MINS were available, there is no GWAS directly corresponding to MINS. The evident difference in predictive accuracy was observed through several methodologies, incorporating both single-trait and multi-trait PRS. Multi-trait PRS outperformed single-trait PRS for a majority of outcomes tested. Thus, multi-trait PRS could be a potential solution to this problem. However, it should be acknowledged that there is not a single optimal approach for all outcomes. This could be due to underlying genetic architecture of the outcome. The main advantage of EX-TERR is that it does not require train/test split in the original sample. This can be especially advantageous when external GWAS are unavailable, as large-scale GWAS with strong associations are

necessary for optimizing PRS. This can also mitigate issues arising from smaller sample sizes. Furthermore, the MARS model underlying EX-TERR is non-parametric which can adapt to the datasets it is provided.

#### 6.5.4 Novel Insights for Genetic and Biological Pathways

While specific pathways were not thoroughly explored within this thesis, PRS methods may potentially reveal novel genetic and pathophysiological interactions between traits. For example, a previous PRS method in the Paré lab focussed on accounting for correlations between genomic regions [15]. Patient genotypes were divided into regions, after which the genetic correlations between these regions were observed for PRS adjustments. The study showed convincing evidence that there exists shared heritability between complex traits, and addresses assumption that all genomic regions will contribute proportionally in all outcomes. This is of particular relevance for multi-trait PRS. In addition to this finding, the study also affirmed various biological pathways, particularly. In particular, there was observed *LIPC* locus. There was observed correlation at the HDLc and CAD at this locus. Combining this with prior knowledge suggests that decreased *LIPC* activity may result in increased HDLc and CAD risk. *LIPC* affects intermediate-density lipoproteins, thus putting into the question the causal HDLc for CAD. Other studies have also uncovered interesting etiologies, such as a schizophrenia PRS being significantly increased for patients with a type 1 bipolar disorder relative to type 2 bipolar disorder.

EX-TERR also has capabilities in detecting genetic and biological associations, due to its consideration of genetic interactions between rotated regional components. The MARS model underlying EX-TERR demonstrates this potential by generating basis functions that directly capture the interactions between input predictors. Furthermore, these interactions can be visualized through built-in earth functions. While these were not explored in detail,

this is another aspect of EX-TERR which holds much potentially for PRS utility. Overall, PRS are not only useful for risk prediction, they can allow for novel insights into disease etiologies and pathophysiologically.

### **6.5.5 Extension to Other Diseases & Traits**

While the elements of this thesis are centralized around cardiovascular diseases, many concepts and techniques can be applied to other complex diseases. Furthermore, the PRS methods can be utilized with any GWAS for any outcome trait/disease. Unique characteristics of certain datasets may result in different performance in PRS methods. The development of EX-TERR establishes a foundation for further PRS development and optimization, especially as novel GWAS data become available. These methods may also be applied across different ethnicities to address the ongoing concern regarding the transferability GWAS to non-matching ancestries. Overall, the findings regarding external GWAS can be considered for future PRS development. The overall objective remains to establish a reliable standard for implementation of genetics into clinical risk prediction which is also applicable to global populations.

## **6.6 Limitations & Considerations**

The general limitations of PRS stems from the many assumptions in their application. The target trait or disease is not only assumed to have high heritability, but also have underlying polygenic characteristics. While many conditions and traits have been proven to be a product of multiple loci, attempting to apply PRS to monogenic or environmentally driven outcome will likely result in unnecessary noise and inaccurate conclusions. PRS may also fail to capture complex genetic interactions, such as gene-environment or gene-gene interactions. The challenges in establishing a clinical standard for PRS stem from how it

will applied and interpreted for a given patient. Since the calculated risk is relative, PRS do not necessarily guarantee disease onset. For instance, a patient with PRS in the top percentile is not guaranteed to develop the disease. This underscores the need for clinicians to apply their full understanding and judgment to use PRS in practice. Furthermore, increased screening resulting in false positives may induce misplaced stress in an individual, and have significant economic cost to society[16]. The complexities make it challenge to standard and adopt PRS in a clinical setting. The nature of the genetic risk captured by PRS must be effectively communicated to patients and their families, as genetic risks can be overemphasized over environmental impacts and lifestyle choices. There are additional challenges to PRS application. Since diseases are inherently complex, the use of PRS in conjunction with clinical risk scores must also be considered. In theory, the cost of calculating PRS for a patient is less than \$100[11]. However, this is offset by the fact that PRS are still in the clinical testing phase, with significant work needed before they become a standardized and easily accessible methodology. This would involve intensive clinical studies with relevant phenotypic data, along with longitudinal follow-up to assess prognosis or treatment outcome[16]. This can be quickly become expensive and challenging, as these studies are expected to be smaller in sample size. Additionally, PRS is a tool which requires acquisition and management of individual genetic information, which has raised ethical and privacy concerns[11, 17]. There are concerns for potential genetic discrimination and uninformed usage of private genetic data. Furthermore, given that no single PRS has proven to be universally optimized, established a standardized approach for specific disease categories may prove challenging. Currently, there remains no established standards or guidelines for clinical application of PRS.

As reinforced by the findings within this thesis, PRS remain dependent on available GWAS studies. There is a clear gap in performance when an external GWAS is unavail-

able, demonstrating the significance of large-scale, high quality GWAS which matches the target outcome. Moreover, with GWAS being regularly published, PRS would need to be continually updated to reflect the most updated findings. Despite being a commonly referenced issue, there is still no specific method or strategy to address the limited ancestry representation in current GWAS. This unresolved issue hinders the global application of PRS in clinical practice, as nearly all ancestries, except for European, are underrepresented in GWAS. Moreover, an underpowered GWAS can lead to limitations in statistical power, particularly when only small, localized samples are available. Within the framework of this thesis, all methods and analyses were likewise conducted purely on European populations (VISION, UK Biobank, European-based GWAS). Thus, the applicability of the current findings to global populations has not yet been established.

There are unique limitations to each PRS methodology. Single-trait PRS, while easy to interpret and simple in nature, does not address the possible interactions between the risk factors. Therefore, it might fail to capture pleiotropy and confounders. Alternatively, multi-trait PRS should enhance predictive power, offering broader applicability and improved generalizability for complex diseases with multiple influences and risk factors. However, the use of multi-trait PRS may be excessive and capture unnecessary noise, especially for outcomes which are highly specific. The structure of the dataset may also impact PRS performance. For example, some methods may work better with continuous outcomes and others with dichotomous outcomes. Naturally, the available sample sizes and scale of GWAS will also influence the choice of methodology, as some methods may be better suited in situations with limited statistical power. As previously noted, there does not appear to be a single PRS method which is optimized for all outcomes. This is logical, considering that every trait is inherently unique, with differing genetic architectures, levels of heritability and related risk factors. It is somewhat unrealistic to take all these factors

into consideration every time a PRS is created for a patient. However, this further highlights the benefits of studies like ours which encompass multiple PRS methodologies, as they help identify the strengths and weaknesses of each method across a range of clinically relevant outcomes.

While EX-TERR, based on the multi-adaptive regression spline (MARS) approach, provides several unique advantages as a PRS methodology, it also has its own set of limitations[18]. Firstly, while EX-TERR's plentitude of parameters allow for high flexibility and adaptation to the user's input dataset, this may actually result in overfitting. MARS fits linear data segments according to basis functions, and while there is a penalty for adding more knots in the model, this can be overridden and is also dependent the inputted dataset. For instance, MARS is not expected to perform as effectively with smaller datasets or datasets with higher dimensionality, which are more susceptible to the influence of outliers. The numerous parameters that must be tuned can lead to inconsistent performance. Furthermore, MARS operates on a greedy forward pass followed by a backwards pass, making it highly computationally intensive especially when compared to alternate methods (e.g. single-trait, random foresting, elastic net). Similar to other multi-trait methods, MARS does variable selection based on importance, excluding ones that it considers irrelevant. However, MARS will not automatically consider correlation between input predictors, which may introduce arbitrariness to results. While the bias-variance trade-off is considered decent in MARS, the rigidity of basis functions may be attributed to lower variance in each separate MARS model. While a benefit of MARS is its ability to determine and showcase interactions between multiple predictors and their respective outcome, it can be difficult to interpret the relative contribution of each predictor to the outcome. This could produce challenges in determining meaningful interactions between variables, especially under the context of biological or genetic etiologies. Finally, since

MARS is a regression model, it is not inherently suited for categorical or binary outcomes. Although it is designed to handle both categorical and continuous data, its design is more suited to numeric or continuous outcomes, due to the nature of the basis functions it is built upon.

## 6.7 Conclusion

Polygenic risk scores have garnered significant attention within the recent years for their potential in clinical applications and precision medicine. This thesis provides an in depth overview with accompanying analyses for PRS in the context of cardiovascular diseases. The breadth and capabilities of PRS were showcased, with promising implications for clinical adaption of PRS. We provide clear evidence that an external GWAS for the matching trait is crucial to optimized PRS performance. The findings this thesis raise important insights for the future development of PRS, and pave the way for future directions in terms of novel biological insights and approaches to clinical implementation. Regardless, there remains a lot of work for PRS before PRS can become standard practice. Future PRS methods should seek to address the limited ancestries covered by GWAS, potential gene-gene or gene-environmental interactions and investigations into how different PRS models perform on various outcomes. Regardless, with additional informed research, the findings in this paper have the potential to eventually guide and influence clinical practice, specifically under the context of cardiovascular diseases.

## References

- [1] R. B. D'Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6):743753, 2008. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.107.699579.
- [2] C. Koz, O. Baysan, A. Hasimi, M. Cihan, et al. Conventional and non-conventional coronary risk factors in male premature coronary artery disease patients already having a low Framingham risk score. *Acta Cardiologica*, 63(5):623628, 2008. ISSN 0001-5385. doi: 10.2143/AC.63.5.2033231.
- [3] S. Koch, J. Schmidtke, M. Krawczak, and A. Caliebe. Clinical utility of polygenic risk scores: a critical 2023 appraisal. *Journal of Community Genetics*, 14(5):471487, October 2023. ISSN 1868-310X. doi: 10.1007/s12687-023-00645-z.
- [4] M. T. Scheuner, W. C. Whitworth, H. McGruder, P. W. Yoon, et al. Expanding the definition of a positive family history for early-onset coronary heart disease. *Genetics in Medicine*, 8(8). ISSN 1098-3600. doi: 10.1097/01.gim.0000232582.91028.03.
- [5] S. Sivapalaratnam, S.M. Boekholdt, M.D. Trip, M.S. Sandhu, et al. Family history of premature coronary heart disease and risk prediction in the EPIC-Norfolk prospective population study. *Heart (British Cardiac Society)*, 96(24):19851989, 2010. ISSN 1355-6037. doi: 10.1136/hrt.2010.210740.
- [6] J. Elliott, B. Bodinier, T. A. Bond, M. Chadeau-Hyam, et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA*, 323(7):636645, 2020. ISSN 0098-7484. doi: 10.1001/jama.2019.22241.

- [7] N. J. Samani, E. Beeston, C. Greengrass, F. Riveros-McKay, et al. Polygenic risk score adds to a clinical risk score in the prediction of cardiovascular disease in a clinical setting. *European Heart Journal*, page ehae342, 2024. ISSN 0195-668X. doi: 10.1093/eurheartj/ehae342.
- [8] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(99):12191224, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0183-z.
- [9] A. M. Small, G. E. M. Melloni, F. K. Kamanu, B. A. Bergmark, et al. Novel polygenic risk score and established clinical risk factors for risk estimation of aortic stenosis. *JAMA Cardiology*, 9(4):357366, April 2024. ISSN 2380-6583. doi: 10.1001/jamacardio.2024.0011.
- [10] D. Klarin and P. Natarajan. Clinical utility of polygenic risk scores for coronary artery disease. *Nature reviews. Cardiology*, 19(5):291301, 2022. ISSN 1759-5002. doi: 10.1038/s41569-021-00638-w.
- [11] A. C. Fahed, A. A. Philippakis, and A. V. Khera. The potential of polygenic scores to improve cost and efficiency of clinical trials. *Nature Communications*, 13(1):2922, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30675-z.
- [12] M. Kiflen, A. Le, S. Mao, R. Lali, et al. Cost-Effectiveness of Polygenic Risk Scores to Guide Statin Therapy for Cardiovascular Disease Prevention. *Circulation: Genomic and Precision Medicine*, 15(5):e003423, 2022. doi: 10.1161/CIRCGEN.121.003423.
- [13] Lipoprotein(a), pcsk9 inhibition, and cardiovascular risk. 139(12). ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.118.037184.

- [14] N. Mars, J. T. Koskela, P. Ripatti, T. T. J. Kiiskinen, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine*, 26(44):549557, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0800-0.
- [15] G. Paré, S. Mao, and W. Q. Deng. A robust method to estimate regional polygenic correlation under misspecified linkage disequilibrium structure. *Genetic Epidemiology*, (7):636647, 2018. ISSN 1098-2272. doi: 10.1002/gepi.22149.
- [16] C. M. Lewis and E. Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, 12(1):44, 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-00742-5.
- [17] A. V. Khera and S. Kathiresan. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature Reviews Genetics*, 18(66):331344, 2017. ISSN 1471-0064. doi: 10.1038/nrg.2016.160.
- [18] Friedman J. H. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1991.