Prediction of coronary heart disease risk using polygenic risk scores in a Mexican population

Nuffield Department of Population Health

Tianshu Liu, BSc(Hons) MSc

St Cross College, University of Oxford

Supervisors

Dr Louisa Gnatiuc Friedrichs, Dr Jason Torres, Professor Jonathan Emberson

Word count: 45,387

Thesis submitted for the degree of Doctor of Philosophy

September, 2025

Abstract

Background: Coronary heart disease (CHD) is the global leading cause of death with a high genetic heritability and polygenic predisposition. Polygenic risk scores (PRSs) have been found to enhance CHD risk prediction in populations of European ancestry but evidence in other populations is limited.

Thesis aims: 1) To conduct a literature review on PRSs published for CHD risk prediction. 2) To evaluate the transferability of eight external CHD PRSs in the Mexico City Prospective Study (MCPS). 3) To construct a novel MCPS-informed CHD PRS with greater representation of admixed-American ancestry. 4) To evaluate the predictive ability of integrated scores (combining genetic and clinical risk) for CHD.

Methods: The Mexico City Prospective Study (MCPS) is a blood-based prospective cohort study with over 150,000 participants recruited between 1998 and 2004, aged over 35 years at the time of recruitment. The participants have been followed up through the National Death Registry. Imputed genotype data are available for over 140,000 participants.

133,207 MCPS participants aged between 35 and 79 years at recruitment and with genetic data available were included in the main analyses. CHD was defined as self-reported diagnosis at baseline or death before age 80 with CHD listed as the primary or as a contributory cause of death. Logistic regression was used to assess the strength of association between each of the eight selected external CHD PRSs (comprising 44 to 6,472,620 single nucleotide polymorphisms [SNPs]) and CHD, adjusted for age, sex and the first seven genetic principal components (PCs). Discrimination was assessed using the area under the receiver-operating-characteristic curve (AUC).

The novel MCPS-informed PRS was trained on 80% of the MCPS data, using a 10-fold cross validation (CV) approach, and evaluated three PRS construction methods with thousands of hyperparameter combinations (Pruning and Thresholding [P+T], LDpred2 and PRS-CSx). Genome-

wide association study (GWAS) analyses on CHD were conducted using MCPS data and meta-analysed with external GWAS results to boost sample size and incorporate genetic information from other ancestries, to be used as input for PRS construction algorithms. The best-performing candidate PRS of each method was then tested on the remaining 20% of the MCPS individuals that were not included for the PRS construction, using logistic regression with no other adjustments.

The best-performing external CHD PRS was combined with three guideline-recommended clinical risk scores (two variations of SCORE2 and the Pooled Cohort Equation [PCE]) into integrated risk scores (clinical risk score x PRS). Each integrated score was evaluated for its performance by comparing it to the original clinical risk score from which it was derived, using logistic regression with no other adjustments and the net reclassification index (NRI).

Results: Over two-thirds of the included participants were women and the mean baseline age was 51 years (standard deviation [SD] 12 years). In the analysis of external PRSs, all eight PRSs were positively and log-linearly associated with CHD, with odds ratios (ORs) per SD ranging from 1.05 (95% confidence interval [CI], 1.03-1.08) to 1.29 (1.25-1.33). Further adjustment for conventional CHD risk factors did not attenuate these associations materially. Multi-ancestry PRSs generally outperformed Eurocentric single-ancestry PRSs. Overall, PRSs only improved model discrimination marginally. Significant sex heterogeneity was identified for six of the eight PRSs, with PRSs consistently showing stronger associations in men compared to women (e.g., OR per SD was 1.37 [95%CI, 1.32-1.43] in men and 1.23 [1.18-1.28] in women for the Patel et al. PRS, which was the PRS with the strongest overall association with CHD).

During the training of the candidate PRSs, multi-ancestry PRSs showed greater association strength and predictive ability. The novel MCPS-informed PRS was constructed using the PRS-CSx method, including 1,180,546 SNPs and leveraging genetic information from three ancestries (i.e., admixed-Americans, Europeans and East-Asians). The MCPS-PRS was strongly associated with CHD when tested on the other 20% of the MCPS data, with an OR per SD of

1.34 (95%CI, 1.26-1.42), which was comparable to the best-performing external CHD PRS. The novel MCPS-informed score included eight times as many admixed-American individuals during its construction as the next most representative admixed-American CHD PRS.

Compared to the clinical risk scores alone, the integrated risk scores predicted a higher CHD risk for participants in the top 20% risk stratum relative to the lowest 20% group (e.g., OR=12.92 for SCORE2-diabetes-recalibrated vs OR=13.66 for SCORE2-diabetes-recalibrated x PRS). In addition, based on a 7.5% CHD risk threshold, this integrated score improved CHD reclassification significantly, with categorical NRI of 1.50% (95% CI, 0.73%-2.27%) and continuous NRI of 17.50% (14.75%-20.24%) compared to SCORE2-diabetes-recalibrated. If extrapolated to the whole of Mexico, this integrated score has the potential to reclassify correctly approximately 30,000 more CHD cases among adults in Mexico.

Conclusion: The selected external PRSs have reasonable transferability among Mexicans and could predict CHD risk independently of conventional CHD risk factors. Sex heterogeneity was identified in genetic predisposition to CHD with men showing greater risk than women. A novel multi-ancestry CHD PRS was developed, leveraging genetic information from three ancestries and including the largest number of admixed-Americans to date. When combining a PRS with a clinical risk score, the integrated risk scores improved CHD risk classification and have the potential to improve early detection of CHD and hence its prevention. Further studies are needed to enhance the representation of women and admixed-Americans in genetic research, and to improve the calibration of clinical risk scores for admixed-American populations.

Statement of contribution to work

The original idea for this thesis was proposed by my supervisors, Professor Jonathan Emberson, Dr Louisa Gnatiuc Friedrichs, and Dr Jason Torres. Under their supervision, I formulated the research questions, carried out the literature review, designed the research plans, and conducted the analyses.

The genetic data used for analyses in this thesis were processed, quality controlled (QCed), and imputed by Dr Jason Torres. The genetic PCs used for analysis in Chapter 4 were also produced by Dr Jason Torres. External PRSs were generated for MCPS participants using a publicly available pipeline developed by the PGS Catalog. The trend test analysis (presented in forest plots) in Chapter 4 was developed from a script provided by Rachel Wade. The GWAS analysis conducted in Chapter 5 (to be used as input for PRS construction) was developed from workflow scripts provided by Dr Jason Torres. For the PRSs generated in Chapter 5, I adapted the software scripts of each PRS algorithm to the MCPS data. The clinical risk scores in Chapter 6 were generated using the R package RiskScorescvd.

All statistical analyses, PRS training and testing, and the generation of tables and figures presented in this thesis were performed independently by myself. I drafted and refined this thesis with guidance from my supervisors.

Publications

1. **Liu T**, Berumen J, Torres J, Alegre-Díaz J, Baca P, González-Carballo C, et al. Polygenic prediction of coronary heart disease among 130,000 Mexican adults. *medRxiv*. 2024:2024.2012.2020.24319332.
2. Bragg F, Kuri-Morales P, Trichia E, Torres JM, Baca P, Garcilazo-Ávila A, **… Liu T,** et al. Type 2 diabetes and cause-specific mortality in Mexico City: a Mendelian randomisation analysis. ***The Lancet Regional Health - Americas***. 2025;45:101082.
3. Gnatiuc Friedrichs L, Kuri-Morales P, Trichia E, Staplin N, Torres J, Alegre-Díaz J, **…, Liu T,** et al. A Mendelian randomization study of the effect of body mass index on 52 causes of death among 125000 Mexican adults with admixed ancestry. ***International Journal of Epidemiology***. 2025;54(4).

Acknowledgements

I would like to begin by expressing my gratitude to the 150,000 MCPS participants, whose commitment to this long-term research made this thesis possible. I am equally thankful to the dedicated fieldworkers who, in the pre-digital era, went door-to-door to recruit participants and collect their information with care. I also would like to extend my thanks to the members of the MCPS management team, who have been dedicated to data verification, conducting resurveys, and ensuring the long-term continuity and quality of this remarkable cohort.

I am also grateful to my supervisors, Jonathan Emberson, Louisa Gnatiuc Friedrichs and Jason Torres, who have supported and guided me throughout my DPhil, offering expert knowledge from different perspectives and providing insightful feedback on this thesis. I would like to express my special thanks to Louisa, who first introduced me to MCPS and this DPhil project in 2021. Thanks are also due to other members of the MCPS team in the Nuffield Department of Population Health (NDPH), who have provided me with both support and joy during my DPhil work. This includes Lisa Holland, Rachel Wade, Eirini Trichia, Michael Turner, and Diego Aguilar Ramirez, as well as Alejandra Vergara-Lope, Tabitha Hubbard, Luisa Fernández Chirino, and Qinfang Lu, who have also become my close friends. I am also deeply thankful to the Clarendon Fund and NDPH for providing me with a three-year full scholarship, enabling me to undertake this DPhil.

Finally, I am forever grateful to my friends and family who have always been supportive throughout the ups and downs of my DPhil journey. To my childhood friend Yuqiu Ye, and to my best university friends Lin Gan and Xingdi Wang, thank you for being constant sources of encouragement and perspective. To my wonderful Oxford Badminton teammates, Yi Liang, Yuying Ding, Hui Min Tay, Ziwei Ye, Chuanqi Wang, and Lily Yang, thank you for pulling me out of my academic bubble and reminding me to enjoy life beyond research. To my parents (Shuren Liu and Hong Zhu) and uncle (Professor Tong Zhang), who sparked my interest in science and supported my decision to pursue a DPhil, your love and support have been my greatest strength. Last but not least, to my late grandparents, you have always inspired me, and I hope I have made you proud.

List of abbreviations

ASCVD	Atherosclerotic Cardiovascular Disease
AUC	Area Under the receiver-operating-characteristic Curve
BMI	Body Mass Index
P+T	Pruning and Thresholding
CABG	Coronary Artery Bypass Graft
CC4D	CARDIoGRAMplusC4D
CHD	Coronary heart disease
CI	Confidence Interval
CKB	China Kadoorie Biobank
CV	Cross Validation
CVd	Cardiovascular Disease
DALYs	Disability-Adjusted Life Year
DBP	Diastolic Blood Pressure
DNA	Deoxyribonucleic Acid
EHR	Electronic Health Record
FE	Fixed Effect
FRS	Framingham Risk Score
GSA	Global Screening Array
GWAS	Genome Wide Association Study
HDL-C	High Density Lipoprotein Cholesterol
HGDP	Human Genome Diversity Project
HIS	Admixed-American GWAS meta-analysis
HR	Hazard Ratio
IBD	Identity-By-Descent
ICD-10	International Classification of Diseases 10th Revision
IHD	Ischaemic Heart Disease (=CHD)
INEGI	the Mexican National Institute of Statistics and Geography
IVW	Inverse-Variance Weighted
JAP	Japanese CHD GWAS
LD	Linkage Disequilibrium
LDL-C	Low Density Lipoprotein Cholesterol
LDSC	LD score regression
MAF	Minor Allele Frequency
MCMC	Markov Chain Monte Carlo
MCPS	Mexico City Prospective Study
MDCS	Malmo Diet and Cancer Study

MeSH	Medical Subject Headings
MEGA	Multi-Ethnic Genotyping Arrays
MI	Myocardial Infarction
MVP	Million Veteran Program
NCD	Non-communicable disease
NRI	Net Reclassification Index
OR	Odds Ratio
PC	Principal Component
PCE	Pooled Cohort Equation
PCI	Percutaneous Coronary Intervention
PRS	Polygenic Risk Score
PTCA	Percutaneous Transluminal Coronary Angioplasty
QC	Quality Control
QALYs	Quality-Adjusted Life Years
RE	Random Effect
RFRS	Framingham Risk Score
RR	Risk Ratio
SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Standard Error
SNP	Single Nucleotide Polymorphisms
TDI	Townsend Deprivation Index
TG	Triglycerides
TOPMed	Trans-Omics for Precision Medicine
Total-C	Total Cholesterol
UKB	UK Biobank
USA	United State of America
WES	Whole Exome Sequencing
WHR	Waist-to-Hip Ratio
YLL	Years of Life Lost

Contents

1 Chapter 1: Introduction **17**

1.1 Coronary heart disease epidemiology and global prevalence 19
1.2 Genetic studies of CHD 20
1.2.1 CHD genetic discovery 20
1.3 Polygenic risk scores for the assessment of CHD risk 22
1.3.1 From GWAS discovery to polygenic risk scores for CHD risk assessment . 22
1.3.2 PRS development 23
1.4 Conventional risk factors for CHD 24
1.4.1 CHD risk scores integrating genetic and clinical information 24
1.5 CHD burden in Mexico 25
1.5.1 Admixed-American populations 25
1.5.2 CHD burden and epidemiology in Mexican populations 26
1.5.3 The Mexico City Prospective Study 27
1.6 DPhil objectives 27
1.6.1 Research aims 27
1.6.2 Thesis structure 28
1.7 References 31

2 Chapter 2: Literature Review **36**

2.1 Background and aims 37
2.2 Method of literature review and search 38
2.2.1 Search strategies 38
2.2.2 Selection of relevant research papers and review 38
2.2.3 Review of eligible studies 40
2.3 Findings from the literature review 40
2.3.1 Literature review results 40
2.3.2 CHD GWAS 41
2.3.3 General characteristics of relevant studies identified from literature review . 42
2.3.4 Performance of CHD PRSs among populations of European ancestry . . . 43
2.3.5 Performance of CHD PRSs among populations of non-European ancestries 45
2.3.6 Admixed-American-sourced CHD PRS 47
2.3.7 PRS construction methodology 48
2.4 Discussion 50
2.4.1 Performance of CHD PRS in population of European ancestry 50
2.4.2 Performance of CHD PRS in populations of non-European ancestry 50
2.4.3 Current methods for PRS construction 51
2.4.4 CHD PRS constructed with admixed-American genetic information 52
2.5 Conclusion 53

2.6 Selection of pre-existing PRSs for evaluation in the Mexico City Prospective Study cohort . . . 53
2.6.1 PRS selection for evaluation in MCPS . . . 54
2.7 References . . . 62

3 Chapter 3: Methods 68
3.1 The Mexico City Prospective Study . . . 69
3.1.1 Design and recruitment . . . 69
3.1.2 Baseline assessment procedure . . . 69
3.1.3 Physical measurements . . . 70
3.1.4 Blood sample collection and processing . . . 70
3.1.5 Follow-up for mortality . . . 71
3.1.6 Genotyping, quality control and imputation . . . 71
3.1.7 Funding and ethics approvals . . . 75
3.2 Epidemiological methods . . . 75
3.2.1 Analysis cohort . . . 75
3.2.2 Primary outcome definition . . . 76
3.2.3 Other CHD-related outcomes for secondary analyses . . . 76
3.3 Polygenic risk scores for CHD . . . 76
3.3.1 Polygenic risk scores computation . . . 76
3.3.2 Genome-wide association study . . . 77
3.4 Statistical methods for investigating the association between polygenic predisposition to CHD and sub-sequent risk for CHD . . . 78
3.4.1 Logistic regression models for the assessment of CHD risk . . . 78
3.4.2 Secondary and sensitivity analysis . . . 79
3.4.3 Handling of missing data . . . 79
3.5 Conclusion . . . 80
3.6 References . . . 84

4 Chapter 4: Transferability of previously published polygenic risk scores for coronary heart disease risk assessment in Mexicans 86
4.1 Background and aims . . . 87
4.2 Methods . . . 88
4.2.1 Selection of PRSs for evaluation . . . 88
4.2.2 Recreation of CHD PRSs for assessment of CHD risk in the MCPS population 92
4.2.3 Statistical analysis . . . 93
4.3 Results . . . 95
4.3.1 Participants included . . . 95
4.3.2 Baseline characteristics of included MCPS participants . . . 95
4.3.3 Fatal and non-fatal CHD cases . . . 96
4.3.4 Characteristics of the study participants by different levels of inherited genetic-risk for CHD . . . 97
4.3.5 Association between CHD PRS and subsequent risk of CHD in MCPS . . . 97
4.3.6 Differences in genetic predisposition to CHD risk by sex . . . 99

4.3.7 Genetic predisposition to CHD risk by other characteristics of the MCPS population . . . 99
4.3.8 Genetic predisposition to CHD, given various definitions and relatedness . 100
4.4 Discussion . . . 101
4.4.1 Summary of findings . . . 101
4.4.2 Comparison of findings in the MCPS with the PRS-source studies . . . 102
4.4.3 Sex heterogeneity in genetic predisposition to CHD risk . . . 102
4.4.4 Genetic predisposition to CHD and other conventional CHD risk factors . . 103
4.4.5 The relevance of multi-ancestry PRS to CHD risk . . . 103
4.4.6 Conclusion . . . 104
4.5 References . . . 144

5 Chapter 5: Developing a novel multi-ancestry coronary heart disease polygenic risk score for an admixed-American population 147
5.1 Background and aims . . . 148
5.2 Methods . . . 150
5.2.1 PRS construction overview . . . 150
5.2.2 Genome-wide association study . . . 152
5.2.3 External datasets used for GWAS meta-analysis . . . 154
5.2.4 Lift-over . . . 163
5.2.5 GWAS meta-analysis . . . 163
5.2.6 PRS construction methods . . . 165
5.2.7 PRS training . . . 171
5.2.8 PRS testing . . . 172
5.3 Results . . . 173
5.3.1 PRS training . . . 173
5.3.2 PRS testing . . . 176
5.4 Discussion . . . 177
5.4.1 Summary of findings . . . 177
5.4.2 Limitations of PRS construction methods developed for single ancestry populations . . . 178
5.4.3 Multi-ancestry PRS construction method . . . 179
5.4.4 Comparing the MCPS-informed PRS with external CHD PRS . . . 180
5.5 Conclusion . . . 181
5.6 References . . . 198

6 Chapter 6: Integration of polygenic and clinical risk scores to improve overall prediction of coronary heart disease risk 203
6.1 Background and aims . . . 204
6.2 Methods . . . 205
6.2.1 Clinical risk score . . . 206
6.2.2 Polygenic risk score for CHD . . . 209
6.2.3 Integration of PRS with clinical risk scores . . . 209
6.2.4 Statistical analysis . . . 211
6.3 Results . . . 213

6.3.1 Association between clinical risk score and CHD risk . . . 214
6.3.2 Association between integrated score and CHD risk . . . 215
6.3.3 Difference in clinical risk by sex and baseline diabetes status . . . 217
6.3.4 Genetic predisposition to CHD risk across clinical risk score strata . . . 218
6.3.5 Sensitivity analysis . . . 218
6.4 Discussion . . . 219
6.4.1 Summary of findings . . . 219
6.4.2 Performance of external non-admixed-American based clinical risk score in MCPS . . . 220
6.4.3 Combining genetic and clinical risk predictors to enhance CHD prediction in clinical settings . . . 221
6.4.4 Potential impact of the implementation of a genetically-enhanced CHD risk score at national level in Mexico . . . 223
6.5 Conclusion . . . 224
6.6 References . . . 252

7 Chapter 7: Discussion 255
7.1 Thesis overview . . . 255
7.2 The relevance of previously-published PRSs to CHD risk in Mexicans . . . 256
7.2.1 Summary of the main findings . . . 256
7.2.2 Finding in context . . . 258
7.3 Developing a novel CHD PRS leveraging genetic data from MCPS . . . 260
7.3.1 Summary of the main findings . . . 260
7.3.2 Findings in context . . . 262
7.4 Clinical and genetically-integrated risk scores and risk for CHD in Mexicans . . . 264
7.4.1 Summary of the main findings . . . 264
7.4.2 Findings in context . . . 266
7.5 Strengths and limitations . . . 268
7.5.1 Strengths . . . 268
7.5.2 Limitations . . . 270
7.6 Implication for public health . . . 271
7.7 Future research . . . 272
7.8 Conclusion . . . 273
7.9 References . . . 275

8 Appendix 278
8.1 References . . . 309

List of Tables

2.1 Systematic Search Strategy . . . 55
2.2 Studies with population of admixed-American ancestry . . . 57
2.3 Associations between popular CHD PRS and CHD reported by external evaluation studies . . . 59
2.4 CHD polygenic risk scores selected for further evaluation . . . 60

3.1 Key data collected in the MCPS . . . 81
3.2 Baseline characteristics of selected socio-demographic characteristics of all recruited participants by sex . . . 82

4.1 CHD polygenic risk score selected for evaluation in MCPS . . . 106
4.2 Baseline Characteristics of 133,207 participants aged 35-79 years (main analysis population) . . . 107
4.3 Baseline characteristics of 137,391 participants aged 35-89 years . . . 108
4.4 CHD case type by ICD-10 codes and diagnosis . . . 109
4.5 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Tada *et al.* . . . 110
4.6 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Oni-Orisan *et al.* . . . 111
4.7 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Koyama *et al.* . . . 112
4.8 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Tcheandjieu *et al.* . . . 113
4.9 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Tamlander *et al.* . . . 114
4.10 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Patel *et al.* . . . 115
4.11 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Inouye *et al.* . . . 116
4.12 Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Khera *et al.* . . . 117

5.1 Cross validation data split case control proportion . . . 183
5.2 Characteristics of external GWAS used in GWAS summary statistics . . . 183
5.3 Summary of GWAS meta-analysis performed . . . 184
5.4 Summary of hyperparameters selected for tuning for each PRS method . . . 185
5.5 LDpred2 convergence model counts . . . 186
5.6 Mean AUC comparison of the best performing PRS of each method . . . 187
5.7 Mean OR per SD comparison of the best performing PRS of each method . . . 188

6.1 Reclassification of cases and controls based on 7.5% risk threshold SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated × Patel *et al.* PRS . . . 225
6.2 NRI and IDI comparing SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated × Patel *et al.* PRS . . . 225
6.3 Reclassification of cases and controls based on 7.5% risk threshold SCORE2-standard with SCORE2-standard × Patel *et al.* PRS . . . 226
6.4 NRI and IDI comparing SCORE2-standard with SCORE2-standard × Patel *et al.* PRS . . . 226
6.5 Reclassification of cases and controls based on 7.5% risk threshold pooled cohort equation with pooled cohort equation × Patel *et al.* PRS . . . 227
6.6 NRI and IDI comparing pooled cohort equation with pooled cohort equation × Patel *et al.* PRS . . . 227
6.7 Baseline characteristics of 26,641 MCPS participants included in the sensitivity analysis of MCPS-PRS . . . 228
6.8 NRI and IDI comparing SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated x MCPS-PRS among the 20% of MCPS participants used for PRS testing . . . 229
6.9 NRI and IDI comparing SCORE2-standard with SCORE2-standard x MCPS-PRS among the 20% of MCPS participants used for PRS testing . . . 229
6.10 NRI and IDI pooled cohort equation with Pooled cohort equation x MCPS-PRS among the 20% of MCPS participants used for PRS testing . . . 230
6.11 Net reclassification index comparing SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated x Patel *et al.* PRS in the first 10 years of follow up . . . 230
6.12 Net reclassification index comparing SCORE2-standard with PCE x Patel *et al.* PRS in the first 10 years of follow up . . . 231
6.13 Net reclassification index comparing SCORE2-diabetes-recalibrated with PCE x Patel *et al.* PRS in the first 10 years of follow up . . . 231
6.14 Categorical NRI comparing each clinical risk score with their corresponding integrated risk score across different risk thresholds . . . 232
6.15 Sensitivity and specificity of risk scores across different risk thresholds . . . 233

S1 Identified papers that constructed a new PRS for CHD risk. . . . 279
S2 Identified papers that applied existing PRSs for CHD risk estimate in their studies. 295
S3 Identified GWAS studies that sourced the development or application of the selected external PRSs for CHD in the identified studies. . . . 302
S4 Equation parameter of SCORE2 (before age 70) for estimation of uncalibrated 10-year risk of CVD . . . 306
S5 Region specific scaling factor of SCORE2 (before age 70) . . . 306
S6 Equation parameter of SCORE2-OP (above age 70) for estimation of uncalibrated 10-year risk of CVD . . . 307
S7 Region specific scaling factor of SCORE2-OP (above age 70) . . . 307
S8 Equation Parameters of the Pooled Cohort Equations for Estimation of 10-Year Risk 308

List of Figures

1.1 Coronary heart disease mortality rate at age 35-79 years . . . 29
1.2 Cumulative number of individuals included in GWAS by year up to 30th Nov 2023 . 30

2.1 Flowchart of literature search and identification . . . 61

3.1 A: Two districts of recruitment. B: Recruitment to MCPS per month, coloured by district . . . 83

4.1 Pairwise correlations between the eight selected PRSs . . . 118
4.2 Odds of CHD by fifth of each PRS . . . 119
4.3 Odds of CHD per 1SD increase in each PRS . . . 120
4.4 Odds of CHD per 1SD increase in each PRS, by sex . . . 121
4.5 Odds of CHD per 1SD increase in each PRS, by baseline age . . . 122
4.6 Odds of CHD per 1SD increase in each PRS, by highest level of education . . . 123
4.7 Odds of CHD per 1SD increase in each PRS, by waist-to-hip ratio . . . 124
4.8 Odds of CHD per 1SD increase in each PRS, by systolic blood pressure . . . 125
4.9 Odds of CHD per 1SD increase in each PRS, by diastolic blood pressure . . . 126
4.10 Odds of CHD per 1SD increase in each PRS, by smoking status . . . 127
4.11 Odds of CHD per 1SD increase in each PRS, by baseline diabetes status . . . 128
4.12 Odds of CHD per 1SD increase in each PRS, by level of Indigenous American ancestry . . . 129
4.13 Odds of CHD before age 90 years per 1SD increase in each PRS, among participants age 35-89 at recruitment . . . 130
4.14 Odds of CHD before age 90 years by fifth of each PRS, among participants age 35-89 years at recruitment . . . 131
4.15 Odds of baseline self-reported angina per 1SD increase in each PRS, for participants aged 35-79 years at recruitment . . . 132
4.16 Odds of baseline self-reported myocardial infarction per 1SD increase in each PRS, for participants aged 35-79 years at recruitment . . . 133
4.17 Odds of baseline self-reported angina or myocardial infarction per 1SD increase in each PRS, for participants aged 35-79 years at recruitment . . . 134
4.18 Odds of CHD (now defined as baseline self-reported angina or myocardial infarction, or death before age 80 with CHD listed as the primary cause of death) per 1SD increase in each PRS, for participants aged 35-79 years at recruitment . . . 135
4.19 Odds of death before age 80 with CHD listed anywhere on the death certificate, per 1SD increase in each PRS . . . 136
4.20 Odds of death before age 80 with CHD listed as the primary cause on the death certificate, per 1SD increase in each PRS . . . 137
4.21 Sensitivity analyses (varying the CHD outcome) with partial adjustments . . . 138
4.22 Sensitivity analyses (varying the CHD outcome) with full adjustments . . . 139

4.23 Odds of CHD (varying CHD outcomes) per 1SD increase for the Patel *et al.* PRS, by sex . . . 140
4.24 Hazard of death before age 80 with CHD listed anywhere on the death certificate, per 1SD increase in each PRS . . . 141
4.25 Hazard of death before age 80 with CHD listed as the primary cause on the death certificate, per 1SD increase in each PRS . . . 142
4.26 Sensitivity analyses (primary CHD definition) among participants unrelated to the 3rd degree . . . 143

5.1 PRS construction flowchart . . . 189
5.2 Mean AUC comparison of P+T algorithms parameters . . . 190
5.3 Mean OR per SD comparison of P+T algorithms parameters . . . 191
5.4 Mean AUC comparison of LDpred2 algorithms parameters . . . 192
5.5 Mean OR per SD comparison of LDpred2 algorithms parameters . . . 193
5.6 Mean AUC comparison of PRS-CSx algorithms parameters . . . 194
5.7 Mean OR per SD comparison of PRS-CSx algorithms parameters . . . 195
5.8 Odds of premature CHD by fifth of each MCPS-informed PRS, in the testing set . . . 196
5.9 Odds of premature CHD per 1SD increase in each PRS, in the testing set . . . 197

6.1 Distribution of clinical risk scores and their components in MCPS . . . 234
6.2 Distribution of SCORE2-diabetes-recalibrated and SCORE2-diabetes-recalibrated x Patel *et al.* PRS (left), and agreement between the two scores (right). . . . 235
6.3 Distribution of SCORE2-standard and SCORE2-standard x Patel *et al.* PRS (left), and agreement between the two scores (right). . . . 236
6.4 Distribution of PCE and PCE x Patel *et al.* PRS (left), and agreement between the two scores (right). . . . 237
6.5 Shape of association between each clinical risk score and CHD . . . 238
6.6 Average strength of association between each standardised clinical risk score and CHD . . . 239
6.7 Shape of association between each clinical risk score and CHD death in the first 10 years of follow-up . . . 240
6.8 Average strength of association between each standardised clinical risk score and CHD death in the first 10 years of follow-up . . . 241
6.9 Shape of association between SCORE2-diabetes-recalibrated integrated score and CHD . . . 242
6.10 Shape of association between SCORE2-standard integrated score and CHD . . . 243
6.11 Shape of association between PCE and CHD . . . 244
6.12 Average strength of association between each standardised integrated score and CHD . . . 245
6.13 Shape of association between each clinical risk score and CHD, by sex . . . 246
6.14 Shape of association between each clinical risk score and CHD, by baseline diabetes status . . . 247
6.15 Odds of premature CHD per 1 SD increase in PRS, by SCORE2-diabetes-recalibrated risk strata . . . 248
6.16 Odds of premature CHD per 1 SD increase in PRS, by SCORE2-standard risk strata 249

6.17 Odds of premature CHD per 1 SD increase in PRS, by PCE risk strata 250
6.18 Average strength of association between each standardised integrated score and CHD, among 20% subset of the MCPS participants that retained for PRS testing . 251

Chapter 1: Introduction

Summary

Coronary heart disease (CHD) is a major cause of death and disability worldwide. It was estimated that in 2021, CHD was responsible for approximately 9 million deaths, representing 13% of all global mortality[1]. While many developed nations have witnessed declines in age-standardised CHD mortality rates over recent decades[1], the disease continues to hold its position as the leading cause of death among adults globally. This persistent burden underscores the critical importance of prioritising CHD prevention as a central focus of global public health initiatives.

CHD is a complex disease with high heritability and can be passed down through family genetics. Several studies have shown that accounting for genetic predisposition can enhance CHD risk stratification and identification beyond conventional vascular risk factors[2–4]. In the past decades, technological advancements have led to a surge in research on the genetic epidemiology of CHD and its translational applications for prevention. With the availability of large-scale genome-wide association studies (GWAS), over 100 CHD genetic risk loci have been discovered, such as the well-known 9p21 locus[5–8]. Complex diseases such as CHD have a polygenic predisposition, due to the joint effect of many common genes. Polygenic CHD risk scores (PRSs) that combine multiple single nucleotide polymorphisms (SNP) effects (e.g., from GWAS summary statistics) into a single genetic score have gained particular interest in epidemiology, and have been shown to identify individuals with CHD risk equivalent to (or even higher than) those with rare monogenic mutations of large effects (e.g., familial hypercholesterolemia)[9]. An increasing number of advanced methods have been developed to improve the predictive value of PRS in the context of chronic disease epidemiology in recent years[10–12].

Like most traits and diseases, GWAS studies of CHD, which provide the relevant genetic informa-

tion for the subsequent construction of CHD-specific PRS, have predominantly focused on populations of European ancestry. As a result, most of the existing CHD PRSs were developed based on the genetic profiles of European ancestry, potentially limiting their effectiveness and applicability to individuals with a different genetic ancestry makeup. For instance, admixed-American populations from central or southern America (hereon referred to as "Latin America") are one of the least studied populations in the field of genetics, comprising less than 0.5% of all the participants in GWAS studies to date[13]. Due to the history of European colonisation, populations in Latin America are highly admixed with complex population structure and are under-represented in genetic epidemiological studies. However, the emergence of large cohort studies with genotype information established in Latin America in the last few decades now enables large studies of genetic epidemiology in such populations. In this thesis, the Mexico City Prospective Study (MCPS) was leveraged for the admixed-American population of interest. MCPS is a blood-based prospective cohort with over 150,000 participants recruited between 1998 and 2004[14]. The study participants have been genotyped and systematically followed up through linkage to the National Death Registry to identify specific causes of death (see **Chapter 3** for detailed information on the MCPS cohort). These features make MCPS a valuable resource for advancing genetic research on CHD risk among admixed-American populations.

The objectives of this thesis are to review existing studies on CHD PRSs, to evaluate the transferability of previously-published CHD PRSs developed in ethnically-diverse populations among Mexicans, to construct a new CHD PRS that more closely represents the genetic variation in admixed-Americans, and to assess whether integrated (genetic x clinical) risk scores improve predictive ability for CHD over and above clinical risk scores. For these purposes, the MCPS cohort data will be used for the main analyses in this thesis. GWAS results from external admixed-American, European, and East-Asian cohorts will be used in **Chapter 5** to improve PRS performance.

1.1 Coronary heart disease epidemiology and global prevalence

Coronary heart disease (CHD) is the most common subtype of cardiovascular diseases (CVD) in adults. CHD occurs when the coronary heart arteries become narrow or occluded[15], typically due to atherosclerosis. This narrowing of the coronary blood vessels restricts the blood flow around the heart, and could lead to ischaemia of the heart tissues and potential complications such as angina and myocardial infarction (MI).

CHD is the global leading cause of morbidity and mortality in adults, affecting over 300 million people in 2022[16]. Between 1950 and 2020, the age-standardised CHD mortality rates have been declining in many countries[17] (see **Figure 1.1**). For instance, the CHD mortality rate at the ages of 35-79 years in the United States of America (USA) decreased from 1,000 per 100,000 person-year in 1950 to 210 per 100,000 person-year in 2020 for men and from 610 per 100,000 per 100,000 person-year in 1950 to 90 per 100,000 person-year in 2020 for women. Although a steep decline in the CHD death rates has been observed in many countries, in 2021, CHD was ranked as the top cause of mortality worldwide and the second leading cause of disability-adjusted life years (DALYs), and premature death and years of life lost (YLL)[1]. However, there is significant geographical variation both in the prevalence and burden of CHD[16]. The region with the highest CHD mortality rate is Eastern Europe (Ukraine, Lithuania, and Belarus), with annual death rates ranging from 490 to 690 per 100,000 person-year[1]. In contrast, South America and most countries in southern Africa report the lowest mortality rates, with fewer than 100 deaths per 100,000 person-year[1].

In contrast to the trends seen in other countries, in Mexico, the CHD mortality rate at the ages of 35-79 years has increased over the past 70 years[1] (in men from about 60 per 100,000 in 1955 to 420 per 100,000 in 2020, and in women from about 40 per 100,000 in 1955 to 220 per 100,000 in 2020)[17] (**Figure 1.1**). The CHD burden in Mexico is shaped by unique epidemiological and genetic factors, which will be discussed in **Section 1.5**.

1.2 Genetic studies of CHD

1.2.1 CHD genetic discovery

CHD is a highly heritable disease, and previous research has estimated that the heritability component of CHD risk ranges from 30 to 60%[18–20], with twin studies yielding the highest heritability (up to 60%) compared to large-scale GWAS studies of the general population (30-50%). These studies underscore the importance of genetic variation in explaining inter-individual differences in CHD risk.

In the past five decades, studies of the genetics of CHD have advanced considerably[21]. Early research focused on individual protein-coding genes with known biological functions for CHD risk. Later, due to assay costs, research shifted towards genotyping common genetic variants with identified or suspected functions, and candidate gene association studies in a population (often CHD case-control designs). However, such studies often faced challenges during the replication stage of their findings, mostly due to weak statistical power caused by inadequate sample size, and population stratification (differences in genetic background between subgroups of a population that are commonly caused by ancestry)[21].

To overcome the challenges related to genetic research of complex diseases, where multiple genes with small effects may be causal to a particular disease, as opposed to a single or rare causal gene, GWAS was introduced to systematically conduct association analyses across all genetic variants available[22]. When conducting a GWAS, each genetic variant (often a single nucleotide polymorphism [SNP]) is individually assessed for its association with the trait of interest, using univariate linear regression (for continuous traits) or logistic regression (for binary traits). Before 2010, GWAS of around 600 to 3,000 individuals in the discovery set identified several important CHD risk loci, such as 9p21, 3p22 and 10p22[5,23,24]. These SNPs were further validated in replication sets of over 10,000 individuals.

In the past decade, large-scale CHD GWAS have become more common, thanks to the cost-

effectiveness and scale of the technological advancements in DNA microarray for genome-wide genotyping in large samples[8]. In addition, the establishment of the 1000 Genomes Project[25] and of the Haplotype Reference Consortium (HRC)[26], provided whole genome sequencing data as reference panels for genotype imputation, and has enabled more genetic variants to be included in CHD GWAS. Moreover, the rise in the setup of large population cohort studies and consortia (e.g., CARDIoGRAMplusC4D) that bring together many medium-sized studies has made it possible to identify and replicate more CHD genetic signals at the population level[6,7,27]. For instance, in 2011, the GWAS conducted by the CARDIoGRAM consortium, involving over 80,000 individuals, identified 13 novel CHD risk loci in one single study[27]. According to a review conducted in 2018, 163 CHD specific risk loci have been identified and replicated at genome-wide significance level (i.e. p-value $\leq 5 \times 10^{-8}$)[8]. Based on more recent large-scale studies[28,29], it is likely that the count of genome-wide significant CHD risk loci now exceeds 200.

However, most CHD GWAS studies to date are predominantly among individuals of European ancestry (see **Figure 1.2**). Based on the statistics published by the GWAS catalog[13], as of July 2024, 91% of the individuals included in a GWAS are of European ancestry and 59% of the GWAS studies published primarily focused on European populations. The under-representation of non-European populations could partially be attributed to the lack of large non-European genotyped cohorts. Nonetheless, due to important differences in minor allele frequency (MAF) or linkage disequilibrium (LD) structure among diverse populations and ancestries, ancestry-specific genetic studies and identification of disease-causing SNPs among non-European populations offer valuable opportunities for genetic discovery. For instance, SNP rs1333049 (G>C), which lies in the 9p21 locus, a region repeatedly shown to be significantly associated with CHD risk in multiple studies[18,28,30,31], has a MAF of 48% in Europeans but only 23% in Africans[32], which may greatly influence its disease association across different ancestries. In recent years, there have been efforts made to address this gap, through the conduct of multi-ancestry CHD GWAS (or meta-analyses) which included individuals from diverse ancestry groups to increase GWAS sam-

ple size and incorporate ancestry information of non-Europeans[30,31]. However, multi-ancestry GWAS reaching sufficiently large sample-sizes remain low compared to that of European ancestry GWAS. Consequently, there is a need for enhanced large-scale CHD GWAS of non-European populations to identify ancestry-specific CHD risk variants that may be missing or under-represented in European cohorts.

1.3 Polygenic risk scores for the assessment of CHD risk

1.3.1 From GWAS discovery to polygenic risk scores for CHD risk assessment

The polygenic architecture that characterises inherited risk of CHD implies that disease susceptibility is influenced by the collective effect of multiple genes. Although each gene identified through a GWAS scan typically contributes only a small portion to the overall disease-specific risk, these modest 'individual' effects can accumulate across many genes, and substantially modify the likelihood of developing the disease in individuals with high genetic propensity for that particular condition[33]. This gives rise to the development and use of PRS, a composite score that combines the weighted genetic information of multiple SNPs together, for the assessment of genetic predisposition to a specific disease. Many previous studies have shown that PRSs have the potential to improve risk stratification and prediction for complex diseases, including CHD[34,35]. For instance, a study conducted in a Chinese population found that including PRS into a model with a CHD clinical risk score could improve model discrimination by 1% and net reclassification index (NRI) by 3.5%[35]. Moreover, a study that developed a genome-wide CHD PRS found that approximately 8% of individuals had more than a threefold increased risk of CHD compared to the rest of the population. These high-risk individuals could not be identified using traditional clinical CHD risk factors alone[9]. The construction of a PRS requires information on genetic variants (i.e., SNPs) that are associated with the outcome of interest. The information includes the estimated effect sizes (or weights) derived from GWAS summary statistics, as well as the risk allele counts for each included SNP in relation to the outcome.

1.3.2 PRS development

Early research using PRSs included only a few SNPs (typically 10-50) that were identified as genome-wide significant (p-value ≤5x10^{-8}) in previously published GWAS studies[36–39]. Although these scores demonstrated the utility of PRS, their disease-risk predicting ability was limited due to low number of included SNPs. Subsequently, there were attempts to include more SNPs in the PRS. However, due to the correlation between SNPs, those in high LD may tag the same disease-causing gene. Including SNPs in high LD in a PRS could potentially lead to an overestimation of genetic risk due to multiple counting of the same underlying genetic risk signal. To account for LD between SNPs, several methods have been developed such as *Pruning and Thresholding* (P+T)[40] which retains SNPs based on the correlation between SNPs and association p-values from GWAS; and *lassosum*[11], which uses a penalised machine learning technique[41] to remove SNPs not favoured by the algorithms due to high collinearity. In recent years, even more advanced methods have been developed that further accommodate diverse genetic patterns and large samples using Bayesian methods such as *LDPred*[10,42], and *PRS-CS(x)*[12,43].

However, due to the under-representation of populations of non-European ancestry in GWAS studies (as described in **Section 1.2.1**), which is a crucial input during PRS construction, most of these previous efforts of CHD PRSs have been limited to populations of European ancestry. PRSs generated from European populations may translate poorly to other ancestries as any ancestry-specific SNPs would be excluded while SNPs may also impact disease differently in different ancestries. Previous findings from studies that have applied European CHD PRSs to non-European cohorts have often been limited by small sample sizes[44,45]. Thus, to enhance the diversity in PRS research, non-European studies with genetic data and large sample sizes are needed[46]. Such efforts will also contribute to the generalisability of CHD PRSs across ancestries[47].

1.4 Conventional risk factors for CHD

Conventional risk factors that are associated with an increased risk for CHD comprise biological, lifestyle, and environmental risk factors. Biological risk factors include systolic and diastolic blood pressure (SBP and DBP), low-density lipoprotein (LDL) cholesterol levels, body mass index (BMI), and diabetes mellitus status[48–50]. Lifestyle risk factors for CHD include smoking, lack of physical activity and dietary factors[48,51–53]. In particular, smoking, high systolic blood pressure and high LDL-cholesterol levels are considered to be the most significant modifiable risk factors for CHD, responsible for over 60% of disability-adjusted life years (DALYs)[1,54,55]. These risk factors play an essential role in guiding the screening and management of CHD. Broad CVD-specific clinical risk scores originally developed to predict long-term risk from any CVD (including, but not restricted to CHD), in clinical settings (e.g., the Framingham Risk Score (FRS), the Pooled Cohort Equation (PCE)[56], QRISK[57,58], ASSIGN[59] and SCORE2[60]) could be applied to CHD, thereby helping to guide management and treatment. These clinical risk scores, together with age and sex, demonstrated over 70% discrimination rate (Harrell's C index) for CHD or CVD cases[3,4,34,61].

Despite their widespread use among countries of different ancestries, it is important to acknowledge that the mentioned clinical scores were mostly developed using populations of European ancestry. However, a recent study has shown that SCORE2, a 10-year CVD risk calculator developed in 2021 using a large cohort of predominantly European populations of 13 countries, predicted CVD well in Asian countries after systematic recalibration to account for differences in risk factor distributions among Asian countries[62].

1.4.1 CHD risk scores integrating genetic and clinical information

As described in **Section 1.4**, traditional clinical risk scores that combine the information of individual conventional CHD risk factors, have proven strong predictive performance for long-term risk of CHD in European populations and translated well to other populations such as Asians

after recalibration. Nonetheless, they explain only a fraction of the total risk for CHD. Hence, it is potentially beneficial to combine polygenic and clinical risk factors in order to boost the overall prediction probability of CHD risk. Several studies have included a clinical risk score (or clinical risk factors) and a PRS in their models as two independent covariates, and have reported contrasting findings for model discrimination and NRI[35,63–65]. For instance, a Chinese study of approximately 5,000 adults found that CHD PRS complemented PCE, improving the model discrimination (Harrell's C index) by 1% when the two were included model additively in the model[35]. However, another study of around 7,000 adults of European ancestry suggested that incorporating a CHD PRS and PCE in to a model may not be sufficient to significantly enhance the predictive power of the models[63]. In contrast, a few studies based on populations of European ancestry found that an integrated score (PRS x PCE) improved the overall risk stratification[3,66–68]. For instance, one study reported modest improvement in model discrimination of 0.03 (95% CI 0.02-0.04) and significant overall net reclassification improvement (NRI) of 5.88% (95% CI 4.73%-7.04%), when a PRS was integrated with PCE in survival analysis, using a CHD risk threshold of 7.5%[3]. In addition, studies assessed the improvement in prediction by an integrated score in populations of non-European ancestries are limited[69]. Due to the uncertainty in the predictive power of combining a PRS and a clinical risk score in under-represented non-European populations in the field, it is therefore important to assess the utility of an integrated score of genetic and clinical information for CHD risk prediction in populations of other ancestries.

1.5 CHD burden in Mexico

1.5.1 Admixed-American populations

There has not been a unified term to describe the populations of Latin America. For instance, "Hispanics", "Latino", and "admixed population" have all been used previously in different literature[30,70,71]. To reflect the admixed nature of their ancestry, populations of Latin America are referred to collectively as "admixed-Americans" in this thesis. Around 8% of the world population

lives in Latin America[72] and around 19% of the US total population are admixed-Americans[73]. However, such admixed-Americans are among the least studied populations in the field of genetic epidemiology. The complex population structure of this ancestry group may yield novel genetic markers associated with disease risk that have remained undiscovered in other populations. Populations in Latin America have a complex admixture pattern, with genetic information inherited from Europeans (mainly Spanish and Italians), indigenous Americans (descendants of tribes like Aztecs and Maya) and a small proportion from Africans (descendants of enslaved Africans brought to the continent by colonisation)[74]. Moreover, the degree of genetic admixture varies greatly within admixed-American populations in Latin America. The complex population structure of Latin American populations and the limited genetic research to date emphasise the importance of conducting admixed-American-specific genetic studies to refine our understanding of how differences in genetic architecture may influence differences in CHD risk.

1.5.2 CHD burden and epidemiology in Mexican populations

Mexico is located at the south part of North America and is the country with the most Spanish-speaking population globally, with over 120 million residents of admixed-American ancestry[74,75]. The increase in CHD mortality rate in Mexico mentioned in **Section 1.1** is likely associated with the substantial increases in prevalence of key major CHD risk factors over this period (in particular, increases in the prevalences of obesity and diabetes). Furthermore, a previous report from the MCPS found that the excess risk of premature CHD death in those with versus without diabetes was about twice as large as that seen in high-income populations of European ancestry (with diabetes conferring a relative risk for CHD of about 4 rather than 2)[76]. In addition, obesity and low-intensity smoking are much more common in Mexico than in many other countries[77,78], while use of effective cardio-protective treatments remains sub-optimal[55]. The impact of varying prevalences (and effects) of conventional CHD risk factors on the performance of a PRS in a population is uncertain, especially for scores derived in one population but applied to another.

Mexico City is the capital of Mexico and the most populous city of the country. The city is located in the Valley of Mexico with an altitude of 2,240 metres. In addition to the Europeans and Africans, in the early 1900s people from other regions of Mexico started to migrate to the city. Consequently, the residents of Mexico City are partly representative of the genetic diversity in Mexico, reflecting the complex admixture pattern of European, African and indigenous American ancestries of the country.

1.5.3 The Mexico City Prospective Study

The Mexico City Prospective Study (MCPS) is a blood-based, fully genotyped prospective cohort with over 150,000 participants aged 35 or older, who were recruited between 1998 and 2004[14], from two districts in Mexico City (Coyoacán and Iztapalapa). Baseline questionnaires were used to collect information on sociodemographic, lifestyle, medication history and anthropometric measures, and blood samples were taken. Genotyping was performed and the genetic data were subsequently imputed[79]. The detailed methods of the study recruitment and assessment will be described in **Chapter 3**. Due to the ancestry pattern in Mexico City, MCPS is an admixed-American cohort with a high degree of Indigenous American ancestry. Therefore, the cohort is well-suited for conducting genetic epidemiology studies on CHD risk to provide valuable insight for both Indigenous-American and admixed-American populations.

1.6 DPhil objectives

1.6.1 Research aims

Using the data from the MCPS[14], the main objective of this thesis is to predict the risk of CHD among Mexicans using PRSs. This objective further divides into four sub-aims. The first aim is to conduct a systematic literature review of studies on CHD PRSs, with the goal of understanding existing methodologies for PRS construction and evaluation. The second aim is to evaluate the transferability of previously-published CHD PRSs among Mexicans. The third aim is to construct a CHD PRS using MCPS genetic data and external GWAS representing other ancestries. This

aim consists two parts, first to run GWAS analyses on CHD in the MCPS population and second to use the MCPS and external GWAS summary statistics to train an MCPS-informed CHD PRS. The fourth and final aim of the thesis is to assess the predictive ability of integrated (PRS x clinical) risk scores for CHD over and above clinical risk scores that incorporate conventional CHD risk factors.

The findings from this thesis could greatly enhance the current understanding of genetically-conferred risk for CHD among admixed-American populations, explore the generalisability of CHD PRS across different ancestries, address the current research gap in ancestry diversity within the field of CHD PRS and, finally, potentially inform and enhance screening for CHD and clinical disease management in this admixed-American population. The analytical methodology from this thesis will also inform future genetic studies of other diseases among this or other populations of admixed-American ancestry.

1.6.2 Thesis structure

This thesis is divided into seven chapters with their pertaining bibliographies and supplemental material appended at the end of the thesis. The first chapter introduces the epidemiological background of CHD, including its population burden and related genetic research, and outlines the thesis objectives. In the second chapter, the systematic literature review procedure and findings are detailed, to give context to the work conducted in this thesis. The third chapter details the methods employed, and describes the data used and general statistical approaches adopted. The fourth to the sixth chapters present the findings from the various statistical and genetic analyses addressing the research aims described in **Section 1.6.1**, and the final chapter discusses the findings in context with results from other studies, including their clinical and public health relevance and provides any emerging conclusions.

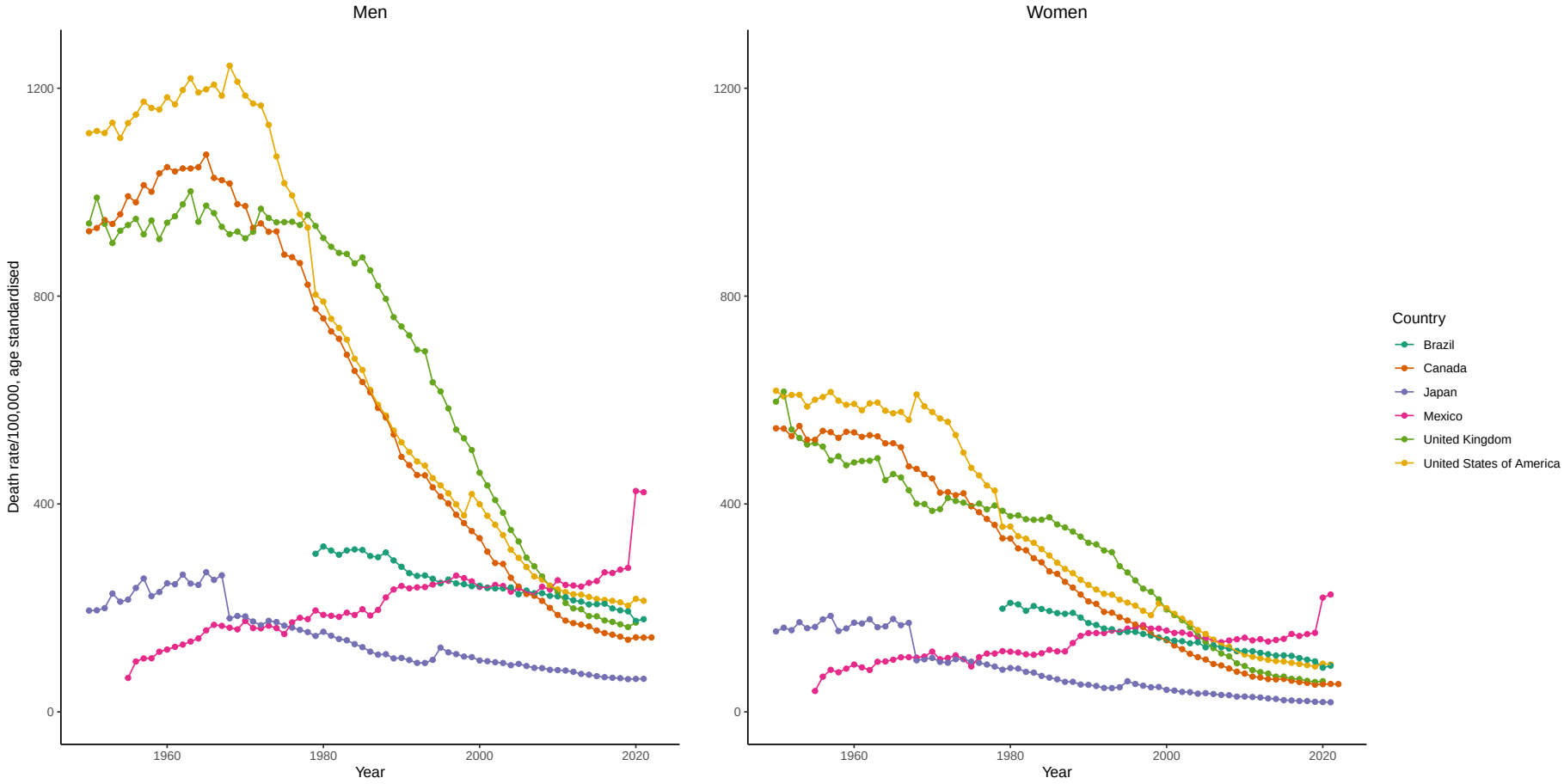


Figure 1.1: Coronary heart disease mortality rate at age 35-79 years

Data retrieved from Richard Doll Consortium[17] (www.richarddollconsortium.org (Hosted by Oxford Population Health) using World Health Organisation mortality and United Nations Population Division data). Death rates standardised by taking the unweighted average of the annual age-specific rates (in 5-year age bands).

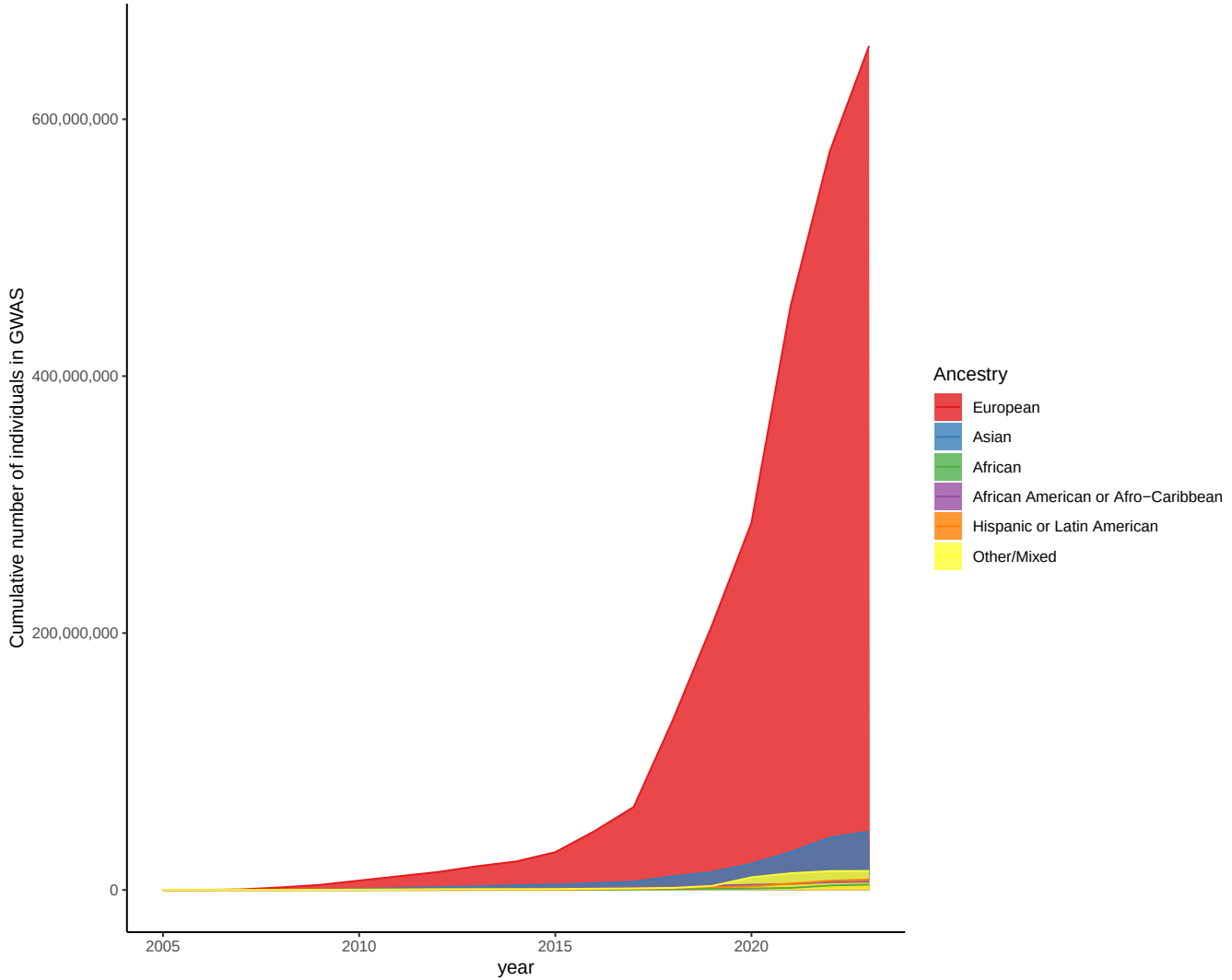


Figure 1.2: Cumulative number of individuals included in GWAS by year up to 30th Nov 2023

Data retrieved from GWAS diversity monitor (https://gwasdiversitymonitor.com/)[80].

1.7 References

1. Ferrari AJ, Santomauro DF, Aali A, Abate YH, Abbafati C, Abbastabar H, et al. “Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021”. *The Lancet* 2024;403(10440):pp. 2133–2161.
2. Cooper RS, Kaufman JS, and Ward R. “Race and Genomics”. *New England Journal of Medicine* 2003;348(12):pp. 1166–1170.
3. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, et al. “Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction”. *Circ Genom Precis Med* 2021;14(2):e003304.
4. King A, Wu L, Deng HW, Shen H, and Wu C. “Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease”. *BMC Medicine* 2022;20(1):p. 385.
5. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. “A common allele on chromosome 9 associated with coronary heart disease”. *Science* 2007;316(5830):pp. 1488–91.
6. Nikpay M, Goel A, Won H.-H, Hall LM, Willenborg C, Kanoni S, et al. “A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease”. *Nature genetics* 2015;47(10):pp. 1121–1130.
7. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. “Large-scale association analysis identifies new risk loci for coronary artery disease”. *Nature Genetics* 2013;45(1):pp. 25–33.
8. Erdmann J, Kessler T, Munoz Venegas L, and Schunkert H. “A decade of genome-wide association studies for coronary artery disease: the challenges ahead”. *Cardiovascular Research* 2018;114(9):pp. 1241–1257.
9. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. *Nat Genet* 2018;50(9):pp. 1219–1224.
10. Privé F, Arbel J, and Vilhjálmsson BJ. “LDpred2: better, faster, stronger”. *Bioinformatics* 2021;36(22-23):pp. 5424–5431.
11. Mak TSH, Porsch RM, Choi SW, Zhou X, and Sham PC. “Polygenic scores via penalized regression on summary statistics”. *Genetic Epidemiology* 2017;41(6):pp. 469–480.
12. Ge T, Chen CY, Ni Y, Feng YA, and Smoller JW. “Polygenic prediction via Bayesian regression and continuous shrinkage priors”. *Nature communications* 2019;10(1):p. 1776.
13. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. “The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource”. *Nucleic Acids Research* 2022;51(D1):pp. D977–D985.
14. Tapia-Conyer R, Kuri-Morales P, Alegre-Díaz J, Whitlock G, Emberson J, Clark S, et al. “Cohort profile: the Mexico City Prospective Study”. *Int J Epidemiol* 2006;35(2):pp. 243–9.
15. Grech ED. “Pathophysiology and investigation of coronary artery disease”. *BMJ* 2003;326(7397):pp. 1027–1030.
16. Stark B, Johnson C, and Roth GA. “Global prevalence of coronary artery disease: an update from the global burden of disease study”. *Journal of the American College of Cardiology* 2024;83(13_Supplement):pp. 2320–2320.
17. Richard Doll Consortium. *www.richarddollconsortium.org (Hosted by Oxford Population Health) using World Health Organization mortality and United Nations Population Division data.* Online Database. June 2024. `https://www.richarddollconsortium.org/projects/mortality-trends`.
18. Aragam KG, Jiang T, Goel A, Kanoni S, Wolford BN, Atri DS, et al. “Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants”. *Nat Genet* 2022;54(12):pp. 1803–1815.
19. Khera AV and Kathiresan S. “Genetics of coronary artery disease: discovery, biology and clinical translation”. *Nature Reviews Genetics* 2017;18(6):pp. 331–344.

20. Zdravkovic S, Wienke A, Pedersen NL, Marenberg ME, Yashin AI, and De Faire U. “Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins”. *J Intern Med* 2002;252(3):pp. 247–54.

21. McPherson R and Tybjaerg-Hansen A. “Genetics of Coronary Artery Disease”. *Circulation Research* 2016;118(4):pp. 564–78.

22. Risch N and Merikangas K. “The future of genetic studies of complex human diseases”. *Science* 1996;273(5281):pp. 1516–7.

23. Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, Hall AS, et al. “New susceptibility locus for coronary artery disease on chromosome 3q22.3”. *Nat Genet* 2009;41(3):pp. 280–2.

24. Erdmann J, Willenborg C, Nahrstaedt J, Preuss M, König IR, Baumert J, et al. “Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23”. *Eur Heart J* 2011;32(2):pp. 158–68.

25. Genomes Project Consortium. “A global reference for human genetic variation”. *Nature* 2015;526(7571):p. 68.

26. Iglesias AI, Van Der Lee SJ, Bonnemaijer PW, Höhn R, Nag A, Gharahkhani P, et al. “Haplotype reference consortium panel: Practical implications of imputations with large reference panels”. *Human mutation* 2017;38(8):pp. 1025–1032.

27. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. “Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease”. *Nat Genet* 2011;43(4):pp. 333–8.

28. van der Harst P and Verweij N. “Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease”. *Circ Res* 2018;122(3):pp. 433–443.

29. Walters RG, Millwood IY, Lin K, Schmidt Valle D, McDonnell P, Hacker A, et al. “Genotyping and population characteristics of the China Kadoorie Biobank”. *Cell Genom* 2023;3(8):p. 100361.

30. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. “Large-scale genome-wide association study of coronary artery disease in genetically diverse populations”. *Nature Medicine* 2022;28(8):pp. 1679–1692.

31. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, et al. “Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease”. *Nature Genetics* 2020;52(11):pp. 1169–1177.

32. Sherry ST, Ward M.-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. “dbSNP: the NCBI database of genetic variation”. *Nucleic Acids Research* 2001;29(1):pp. 308–311.

33. Abraham G, Rutten-Jacobs L, and Inouye M. “Risk Prediction Using Polygenic Risk Scores for Prevention of Stroke and Other Cardiovascular Diseases”. *Stroke* 2021;52(9):pp. 2983–2991.

34. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. “Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 636–645.

35. Lu X, Liu Z, Cui Q, Liu F, Li J, Niu X, et al. “A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study”. *European Heart Journal* 2022;43(18):pp. 1702–1711.

36. Tada H, Melander O, Louie JZ, Catanese JJ, Rowland CM, Devlin JJ, et al. “Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history”. *Eur Heart J* 2016;37(6):pp. 561–7.

37. Mega JL, Stitziel NO, Smith JG, Chasman DI, Caulfield MJ, Devlin JJ, 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* 2015;385(9984):pp. 2264–2271.

38. Tikkanen E, Havulinna AS, Palotie A, Salomaa V, and Ripatti S. “Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease”. *Arteriosclerosis, Thrombosis and Vascular Biology* 2013;33(9):pp. 2261–2266.

39. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. “A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses”. *Lancet* 2010;376(9750):pp. 1393–400.

40. Privé F, Vilhjálmsson BJ, Aschard H, and Blum MGB. “Making the Most of Clumping and Thresholding for Polygenic Scores”. *Am J Hum Genet* 2019;105(6):pp. 1213–1221.

41. Tibshirani R. “Regression shrinkage and selection via the lasso: a retrospective”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011;73(3):pp. 273–282.

42. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores”. *The American Journal of Human Genetics* 2015;97(4):pp. 576–592.

43. Ruan Y, Lin Y.-F, Feng Y.-CA, Chen C.-Y, Lam M, Guo Z, et al. “Improving polygenic prediction in ancestrally diverse populations”. *Nature Genetics* 2022;54(5):pp. 573–580.

44. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, et al. “Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups”. *American Journal of Human Genetics* 2020;106(5):pp. 707–716.

45. Saad M, El-Menyar A, Kunji K, Ullah E, Al Suwaidi J, and Kullo IJ. “Validation of Polygenic Risk Scores for Coronary Heart Disease in a Middle Eastern Cohort Using Whole Genome Sequencing”. *Circulation* 2022;Genomic and precision medicine.e003712.

46. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. “Analysis of polygenic risk score usage and performance in diverse human populations”. *Nature Communications* 2019;10(1):p. 3328.

47. Graham SE, Clarke SL, Wu K.-HH, Kanoni S, Zajac GJM, Ramdas S, et al. “The power of genetic diversity in genome-wide association studies of lipids”. *Nature* 2021;600(7890):pp. 675–679.

48. Hajar R. “Risk Factors for Coronary Artery Disease: Historical Perspectives”. *Heart Views* 2017;18(3):pp. 109–114.

49. Fuchs FD and Whelton PK. “High Blood Pressure and Cardiovascular Disease”. *Hypertension* 2020;75(2):pp. 285–292.

50. Grant PJ, Cosentino F, and Marx N. “Diabetes and coronary artery disease: not just a risk factor”. *Heart* 2020;106(17):pp. 1357–1364.

51. Gallucci G, Tartarone A, Lerose R, Lalinga AV, and Capobianco AM. “Cardiovascular risk of smoking and benefits of smoking cessation”. *J Thorac Dis* 2020;12(7):pp. 3866–3876.

52. Winzer EB, Woitek F, and Linke A. “Physical Activity in the Prevention and Treatment of Coronary Artery Disease”. *J Am Heart Assoc* 2018;7(4).

53. Schnohr P, Lange P, Scharling H, and Jensen JS. “Long-term physical activity in leisure time and mortality from coronary heart disease, stroke, respiratory diseases, and cancer. The Copenhagen City Heart Study”. *Eur J Cardiovasc Prev Rehabil* 2006;13(2):pp. 173–9.

54. Stanaway JD, Afshin A, Gakidou E, Lim SS, Abate D, Abate KH, et al. “Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017”. *The Lancet* 2018;392(10159):pp. 1923–1994.

55. Tapia-Conyer R, Alegre-Díaz J, Gnatiuc L, Wade R, Ramirez-Reyes R, Herrington WG, et al. “Association of Blood Pressure With Cause-Specific Mortality in Mexican Adults”. *JAMA Network Open* 2020;3(9):e2018141–e2018141.

56. D’Agostino R. B. S, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. “General cardiovascular risk profile for use in primary care: the Framingham Heart Study”. *Circulation* 2008;117(6):pp. 743–53.

57. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. “Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2”. *BMJ* 2008;336(7659):pp. 1475–1482.

58. Hippisley-Cox J, Coupland C, and Brindle P. “Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study”. *BMJ* 2017;357:j2099.

59. Woodward M, Brindle P, and Tunstall-Pedoe H. “Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC)”. *Heart* 2007;93(2):pp. 172–176.

60. SCORE2 working group ESC Cardiovascular risk collaboration. “SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe”. *European Heart Journal* 2021;42(25):pp. 2439–2454.

61. Qiu J, Chang Z, Wang K, Chen K, Wang Q, Zhang J, et al. “The predictive accuracy of coronary heart disease risk prediction models in rural Northwestern China”. *Preventive Medicine Reports* 2023;36:p. 102503.

62. SCORE2 Asia-Pacific writing group, Hageman SHJ, Huang Z, Lee H, Kaptoge S, Dorresteijn JAN, et al. “Risk prediction of cardiovascular disease in the Asia-Pacific region: the SCORE2 Asia-Pacific model”. *European Heart Journal* 2024;46(8):pp. 702–715.

63. Mosley JD, Gupta DK, Tan J, Yao J, Wells QS, Shaffer CM, et al. “Predictive Accuracy of a Polygenic Risk Score Compared with a Clinical Risk Score for Incident Coronary Heart Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 627–635.

64. de La Harpe R, Thorball CW, Redin C, Fournier S, Müller O, Strambo D, et al. “Combining European and U.S. risk prediction models with polygenic risk scores to refine cardiovascular prevention: the CoLaus|PsyCoLaus Study”. *Eur J Prev Cardiol* 2023;30(7):pp. 561–571.

65. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, et al. “Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses”. *PLoS Medicine* 2021;18(1) (no pagination).

66. Li L, Pang S, Starnecker F, Mueller-Myhsok B, and Schunkert H. “Integration of a polygenic score into guideline-recommended prediction of cardiovascular disease”. *Eur Heart J* 2024.

67. Samani NJ, Beeston E, Greengrass C, Riveros-McKay F, Debiec R, Lawday D, et al. “Polygenic risk score adds to a clinical risk score in the prediction of cardiovascular disease in a clinical setting”. *European Heart Journal* 2024;45(34):pp. 3152–3160.

68. Ramírez J, Duijvenboden S van, Young WJ, Tinker A, Lambiase PD, Orini M, et al. “Prediction of Coronary Artery Disease and Major Adverse Cardiovascular Events Using Clinical and Genetic Risk Scores for Cardiovascular Risk Factors”. *Circ Genom Precis Med* 2022;15(5):e003441.

69. Weale ME, Riveros-Mckay F, Selzam S, Seth P, Moore R, Tarran WA, et al. “Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries”. *American Journal of Cardiology* 2021;148:pp. 157–164.

70. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. *Nature Medicine* 2023;29(7):pp. 1793–1803.

71. Iribarren C, Lu M, Jorgenson E, Martínez M, Lluis-Ganella C, Subirana I, et al. “Weighted Multi-marker Genetic Risk Scores for Incident Coronary Heart Disease among Individuals of African, Latino and East-Asian Ancestry”. *Sci Rep* 2018;8(1):p. 6853.

72. United Nations. “World Population Prospects 2022: Summary of Results”. *Department of Economiv and Social Affairs, Population Division (2022)* 2022.

73. United States Census Bureau. “Census Bureau QuickFacts: United States”. In: 2023. https://www.census.gov/quickfacts/fact/table/US/RHI725223#qf-headnote-b.

74. “Socioeconomic indices, demography and population structure”. In: *The Evolution and Genetics of Latin American Populations*. Ed. by Francisco M. Salzano and Maria C. Bortolini. Cambridge Studies in Biological and Evolutionary Anthropology. Cambridge: Cambridge University Press, 2001, pp. 55–102. doi: DOI:10.1017/CBO9780511666100.004. https://www.cambridge.org/core/product/99516DDC896C6D7E0A0C12D333A19858.

75. National Institute of Statistics and Geography. *Demographic and Social Information*. https://en.www.inegi.org.mx/programas/ccpv/2020/. 2024.

76. Alegre-Díaz J, Herrington W, López-Cervantes M, Gnatiuc L, Ramirez R, Hill M, et al. “Diabetes and Cause-Specific Mortality in Mexico City”. *N Engl J Med* 2016;375(20):pp. 1961–1971.

77. Abarca-Gómez L, Abdeen ZA, Hamid ZA, Abu-Rmeileh NM, Acosta-Cazares B, Acuin C, et al. “Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults”. *The Lancet* 2017;390(10113):pp. 2627–2642.

78. Thomson B, Tapia-Conyer R, Lacey B, Lewington S, Ramirez-Reyes R, Aguilar-Ramirez D, et al. “Low-intensity daily smoking and cause-specific mortality in Mexico: prospective study of 150 000 adults”. *Int J Epidemiol* 2021;50(3):pp. 955–964.

79. Ziyatdinov A, Torres J, Alegre-Díaz J, Backman J, Mbatchou J, Turner M, et al. “Genotyping, sequencing and analysis of 140,000 adults from Mexico City”. *Nature* 2023;622(7984):pp. 784–793.

80. Mills MC and Rahal C. “The GWAS Diversity Monitor tracks diversity by disease in real time”. *Nature Genetics* 2020;52(3):pp. 242–243.

Chapter 2: Literature Review

Summary

This chapter describes the systematic literature review conducted to search for prior evidence on the construction of polygenic instruments for the prediction of CHD risk in diverse populations. CHD GWAS conducted in large cohorts and subsequently used for PRS construction were documented as part of the broader investigation. The search was conducted using three databases (PubMed, Embase and Medline) and identified 71 eligible studies, including studies that constructed a new PRS for CHD (n=47) and studies that applied an existing PRS in the assessment of subsequent CHD risk (n=24). Only six of the 71 identified studies included populations of admixed-American ancestry.

Most of the studies identified were based solely on European populations, with most involving large sample sizes and advanced methods for the development of the PRSs[1–4]. The key findings from these specific papers suggested that the PRSs generated from European populations were transferable externally among Europeans, however, their transferability was diminished when applied to non-European populations. Among the studies identified there was only one non-European study with a sample size of over 100,000 participants[5]. Due to the low representation in genetic studies of CHD, most PRSs used in non-European studies were derived from European GWAS summary statistics[6–14]. For studies employing a European sourced PRS, the PRS was a weaker predictor of CHD risk among admixed-American and African populations than it was in their respective source populations. While no admixed-American-specific CHD-PRS association studies were found, two identified studies developed a multi-ancestry PRS using meta-analysed GWAS of multiple ancestries, including admixed-Americans. Although only a small proportion of the included participants were admixed-Americans, the strength of the associations of the multi-ancestry PRSs with subsequent CHD risk was greater in the admixed-American sub-cohort

compared to the Eurocentric PRSs.

Through the systematic literature review, the under-representation of admixed-American populations in GWAS and CHD PRSs research was further confirmed. Due to the scarce evidence in admixed-American populations, eight PRSs were selected for evaluation in the Mexico City Prospective Study (MCPS) cohort based on their construction methodology, ancestry involvement and popularity.

2.1 Background and aims

As introduced in **Chapter 1**, CHD is the leading cause of premature deaths globally, and its physiopathology involves high genetic heritability and preventable biological and lifestyle determinants. Due to the polygenic nature of complex diseases such as CHD, PRSs have been developed to assess CHD risk, using information from GWAS, and previous evidence has suggested that PRSs have the potential to improve risk stratification and prediction of CHD[4,15]. Non-Europeans such as admixed-Americans are highly under-represented in genetic studies, accounting for only about 5% of participants in GWAS studies[16]. Since GWAS serves as an important input during PRS construction, this disparity has similarly affected the diversity in CHD PRS research.

The main aim of this systematic literature review was to identify and summarise studies of CHD PRSs by ancestry, with a specific goal to identify any admixed-American studies. Additionally, the review aimed to provide a comprehensive overview on the current state-of-the-art methods for the development and application of PRSs for CHD-risk assessment across ethnically different populations, and to identify potential research gaps and aspects for future work.

2.2 Method of literature review and search

2.2.1 Search strategies

The literature search strategies followed the PRISMA 2020 guideline[17]. Three literature databases were searched systematically (PubMed, Embase and Medline) to identify relevant published research papers on PRSs for CHD. **Table 2.1** presents the search strategy and Medical Subject Headings (MeSH) terms used across the three databases to identify relevant articles in a consistent manner. The primary searches were conducted among full research papers published in English and excluded any preprints. The polygenic exposure was captured by terms “polygenic risk scores”, “polygenic scores”, and “genetic risk scores”. The CHD outcome was defined as ”ischemic heart disease” in Embase and ”myocardial ischemia” in Medline and PubMed, and included either prevalent, incident or mortality cases. Keyword searches were restricted to paper abstracts and titles, and there were no time period restrictions (all the relevant papers were published within the past 20 years). ”Prediction” was not included as a keyword as the strength of association between PRSs and CHD across different ancestries was also the focus of the review.

The systematic search across the three databases was undertaken on 24th November 2022.

2.2.2 Selection of relevant research papers and review

Duplicated records were removed using Endnote and manual screening. To further refine the search results, the systematic literature review screening was divided into two stages: initial screening based on titles and abstracts, and second stage screening based on full text. Both the initial and second stages of the screening aimed to select only relevant reports that assessed the relationship between a CHD PRS and a CHD outcome. Papers included in the final review underwent detailed assessment of their relevance to the topic and findings.

2.2.2.1 Initial screening

This initial stage aimed to efficiently exclude any irrelevant studies through screening of titles and abstracts. The criteria for inclusion required studies to evaluate the association between polygenic predisposition to CHD (measured by a CHD PRS) and risk of CHD (fatal or non-fatal event). For instance, a study using a blood pressure PRS to predict CHD risk would be ineligible for inclusion. Secondly, to ensure that relevant studies were well-powered to reliably detect the associations of interest, any studies with a sample size (to train a PRS or to evaluate a pre-existing PRS) smaller than 5,000 were further excluded. Studies that were not conducted in a general population were also excluded, to ensure unbiased comparison later on in the review. Mendelian Randomisation (MR) studies were excluded as the consideration involved in building a PRS to act as an instrument for MR studies is different from that used in risk prediction studies. For instance, to avoid weak instrument bias, an MR study would only select variants that are strongly associated with the outcome of interest. The methodology is more relaxed for PRSs used in an observational analysis framework aiming to maximise prediction or to assess associations, with millions of SNPs allowed for inclusion in the PRSs, regardless of the significance of their association to the disease of interest. Finally, any retrieved author replies, letters, editorial comments, reviews, methodological papers, or any qualitative, case report and cost-effectiveness studies were excluded at this initial stage of the review.

2.2.2.2 Second-stage screening

The second round of screening was based on full-text assessment. In addition to the criteria set in the initial stage of screening, three additional criteria for inclusion were applied. Firstly, studies that evaluated unweighted PRSs for CHD (mostly published before 2010) were excluded, due to the fact that an unweighted PRS disregards the effect size of each genetic variant on CHD, which may undermine the ability of the PRS to reliably predict CHD risk. Secondly, studies that did not evaluate the direct association between a PRS and CHD risk were further excluded, as it would

be difficult to assess their performance and subsequently compare across studies. For instance, if a study only assessed an interaction effect (such as between a CHD PRS and a dietary factor) but did not report the main association with CHD risk, it would be excluded. Thirdly, short reports containing only intermediate results of ongoing studies were also excluded.

2.2.3 Review of eligible studies

Any studies remaining after the two-stage screening process were thoroughly reviewed and evaluated. These studies were divided into two groups of interest based on: 1) studies that developed and assessed at least one new PRS on CHD risk, and 2) studies that only evaluated the performance of an existing PRS on CHD risk. For all the eligible studies identified, a summary of each study and their results was extracted and reviewed, specifically, year of publication, GWAS data sources, follow-up period, included ancestries (Eurocentric, Admixed-Americans, Asian, African), number of SNPs included, CHD outcome definition, training and evaluation sample sizes, statistical models applied, confounder adjustments, and key findings (strength of association and predictive power). Studies that evaluated a CHD PRS derived predominantly from European populations (GWAS input and training sample) were classified as a European PRS. All others were classified as non-European PRSs. In the group of studies that developed a new PRS, the methodology that they employed was additionally extracted. In addition, the GWAS sources used for the PRS construction in all eligible studies were summarised and reviewed as part of the broader investigation.

2.3 Findings from the literature review

2.3.1 Literature review results

Figure 2.1 displays the process of identifying eligible literature through database searches, screening and review. After applying the criteria described above, a total of 1,718 publications were retrieved from the Medline, Embase, and PubMed databases. After the removal of 955 duplicated publications, 763 unique reports remained.

After the initial screening based on titles and abstracts, 637 studies were excluded and 126 studies remained. The top three ineligible reasons for inclusion were: papers that included a PRS not computed for CHD (n=290); non-epidemiological studies (i.e., reviews, methodology papers; n=174); and a sample size ≤5000 participants (n=67).

Two extra papers were additionally identified during the abstract screening process, which were not identified in the initial literature search. Hence, a subset of 128 papers entered the second stage of the screening, during which another 57 publications were removed (based on the exclusion criteria listed in the **Sections 2.2.2.1** and **2.2.2.2**), leaving 71 eligible studies for final review.

Of the remaining 71 articles included in the extensive review, 47 studies (*Table S1*) constructed at least one new CHD PRS and tested its association with one or more CHD related outcomes, 24 studies (*Table S2*) applied an existing PRS in their association studies of CHD risk, and 26 CHD GWAS studies were identified (*Table S3*).

2.3.2 CHD GWAS

Among the 26 CHD GWAS studies identified, 18 were predominantly based on populations of European ancestries[18–35], five were based on populations of Asian ancestry[5,9,36–38] and three were based on populations of multiple ancestries[34,39,40]. Studies published before the setup of the first CHD GWAS consortium[34] in 2011 generally had a small sample size with fewer than 10,000 participants[28–33,36,37]. Consequently, most eligible PRS studies of CHD were published after 2011, with only one study published before[41].

The GWAS that involved the largest number of admixed-American individuals was the multi-ancestry GWAS meta-analysis conducted by Tcheandjieu *et al.*[39]. This report included over 30,000 admixed-American participants from the Million Veteran Program (MVP)[42], and an additional 25,000 admixed-American participants from external cohorts, making the total number of included admixed-American individuals to 55,000. However, the study only included GWAS

on MVP samples for their subsequent PRS analyses. No GWAS identified in the review was conducted solely on a population of admixed-American ancestry.

2.3.3 General characteristics of relevant studies identified from literature review

A majority of the 71 eligible studies were based on European-dominant populations. Most non-European studies identified either developed their PRSs using European-based GWAS data or evaluated a pre-existing PRS derived from European-dominant cohorts[6–14]. Among the included studies that constructed a new CHD PRS (n=47)[1–10,39,41,43–77], over 80% of them were based on European populations, both in terms of the GWAS used and the populations included for PRS evaluation. Eight studies developed a new CHD PRS[5–10,39,57] involving populations of ancestries beyond Europeans. Studies based on other ancestry groups were generally small in size, or constructed a multi-ancestry PRS that combined multiple GWAS of several ancestry groups. Only one Asian study had a GWAS and training sample size of over 100,000 participants[5]. Two multi-ancestry studies included over 60,000 individuals of African ancestry and more than 30,000 individuals of admixed-American ancestry from MVP[42] in their GWAS samples but used different methods for PRS construction[39,57]. However, one study included only European population in their PRS evaluation samples[57]. Notably, the top 10 largest studies were based solely on European populations[2,43–51], with sample sizes over 300,000 (*Table S1*). The same applied for studies that evaluated previously published CHD PRSs (n=24)[11–14,78–97]. Only six studies evaluated PRSs using populations of multiple ancestries[11–14,87,95], with the largest including approximately 18,000 non-European individuals. In contrast, there were six studies that evaluated PRSs in European-dominant populations with sample sizes exceeding 100,000 (*Table S2*). Only three studies evaluated a pre-existing PRS constructed incorporating genetic information from non-European populations[7,9,57].

Among all the eligible studies identified, six included populations of admixed-American ancestry alongside other ancestries (see **Table 2.2**). Three studies constructed a new PRS[6,39,57],

and three studies evaluated pre-existing CHD PRSs[11–13]. Among three studies that constructed a new PRS, two incorporated admixed-American GWAS data (from Million Veteran Program (MVP)[42]) and created a multi-ancestry CHD PRS[39,57], while the third study used only European GWAS information for PRS construction and then evaluated the PRS in a multi-ancestry cohort. Apart from two studies that included 30,000 admixed-American samples in their GWAS[39,57], the number of admixed-American participants included in the GWAS, training or evaluation samples of the other four studies were relatively small, ranging from 1,000 to 7,000[6,11–13,39].

Among the 24 studies that evaluated the performance of a pre-existing PRS for CHD risk assessment, the three most commonly evaluated CHD PRSs were Khera *et al.*[1] (applied in 14 studies[11,12,14,57,59,67,78–80,83,88,93,96,97]), Inouye *et al.*[2] (applied in 10 studies[7,9,11,57,59,67,79,80,82,93]) and Tada *et al.*[64] (applied in seven studies[11,57,80,89–92]). Furthermore, all these three common PRSs were constructed based on predominantly European CHD GWAS.

The common statistical method for assessing the relevance between a PRS and CHD risk involved standardising the PRS and then performing logistic (for prevalent CHD) or Cox regression (for incident CHD or CHD mortality)[1–8,11–13,39,44,45,49–53,56,57,59,61–64,69,73–75,77,78,82,83,86,88,93,96,97]. For studies that assessed the PRS predictability of CHD, 37 studies reported model discrimination using the area under the receiver-operating-characteristic curve (AUC) for logistic models, or the Harrell's C-index for Cox models[1–5,7–9,11–13,41,49,56,58,59,61,63–65,67–72,74–77,79,80,82,88,89,93,97]. In addition, 16 studies[4,8,9,13,41,56,65,69,70,74,75,77,82,83,88,89] reported net reclassification improvement index (NRI).

2.3.4 Performance of CHD PRSs among populations of European ancestry

2.3.4.1 Magnitude of associations

Among the studies that constructed a new European PRS, 36 of them[1–4,41,43,44,46–52,54–66,69–77] reported the strength of the associations, with all studies showing a significant association between the PRS and risk of CHD. For studies that analysed the effect of a PRS after adjustment

for age, sex and population structure (such as genetic principal component [PCs]) as covariates, the strengths of the association per one standard deviation (SD) higher genetic predisposition ranged from an odds ratio (OR) of 1.22 to 1.77[1,3,44,45,51,56,74,75,77] or a hazard ratio (HR) of 1.07 to 1.73, respectively[2,4,56,59,62,74,75,77].

When a PRS derived from a European population was evaluated in another European population in 18 studies, the associations were all significant and positive[78–86,88–94,96,97]. For example, in its training set assessment, the Khera *et al.*[1] PRS yielded an OR per SD for CHD risk of 1.72 (95% CI 1.67-1.78). When evaluated externally among other European populations, the same PRS yielded similar ORs, with association strength ranging from 1.46 to 1.89[11,39,78,93,96]. In other survival analyses on incident CHD, each SD higher level of the Khera *et al.* PRS was associated with a 15% to 50% higher hazard of incident CHD[11,88,97], a magnitude of effect comparable to that from non-survival analyses. Of the other popular PRSs used for external evaluation, the Inouye *et al.*[2] and the Tada *et al.*[64] PRSs also showed strong magnitudes of associations when evaluated among European populations irrespective of the statistical model applied (see **Table 2.3**).

2.3.4.2 Predictive ability

Model discrimination varied greatly between identified studies due to differences in model adjustments in addition to the PRSs employed. With age, sex and population structure as covariates in addition to the PRS, the model discrimination ranged from an AUC of 0.63 to 0.87 in studies based on logistic regression[1–3,12,58,75,77,79,97], while the Harrell's C index ranged from 0.62 to 0.76 in Cox regression based studies[2,4,56,59,77,88].

Generally, when evaluated externally, the model discrimination with a PRS tended to decrease slightly compared to the model discrimination reported by the source studies that constructed the CHD PRSs[64,67,79,80,88,89,93,97]. For instance, the model AUC reported by the Khera *et al.*[1] PRS was 0.81 (95% CI 0.81-0.81) when the PRS was evaluated internally among its own testing sam-

ples, but decreased to around 0.75 when evaluated externally in other European cohorts[11,88,93] given similar covariate adjustments (i.e. age, sex and genetic PC]).

When a specific PRS was included in a CHD risk-prediction model, the change in discrimination and the NRI were mostly positive, but modest. For example, Inouye *et al.* reported a 2.8% improvement in the AUC by incorporating their CHD PRS in a logistic model[2]. When this PRS was validated externally using a Cox model, the increment in Harrell' C index was 2.2% (95% CI 1.7%-2.6%)[82]. In the case of the Khera *et al.*[1] PRS study, an NRI improvement of 7.7% (95% CI 7.5%-7.9%) and of 8.5% (95% CI 7.1%-9.8%) were reported, when the additive value of the PRS was compared to that of the clinical QRISK3 score[83], and the pooled cohort equations (PCE)[88], respectively, using logistic models. In another Cox-based study that compared a PRS with the relevance of the PCE to predicting CHD risk, PRS improved the Harrell's C index by 2% (95% CI 1%-3%) and NRI by 4.0% (95% CI 3.1%-4.9%)[4]. One other study also reported a minor improvement in the AUC of 0.9% (95% CI 0.3%-1.5%) and an insignificant change in the NRI of 0.4% (95% CI -3.4%-4.1%) after including a 152-SNP PRS in a logistic model to predict risk for prevalent CHD[74].

2.3.5 Performance of CHD PRSs among populations of non-European ancestries

2.3.5.1 Magnitude of associations

Of the eight studies that developed a new non-European CHD PRS, they either used CHD GWAS sourced from non-Europeans or trained in a sample that included (but not exclusively restricted to) non-European populations[5–10,39,57]. For studies that internally evaluated the study-specific derived CHD PRSs in models mainly adjusted for age, sex and population structure, the strengths of association with prevalent CHD risk ranged from an OR per SD of 1.43 to 1.84[5,7,8,39,57]. One study evaluated the strength of association with incident CHD risk and reported an HR per SD of 1.22 (95% CI 1.09-1.33)[5]. Recently, a new multi-ancestry PRS was developed using genetic information from four different ancestry groups (specifically, European, admixed-American,

South-Asian and Africans). This multi-ancestry score demonstrated good performance across all ancestry groups with an OR per SD higher polygenic predisposition to CHD of 1.61 (95% CI 1.53-1.70) among admixed-Americans, 1.72 (1.69-1.75) among Europeans, 1.83 (1.69-1.99) among South Asians and 1.25 (1.21-1.29) among Africans, respectively[57].

Only two non-European PRSs were evaluated by other studies[5,8]. One study first conducted a CHD GWAS in a cohort of 150,000 Japanese adults, and subsequently developed a multi-ancestry (Japanese- European) CHD PRS that yielded an OR per SD of 1.84 (1.74-1.94) when evaluated internally among its own testing population subset[5]. The performance of this multi-ancestry score remained significant when evaluated externally in a Middle-Eastern cohort of different ancestry (i.e., from West Asia) of 7,000 participants, and yielded an OR per SD of 1.81 (1.66-1.98)[7]. When a European-derived PRS was evaluated in the same Middle-Eastern cohort, it yielded an OR per SD of 1.54 (1.43-1.66)[7]. The same study also assessed the performance of another CHD PRS trained using a South Asian cohort which reported an OR per SD of 1.60 (1.32-1.94)[8] in the source study, and yielded an OR per SD of 1.53 (1.42-1.64) in this Middle-Eastern sample[7].

When the European-derived PRSs were evaluated in non-European populations, their performance was mostly attenuated[11,12,39], with wide confidence interval due to the modest sample sizes available. For instance, when the Eurocentric PRS by Khera *et al.*[1] was evaluated separately in three different ancestry groups (i.e., European, African and admixed-American), its performance in the European group was comparable to that reported in its source study with OR per SD of 1.66 (1.62-1.71), but much weaker for admixed-Americans or Africans, with OR per SD of 1.42 (1.25-1.61) and 1.30 (1.21-1.41)[11], respectively (**Table 2.3**).

It was difficult to compare the performance of the various PRSs across the six studies that included populations of admixed-American ancestry, largely due to the high variability in results driven by the small sample sizes of these admixed-American studies. For instance, when a study evaluated a 163-SNP CHD PRS only in its admixed-American subgroup of 1,500 individuals us-

ing logistic regression, a high effect size but wide confidence interval was observed, indicating great uncertainty (OR per SD= 1.87 [95% CI 1.19-2.95])[6].

2.3.5.2 Predictive ability

Four studies evaluating a non-European CHD PRS reported model discrimination or NRI[5,7–9]. The Japanese score described in **Section 2.3.5.1** yielded a model AUC of 0.67 (95% CI 0.66-0.69) when evaluated internally and 0.67 (95% CI 0.65-0.69) when externally validated in a Middle Eastern cohort[7]. Similarly, a PRS derived from South Asians[8] showed an AUC of 0.66 during internal validation and an AUC of 0.68 (95% CI 0.67-0.70) when externally evaluated in the same Middle Eastern cohort[7]. Moreover, the South Asian score proved to be the best performing PRS when evaluated in a group of South Asians in Britain (British Pakistan and Bangladeshi), with a modest incremental AUC of 0.9% (95% CI 0.6%-1.2%)[9].

Three studies reported model AUC when evaluating a European PRS in a non-European cohort[7,11,12]. The Inouye *et al.*[2] PRS yielded a model AUC of 0.69 (95% CI 0.67-0.70) in the Middle Eastern cohort described above[7] and a model AUC of 0.78 and 0.79 when separately evaluated in the African and admixed-American subgroup of a mixed cohort[11]. The Khera et al.[1] PRS was also evaluated in the same report and yielded a model AUC of 0.78 and 0.77 for the African and admixed-American subgroup, respectively[11]. However, the sample sizes of the evaluation subgroups were small (7,500 and 2,500), so these results were subject to high uncertainty.

2.3.6 Admixed-American-sourced CHD PRS

Two studies developed a new multi-ancestry PRS incorporating genetic information from admixed-American populations[39,57]. Both studies demonstrated that their CHD PRSs outperformed the European-sourced CHD PRSs. One study that constructed a multi-ancestry PRS using GWAS data and training samples with populations of admixed-Americans, African and European ancestries showed a considerably greater OR per SD for CHD of 1.43 (95% CI 1.27-1.61)[39] compared to the OR yielded by an European PRS[2], of 1.18 (1.07-1.30), when evaluated in a subgroup of

admixed-American individuals. Similarly, another multi-ancestry PRS constructed using GWAS information for four ancestries (admixed-American, European, Asian and African)[57] and trained on a European cohort, demonstrated the greatest association with CHD risk, with an OR per SD of 1.61 (1.53-1.70) when internally evaluated among 16,000 held-out validation samples of admixed-Americans. This was notably higher than the second-best performing (European-sourced) PRS in the study, which yielded an OR per SD of 1.52 (1.44-1.60). Neither of these two studies reported on the predictive performance of their respective CHD PRSs.

The performance of CHD PRSs was closely associated with the methodology employed during PRS construction. The next section will summarise the PRS methodologies employed across eligible studies.

2.3.7 PRS construction methodology

Through the review of the relevant papers identified, five main methods for constructing a PRS for CHD risk were identified: selection of significant variants; *pruning and thresholding* (P+T); machine learning (ML) methods; Bayesian methods; and the *ensemble* method.

2.3.7.1 The significant variants method for PRS construct

This method uses significant variants identified from previous GWAS (i.e., genome-wide, false discovery rate (FDR), nominal). A PRS is constructed as the weighted sum of the significant variants effect sizes and their count of risk alleles. Among the 47 studies that developed a new CHD PRS, 24 studies[6,10,41,43,44,47,48,50–55,60–64,66,69,70,72–74] utilised this particular method for PRS construction. The number of SNPs included in those PRSs is generally small, ranging from fewer than 10 to a few hundred[44,61].

2.3.7.2 Pruning and thresholding

This method involves two consecutive steps: 1) pruning; and 2) thresholding. Genetic variants are highly correlated to one another due to linkage disequilibrium (LD). The pruning step

subsets the relevant variants to a group of uncorrelated variants based on their GWAS effect sizes and SNP-SNP pairwise correlation[98] . The thresholding step then sets a particular p-value significance threshold and variants with a p-value significance below that 'threshold' value will be removed (similar to **Section 2.3.7.1**). Eight identified studies[4,9,39,45,46,65,67,75] adopted this methodology for the creation of a PRS for CHD.

2.3.7.3 Machine learning methods

Two studies identified from the literature review[4,65] developed their CHD PRSs using a machine learning method, specifically *lassosum*[99], which employs Least Absolute Shrinkage and Selection Operator (LASSO)[100] to account for LD among genetic variants. This method requires a large training sample for hyper-parameter tuning. It has been shown that *lassosum* outperforms pruning and thresholding in terms of predictive accuracy[4,99].

2.3.7.4 Bayesian methods

There were 14 studies[1,3,5,8,9,49,50,56–59,65,76,77] that evaluated PRSs for CHD developed using various Bayesian methods. The general approach involves estimating posterior SNP weights based on a predefined SNP effect size prior and effect sizes derived from GWAS results. The prior and estimation algorithms of posterior SNP weights vary across different methods. PRS-CS(x)[76,101], which uses a continuous shrinkage prior, and LDpred[102,103], which uses a normal mixture prior, are the most commonly adopted methods. These methods generally require a large training sample for hyper-parameter tuning and a testing sample for PRS evaluation. The PRSs generated by this approach generally demonstrated strong performance when validated internally and externally, showing robustness in association strengths and predictive ability[3,11,78,80,88,96].

2.3.7.5 Ensemble method

This method is not technically a PRS construction method as it does not involve calculations of SNP effect sizes. Instead, it integrates PRSs constructed from other methods into a meta-PRS.

The integration could be simple addition or could involve more complicated algorithms[2]. The predictive ability of the meta-score is therefore highly dependent upon the quality of the PRSs included. Two studies[2,7] adopted this method.

2.4 Discussion

2.4.1 Performance of CHD PRS in population of European ancestry

As expected, the majority of the studies identified were based predominantly on populations of European ancestry. Findings from these studies consistently suggested that CHD-specific PRSs generated from European populations showed comparable performance when evaluated externally among Europeans. The magnitudes of the associations were typically comparable to those reported in the source studies with similar model adjustments[11,82,83,93,96] and the predictability stayed strong[67,79,80,88,93,97]. There was also consistent evidence suggesting that including a CHD PRS into a model with a clinical risk score could improve model discrimination among European populations moderately[2,4,65].

Among European populations, there has been a rapid development in CHD PRSs using more advanced (i.e., Bayesian and machine learning) methods, which has yielded promising results. This progress can be attributed to the availability of large European cohorts for PRS training. As detailed in **Section 2.3.3**, most studies with large sample sizes (and thus study power) were based predominantly on populations of European ancestry[2,43–51]. This aligns with the finding that most studies developed a new PRS using more advanced methods[1,3,9,49,50,56,58,59,65] trained on large European cohorts, as these methods (e.g. lassosum[99], PRS-CS[76]) typically require a large tuning sample in order to achieve optimal performance.

2.4.2 Performance of CHD PRS in populations of non-European ancestry

Despite rapid growth in PRS construction methodology, non-Europeans remained highly underrepresented in PRS studies on CHD. As shown in **Section 2.3.2**, only a few GWAS studies

were conducted among non-European populations, and most of them were focused on Asian populations. This resulted in most non-European studies evaluating the performance of CHD PRSs using European-derived PRSs[6–14]. For studies employing European-derived PRSs, the association between a PRS and CHD risk was generally weaker among admixed-American and African populations compared to European populations from the same cohort[11,12,39]. However, the results and their credibility were highly variable due to modest sample sizes. One study showed a stronger association between the PRS and CHD when evaluated in a sub-cohort of admixed-American individuals, however this result was subject to high uncertainty due to the small sample size of 1,500 people[6]. This highlights the potential limitations in the transferability of external CHD PRSs across ancestries and the limitations of generalising findings beyond ancestries similar to the source populations.

For studies that constructed a new non-European PRS[5–10,39,57], most of them trained the PRSs in multi-ancestry cohorts, or sourced multi-ancestry GWAS that included more participants of European than of non-European ancestries. Nonetheless, in these studies, the newly-derived PRSs associated more strongly with CHD risk, highlighting the importance of incorporating ancestry-specific information for enhancing the PRS performance. In such studies, however, evidence on the predictive performance of either a European or a non-European PRS was limited. Non-European PRSs generally exhibited modest or uncertain performance (often due to small sample size) when evaluated externally in the sub-cohorts of non-European ancestries[7,9,11,12].

2.4.3 Current methods for PRS construction

The strong performance of a PRS involving millions of SNPs and constructed using more advanced methods (i.e., ML, Bayesian)[1–4,56,59], has demonstrated important superiority in predicting subsequent CHD risk over simpler methods (*Pruning +Thresholding* or significant variants)[1,2] when applied in European populations. However, the evidence of the utility of PRSs constructed using more advanced methods in populations of non-European ancestry is insufficient, due to

the small number of studies conducted among non-Europeans.

In addition, it has also been shown that PRSs constructed using GWAS meta-analysis of multiple ancestries have comparable or even better performance than single-ancestry non-European PRSs among non-European populations, as demonstrated by Patel *et al.*[57], Koyama *et al.*[5] and Tcheandjieu *et al.*[39]. Moreover, the Koyama *et al.*[5] PRS has demonstrated greater association with CHD risk than other European PRSs when evaluated externally, although their model AUCs were comparable[7]. Due to the scarcity in GWAS for non-Europeans, small GWAS were not sufficiently powered to identify significant variants. GWAS meta-analysis could boost sample size, thereby providing more precise SNP effect estimates. The PRSs constructed using GWAS of multiple ancestries could also be more generalisable to diverse populations.

2.4.4 CHD PRS constructed with admixed-American genetic information

Only three studies including admixed-American individuals constructed a new PRS (**Table 2.2**) and only two of these included genetic information for admixed-Americans. No CHD PRS was constructed using an exclusive or dominant "admixed-American GWAS". Both studies constructed a CHD multi-ancestry PRS involving multi-ancestry GWAS during their construction. Although only a small proportion of the included participants were of admixed-American ancestry, both studies showed that the strength of association of their multi-ancestry PRS with CHD risk was greater in their admixed-American sub-cohort than when using the European-only PRS. However, neither of the two PRSs was externally evaluated. Furthermore, one multi-ancestry PRS was trained solely in MVP[39,42], a cohort with over 90% of participants being men and the other PRS was trained using participants of the UK Biobank[57,104], a cohort with most participants of European ancestry. Therefore, the transferability of these two multi-ancestry PRSs remains questionable.

2.5 Conclusion

The systematic literature review identified 71 relevant studies of CHD risk prediction with polygenic predisposition in an observational epidemiological framework. Of those, 47 studies constructed a new PRS and 24 evaluated an existing PRS in a different population. 26 GWAS were identified, with most of these based solely on Europeans. Studies of other ancestries were mostly multi-ancestry studies incorporating small sample sizes of non-European populations and European samples were included to boost the overall sample size and enhance statistical power. There was no GWAS or PRS study conducted exclusively in an admixed-American-dominant or admixed-American-only cohort. The most commonly evaluated PRS was derived from Europeans. Most studies assessed the performance of specific PRSs on CHD risk using either logistic or Cox regression models, and reported the predictive power of the PRS using model discrimination or NRI. European PRSs derived using advanced methods (e.g. LDpred[102]) consistently showed strong performance among Europeans, both in association strengths and predictive power. Their performance in other ancestries including admixed-Americans became weaker and was subject to high uncertainty, warranting further research.

To account for problems of small sample size, three studies[5,39,57] utilised multi-ancestry GWAS during the construction of the PRS. Although these PRSs demonstrated stronger association with CHD than European-only PRSs, there is insufficient evidence of their predictive power and generalisability. Moreover, increasing the included non-European ancestry sample size (i.e., training sample and input GWAS) may further improve the CHD PRS performance.

2.6 Selection of pre-existing PRSs for evaluation in the Mexico City Prospective Study cohort

The MCPS cohort is a unique resource for performing genetic studies of CHD, given the scarcity of admixed-American studies (as shown above). As one of the largest blood-based cohorts of admixed-American people (Mexicans), MCPS has genetic data available for over 140,000

admixed-American individuals, which is more than four times the size of the largest admixed-American cohort included in the multi-ancestry GWAS identified through review[39] (see **Section 2.3.6**), making it a valuable resource for PRS studies. For evaluation studies, MCPS will enable more robust findings and further clarify the transferability of previously published PRSs. Beyond evaluation, the genetic data that MCPS possesses will aid the development of a CHD PRS that more closely represents the complex genetic pattern in admixed-Americans.

2.6.1 PRS selection for evaluation in MCPS

Eight PRSs were selected from the 47 eligible studies in this review to be further evaluated for their transferability in MCPS (in **Chapter 4**). The PRSs were selected based either on the methodology they employed, the relevance to the ancestry represented in the MCPS cohort and/or their extensive evaluation in the existing literature. The aim was to include a spectrum of CHD PRSs to evaluate which method and ancestry combination generates the PRS that would result in the best performance in admixed-Americans. Overall, four PRSs selected[1–3,64] were predominantly derived from European ancestry populations using a variety of methods. One PRS selected used GWAS sources of European ancestries but was evaluated in a population of multiple ancestries[6]. Three PRSs selected involved multiple ancestries in their GWAS input and evaluation samples[5,39,57]. Their detailed characteristics are presented in **Table 2.4**.

Table 2.1: Systematic Search Strategy

Searched database	Search process
Medline (Ovid MEDLINE® Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE® Daily and Ovid MEDLINE®) 1946 to present	1 polygenic risk score*(abstract/title) 2 genetic risk score*(abstract/title) 3 polygenic score*(abstract/title) 4 1 or 2 or 3 5 cardiovascular disease*(abstract/title). 6 cardiovascular diseases/ or exp myocardial ischemia/ 7 5 or 6 8 4 and 7
	Myocardial ischemia includes: • *Acute coronary syndrome* • *Angina pectoris* • *Coronary disease (including CHD)* • *Kounis Syndrome* • *Myocardial Infarction* • *Myocardial Reperfusion injury*
Embase (Ovid) 1974 to present	1 polygenic risk score*(abstract/title) 2 polygenic score*(abstract/title) 3 genetic risk score*(abstract/title) 4 1 or 2 or 3 5 cardiovascular disease/ or exp ischemic heart disease/ 6 cardiovascular disease*(abstract/title) 7 5 or 6 8 4 and 7 9 limit 8 to (conference abstract or conference paper or "conference review" or "preprint (unpublished, non-peer reviewed)" or "review") 10 8 not 9
	Ischemia heart disease includes: • *Acute coronary syndrome* • *Angina pectoris* • *Cardiac allograft vasculopathy*

	• *Coronary artery atherosclerosis* • *Coronary artery constriction* • *Coronary artery thrombosis* • *Coronary subclavian steal syndrome* • *Heart infarction* • *Heart muscle ischemia* • *Kounis syndrome* • *Myocardial hibernation*
PubMed 2005 to present	**("polygenic score*"[Title/Abstract] OR "polygenic risk score*"[Title/Abstract] OR "genetic risk score*"[Title/Abstract]) AND ("myocardial ischemia*"[MeSH Terms] OR "cardiovascular disease*"[Title/Abstract])**
	Myocardial ischemia includes: • *Acute coronary syndrome* • *Angina pectoris* • *Coronary disease (including CHD)* • *Kounis Syndrome* • *Myocardial Infarction* • *Myocardial Reperfusion injury*

Table 2.2: Studies with population of admixed-American ancestry

Study population, (Author/Year)	Ethnicity and Sample size	Follow-up time (median)	Validation sample size	Outcome definition	PRS used/constructed	GWAS source For new PRS	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
UK Biobank, Million veteran program, Gene&Health (**Patel *et al.*, 2023**[57])	European: 4,412/112,237(3.78%)	NA	In MVP: **2,140/14,293 (Admixed-American)** 29,171/95,296(European) 4,831/28,265(African) In G&H: 853/16,021(South Asian)	Prevalent CHD Outcome definition: angina, MI, coronary bypass or revascularisation (CABG, PTCA), coronary thrombolysis, chronic IHD, acute MI discharge	For each of 9 CHD trait, a multi-ancestry PRS was created by summing over ancestry specific PRS. The CHD PRS summed over all mult-ancestry PRS of all traits.	Genes & Health, FinnGen, Million Veteran Program, Biobank Japan and CC4D excluding UK Biobank samples	Logistic	Age, sex, genotyping array, 10 genetic PCs	OR per SD: **Admixed-American: 1.61 (1.53-1.70)** European: 1.72 (1.69-1.75) South Asian: 1.83 (1.69-1.99) African: 1.25 (1.21-1.29)	
Million veteran program (MVP) , (**Tcheandjieu *et al.*, 2022**[39])	**Admixed-American: 574/6,314 (9.09%),** European: 6,158 / 67,738 African American: 1,552/17,072	NA	Testing: **6,378/30,648(Admixed-American)** 95,151/292,438(European), 17,202/76,709 (African American),	Incident CHD and Prevalent CHD Outcome definition: acute MI discharge, coronary revascularization, acute or old MI, coronary syndrome, angina, chronic IHD	4 existing PRS , 1 new PRS created using P+T on MVP GWAS meta-analysis coefficients, with randomly selected MVP participants as LD reference panel	GWAS meta-analysis (White) MVP+Van der Harst *et al.*+Nikpay *et al.* ; GWAS meta-analysis(Admixed-American and Black) MVP +Biobank Japan)	Logistic	batch, age, sex, 10 genetic PCs	OR per SD (prevalent CHD):, meta GRS: **Admixed-American 1.18 (1.07-1.30),** White 1.24 (1.21-1.27), Black 1.16 (1.10-1.23), population specific PRS: **Admixed-American 1.43 (1.27-1.61),** White 1.35 (1.31-1.38), Black 1.21 (1.15-.28)	
GERA (**Oni-Orisan *et al.*, 2022**[6])	**Admixed Americans 71/1,538 (4.62%),** European 1,185 /21,284, Black 28/540, East Asian 61/1,489	8.2 yrs	NA	Prevalent MI Outcome definition: fatal or non-fatal MI	164 SNPs from a review, used coefficients in reviewed GWASs as weights.	Erdmann review 2018	Cox	sex, age, hypertension, diabetes, smoking , restricted to statins non-users	HR per SD: **Admixed Americans 1.87 (1.19–2.95),** White 1.59 (1.42-1.78), Black 0.77 (0.32-1.85), East-Asians 2.05 (1.27-3.31)	
Partners Healthcare Biobank (1), Penn Medicine Biobank (2), Mount Sinai BioMe Biobank (3) , **(Aragam *et al.*, 2020**[12])	**Admixed Americans (n=7,847)** Overall: 11,020 / 47,108, African (n=9,773) East Asian (n=167) European (n=29,212), South Asian (n=109)	NA	NA	Prevalent CHD Outcome definition: MI, acute or chronic CHD, coronary bypass or revascularisation (CABG, PTCA), coronary thrombolysis	Khera *et al.*[1], mean centred and normalised by genetic ancestry within each cohort	NA	Logistic (cohort-specific) and FE meta-analysis (all)	age, sex, 10 ancestry PCs, genotyping array (Partners and Penn Medicine Biobanks only).	OR per SD: **Admixed Americans (Mt. Sinai only, n=7048) 1.50 (1.44 - 1.57),** Overall 1.42 (1.38-1.46), top 5% vs rest: 2.28 (2.04-2.53), top 20% vs rest: 1.92 (1.80-2.04),	AUC - **Admixed Americans= 0.63,** Partners Biobank=0.59, Penn Biobank=0.60, BioMe= 0.61,

eMERGE network **(Dikilitas *et al.*, 2020[11])**	**Admixed-American: 120/2,493 (4.81%),** European: 2,221 / 45,645, African: 311/7,597	**10.4 yrs. (HE),** 11.7 yrs. (EA), 9.2 yrs. (AA),	NA	Incident (primary) and prevalent CHD Outcome definition: MI, Coronary Revascularisation (PCI, CABG)	2 restricted PRS: 1. Tikkanen *et al.*[63] and 2. Tada *et al.*[64], 2 genome-wide PRSs 3. Inouye *et al.*[2] and 4. Khera *et al.*[1]	NA	Cox and Logistic	sex, eMERGE site, 5 genetic PCs and age at first EHR record, duration of HER (logistic only)	HR per SD : Admixed-Americans HRs **1) 1.14 (0.94–1.37) 2) 1.13 (0.93–1.36) 3) 1.53 (1.23–1.90) 4) 1.16 (0.96–1.41)** EA 1) 1.18 (1.13–1.23) 2) 1.20 (1.15–1.25) 3) 1.53 (1.46–1.60) 4) 1.50 (1.43–1.56) AA 1) 1.11 (0.99–1.24), 2. 1.05 (0.94–1.17), 3. 1.27 (1.13–1.43), 4. 1.19 (1.07–1.33); ORs: **HE: 1. 1.27 (1.12–1.42), 2. 1.20 (1.06–1.35), 3. 1.93 (1.67–2.22), 4. 1.42 (1.25–1.61),** EA: 1. 1.24 (1.21–1.28), 2. 1.28 (1.25–1.32), 3.1.73 (1.68–1.78), 4. 1.66 (1.62–1.71); AA:1. 1.07 (0.99–1.16), 2. 1.05 (0.98–1.14), 3. 1.40 (1.30–1.52), 4. 1.30 (1.21–1.41);	AUC/Cox: **HE=0.65-0.68,** EA=0.67- 0.72; AA= 0.65- 0.66; AUC/Logistic **HE=0.77-0.79,** EA=0.75-0.77; AA= 0.76- 0.77
GERA **(Iribarren *et al.* 2 , 2018[13])**	**Latinos: 316/4,349 (7.26%)**, African (95/2,089), East Asians (39/4,804), Overall: 450 / 11,242	8.7 yrs	NA	Incident CHD Outcome definition: MI hospital discharge, un/stable angina, coronary revascularization procedures (CABG, PCI), and death from angina, MI, or hypertensive heart disease	2 PRSs from Iribarren *et al.*[68] (GRS_12, GRS_51)	NA	Cox and FE meta-analysis combining ethnic groups	a) 6 genetic PCs b)+age, sex, TC, HDL-C, SBP, DBP, smoking, DM c)+family heart disease d)education , BMI, antihypertensive, statins, drinking	OR per SD: GRS_12 a) 1.15(1.04-1.26), b)1.17(1.06-1.28), c) 1.17(1.06-1.28) d)1.15(1.04-1.27); GRS_51: a) 1.17(1.06-1.29), b) 1.18 (1.07-1.30), c) 1.18 (1.07-1.30), d) 1.16(1.05-1.28)	Harrell's C -- adjusted for FRS: GRS_12: 0.72; GRS_51: 0.72 NRI -- include FRS: GRS_12: 0.06 (0.02–0.11), GRS_51: 0.03 (−0.00–0.06)

MI: Myocardial Infarction, CHD: Coronary Artery Disease, CVD: Cardiovascular Disease, BMI: Body Mass Index, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, OR: Odds Ratio, RR: Risk Ratio, HR: Hazard Ratio, CI: Confidence interval, PRS: Polygenic Risk Score, PC: Principal Components, CC4D: CARDIoGRAMplusC4D consortium, GWAS: Genome-wide Association Study, AUC: Area under the ROC curve, NRI: Net reclassification Index, P+T: Pruning and Thresholding, SD: standard deviation, TC: Total cholesterol, HDL-C: High Density Lipoprotein Cholesterol, LDL-C: Low Density Lipoprotein Cholesterol, TDI: Townsend Deprivation Index, IPAQ: The international Physical Activity Questionnaire, CABG: Coronary Artery bypass graft, PTCA: Percutaneous Transluminal coronary angioplasty, PCI: Percutaneous Coronary Intervention, SNP: Single Nucleotide Polymorphisms, BG: Blood Glucose, TG: Triglycerides, PCE: pooled cohort equation, SD: Standard deviation, FRS: Framingham Risk Score, MDCS: Malmo Diet and Cancer Study, FE: Fixed effect, RE: Random effect, ASCVD: atherosclerotic cardiovascular disease

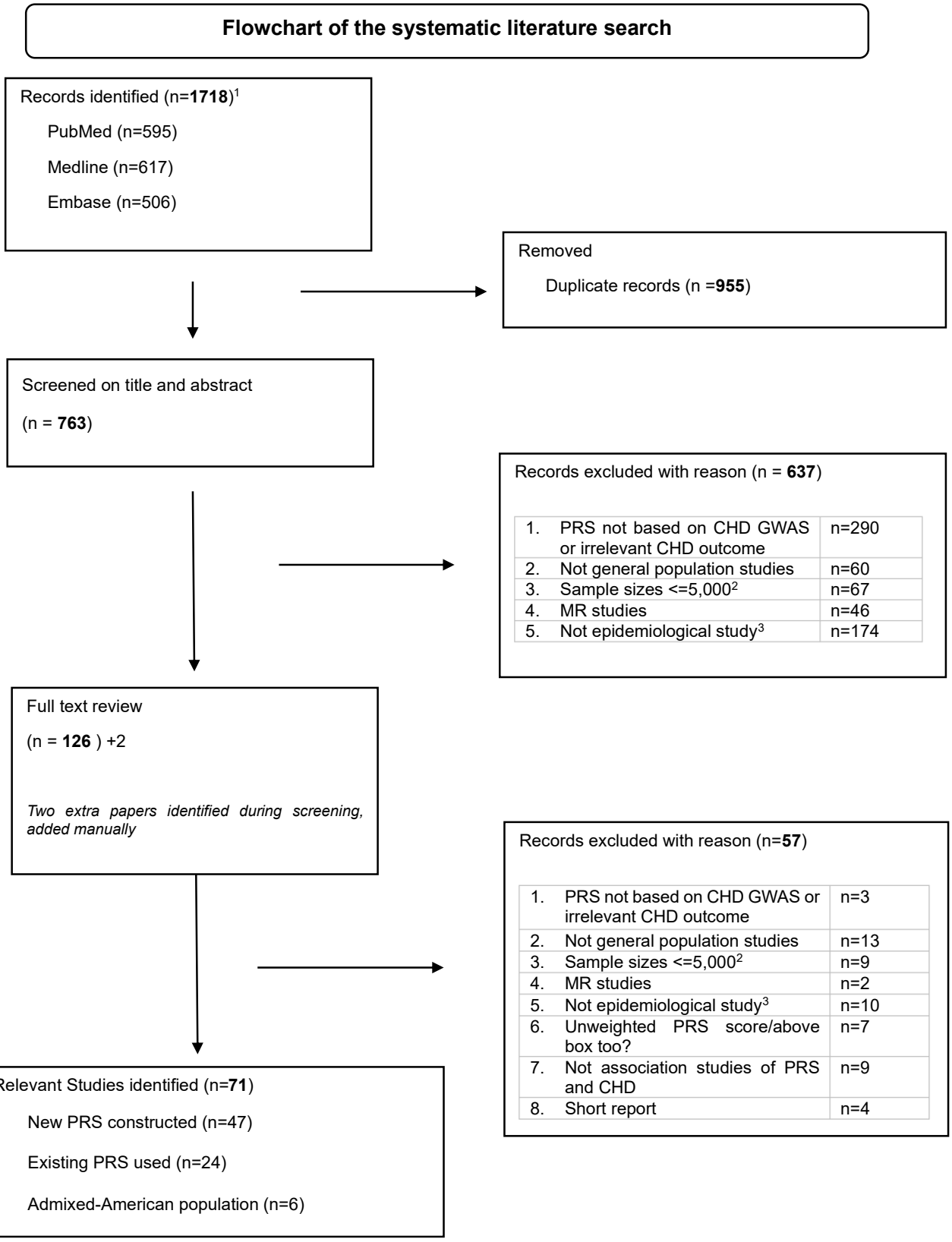
Table 2.3: Associations between popular CHD PRS and CHD reported by external evaluation studies

PRS name	European population		Non-European population	
	OR per SD	HR per SD	OR per SD	HR per SD
Khera *et al.*[1]	Khera *et al.* [1] (source): 1.72 (1.67-1.78) Wunnemann *et al.*[93]: 1.61 (1.51–1.73), Dikilitas *et al.*[11]: 1.66 (1.62–1.71), Aragam *et al.*[12]: 1.52 (1.46-1.58), Mosley *et al.*[96]: 1.89(1.75-2.03), Tcheandjieu *et al.*[39]: 1.36 (1.35-1.37) Goodman *et al.*[78]: 1.695(1.672-1.718) Ramirez *et al.*[83]: 1.559 (1.519-1.598)	Dikilitas *et al.*[11]: 1.50 (1.43–1.56), Hindy *et al.*[88]: 1.45 (1.40–1.49) (MDCS), 1.53 (1.49–1.56) (UKB) Khan *et al.*[97]: 1.22 (1.15--1.30)	Admixed-American: Aragam *et al.*[12]: 1.50 (1.44 - 1.57) Dikilitas *et al.*[11]: 1.42 (1.25–1.61) Tcheandjieu *et al.*[39]:1.32 (1.28-1.36) African: Aragam *et al.*[12]: 1.29 (1.23 - 1.34) Dikilitas *et al.*[11]: 1.30 (1.21–1.41) Tcheandjieu *et al.*[39]:1.10(1.08-1.12)	Admixed-American: Dikilitas *et al.*[11]: 1.16 (0.96–1.41) African: Dikilitas *et al.*[11]: 1.19 (1.07–1.33)
Inouye *et al.*[2]	Dikilitas *et al.*[11]: 1.73 (1.68–1.78), Wunnemann *et al.*[93]: 1.69 (1.58-1.81) Tcheandjieu *et al.*[39]: 1.38 (1.36-1.39)	Inouye *et al.*[2] (source): 1.71 (1.68-1.73) Sun, L *et al.*[82]: 1.56 (1.51-1.62), Dikilitas *et al.*[11]: 1.53 (1.46–1.60),	Admixed-American: Dikilitas *et al.*[11]: 1.93 (1.67–2.22) Tcheandjieu *et al.*[39]: 1.39(1.34-1.43) African: Dikilitas *et al.*[11]: 1.40 (1.30–1.52), Tcheandjieu *et al.*[39]: 1.12(1.1-1.14) Middle eastern: Saad *et al.*[7]: 1.54 (1.43–1.66)	Admixed-American: Dikilitas *et al.*[11]: 1.53 (1.23–1.90), African: Dikilitas *et al.*[11]: 1.27 (1.13–1.43)
Tada *et al.*[64]	Dikilitas *et al.*[11]: 1.28 (1.25–1.32),	Tada *et al.* I[64] (source): 1.23(1.18-1.28) Dikilitas *et al.*[11]: 1.20 (1.15–1.25)	Admixed-American: Dikilitas *et al.*[11]: 1.20 (1.06–1.35) African: Dikilitas *et al.*[11]: 1.05 (0.98–1.14)	Admixed-American: Dikilitas *et al.*[11] : 1.13 (0.93–1.36) African: Dikilitas *et al.*[11] :1.05 (0.94–1.17)

OR: Odds Ratio, RR: Risk Ratio, HR: Hazard Ratio, CI: Confidence interval, PRS: Polygenic Risk Score, SD: Standard Deviation

Table 2.4: CHD polygenic risk scores selected for further evaluation

Authors / PRS ID in PGS catalogue/Ancestry	PRS characteristics	CHD odds or hazard ratio (95% CI)*	AUC (95%CI)
Tada *et al.*[64] /PGS000011/European	A PRS constructed using 50 SNPs and European GWAS. **GWAS:** European dominant (~95%) with a small proportion of Asian samples **Evaluation sample:** 100% European	HR per SD: 1.23 (1.18-1.28)	0.75 (no CI reported)
Oni-orisan *et al.*[6] /PGS004595/European	A PRS constructed using European GWAS and 164 genome-wide significant CHD risk SNPs from UKB and CC4D GWAS. It was evaluated among datasets that contains admixed-American participants (n=1538). **GWAS:** European dominant **Evaluation sample:** Multi-ancestry	HR: 1.87 (1.19-2.95)	NR
Koyama *et al.*[5] /PGS000337/Japanese and European	A multi-ancestry PRS that used Japanese and European genetic information during construction. **GWAS:** European and Japanese dominant **Training sample:** Japanese dominant **Evaluation sample:** Japanese dominant	OR per SD: 1.84 (1.74-1.94)	0.67 (0.66–0.69)
Tcheandjieu *et al.*[39] /PGS003446/Multi-ancestry	A multi-ancestry PRS that used admixed-American, European and Japanese genetic information during construction using pruning and thresholding. **GWAS:** Multi-ancestry (Admixed-Americans, African, Japanese, European) **Training sample:** European, African and admixed-Americans **Evaluation sample:** Admixed-American	OR per SD: 1.43 (1.27-1.61)	NR
Tamlander *et al.*[3] / PGS001780/European	A PRS developed using a novel method, PRS-CS-auto. **GWAS:** European dominant (~85%) with a small proportion of Asian samples **Evaluation sample:** European	OR per SD (in UK Biobank): 1.72 (1.70-1.75)	0.79 (0.79–0.80)
Patel *et al.*[57] /PGS003725/Multi-ancestry	A multi-ancestry and multi-trait PRS derived using Admixed-American, European, African and Asian genetic information. **GWAS:** Multi-ancestry (admixed-Americans, African, Japanese, European) **Training sample:** European **Evaluation sample:** Admixed-American	OR per SD: 1.61 (1.53-1.70)	NR
Inouye *et al.*[2] /PGS000018/European	A meta-score based on the weighted average of 3 PRSs. **GWAS:** European dominant (~90%) with a small proportion of Asian samples **Training sample:** European dominant **Evaluation sample**: European dominant	HR per SD: 1.71 (1.68-1.73)	0.79 (no CI reported)
Khera *et al.*[1] /PGS000013/European	A PRS developed using LDpred2[102] with a large European dominant GWAS input. **GWAS:** European dominant (~85%) with a small proportion of Asian samples **Training sample:** European dominant **Evaluation sample:** European dominant	OR per SD (in training sample): 1.72 (1.67-1.78)	0.81 (0.81-0.81)



1. Date of literature search: 24th Nov 2022

2. Sample size for the primary cohort that used to derive the polygenic risk score (mentioned as training set in some papers).

3. This category of exclusion includes: Author reply, letter, Review, Methodology paper, case report, qualitative study, cost-effectiveness study

CVD: Cardiovascular disease, CHD: Coronary Artery Disease, PRS: Polygenic Risk Score

Figure 2.1: Flowchart of literature search and identification

2.7 References

1. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. *Nat Genet* 2018;50(9):pp. 1219–1224.
2. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. “Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention”. *J Am Coll Cardiol* 2018;72(16):pp. 1883–1893.
3. Tamlander M, Mars N, Pirinen M, FinnGen, Widen E, and Ripatti S. “Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes”. *Communications Biology* 2022;5(1):p. 158.
4. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. “Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 636–645.
5. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, et al. “Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease”. *Nature Genetics* 2020;52(11):pp. 1169–1177.
6. Oni-Orisan A, Haldar T, Cayabyab MAS, Ranatunga DK, Hoffmann TJ, Iribarren C, et al. “Polygenic Risk Score and Statin Relative Risk Reduction for Primary Prevention of Myocardial Infarction in a Real-World Population”. *Clinical Pharmacology & Therapeutics* 2022;112(5):pp. 1070–1078.
7. Saad M, El-Menyar A, Kunji K, Ullah E, Al Suwaidi J, and Kullo IJ. “Validation of Polygenic Risk Scores for Coronary Heart Disease in a Middle Eastern Cohort Using Whole Genome Sequencing”. *Circulation* 2022;Genomic and precision medicine.e003712.
8. Wang M, Menon R, Mishra S, Patel AP, Chaffin M, Tanneeru D, et al. “Validation of a Genome-Wide Polygenic Score for Coronary Artery Disease in South Asians”. *J Am Coll Cardiol* 2020;76(6):pp. 703–714.
9. Huang QQ, Sallah N, Dunca D, Trivedi B, Hunt KA, Hodgson S, et al. “Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals”. *Nature communications* 2022;13(1):p. 4664.
10. Tragante V, Doevendans PA, Nathoe HM, Graaf Y van der, Spiering W, Algra A, et al. “The impact of susceptibility loci for coronary artery disease on other vascular domains and recurrence risk”. *Eur Heart J* 2013;34(37):pp. 2896–904.
11. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, et al. “Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups”. *American Journal of Human Genetics* 2020;106(5):pp. 707–716.
12. Aragam KG, Dobbyn A, Judy R, Chaffin M, Chaudhary K, Hindy G, et al. “Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease”. *Journal of the American College of Cardiology* 2020;75(22):pp. 2769–2780.
13. Iribarren C, Lu M, Jorgenson E, Martínez M, Lluis-Ganella C, Subirana I, et al. “Weighted Multi-marker Genetic Risk Scores for Incident Coronary Heart Disease among Individuals of African, Latino and East-Asian Ancestry”. *Sci Rep* 2018;8(1):p. 6853.
14. Hasbani NR, Ligthart S, Brown MR, Heath AS, Bebo A, Ashley KE, et al. “American Heart Association’s Life’s Simple 7: Lifestyle Recommendations, Polygenic Risk, and Lifetime Risk of Coronary Heart Disease”. *Circulation* 2022;145(11):pp. 808–818.
15. Lu X, Liu Z, Cui Q, Liu F, Li J, Niu X, et al. “A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study”. *European Heart Journal* 2022;43(18):pp. 1702–1711.
16. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. “The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource”. *Nucleic Acids Research* 2022;51(D1):pp. D977–D985.
17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”. *BMJ* 2021;372:n71.

18. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner K, et al. “FinnGen: Unique genetic insights from combining isolated population and national health register data”. *medRxiv* 2022:p. 2022.03.03.22271360.

19. van der Harst P and Verweij N. “Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease”. *Circ Res* 2018;122(3):pp. 433–443.

20. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, et al. “Association analyses based on false discovery rate implicate new loci for coronary artery disease”. *Nature Genetics* 2017;49(9):pp. 1385–1391.

21. Webb TR, Erdmann J, Stirrups KE, Stitziel NO, Masca NG, Jansen H, et al. “Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease”. *J Am Coll Cardiol* 2017;69(7):pp. 823–836.

22. Klarin D, Zhu QM, Emdin CA, Chaffin M, Horner S, McMillan BJ, et al. “Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease”. *Nature Genetics* 2017;49(9):pp. 1392–1397.

23. Stitziel NO, Stirrups KE, Masca NG, Erdmann J, Ferrario PG, König IR, et al. “Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease”. *N Engl J Med* 2016;374(12):pp. 1134–44.

24. Nikpay M, Goel A, Won H.-H, Hall LM, Willenborg C, Kanoni S, et al. “A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease”. *Nature genetics* 2015;47(10):pp. 1121–1130.

25. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. “Large-scale association analysis identifies new risk loci for coronary artery disease”. *Nature Genetics* 2013;45(1):pp. 25–33.

26. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. “Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease”. *Nat Genet* 2011;43(4):pp. 333–8.

27. Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, Hall AS, et al. “New susceptibility locus for coronary artery disease on chromosome 3q22.3”. *Nat Genet* 2009;41(3):pp. 280–2.

28. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, et al. “Genetic variants associated with Lp(a) lipoprotein level and coronary disease”. *N Engl J Med* 2009;361(26):pp. 2518–28.

29. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, et al. “Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants”. *Nat Genet* 2009;41(3):pp. 334–41.

30. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. “A common variant on chromosome 9p21 affects the risk of myocardial infarction”. *Science* 2007;316(5830):pp. 1491–3.

31. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. “Genomewide Association Analysis of Coronary Artery Disease”. *New England Journal of Medicine* 2007;357(5):pp. 443–453.

32. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. “A common allele on chromosome 9 associated with coronary heart disease”. *Science* 2007;316(5830):pp. 1488–91.

33. Shiffman D, Rowland CM, Louie JZ, Luke MM, Bare LA, Bolonick JI, et al. “Gene variants of VAMP8 and HNRPUL1 are associated with early-onset myocardial infarction”. *Arterioscler Thromb Vasc Biol* 2006;26(7):pp. 1613–8.

34. Coronary Artery Disease (C4D) Genetics Consortium. “A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease”. *Nat Genet* 2011;43(4):pp. 339–44.

35. Gudbjartsson DF, Bjornsdottir US, Halapi E, Helgadottir A, Sulem P, Jonsdottir GM, et al. “Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction”. *Nat Genet* 2009;41(3):pp. 342–7.

36. Aoki A, Ozaki K, Sato H, Takahashi A, Kubo M, Sakata Y, et al. “SNPs on chromosome 5p15.3 associated with myocardial infarction in Japanese population”. *J Hum Genet* 2011;56(1):pp. 47–51.

37. Wang F, Xu CQ, He Q, Cai JP, Li XC, Wang D, et al. “Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population”. *Nat Genet* 2011;43(4):pp. 345–9.

38. Ishigaki K, Akiyama M, Kanai M, Takahashi A, Kawakami E, Sugishita H, et al. “Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases”. *Nature Genetics* 2020;52(7):pp. 669–679.

39. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. “Large-scale genome-wide association study of coronary artery disease in genetically diverse populations”. *Nature Medicine* 2022;28(8):pp. 1679–1692.

40. Howson JMM, Zhao W, Barnes DR, Ho W.-K, Young R, Paul DS, et al. “Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms”. *Nature Genetics* 2017;49(7):pp. 1113–1119.

41. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. “A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses”. *Lancet* 2010;376(9750):pp. 1393–400.

42. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. “Million Veteran Program: A mega-biobank to study genetic influences on health and disease”. *J Clin Epidemiol* 2016;70:pp. 214–23.

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

44. Ntalla I, Kanoni S, Zeng L, Giannakopoulou O, Danesh J, Watkins H, et al. “Genetic Risk Score for Coronary Disease Identifies Predispositions to Cardiovascular and Noncardiovascular Diseases”. *Journal of the American College of Cardiology* 2019;73(23):pp. 2932–2942.

45. Patel RS, Denaxas S, Howe LJ, Eggo RM, Shah AD, Allen NE, et al. “Reproducible disease phenotyping at scale: Example of coronary artery disease in UK Biobank”. *PLoS One* 2022;17(4):e0264828.

46. Howe LJ, Dudbridge F, Schmidt AF, Finan C, Denaxas S, Asselbergs FW, et al. “Polygenic risk scores for coronary artery disease and subsequent event risk amongst established cases”. *Human Molecular Genetics* 2020;29(8):pp. 1388–1395.

47. Yuan S, Huang X, Ma W, Yang R, Xu F, Han D, et al. “Associations of HDL-C/LDL-C with myocardial infarction, all-cause mortality, haemorrhagic stroke and ischaemic stroke: A longitudinal study based on 384 093 participants from the UK Biobank”. *Stroke and Vascular Neurology* 2022;(no pagination).

48. Fan M, Sun D, Zhou T, Heianza Y, Lv J, Li L, et al. “Sleep patterns, genetic susceptibility, and incident cardiovascular disease: a prospective study of 385 292 UK biobank participants”. *Eur Heart J* 2020;41(11):pp. 1182–1189.

49. Zhang H, Zeng Y, Yang H, Hu Y, Chen W, Ying Z, et al. “Familial factors, diet, and risk of cardiovascular disease: a cohort analysis of the UK Biobank”. *The American journal of clinical nutrition* 2021;114(5):pp. 1837–1846.

50. Huang Y, Hui Q, Gwinn M, Hu YJ, Quyyumi AA, Vaccarino V, et al. “Sexual Differences in Genetic Predisposition of Coronary Artery Disease”. *Circulation. Genomic and Precision Medicine* 2021;14(1):e003147.

51. Carter AR, Harrison S, Gill D, Davey Smith G, Taylor AE, Howe LD, et al. “Educational attainment as a modifier for the effect of polygenic scores for cardiovascular risk factors: Cross-sectional and prospective analysis of UK Biobank”. *International Journal of Epidemiology* 2022;51(3):pp. 885–897.

52. Daghlas I, Dashti HS, Lane J, Aragam KG, Rutter MK, Saxena R, et al. “Sleep Duration and Myocardial Infarction”. *Journal of the American College of Cardiology* 2019;74(10):pp. 1304–1314.

53. Huang Y, Hui Q, Gwinn M, Hu YJ, Quyyumi AA, Vaccarino V, et al. “Interaction between genetics and smoking in determining risk of coronary artery diseases”. *Genetic Epidemiology* 2022;46(3-4):pp. 199–212.

54. Heianza Y, Zhou T, Sun D, Hu FB, Manson JE, and Qi L. “Genetic susceptibility, plant-based dietary patterns, and risk of cardiovascular disease”. *American Journal of Clinical Nutrition* 2020;112(1):pp. 220–228.

55. Verweij N, Eppinga RN, Hagemeijer Y, and van der Harst P. “Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure”. *Scientific Reports* 2017;7(1):p. 2761.

56. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, et al. “Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers”. *Nature Medicine* 2020;26(4):pp. 549–557.

57. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. *Nature Medicine* 2023;29(7):pp. 1793–1803.

58. Ye Y, Chen X, Han J, Jiang W, Natarajan P, and Zhao H. “Interactions Between Enhanced Polygenic Risk Scores and Lifestyle for Cardiovascular Disease, Diabetes, and Lipid Levels”. *Circulation. Genomic and Precision Medicine* 2021;14(1):e003128.

59. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, et al. “Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction”. *Circ Genom Precis Med* 2021;14(2):e003304.

60. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. “Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease”. *New England Journal of Medicine* 2016;375(24):pp. 2349–2358.

61. Iribarren C, Lu M, Jorgenson E, Martinez M, Lluis-Ganella C, Subirana I, et al. “Clinical Utility of Multimarker Genetic Risk Scores for Prediction of Incident Coronary Heart Disease: A Cohort Study among over 51 Thousand Individuals of European Ancestry”. *Circulation: Cardiovascular Genetics* 2016;9(6):pp. 531–540.

62. Mega JL, Stitziel NO, Smith JG, Chasman DI, Caulfield MJ, Devlin JJ, 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* 2015;385(9984):pp. 2264–2271.

63. Tikkanen E, Havulinna AS, Palotie A, Salomaa V, and Ripatti S. “Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease”. *Arteriosclerosis, Thrombosis and Vascular Biology* 2013;33(9):pp. 2261–2266.

64. Tada H, Melander O, Louie JZ, Catanese JJ, Rowland CM, Devlin JJ, et al. “Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history”. *Eur Heart J* 2016;37(6):pp. 561–7.

65. King A, Wu L, Deng HW, Shen H, and Wu C. “Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease”. *BMC Medicine* 2022;20(1):p. 385.

66. Di Narzo A, Frades I, Crane HM, Crane PK, Hulot JS, Kasarskis A, et al. “Meta-analysis of sample-level dbGaP data reveals novel shared genetic link between body height and Crohn’s disease”. *Human Genetics* 2021;140(6):pp. 865–877.

67. Gola D, Erdmann J, Läll K, Mägi R, Müller-Myhsok B, Schunkert H, et al. “Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease”. *Circ Genom Precis Med* 2020;13(6):e002932.

68. Gola D, Erdmann J, Müller-Myhsok B, Schunkert H, and König IR. “Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status”. *Genet Epidemiol* 2020;44(2):pp. 125–138.

69. Morris RW, Cooper JA, Shah T, Wong A, Drenos F, Engmann J, et al. “Marginal role for 53 common genetic variants in cardiovascular disease prediction”. *Heart* 2016;102(20):pp. 1640–7.

70. Brautbar A, Pompeii LA, Dehghan A, Ngwa JS, Nambi V, Virani SS, et al. “A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies”. *Atherosclerosis* 2012;223(2):pp. 421–426.

71. Goldstein BA, Knowles JW, Salfati E, Ioannidis JPA, and Assimes TL. “Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: Coronary heart disease as an example”. *Frontiers in Genetics* 2014;5(AUG) (no pagination).

72. Goldstein BA, Yang L, Salfati E, and Assimes TL. “Contemporary Considerations for Constructing a Genetic Risk Score: An Empirical Approach”. *Genetic Epidemiology* 2015;39(6):pp. 439–45.

73. Powell KL, Stephens SR, and Stephens AS. “Cardiovascular risk factor mediation of the effects of education and Genetic Risk Score on cardiovascular disease: a prospective observational cohort study of the Framingham Heart Study”. *BMJ Open* 2021;11(1):e045210.

74. De Vries PS, Kavousi M, Ligthart S, Uitterlinden AG, Hofman A, Franco OH, et al. “Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: The Rotterdam Study”. *International Journal of Epidemiology* 2015;44(2):pp. 682–688.

75. Abraham G, Havulinna AS, Bhalala OG, Byars SG, De Livera AM, Yetukuri L, et al. “Genomic prediction of coronary heart disease”. *Eur Heart J* 2016;37(43):pp. 3267–3278.

76. Ge T, Chen CY, Ni Y, Feng YA, and Smoller JW. “Polygenic prediction via Bayesian regression and continuous shrinkage priors”. *Nature communications* 2019;10(1):p. 1776.

77. Manikpurage HD, Eslami A, Perrot N, Li Z, Couture C, Mathieu P, et al. “Polygenic Risk Score for Coronary Artery Disease Improves the Prediction of Early-Onset Myocardial Infarction and Mortality in Men”. *Circ Genom Precis Med* 2021;14(6):e003452.

78. Goodman MO, Cade BE, Shah NA, Huang T, Dashti HS, Saxena R, et al. “Pathway-Specific Polygenic Risk Scores Identify Obstructive Sleep Apnea-Related Pathways Differentially Moderating Genetic Susceptibility to Coronary Artery Disease”. *Circulation. Genomic and Precision Medicine* 2022;15(5):e003535.

79. Isgut M, Sun J, Quyyumi AA, and Gibson G. “Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later”. *Genome Medicine* 2021;13(1):p. 13.

80. Zaccardi F, Timmins IR, Goldney J, Dudbridge F, Dempsey PC, Davies MJ, et al. “Self-reported walking pace, polygenic risk scores and risk of coronary artery disease in UK biobank”. *Nutrition, Metabolism and Cardiovascular Diseases* 2022;32(11):pp. 2630–2637.

81. Kim Y, Yeung SLA, Sharp SJ, Wang M, Jang H, Luo S, et al. “Genetic susceptibility, screen-based sedentary activities and incidence of coronary heart disease”. *BMC Medicine* 2022;20(1):p. 188.

82. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, et al. “Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses”. *PLoS Medicine* 2021;18(1) (no pagination).

83. Ramírez J, Duijvenboden S van, Young WJ, Tinker A, Lambiase PD, Orini M, et al. “Prediction of Coronary Artery Disease and Major Adverse Cardiovascular Events Using Clinical and Genetic Risk Scores for Cardiovascular Risk Factors”. *Circ Genom Precis Med* 2022;15(5):e003441.

84. Livingstone KM, Abbott G, Bowe SJ, Ward J, Milte C, and McNaughton SA. “Diet quality indices, genetic risk and risk of cardiovascular disease and mortality: a longitudinal analysis of 77 004 UK Biobank participants”. *BMJ Open* 2021;11(4):e045362.

85. Livingstone KM, Abbott G, Ward J, and Bowe SJ. “Unhealthy Lifestyle, Genetics and Risk of Cardiovascular Disease and Mortality in 76,958 Individuals from the UK Biobank Cohort Study”. *Nutrients* 2021;13(12).

86. Rostami S, Hoff M, Dalen H, Hveem K, and Videm V. “Genetic risk score associations for myocardial infarction are comparable in persons with and without rheumatoid arthritis: the population-based HUNT study”. *Scientific reports* 2020;10(1):p. 20416.

87. Liang F, Liu F, Li J, Huang K, Yang X, Chen S, et al. “Genetic risk modifies the effect of long-term fine particulate matter exposure on coronary artery disease”. *Environment International* 2022;170 (no pagination).

88. Hindy G, Aragam KG, Ng K, Chaffin M, Lotta LA, Baras A, et al. “Genome-Wide Polygenic Score, Clinical Risk Factors, and Long-Term Trajectories of Coronary Artery Disease”. *Arteriosclerosis, Thrombosis and Vascular Biology* 2020;40(11):pp. 2738–2746.

89. Hindy G, Wiberg F, Almgren P, Melander O, and Orho-Melander M. “Polygenic Risk Score for Coronary Heart Disease Modifies the Elevated Risk by Cigarette Smoking for Disease Incidence”. *Circulation: Genomic and Precision Medicine* 2018;11(1):E001856.

90. Martikainen P, Korhonen K, Jelenkovic A, Lahtinen H, Havulinna A, Ripatti S, et al. “Joint association between education and polygenic risk score for incident coronary heart disease events: a longitudinal population-based study of 26 203 men and women”. *Journal of epidemiology and community health.* 2021;06.

91. Fritz J, Shiffman D, Melander O, Tada H, and Ulmer H. “Metabolic Mediators of the Effects of Family History and Genetic Risk Score on Coronary Heart Disease-Findings From the Malmo Diet and Cancer Study”. *Journal of the American Heart Association* 2017;6(3):p. 20.

92. Svensson T, Kitlinski M, Engström G, and Melander O. “A genetic risk score for CAD, psychological stress, and their interaction as predictors of CAD, fatal MI, non-fatal MI and cardiovascular death”. *PLoS One* 2017;12(4):e0176029.

93. Wünnemann F, Sin Lo K, Langford-Avelar A, Busseuil D, Dubé MP, Tardif JC, et al. “Validation of Genome-Wide Polygenic Risk Scores for Coronary Artery Disease in French Canadians”. *Circ Genom Precis Med* 2019;12(6):e002481.

94. Menniti G, Paquet C, Han HY, Dube L, and Nielsen DE. “Multiscale Risk Factors of Cardiovascular Disease: CLSA Analysis of Genetic and Psychosocial Factors”. *Frontiers in Cardiovascular Medicine* 2021;8:p. 599671.

95. Hamad R, Glymour MM, Calmasini C, Nguyen TT, Walter S, and Rehkopf DH. “Explaining the Variance in Cardiovascular Disease Risk Factors: A Comparison of Demographic, Socioeconomic, and Genetic Predictors”. *Epidemiology* 2022;33(1):pp. 25–33.

96. Mosley JD, Gupta DK, Tan J, Yao J, Wells QS, Shaffer CM, et al. “Predictive Accuracy of a Polygenic Risk Score Compared with a Clinical Risk Score for Incident Coronary Heart Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 627–635.

97. Khan SS, Page C, Wojdyla DM, Schwartz YY, Greenland P, and Pencina MJ. “Predictive Utility of a Validated Polygenic Risk Score for Long-Term Risk of Coronary Heart Disease in Young and Middle-Aged Adults”. *Circulation* 2022;146(8):pp. 587–596.

98. Privé F, Vilhjálmsson BJ, Aschard H, and Blum MGB. “Making the Most of Clumping and Thresholding for Polygenic Scores”. *Am J Hum Genet* 2019;105(6):pp. 1213–1221.

99. Mak TSH, Porsch RM, Choi SW, Zhou X, and Sham PC. “Polygenic scores via penalized regression on summary statistics”. *Genetic Epidemiology* 2017;41(6):pp. 469–480.

100. Tibshirani R. “Regression shrinkage and selection via the lasso: a retrospective”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011;73(3):pp. 273–282.

101. Ruan Y, Lin Y.-F, Feng Y.-CA, Chen C.-Y, Lam M, Guo Z, et al. “Improving polygenic prediction in ancestrally diverse populations”. *Nature Genetics* 2022;54(5):pp. 573–580.

102. Privé F, Arbel J, and Vilhjálmsson BJ. “LDpred2: better, faster, stronger”. *Bioinformatics* 2021;36(22-23):pp. 5424–5431.

103. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores”. *The American Journal of Human Genetics* 2015;97(4):pp. 576–592.

104. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. “The UK Biobank resource with deep phenotyping and genomic data”. *Nature* 2018;562(7726):pp. 203–209.

Chapter 3: Methods

Summary

This chapter describes the study design of the Mexico City Prospective Study (MCPS), including details on participant recruitment, baseline assessment, genotyping, and the analysis of genetic data for investigating polygenic risk scores (PRSs) for coronary heart disease (CHD) risk within the cohort. The primary and secondary outcome definitions for CHD and the construction of PRSs, as well as the general epidemiological and statistical methods used for analyses throughout this thesis, are described. More detailed descriptions of the specific analyses used for the evaluation of pre-existing PRSs will be provided in **Chapter 4**. The derivation and validation of an MCPS-informed PRS for CHD will be detailed in **Chapter 5** and the assessment of the predictive ability of integrating a CHD PRS with a clinical risk score will be presented in **Chapter 6**.

3.1 The Mexico City Prospective Study

3.1.1 Design and recruitment

The dataset used for the main analyses in this thesis is from the Mexico City Prospective Study (MCPS)[1,2]. The study was established in the early 1990s as a collaboration between the Mexican Ministry of Health (Secretaria de Salud SSA) and the Clinical Trial Service Unit and Epidemiological Studies Unit at the University of Oxford, aiming to investigate the major determinants of morbidity and premature mortality in Mexican adults. The study participants were recruited from two neighbouring districts (Coyoacán and Iztapalapa) in central and north-east Mexico City. The two selected districts contain a diverse population, with a mix of long-term residents and recent migrants from other parts of Mexico (**Figure 3.1**). From 1995 to 1997, household information from these two districts was compiled through census-style door-to-door interviews, to collect basic information on household members (e.g., age, name and address) and to identify potentially eligible participants. Between 1998 and 2004, 112,333 eligible households were identified and adults aged 35 or older were systematically invited to join the study. Of those, 106,069 (95%) of the households consented and 159,755 individual participants were successfully recruited into the study. Of the recruited participants, two-thirds were women (as women were more likely to be at home during the working hours when the home-based survey was largely conducted).

3.1.2 Baseline assessment procedure

During recruitment, each participant answered an interviewer-administered standardised questionnaire that collected information on lifestyle characteristics, socio-economic status, medical histories and medication use (**Table 3.1**), and had physical measurements (including height, weight and blood pressure) taken. Nearly all participants provided a 10-ml non-fasting venous blood sample, which was stored long term at -150°C in accredited laboratory facilities.

The information collected was entered directly into a hand-held data recorder either via a touch sensitive screen (SPT 500 personal digital assistant with palm operating system [Symbol], used

from November 2001 to December 2004, or by a barcode and alphanumeric keypad combination [Microwand 32ES barcode optical reading device Handheld Products], used from April 1998 to November 2001). The data recorder was pre-programmed to query extreme or implausible measurements, prompting the fieldworkers to check such values and preventing them from being recorded. After the data was collected, further possible data errors were identified by a database checking programme and reviewed by the Data Monitoring Team.

3.1.3 Physical measurements

Anthropometric parameters (height, weight, waist and hip circumference) were measured by trained nurses. Standing height without wearing shoes was measured (to the nearest millimetre) with a wooden triangle and a non-stretchable tape. Waist and hip circumferences were also measured (to the nearest millimetre) using a non-stretchable tape with the participant wearing light clothing. Waist circumference was measured at the midpoint between the iliac crest and the costal margin and hip circumference was measured around the femoral epicondyles. Weight was measured using portable analogue scales between 1998-2003 (to the nearest kg) and portable electronic scales between 2003-2004 (to the nearest 100g), and the equipment was recalibrated twice a week. Body mass index (BMI) was calculated as the weight in kilograms divided by the square of height in metres. Systolic and diastolic blood pressure in the sitting position (SBP, DBP) were measured using a standard mercury sphygmomanometer that was recalibrated once a month. The measurements were taken three times to the nearest mmHg, with three minutes between each measurement. **Table 3.2** shows key baseline characteristics of the MCPS participants.

3.1.4 Blood sample collection and processing

The blood sample was available for 155,487 (98%) participants. For each participant, a 10-ml non-fasting venous blood sample was collected into a single EDTA vacutainer with a barcode label that was unique to each participant. The blood samples were first stored in an insulated box

containing several chilled packs maintaining a temperature around 4-10ºC immediately after collection. Subsequently, the samples were transferred to a central laboratory in Mexico City, where they were refrigerated overnight at 4°C. The next morning, the samples were centrifuged (2100g at 4°C for 15 minutes) and separated into two plasma aliquots and one buffy coat aliquot, both stored in 1.8 ml cryovials (Nunc A/S, Denmark). Each of the three cryovials was subsequently stored separately in three boxes and then placed in three separate freezers for storage at -80°C before shipping to Oxford in insulated boxes with dry ice for long-term storage. Upon arrival at Oxford, the samples were placed into tanks with liquid nitrogen vapor for long-term storage at -150°C.

3.1.5 Follow-up for mortality

Participants are being followed up indefinitely for cause-specific mortality through probabilistic linkage (based on name, including phonetic coding of names, age, and sex) to the Mexican System for Epidemiologic Death Statistics (Subsistema Epidemiológico y Estadístico de Defunciones or SEED) electronic death registry in Mexico City, administered by the Ministry of Health. Field validation of more than 7,000 matched deaths confirmed the reliability of the matching algorithm in over 95% of cases. By September 30, 2022, 34,079 medically certified deaths were linked to MCPS participants. Death registration in Mexico City is reliable and complete, with the causes of almost all deaths certified medically[3]. Diseases recorded on the death certificates are routinely coded using the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10), with subsequent review by the MCPS study clinicians (blinded to any baseline information) to recode, when necessary, the underlying cause of death[4].

3.1.6 Genotyping, quality control and imputation

3.1.6.1 DNA extraction

From the buffy coat biological samples collected, the DNA was extracted at the UK Biocentre using Perkin Elmer Chemagic 360 systems and suspended in Tris-EDTA (TE) buffer. To determine

the yield and quality of the DNA, ultraviolet-visible (UV-VIS) spectroscopy (Trinean DropSense96) was used to check the normalised samples, and samples were deemed uncontaminated if 2 µg DNA at 20 ng μl^{-1} concentration had a 260:280 nm ratio of >1.8 and a 260:230 nm ratio of 2.0-2.2. Of the participants with a blood sample available, 146,068 (94%) had DNA successfully extracted and transferred to the Regeneron Genetics Centre (RGC) for genotyping, exome sequencing and whole genome sequencing (on a subset of 10,008 samples). Only the genotype data (or imputed-genotype data: see below) was used for analysis in the thesis.

3.1.6.2 Sample and genotyping quality control (QC)

The blood samples were genotyped at RGC sequencing lab using the Illumina Global Screening Array (GSA) v2 beadchip for 650,380 polymorphic variants and were mapped to genome build GRCh38 (hg38)[2]. Of the 146,068 samples available for genotyping, 145,266 (99.5%) were successfully processed and 802 (0.5%) samples did not meet the technical requirements for processing. The average genotype call rate per sample was 98.4%, and 98.4% of samples had a call rate over 90%. Of all samples that were processed, 4,435 (3.1%) were excluded as they failed one or more of the initial RGC QC metrics. Specifically, 1,827 (1.3%) samples were discordant between the participant-reported sex and the genetically determined sex, 2,276 (1.6%) had a sample call rate below 90%, 44 (0.3%) had genotyped variants different from their exome data, 1,063 (0.7%) were duplicated samples, 268 (0.2%) had uncertain linkage to the donor-participants and 6 (0.004%) encountered instrument issues during DNA extraction.

A total of 140,831 samples with 650,380 variants passed the initial RGC QC and were sent back to Oxford from RGC, from which they underwent Oxford QC procedure on variants, using a modified QC workflow as described in Ziyatdinov *et al.*[2]. The Oxford QC procedure consisted of consistency checks of genotypes in sex chromosomes (steps 1-4), filtering based on variant-level and sample-level missingness using the threshold of 0.10 (steps 5-6), adherence to HWE for a subset of participants that were unrelated to the 3rd degree (step 7), identifying Mendel errors of

genotypes in nuclear families and setting them to missing (step 8) and repeating steps 5-7 (steps 9-11)[2]. After the Oxford QC, 140,829 samples with 558,008 variants (539,448 autosomal and 18,560 X variants) were retained.

3.1.6.3 Genotype Imputation

The RGC QCed genotyped data (see **Section 3.1.6.2**) of the 140,831 samples underwent subsequent imputation for additional variants using the Trans-Omics for Precision Medicine (TOPMed) version R2 in GRCh38 imputation server[5]. TOPMed has a diverse reference panel, which includes 47,159 Europeans, 24,267 African and 17,085 admixed American genomes. This process resulted in an enhanced repertoire of over 308 million variants. After applying an INFO score > 0.3 threshold, 9,400,000 variants ultimately retained.

3.1.6.4 Relatedness mapping

The MCPS has a study population with high relatedness due to the fact that all the eligible participants from the same household that were selected were invited to take part in the study. The genetic relatedness of the study participants (with genetic data that passed Oxford QC) was calculated based on kinship coefficients and probability of zero identity by descent (IBD) sharing, using the Kinship-based Inference for GWAS (KING) software[6]. First-, second-, and third-degree relatedness (e.g., siblings, half-siblings, and first cousins) correspond to expected kinship coefficients of 0.25, 0.125, and 0.0625, respectively. After these calculations, 40,695 participants were unrelated to any other participants (to the 3rd degree). Then an independent vertex set (IVS) analysis was performed on the remaining 100,134 participants (that are related to at least one other participant) to retain a subset of unrelated samples[2]. The IVS-selected participants and unrelated participants together formed a maximally unrelated set of 79,612 (57%) participants unrelated to the third degree (heron referred to as “unrelated set”).

3.1.6.5 Genetic principal component analysis (PCA)

PCA analysis was performed on the genotyping array data of the 40,695 unrelated individuals (up to the 3rd degree) that passed Oxford QC to identify population structure in MCPS, following the workflow implemented in bigsnpr[7]. Genetic variants with minor allele frequency (MAF) <0.01 were first removed, then the samples were clumped based on linkage disequilibrium (LD) r^2 (pairwise correlation between variants) threshold of 0.2, and then any long-range LD regions were removed. Outliers of these unrelated samples were detected using a procedure based on machine learning algorithms K-nearest neighbours and removed. PCA was performed on the remaining samples, and PC scores and loadings for the first 20 PCs were estimated using truncated singular value decomposition (SVD) of the scaled genotype matrix. These estimated PC scores and loadings were then projected onto the withheld 100,134 samples (related participants and outliers that were removed earlier).

3.1.6.6 Mapping of genetic ancestry

Due to a history of European colonisation, the contemporary Mexican population has a complex pattern of ancestral admixture. To characterise specific proportions of genomic segments inherited from inferred ancestral populations among MCPS participants (hereon referred to as "global ancestry"), the software *Admixture*[8] was used. First, a set of reference samples containing genetic information for 3,964 individuals representing European, East Asian, African and American ancestries, was prepared using samples from the 1000 Genomes Project[9], the Human Genome Diversity Project (HGDP)[10], the Metabolic Analysis of an Indigenous Sample (MAIS) study[11], and the MCPS (wherein 1,000 unrelated samples were randomly selected and included in the reference set). The reference samples were input into the Admixture software and five-fold cross validation was applied to determine the optimal number of inferred ancestral populations and per-individual ancestry proportions for each reference sample. The remaining MCPS samples not used as the reference samples were then projected into the Admixture model, with model

parameters for population-specific allele frequencies fixed to obtain their ancestry proportions.

3.1.7 Funding and ethics approvals

The MCPS has received support from the Mexican Ministry of Health, the National Council of Science and Technology for Mexico, the Wellcome Trust [058299/Z/99], Cancer Research UK, the British Heart Foundation [RE/13/1/30181], Kidney Research UK [MR/R007764/1], and the UK Medical Research Council [MC_UU_00017/2, MR/Z504543/1]. The genotyping was funded through an academic partnership between the National Autonomous University of Mexico, the University of Oxford, Regeneron and AstraZeneca. The Wellcome Trust Core Award Grant Number [203141/Z/16/Z] and the NIHR Oxford BRC supported the computational aspects of this research. The funding sources had no role in the design, conduct or analyses in this thesis.

Ethics approvals were obtained from the Mexican National Council of Science and Technology, the Mexican Ministry of Health, the Ethics and Research Commissions of the Medicine Faculty at the National Autonomous University of Mexico (*Universidad Nacional Autónoma de México, UNAM*) and of the University of Oxford [135/2014]. The Mexican Ministry of Health approved the transport of the biological samples from Mexico City to the UK. The Medical Ethics Committee of UNAM [FMED/CEI/MHU/001/2020] granted approval for data sharing with Regeneron for genotyping and subsequent analysis.

All study participants provided written informed consent to participate in the baseline assessment, resurveys and to donate blood samples within the remits specified for research, including anonymised genetic research.

3.2 Epidemiological methods

3.2.1 Analysis cohort

The analyses in this thesis focused on fatal or non-fatal events from coronary heart disease (CHD) occurring before the age of 80 years. Based on the average life expectancy estimated

for people living in Mexico City[12], and the difficulty of correctly attributing causes of death at very old ages, the age of 80 years was deemed appropriate for the purposes of this project. Analyses therefore focus on MCPS participants aged 35-79 at baseline and CHD deaths before age 80 years.

3.2.2 Primary outcome definition

At baseline, the MCPS participants were asked to report specific chronic diseases that they had been diagnosed with by a medical professional prior to enrolment. Among the medical histories collected, the CHD-specific diseases included heart attack (i.e., myocardial infarction) and angina. Therefore, for the purposes of this thesis, "CHD" cases were defined as a composite outcome including either a self-reported history of heart attack or angina at the baseline assessment (for individuals aged <80 years at recruitment) or death from CHD before the age of 80 years (with the specific causes defined by the ICD-10 codes I20-I25) listed anywhere on the death certificate (i.e., underlying or contributory cause).

3.2.3 Other CHD-related outcomes for secondary analyses

Secondary outcomes included the primary CHD outcome excluding deaths where CHD was not the underlying (i.e., the primary) cause, restricting the outcome to non-fatal self-reported myocardial infarction, angina or both at baseline, and restricting it to the CHD death component (as listed anywhere on the death certificate and, separately, as listed as the underlying cause).

3.3 Polygenic risk scores for CHD

3.3.1 Polygenic risk scores computation

A polygenic risk score (PRS), representing the genetic liability for a specific disease, combines the genetic effects of multiple single nucleotide polymorphisms (SNPs) associated with a partic-

ular disease into one composite score, using the equation below:

$$PRS = \sum_{i=1}^{n} X_i \beta_i \tag{3.1}$$

For SNP_i of the *n* SNPs included in a PRS, X_i indicates its disease risk-increasing allele count (i.e., 0,1 or 2) and β_i indicates its weight. The overall estimated genetic score is then calculated as the sum of the products of each SNP-specific allele count that associated with the outcome, and its corresponding weight. These weights may represent SNP effect sizes derived from a genome-wide association study (GWAS), or may be estimated using more advanced algorithms that utilise GWAS summary statistics as input for the algorithms. To ensure fair comparison between PRSs with different numbers of SNPs included, all the PRSs used for analyses in this thesis were standardised to a mean of 0 and a standard deviation (SD) of 1 (i.e., representing an increase in genetic liability for a given outcome relative to the mean of each PRS considered).

In **Chapter 4**, previously-published PRSs for CHD are re-created for the MCPS participants using external weights, and the computation is conducted using a publicly-available pipeline developed by the PGS Catalog[13,14]. In **Chapter 5**, the novel PRS is trained using MCPS and external GWAS data, applying three methods: Pruning and Thresholding[15,16], LDpred2[17] and PRS-CSx[18], to determine the SNPs to retain and their corresponding weights for PRS construction. The computation of the PRSs was either done by the mentioned software or with PLINK-2.0[19]. The external and MCPS-informed PRSs computations will be explained in more detail in the relevant chapters. Throughout the thesis, PRSs were computed using the TOPMed imputed genotype data that passed RGC QC as detailed in **Section 3.1.6**.

3.3.2 Genome-wide association study

As shown in **Section 3.3.1**, Genome-wide association study(GWAS) data is an important input for PRS construction. GWAS analyses are typically adjusted for age, sex and, genetic principal components (PCs), to account for population structure. As described in **Section 3.1.6.4**, only half

of the study participants are unrelated up to the third degree based on IBD estimation. Due to the high relatedness and complex population structure present in the MCPS population, a standard GWAS involving a simultaneous logistic regression across all SNPs, may lead to spurious results. The REGENIE[20] software employs a machine-learning based whole-genome regression model that accounts for population structure and within-sample relatedness, making this alternative approach more suitable for performing GWAS in MCPS. The method will be described in more detail in **Chapter 5**.

3.4 Statistical methods for investigating the association between polygenic predisposition to CHD and sub-sequent risk for CHD

Aims two to four that will be addressed in the following chapters all focus on evaluating the performance of CHD PRSs for CHD prediction. While each aim explores a different aspect of risk score evaluation (i.e., assessing pre-existing PRSs, construction of a new PRS, integrating PRSs and clinical risk scores), the methodology used to assess risk score predictability follows a consistent overarching approach, which is described in this section.

3.4.1 Logistic regression models for the assessment of CHD risk

Throughout the thesis, conventional logistic regression models are used to first assess the shape and the strength of the association between each candidate PRS (external or newly derived) and the odds of CHD in the MCPS population. The shape of association is first assessed by categorising each PRS into quintile groups, with the lowest group as the reference. The overall strength of association is subsequently assessed with each PRS as a continuous variable (per one SD increase across the full range of each PRS). The primary models are adjusted for age at recruitment, sex and the first seven genetic PCs to account for population structure (i.e., 'partial adjustment'). Only the first seven PCs are used because the SNP loadings of the higher PCs started to capture long-range LDs, as indicated by the non-normally distributed SNP loadings[2]. Subsequent analyses additionally adjusted for highest level of educational attainment (university

or high school, middle school, elementary school, other), diabetes at baseline (self-reported diagnosis, HbA1c≥6.5% or anti-diabetes medication use), waist-to-hip ratio (WHR), systolic and diastolic blood pressure (SBP, DBP) and smoking status (never smoker, ex-smoker, and current smoker) ('full adjustment'). These factors are often considered to be confounders in traditional observational analyses but in the context of the association between a PRS and CHD they could plausibly be effect mediators. Adjustment was made for WHR rather than BMI as previous analyses in the cohort have demonstrated the former to be more strongly related to CHD mortality risk than the latter[21,22]. Adjustment for district was not performed as it was felt that it could neither be a confounder or a mediator of any association (and the adjustment for PCs already accounted for population structure) while adjustment for lipid measurements was not done because of incomplete data on these measurements at the time of the analyses. The area under the receiver-operating-characteristic curve (AUC) is used to provide an estimate of model discrimination.

3.4.2 Secondary and sensitivity analysis

Secondary analyses in this thesis include: separate analyses by covariates included in the models, and by ancestry proportion, to investigate potential effect modification by these characteristics. To assess analyses robustness, sensitivity analyses include different CHD outcome definitions (as described in **Section 3.2.3**) and analyses restricted to participants unrelated to the 3rd degree (as defined in **Section 3.1.6.4**).

Further details of specific secondary and sensitivity analyses performed will be described in later chapters.

3.4.3 Handling of missing data

For all the analyses in this thesis, participants are excluded if they had genetic data missing or if their genetic data did not pass the QC steps, with no exclusions. Primary analyses (partial adjustment) in this thesis follow complete case analysis procedure.

For secondary analyses with more adjustments, missing continuous variables are imputed using the median of the respective variable. Missing categorical variables are imputed using the highest selected category of the respective variable.

3.5 Conclusion

As introduced in **Chapter 1** and reviewed in **Chapter 2**, populations of admixed-American ancestry are highly under-represented in genetic studies. Research on CHD PRSs in this population could address the current research gap in ancestry diversity and be clinically useful to CHD management in this region. The characteristics of MCPS as described in this chapter make it well-placed for conducting CHD PRS research among admixed-American populations.

The following chapters will present the results for my DPhil aims two to four, specifically, the evaluation of transferability of previously-published CHD PRSs will be described in **Chapter 4**; the construction, training and testing of MCPS-informed CHD PRSs will be explained in **Chapter 5** and the assessment of integrated (genetic and clinical) risk scores will be detailed in **Chapter 6**. Methods relevant to a specific chapter will be described within that chapter.

Table 3.1: Key data collected in the MCPS

Socio-demographics
- Age
- Sex
- Area of residence
- Marital status
- Education
- Occupation
- Income
- Health service provider

Lifestyle characteristics
- Smoking
- Passive smoking
- Alcohol consumption
- Physical activity
- Sleep duration
- Fruit/vegetable intake
- Fried food intake
- Type of cooking oil used

Prior diseases and medication

Reproductive history
- Menopausal status
- Hysterectomy
- Oophorectomy
- Hormone replacement therapy
- Contraceptive use
- Age at first sexual relationship
- Age at first pregnancy
- Number of pregnancies

Physical measurements
- Height
- Weight
- Waist circumference
- Hip circumference
- Systolic blood pressure
- Diastolic blood pressure

Blood sampling data
- Time of blood sampling
- Time since last meal

Glycosylated haemoglobin

NMR metabolomics data

14 lipoprotein subclasses
- XXL VLDL
- XL VLDL
- L VLDL
- M VLDL
- S VLDL
- XS VLDL
- IDL
- L LDL
- M LDL
- S LDL
- XL HDL
- L HDL
- M HDL
- S HDL

7 lipid measures for each subclass
- Particle number
- Cholesterol
- Free cholesterol
- Esterified cholesterol
- Triglycerides
- Phospholipids
- Total lipids

Lipoprotein mean particle sizes & apolipoproteins
- VLDL-D
- LDL-D
- HDL-D
- Apo A1
- Apo B

Fatty acids
- Polyunsaturated fatty acids
- Monounsaturated fatty acids
- Saturated fatty acids
- Docosahexaenoic acid
- Linoleic acid
- Omega-3
- Omega-6
- Total fatty acids

Cholines and glycolysis-related
- Total cholines
- Phosphatidylcholine
- Sphingomyelin
- Lactate
- Citrate
- Glucose

Amino acids
- Alanine
- Glutamine
- Histidine
- Isoleucine
- Leucine
- Valine
- Phenylalanine
- Tyrosine

Ketone bodies, inflammation and kidney function
- Acetate
- Aceto-acetate
- β-hydroxy-butyrate
- Albumin
- Creatinine
- Glycoprotein acetyls

Genetic data
- Directly genotyped variants
- Imputed variants
- Exome sequence data
- Whole genome sequence data (10k subset)

Clinical chemistry (1k subset) *
- ALT, AST and GGT
- HDL-C and LDL-C
- Total cholesterol
- Triglycerides
- Apo A1 and Apo B
- Lp(a)
- Albumin
- hs C-reactive protein
- Creatinine
- Uric acid

Resurvey (<u>additional</u> questions / measurements); 10k subset
- Diabetes control questions
- Diabetes consequences *(eg, eyes, amputations, dialysis)*
- Cognitive function (MMSE)
- Fractures/fall questions
- Additional dietary questions *(eg, sugary drinks, added salt, meat, fish, desserts, diets)*
- Bioimpedance *(fat mass, fat free mass, muscle mass, muscle score, bone mass, body water, degree of obesity, visceral fat rating, basal metabolic rate, metabolic age, Rohrer's index)*
- Urine sample (albumin:creatinine ratio)

Cause-specific mortality data (35k dead by mid-2022)
- Date of death
- ICD-10 underlying cause
- ICD-10 contributory causes
- Timing/duration of diseases
- Location of death
- Seen by doctor before death

ALT=Alanine aminotransferase, AST=Aspartate aminotransferase, GGT=Gamma-glutamyl transpeptidase, HDL=High-density lipoprotein, IDL=Intermediate-density lipoprotein, LDL=Low-density lipoprotein, Lp(a)=Lipoprotein(a), TG=Triglycerides, VLDL=Very low-density lipoprotein. All data can be visualised in detail at the study's online Data Showcase, available at: https://datashare.ndph.ox.ac.uk/mexico/index.cgi. * Pilot data only.

Table 3.2: Baseline characteristics of selected socio-demographic characteristics of all recruited participants by sex

	Men n=52,644 (33%)	Women n=107,111 (67%)	All n=159,755
Age, years	53.5 (13.5)	52.2 (13.2)	52.6 (13.3)
Resident of Coyoacán	22,652 (43.0%)	41,181 (38.4%)	63,833 (40.0%)
Highest attained educational level			
University/high school	12,239 (23.3%)	12,179 (11.4%)	24,418 (15.3%)
Middle school	13,356 (25.4%)	25,237 (23.6%)	38,593 (24.2%)
Elementary	21,918 (41.7%)	53,114 (49.6%)	75,032 (47.0%)
Other	5,110 (9.7%)	16,518 (15.4%)	21,628 (13.5%)
Missing	21 (0.0%)	63 (0.1%)	84 (0.1%)
Smoking status			
Never	10,824 (20.6%)	67,335 (62.9%)	78,159 (48.9%)
Ex-smoker	16,271 (30.9%)	15,651 (14.6%)	31,922 (20.0%)
Current (<daily)	8,023 (15.3%)	9,306 (8.7%)	17,329 (10.8%)
Current (<10 cigarettes/day)	10,763 (20.5%)	11,233 (10.5%)	21,996 (13.8%)
Current (>=10 cigarettes/day)	6,712 (12.8%)	3,497 (3.3%)	10,209 (6.4%)
Missing	51 (0.1%)	89 (0.1%)	140 (0.1%)
Alcohol intake			
Never	3,387 (6.4%)	28,709 (26.8%)	32,096 (20.1%)
Former	9,755 (18.5%)	13,062 (12.2%)	22,817 (14.3%)
up to 3 times a month	31,737 (60.3%)	62,984 (58.8%)	94,721 (59.3%)
up to 2 times a week	5,445 (10.3%)	1,704 (1.6%)	7,149 (4.5%)
3+ times a week	2,294 (4.4%)	612 (0.6%)	2,906 (1.8%)
Missing	26 (0.0%)	40 (0.0%)	66 (0.0%)
Physical measures			
SBP, mmHg	128.9 (16.0)	127.2 (17.3)	127.7 (16.9)
Missing	169 (0.3%)	153 (0.1%)	322 (0.2%)
DBP, mmHg	84.4 (10.0)	82.6 (10.3)	83.2 (10.2)
Missing	169 (0.3%)	153 (0.1%)	322 (0.2%)
BMI, kg/m²	27.9 (4.3)	29.5 (5.3)	29.0 (5.1)
Missing	871 (1.7%)	1,260 (1.2%)	2,131 (1.3%)
Waist-to-hip Ratio	1.0 (0.1)	0.9 (0.1)	0.9 (0.1)
Missing	873 (1.7%)	1,245 (1.2%)	2,118 (1.3%)
Prior disease*			
Coronary heart disease	1,121 (2.1%)	1,385 (1.3%)	2,506 (1.6%)
Stroke	696 (1.3%)	1,216 (1.1%)	1,912 (1.2%)
Cancer	7,223 (13.7%)	14,620 (13.6%)	21,843 (13.7%)
Diabetes	377 (0.7%)	1,649 (1.5%)	2,026 (1.3%)
Other‡	2,904 (5.5%)	10,616 (9.9%)	13,520 (8.5%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index

*Self-reported previous diagnoses unless otherwise stated.

‡Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

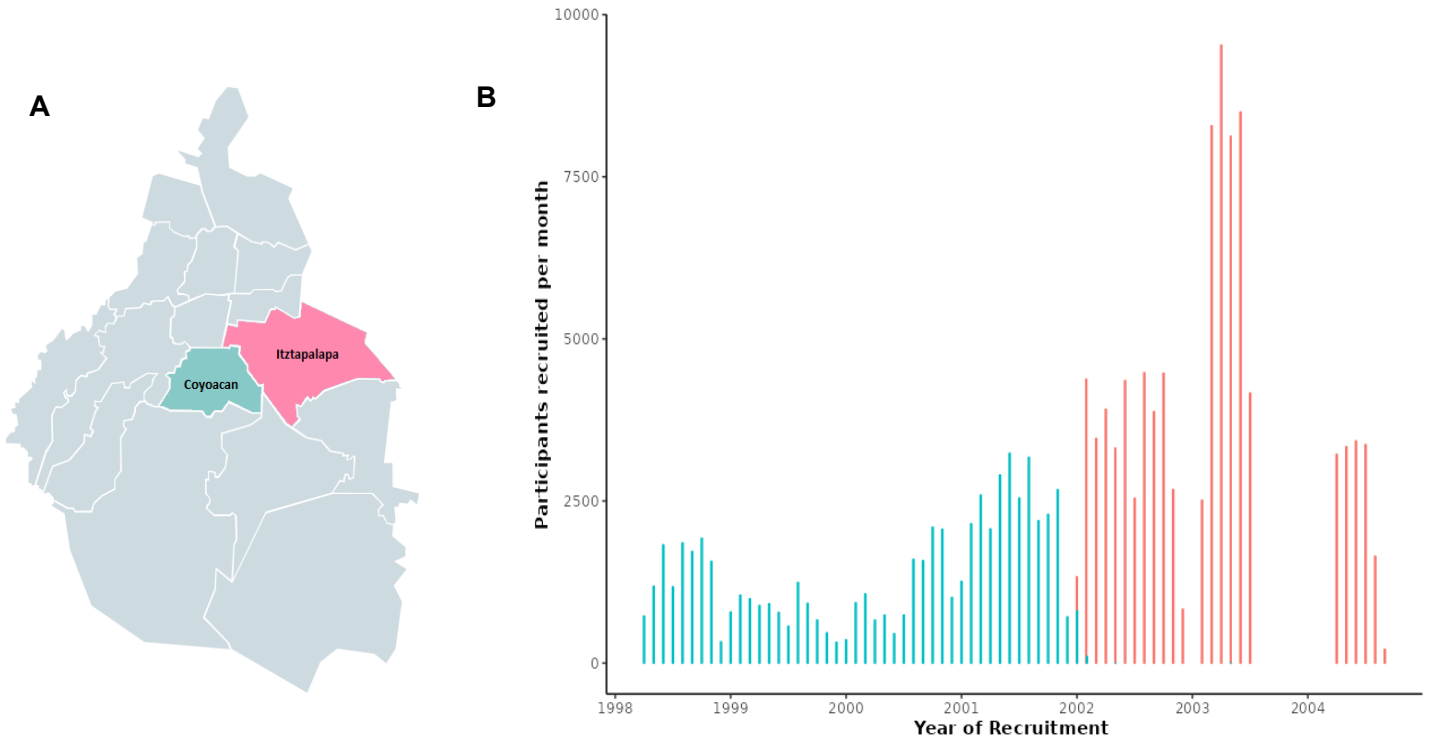


Figure 3.1: A: Two districts of recruitment. B: Recruitment to MCPS per month, coloured by district

3.6 References

1. Tapia-Conyer R, Kuri-Morales P, Alegre-Díaz J, Whitlock G, Emberson J, Clark S, et al. “Cohort profile: the Mexico City Prospective Study”. *Int J Epidemiol* 2006;35(2):pp. 243–9.
2. Ziyatdinov A, Torres J, Alegre-Díaz J, Backman J, Mbatchou J, Turner M, et al. “Genotyping, sequencing and analysis of 140,000 adults from Mexico City”. *Nature* 2023;622(7984):pp. 784–793.
3. Mikkelsen L, Phillips DE, AbouZahr C, Setel PW, Savigny D de, Lozano R, et al. “A global assessment of civil registration and vital statistics systems: monitoring data quality and progress”. *Lancet* 2015;386(10001):pp. 1395–1406.
4. Alegre-Díaz J, Herrington W, López-Cervantes M, Gnatiuc L, Ramirez R, Hill M, et al. “Diabetes and Cause-Specific Mortality in Mexico City”. *N Engl J Med* 2016;375(20):pp. 1961–1971.
5. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. “Next-generation genotype imputation service and methods”. *Nat Genet* 2016;48(10):pp. 1284–1287.
6. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, and Chen W.-M. “Robust relationship inference in genome-wide association studies”. *Bioinformatics* 2010;26(22):pp. 2867–2873.
7. Privé F, Luu K, Blum MGB, McGrath JJ, and Vilhjálmsson BJ. “Efficient toolkit implementing best practices for principal component analysis of population genetic data”. *Bioinformatics* 2020;36(16):pp. 4449–4457.
8. Alexander DH, Novembre J, and Lange K. “Fast model-based estimation of ancestry in unrelated individuals”. *Genome Res* 2009;19(9):pp. 1655–64.
9. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. “A map of human genome variation from population-scale sequencing”. *Nature* 2010;467(7319):pp. 1061–73.
10. Cavalli-Sforza LL. “The Human Genome Diversity Project: past, present and future”. *Nature Reviews Genetics* 2005;6(4):pp. 333–340.
11. García-Ortiz H, Barajas-Olmos F, Contreras-Cubas C, Cid-Soto MÁ, Córdova EJ, Centeno-Cruz F, et al. “The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas”. *Nature Communications* 2021;12(1):p. 5942.
12. National Institute of Statistics and Geography. *Life expectancy at birth by federal entity according to sex, annual series from 2010 to 2024*. Web Page. 2024. `https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Mortalidad_Mortalidad_09_b87a4bf1-9b47-442a-a5fc-ee5c65e37648`.
13. Lambert SA, Wingfield B, Gibson JT, Gil L, Ramachandran S, Yvon F, et al. “The Polygenic Score Catalog: new functionality and tools to enable FAIR research”. *medRxiv* 2024:p. 2024.05.29.24307783.
14. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. “The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation”. *Nat Genet* 2021;53(4):pp. 420–425.
15. Choi SW and O’Reilly PF. “PRSice-2: Polygenic Risk Score software for biobank-scale data”. *GigaScience* 2019;8(7).
16. Privé F, Vilhjálmsson BJ, Aschard H, and Blum MGB. “Making the Most of Clumping and Thresholding for Polygenic Scores”. *Am J Hum Genet* 2019;105(6):pp. 1213–1221.
17. Privé F, Arbel J, and Vilhjálmsson BJ. “LDpred2: better, faster, stronger”. *Bioinformatics* 2021;36(22-23):pp. 5424–5431.
18. Ruan Y, Lin Y.-F, Feng Y.-CA, Chen C.-Y, Lam M, Guo Z, et al. “Improving polygenic prediction in ancestrally diverse populations”. *Nature Genetics* 2022;54(5):pp. 573–580.
19. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. *GigaScience* 2015;4(1).
20. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. *Nature Genetics* 2021;53(7):pp. 1097–1103.

21. Gnatiuc L, Tapia-Conyer R, Wade R, Ramirez-Reyes R, Aguilar-Ramirez D, Herrington W, et al. “Abdominal and gluteo-femoral markers of adiposity and risk of vascular-metabolic mortality in a prospective study of 150 000 Mexican adults”. *Eur J Prev Cardiol* 2022;29(5):pp. 730–738.

22. Gnatiuc L, Alegre-Díaz J, Wade R, Ramirez-Reyes R, Tapia-Conyer R, Garcilazo-Ávila A, et al. “General and Abdominal Adiposity and Mortality in Mexico City: A Prospective Study of 150 000 Adults”. *Ann Intern Med* 2019;171(6):pp. 397–405.

Chapter 4: Transferability of previously published polygenic risk scores for coronary heart disease risk assessment in Mexicans

The results presented in this chapter have been published on medRxiv:

Liu T, Berumen J, Torres J, Alegre-Díaz J, Baca P, González-Carballo C, et al. Polygenic prediction of coronary heart disease among 130,000 Mexican adults. medRxiv. 2024:2024.12.20.24319332.

Summary

Polygenic risk scores (PRSs) have been found to enhance coronary heart disease (CHD) prediction in populations of European ancestry, but evidence on their performance in non-European populations, such as admixed-Americans, is extremely limited. In this chapter the transferability of eight previously-published PRSs, selected from the systematic literature review detailed in **Chapter 2**, among 133,207 genotyped participants aged 35-79 years at recruitment from the Mexico City Prospective Study (MCPS), is evaluated. The outcome of interest is CHD, defined as a history of prior doctor-diagnosed CHD at recruitment or CHD-specific death before the age of 80 years. All eight PRSs were positively and log-linearly associated with CHD, with each one standard deviation (SD) increase in PRS level associated with between a 5% (95% confidence interval (CI), 3%-8%) and a 29% (95% CI, 25%-33%) increase in the odds of CHD. Findings were consistent across different ages, levels of Indigenous ancestry and among unrelated participants. Additional adjustment for conventional CHD risk factors (adiposity, blood-pressure, education level, diabetes status and smoking status) did not significantly alter the strength of the main associations with each PRS. That is, each PRS predicted risk independently of conventional risk factors. Sensitivity analyses confirmed the robustness of the main findings. Notably, significant heterogeneity by sex was observed for six of the eight selected CHD PRSs, with men consistently exhibiting stronger genetic-predisposition to CHD risk compared to women (e.g.,

odds ratio [OR] per SD higher PRS 1.37 [95% CI 1.32–1.43] in men vs. 1.23 [95% CI 1.18–1.28] in women for the PRS with the strongest overall association with CHD). The PRSs constructed using multi-ancestry GWAS information displayed stronger associations with CHD in this Mexican cohort than the PRSs constructed using Euro-centric GWAS information.

4.1 Background and aims

CHD is the global leading cause of death and disability, affecting over 300 million individuals in 2022[1] and preventing it has therefore been a major focus of health disease prevention policies worldwide. With its estimated heritability of 40% to 60%[2,3] and its polygenic nature, combining the effects of multiple single nucleotide polymorphisms (SNPs) on CHD into a single score has the potential to enhance risk prediction and stratification beyond conventional risk factors[4–6]. Genome-wide association studies (GWAS) are a key resource for the construction of a useful PRS and there has been a rapid recent increase in the number of such studies (and methods to develop PRSs), at least among European populations[7–9].

However, as shown in **Chapter 2**, there have been no PRS studies focused on admixed-American-only (or even admixed-American-dominant) populations. PRSs generated from European populations may translate poorly to other populations (not least as any important ancestry-specific SNPs would likely be excluded) but few previous studies have aimed to assess this. Among admixed-Americans in Mexico, there has been an increase in age-specific CHD mortality rates in the past few decades[10], perhaps due to increases in obesity and diabetes. The performance of existing PRSs (and potential development of new PRSs) in such a population is therefore potentially extremely valuable. Using data from MCPS, the aim of this chapter is to take the eight PRSs[11–18] identified in **Chapter 2** and evaluate the shape and strength of their associations with CHD. The development of a new CHD PRS that more closely represents the genetic information of admixed-Americans in MCPS is presented in **Chapter 5**.

4.2 Methods

The dataset used for analyses was MCPS. Details of the MCPS study design, methods and the population are described in **Chapter 3**. In this chapter, all participants aged 35-79 years at recruitment with genotyped data available were included. The main outcome of interest for this population was CHD, defined as self-reported heart attack or angina at baseline and/or death before age 80 with CHD listed anywhere on the death certificate. Other CHD outcomes are also explored however (see **Section 4.2.3.3**).

4.2.1 Selection of PRSs for evaluation

From the systematic literature review described in **Chapter 2**, eight CHD PRSs were selected from 47 published studies that constructed a new PRS. The selection was based on the ancestry relevance of the PRS to the MCPS population, the methodology employed for the composition of the PRS, and the degree of evaluation received in the previous literature. The aim of the selection criteria was to cover a wide spectrum of CHD PRSs, in order to evaluate which statistical method and ancestry combination resulted in better PRS performance in the MCPS. The characteristics of the selected PRSs are presented in **Table 4.1**. A brief description of each PRS is provided below, by the order of the number of SNPs included in each score.

4.2.1.1 Tada *et al.*[17] (50 SNPs)

This is a relative early study in the field on PRS for CHD research. The PRS comprises 50 genome-wide significant variants published by two CHD GWAS conducted in Europeans[19,20]. The score was internally evaluated in a cohort of predominantly European ancestry participants[21] using Cox regression models adjusted for age, sex and established CHD risk factors (i.e., SBP, smoking, prevalent diabetes, ApoB an apoA-I), and yielded a hazard ratio (HR) per SD of 1.23 (95% CI 1.18-1.28) against CHD risk and a Harrell's C-index for discrimination of 0.75.

4.2.1.2 Oni-Orisan *et al.*[11] (164 SNPs)

The original aim of the source study was to evaluate the utility of the CHD PRS in enhancing CHD risk prediction and to assess the effectiveness of statin treatment. The PRS constructed in this study contained all the genome-wide significant CHD SNPs documented up to the year 2018[22] from European GWAS and was evaluated in a multi-ancestry population with 84% European, 7% admixed-American, 6% Asian and 3% African participants. The CHD risk was assessed using a Cox model adjusted for sex, age, smoking and presence of hypertension or diabetes. Among non-statin users (reflecting the general population), the PRS yielded a HR per SD of 1.87 (95% CI 1.19-2.25) in the admixed-American sub-population, 1.59 (1.42-1.78) in the European sub-population, 0.77 (0.32-1.85) in the African sub-population and 2.05 (1.27-3.31) in the East-Asian sub-population.

4.2.1.3 Koyama *et al.*[12] (75,028 SNPs)

This study was one of the largest CHD GWAS among Japanese adults, that subsequently sourced the development of a CHD PRS. The newly-developed PRS was trained using a wide range of methods and GWAS combinations (to integrate ancestry information of Japanese and Europeans) through a 10-fold cross validation approach. The best-performing PRS for predicting CHD risk was determined based on predictive performance (i.e., discrimination) in a logistic model adjusted for age and sex. The best-performing PRS used the GWAS meta-analysis of three cohorts (the study GWAS, UK Biobank[23] and CARDIoGRAMplusC4D[24]) and employed the pruning and thresholding (P+T) method, resulting in an OR per SD of 1.84 (95% CI 1.74-1.94) and an area under the receiver-operating-characteristic curve (AUC) of 0.67 (95% CI 0.66-0.69) during the training stage. In their internal evaluation (on an independent testing sample), the PRS was evaluated in a Cox model adjusted for sex, age, age^2 and top 10 genetic principal components (PCs) against CHD mortality, leading to an HR per SD of 1.22 (95% CI 1.11-1.33).

4.2.1.4 Tcheandjieu *et al.*[14] (538,084 SNPs)

This study first conducted a multi-ancestry GWAS meta-analysis using the three sub-population cohorts (admixed-American, African, and European) from the Million Veteran Program[25] (MVP), along with external GWAS data from European and East-Asian ancestries[23,24,26]. It then developed a CHD PRS using the P+T method, based on the summary statistics from the multi-ancestry GWAS. After tuning for the best-performing PRS, the selected PRS was evaluated separately in the admixed-American, African and European sub-cohorts (heldout from the GWAS) of MVP for prediction of CHD. In the admixed-American sub-cohort, the PRS yielded an OR per SD of 1.43 (95% CI 1.27-1.61) which outperformed the European (only) CHD PRS[15] for which the OR was 1.18 (1.07-1.30).

4.2.1.5 Tamlander *et al.*[13] (1,090,048 SNPs)

Utilising GWAS information sourced from populations of European ancestry[24], this study developed a new PRS for CHD using a novel Bayesian algorithm, PRS-CS (auto). Details of the method applied are described in **Chapter 2 and 5**. The PRS yielded an OR per SD of 1.72 (95% CI 1.70-1.75) and an AUC of 0.79 (95% CI 0.79-0.80) when evaluated in a logistic regression adjusted for year of birth, sex, and the first 10 genetic PCs among participants of UK Biobank[13,27].

4.2.1.6 Patel *et al.*[16] (1,296,172 SNPs)

A multi-trait and multi-ancestry PRS was developed in this study. The development consisted of two stages. In the first stage, a multi-ancestry PRS was created for CHD as well as each of the selected CHD associated trait and risk factors (e.g., diabetes status, blood pressure, ischemic stroke) using a Bayesian algorithm, LDpred2[7] (details of the method are described in **Chapter 2 and 5**), and a GWAS meta-analysis of five ancestries (South and East-Asian, African, admixed-Americans and European). Ancestry-specific scores for each trait (CHD and its associated trait and risk factors) were created. The optimal combination of ancestry-specific PRSs, used to

construct each trait-specific PRS, was identified based on their discriminative performance using the Akaike Information Criterion (AIC) in logistic regression models. In the second stage, these trait-specific PRSs were combined together and the best combination was determined by their logistic regression AIC (similar to stage one). In both stages, a cohort predominately of European ancestry was used for tuning the LDpred2 parameters and for selecting the optimal PRS combinations. The PRS was evaluated within ancestry-specific sub-population groups in MVP[25] (heldout from the GWAS meta-analysis), using logistic regression adjusted for sex, age, genotyping array and the first 10 genetic PCs. The PRS demonstrated an OR per SD for CHD of 1.61 (95% CI 1.53-1.70) in admixed-American sub-population, outperforming other external Eurocentric CHD PRSs.

4.2.1.7 Inouye *et al.*[15] (1,745,179 SNPs)

This study aimed to improve PRS prediction for CHD by combining three CHD PRSs together into a metaGRS (Ensemble method). The three candidate PRSs included one previously published European score[28] and two PRSs generated "in house", with SNP selection based either on false discovery rate (FDR) significant variants or the P+T method using summary statistics from a European GWAS[24]. The weight of each candidate PRS for combination was determined based on their association with CHD (i.e., HR from a Cox model) and pairwise correlation with the other two candidate PRSs. The metaGRS was then evaluated in 480,000 predominately European individuals[27] using a logistic model adjusted for sex, age, genotyping array and 10 genetic PCs and yielded a model AUC of 0.79, representing a 2.8% improvement in model discrimination compared to a model without the PRS. When evaluated in a Cox model, the PRS demonstrated a HR per SD of 1.71 (95% CI 1.68-1.73). This PRS has been extensively evaluated by many studies[5,16,29–36] and demonstrated strong association with CHD risk.

4.2.1.8 Khera *et al.*[18] (6,630,150 SNPs)

The PRS in this study contains the greatest number of SNPs among all eight external candidate PRSs evaluated in this chapter. The study developed a CHD PRS using a European CHD GWAS[24] and was trained using both the LDpred2[7] and P+T methods. The best-performing score was selected based on their logistic regression model performance in a subset of participants from the UK Biobank[27]. The PRS generated through the LDpred2 method showed better performance, and yielded an OR per SD of 1.72 (95% CI 1.67-1.78) and an AUC of 0.81 in a logistic regression model adjusted for age, sex and the first four genetic PCs during training. The model discrimination power of the PRS was comparably strong (AUC=0.81) when evaluated in the UK Biobank sample retained for testing (i.e., different from the training subset of the UKB population). This score was one of the most commonly evaluated PRS in external studies of CHD risk (see **Chapter 2**).

Of all the eight candidate PRSs selected for evaluation in MCPS, five scores used genetic information restricted solely to European ancestries[11,13,15,17,18] and three used combined genetic information from multiple ancestries[12,14,16].

4.2.2 Recreation of CHD PRSs for assessment of CHD risk in the MCPS population

The eight CHD-specific PRS (above) were recreated for each MCPS participant using study-specific genetic information and a validated pipeline developed by the PGS Catalog[37,38]. The PGS Catalog is a publicly available online database which collects standardised information of published PRS, such as SNP name, genome build, outcome definition, effect sizes and GWAS-specific ancestries that sourced the PRS. PRSs stored in the catalog are annotated with a unique ID, which can be used to retrieve detailed PRS-specific information. The PGS catalogue developed a reproducible workflow for PRS calculation called pgsc_calc[39], which takes the genetic information for the analysis population and the PRS annotation IDs as input variables. During the computation process of the score within the MCPS study population, all of the eight selected

PRSs had their score-specific information stored in the Catalog and their IDs were used to retrieve their PRS information stored in the Catalog. The individual SNPs and their associated weights varied substantially across the eight external PRSs selected for evaluation in MCPS. Consequently, each PRS was subsequently standardised to have mean 0 and standard deviation 1 to allow comparisons with each other in the subsequent analysis. The SNP variant matching rates for all of the selected PRSs within the MCPS were over 85%, based on exact matching with no use of proxy SNPs. The PGS Catalog program pgsc_calc (v1.3.2) was performed using the Nextflow-23.04.0[40], PLINK-2.0[41], and Anaconda Distribution 3 software.

4.2.3 Statistical analysis

4.2.3.1 Primary analysis

The general statistical approaches were provided in **Chapter 3**. Briefly, logistic regression was used to assess the association between each candidate PRS and the odds of CHD. First, to assess the shape of the associations, each PRS was categorised into five equally-sized groups with the lowest group as the reference category. Subsequently, to assess the average strength of the associations, each PRS was entered into the model as a continuous (standardised) variable. These logistic regression models were initially adjusted just for age, sex and the first seven genetic PCs ('partial adjustments') to keep consistent with most of the PRS source studies. Subsequently, to assess the extent to which each PRS predicted CHD *independently* of established vascular risk factors (noting that some of the variants contributing to each PRS are included precisely *because* they predict such risk factors), models were further adjusted for highest level of educational attainment (university or high school, middle school, elementary school, other/none), waist-to-hip ratio (WHR), systolic and diastolic blood pressure (SBP and DBP), smoking status (never, former, current), and presence of diabetes at recruitment into the study (previously-diagnosed or HbA1c ≥6.5%, vs not) ('full adjustment'). Missing values were either median imputed for continuous variables or imputed using the highest selected category for categorical variables. The AUC was estimated to assess model discrimination.

4.2.3.2 Secondary analysis

To investigate the potential effect modification of genetic predisposition to CHD risk by conventional risk factors, analyses were also done separately by sex, baseline age (35-54, 55-64 and 65-79 years, respectively), proportion of Indigenous American ancestry (<60%, 60 to <80% and ≥80%, respectively), as well as the other factors included in the fully adjusted model (described above). The continuous variables used in the stratified analysis were categorised as follows, WHR was grouped as <0.85, 0.85 to <0.9 and ≥0.9 units, SBP as <120, 120-139 and ≥140 mmHg, and DBP as <80, 80-89 and ≥90 mmHg. For each stratified analysis a test of heterogeneity or trend was performed (as appropriate) in order to assess whether the strata-specific ORs varied significantly between the subgroups considered.

4.2.3.3 Sensitivity analysis

Sensitivity analyses included redefining the primary CHD outcome excluding deaths where CHD was not the underlying (i.e., the primary) cause, restricting to non-fatal self-reported myocardial infarction, angina or both at baseline, restricting to non-fatal self-reported myocardial infarction, angina or both at baseline, restricting to the CHD death component (as listed anywhere on the death certificate and as listed as the underlying cause). In addition, Cox regression was used to model time to CHD death in the full MCPS study cohort, with participants who had CHD at baseline retained to ensure consistency with the primary analysis cohort. Analyses of the primary CHD outcome were also repeated limited to participants who were unrelated to the 3rd degree (based on identity-by-descent [IBD] estimates, see **Chapter 3**). Finally, additional analyses extended the age range studied (participants and CHD events) up to age 90 years.

4.2.3.4 Visualisation of results

When assessing the shape of the associations between each PRS and CHD across five equally-sized groups of each PRS, a group-specific 95% confidence interval was estimated around each

odds ratio [OR], including the reference group with OR of 1.0[42]. The ORs and their CIs for each PRS category were plotted on the y-axis and each PRS category represented by the mean values within that specific category was plotted on the x-axis. A line of best fit was also plotted across the five groups, which was calculated based on the log ORs and their standard errors (SEs), using the inverse-variance weighted (IVW) method.

Data cleaning, analyses and figures were performed in R 4.2.1. Statistical significance was determined at a 2-sided alpha of 0.05.

4.3 Results

4.3.1 Participants included

Of the 159,755 participants recruited in MCPS, 22,364 (14%) were excluded. The reason of exclusion comprised 18,924 (11.8%) individuals without genetic data available or their genetic data failed one of the QC thresholds described in **Chapter 3**, 2,221 (1.4%) individuals with unverified mortality linkage, and a further 1,219 (0.7%) individuals aged ≥90 years at recruitment. Of the remaining 137,391 participants, 133,207 were aged 35-79 years at recruitment (included in the primary analysis) and a further 4,184 participants were aged 80-89 years (included in secondary analyses).

4.3.2 Baseline characteristics of included MCPS participants

Table 4.2 displays the baseline characteristics of the 133,207 participants included in the primary analysis. Among these, over two-thirds were women, the mean age at recruitment was 51 years (SD 12 years), and 57% participants were unrelated to the 3rd degree. On average, the ancestry admixture comprised a 67% inherited Indigenous American ancestry, 28% European, 3% African, and 1% East Asian. Overall, the highest level of educational attainment was 16% for university or high school, 25% for middle school, 47% for elementary school and 12% other. A lower proportion of women than men had completed university or high school. At baseline, 32% of the

participants were current smokers and 20% were former smokers, with a much higher proportion of men being ever smokers, compared to women. Mean SBP and DBP were 127 mmHg (SD 17 mmHg) and 83 mmHg (SD 10 mmHg), respectively. A history of CHD (i.e., myocardial infarction or angina) at baseline was reported by 1,901 (1.4%) participants and a history of stroke was reported by 1,414 (1.1%) participants, respectively. 24,796 participants (19%) self-reported a history of diagnosed diabetes or had an HbA1c concentration indicative of undiagnosed diabetes (≥6.5%), while 8% of the participants reported having at least one other chronic disease (e.g., hepatic, renal, neoplastic, or respiratory).

Overall, the baseline characteristics for all the 137,391 participants aged 35-79 years at recruitment were similar to those aged 35-74 years described above, and they are presented in **Table 4.3**.

4.3.3 Fatal and non-fatal CHD cases

The median (IQR) follow-up among those still alive as of September 30, 2022 was 20.3 years (19.4 to 21.5 years). For the participants aged 35-79 years at recruitment, there were 17,737 deaths before age 80 including 3,479 that were attributed to CHD (according to the medical ICD-10 classification recorded in the death registries). Of these 3,479 deaths, CHD was listed as the primary underlying cause for 2,927 and as a contributing cause for 552 deaths. Of those who died before age 80 from CHD, 217 also reported having doctor-diagnosed CHD at baseline assessment, and 3,262 did not. Consequently, 5,163 participants were considered to have CHD (i.e., the primary outcome definition), based on either self-reported medical history CHD or having died before the age of 80 years, from a CHD cause (i.e., 3479 with CHD death before age 80 plus 1,901 with pre-existing CHD at recruitment minus 217 with both). **Table 4.4** presents the subgroups of CHD, with myocardial infarction identified as the most common subtype.

4.3.4 Characteristics of the study participants by different levels of inherited genetic-risk for CHD

The eight evaluated CHD PRSs included between 44 and 6,472,620 unique SNPs (based exclusively on exact SNP matching, without proxies). **Figure 4.1** displays the pairwise correlation heatmap between all the eight PRSs. The European-based PRSs that included thousands of SNPs (Khera *et al.*[18], Inouye *et al.*[15] and Tamlander *et al.*[13]) were highly correlated with each other, with correlation coefficients > 0.5. Two of the multi-ancestry PRS (Koyama *et al.*[12] and Patel *et al.*[16]) were also strongly correlated with each other and also with the European PRS that included thousands of SNPs ($r^2 > 0.5$).

Table 4.5 to 4.12 show the baseline characteristics of the participants included in the analyses, by five equally-sized PRS groups of the eight evaluated CHD PRSs. Six of all PRSs[11–16] showed significant differences by PRS level in the mean inherited Indigenous American ancestry proportion. Variations in the highest education attainment levels were observed for the PRSs Inouye *et al.*[15], Tcheandjieu *et al.*[14], and Tamlander *et al.*[13]. Differences in smoking status prevalence were noted for the PRSs Tcheandjieu *et al.*[14], Patel *et al.*[16], and Inouye *et al.*[15]. Differences in LDL-C levels were associated with the PRSs Tcheandjieu *et al.*[14], Patel *et al.*[16], Inouye *et al.*[15], and Khera *et al.*[18]. Similarly, triglycerides (TG) and HbA1c levels varied for the PRSs Koyama *et al.*[12], Tamlander *et al.*[13], Patel *et al.*[16], and Inouye *et al.*[15]. Finally, differences in diabetes prevalence at baseline assessment were observed for the PRSs Koyama *et al.*[12], Patel *et al.*[16], Inouye *et al.*[15], and Tamlander *et al.*[13]. No significant differences across PRS levels were observed for alcohol intake, BMI, WHR, blood pressure levels or estimated glomerular filtration rate (eGFR) across the various groups of genetic predisposition to CHD.

4.3.5 Association between CHD PRS and subsequent risk of CHD in MCPS

For all the external PRSs studied, the association between a higher level of genetic predisposition and the odds of having CHD was positive and largely log-linear (**Figure 4.2 and Figure 4.3**).

When treating each PRS as a categorical variable, those within the 20% highest level of genetic predisposition to CHD had a 1.14 to 2.02 times higher odds of CHD compared with those in the lowest (reference) 20% group (**Figure 4.2**). Among the eight PRSs selected, the strength of the association between the PRS and CHD was somewhat weaker for the PRSs that were constructed using only genome-wide significant variants (and thus included fewer SNPs[11,17]), compared with the PRSs that were constructed using more advanced methods (which included thousands of SNPs). In the primary analysis adjusted for age, sex and the first seven genetic PCs for ancestry admixture, the strongest overall association with CHD risk in MCPS was found for the Patel *et al.*[16] PRS, for which each 1SD higher level of genetic predisposition was associated with a 29% increase in the odds of subsequent CHD (OR per SD=1.29, 95% CI 1.25-1.33) (**Figure 4.3**). In contrast, the weakest overall association was seen for the Tada *et al.* PRS[17], for which each 1SD higher level of genetic predisposition was associated with only a 5% increase in the odds of CHD (OR per SD=1.05, 1.03-1.08). After additional adjustment for conventional vascular risk factors (i.e., highest education attainment, WHR, blood-pressure, smoking status, and pre-existing diabetes at baseline), the strength of the association with CHD was slightly attenuated for each of the eight PRSs (**Figure 4.3**). In addition, the multi-ancestry PRS (Patel *et al.*[16], Koyama *et al.*[12] and Tcheandjieu *et al.*[14]) showed relatively stronger associations with CHD risk compared with the European-sourced PRS.

The model AUC for a logistic regression including age, sex and the first seven PCs (without PRS) was 0.716 (0.709-0.722) and further accounting for conventional CHD risk factors increased it to 0.745 (0.739-0.751). Including a PRS into the models only improved the AUC marginally. For instance, including the Patel *et al.* PRS[16], the PRS with the greatest association with CHD, into the models improved the model AUC to 0.724 (0.717-0.730) and 0.749 (0.743-0.755), respectively (**Figure 4.3**).

4.3.6 Differences in genetic predisposition to CHD risk by sex

In the sex stratified analyses, two of the eight PRSs with around a hundred genome-wide significant SNPs included (Tada *et al.*[17] [44 SNPs] and Oni-Orisan *et al.*[11] [141 SNPs]) displayed similar strengths of association with CHD risk in both men and women (**Figure 4.4**). By contrast, the other six PRSs (with thousands to millions of SNPs included) (Koyama *et al.*[12], Tcheandjieu *et al.*[14], Tamlander *et al.*[13], Patel *et al.*[16], Inouye *et al.*[15] and Khera *et al.*[18]), displayed significant sex-specific heterogeneity in genetic predisposition to subsequent CHD risk, with a significantly stronger association observed in men compared to women. The multi-ancestry PRS Tcheandjieu *et al.*[14] exhibited the greatest sex heterogeneity, with an OR of 1.30 (95% CI 1.24-1.37) per 1SD higher genetic predisposition in men versus an OR of 1.10 (95% CI 1.05-1.15) in women. For both men and women, the Patel *et al.*[16] PRS showed the strongest association with CHD risk, with an OR per SD of 1.37 (95% CI 1.32-1.43) and 1.23 (1.18-1.28) in men and women, respectively. Further adjustments for conventional CHD risk factors did not essentially alter these sex differences, however, the model discrimination was slightly higher for women than for men, for all the eight PRSs evaluated (reflecting the slightly better ability of the other risk factors to predict CHD in women than men).

4.3.7 Genetic predisposition to CHD risk by other characteristics of the MCPS population

4.3.7.1 Differences in genetic predisposition to subsequent CHD risk by various biological characteristics

Subgroup analyses aiming to investigate potential effect modification by different levels of age, educational attainment, waist-hip ratio, blood pressure, smoking status, pre-existing diagnosed or undiagnosed diabetes, and inherited proportion of Indigenous ancestry are shown in **Figure 4.5 to Figure 4.12**. Overall, results for each PRS were broadly consistent by levels of these factors. Furthermore, for most of the evaluated PRSs, the associations appeared *slightly* stronger among those with a higher waist-to-hip ratio, or a higher level of education, and among those

with lower blood pressure (SBP and DBP), but did not materially differ by diabetes status. A slight U-shaped trend was observed for smoking status, as the association with CHD risk appeared strongest among former smokers. Similar trend was also observed for levels of Indigenous American ancestry proportion, with individuals with 60%-80% proportions of Indigenous ancestry showing a greater genetic predisposition to CHD risk. However, the confidence intervals for these mentioned associations overlapped.

4.3.7.2 Genetic predisposition to CHD risk including older participants in MCPS

Extending the analysis to MCPS individuals within the age range of 35-89 years yielded similar results to the main analyses (**Figure 4.13 to Figure 4.14**). Among 137,391 participants aged 35-89 years, 7,155 either self-reported pre-existing CHD or died before the age of 90 years from CHD listed on their death certificate. All PRSs exhibited a positive and log-linear relationship with CHD, although the association was slightly attenuated for most of the PRSs. The Patel *et al.*[16] PRS remained the most strongly associated PRS, with an OR per SD of 1.24 (95% CI 1.21-1.27) when adjusted for age, sex and the first seven genetic PCs.

4.3.8 Genetic predisposition to CHD, given various definitions and relatedness

Findings of the shapes and strength of the associations were broadly consistent for the alternative definitions of CHD considered (**Figure 4.15 to Figure 4.22**). For all the eight candidate PRSs, the weakest association was seen for CHD defined solely by self-reported angina at baseline, while fatal CHD resulted in stronger associations. The Patel *et al.*[16] PRS consistently demonstrated the strongest association with CHD, irrespective of the exact CHD definition. This PRS also showed significant sex heterogeneity for all definitions of CHD apart from CHD deaths, and the differences were the most apparent for self-reported CHD at baseline (**Figure 4.23**)).

The results from the Cox models were largely consistent with the main analysis findings (see **Figure 4.24 and Figure 4.25**). The Patel *et al.*[16] PRS continued to show the strongest association with CHD mortality risk, with hazard ratios (HRs) per SD of 1.29 (95%CI 1.25-1.33) and 1.30

(1.26-1.35) when adjusted for age, sex and the first seven genetic PCs, for deaths with CHD listed anywhere or as the primary cause on the death certificate, respectively.

When the analyses were restricted to individuals unrelated to the 3rd degree, 75,875 participants were retained, of which 2,947 met the primary CHD definition. The genetic predisposition to CHD showed similar results among the unrelated participants (**Figure 4.26**). When adjusted for age, sex and the first seven genetic PCs, the PRS by Patel *et al.*[16] showed the greatest association with CHD in the unrelated subset, with an OR per SD of 1.29 (95%CI 1.25-1.34), while the Tada *et al.*[17] PRS showed the weakest association, with an OR per SD of 1.06 (1.02-1.10).

4.4 Discussion

4.4.1 Summary of findings

In this large study of Mexican adults, eight previously-published CHD PRSs derived from either European or multi-ancestry GWAS sources were evaluated for their relevance to CHD. The results demonstrated reasonable transferability and good potential for reliably predicting CHD risk using the selected external PRS in this Mexican population. All the evaluated PRSs showed significant and positive log-linear associations with CHD risk at ages 35-79 years, with patterns robust to more stringent CHD definitions and genetic (un)relatedness, reinforcing the validity of the findings. The multi-ancestry sourced PRSs consistently showed stronger associations with CHD risk compared to the European-sourced PRSs. However, there was only slight difference in model discrimination among the candidate PRSs. One key finding was the sex heterogeneity among six of the candidate PRSs, with men having a significantly stronger genetically-predicted CHD risk compared to women. These findings were among the PRSs including thousands of SNPs, potentially suggesting that PRSs with a larger number of SNPs may better capture sex-specific genetic predisposition to CHD risk, compared to PRSs with only genome-wide significant SNPs, although this could also reflect an important gap in GWAS studies of CHD in women compared with men.

4.4.2 Comparison of findings in the MCPS with the PRS-source studies

The estimated effect sizes of all of the eight PRSs on CHD in the current study are smaller than the effect sizes reported in their source studies (**Table 4.1**) or the estimates from the external validations in admixed-American populations conducted by the source studies[14,35,43]. This could be partly attributed to lack of non-fatal incident CHD data, resulting in a relatively lower CHD case rate in MCPS. In the source-studies, the CHD event rates were notably higher during the training process of the selected PRS. For instance, the CHD case rates were 20%, 4.6%, 9.1%, 15.2%, respectively for Patel *et al.*[16], Inouye *et al.*[15], Tcheandjieu *et al.*[14] and Koyama *et al.*[12], compared to only 3.9% in the MCPS cohort. These differences could be further investigated when the accruing non-fatal data becomes available in future analyses of MCPS.

4.4.3 Sex heterogeneity in genetic predisposition to CHD risk

Similar sex heterogeneity in CHD risk, given genetic predisposition to CHD, have been reported by several other studies conducted in populations of European ancestry[15,16,44], with all previous studies having reported significantly greater CHD risk in men than women. One previous study investigated sex differences using a CHD PRS and three subscores based on CHD mediating factors (blood pressure, blood lipids and body mass index) by including a sex-PRS interaction term in their model originally adjusted for age, population structure and other conventional CHD risk factors[44]. The study identified a significant gene-sex interaction for the blood pressure score in its association with incident CHD, both overall and across different levels of genetic predisposition to CHD. The study further identified a novel genetic locus at 21q22.11 which showed significant association with CHD only in men. This locus has been found to link with bone mineral density, body fat distribution, and pulse pressure in previous research. Poorer genetic prediction of CHD in women may also reflect sex-biases that exist in clinical definitions of CHD outcomes (e.g., progression, presentation, age at onset). For instance, the age of onset for women tends to be later than men. However, most of these definitions are dominated by studies of men, resulting in

poorer performance in women[45]. Further sex-specific research is needed for CHD genetics (e.g., GWAS and PRS studies) to address these gaps, allowing better understanding of the genetic risk for CHD in both men and women.

4.4.4 Genetic predisposition to CHD and other conventional CHD risk factors

The CHD risk conferred by each of the eight evaluated PRS and the conventional vascular risk factors in MCPS were largely independent of each-other, consistent with reports from previous studies conducted in European populations[15,46–48]. For instance, Inouye *et al.*[15] conducted a competing risk analysis between a CHD PRS and CHD risk factors and found no evidence of competing risk. Moreover, they found that their PRS could still predict CHD among patients receiving CHD treatments, suggesting the additive and independent value of genetic risk in disease prediction beyond conventional clinical measures. This suggests that combining genetic and clinical information has the potential to further improve CHD risk prediction and stratification, and that addressing modifiable conventional risk factors for preventing CHD events remains important clinically. Indeed, several studies that combined PRSs constructed using the latest state-of-the-art methods and clinical risk scores or conventional risk factors have shown that the integrated score could enhance CHD screening in addition to existing clinical tools, and inform precision medicine efforts for better stratification of CHD risk among otherwise 'healthy' individuals[5,46,49].

4.4.5 The relevance of multi-ancestry PRS to CHD risk

Beyond the cohort-specific factors described in **Section 4.4.2** the diminished performance of PRSs derived from European populations in the Mexican cohort aligns with broader findings across diverse ancestry groups. As highlighted in **Chapter 2**, PRSs derived from European populations demonstrated strong performance when evaluated in populations of the same ancestry, with AUC values ranging from 0.75 to 0.81 (**Table 4.1**). However, several studies found that the strength of association for European GWAS-sourced PRSs remained positive but became weaker when applied to populations of admixed-American ancestry, as well as populations of

other ancestries (e.g., African, Middle Eastern)[14,35,43,50].

The ability of a PRS to predict CHD depends on several factors such as the sample size of the GWAS input used in the PRS algorithms, and the accuracy of the effect estimates for the individual genetic variants included. The genetic effects estimated by GWAS are influenced by population-specific environmental factors, which also vary significantly across different ancestries. In an ideal scenario, the best PRS to predicting the risk of CHD would be derived using large-scale admixed-American cohorts that could best capture the shared and unique genetic and non-genetic architectures for the risk factors that are specific to admixed-American populations. However, as admixed-Americans are highly underrepresented in GWAS, due to lack of large population-specific cohorts, there is a critical gap in the development of an 'admixed-American' PRS. It has been suggested that leveraging diverse ancestry sources for GWAS inputs may improve the generalisability of PRS across diverse ancestries, as the true causal variants for CHD risk are likely shared across populations[48]. This has been proved by the present findings on the Patel *et al.*[16] and the Koyama *et al.*[12] PRS, which demonstrated the strongest association with CHD risk in this chapter, and are both multi-ancestry GWAS-sourced instruments. These two selected PRSs outperformed all other candidate PRS derived from large European-dominant GWAS that included even more SNPs. Nevertheless, multi-ancestry PRSs have limitations as they do not fully account for population-specific environmental influences, and may overlook risk variants that are unique to under-represented populations. Therefore, there is urgent need for large-scale GWAS in admixed-American populations to ensure their fair representation in genetic research and to enhance the accuracy and transferability of PRSs in diverse populations.

4.4.6 Conclusion

In summary, this chapter evaluated the performance of eight previously-published CHD PRSs identified in **Chapter 2** in a large cohort of 130,000 Mexican adults, a population of predominantly admixed-American ancestry that has been understudied in genetic studies. The findings showed

that externally derived PRSs have reasonable transferability and could predict CHD risk independent of conventional vascular risk factors. Sex heterogeneity was identified for six PRSs (those with thousands of SNPs included) with men showing greater associations with CHD risk than women. Overall, multi-ancestry PRSs generally outperformed European-only PRSs in terms of strengths of associations with risk. However, the PRSs selected for assessment in MCPS do not fully account for the genetic architecture of CHD in Mexico. PRSs derived from large studies with a greater proportion of admixed-Americans could more closely represent genetic variation in Latin America, and may further enhance accuracy and polygenic prediction of CHD risk in admixed-American adults.

The next chapter will provide a detailed description on the computation of GWAS and development of a CHD PRS utilising the largest prospective cohort of Mexicans to date and leveraging advanced algorithms.

Table 4.1: CHD polygenic risk score selected for evaluation in MCPS

Authors / PRS ID in PGS catalogue/Ancestry	PRS characteristics	CHD odds or hazard ratio (95% CI)*	AUC (95%CI)
Tada *et al*[17] /PGS000011/European	A PRS constructed using 50 SNPs and European GWAS. **GWAS:** European dominant (~95%) with a small proportion of Asian samples **Evaluation sample:** 100% European	HR per SD: 1.23 (1.18-1.28)	0.75 (no CI reported)
Oni-orisan *et al*[11] / PGS004595/European	A PRS constructed using European GWAS and 164 genome-wide significant CHD risk SNPs from UKB and CC4D GWAS. It was evaluated among datasets that contains admixed-American participants (n=1538). **GWAS:** European dominant **Evaluation sample:** Multi-ancestry	HR per SD: 1.87 (1.19-2.95)	NR
Koyama *et al*[12] /PGS000337/Japanese and European	A trans-ancestry PRS that used Japanese and European genetic information during construction. **GWAS:** European and Japanese dominant **Training sample:** Japanese dominant **Evaluation sample:** Japanese dominant	OR per SD: 1.84 (1.74-1.94)	0.67 (0.66–0.69)
Tcheandjieu *et al*[14] /PGS003446/Multi-ancestry	A trans-ancestry PRS that used admixed-American, European and Japanese genetic information during construction using pruning and thresholding[51]. **GWAS:** Multi-ancestry (admixed-Americans, African, Japanese, European) **Training sample:** European, African and admixed-Americans **Evaluation sample:** Admixed-American	OR per SD: 1.43 (1.27-1.61)	NR
Tamlander *et al*[13] / PGS001780/European	A PRS developed using a novel method, PRS-CS-auto. **GWAS:** European dominant (~85%) with a small proportion of Asian samples **Evaluation sample:** European	OR per SD (in UK Biobank): 1.72 (1.70-1.75)	0.79 (0.79–0.80)
Patel *et al*[16] /PGS003725/Multi-ancestry	A trans-ancestry and multi-trait PRS derived using Admixed-American, European, African and Asian genetic information. **GWAS:** Multi-ancestry (admixed-Americans, African, Japanese, European) **Training sample:** European **Evaluation sample:** Admixed-American	OR per SD: 1.61 (1.53-1.70)	NR
Inouye *et al*[15] /PGS000018/European	A meta-score based on the weighted average of 3 PRSs. **GWAS:** European dominant (~90%) with a small proportion of Asian samples **Training sample:** European dominant **Evaluation sample**: European dominant	HR per SD: 1.71 (1.68-1.73)	0.79 (no CI reported)
Khera *et al*[18] /PGS000013/European	A PRS developed using LDPred2[7] with a large European dominant GWAS input. **GWAS:** European dominant (~85%) with a small proportion of Asian samples **Training sample:** European dominant **Evaluation sample:** European dominant	OR per SD (in training sample): 1.72 (1.67-1.78)	0.81 (0.81-0.81)

* In the 'evaluation sample' unless otherwise stated. AUC=Area under the receiver operating characteristic curve. CHD=Coronary heart disease. CI=Confidence interval. OR= Odds ratio. HR=Hazard ratio. NR=Not reported

Table 4.2: Baseline Characteristics of 133,207 participants aged 35-79 years (main analysis population)

	Men n=43,338 (33%)	Women n=89,869 (67%)	All n=133,207
Age, years	52.0 (12.1)	50.8 (11.7)	51.2 (11.8)
Resident of Coyoacán	17,958 (41%)	33,057 (37%)	51,015 (38%)
Unrelated participants	24,145 (56%)	51,730 (58%)	75,875 (57%)
Ancestry admixture percentage			
Indigenous American	66.5 (18.0)	67.0 (17.8)	66.8 (17.9)
African	3.4 (2.8)	3.5 (2.8)	3.4 (2.8)
East Asian	1.4 (2.0)	1.4 (1.7)	1.4 (1.8)
European	28.7 (16.3)	28.1 (16.1)	28.3 (16.2)
Highest attained educational level			
University/high school	10,340 (24%)	10,369 (12%)	20,709 (16%)
Middle school	11,468 (26%)	21,964 (24%)	33,432 (25%)
Elementary	17,902 (41%)	45,005 (50%)	62,907 (47%)
Other	3,616 (8%)	12,475 (14%)	16,091 (12%)
Missing	12 (0%)	56 (0%)	68 (0%)
Smoking status			
Never	8,719 (20%)	55,872 (62%)	64,591 (48%)
Former	13,060 (30%)	13,116 (15%)	26,176 (20%)
Current	21,521 (50%)	20,809 (23%)	42,330 (32%)
Missing	38 (0%)	72 (0%)	110 (0%)
Alcohol intake			
Never	2,660 (6%)	23,443 (26%)	26,103 (20%)
Former	7,598 (18%)	10,653 (12%)	18,251 (14%)
Current	33,062 (76%)	55,741 (62%)	88,803 (67%)
Missing	18 (0%)	32 (0%)	50 (0%)
Physical measures			
SBP, mmHg	128.5 (15.8)	126.6 (16.9)	127.2 (16.6)
DBP, mmHg	84.4 (9.9)	82.5 (10.2)	83.1 (10.2)
BMI, kg/m²	28.0 (4.3)	29.6 (5.3)	29.1 (5.1)
Waist-to-hip Ratio	0.95 (0.07)	0.88 (0.07)	0.90 (0.08)
Laboratory measurements			
HDL-C, mmol/L	0.93 (0.19)	1.03 (0.22)	1.00 (0.21)
LDL-C, mmol/L	2.39 (0.79)	2.49 (0.79)	2.46 (0.79)
Triglycerides, mmol/L	1.65 (0.68)	1.54 (0.64)	1.57 (0.66)
HbA1c, %	6.10 (1.72)	6.10 (1.71)	6.10 (1.71)
eGFR, ml/min/1.73m²*	101.0 (15.9)	102.0 (16.1)	101.7 (16.0)
Prior disease †			
Coronary heart disease	848 (2%)	1,053 (1%)	1,901 (1%)
Stroke	485 (1%)	929 (1%)	1,414 (1%)
Cancer	266 (1%)	1,314 (1%)	1,580 (1%)
Diabetes ‡	8,250 (19%)	16,546 (18%)	24,796 (19%)
Other §	2,276 (5%)	8,837 (10%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

* Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

† Self-reported previous diagnoses unless otherwise stated.

‡ Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

§ Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

Table 4.3: Baseline characteristics of 137,391 participants aged 35-89 years

	Men n=44,853 (33%)	Women n=92,538 (67%)	All n=137,391
Age, years	53.1 (13.1)	51.8 (12.8)	52.2 (12.9)
Resident of Coyoacán	18,425 (41%)	33,947 (37%)	52,372 (38%)
Unrelated participants	24,979 (56%)	52,896 (57%)	77,875 (57%)
Ancestry admixture percentage			
Indigenous American	66.4 (18.1)	67.0 (17.9)	66.8 (17.9)
African	3.4 (2.8)	3.5 (2.8)	3.4 (2.8)
East Asian	1.4 (2.0)	1.4 (1.7)	1.4 (1.8)
European	28.7 (16.4)	28.2 (16.2)	28.4 (16.2)
Highest attained educational level			
University/high school	10,443 (23%)	10,421 (11%)	20,864 (15%)
Middle school	11,596 (26%)	22,116 (24%)	33,712 (25%)
Elementary	18,694 (42%)	46,226 (50%)	64,920 (47%)
Other	4,104 (9%)	13,716 (15%)	17,820 (13%)
Missing	16 (0%)	59 (0%)	75 (0%)
Smoking status			
Never	9,084 (20%)	57,951 (63%)	67,035 (49%)
Former	13,876 (31%)	13,553 (15%)	27,429 (20%)
Current	21,852 (49%)	20,958 (23%)	42,810 (31%)
Missing	41 (0%)	76 (0%)	117 (0%)
Alcohol intake			
Never	2,859 (6%)	24,582 (27%)	27,441 (20%)
Former	8,179 (18%)	11,160 (12%)	19,339 (14%)
Current	33,793 (75%)	56,763 (61%)	90,556 (66%)
Missing	22 (0%)	33 (0%)	55 (0%)
Physical measures			
SBP, mmHg	128.8 (15.9)	127.0 (17.2)	127.6 (16.8)
DBP, mmHg	84.4 (10.0)	82.5 (10.3)	83.1 (10.2)
BMI, kg/m^2	27.9 (4.3)	29.6 (5.3)	29.0 (5.1)
Waist-to-hip Ratio	0.95 (0.07)	0.88 (0.07)	0.90 (0.08)
Laboratory measurements			
HDL-C, mmol/L	0.93 (0.19)	1.04 (0.22)	1.00 (0.21)
LDL-C, mmol/L	2.38 (0.79)	2.49 (0.79)	2.45 (0.79)
Triglycerides, mmol/L	1.64 (0.68)	1.53 (0.64)	1.57 (0.66)
HbA1c, %	6.10 (1.70)	6.10 (1.70)	6.10 (1.70)
eGFR, ml/min/1.73m^2[§]	100.3 (16.4)	101.3 (16.6)	100.9 (16.5)
Prior disease[*]			
Coronary heart disease	937 (2%)	1,152 (1%)	2,089 (2%)
Stroke	547 (1%)	1,011 (1%)	1,558 (1%)
Cancer	311 (1%)	1,377 (1%)	1,688 (1%)
Diabetes[†]	8,599 (19%)	17,242 (19%)	25,841 (19%)
Other[‡]	2,449 (5%)	9,128 (10%)	11,577 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.4: CHD case type by ICD-10 codes and diagnosis

Outcomes	Men	Women	Overall
Primary definition of CHD	2,415	2,748	5,163
Baseline CHD	848	1,053	1,901
Baseline angina	136	358	494
Baseline myocardial infarction	747	729	1,476
CHD deaths (anywhere on the death certificate, I20-I25)	1,717	1,762	3,479
I20 death (angina pectoris)	19	18	37
I21 death (acute myocardial infarction)	1,492	1,539	3,031
I22 death (subsequent myocardial infarction)	4	2	6
I23 death (complications following myocardial infarction)	0	0	0
I24 death (other acute CHD)	53	58	111
I25 death (chronic CHD)	407	340	747
CHD deaths (primary cause on the death certificate, I20-I25)	1,437	1,490	2,927
I20 death (angina pectoris)	3	4	7
I21 death (acute myocardial infarction)	1,304	1,329	2,633
I22 death (subsequent myocardial infarction)	4	2	6
I23 death (complications following myocardial infarction)	0	0	0
I24 death (other acute CHD)	24	29	53
I25 death (chronic CHD)	102	126	228

Table 4.5: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Tada *et al.*

Tada *et al.*/PGS000011/ European/44 SNPs[17]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	51.3 (11.9)	51.3 (11.9)	51.2 (11.9)	51.2 (11.8)	51.2 (11.7)	51.2 (11.8)
Resident of Coyoacán	10,079 (38%)	10,295 (39%)	10,077 (38%)	10,257 (39%)	10,307 (39%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	67.1 (17.5)	66.8 (17.9)	66.7 (18.0)	67.0 (18.0)	66.6 (17.8)	66.8 (17.9)
African	3.5 (3.0)	3.5 (2.9)	3.4 (2.8)	3.4 (2.7)	3.4 (2.7)	3.4 (2.8)
East Asian	1.4 (1.8)	1.4 (1.6)	1.4 (2.0)	1.4 (1.8)	1.4 (1.9)	1.4 (1.8)
European	27.9 (15.8)	28.4 (16.3)	28.4 (16.3)	28.2 (16.3)	28.6 (16.1)	28.3 (16.2)
Highest attained educational level						
University/high school	4,091 (15%)	4,164 (16%)	4,143 (16%)	4,183 (16%)	4,128 (16%)	20,709 (16%)
Middle school	6,695 (25%)	6,776 (25%)	6,560 (25%)	6,613 (25%)	6,788 (25%)	33,432 (25%)
Elementary	12,557 (47%)	12,444 (47%)	12,679 (48%)	12,668 (48%)	12,559 (47%)	62,907 (47%)
Other	3,282 (12%)	3,244 (12%)	3,244 (12%)	3,167 (12%)	3,154 (12%)	16,091 (12%)
Missing	17 (0%)	13 (0%)	15 (0%)	10 (0%)	13 (0%)	68 (0%)
Smoking status						
Never	12,874 (48%)	12,888 (48%)	12,929 (49%)	13,051 (49%)	12,849 (48%)	64,591 (48%)
Former	5,199 (20%)	5,156 (19%)	5,225 (20%)	5,213 (20%)	5,383 (20%)	26,176 (20%)
Current	8,545 (32%)	8,576 (32%)	8,462 (32%)	8,362 (31%)	8,385 (32%)	42,330 (32%)
Missing	24 (0%)	21 (0%)	25 (0%)	15 (0%)	25 (0%)	110 (0%)
Alcohol intake						
Never	5,082 (19%)	5,384 (20%)	5,245 (20%)	5,268 (20%)	5,124 (19%)	26,103 (20%)
Former	3,701 (14%)	3,571 (13%)	3,683 (14%)	3,605 (14%)	3,691 (14%)	18,251 (14%)
Current	17,847 (67%)	17,680 (66%)	17,699 (66%)	17,759 (67%)	17,818 (67%)	88,803 (67%)
Missing	12 (0%)	6 (0%)	14 (0%)	9 (0%)	9 (0%)	50 (0%)
Physical measures						
SBP, mmHg	127.0 (16.5)	127.1 (16.6)	127.3 (16.6)	127.2 (16.6)	127.6 (16.7)	127.2 (16.6)
DBP, mmHg	83.0 (10.2)	83.0 (10.1)	83.1 (10.1)	83.2 (10.2)	83.2 (10.3)	83.1 (10.2)
BMI, kg/m²	29.2 (5.1)	29.1 (5.1)	29.1 (5.1)	29.1 (5.0)	29.1 (5.0)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)
LDL-C, mmol/L	2.45 (0.79)	2.47 (0.79)	2.46 (0.79)	2.46 (0.79)	2.45 (0.79)	2.46 (0.79)
Triglycerides, mmol/L	1.56 (0.65)	1.57 (0.65)	1.57 (0.66)	1.58 (0.66)	1.58 (0.66)	1.57 (0.66)
HbA1c, %	6.09 (1.70)	6.09 (1.70)	6.09 (1.71)	6.09 (1.70)	6.12 (1.75)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	101.7 (16.1)	101.6 (16.2)	101.7 (15.9)	101.6 (15.9)	101.6 (16.1)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	375 (1%)	380 (1%)	346 (1%)	386 (1%)	414 (2%)	1,901 (1%)
Stroke	249 (1%)	292 (1%)	295 (1%)	287 (1%)	291 (1%)	1,414 (1%)
Cancer	335 (1%)	318 (1%)	317 (1%)	315 (1%)	295 (1%)	1,580 (1%)
Diabetes[†]	4,955 (19%)	4,901 (18%)	4,900 (18%)	4,938 (19%)	5,102 (19%)	24,796 (19%)
Other[‡]	2,197 (8%)	2,204 (8%)	2,166 (8%)	2,282 (9%)	2,264 (8%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.6: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Oni-Orisan *et al.*

Oni-Orisan *et al.*/PGS004595/ European/141 SNPs[11]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	51.3 (11.9)	51.3 (11.8)	51.2 (11.8)	51.1 (11.8)	51.3 (11.8)	51.2 (11.8)
Resident of Coyoacán	9,883 (37%)	10,000 (38%)	10,221 (38%)	10,280 (39%)	10,631 (40%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	69.0 (17.5)	68.4 (17.8)	67.5 (17.8)	66.2 (17.9)	63.1 (17.7)	66.8 (17.9)
African	3.2 (2.6)	3.3 (2.7)	3.4 (2.8)	3.5 (2.9)	3.8 (3.0)	3.4 (2.8)
East Asian	1.4 (1.5)	1.4 (1.5)	1.4 (1.4)	1.4 (2.1)	1.5 (2.3)	1.4 (1.8)
European	26.4 (15.8)	26.9 (16.1)	27.7 (16.0)	28.8 (16.2)	31.6 (16.1)	28.3 (16.2)
Highest attained educational level						
University/high school	3,932 (15%)	4,007 (15%)	4,029 (15%)	4,224 (16%)	4,517 (17%)	20,709 (16%)
Middle school	6,629 (25%)	6,577 (25%)	6,696 (25%)	6,659 (25%)	6,871 (26%)	33,432 (25%)
Elementary	12,677 (48%)	12,718 (48%)	12,582 (47%)	12,651 (48%)	12,279 (46%)	62,907 (47%)
Other	3,391 (13%)	3,323 (12%)	3,320 (12%)	3,094 (12%)	2,963 (11%)	16,091 (12%)
Missing	13 (0%)	16 (0%)	14 (0%)	13 (0%)	12 (0%)	68 (0%)
Smoking status						
Never	13,096 (49%)	13,115 (49%)	13,046 (49%)	12,857 (48%)	12,477 (47%)	64,591 (48%)
Former	5,300 (20%)	5,223 (20%)	5,202 (20%)	5,164 (19%)	5,287 (20%)	26,176 (20%)
Current	8,232 (31%)	8,285 (31%)	8,368 (31%)	8,593 (32%)	8,852 (33%)	42,330 (32%)
Missing	14 (0%)	18 (0%)	25 (0%)	27 (0%)	26 (0%)	110 (0%)
Alcohol intake						
Never	5,202 (20%)	5,271 (20%)	5,261 (20%)	5,325 (20%)	5,044 (19%)	26,103 (20%)
Former	3,681 (14%)	3,747 (14%)	3,625 (14%)	3,563 (13%)	3,635 (14%)	18,251 (14%)
Current	17,745 (67%)	17,614 (66%)	17,744 (67%)	17,748 (67%)	17,952 (67%)	88,803 (67%)
Missing	14 (0%)	9 (0%)	11 (0%)	5 (0%)	11 (0%)	50 (0%)
Physical measures						
SBP, mmHg	126.6 (16.6)	127.0 (16.6)	127.4 (16.6)	127.3 (16.5)	127.8 (16.6)	127.2 (16.6)
DBP, mmHg	82.8 (10.2)	83.0 (10.1)	83.2 (10.3)	83.2 (10.2)	83.4 (10.2)	83.1 (10.2)
BMI, kg/m²	29.2 (5.1)	29.2 (5.1)	29.1 (5.1)	29.1 (5.1)	29.0 (5.0)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)
LDL-C, mmol/L	2.43 (0.78)	2.45 (0.79)	2.45 (0.79)	2.47 (0.79)	2.49 (0.80)	2.46 (0.79)
Triglycerides, mmol/L	1.56 (0.65)	1.57 (0.66)	1.57 (0.66)	1.58 (0.66)	1.59 (0.67)	1.57 (0.66)
HbA1c, %	6.09 (1.69)	6.12 (1.73)	6.10 (1.72)	6.09 (1.72)	6.08 (1.70)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	101.9 (16.1)	101.8 (16.0)	101.6 (16.3)	101.6 (16.0)	101.4 (15.8)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	345 (1%)	306 (1%)	385 (1%)	396 (1%)	469 (2%)	1,901 (1%)
Stroke	259 (1%)	245 (1%)	322 (1%)	308 (1%)	280 (1%)	1,414 (1%)
Cancer	303 (1%)	326 (1%)	312 (1%)	320 (1%)	319 (1%)	1,580 (1%)
Diabetes[†]	4,903 (18%)	5,078 (19%)	5,027 (19%)	4,895 (18%)	4,893 (18%)	24,796 (19%)
Other[‡]	2,202 (8%)	2,151 (8%)	2,309 (9%)	2,245 (8%)	2,206 (8%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.7: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Koyama *et al.*

Koyama *et al.*/PGS000337/ Japanese and European/64,185 SNPs[12]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	51.7 (12.0)	51.4 (11.9)	51.2 (11.8)	51.0 (11.8)	50.9 (11.6)	51.2 (11.8)
Resident of Coyoacán	10,764 (40%)	10,244 (38%)	10,137 (38%)	10,027 (38%)	9,843 (37%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	62.6 (18.2)	65.9 (17.8)	67.4 (17.7)	68.6 (17.5)	69.7 (17.3)	66.8 (17.9)
African	3.7 (2.7)	3.5 (2.8)	3.4 (2.7)	3.3 (2.8)	3.3 (3.0)	3.4 (2.8)
East Asian	1.4 (1.7)	1.4 (1.6)	1.4 (1.9)	1.4 (1.7)	1.4 (2.1)	1.4 (1.8)
European	32.3 (16.7)	29.2 (16.2)	27.7 (16.0)	26.6 (15.7)	25.6 (15.4)	28.3 (16.2)
Highest attained educational level						
University/high school	4,846 (18%)	4,238 (16%)	4,102 (15%)	3,910 (15%)	3,613 (14%)	20,709 (16%)
Middle school	6,753 (25%)	6,774 (25%)	6,733 (25%)	6,641 (25%)	6,531 (25%)	33,432 (25%)
Elementary	12,110 (45%)	12,472 (47%)	12,567 (47%)	12,775 (48%)	12,983 (49%)	62,907 (47%)
Other	2,921 (11%)	3,146 (12%)	3,225 (12%)	3,296 (12%)	3,503 (13%)	16,091 (12%)
Missing	12 (0%)	11 (0%)	14 (0%)	19 (0%)	12 (0%)	68 (0%)
Smoking status						
Never	12,947 (49%)	13,096 (49%)	12,831 (48%)	12,915 (49%)	12,802 (48%)	64,591 (48%)
Former	5,258 (20%)	5,191 (20%)	5,356 (20%)	5,162 (19%)	5,209 (20%)	26,176 (20%)
Current	8,419 (32%)	8,332 (31%)	8,429 (32%)	8,539 (32%)	8,611 (32%)	42,330 (32%)
Missing	18 (0%)	22 (0%)	25 (0%)	25 (0%)	20 (0%)	110 (0%)
Alcohol intake						
Never	5,023 (19%)	5,300 (20%)	5,166 (19%)	5,293 (20%)	5,321 (20%)	26,103 (20%)
Former	3,495 (13%)	3,583 (13%)	3,590 (13%)	3,732 (14%)	3,851 (14%)	18,251 (14%)
Current	18,115 (68%)	17,746 (67%)	17,876 (67%)	17,608 (66%)	17,458 (66%)	88,803 (67%)
Missing	9 (0%)	12 (0%)	9 (0%)	8 (0%)	12 (0%)	50 (0%)
Physical measures						
SBP, mmHg	126.2 (16.2)	126.9 (16.6)	127.3 (16.5)	127.6 (16.8)	128.2 (16.9)	127.2 (16.6)
DBP, mmHg	82.5 (10.0)	82.9 (10.2)	83.1 (10.2)	83.3 (10.2)	83.6 (10.3)	83.1 (10.2)
BMI, kg/m²	29.0 (5.1)	29.1 (5.1)	29.1 (5.0)	29.2 (5.1)	29.2 (5.1)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.01 (0.22)	1.01 (0.21)	1.00 (0.21)	1.00 (0.21)	0.99 (0.21)	1.00 (0.21)
LDL-C, mmol/L	2.45 (0.78)	2.45 (0.78)	2.46 (0.79)	2.46 (0.80)	2.46 (0.80)	2.46 (0.79)
Triglycerides, mmol/L	1.52 (0.63)	1.56 (0.65)	1.58 (0.66)	1.59 (0.66)	1.62 (0.68)	1.57 (0.66)
HbA1c, %	5.98 (1.58)	6.06 (1.66)	6.10 (1.72)	6.13 (1.75)	6.22 (1.83)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	101.2 (15.9)	101.5 (16.0)	101.7 (16.1)	101.9 (16.1)	102.0 (16.0)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	325 (1%)	330 (1%)	357 (1%)	396 (1%)	493 (2%)	1,901 (1%)
Stroke	252 (1%)	279 (1%)	284 (1%)	263 (1%)	336 (1%)	1,414 (1%)
Cancer	336 (1%)	357 (1%)	285 (1%)	318 (1%)	284 (1%)	1,580 (1%)
Diabetes[†]	4,208 (16%)	4,731 (18%)	4,980 (19%)	5,164 (19%)	5,713 (21%)	24,796 (19%)
Other[‡]	2,392 (9%)	2,242 (8%)	2,237 (8%)	2,151 (8%)	2,091 (8%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.8: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Tcheandjieu *et al.*

Tcheandjieu *et al.*/PGS003446/ Multi-ancestry/485,464 SNPs[14]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	51.0 (11.8)	51.0 (11.8)	51.0 (11.8)	51.3 (11.8)	51.9 (12.0)	51.2 (11.8)
Resident of Coyoacán	8,908 (33%)	9,590 (36%)	10,094 (38%)	10,650 (40%)	11,773 (44%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	80.2 (14.4)	73.0 (15.4)	67.0 (15.5)	61.3 (15.2)	52.7 (15.4)	66.8 (17.9)
African	2.4 (2.9)	3.0 (2.8)	3.5 (2.6)	3.9 (2.6)	4.4 (2.6)	3.4 (2.8)
East Asian	1.3 (2.4)	1.4 (1.9)	1.5 (2.0)	1.5 (1.3)	1.5 (1.3)	1.4 (1.8)
European	16.1 (12.2)	22.6 (13.5)	28.0 (13.8)	33.4 (13.9)	41.3 (14.8)	28.3 (16.2)
Highest attained educational level						
University/high school	3,238 (12%)	3,697 (14%)	4,117 (15%)	4,400 (17%)	5,257 (20%)	20,709 (16%)
Middle school	6,051 (23%)	6,469 (24%)	6,823 (26%)	7,013 (26%)	7,076 (27%)	33,432 (25%)
Elementary	13,386 (50%)	13,054 (49%)	12,564 (47%)	12,299 (46%)	11,604 (44%)	62,907 (47%)
Other	3,951 (15%)	3,411 (13%)	3,121 (12%)	2,915 (11%)	2,693 (10%)	16,091 (12%)
Missing	16 (0%)	10 (0%)	16 (0%)	14 (0%)	12 (0%)	68 (0%)
Smoking status						
Never	14,516 (55%)	13,488 (51%)	12,956 (49%)	12,215 (46%)	11,416 (43%)	64,591 (48%)
Former	5,093 (19%)	5,151 (19%)	5,305 (20%)	5,260 (20%)	5,367 (20%)	26,176 (20%)
Current	7,011 (26%)	7,984 (30%)	8,359 (31%)	9,145 (34%)	9,831 (37%)	42,330 (32%)
Missing	22 (0%)	18 (0%)	21 (0%)	21 (0%)	28 (0%)	110 (0%)
Alcohol intake						
Never	5,649 (21%)	5,233 (20%)	5,163 (19%)	5,127 (19%)	4,931 (19%)	26,103 (20%)
Former	3,658 (14%)	3,712 (14%)	3,668 (14%)	3,548 (13%)	3,665 (14%)	18,251 (14%)
Current	17,330 (65%)	17,685 (66%)	17,799 (67%)	17,955 (67%)	18,034 (68%)	88,803 (67%)
Missing	5 (0%)	11 (0%)	11 (0%)	11 (0%)	12 (0%)	50 (0%)
Physical measures						
SBP, mmHg	126.2 (16.2)	126.8 (16.4)	127.3 (16.6)	127.6 (16.7)	128.3 (17.0)	127.2 (16.6)
DBP, mmHg	82.5 (9.9)	82.9 (10.1)	83.2 (10.2)	83.3 (10.3)	83.6 (10.4)	83.1 (10.2)
BMI, kg/m²	29.0 (4.9)	29.1 (5.0)	29.2 (5.0)	29.2 (5.1)	29.1 (5.3)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.07)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.22)	1.00 (0.21)
LDL-C, mmol/L	2.39 (0.77)	2.43 (0.78)	2.45 (0.78)	2.48 (0.78)	2.53 (0.82)	2.46 (0.79)
Triglycerides, mmol/L	1.57 (0.64)	1.57 (0.66)	1.58 (0.66)	1.57 (0.65)	1.58 (0.67)	1.57 (0.66)
HbA1c, %	6.16 (1.77)	6.12 (1.72)	6.10 (1.71)	6.08 (1.71)	6.04 (1.66)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	102.9 (15.6)	102.4 (15.9)	101.8 (16.0)	101.2 (16.1)	99.9 (16.4)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	249 (1%)	275 (1%)	355 (1%)	450 (2%)	572 (2%)	1,901 (1%)
Stroke	267 (1%)	289 (1%)	269 (1%)	276 (1%)	313 (1%)	1,414 (1%)
Cancer	278 (1%)	295 (1%)	282 (1%)	335 (1%)	390 (1%)	1,580 (1%)
Diabetes[†]	5,066 (19%)	5,053 (19%)	5,033 (19%)	4,816 (18%)	4,828 (18%)	24,796 (19%)
Other[‡]	1,880 (7%)	2,106 (8%)	2,180 (8%)	2,346 (9%)	2,601 (10%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.9: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Tamlander *et al.*

Tamlander *et al.*/ PGS001780/ European/1,087,958 SNPs[13]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	52.0 (12.0)	51.4 (11.9)	51.0 (11.8)	51.0 (11.8)	50.7 (11.7)	51.2 (11.8)
Resident of Coyoacán	11,363 (43%)	10,657 (40%)	10,099 (38%)	9,634 (36%)	9,262 (35%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	56.7 (16.9)	63.7 (17.0)	67.6 (17.0)	71.1 (16.8)	75.1 (15.9)	66.8 (17.9)
African	4.3 (3.1)	3.8 (2.8)	3.4 (2.7)	3.1 (2.6)	2.6 (2.4)	3.4 (2.8)
East Asian	1.5 (1.5)	1.4 (1.7)	1.4 (1.8)	1.4 (2.0)	1.4 (2.1)	1.4 (1.8)
European	37.5 (15.8)	31.1 (15.5)	27.5 (15.3)	24.5 (14.9)	20.9 (14.1)	28.3 (16.2)
Highest attained educational level						
University/high school	5,318 (20%)	4,523 (17%)	4,002 (15%)	3,605 (14%)	3,261 (12%)	20,709 (16%)
Middle school	7,075 (27%)	6,861 (26%)	6,787 (25%)	6,478 (24%)	6,231 (23%)	33,432 (25%)
Elementary	11,578 (43%)	12,270 (46%)	12,655 (48%)	13,087 (49%)	13,317 (50%)	62,907 (47%)
Other	2,654 (10%)	2,975 (11%)	3,190 (12%)	3,459 (13%)	3,813 (14%)	16,091 (12%)
Missing	17 (0%)	12 (0%)	7 (0%)	12 (0%)	20 (0%)	68 (0%)
Smoking status						
Never	12,389 (47%)	12,766 (48%)	12,950 (49%)	13,056 (49%)	13,430 (50%)	64,591 (48%)
Former	5,382 (20%)	5,189 (19%)	5,207 (20%)	5,305 (20%)	5,093 (19%)	26,176 (20%)
Current	8,846 (33%)	8,661 (33%)	8,464 (32%)	8,259 (31%)	8,100 (30%)	42,330 (32%)
Missing	25 (0%)	25 (0%)	20 (0%)	21 (0%)	19 (0%)	110 (0%)
Alcohol intake						
Never	4,887 (18%)	5,091 (19%)	5,212 (20%)	5,389 (20%)	5,524 (21%)	26,103 (20%)
Former	3,494 (13%)	3,527 (13%)	3,621 (14%)	3,769 (14%)	3,840 (14%)	18,251 (14%)
Current	18,249 (69%)	18,012 (68%)	17,800 (67%)	17,477 (66%)	17,265 (65%)	88,803 (67%)
Missing	12 (0%)	11 (0%)	8 (0%)	6 (0%)	13 (0%)	50 (0%)
Physical measures						
SBP, mmHg	126.7 (16.4)	127.1 (16.4)	127.0 (16.6)	127.6 (16.7)	127.9 (16.8)	127.2 (16.6)
DBP, mmHg	82.9 (10.1)	83.0 (10.2)	83.0 (10.2)	83.3 (10.2)	83.4 (10.2)	83.1 (10.2)
BMI, kg/m²	29.0 (5.2)	29.1 (5.1)	29.1 (5.1)	29.2 (5.0)	29.2 (5.0)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.91 (0.07)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.01 (0.22)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	0.99 (0.21)	1.00 (0.21)
LDL-C, mmol/L	2.47 (0.78)	2.47 (0.78)	2.46 (0.79)	2.45 (0.80)	2.45 (0.80)	2.46 (0.79)
Triglycerides, mmol/L	1.53 (0.64)	1.56 (0.65)	1.57 (0.66)	1.59 (0.66)	1.63 (0.68)	1.57 (0.66)
HbA1c, %	5.95 (1.55)	6.03 (1.63)	6.10 (1.72)	6.15 (1.76)	6.26 (1.87)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	100.5 (16.0)	101.4 (16.0)	101.8 (16.0)	102.1 (16.1)	102.4 (16.0)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	379 (1%)	367 (1%)	357 (1%)	384 (1%)	414 (2%)	1,901 (1%)
Stroke	255 (1%)	279 (1%)	300 (1%)	278 (1%)	302 (1%)	1,414 (1%)
Cancer	367 (1%)	329 (1%)	301 (1%)	308 (1%)	275 (1%)	1,580 (1%)
Diabetes[†]	4,096 (15%)	4,624 (17%)	4,952 (19%)	5,287 (20%)	5,837 (22%)	24,796 (19%)
Other[‡]	2,504 (9%)	2,352 (9%)	2,233 (8%)	2,135 (8%)	1,889 (7%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.10: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Patel *et al.*

Patel *et al.*/PGS003725/ Multi-ancestry/1,273,824 SNPs[16]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	51.5 (12.0)	51.4 (11.9)	51.3 (11.9)	51.1 (11.7)	50.9 (11.6)	51.2 (11.8)
Resident of Coyoacán	9,980 (37%)	10,079 (38%)	10,204 (38%)	10,162 (38%)	10,590 (40%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	70.4 (17.7)	68.8 (17.7)	67.5 (17.7)	65.6 (17.7)	62.0 (17.4)	66.8 (17.9)
African	2.9 (2.3)	3.2 (2.5)	3.4 (2.7)	3.6 (2.8)	4.1 (3.4)	3.4 (2.8)
East Asian	1.3 (1.5)	1.4 (1.7)	1.4 (1.9)	1.4 (1.7)	1.5 (2.3)	1.4 (1.8)
European	25.4 (16.0)	26.7 (16.1)	27.7 (16.0)	29.4 (16.0)	32.3 (15.9)	28.3 (16.2)
Highest attained educational level						
University/high school	4,224 (16%)	4,076 (15%)	4,110 (15%)	4,080 (15%)	4,219 (16%)	20,709 (16%)
Middle school	6,441 (24%)	6,761 (25%)	6,658 (25%)	6,691 (25%)	6,881 (26%)	33,432 (25%)
Elementary	12,709 (48%)	12,498 (47%)	12,576 (47%)	12,642 (47%)	12,482 (47%)	62,907 (47%)
Other	3,256 (12%)	3,292 (12%)	3,278 (12%)	3,218 (12%)	3,047 (11%)	16,091 (12%)
Missing	12 (0%)	14 (0%)	19 (0%)	10 (0%)	13 (0%)	68 (0%)
Smoking status						
Never	13,720 (52%)	13,227 (50%)	12,869 (48%)	12,634 (47%)	12,141 (46%)	64,591 (48%)
Former	5,166 (19%)	5,245 (20%)	5,302 (20%)	5,224 (20%)	5,239 (20%)	26,176 (20%)
Current	7,732 (29%)	8,149 (31%)	8,450 (32%)	8,758 (33%)	9,241 (35%)	42,330 (32%)
Missing	24 (0%)	20 (0%)	20 (0%)	25 (0%)	21 (0%)	110 (0%)
Alcohol intake						
Never	5,127 (19%)	5,196 (20%)	5,336 (20%)	5,331 (20%)	5,113 (19%)	26,103 (20%)
Former	3,616 (14%)	3,599 (14%)	3,608 (14%)	3,688 (14%)	3,740 (14%)	18,251 (14%)
Current	17,890 (67%)	17,834 (67%)	17,690 (66%)	17,611 (66%)	17,778 (67%)	88,803 (67%)
Missing	9 (0%)	12 (0%)	7 (0%)	11 (0%)	11 (0%)	50 (0%)
Physical measures						
SBP, mmHg	125.8 (16.1)	126.6 (16.4)	127.2 (16.4)	127.9 (16.9)	128.6 (17.0)	127.2 (16.6)
DBP, mmHg	82.3 (10.1)	82.7 (10.1)	83.2 (10.1)	83.4 (10.3)	83.9 (10.4)	83.1 (10.2)
BMI, kg/m²	28.9 (5.0)	29.1 (5.0)	29.1 (5.1)	29.2 (5.1)	29.3 (5.2)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.01 (0.22)	1.01 (0.21)	1.00 (0.21)	1.00 (0.21)	0.99 (0.21)	1.00 (0.21)
LDL-C, mmol/L	2.40 (0.76)	2.44 (0.78)	2.46 (0.79)	2.48 (0.80)	2.51 (0.82)	2.46 (0.79)
Triglycerides, mmol/L	1.50 (0.61)	1.55 (0.64)	1.58 (0.66)	1.60 (0.68)	1.63 (0.70)	1.57 (0.66)
HbA1c, %	5.98 (1.58)	6.05 (1.64)	6.09 (1.69)	6.16 (1.78)	6.22 (1.84)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	102.0 (15.7)	101.9 (15.8)	101.5 (16.2)	101.6 (16.1)	101.3 (16.4)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	285 (1%)	324 (1%)	332 (1%)	389 (1%)	571 (2%)	1,901 (1%)
Stroke	233 (1%)	259 (1%)	310 (1%)	280 (1%)	332 (1%)	1,414 (1%)
Cancer	312 (1%)	321 (1%)	316 (1%)	319 (1%)	312 (1%)	1,580 (1%)
Diabetes[†]	4,195 (16%)	4,583 (17%)	4,922 (18%)	5,353 (20%)	5,743 (22%)	24,796 (19%)
Other[‡]	2,253 (8%)	2,208 (8%)	2,193 (8%)	2,269 (9%)	2,190 (8%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.11: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Inouye *et al.*

Inouye *et al.*/PGS000018/ European/1,720,068 SNPs[15]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	52.4 (12.1)	51.4 (11.8)	51.0 (11.8)	50.8 (11.8)	50.7 (11.7)	51.2 (11.8)
Resident of Coyoacán	11,967 (45%)	10,593 (40%)	10,046 (38%)	9,484 (36%)	8,925 (33%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	51.8 (15.1)	61.7 (15.2)	67.8 (15.6)	73.3 (15.3)	79.6 (14.3)	66.8 (17.9)
African	4.6 (2.8)	4.0 (2.8)	3.5 (2.8)	2.9 (2.6)	2.3 (2.3)	3.4 (2.8)
East Asian	1.5 (1.5)	1.5 (1.6)	1.5 (1.9)	1.4 (1.8)	1.3 (2.2)	1.4 (1.8)
European	42.1 (14.4)	32.8 (13.8)	27.3 (13.9)	22.4 (13.5)	16.9 (12.4)	28.3 (16.2)
Highest attained educational level						
University/high school	5,825 (22%)	4,492 (17%)	4,008 (15%)	3,480 (13%)	2,904 (11%)	20,709 (16%)
Middle school	7,149 (27%)	7,077 (27%)	6,792 (26%)	6,366 (24%)	6,048 (23%)	33,432 (25%)
Elementary	11,220 (42%)	12,188 (46%)	12,757 (48%)	13,209 (50%)	13,533 (51%)	62,907 (47%)
Other	2,440 (9%)	2,872 (11%)	3,072 (12%)	3,568 (13%)	4,139 (16%)	16,091 (12%)
Missing	8 (0%)	12 (0%)	12 (0%)	18 (0%)	18 (0%)	68 (0%)
Smoking status						
Never	11,982 (45%)	12,541 (47%)	12,966 (49%)	13,351 (50%)	13,751 (52%)	64,591 (48%)
Former	5,364 (20%)	5,314 (20%)	5,198 (20%)	5,200 (20%)	5,100 (19%)	26,176 (20%)
Current	9,268 (35%)	8,771 (33%)	8,454 (32%)	8,067 (30%)	7,770 (29%)	42,330 (32%)
Missing	28 (0%)	15 (0%)	23 (0%)	23 (0%)	21 (0%)	110 (0%)
Alcohol intake						
Never	4,730 (18%)	5,224 (20%)	5,215 (20%)	5,404 (20%)	5,530 (21%)	26,103 (20%)
Former	3,379 (13%)	3,567 (13%)	3,661 (14%)	3,708 (14%)	3,936 (15%)	18,251 (14%)
Current	18,520 (70%)	17,840 (67%)	17,757 (67%)	17,519 (66%)	17,167 (64%)	88,803 (67%)
Missing	13 (0%)	10 (0%)	8 (0%)	10 (0%)	9 (0%)	50 (0%)
Physical measures						
SBP, mmHg	126.9 (16.5)	127.1 (16.4)	127.0 (16.6)	127.5 (16.8)	127.7 (16.7)	127.2 (16.6)
DBP, mmHg	83.0 (10.1)	83.1 (10.2)	83.1 (10.2)	83.2 (10.2)	83.3 (10.2)	83.1 (10.2)
BMI, kg/m²	28.9 (5.2)	29.1 (5.1)	29.1 (5.1)	29.2 (5.0)	29.2 (4.9)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.91 (0.07)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.01 (0.22)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	0.99 (0.21)	1.00 (0.21)
LDL-C, mmol/L	2.50 (0.78)	2.48 (0.80)	2.45 (0.79)	2.44 (0.78)	2.42 (0.79)	2.46 (0.79)
Triglycerides, mmol/L	1.52 (0.63)	1.56 (0.66)	1.57 (0.66)	1.60 (0.67)	1.62 (0.67)	1.57 (0.66)
HbA1c, %	5.90 (1.48)	6.04 (1.66)	6.10 (1.72)	6.16 (1.76)	6.28 (1.89)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	99.8 (16.1)	101.2 (16.0)	102.0 (15.9)	102.3 (16.1)	103.0 (15.9)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	432 (2%)	382 (1%)	364 (1%)	332 (1%)	391 (1%)	1,901 (1%)
Stroke	255 (1%)	267 (1%)	298 (1%)	285 (1%)	309 (1%)	1,414 (1%)
Cancer	384 (1%)	348 (1%)	318 (1%)	263 (1%)	267 (1%)	1,580 (1%)
Diabetes[†]	3,898 (15%)	4,589 (17%)	4,959 (19%)	5,418 (20%)	5,932 (22%)	24,796 (19%)
Other[‡]	2,664 (10%)	2,319 (9%)	2,231 (8%)	2,029 (8%)	1,870 (7%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 4.12: Baseline characteristics of 133,207 participants aged 35-79 by fifth of PRS Khera *et al.*

Khera *et al.*/PGS000013/ European/6,472,620 SNPs[18]	Fifth of the PRS distribution					All n=133,207
	1 n=26,642 (20%)	2 n=26,641 (20%)	3 n=26,641 (20%)	4 n=26,641 (20%)	5 n=26,642 (20%)	
Age, years	51.4 (11.9)	51.2 (11.8)	51.3 (11.8)	51.1 (11.8)	51.2 (11.8)	51.2 (11.8)
Resident of Coyoacán	10,270 (39%)	10,129 (38%)	10,096 (38%)	10,227 (38%)	10,293 (39%)	51,015 (38%)
Ancestry admixture percentage						
Indigenous American	66.1 (17.7)	67.9 (17.9)	67.9 (17.8)	67.2 (17.9)	65.1 (17.8)	66.8 (17.9)
African	3.7 (3.4)	3.3 (2.7)	3.3 (2.6)	3.4 (2.6)	3.5 (2.6)	3.4 (2.8)
East Asian	1.4 (1.4)	1.4 (1.6)	1.4 (1.6)	1.4 (1.9)	1.5 (2.4)	1.4 (1.8)
European	28.7 (16.0)	27.4 (16.2)	27.4 (16.1)	28.0 (16.2)	29.9 (16.2)	28.3 (16.2)
Highest attained educational level						
University/high school	4,272 (16%)	4,068 (15%)	4,091 (15%)	4,108 (15%)	4,170 (16%)	20,709 (16%)
Middle school	6,707 (25%)	6,746 (25%)	6,642 (25%)	6,683 (25%)	6,654 (25%)	33,432 (25%)
Elementary	12,395 (47%)	12,676 (48%)	12,604 (47%)	12,657 (48%)	12,575 (47%)	62,907 (47%)
Other	3,251 (12%)	3,143 (12%)	3,291 (12%)	3,176 (12%)	3,230 (12%)	16,091 (12%)
Missing	17 (0%)	8 (0%)	13 (0%)	17 (0%)	13 (0%)	68 (0%)
Smoking status						
Never	13,064 (49%)	13,057 (49%)	13,048 (49%)	12,703 (48%)	12,719 (48%)	64,591 (48%)
Former	5,294 (20%)	5,132 (19%)	5,187 (19%)	5,256 (20%)	5,307 (20%)	26,176 (20%)
Current	8,264 (31%)	8,437 (32%)	8,382 (31%)	8,654 (33%)	8,593 (32%)	42,330 (32%)
Missing	20 (0%)	15 (0%)	24 (0%)	28 (0%)	23 (0%)	110 (0%)
Alcohol intake						
Never	5,175 (19%)	5,122 (19%)	5,258 (20%)	5,300 (20%)	5,248 (20%)	26,103 (20%)
Former	3,641 (14%)	3,617 (14%)	3,635 (14%)	3,659 (14%)	3,699 (14%)	18,251 (14%)
Current	17,815 (67%)	17,896 (67%)	17,737 (67%)	17,670 (66%)	17,685 (66%)	88,803 (67%)
Missing	11 (0%)	6 (0%)	11 (0%)	12 (0%)	10 (0%)	50 (0%)
Physical measures						
SBP, mmHg	126.7 (16.6)	126.9 (16.3)	127.3 (16.6)	127.3 (16.7)	128.0 (16.8)	127.2 (16.6)
DBP, mmHg	82.9 (10.1)	83.0 (10.2)	83.1 (10.2)	83.2 (10.2)	83.4 (10.3)	83.1 (10.2)
BMI, kg/m²	29.2 (5.2)	29.1 (5.0)	29.1 (5.0)	29.1 (5.1)	29.0 (5.1)	29.1 (5.1)
Waist-to-hip Ratio	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)
Laboratory measurements						
HDL-C, mmol/L	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.21)
LDL-C, mmol/L	2.42 (0.77)	2.44 (0.79)	2.45 (0.79)	2.46 (0.80)	2.51 (0.80)	2.46 (0.79)
Triglycerides, mmol/L	1.55 (0.64)	1.57 (0.65)	1.58 (0.67)	1.58 (0.66)	1.59 (0.67)	1.57 (0.66)
HbA1c, %	6.05 (1.66)	6.09 (1.70)	6.10 (1.71)	6.11 (1.73)	6.13 (1.76)	6.10 (1.71)
eGFR, ml/min/1.73m²[§]	101.5 (15.9)	101.8 (16.2)	101.6 (16.4)	101.9 (15.8)	101.5 (15.9)	101.7 (16.0)
Prior disease[*]						
Coronary heart disease	314 (1%)	334 (1%)	328 (1%)	414 (2%)	511 (2%)	1,901 (1%)
Stroke	261 (1%)	269 (1%)	302 (1%)	284 (1%)	298 (1%)	1,414 (1%)
Cancer	318 (1%)	326 (1%)	314 (1%)	304 (1%)	318 (1%)	1,580 (1%)
Diabetes[†]	4,600 (17%)	4,969 (19%)	4,966 (19%)	5,037 (19%)	5,224 (20%)	24,796 (19%)
Other[‡]	2,308 (9%)	2,195 (8%)	2,200 (8%)	2,202 (8%)	2,208 (8%)	11,113 (8%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

[*]Self-reported previous diagnoses unless otherwise stated.

[†]Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

[‡]Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

[§]Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

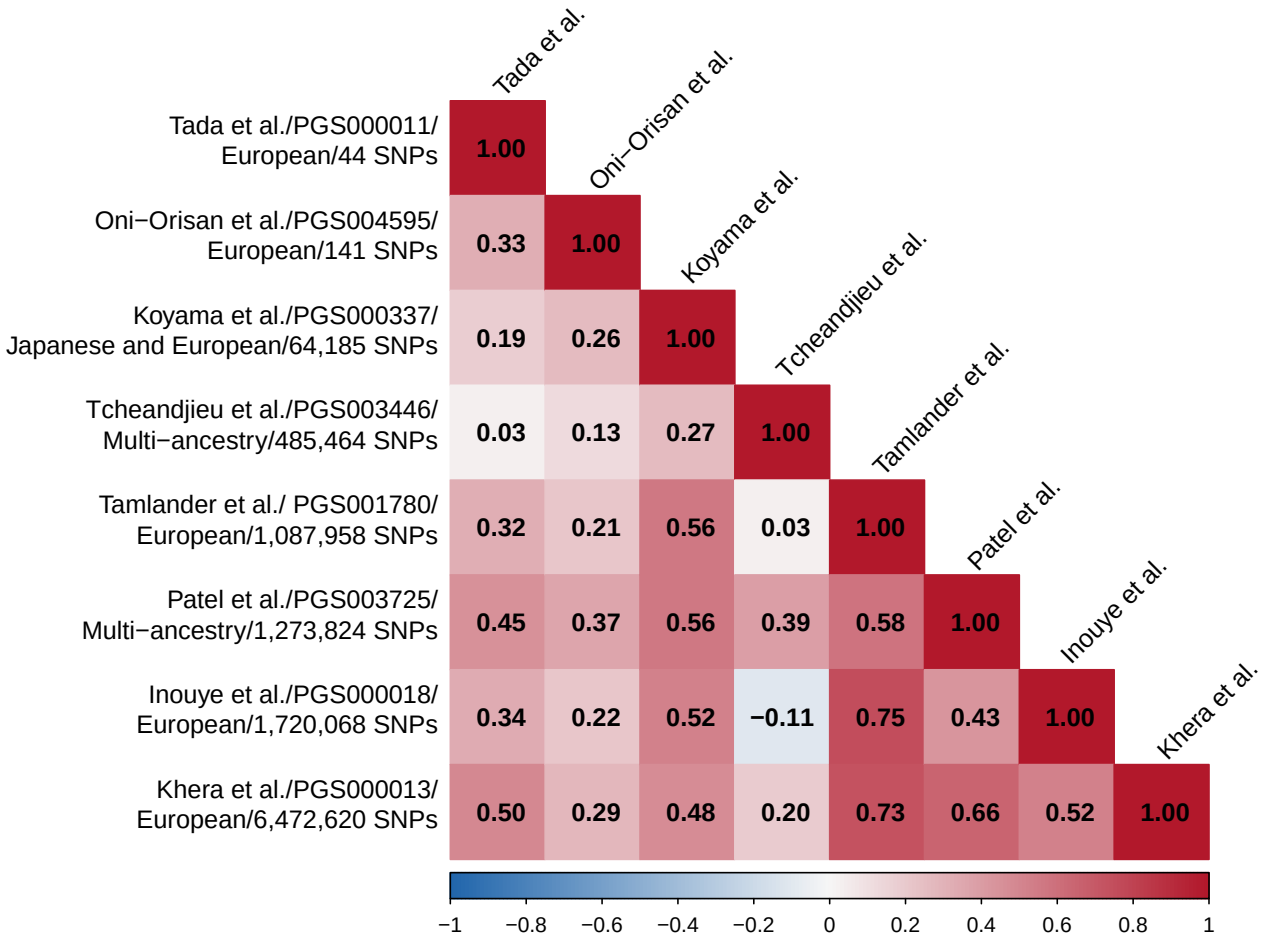


Figure 4.1: Pairwise correlations between the eight selected PRSs

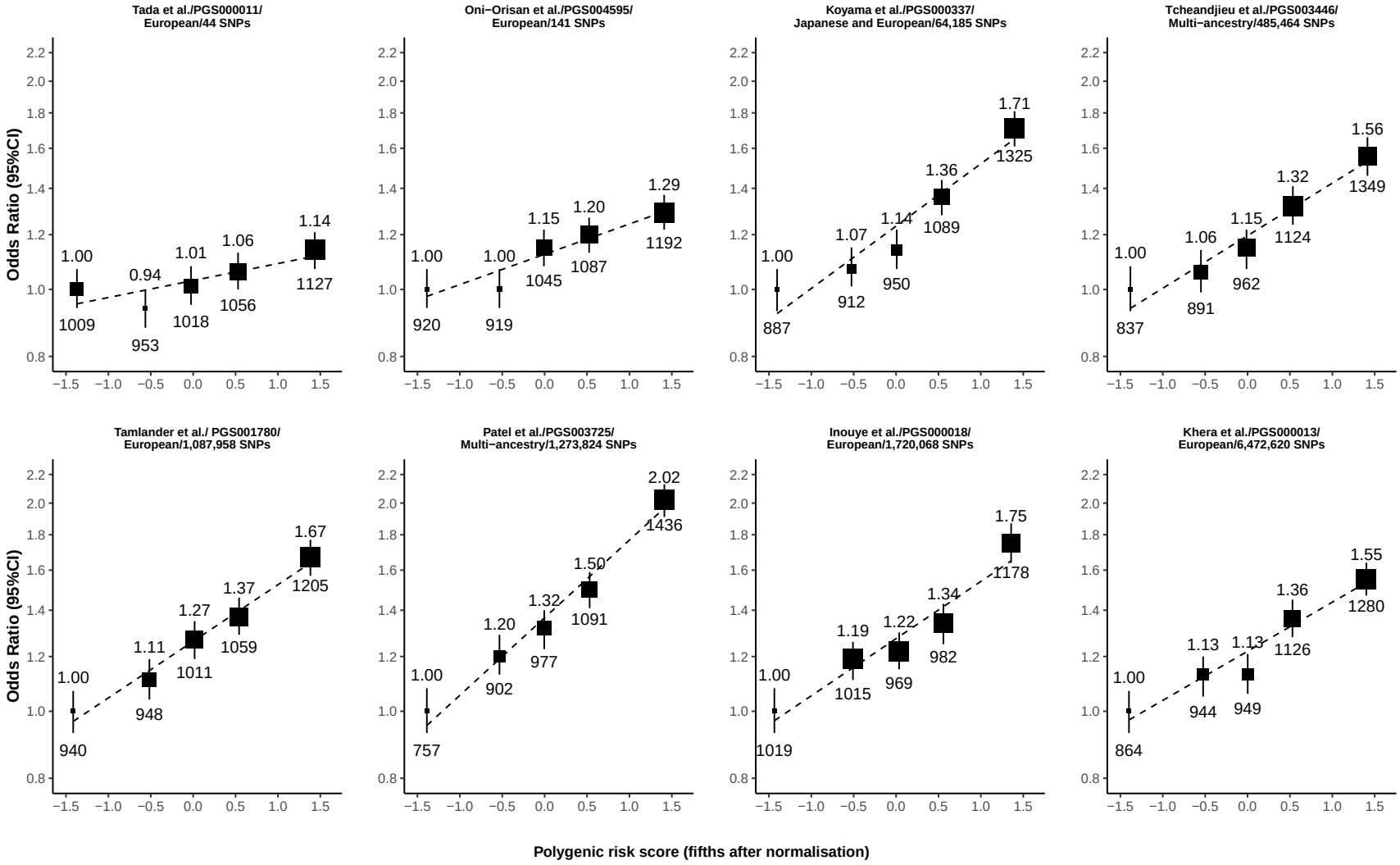


Figure 4.2: Odds of CHD by fifth of each PRS

Analyses are adjusted for age, sex and the first 7 principal components. Each group is plotted against the mean of the normalised PRS. The vertical lines through each point represent 95% confidence intervals and are shown for each category (including the reference category with RR=1.0). The area of each square is inversely proportional to the square of the standard error of the log odds ratio (i.e., it is proportional to the amount of statistical information). ORs are shown above each point and the number of CHD cases below each point. PGS IDs refer to the ID number of PRSs on the PGS catalogue.

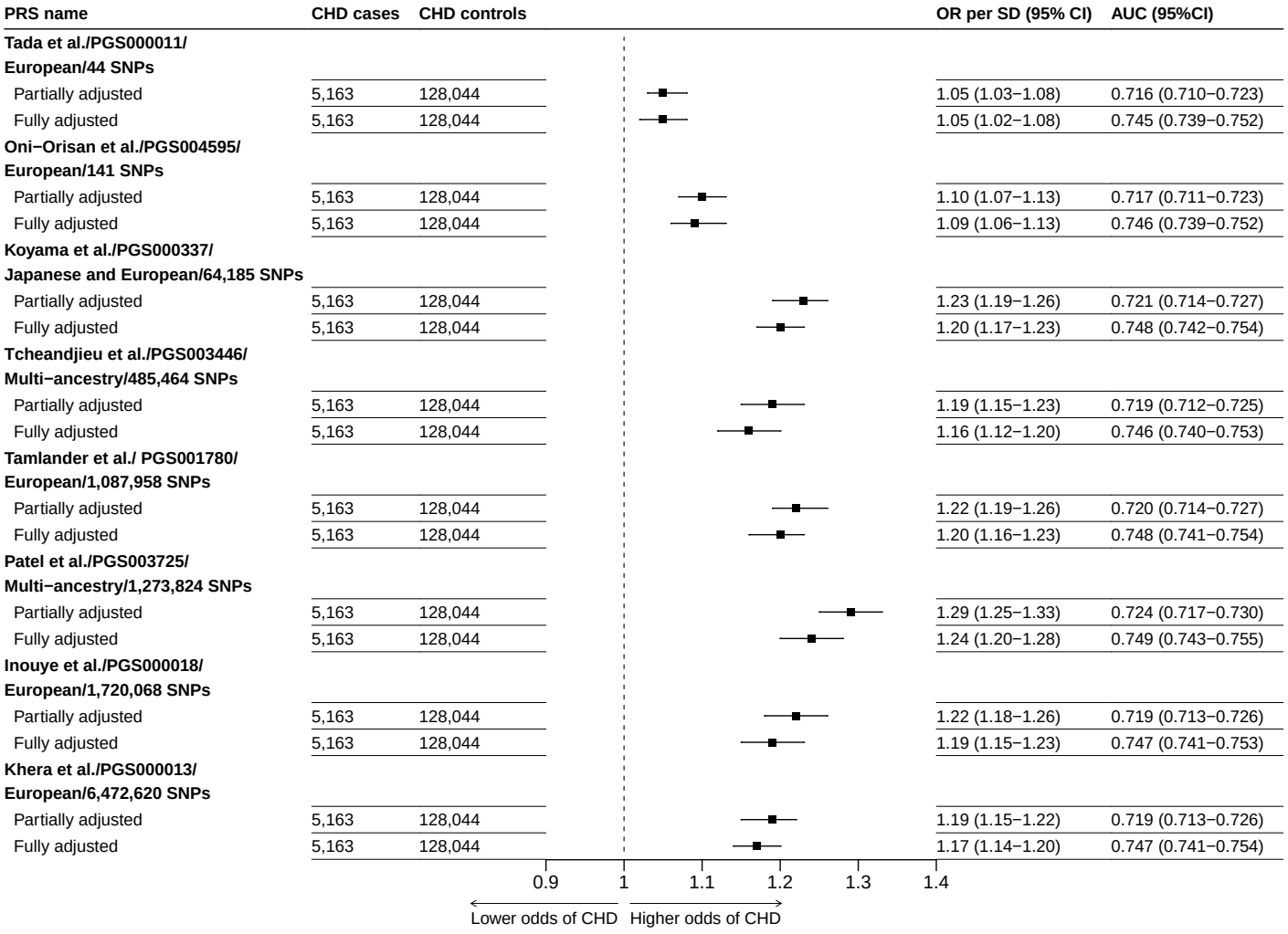


Figure 4.3: Odds of CHD per 1SD increase in each PRS

In the partially adjusted model adjustment is for age, sex and the first 7 genetic principal components. The fully adjusted model is further adjusted for baseline waist-to-hip ratio, systolic and diastolic blood pressure, smoking status, level of education, and diabetes. Analyses are restricted to eligible participants with complete data on all covariates. PGS IDs refer to the ID number of PRSs on the PGS catalogue.

	Men			Women				
PRS name	**CHD case / control**	**OR per SD (95%CI)**	**AUC (95%CI)**	**CHD case / control**	**OR per SD (95%CI)**	**AUC (95%CI)**	**OR comparison between sex**	**Sex interaction p−value**
Tada et al./PGS000011/ European/44 SNPs								
Partially adjusted	2,415 / 40,923	1.06 (1.02−1.10)	0.681 (0.671−0.691)	2,748 / 87,121	1.05 (1.01−1.09)	0.717 (0.708−0.726)		0.65
Fully adjusted	2,415 / 40,923	1.05 (1.01−1.10)	0.719 (0.709−0.729)	2,748 / 87,121	1.05 (1.01−1.09)	0.746 (0.737−0.755)		0.98
Oni−Orisan et al./PGS004595/ European/141 SNPs								
Partially adjusted	2,415 / 40,923	1.12 (1.08−1.17)	0.682 (0.671−0.692)	2,748 / 87,121	1.08 (1.04−1.13)	0.717 (0.708−0.726)		0.09
Fully adjusted	2,415 / 40,923	1.12 (1.07−1.16)	0.720 (0.710−0.730)	2,748 / 87,121	1.08 (1.03−1.12)	0.746 (0.738−0.755)		0.09
Koyama et al./PGS000337/ Japanese and European/64,185 SNPs								
Partially adjusted	2,415 / 40,923	1.29 (1.24−1.35)	0.691 (0.681−0.701)	2,748 / 87,121	1.18 (1.13−1.22)	0.719 (0.711−0.728)		<0.0001
Fully adjusted	2,415 / 40,923	1.27 (1.22−1.33)	0.726 (0.716−0.736)	2,748 / 87,121	1.14 (1.10−1.19)	0.747 (0.738−0.756)		<0.0001
Tcheandjieu et al./PGS003446/ Multi−ancestry/485,464 SNPs								
Partially adjusted	2,415 / 40,923	1.30 (1.24−1.37)	0.689 (0.679−0.699)	2,748 / 87,121	1.10 (1.05−1.15)	0.717 (0.708−0.726)		<0.0001
Fully adjusted	2,415 / 40,923	1.28 (1.21−1.34)	0.724 (0.714−0.734)	2,748 / 87,121	1.07 (1.02−1.12)	0.746 (0.737−0.754)		<0.0001
Tamlander et al./ PGS001780/ European/1,087,958 SNPs								
Partially adjusted	2,415 / 40,923	1.29 (1.23−1.34)	0.689 (0.679−0.699)	2,748 / 87,121	1.17 (1.12−1.22)	0.719 (0.710−0.728)		<0.0001
Fully adjusted	2,415 / 40,923	1.26 (1.21−1.32)	0.725 (0.715−0.735)	2,748 / 87,121	1.15 (1.10−1.19)	0.747 (0.738−0.756)		<0.0001
Patel et al./PGS003725/ Multi−ancestry/1,273,824 SNPs								
Partially adjusted	2,415 / 40,923	1.37 (1.32−1.43)	0.696 (0.686−0.706)	2,748 / 87,121	1.23 (1.18−1.28)	0.721 (0.712−0.730)		<0.0001
Fully adjusted	2,415 / 40,923	1.33 (1.27−1.39)	0.729 (0.719−0.738)	2,748 / 87,121	1.17 (1.13−1.22)	0.747 (0.739−0.756)		<0.0001
Inouye et al./PGS000018/ European/1,720,068 SNPs								
Partially adjusted	2,415 / 40,923	1.29 (1.23−1.36)	0.688 (0.678−0.698)	2,748 / 87,121	1.16 (1.11−1.22)	0.718 (0.710−0.727)		<0.0001
Fully adjusted	2,415 / 40,923	1.27 (1.21−1.34)	0.724 (0.714−0.734)	2,748 / 87,121	1.13 (1.08−1.18)	0.747 (0.738−0.755)		<0.0001
Khera et al./PGS000013/ European/6,472,620 SNPs								
Partially adjusted	2,415 / 40,923	1.25 (1.20−1.30)	0.688 (0.678−0.698)	2,748 / 87,121	1.14 (1.09−1.18)	0.718 (0.709−0.727)		<0.0001
Fully adjusted	2,415 / 40,923	1.24 (1.19−1.29)	0.724 (0.715−0.734)	2,748 / 87,121	1.12 (1.08−1.16)	0.747 (0.738−0.755)		<0.0001

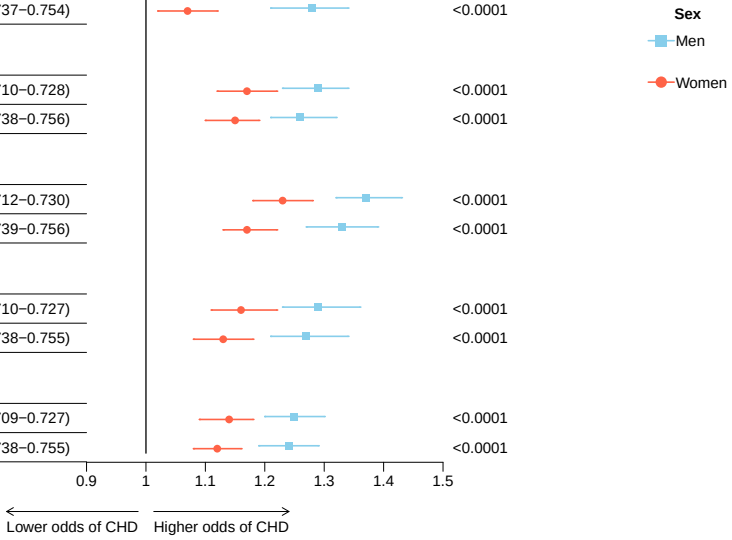


Figure 4.4: Odds of CHD per 1SD increase in each PRS, by sex

Analyses as for Figure 4.3, now presented separately for men and women.

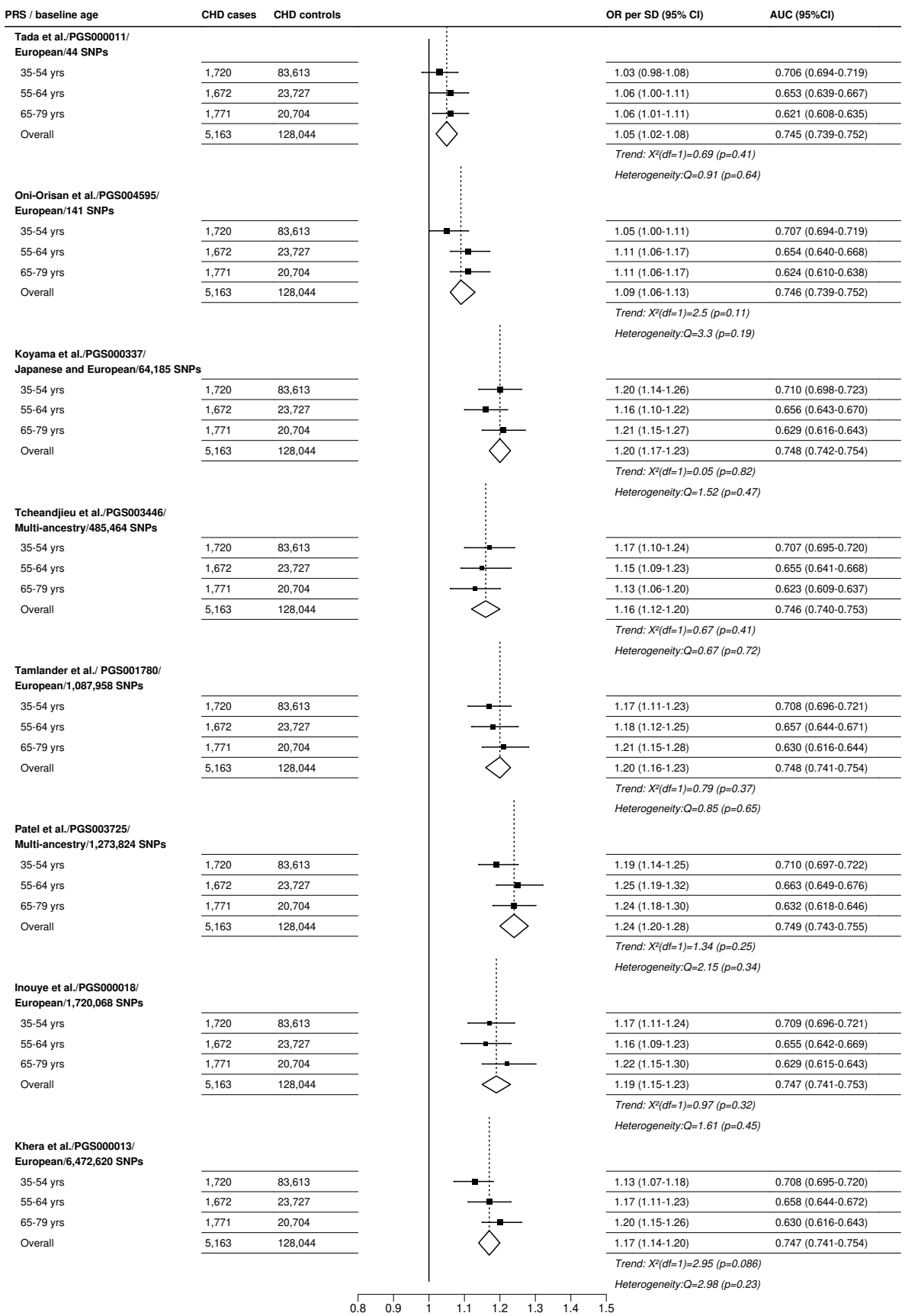


Figure 4.5: Odds of CHD per 1SD increase in each PRS, by baseline age

Subgroup-specific effects are from the fully adjusted model excluding age. The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

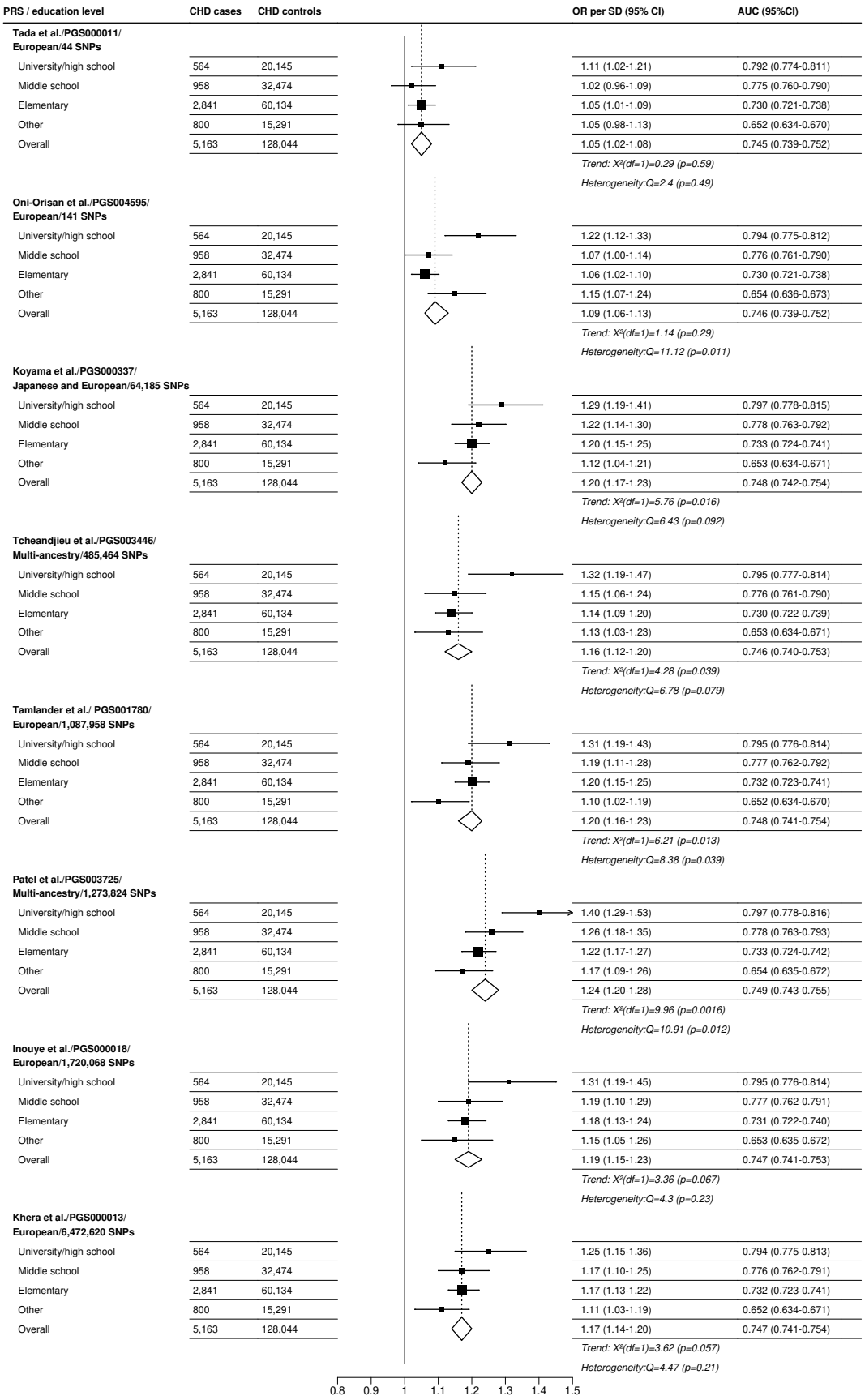


Figure 4.6: Odds of CHD per 1SD increase in each PRS, by highest level of education

Subgroup-specific effects are from the fully adjusted model excluding highest level of education. The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

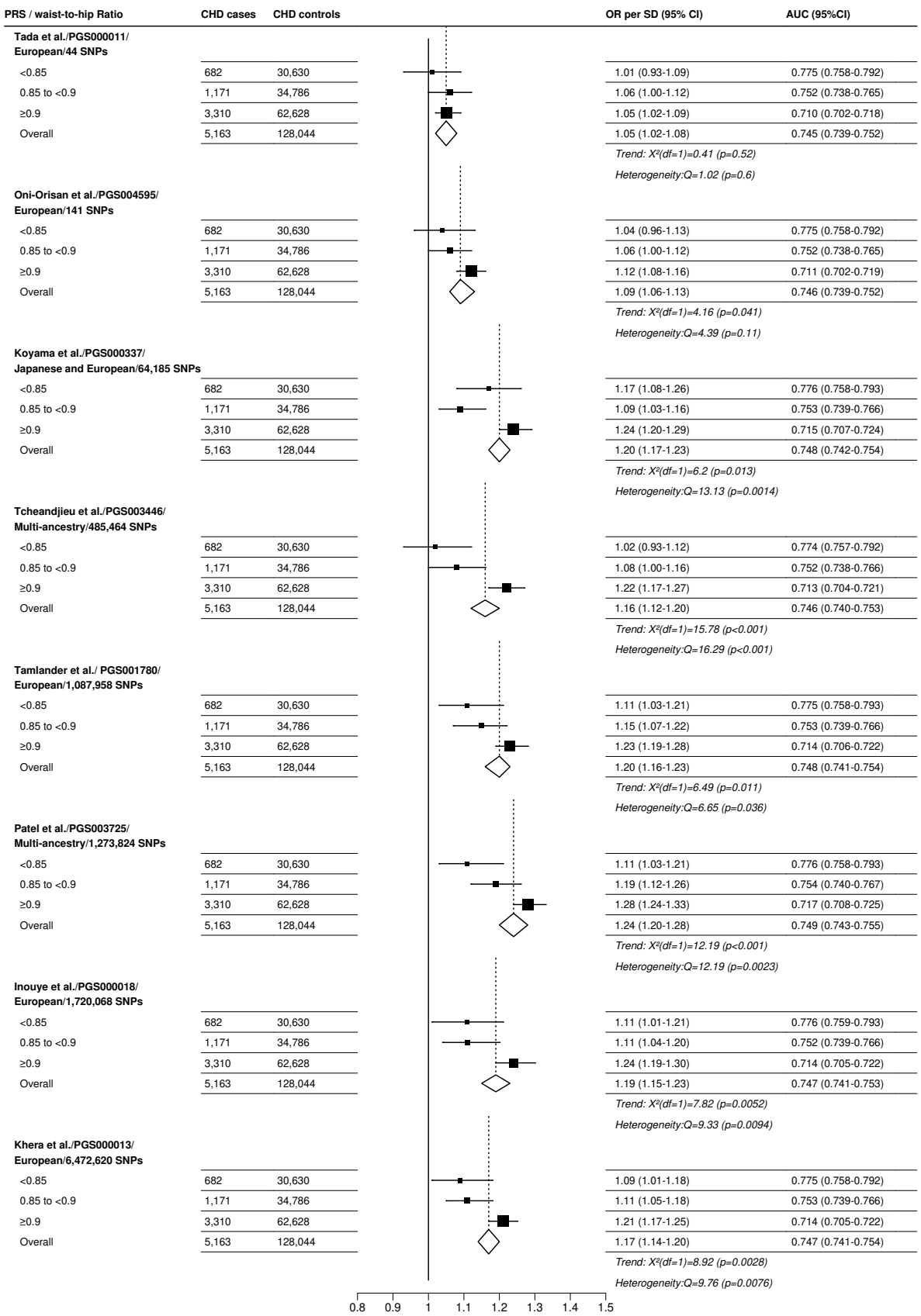


Figure 4.7: Odds of CHD per 1SD increase in each PRS, by waist-to-hip ratio

Subgroup-specific effects are from the fully adjusted model excluding waist-to-hip ratio. The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

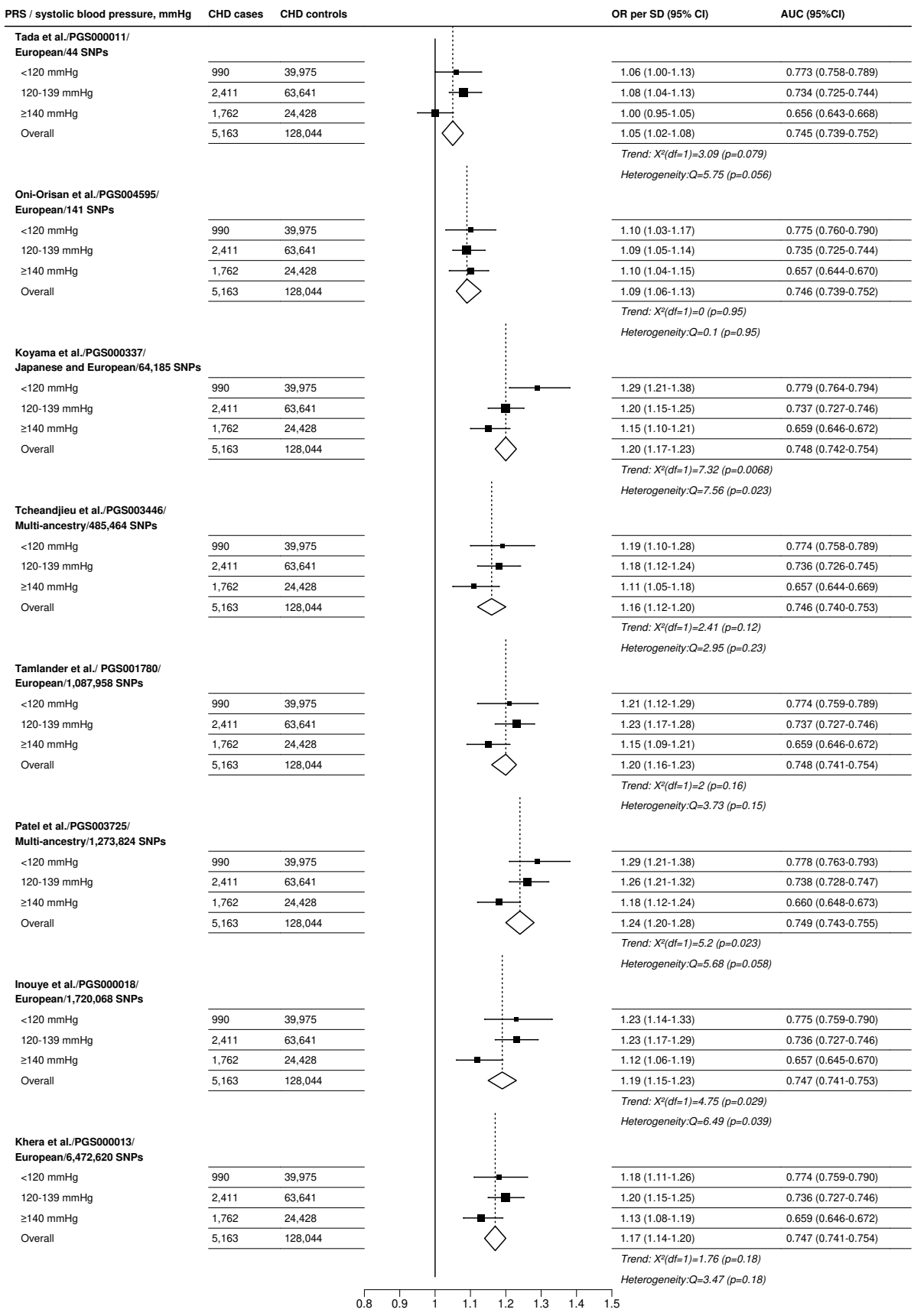


Figure 4.8: Odds of CHD per 1SD increase in each PRS, by systolic blood pressure

Subgroup-specific effects are from the fully adjusted model excluding systolic blood pressure. The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

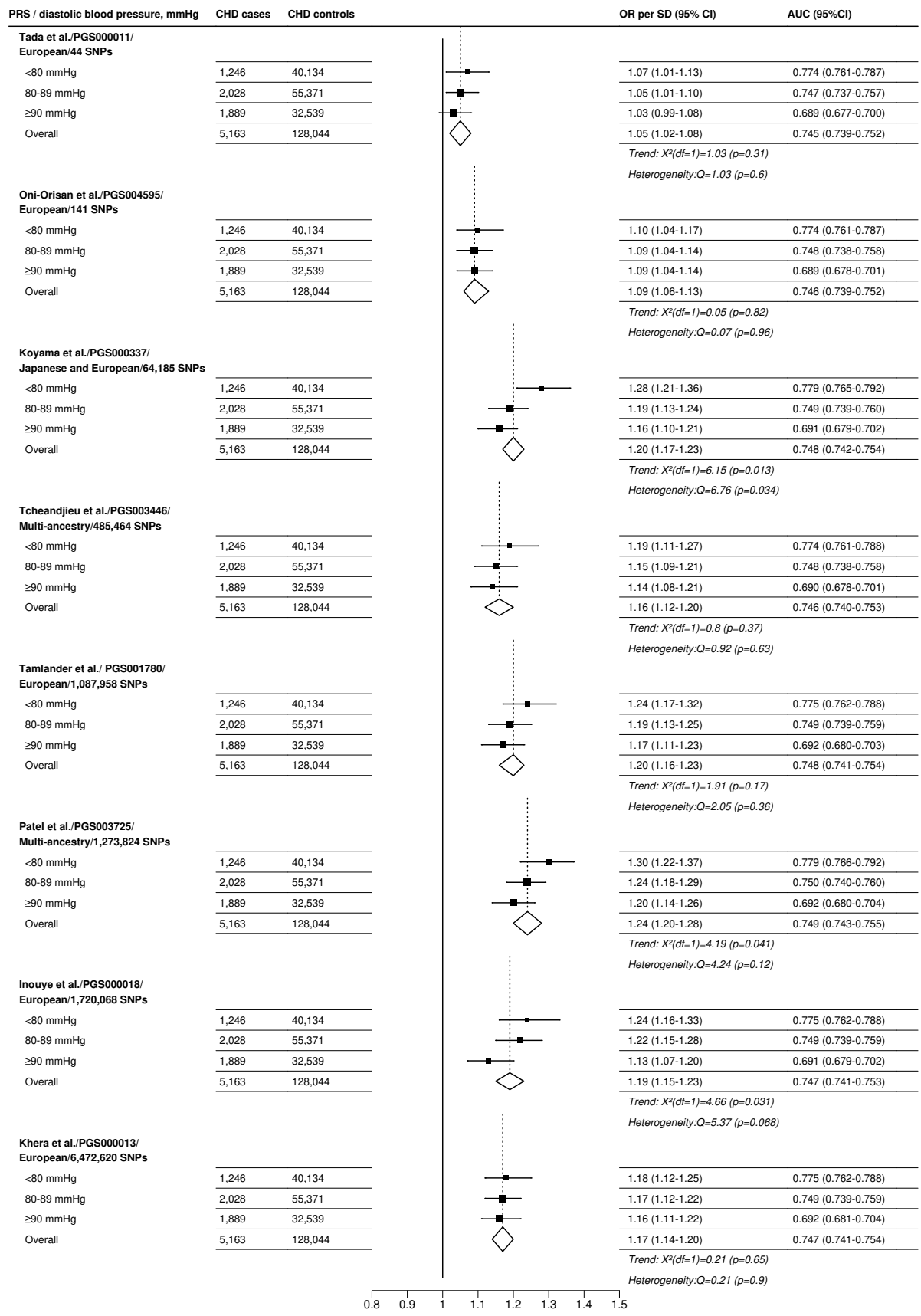


Figure 4.9: Odds of CHD per 1SD increase in each PRS, by diastolic blood pressure

Subgroup-specific effects are from the fully adjusted model excluding diastolic blood pressure. The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

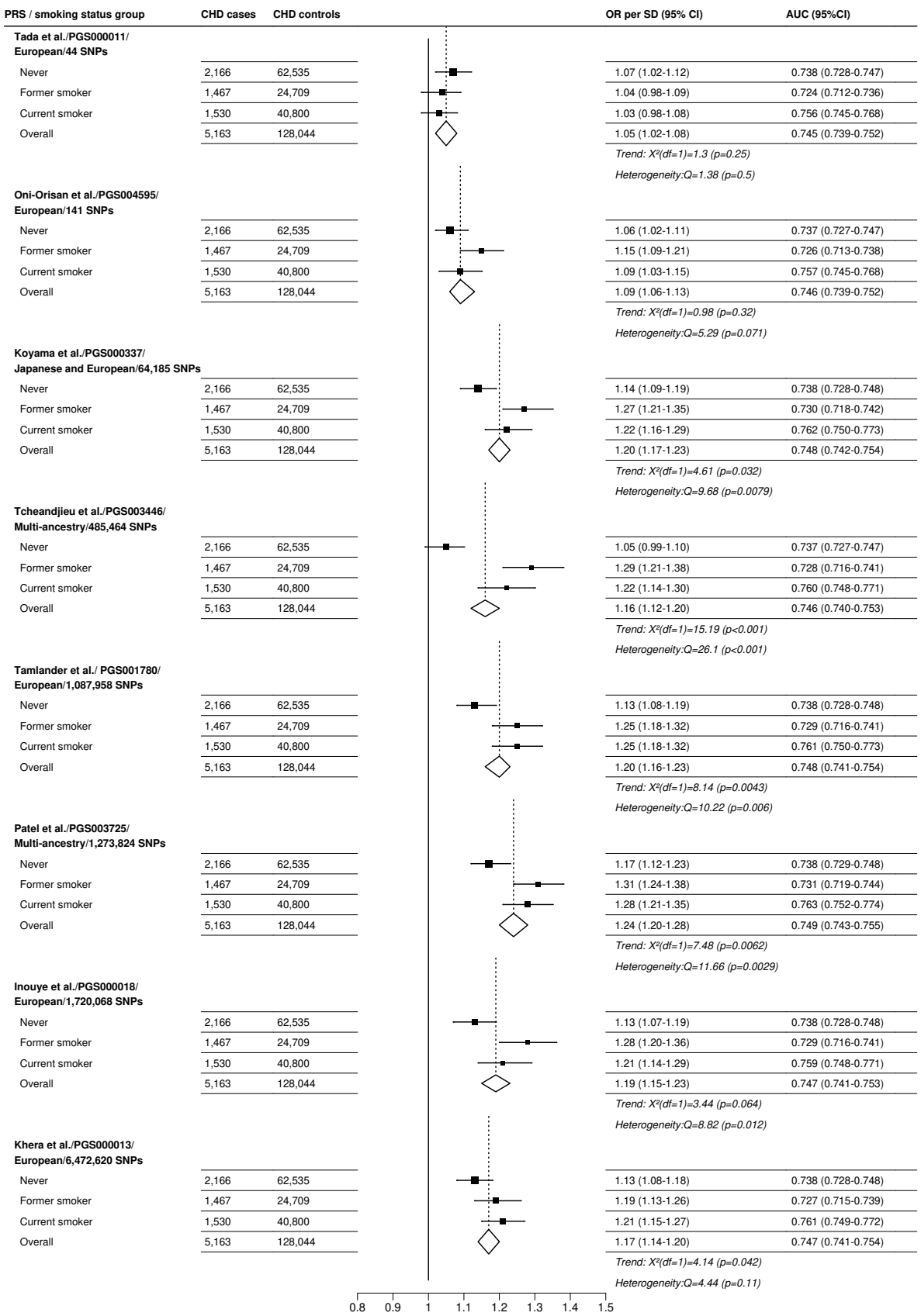


Figure 4.10: Odds of CHD per 1SD increase in each PRS, by smoking status

Subgroup-specific effects are from the fully adjusted model excluding smoking status. The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

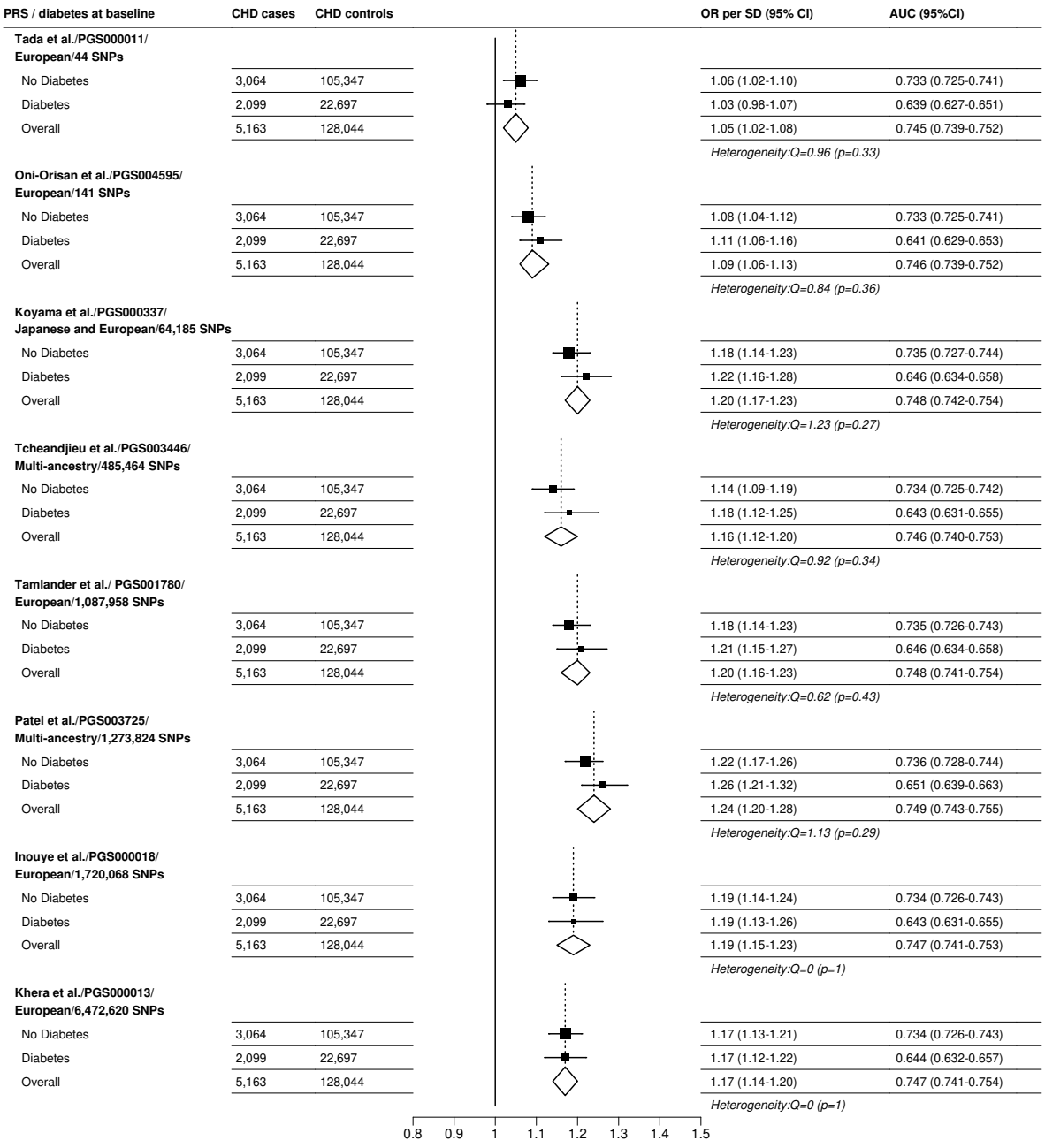


Figure 4.11: Odds of CHD per 1SD increase in each PRS, by baseline diabetes status

Subgroup-specific effects are from the fully adjusted model excluding baseline diabetes status. The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

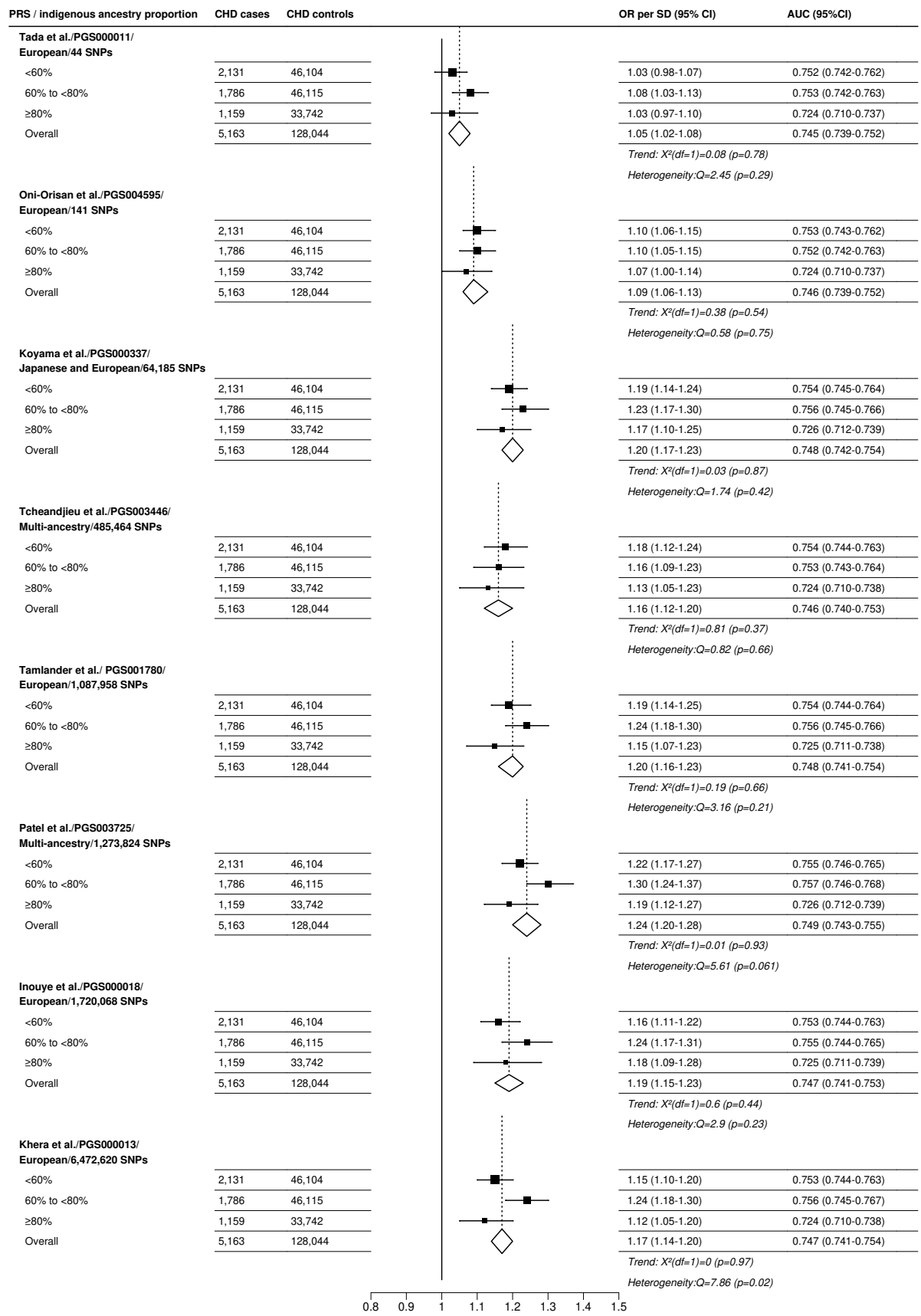


Figure 4.12: Odds of CHD per 1SD increase in each PRS, by level of Indigenous American ancestry

Subgroup-specific effects are from the fully adjusted model.The overall effect shown is from the fully adjusted model as shown in Figure 4.3.

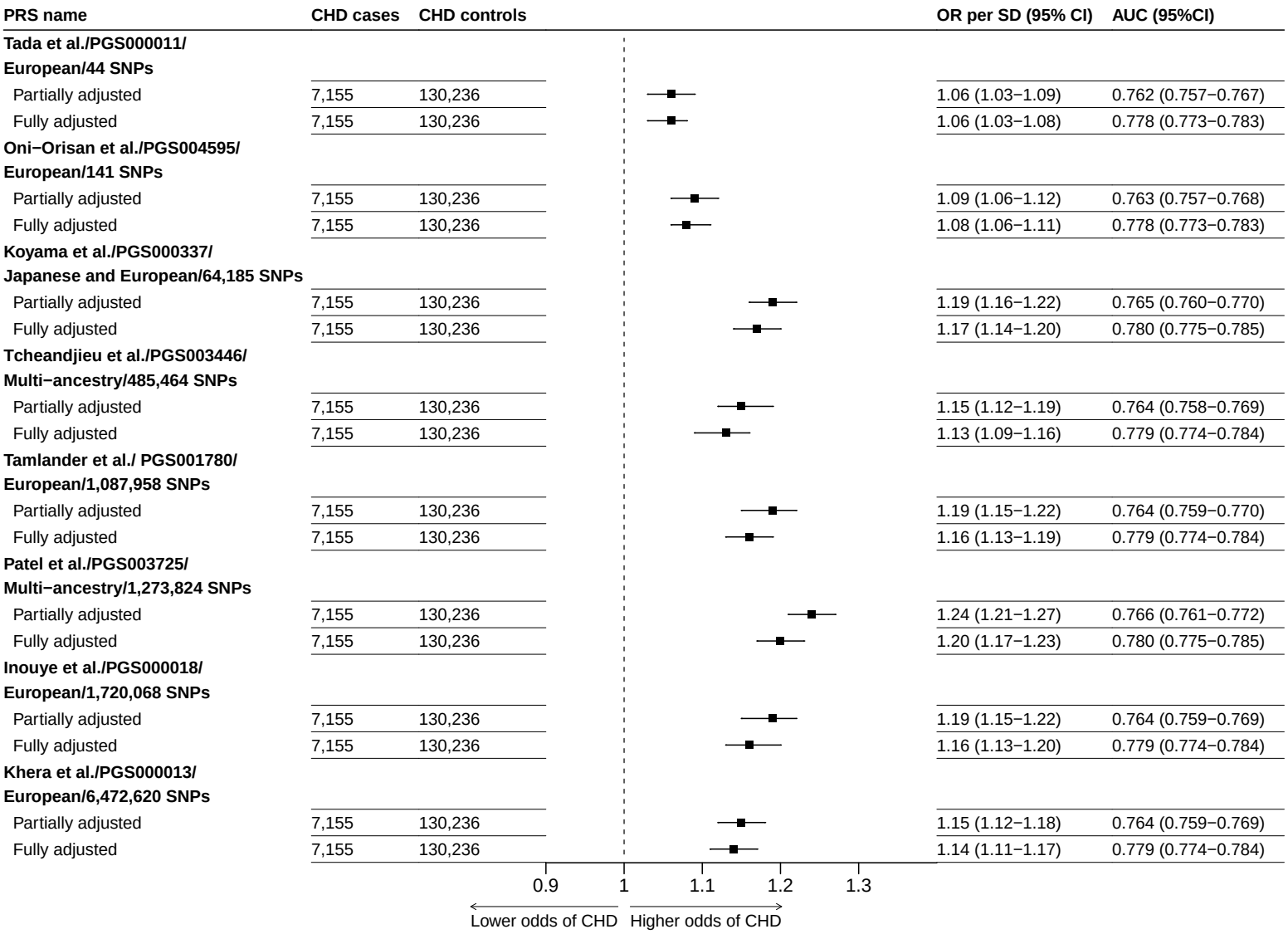


Figure 4.13: Odds of CHD before age 90 years per 1SD increase in each PRS, among participants age 35-89 at recruitment

Analyses as for Figure 4.3.

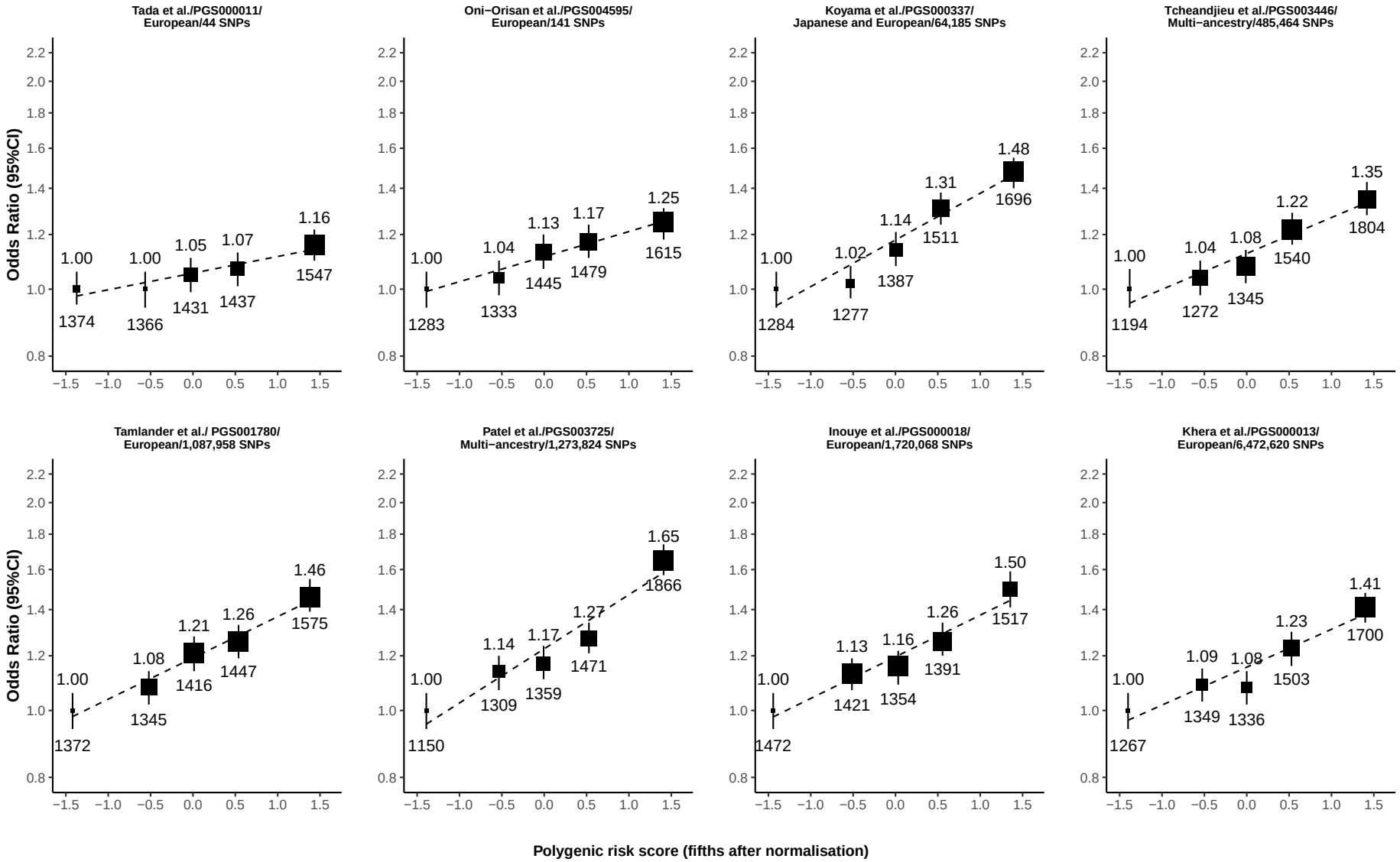


Figure 4.14: Odds of CHD before age 90 years by fifth of each PRS, among participants age 35-89 years at recruitment

Analyses as for Figure 4.2.

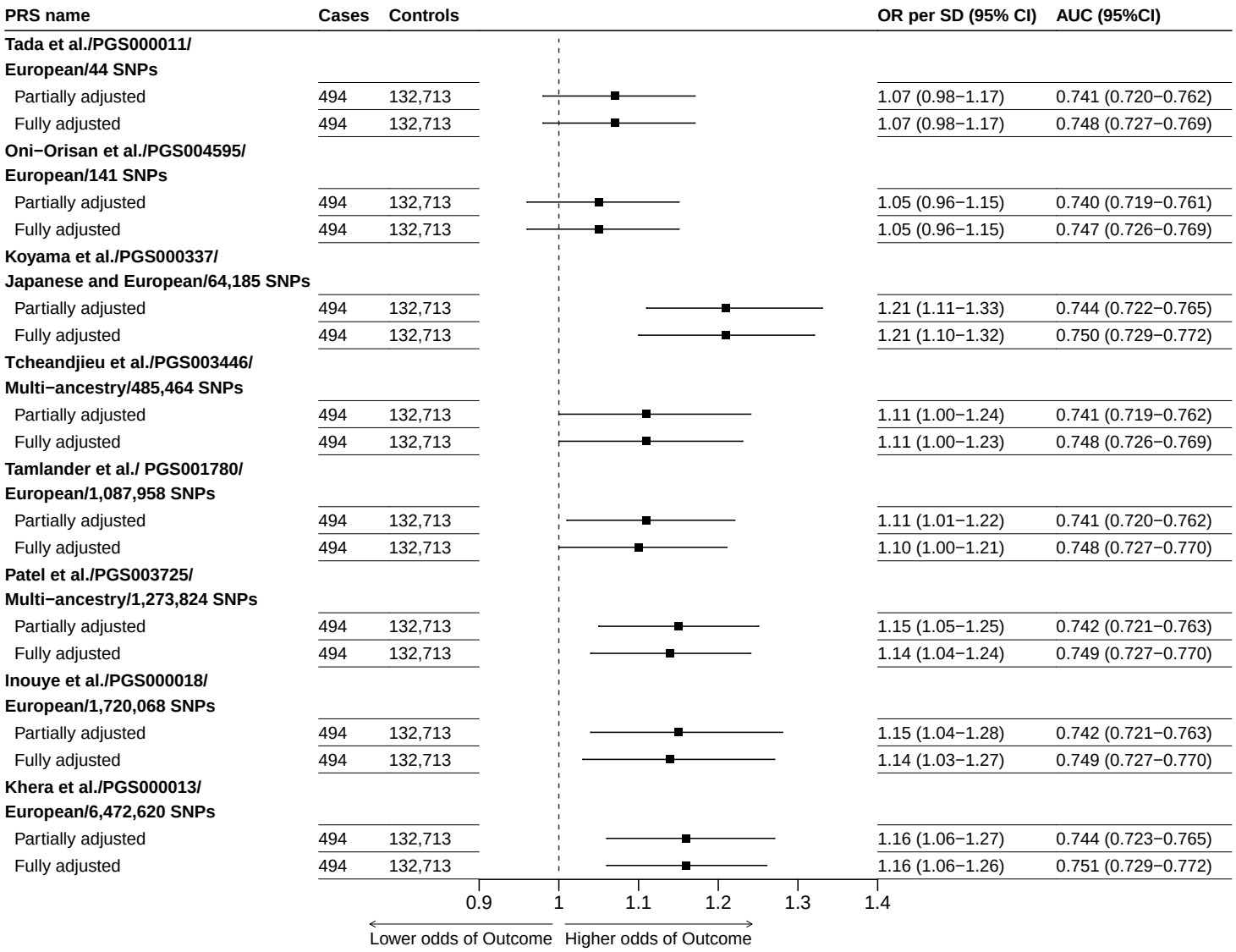


Figure 4.15: Odds of baseline self-reported angina per 1SD increase in each PRS, for participants aged 35-79 years at recruitment

Analyses as for Figure 4.3.

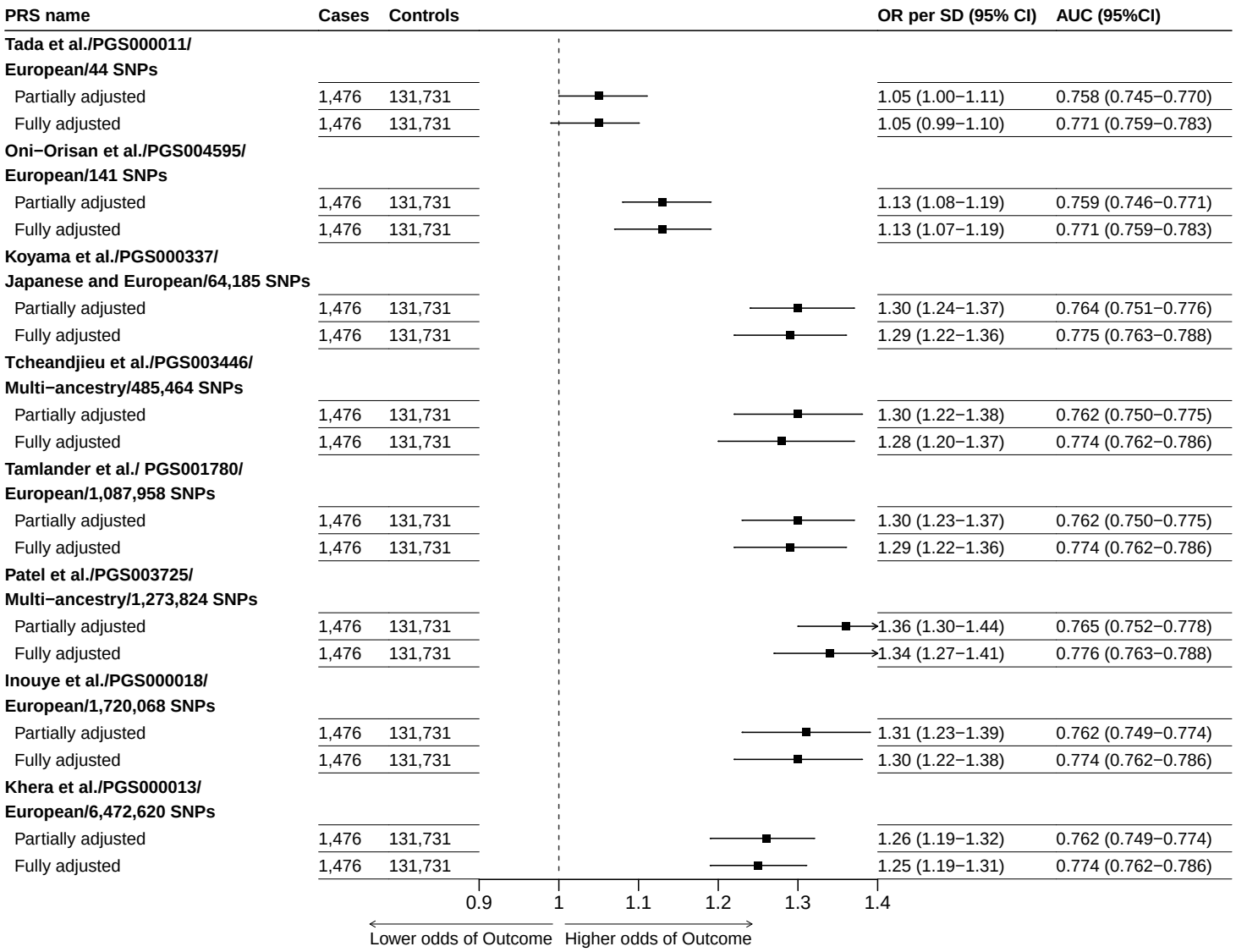


Figure 4.16: Odds of baseline self-reported myocardial infarction per 1SD increase in each PRS, for participants aged 35-79 years at recruitment

Analyses as for Figure 4.3.

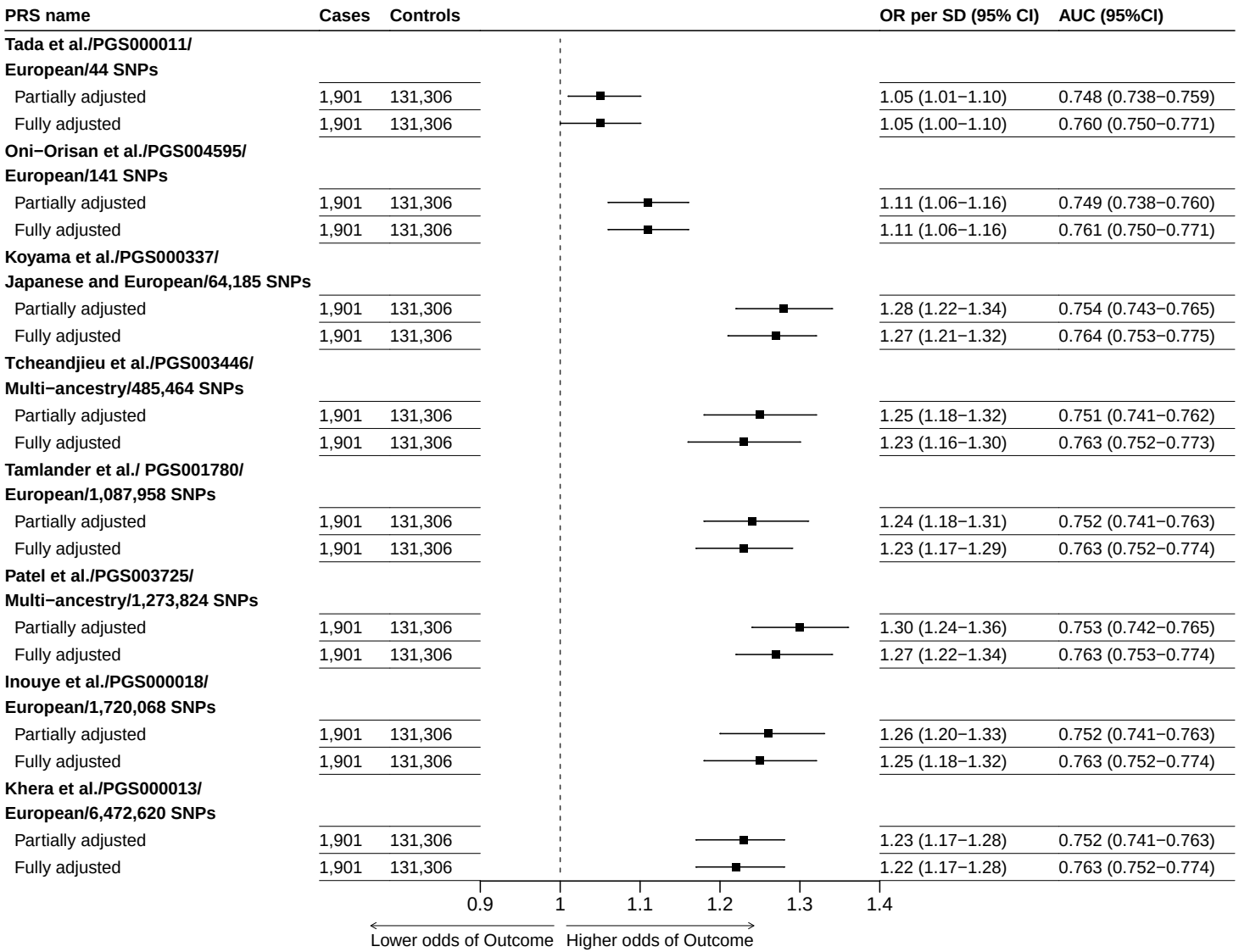


Figure 4.17: Odds of baseline self-reported angina or myocardial infarction per 1SD increase in each PRS, for participants aged 35-79 years at recruitment

Analyses as for Figure 4.3.

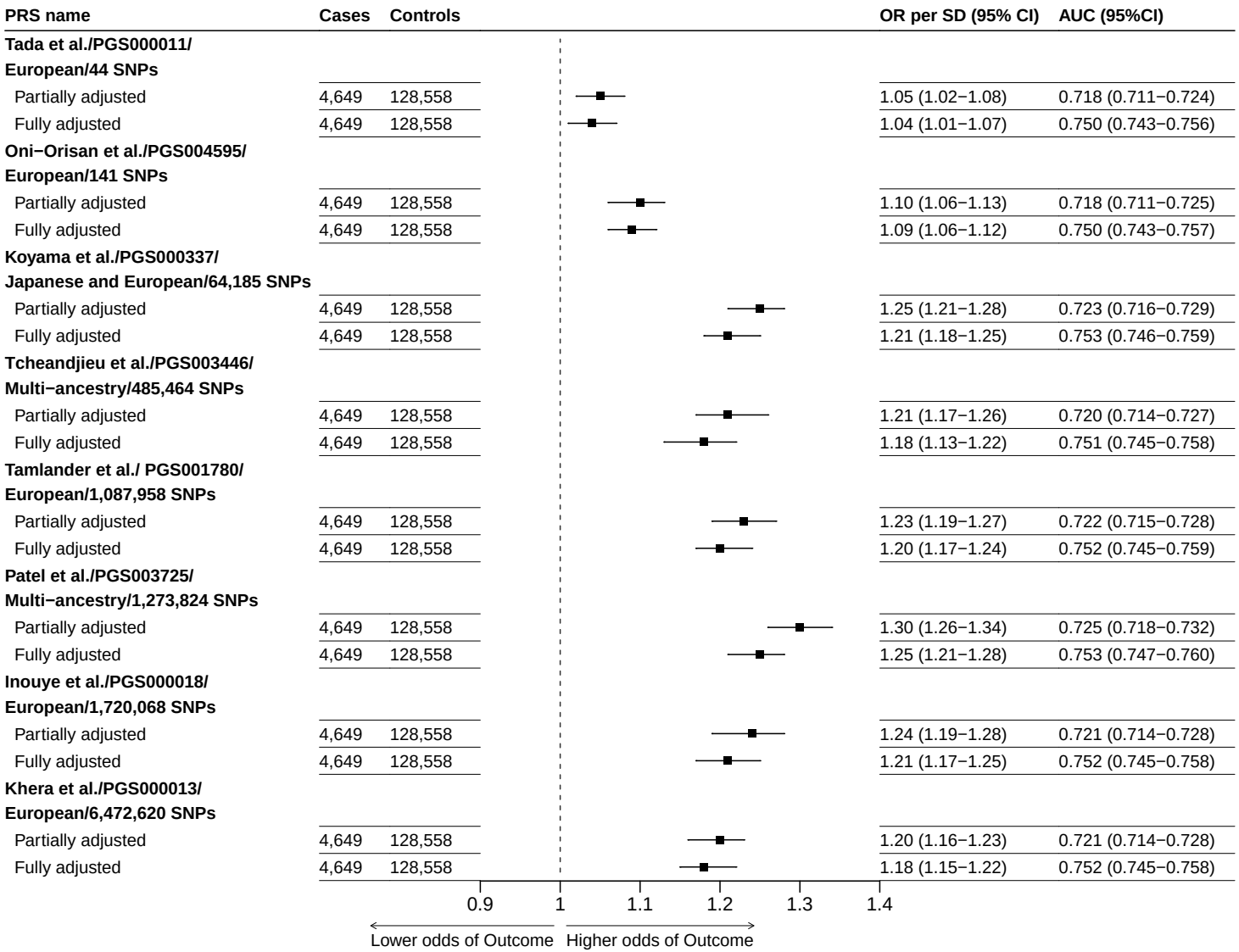


Figure 4.18: Odds of CHD (now defined as baseline self-reported angina or myocardial infarction, or death before age 80 with CHD listed as the primary cause of death) per 1SD increase in each PRS, for participants aged 35-79 years at recruitment

Analyses as for Figure 4.3.

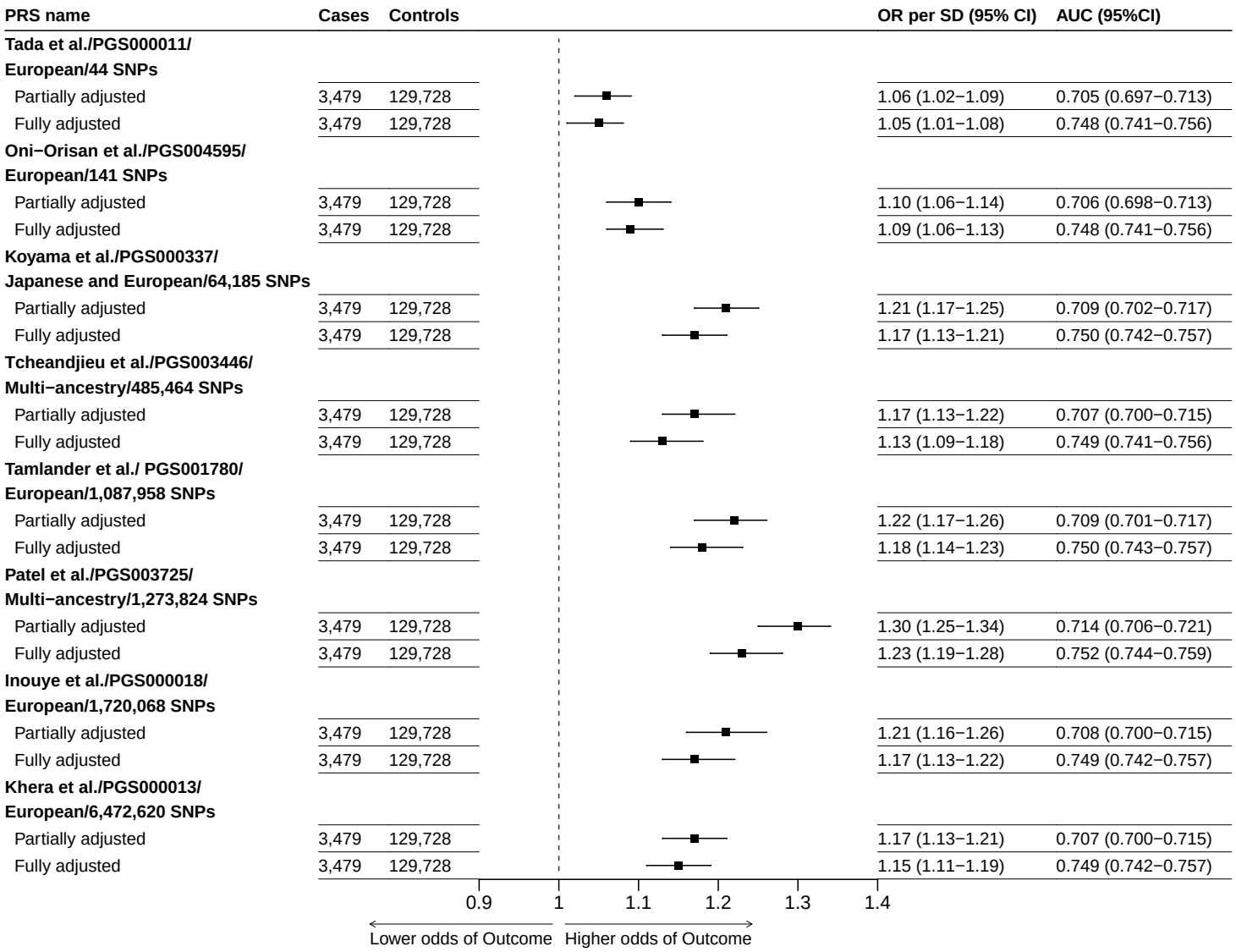


Figure 4.19: Odds of death before age 80 with CHD listed anywhere on the death certificate, per 1SD increase in each PRS

Analyses as for Figure 4.3.

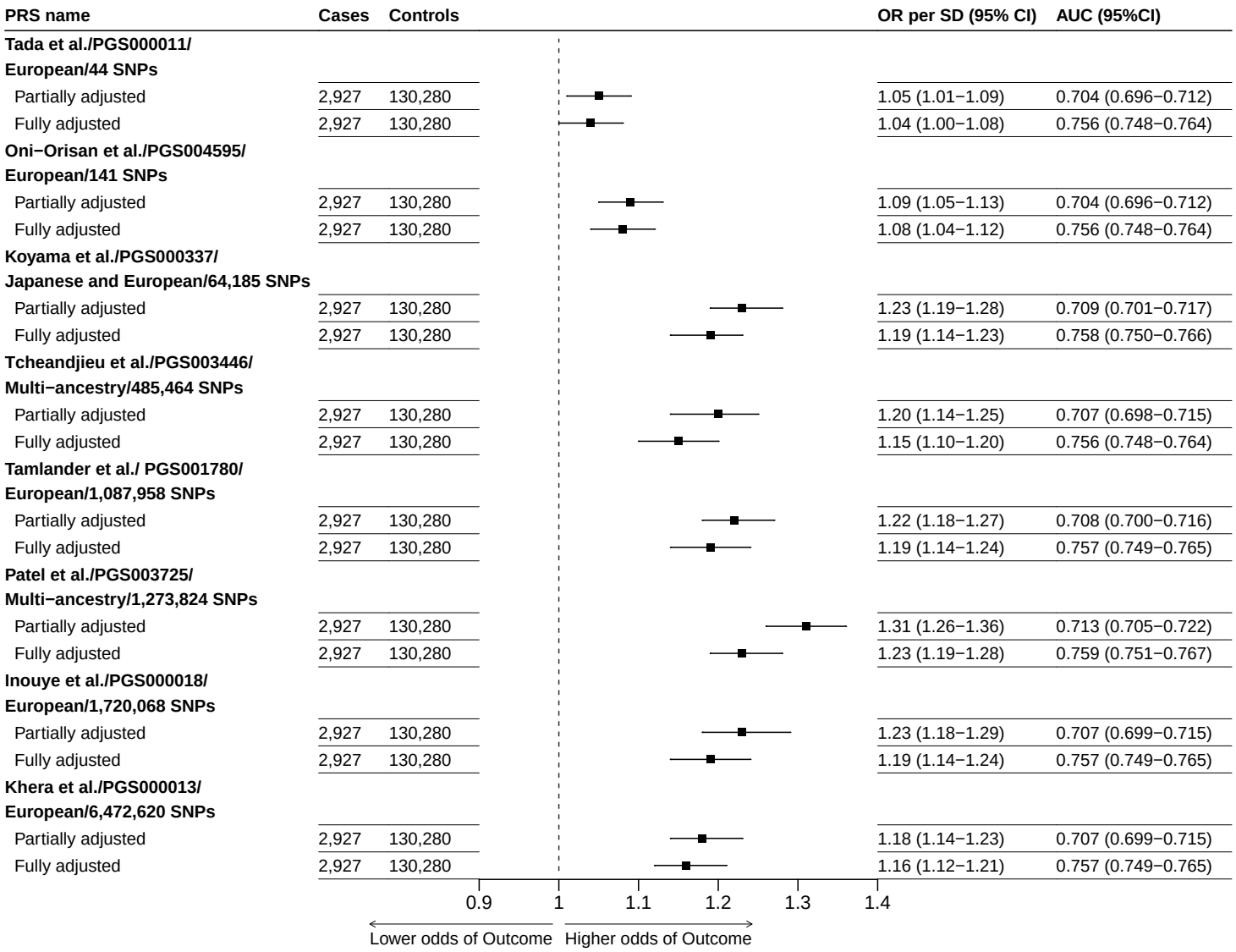


Figure 4.20: Odds of death before age 80 with CHD listed as the primary cause on the death certificate, per 1SD increase in each PRS

Analyses as for Figure 4.3.

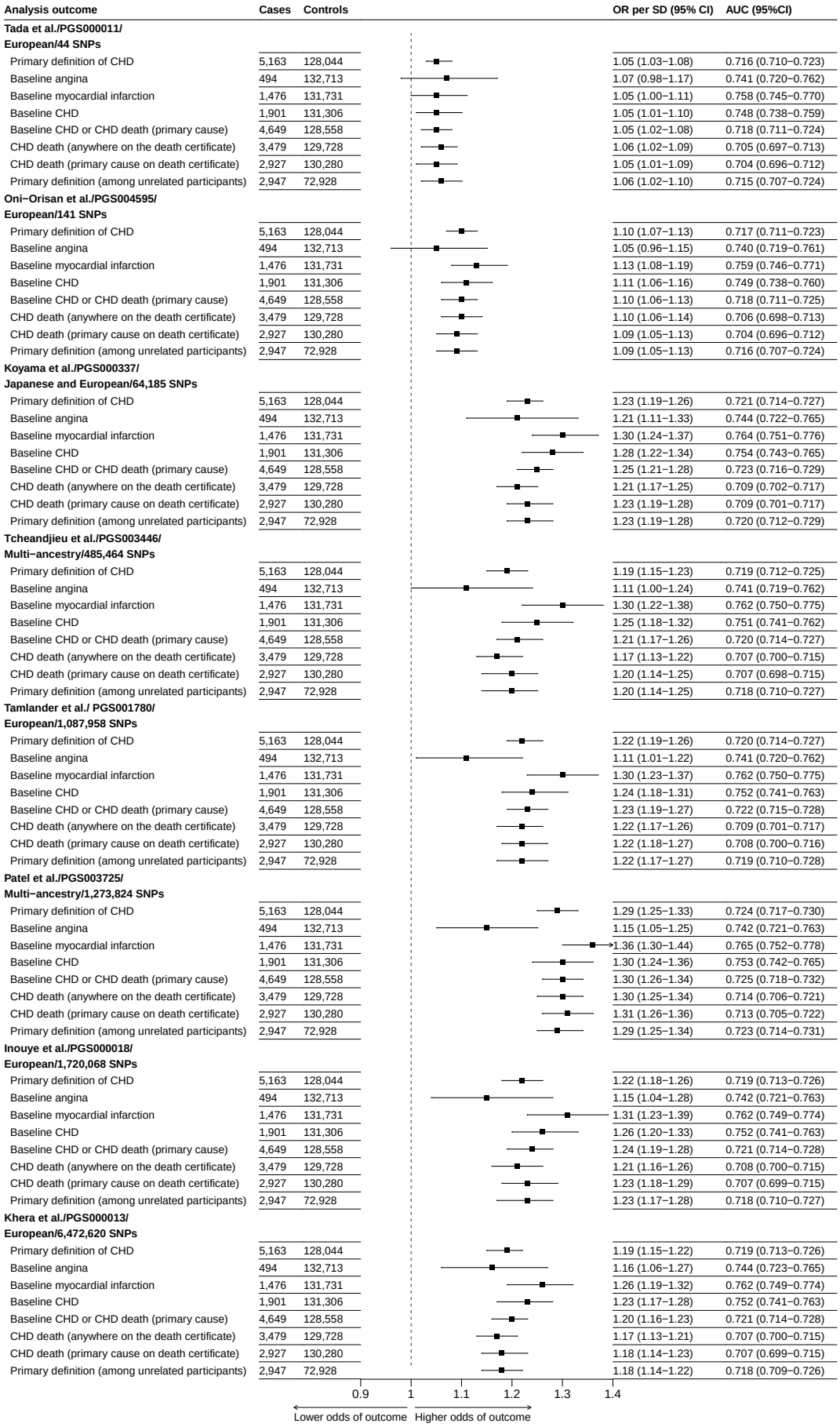


Figure 4.21: Sensitivity analyses (varying the CHD outcome) with partial adjustments

Odds ratios (ORs) were estimated with regression models adjusted for baseline age, sex and seven genetic principal components.

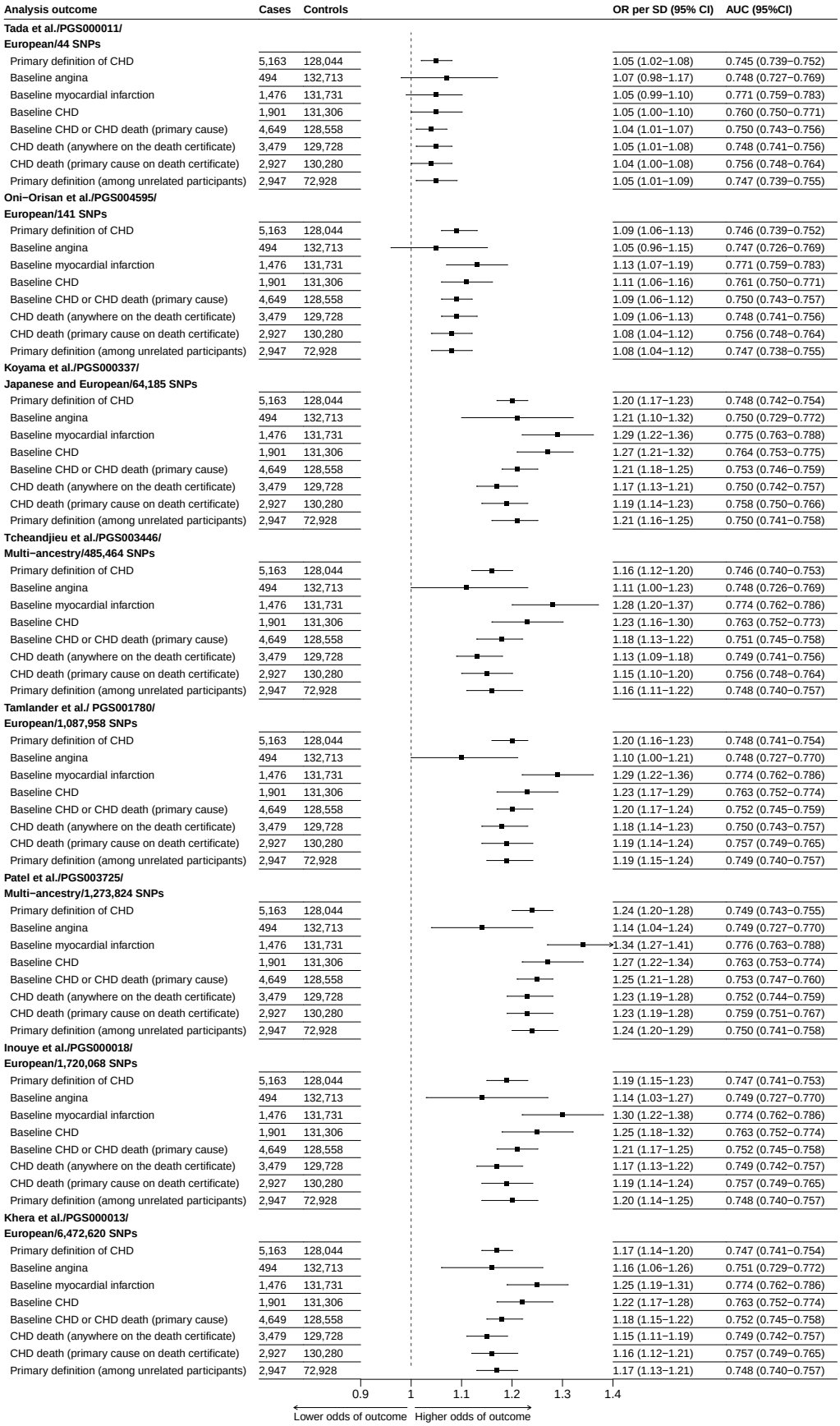


Figure 4.22: Sensitivity analyses (varying the CHD outcome) with full adjustments

Odds ratios (ORs) were estimated with regression models adjusted for baseline age, sex, seven genetic principal components, waist-to-hip ratio, systolic and diastolic blood pressures, education attainment level, smoking status, and diabetes at baseline.

Patel et al./PGS003725/ Multi−ancestry/1,273,824 SNPs	Men			Women				
	CHD case / control	OR per SD (95%CI)	AUC (95%CI)	CHD case / control	OR per SD (95%CI)	AUC (95%CI)	OR comparison between sex	Sex interaction p−value
Partially adjusted								
Primary definition of CHD	2,415 / 40,923	1.37 (1.32−1.43)	0.696 (0.686−0.706)	2,748 / 87,121	1.23 (1.18−1.28)	0.721 (0.712−0.730)		<0.0001
Baseline angina	136 / 43,202	1.43 (1.21−1.70)	0.718 (0.674−0.763)	358 / 89,511	1.05 (0.95−1.17)	0.756 (0.733−0.780)		4e−04
Baseline myocardial infarction	747 / 42,591	1.51 (1.40−1.62)	0.781 (0.765−0.796)	729 / 89,140	1.24 (1.15−1.33)	0.728 (0.709−0.747)		<0.0001
Baseline CHD	848 / 42,490	1.48 (1.38−1.59)	0.770 (0.755−0.786)	1,053 / 88,816	1.17 (1.10−1.25)	0.735 (0.720−0.750)		<0.0001
Baseline CHD or CHD death (primary cause)	2,165 / 41,173	1.39 (1.33−1.46)	0.698 (0.687−0.708)	2,484 / 87,385	1.23 (1.18−1.28)	0.724 (0.715−0.734)		<0.0001
CHD death (anywhere on the death certificate)	1,717 / 41,621	1.33 (1.27−1.40)	0.668 (0.656−0.679)	1,762 / 88,107	1.27 (1.21−1.33)	0.715 (0.704−0.725)		0.07
CHD death (primary cause on death certificate)	1,437 / 41,901	1.34 (1.27−1.42)	0.663 (0.650−0.675)	1,490 / 88,379	1.28 (1.21−1.34)	0.720 (0.709−0.731)		0.10
Fully adjusted								
Primary definition of CHD	2,415 / 40,923	1.33 (1.27−1.39)	0.729 (0.719−0.738)	2,748 / 87,121	1.17 (1.13−1.22)	0.747 (0.739−0.756)		<0.0001
Baseline angina	136 / 43,202	1.43 (1.21−1.70)	0.736 (0.693−0.779)	358 / 89,511	1.04 (0.94−1.16)	0.763 (0.739−0.787)		3e−04
Baseline myocardial infarction	747 / 42,591	1.49 (1.39−1.61)	0.792 (0.777−0.807)	729 / 89,140	1.21 (1.12−1.30)	0.744 (0.726−0.763)		<0.0001
Baseline CHD	848 / 42,490	1.47 (1.37−1.58)	0.781 (0.766−0.796)	1,053 / 88,816	1.14 (1.08−1.22)	0.748 (0.733−0.763)		<0.0001
Baseline CHD or CHD death (primary cause)	2,165 / 41,173	1.35 (1.29−1.41)	0.733 (0.723−0.744)	2,484 / 87,385	1.17 (1.12−1.22)	0.754 (0.745−0.763)		<0.0001
CHD death (anywhere on the death certificate)	1,717 / 41,621	1.27 (1.21−1.34)	0.717 (0.706−0.729)	1,762 / 88,107	1.19 (1.14−1.25)	0.754 (0.744−0.765)		0.01
CHD death (primary cause on death certificate)	1,437 / 41,901	1.28 (1.21−1.35)	0.722 (0.710−0.735)	1,490 / 88,379	1.20 (1.13−1.26)	0.767 (0.756−0.778)		0.02

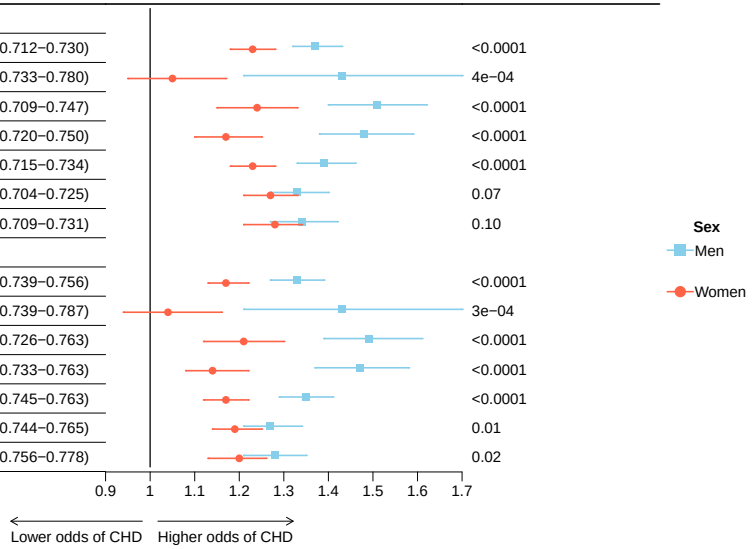


Figure 4.23: Odds of CHD (varying CHD outcomes) per 1SD increase for the Patel *et al.* PRS, by sex

Analyses as for Figure 4.3.

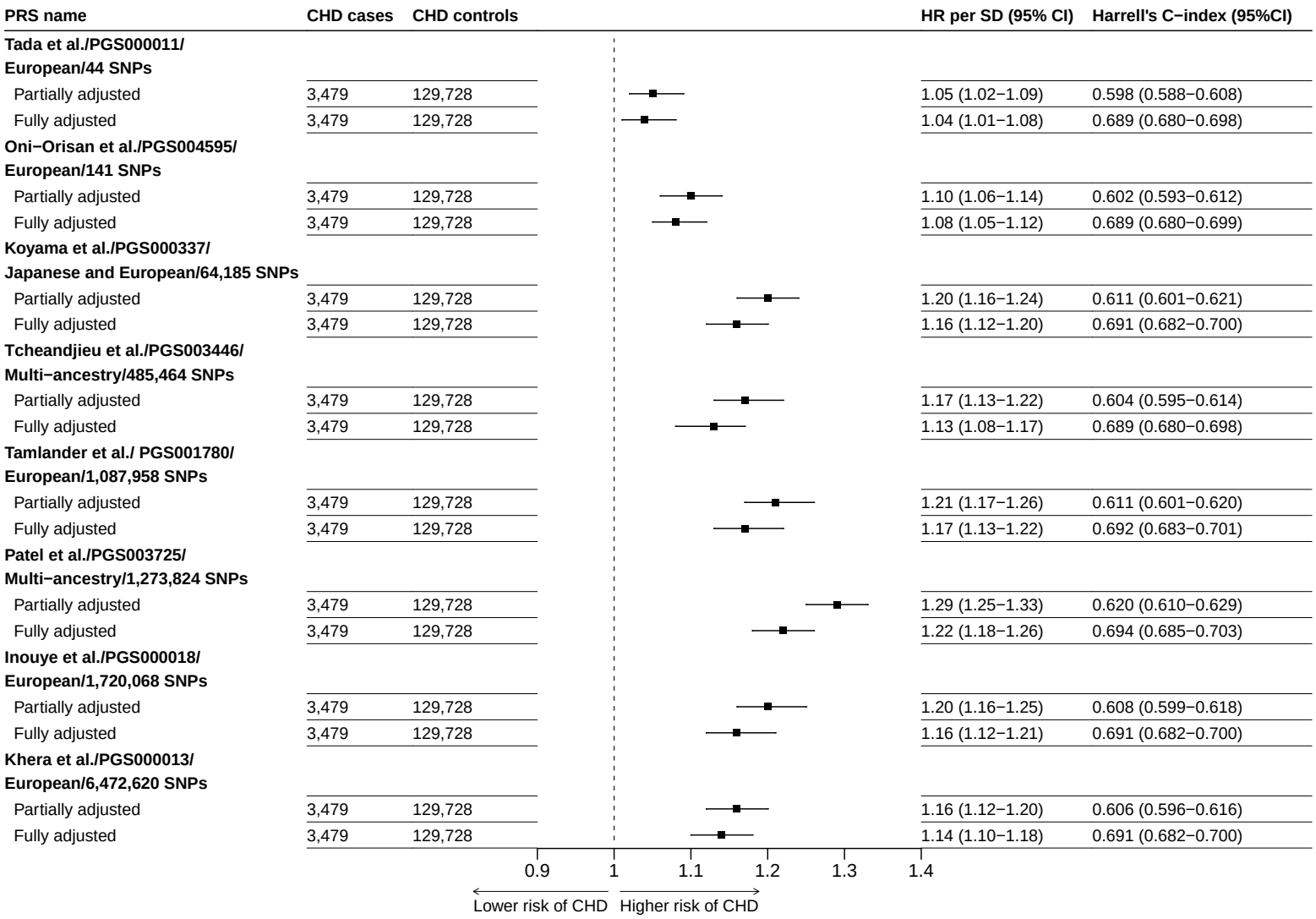


Figure 4.24: Hazard of death before age 80 with CHD listed anywhere on the death certificate, per 1SD increase in each PRS

Hazard ratios per 1SD increase in each PRS were estimated using Cox regression with adjustment for the same covariates included in the main logistic regression analyses.

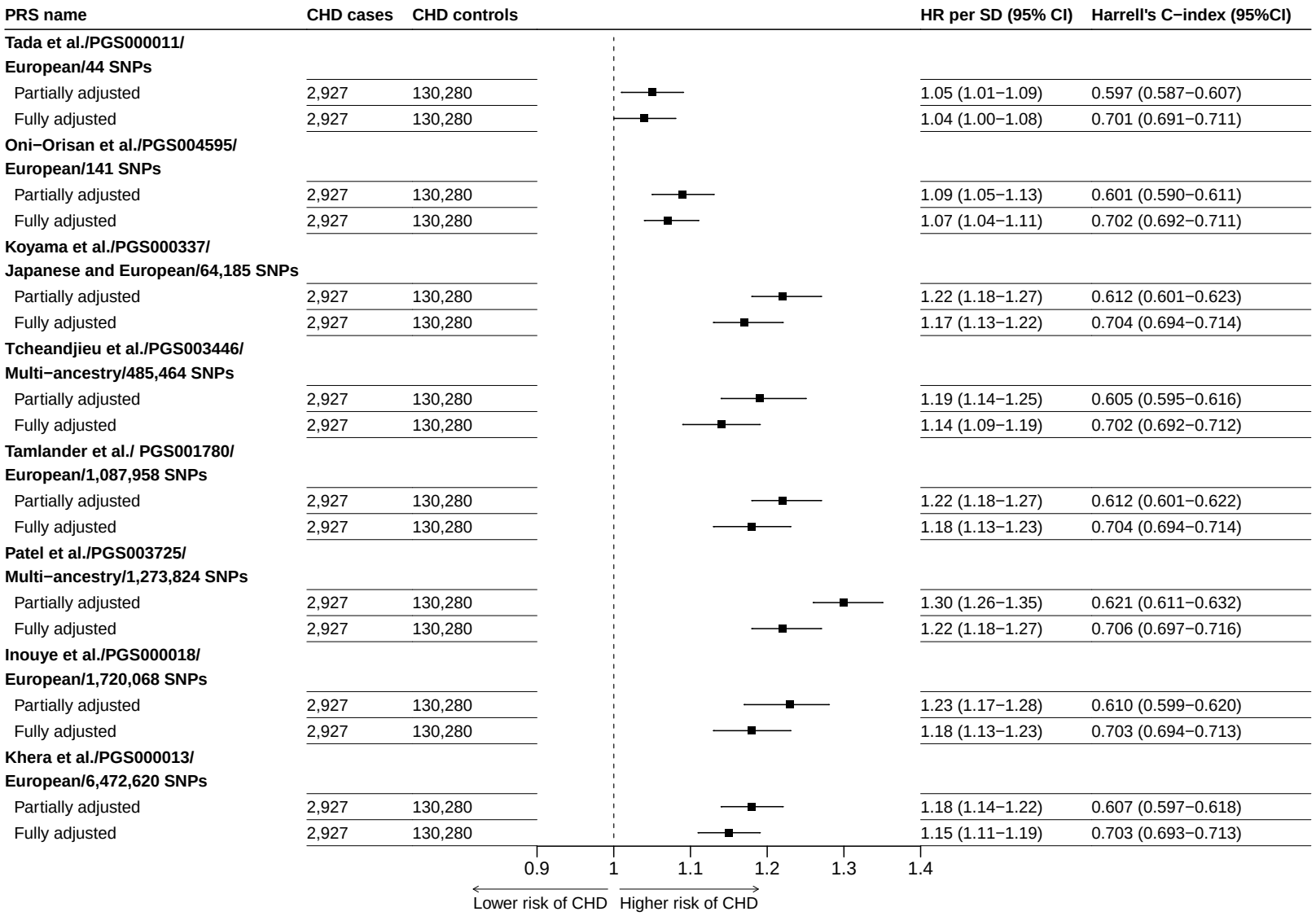


Figure 4.25: Hazard of death before age 80 with CHD listed as the primary cause on the death certificate, per 1SD increase in each PRS

Analyses as for Figure 4.24.

PRS name	CHD cases	CHD controls	OR per SD (95% CI)	AUC (95%CI)
Tada et al./PGS000011/ European/44 SNPs				
Partially adjusted	2,947	72,928	1.06 (1.02−1.10)	0.715 (0.707−0.724)
Fully adjusted	2,947	72,928	1.05 (1.01−1.09)	0.747 (0.739−0.755)
Oni−Orisan et al./PGS004595/ European/141 SNPs				
Partially adjusted	2,947	72,928	1.09 (1.05−1.13)	0.716 (0.707−0.724)
Fully adjusted	2,947	72,928	1.08 (1.04−1.12)	0.747 (0.738−0.755)
Koyama et al./PGS000337/ Japanese and European/64,185 SNPs				
Partially adjusted	2,947	72,928	1.23 (1.19−1.28)	0.720 (0.712−0.729)
Fully adjusted	2,947	72,928	1.21 (1.16−1.25)	0.750 (0.741−0.758)
Tcheandjieu et al./PGS003446/ Multi−ancestry/485,464 SNPs				
Partially adjusted	2,947	72,928	1.20 (1.14−1.25)	0.718 (0.710−0.727)
Fully adjusted	2,947	72,928	1.16 (1.11−1.22)	0.748 (0.740−0.757)
Tamlander et al./ PGS001780/ European/1,087,958 SNPs				
Partially adjusted	2,947	72,928	1.22 (1.17−1.27)	0.719 (0.710−0.728)
Fully adjusted	2,947	72,928	1.19 (1.15−1.24)	0.749 (0.740−0.757)
Patel et al./PGS003725/ Multi−ancestry/1,273,824 SNPs				
Partially adjusted	2,947	72,928	1.29 (1.25−1.34)	0.723 (0.714−0.731)
Fully adjusted	2,947	72,928	1.24 (1.20−1.29)	0.750 (0.741−0.758)
Inouye et al./PGS000018/ European/1,720,068 SNPs				
Partially adjusted	2,947	72,928	1.23 (1.17−1.28)	0.718 (0.710−0.727)
Fully adjusted	2,947	72,928	1.20 (1.14−1.25)	0.748 (0.740−0.757)
Khera et al./PGS000013/ European/6,472,620 SNPs				
Partially adjusted	2,947	72,928	1.18 (1.14−1.22)	0.718 (0.709−0.726)
Fully adjusted	2,947	72,928	1.17 (1.13−1.21)	0.748 (0.740−0.757)

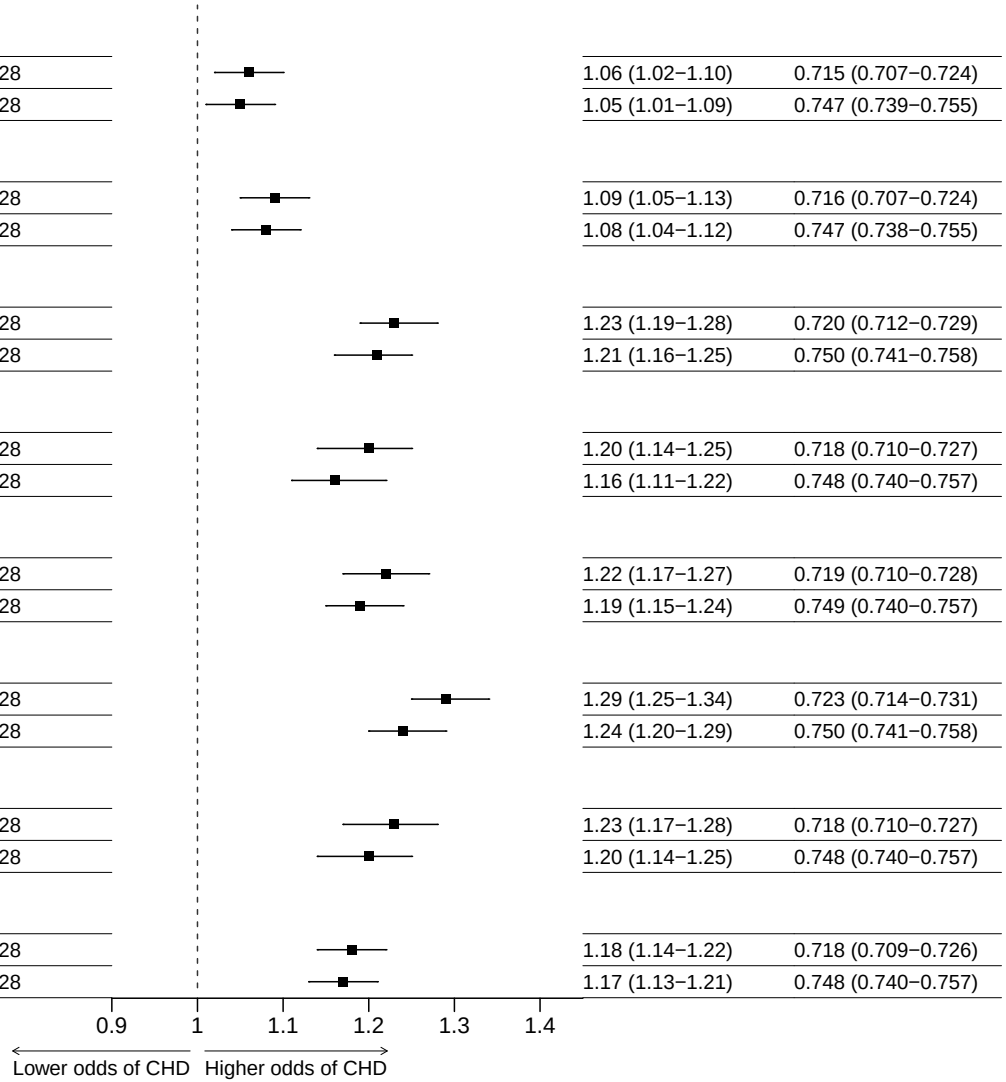


Figure 4.26: Sensitivity analyses (primary CHD definition) among participants unrelated to the 3rd degree

Analyses as for Figure 4.3.

4.5 References

1. Stark B, Johnson C, and Roth GA. “Global prevalence of coronary artery disease: an update from the global burden of disease study”. *Journal of the American College of Cardiology* 2024;83(13_Supplement):pp. 2320–2320.
2. McPherson R and Tybjaerg-Hansen A. “Genetics of Coronary Artery Disease”. *Circulation Research* 2016;118(4):pp. 564–78.
3. Zdravkovic S, Wienke A, Pedersen NL, Marenberg ME, Yashin AI, and De Faire U. “Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins”. *J Intern Med* 2002;252(3):pp. 247–54.
4. Cooper RS, Kaufman JS, and Ward R. “Race and Genomics”. *New England Journal of Medicine* 2003;348(12):pp. 1166–1170.
5. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, et al. “Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction”. *Circ Genom Precis Med* 2021;14(2):e003304.
6. King A, Wu L, Deng HW, Shen H, and Wu C. “Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease”. *BMC Medicine* 2022;20(1):p. 385.
7. Privé F, Arbel J, and Vilhjálmsson BJ. “LDpred2: better, faster, stronger”. *Bioinformatics* 2021;36(22-23):pp. 5424–5431.
8. Mak TSH, Porsch RM, Choi SW, Zhou X, and Sham PC. “Polygenic scores via penalized regression on summary statistics”. *Genetic Epidemiology* 2017;41(6):pp. 469–480.
9. Ge T, Chen CY, Ni Y, Feng YA, and Smoller JW. “Polygenic prediction via Bayesian regression and continuous shrinkage priors”. *Nature communications* 2019;10(1):p. 1776.
10. Richard Doll Consortium. *www.richarddollconsortium.org (Hosted by Oxford Population Health) using World Health Organization mortality and United Nations Population Division data.* Online Database. June 2024. `https://www.richarddollconsortium.org/projects/mortality-trends`.
11. Oni-Orisan A, Haldar T, Cayabyab MAS, Ranatunga DK, Hoffmann TJ, Iribarren C, et al. “Polygenic Risk Score and Statin Relative Risk Reduction for Primary Prevention of Myocardial Infarction in a Real-World Population”. *Clinical Pharmacology & Therapeutics* 2022;112(5):pp. 1070–1078.
12. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, et al. “Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease”. *Nature Genetics* 2020;52(11):pp. 1169–1177.
13. Tamlander M, Mars N, Pirinen M, FinnGen, Widen E, and Ripatti S. “Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes”. *Communications Biology* 2022;5(1):p. 158.
14. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. “Large-scale genome-wide association study of coronary artery disease in genetically diverse populations”. *Nature Medicine* 2022;28(8):pp. 1679–1692.
15. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. “Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention”. *J Am Coll Cardiol* 2018;72(16):pp. 1883–1893.
16. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. *Nature Medicine* 2023;29(7):pp. 1793–1803.
17. Tada H, Melander O, Louie JZ, Catanese JJ, Rowland CM, Devlin JJ, et al. “Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history”. *Eur Heart J* 2016;37(6):pp. 561–7.
18. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. *Nat Genet* 2018;50(9):pp. 1219–1224.

19. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. “Large-scale association analysis identifies new risk loci for coronary artery disease”. *Nature Genetics* 2013;45(1):pp. 25–33.

20. Coronary Artery Disease (C4D) Genetics Consortium. “A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease”. *Nat Genet* 2011;43(4):pp. 339–44.

21. Berglund G, Elmstähl S, Janzon L, and Larsson SA. “The Malmo Diet and Cancer Study. Design and feasibility”. *J Intern Med* 1993;233(1):pp. 45–51.

22. Erdmann J, Kessler T, Munoz Venegas L, and Schunkert H. “A decade of genome-wide association studies for coronary artery disease: the challenges ahead”. *Cardiovascular Research* 2018;114(9):pp. 1241–1257.

23. van der Harst P and Verweij N. “Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease”. *Circ Res* 2018;122(3):pp. 433–443.

24. Nikpay M, Goel A, Won H.-H, Hall LM, Willenborg C, Kanoni S, et al. “A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease”. *Nature genetics* 2015;47(10):pp. 1121–1130.

25. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. “Million Veteran Program: A mega-biobank to study genetic influences on health and disease”. *J Clin Epidemiol* 2016;70:pp. 214–23.

26. Ishigaki K, Akiyama M, Kanai M, Takahashi A, Kawakami E, Sugishita H, et al. “Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases”. *Nature Genetics* 2020;52(7):pp. 669–679.

27. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. “The UK Biobank resource with deep phenotyping and genomic data”. *Nature* 2018;562(7726):pp. 203–209.

28. Abraham G, Havulinna AS, Bhalala OG, Byars SG, De Livera AM, Yetukuri L, et al. “Genomic prediction of coronary heart disease”. *Eur Heart J* 2016;37(43):pp. 3267–3278.

29. Huang QQ, Sallah N, Dunca D, Trivedi B, Hunt KA, Hodgson S, et al. “Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals”. *Nature communications* 2022;13(1):p. 4664.

30. Gola D, Erdmann J, Läll K, Mägi R, Müller-Myhsok B, Schunkert H, et al. “Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease”. *Circ Genom Precis Med* 2020;13(6):e002932.

31. Saad M, El-Menyar A, Kunji K, Ullah E, Al Suwaidi J, and Kullo IJ. “Validation of Polygenic Risk Scores for Coronary Heart Disease in a Middle Eastern Cohort Using Whole Genome Sequencing”. *Circulation* 2022;Genomic and precision medicine.e003712.

32. Isgut M, Sun J, Quyyumi AA, and Gibson G. “Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later”. *Genome Medicine* 2021;13(1):p. 13.

33. Zaccardi F, Timmins IR, Goldney J, Dudbridge F, Dempsey PC, Davies MJ, et al. “Self-reported walking pace, polygenic risk scores and risk of coronary artery disease in UK biobank”. *Nutrition, Metabolism and Cardiovascular Diseases* 2022;32(11):pp. 2630–2637.

34. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, et al. “Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses”. *PLoS Medicine* 2021;18(1) (no pagination).

35. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, et al. “Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups”. *American Journal of Human Genetics* 2020;106(5):pp. 707–716.

36. Wünnemann F, Sin Lo K, Langford-Avelar A, Busseuil D, Dubé MP, Tardif JC, et al. “Validation of Genome-Wide Polygenic Risk Scores for Coronary Artery Disease in French Canadians”. *Circ Genom Precis Med* 2019;12(6):e002481.

37. Lambert SA, Wingfield B, Gibson JT, Gil L, Ramachandran S, Yvon F, et al. “The Polygenic Score Catalog: new functionality and tools to enable FAIR research”. *medRxiv* 2024:p. 2024.05.29.24307783.

38. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. “The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation”. *Nat Genet* 2021;53(4):pp. 420–425.

39. PGS Catalog Team. *PGS Catalog Calculator (in preparation)*. Online Database. June 2022. `https://github.com/PGScatalog/pgsc_calc`.

40. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, and Notredame C. “Nextflow enables reproducible computational workflows”. *Nature Biotechnology* 2017;35(4):pp. 316–319.

41. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. *GigaScience* 2015;4(1).

42. Plummer M. “Improved estimates of floating absolute risk”. *Stat Med* 2004;23(1):pp. 93–104.

43. Aragam KG, Dobbyn A, Judy R, Chaffin M, Chaudhary K, Hindy G, et al. “Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease”. *Journal of the American College of Cardiology* 2020;75(22):pp. 2769–2780.

44. Huang Y, Hui Q, Gwinn M, Hu YJ, Quyyumi AA, Vaccarino V, et al. “Sexual Differences in Genetic Predisposition of Coronary Artery Disease”. *Circulation. Genomic and Precision Medicine* 2021;14(1):e003147.

45. Byars SG and Inouye M. “Genome-Wide Association Studies and Risk Scores for Coronary Artery Disease: Sex Biases”. *Adv Exp Med Biol* 2018;1065:pp. 627–642.

46. Li L, Pang S, Starnecker F, Mueller-Myhsok B, and Schunkert H. “Integration of a polygenic score into guideline-recommended prediction of cardiovascular disease”. *Eur Heart J* 2024.

47. Lambert SA, Abraham G, and Inouye M. “Towards clinical utility of polygenic risk scores”. *Hum Mol Genet* 2019;28(R2):R133–r142.

48. Xiang R, Kelemen M, Xu Y, Harris LW, Parkinson H, Inouye M, et al. “Recent advances in polygenic scores: translation, equitability, methods and FAIR tools”. *Genome Medicine* 2024;16(1):p. 33.

49. de La Harpe R, Thorball CW, Redin C, Fournier S, Müller O, Strambo D, et al. “Combining European and U.S. risk prediction models with polygenic risk scores to refine cardiovascular prevention: the CoLaus|PsyCoLaus Study”. *Eur J Prev Cardiol* 2023;30(7):pp. 561–571.

50. Vassy JL, Posner DC, Ho YL, Gagnon DR, Galloway A, Tanukonda V, et al. “Cardiovascular Disease Risk Assessment Using Traditional Risk Factors and Polygenic Risk Scores in the Million Veteran Program”. *JAMA Cardiol* 2023;8(6):pp. 564–574.

51. Choi SW and O’Reilly PF. “PRSice-2: Polygenic Risk Score software for biobank-scale data”. *GigaScience* 2019;8(7).

Chapter 5: Developing a novel multi-ancestry coronary heart disease polygenic risk score for an admixed-American population

Summary

Due to the lack of large cohorts for people in Latin America, no coronary heart disease (CHD) polygenic risk score (PRS) has been identified as being derived from admixed-American-only or highly admixed-American dominant cohorts. The performance of PRSs developed from other ancestries would be attenuated when applied to Latin American populations, as unique genetic and non-genetic risk factors could not be best captured. As shown in **Chapter 2**, the association strengths of highly European dominant PRSs diminished when applied to non-European cohorts. The multi-ancestry CHD PRSs that incorporated genetic information from multiple ancestries during their construction showed stronger performance in **Chapter 4**, potentially due to improved generalisability, as causal variants are likely shared across populations. In addition, multi-ancestry studies allow for more accurate and precise estimates of the effects of this class of variants. Although a few earlier CHD PRSs incorporated genetic information from admixed-American cohorts during their construction, the representation of individuals from Latin America remained low. In this chapter, a novel multi-ancestry CHD PRS was developed by leveraging admixed-American genetic information from the Mexico City Prospective Study (MCPS) and another large admixed-American GWAS meta-analysis, marking the largest inclusion of admixed-American participants in a CHD PRS to date.

The MCPS data were split into training and testing sets based on the ratio of 4:1. The training set was used for PRS hyper-parameter tuning via a 10-fold cross validation (CV) approach. Three PRS construction algorithms were assessed (Pruning and Thresholding (P+T), LDpred2[1], and PRS-CSx[2]) and the best-performing hyper-parameters of each method were taken forward for evaluation in the testing set. During training, multi-ancestry candidate PRSs always outper-

formed admixed-American-only candidate PRSs. The best-performing MCPS-informed candidate PRS was constructed using the PRS-CSx[2] method and included 1,180,546 single nucleotide polymorphisms (SNPs). This PRS was selected as the novel MCPS-informed PRS. Leveraging diverse genetic information from admixed-Americans, Europeans and East-Asians, the novel PRS demonstrated strong association with CHD risk, with an odds ratio (OR) per standard deviation (SD) of 1.34 (95% CI, 1.26-1.42). This was comparable to the best performing external PRS in **Chapter 4** when applied to the same testing set (OR per SD=1.34, 1.26-1.43). However, the improvement in model discrimination conferred by the novel MCPS-informed PRS was minimal.

5.1 Background and aims

Given the under-representation of admixed-Americans in genome-wide association studies (GWAS) (less than 0.5%[3]), **Chapter 2** (literature review) did not identify any CHD PRS that was developed from admixed-American-only or highly admixed-American-dominant populations. It has also been shown in **Chapter 2** that PRSs derived from more advanced methods (i.e., machine learning or Bayesian) tend to yield more promising results[4–7]. However, these advanced PRS methods typically require a large sample size for PRS hyper-parameter tuning in order to ensure optimal performance and hence yield precise SNP estimates[1,8]. As a result, the development of CHD PRSs has progressed more rapidly among populations of European ancestries, as currently only European-descent cohorts are large enough to fully leverage these advanced PRS algorithms[1,2,8,9].

Due to the lack of non-European PRSs, previous studies have evaluated the genetic predisposition to CHD among populations of non-European ancestry using European-derived PRSs[4,10–12]. The performance of PRSs could be diminished when applied to a population that is different from the one that they were derived from. As shown in **Chapter 2 and 4**, the association strengths of European CHD PRSs became weaker when applied in populations of admixed-American ancestry. The differences in genetic and non-genetic (i.e., lifestyle, environmental, and biological)

CHD risk factors between Europeans and admixed-Americans suggest that a European CHD PRS may not accurately estimate the SNP effects in admixed American individuals. Several non-European studies have constructed CHD PRSs incorporating genetic data from diverse ancestries[10,13,14]. These PRSs demonstrated better generalisability compared to single-ancestry PRSs. As shown in **Chapter 4**, multi-ancestry PRSs (that included admixed-American genetic information) exhibited stronger associations with CHD risk among Mexicans than PRSs derived solely from European populations, although they did not significantly improve the predictive performance of CHD compared to other evaluated PRSs. Existing multi-ancestry CHD PRSs that incorporate admixed-American genetic information have been constructed using relatively small admixed-American sample sizes. The Patel *et al.* PRS[14] evaluated in MCPS has the largest admixed-American sample inclusion to date. The score included approximately 30,000 admixed-American participants from the Million Veterans Program (MVP)[15], a cohort in the United States of America (USA), in its input GWAS, but was trained solely among Europeans. Moreover, as introduced in **Chapter 1**, there is substantial variation in admixture patterns within admixed-American populations, and there has been no CHD PRS constructed incorporating genetic information specifically from Mexicans. Therefore, the assessed PRSs in **Chapter 4** may not fully account for the genetic architecture of admixed-American populations, particularly that of Mexico.

Over 130,000 participants in MCPS have both genetic and verified mortality data available, which is over four times more than the admixed-American samples used to derive the Patel *et al.* PRS[14]. Using data from MCPS, the largest available Mexican cohort and large external GWAS summary statistics of admixed-American, European, and East-Asian populations, a CHD PRS could be developed with a higher proportion of admixed-American representation, more accurately reflecting the complex genetic patterns in this population . This may also further enhance CHD risk prediction among Mexican adults. The aim of this chapter is therefore to develop a novel multi-ancestry PRS for CHD risk prediction by incorporating genetic information from MCPS and external GWAS sources of other ancestries, leveraging advanced PRS construction methods, and evaluating its

performance against the best performing external CHD PRS presented in **Chapter 4**. This novel MCPS-informed PRS includes a larger number of admixed-American individuals during its construction than any other existing CHD PRS.

5.2 Methods

MCPS data were used for PRS construction, the details of the cohort have been described in **Chapter 3**. All participants aged 35-79 at baseline, with genetic data available and verified mortality information were included in the analysis. The main outcome of interest for the analyses in the chapter was fatal or non-fatal CHD, with detailed definitions explained in **Chapter 3**.

5.2.1 PRS construction overview

Figure 5.1 illustrates the workflow for PRS construction. MCPS participants were initially divided into training and testing sets randomly based on a 4:1 ratio, with no sample overlap. The testing set was withheld and not included in any stages of the training process to ensure sample independence and avoid overfitting in the final internal testing stage.

5.2.1.1 PRS training stage

The training dataset, comprising 80% of the full MCPS cohort, was then used to run GWAS analyses to extract SNP information and to tune PRS hyper-parameters. To ensure all samples in the training set were used for both PRS construction and tuning validation, a 10-fold CV approach was adopted. Participants in the training set were randomly assigned to 10 groups using simple random sampling with no sample overlap. In each round of training, one group served as the validation set, while the remaining nine groups were used to perform a CHD GWAS. As a result, the GWAS set comprised 72% of the full MCPS dataset, while the validation set comprised 8%. **Table 5.1** displays the number and proportion of CHD cases and controls used in the GWAS and validation sets. The case rates remained relatively stable across each fold of the CV. As shown in **Chapter 2** (literature review) and **Chapter 4**, multi-ancestry PRSs outper-

formed single-ancestry PRSs among non-European cohorts. Two of the three PRS construction methods evaluated in this chapter (i.e. Pruning and Thresholding [P+T] and LDpred2[1]) only accept a single GWAS summary statistics file as input. The other method (PRS-CSx[2]) allows one ancestry-specific GWAS file for each of the included ancestries during PRS construction. Therefore, to compute multi-ancestry PRSs, multi-ancestry GWAS meta-analyses were performed. After the GWAS was conducted in MCPS, its summary statistics were then meta-analysed with external GWAS summary statistics to boost sample size, thereby providing more precise SNP effect estimates. Moreover, the GWAS meta-analysis enabled genetic information from diverse ancestries to be integrated, which could enhance the generalisability and transferability of the resulting PRS. Several multi-ancestry combinations were evaluated in this chapter (details in **Section 5.2.3 to 5.2.5**). The meta-analysed GWAS summary statistics were then used for the PRS construction algorithms.

For each PRS construction method, hyper-parameters needed to be tuned to identify the values that yield the best CHD predictive performance. Each hyper-parameter set was input into its corresponding algorithms, along with the meta-analysed GWAS summary statistics. The algorithms then output a list of selected SNPs and their corresponding weights, which were used to calculate a PRS for participants in the validation set. Each candidate PRS was standardised and evaluated for its performance within the validation set using a logistic regression model (details in **Section 5.2.7**). The performance of each PRS was assessed using the area under the receiver operating characteristic curve (AUC) for prediction and the odds ratio (OR) per standard deviation (SD) for association strength. This process was repeated 10 times, with each of the 10 groups taking a turn as the validation set and the remaining nine groups were used as the GWAS set. Therefore, for each hyper-parameter, 10 performance measures (AUC or OR per SD) were obtained, one from each fold of validation. The overall performance of each hyper-parameter set of each PRS method was then averaged across the 10 folds for comparison with each other to determine the best performing hyper-parameter (and the optimal multi-ancestry GWAS com-

binations) of each PRS method. The best performing PRS hyper-parameters of each method were then taken forward for evaluation in the testing set.

5.2.1.2 Testing set evaluation

A CHD GWAS was performed on the total training set data (80% of the full MCPS participants). The GWAS summary statistics were then meta-analysed with external GWAS summary statistics, using the optimal GWAS combinations that were determined during the training stage. These meta-analysed GWAS summary statistics were then input into the PRS construction algorithms of each method, along with the best performing hyper-parameters identified during PRS training. Using the SNP weights output by each algorithm, PRSs were created for participants in the testing set. These PRSs were then evaluated for their performance in the testing set, using logistic regression (details in **Section 5.2.8**). The best-performing candidate PRS in the testing set was selected as the novel MCPS-informed PRS.

5.2.2 Genome-wide association study

Details of the genotyping and imputation procedures for the MCPS genetic data are described in **Chapter 3**. Briefly, variants were genotyped using the Illumina GSAv2 array with positions mapped to the GRCh38 genome build[16]. After quality control (QC) steps, 140,831 samples were retained and underwent genotype imputation. Imputation was performed using the Trans-Omics for Precision Medicine (TOPMed) imputation server[17]. As mentioned in **Section 5.2.1**, a CHD GWAS was conducted on 90% of the training set data in each round of the 10-fold CV procedure. In addition, a CHD GWAS was performed on the full training set to construct candidate PRSs to be evaluated in the testing set. These 11 GWAS analyses were conducted using a machine learning based method called REGENIE[18]. Before performing each GWAS, the SNPs underwent per-SNP effective sample size filtering, based on the CHD case numbers in the corresponding GWAS set.

5.2.2.1 Per-SNP effective sample size filtering

The procedures used in a previous large-scale multi-ancestry GWAS[19] were followed for running the filtering steps. SNPs with imputation information $r^2 > 0.4$ and per-SNP effective sample size (N_{eff}) > 50 were retained for GWAS analysis. N_{eff} was calculated for each SNP using the following equation:

$$N_{\text{eff}} = 2 \times r^2 \times N_{\text{info}} \times \text{MAF}(1 - \text{MAF}) \tag{5.1}$$

Where r^2 is the imputation information, N_{info} is the number of informative samples which corresponds to the number of CHD cases in this study, and MAF is the minor allele frequency of each SNP.

5.2.2.2 REGENIE

REGENIE runs in two steps[18]. In the first step, SNPs from each chromosome are partitioned into consecutive blocks of a user-defined size B. Then, within each block, the algorithm fits a ridge regression model[20] with CHD as the outcome and all SNPs in the block as predictors, with an additional penalty term, λ, applied. With the penalty term included in the ridge regression, the estimated effect sizes for all SNPs are shrunk towards zero to control for overfitting. The optimal penalty is determined via J-fold CV, minimising the mean squared error between the predicted and observed disease risk. Once the optimal λ is selected, REGENIE refits the ridge regression model with the chosen penalty and the resulting effect size estimates are used to compute a single score per block. This score is calculated by linearly combining the product of the allele count of each SNP and its estimated effect size. SNPs with larger estimated effects are pulled less towards zero, therefore the resulting block score emphasise SNPs more strongly associated with the outcome. Performing this for every block reduces the hundreds of thousands of SNPs to one score per block.

The scores representing blocks across the whole genome are then combined into 23 leave-

one-chromosome-out (LOCO) scores. For each chromosome c, REGENIE fits a second ridge regression on block scores from all chromosomes except the chromosome c[18], again using J-fold CV to choose the optimal λ and thus determine weights for each input block score. The LOCO score for chromosome c is then computed by linearly combining the block scores from the other 22 chromosomes using their estimated weights from the second ridge regression. In the second step, the standard GWAS approach is carried out, where a univariate logistic regression is run between each SNP and the CHD outcome. In addition to adjusting for age, sex, and the first seven genetic principal components (PCs) in the GWAS, the model additionally adjusts for the LOCO score of the chromosome to which each SNP belongs to. Since this LOCO score excludes data from the chromosome that contains the SNP, the analysis accounts for relatedness and population structure without overfitting[18]. REGENIE has been shown to work efficiently for samples with high relatedness and population structure.

The GWAS analyses were conducted using the REGENIE software (version 3.1.3) in this chapter, with the block size B set to 1,000, and the number of CV folds J set to the default value of 5.

5.2.3 External datasets used for GWAS meta-analysis

As mentioned in **Section 5.2.1**, additional genetic study results were obtained to perform multi-ancestry GWAS meta-analysis during PRS training and testing. The studies and the cohorts included in each GWAS are briefly described below. For the meta-analysis, only GWAS summary statistics were used, and no individual-level data were accessed from any of the cohorts mentioned. The basic information for each external GWAS included is summarised in **Table 5.2**.

5.2.3.1 UK Biobank CHD GWAS

UK Biobank (UKB) is a population-based prospective study with over half a million participants recruited across the UK between 2004 and 2010, aged 40 to 69 years at the time of recruitment[21]. The majority of the participants are of European ancestry. Each participant was asked to complete a lifestyle and health questionnaire and to have their physical measurements taken.

Blood, urine, and saliva were collected and stored in a UK Biobank automated laboratory. Participants are being routinely followed up through linkage to national datasets for deaths, incident diseases, and hospital admissions. The blood samples taken were genotyped using either the Affymetrix Applied Biosystems UK BiLEVE Axiom Array (49,950 participants) or the Affymetrix Applied Biosystems UK Biobank Axiom Array (438,427 participants), and SNPs were mapped to the GRCh37 genome build. After quality control, genotype imputation was performed first using the Haplotype Reference Consortium (HRC)[22] as the reference panel and then using a merged UK10K[23] and 1000 Genomes phase 3 reference panel[24]. The two sets of imputed genotypes were then combined. If a SNP existed in both reference panels, the HRC imputation result was used[25].

At baseline, UKB participants could self-report CHD if they had experienced any of the following: myocardial infarction (MI), coronary angioplasty with stenting, coronary artery bypass grafting (CABG), or triple heart bypass surgery[26]. Incident CHD cases were also identified through linkage to Hospital Episode Statistics (HES), using ICD-10 codes I21–I25 and OPCS-4 procedure codes K40-K46, K49, K50, and K75[26].

The CHD GWAS on UKB participants was conducted in 2017[26]. Controls were defined as UKB participants with neither a personal nor a family history of CHD. The analysis included 34,541 CHD cases and 261,984 controls from the UKB cohort. Due to concerns about the reliability of the UK10K reference panel at the time of analysis, only SNPs imputed using the HRC panel were included. The GWAS analysis was conducted using mixed linear models implemented in BOLT-LMM[27], adjusted for age, sex, genotyping array, and the first 30 genetic PCs. The UKB CHD GWAS results are publicly available[26].

5.2.3.2 CARDIoGRAMplusC4D Consortium GWAS meta-analysis

The CARDIoGRAMplusC4D (CC4D) Consortium is a collaboration across multiple studies aimed at combining genetic data to identify risk loci for CHD and MI[28,29]. In 2015, the consortium con-

ducted a GWAS meta-analysis using 60,801 CHD cases and 123,504 controls from 48 studies[28]. Around 80% of all the included participants were of European ancestry and the rest were mostly of South and East Asian ancestry. CHD cases were included if the CHD definition used in the included studies encompassed any of the following: MI, acute coronary syndrome, chronic stable angina, or coronary stenosis >50%. The genotyping data from each study were mapped to genome build GRCh37 and were imputed to a 1000 Genomes phase 1v3 reference panel[24]. After QC, a GWAS was performed in each study using logistic regression (additive model) to evaluate the relationship between each SNP risk allele count and CHD risk. Variants that were present in at least 60% of the studies were subsequently meta-analysed across studies using an inverse-variance weighted (IVW) 'fixed-effect' meta-analysis, implemented with GWAMA[30]. If the variants showed significant heterogeneity, as indicated by Cochran Q statistics[31] and the I^2 index[32], they were reanalysed using a random-effects model. The CC4D GWAS meta-analysis results are publicly available[28].

5.2.3.3 Japanese CHD GWAS

The samples used in the Japanese CHD GWAS (JAP) came from multiple prospective cohorts in Japan. All the case samples analysed in the CHD GWAS were from Biobank Japan (BBJ). The controls used for the analysis came from the BBJ samples that did not have CHD and from other population-based prospective cohorts in Japan: the Tohoku Medical Megabank Project[33] (TMM), the Japan Public Health Center-based Prospective Study (JPHC)[34,35], and the Japan Multi-Institutional Collaborative Cohort Study (J-MICC Study)[36].

5.2.3.3.1 Biobank Japan

BBJ is a patient-based cohort that recruited approximately 200,000 individuals with one or more of the 47 prevalent or newly diagnosed common diseases between 2003 and 2008, from 66 hospitals across Japan[37]. At recruitment, clinical information and biological samples were collected from each participant. Clinical information was collected through reviewing medical records and

using a standardised questionnaire that covers lifestyle habits, physical and clinical measurements, and family history of common diseases. Medical records were recollected once a year until 2013 to review newly developed diseases. Mortality information are being collected for participants through linkage to the national death registry and deaths are certified using ICD-10 codes. For biological samples, blood was taken and DNA was subsequently extracted. The CHD outcome used for GWAS was defined as having either prevalent MI or stable or unstable angina.

5.2.3.3.2 Tohoku Medical Megabank Project

TMM was originally set up to tackle health-related issues after the tsunami and earthquake events. The study recruited approximately 80,000 participants aged 20 years or older from Miyagi and Iwate Prefectures of Japan between 2013 and 2016[33]. At recruitment, all participants were asked to fill out a study questionnaire that covered demographic factors, lifestyle habits, and medical history. Physiological measurements such as height and blood pressure were assessed and blood and urine samples were taken during recruitment. Participants are being followed up for morbidity using resurveys and by reviewing electronic health records (EHRs), the healthcare insurance system, and the cancer registry system.

5.2.3.3.3 Japan Public Health Center-based Prospective Study

JPHC is a population-based cohort that recruited over 140,000 participants aged 40 to 69 years across 11 health centres in Japan between 1990 and 1994[34,35]. The study mainly focused on studying risk factors associated with cancer and cardiovascular disease (CVD). At baseline, participants were asked to complete a standardised questionnaire on lifestyle, diet, and health conditions. Blood samples and health check-up data were taken from 48,000 participants. Resurveys on lifestyle and health conditions were conducted after five and 10 years of follow up. Participants are being followed up for mortality through linkage to the national death registry and for cancer and CVD incidence through linkage to local hospitals and population-based registries.

5.2.3.3.4 Japan Multi-Institutional Collaborative Cohort Study

J-MICC study is a prospective study with over 100,000 participants recruited between 2005 and 2010, aged 35-69 at recruitment[36]. The study was originally set up to study the prevention of cancer, using genetic determinants. At baseline, demographic, lifestyle, physiological, and health-related data were obtained from each participant through a standardised questionnaire. Blood samples were taken from all participants. Participants are followed up for incident cancer and death through follow-up surveys and linkage to cancer registries, the health insurance system, and the national death registry.

After DNA was extracted for the control participants, together with BBJ case participants, eligible samples were subsequently genotyped using the Illumina HumanOmniExpressExome BeadChip or a combination of the Illumina HumanOmniExpress and HumanExome BeadChips[38] with SNPs mapped to genome build GRCh37[39]. After QC, the genotype data were imputed using the 1000 Genomes phase 3 reference panel[39]. GWAS on CHD was performed using a generalised mixed model implemented in SAIGE[40], with adjustment for sex, age, and the first five genetic PCs. The analysis included 168,228 individuals (25,892 cases and 142,336 controls). GWAS results were made publicly available[38].

5.2.3.4 China Kadoorie Biobank CHD GWAS

China Kadoorie Biobank (CKB) is a blood-based prospective study that recruited approximately 0.5 million people between 2004 and 2008, aged 30 to 79[41]. Participants were recruited from five rural and five urban regions of China. At recruitment, demographic, socio-economic status, dietary and lifestyle information, and physical measurements were obtained from each of the participants. Participants have been followed up for the incidence of stroke, coronary heart disease (CHD), cancers, and diabetes through linkage to disease registries in eight study areas, and for mortality through linkage to regional death registries. Diseases and deaths were classified using ICD-10 codes, with CHD defined using ICD-10 codes I20-I25.

Blood samples were collected from all participants, and DNA was extracted and genotyped for 105,408 individuals using a custom-designed Affymetrix array[42], with SNP positions mapped to genome build GRCh38. Imputation was performed, following QC, using the 1000 Genomes phase 3 reference panel[39]. To address issues of case enrichment, GWAS for CHD was performed on a subset of 75,855 participants (13,748 CHD cases and 62,107 controls) that was representative of the CHD prevalence in CKB[42]. A GWAS was performed for each of the 10 regions, using generalised mixed models implemented in SAIGE[40], adjusted for sex, age, age^2, genotyping array version, and the first 11 genetic PCs[42]. The 10 region-specific GWAS results were then combined using IVW 'fixed effect' meta-analysis implemented in METAL[43].

5.2.3.5 Admixed-American GWAS meta-analysis

The admixed-American GWAS meta-analysis (HIS) summary statistics were obtained from unpublished results provided by our collaborator, with interim results previously presented at the American Society of Human Genetics (ASHG) annual conference[44]. The meta-analysis involved 144,882 admixed-American participants (20,450 cases and 124,432 controls) from nine cohorts in the US, described briefly below. For each cohort, unless stated otherwise, participant information on lifestyle, socio-economic status, and medical history was collected via a standardised questionnaire, along with physiological measurements and blood samples taken at the time of recruitment.

5.2.3.5.1 Hispanic Community Health Study/Study of Latinos

The Hispanic Community Health Study (HCHS)/Study of Latinos (SOL) is a large long-term study specifically aimed at researching health and disease problems among admixed-Americans living in the US[45,46]. From 2008 to 2011, the study recruited over 16,000 admixed-American individuals aged 18 to 74 at the time of recruitment. Participants are being followed up through follow-up interviews to identify health outcomes of interest. Genotyping data were obtained for approximately 13,000 participants using the Applied Biosystems TaqMan or the Sequenom iPLEX assays[47,48],

and of them 11,784 participants were included in the meta-analysis.

5.2.3.5.2 Women's Health Initiative

The Women's Health Initiative (WHI) is a study on the prevention and control of major complex diseases among post-menopausal women, established in 1993. The study is divided into a clinical trial and an observational study[49]. The observational study branch recruited over 93,000 women aged 50 to 79 at baseline, with 4.5% of them identified as of admixed-American ancestry. Participants have been followed up annually for health-related outcomes by mail. A subset of approximately 12,400 participants had their DNA data genotyped using the Illumina Multi-ethnic Genotyping Array (MEGA)[50]. Of them, 4,626 participants were of admixed-American ancestry and hence included in the meta-analysis.

5.2.3.5.3 BioMe Biobank Program

The BioMe Biobank is an EHR-linked prospective study founded in 2007 in New York aiming to optimise personal health care[51]. The study has been enrolling participants from hospitals regardless of ancestry, sex, age, or medical status. As of February 2021, the study had recruited over 56,000 adults, including more than 19,000 individuals of admixed-American ancestry. Participants are being constantly followed up using EHRs. Nearly all participants had DNA extracted and genotyped using one of the following platforms: Affymetrix Genome-Wide Human SNP arrays, OmniExpress Exome Array, MEGA, or Global Screening Arrays (the latter performed by the Regeneron Sequencing Center). In total, 18,279 participants from BioMe were included in the meta-analysis. Although CC4D also included participants from BioMe[28], they represented only approximately 2% of the study samples. Moreover, ongoing recruitment and follow-up led to a substantial increase in the number of participants and also CHD cases by 2021 compared to 2015. As a result, the overlap in participants and CHD cases between the admixed-American meta-analysis and CC4D is estimated to be approximately 20 to 25% of the BioMe samples included, representing less than 1% of the total participants in both CC4D and the admixed-

American meta-analysis.

5.2.3.5.4 Multiethnic Cohort

The Multiethnic Cohort (MEC) is a cohort established in 1993 to study the differences in cancer incidence among cohorts of five different ancestries (Japanese American, Native Hawaiians, African American, admixed-American, and white American) in Hawaii and Los Angeles in the US[52,53]. Between 1993 and 1996, the study recruited over 215,000 participants aged 45 to 75 at baseline, including 47,438 individuals of admixed-American ancestry. Participants were followed up through linkage to cancer and regional death registries, as well as follow-up surveys, which have been conducted every five years. Blood samples were collected for a subset of 70,000 MEC participants in 2004[53], from which DNA was extracted and genotyped using the Applied Biosystems TaqMan or the TaqMan OpenArray platforms[48]. Overall, 6,911 MEC admixed-American participants were included in the meta-analysis.

5.2.3.5.5 All of Us

The All of Us (AoU) program is a US-based prospective cohort established in 2015, with the goal of improving personalised disease prevention, treatment, and care[54,55]. The study began enrolment in 2018 with the aim of recruiting one million participants aged over 18 years across the US from diverse ancestries and backgrounds. As of 2023, over 413,000 individuals had been recruited to the study[56] and approximately 70% of participants could be followed up through linkage to EHRs. After DNA extraction from blood samples, a subset of them underwent whole-genome sequencing using the Illumina NovaSeq 6000 instrument[56]. A total of 45,590 admixed-American participants from AoU were included in the meta-analysis.

5.2.3.5.6 Vanderbilt DNA Databank

The Vanderbilt DNA Databank (BioVU) is a biobank established in 2007 by Vanderbilt University Medical Center that used an opt-out consent model, where patients receiving care at the

medical centre were automatically enrolled into the study unless they specifically requested to be excluded[57]. DNA was extracted from unused blood from clinical tests, and linked to detailed health data via de-identified EHRs. The DNA samples were subsequently genotyped using the MEGA array[58]. Overall, 1,856 participants from BioVU were of admixed-American ancestry and included in the meta-analysis.

5.2.3.5.7 Cameron County Hispanic Cohort

The Cameron County Hispanic Cohort (CCHC) is a community-based cohort established in 2004, aiming to study the risk factors for obesity and diabetes among Mexican Americans[59]. The study aimed to recruit participants aged over 18 years from Cameron County in the state of Texas on the US-Mexico border. By 2023, the study had recruited approximately 5,000 participants[60]. EHRs were used to obtain follow-up information for participants. DNA information was extracted from blood samples and subsequently genotyped using the Illumina MEGA array[61]. The meta-analysis included 4,306 participants from CCHC.

5.2.3.5.8 Million Veteran Program

The Million Veteran Program (MVP) aims to improve care for US military veterans by studying how health is associated with genetic, environmental, and lifestyle risk factors[15]. The study started recruitment in 2011 and has recruited over 885,000 veteran participants, with over 90% of them being men[15]. Participants are being followed up for clinical outcomes through medical records and other health administrative data. Participants who provided blood samples had their DNA extracted and genotyped using the Affymetrix Axiom Biobank Array. Overall, 47,269 admixed-American participants from MVP were included in the meta-analysis.

5.2.3.5.9 Mycode Community Health Initiative

In 2007, Geisinger Health System (GHS) in Pennsylvania of the US, launched the Mycode Community Health Initiative (hereon referred to as ”Mycode”) to link blood and DNA samples with

its EHR records[62]. Participants aged over 18 were recruited from Geisinger clinics, with blood samples taken. DNA was extracted from blood samples and subsequently genotyped using the Illumina OmniExpress array. Overall, 4,261 study participants from Mycode were of admixed-American ancestry and included in the meta-analysis.

CHD cases were identified from the nine cohorts (mentioned in **Sections 5.2.3.5.1 to 5.2.3.5.9**) included in the GWAS meta-analysis, either at baseline or through follow-up. Cases identified during follow-up were defined using ICD-9 codes 410-414 and 997.1, and ICD-10 codes I20-I25 (based on two or more occurrences), or evidence of having undergone percutaneous coronary intervention (PCI) or CABG. Genotype data for each cohort, except AoU, which has whole genome sequenced data available, were imputed using the TOPMed reference panel[17], with SNPs mapped to GRCh38 genome build positions. A CHD GWAS was performed in each study using SAIGE to fit generalised mixed models[40]. Variants with MAF<0.01 were excluded from the analysis. The GWAS results from each study were then meta-analysed together using IVW 'fixed effect' meta-analysis implemented in METAL[43].

5.2.4 Lift-over

For the UKB, CC4D, and Japanese cohorts, SNP positions were mapped to the GRCh37 reference genome. In contrast, the MCPS, CKB, and admixed-American GWAS meta-analysis cohorts used GRCh38 for SNP mapping (see **Table 5.2**). Therefore, to achieve alignment of the SNP mappings in all cohorts in order to conduct GWAS meta-analysis in this chapter, lift-over was performed for SNPs recorded in UKB, CC4D, and Japanese CHD GWAS summary statistics from GRCh37 to GRCh38 using the UCSC genome browser tool and chain file[63].

5.2.5 GWAS meta-analysis

CHD GWAS summary statistics from MCPS participants were meta-analysed with external GWAS summary statistics in each round of the 10-fold CV using IVW 'fixed effect' meta-analysis implemented in METAL[43]. For each SNP, the software extracted its associated effect size and stan-

dard error from each GWAS and performed the meta-analysis using the equations as follows:

$$\beta_{\text{meta}} = \frac{\sum_i \beta_i w_i}{\sum_i w_i} \quad \text{and} \quad \text{SE}_{\text{meta}} = \sqrt{\frac{1}{\sum_i w_i}}, \qquad w_i = \frac{1}{\text{SE}_i^2} \tag{5.2}$$

Where β_i is the effect size reported by the i^{th} included GWAS for that SNP, SE_i is the standard error associated with β_i, and w_i is the weight associated with β_i calculated by taking the reciprocal of the squared standard error SE_i (inverse variance). β_{meta} and SE_{meta} are the meta-analysed effect size and standard error for that SNP, respectively. If the SNP is missing in one of the included GWAS, the software still performs the meta-analysis using the remaining available GWAS. If a SNP exists in only one GWAS, β_{meta} and SE_{meta} will use the SNP effect size and standard error from that GWAS.

As briefly mentioned in **Section 5.2.1**, PRS algorithms P+T[64] and LDpred2[1] take a single set of GWAS summary statistics as input and PRS-CSx[2] takes ancestry-specific GWAS summary statistics for each of the included ancestries. Using the MCPS CHD GWAS and five external GWAS (described in **Section 5.2.3** and **Table 5.2**), four GWAS meta-analyses were performed in each round of the 10-fold CV (see **Table 5.3** for details), with the aim of assessing the best ancestry combinations during PRS training. These GWAS meta-analyses were used as input for P+T[64] and LDpred2[1].

For PRS-CSx[2], the admixed-American-specific GWAS input (HIS-meta in **Table 5.3**) was created by meta-analysing the external admixed-American GWAS (described in **Section 5.2.3.5**) with the MCPS GWAS. Additionally, the CC4D[28] and the UKB[26] GWAS results were meta-analysed to form a European-specific CHD GWAS (EUR-meta), and the JAP[38] and the CKB[42] GWAS results were combined into an East Asian-specific CHD GWAS (EAS-meta) (see **Table 5.3**). The details of each PRS algorithm are provided in **Section 5.2.6**.

5.2.6 PRS construction methods

Three PRS construction algorithms (P+T[64], LDpred2[1], and PRS-CSx[2]) were evaluated in this chapter, each requiring input of several hyper-parameters that needed tuning along with the aforementioned GWAS meta-analysis summary statistics. For each method and hyper-parameter set, the models were run 10 times using a 10-fold CV approach. This CV procedure identified the optimal hyper-parameter set for each method, with the aim of maximising predictive performance while ensuring generalisability. Since the number of included SNPs varied for different PRSs, each raw score had a different score range, which would impact the association strength estimated in later analyses. To avoid bias, each candicate PRS output by each method and hyper-parameter set was subsequently standardised. **Table 5.4** provides a summary of hyper-parameter sets used to tune each of the PRS methods evaluated in this chapter.

5.2.6.1 Pruning and Thresholding

Pruning and Thresholding (P+T) is one of the most widely adopted methods for constructing PRSs[65], and was implemented here using the PRSice[64] software. The method filters out SNPs from the input GWAS summary statistics that are in high linkage disequilibrium (LD) and less associated with the outcome based on user-defined thresholds, denoted as r^2 and p, respectively. Hence, the software performs two steps. Firstly, in each genomic window (the software used ± 250kb), SNPs are sorted in ascending order based on their significance level in the input GWAS. Then, using the genetic data of the validation set, maximum-likelihood haplotype frequencies are estimated for each SNP pair to compute Pearson correlation r, which is then squared to obtain r^2 for SNPs in that window (to estimate LD). Starting from the most significant SNP in the GWAS (with the lowest p-value), the algorithm removes any other SNPs whose r^2 is above the user-defined threshold . Secondly, among the remaining SNPs after LD pruning, any SNPs with a p-value exceeding the user-defined significance threshold, p, are removed. The resulting set of SNPs and their effect sizes in the input GWAS are linearly combined to construct the

PRS (see **Chapter 3, Section 3.3.1**) for participants in the validation set. The user-defined hyper-parameters r^2 and p that were evaluated in the training stage are detailed as follows (also summarised in **Table 5.4**):

1. r^2:0.2 to 1 in intervals of 0.1 (9 values)
2. p: a total of 266 values, comprising
    - Discrete values: $5x10^{-8}$, $5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, 1 (5 values)
    - Continuous range: $5x10^{-4}$ to 0.05 in the interval of $5x10^{-4}$ (100 values)
    - Continuous range: 0.1-0.9 in the interval of 0.005 (161 values)

In total, 2,394 hyper-parameter combinations were input into the algorithm, alongside the four GWAS meta-analysis summary statistics (**Table 5.3**), resulting in 9,576 PRSs computed in the validation set during each round of the 10-fold CV.

5.2.6.2 LDpred2

LDpred[66] and LDpred2[1] have been widely used for PRS construction in recent years[6,13,14] and have demonstrated strong performance in disease risk prediction as shown in **Chapter 2**. The method uses the input GWAS summary statistics and LD information between SNPs to infer SNP weights through a Bayesian framework. First, a point-normal mixture prior is assumed for each SNP effect sizes, which is a weighted sum of a fixed probability of the SNP effect size being zero and a normal distribution. More precisely, for each effect size β_i of SNP_i, it is assumed that they are distributed as follows:

1. Point mass at zero: P(β_i=0)=1-p
2. Normal distribution: $\beta_i \sim$ iid $N(0, \sigma^2)$, where $\sigma^2 = \frac{h^2}{Mp}$

Where p is the probability that SNP_i is causal and is drawn from a normal distribution. In other words, p is the fraction of causal markers in the input GWAS. M is the total number of SNPs in

the GWAS and h^2 is the heritability captured by the M SNPs. Thus, Mp is the expected number of causal SNPs. Combining the two distributions together yields:

$$\beta_i \sim \text{iid} \begin{cases} \text{N}\left(0, \frac{h^2}{Mp}\right), & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases} \tag{5.3}$$

Hence, the prior of the model depends on two user-defined hyper-parameters, heritability h^2 and the fraction of causal markers p. Although heritability h^2 can be estimated using LD score regression (LDSC)[67], the method developer suggested tuning for the heritability h^2 to maximise predictive accuracy. LDpred2 includes an additional hyper-parameter, sparsity, which sets the posterior β_i to zero if the posterior causal probability p_i, for SNP_i, is smaller than the user-defined threshold p[1]. In other words, if p_i<p, the algorithm assumes that SNP_i has no causal effect on the outcome of interest.

The algorithm allows for a special case where p=1 (i.e., assuming all SNPs are causal). This changes the model to an infinitesimal model (LDpred-inf) with the prior as follows:

$$\beta_i \sim \text{iid N}(0, \sigma^2), \text{ where } \sigma^2 = \frac{h^2}{M} \tag{5.4}$$

This LDpred-inf model was also evaluated in the analysis and used as a reference. Since all markers are assumed to be causal (p=1), LDpred-inf does not have the sparsity option. For simplicity, h^2 for the LDpred-inf model was estimated using LDSC.

Using the genetic data of the validation set, LDpred2 first calculates the LD matrix and then together with the GWAS summary statistics, LDpred2 uses Gibbs sampling through Markov chain Monte Carlo (MCMC)[68] to compute the posterior mean effect size for each SNP. The PRS is then calculated for each participant in the validation set through a linear combination of the product of each SNP allele count and its associated posterior effect size (see **Chapter 3, Section 3.3.1**). The algorithm was implemented using the bigsnpr package in R[69].

During the analysis, only SNPs present in the HapMap3+[70] variants list (1,444,196 SNPs) were used, as recommended by the method developers, so that the software could run efficiently while enabling good coverage of the genome. In addition, the algorithm requires an effective sample size for each input SNP. The effective sample sizes for SNPs were calculated based on the equation below:

$$N_{\text{eff(LDpred)}} = \frac{4}{\frac{1}{N_{\text{case}}} + \frac{1}{N_{\text{control}}}} \tag{5.5}$$

Where N_{case} represents the number of participants with the disease, and $N_{control}$ represents the number of participants without the disease. For each SNP, N_{case} and $N_{control}$ were calculated by summing the numbers of CHD cases and controls only from those GWAS studies where the SNP was available during GWAS meta-analysis. For example, if a SNP was existed in three out of the four GWAS studies included in the GWAS meta-analysis, the cases and controls were summed across those three studies only.

If the input GWAS was a GWAS meta-analysis, another way of computing the effective sample size was recommended by the method developer[1]:

$$\begin{aligned} T_i &= \frac{8}{\text{SE}_i^2}, \\ N_{\text{eff(LDpred)}}^{\text{meta-analysis}} &= \text{quantile}(T, 0.999) \end{aligned} \tag{5.6}$$

Where SE_i is the standard error associated with the effect size of each SNP_i, and the effective sample size is calculated as the 99.9th percentile of all T_i. Therefore, for LDpred2, three hyper-parameters were tuned and two methods for calculating the effective sample size were evaluated. The tuning grid was defined as follows (also summarised in **Table 5.4**):

1. h^2: 0.01 to 0.2 in the interval of 0.01 (20 values)
2. p: a total of 108 values, comprising
    - Discrete values: $5x10^{-10}$, $5x10^{-9}$, $5x10^{-8}$, $5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, $5x10^{-4}$, $5x10^{-3}$ (8

values)

- Continuous range:0.01 to 1 in the interval of 0.01 (100 values)

3. Sparsity: True or False (2 values)

4. Effective sample size calculation: standard and meta-analysis-specific (2 methods)

Two LDpred-inf models were also evaluated, one for each effective sample size calculation method. In total, 8,640 hyper-parameter combinations and two LDpred-inf models were input into the algorithm, alongside four sets of GWAS meta-analysis summary statistics (**Table 5.3**) in each round of the 10-fold CV. However, for those hyper-parameter combinations for which the MCMC model failed to converge, no SNP posterior effect size (i.e., weight) was produced, and thus no PRS was computed for these combinations in the validation set.

5.2.6.3 PRS-CSx

PRS-CSx is a relatively new method for constructing PRS that was developed to account for genetic variation from diverse ancestries using a Bayesian approach[2]. The method is an extension of PRS-CS[9], a method that has been shown to construct PRS with high predictive performance for disease outcomes among populations of European ancestry[71]. PRS-CSx uses a similar algorithm to PRS-CS, but allows for more than one input GWAS. The method first models GWAS summary statistics from each ancestry separately before integrating the genetic information across ancestries to improve polygenic risk prediction in diverse populations. For each ancestry-specific GWAS, the method assumes an ancestry-shared continuous shrinkage prior on the SNP effect sizes. In the GWAS of ancestry k, for each effect size β_{ik} of SNP_i, it is assumed that:

$$\beta_{ik} \sim \mathrm{N}\left(0, \frac{\sigma_k^2}{N_k}\psi_i\right), \quad \psi_i \sim \mathrm{Gamma}(a, \delta_i), \quad \delta_i \sim \mathrm{Gamma}(b, \phi), \tag{5.7}$$

$$\text{In default setting, } a = 1,\ b = \frac{1}{2}$$

Where N_k is the sample size of the GWAS and σ_k^2 is the residual variance for ancestry k. ϕ is a user-defined hyper-parameter that controls the global shrinkage of all SNPs, ψ_i is the local parameter that controls SNP-specific shrinkage, depending on ϕ. Both ϕ and ψ_i are independent of k and shared across ancestries. Hence this continuous shrinkage prior allows correlated (similar distribution shape) but varying SNP effect sizes across ancestries. The two-level Gamma-Gamma prior on each local shrinkage parameter ψ_i concentrates much of its mass around zero, allowing strong shrinkage of small effect sizes, while its heavy tail allows truly large SNP effects to avoid a strong penalty. SNPs that have strong effect sizes in one ancestry will receive less penalisation in the others due to the shared ψ_i across ancestries. The algorithm has an 'auto' option, which estimates the user-defined ϕ automatically using input GWAS summary statistics.

With a GWAS and LD reference panel for each ancestry, the algorithm estimates the ancestry-specific posterior mean effect sizes for each SNP using MCMC. The multi-ancestry SNP effect sizes are obtained by combining ancestry-specific posterior effects using IVW 'fixed-effect' meta-analysis. The PRS-CSx algorithm was implemented using the Python package of the same name[2].

During the analysis, European, admixed-American, and East-Asian ancestry-specific LD reference panels constructed using 1000 Genomes phase 3[39] were used (provided by the method developer). For SNP identification, the PRS-CSx software requires SNP rsIDs, rather than chromosome and position identifiers, which were used in the GWAS and other PRS construction methods. To enable this conversion, the R packages GenomicRanges[72] and SNPlocs.Hsapiens.dbSNP155.GRCh38[73] were used to map chromosome and position identifiers to SNP rsIDs for all SNPs in the input GWAS files and the genetic files of the validation set. After SNP effect sizes (ancestry-specific and meta-analysed) were estimated by the method software, they were linearly combined, first within each chromosome then across all chromosomes using PLINK2[74] (as described in **Chapter 3, Section 3.3.1**) to compute ancestry-specific and multi-ancestry candidate PRSs for participants in the validation set. Both ancestry-specific and multi-ancestry PRSs

were evaluated.

As illustrated above, PRS-CSx requires one ancestry-specific GWAS for each ancestry input. GWAS results of the same ancestry have been combined together using meta-analysis (see **Section 5.2.5**). For the analysis, three ancestry combinations were evaluated (see **Table 5.4** for details), with the aim of maintaining consistency with the multi-ancestry GWAS input used for the P+T and LDpred2 methods.

There was one hyper-parameter, the global scaling ϕ, that required tuning. Based on the recommendation of the method developer, four values were tested: $1x10^{-6}$, $1x10^{-4}$, 0.01, 1. In addition, the 'auto' option for ϕ was also evaluated.

In total, five hyper-parameter values and three ancestry-specific GWAS combinations were input into the PRS-CSx algorithm, resulting in 15 multi-ancestry and 40 ancestry-specific PRSs generated in each round of the 10-fold CV.

5.2.7 PRS training

In each round of the 10-fold CV, standardised PRSs were computed for participants in the validation set using the various methods and hyper-parameter combinations detailed in **Section 5.2.6**. To evaluate the performance of each PRS alone, logistic regression was performed using the (standardised) PRS as the sole predictor and CHD as the outcome, with no additional covariate adjustments. The performance of each PRS was evaluated using two metrics. First, the OR per SD of the PRS was used to assess its association strength with CHD risk. Second, the model AUC was used to measure the predictive ability of each PRS. After completing the 10-fold CV, each hyper-parameter set of each method produced 10 performance measurements. The overall performance of each hyper-parameter set of each method was calculated by averaging the results across the 10 CV folds. For mean AUC, this was done by computing the mean of the 10 AUC values. OR per SD was obtained by exponentiating the regression coefficient of the PRS. To calculate the mean OR per SD across the 10 validation folds, the mean

of the original regression coefficients was first computed, and then the exponential of that mean was taken. These averages were then compared across hyper-parameter sets to identify the best-performing hyper-parameter set (including the optimal GWAS combination) for each PRS method, which was taken forward to the testing stage.

5.2.8 PRS testing

In the testing stage, a CHD GWAS was performed on the full training data (see **Section 5.2.1** and **Figure 5.1**) and the resulting summary statistics were meta-analysed with external GWAS datasets using the optimal GWAS combination identified for each PRS construction method. For each PRS method, the best-performing hyper-parameter set (identified during training based on mean AUC or based on OR per SD) was input into the corresponding algorithm. The output SNP weights were then used to construct PRS for participants in the testing set, which was withheld and not used during any stages of the training process.

To evaluate the performance of these candidate PRSs, we used logistic regression with each PRS as the predictor and CHD as the outcome. Models were first run without any covariate adjustments ('no adjustment'). The rest of the analyses followed the analytical approaches outlined in **Chapter 3, Section 3.3.1**. Briefly, PRSs were first treated as categorical variables (with five equally-sized groups and the lowest group as the reference) to assess the shape of the associations. Then PRSs were treated as continuous variables to assess the average strengths of associations. Under partial adjustment, models were adjusted for age, sex, and the first seven genetic PCs. Under full adjustment, models additionally included covariates that were risk factors for CHD: the highest level of education attainment, systolic and diastolic blood pressure (SBP, DBP), smoking status, waist-to-hip ratio (WHR), and diabetes status at baseline. Missing values were imputed using strategies listed in **Chapter 3**. AUC was estimated to assess model discrimination. The PRS with the strongest performance based on AUC was retained as the novel MCPS-informed PRS and carried over to **Chapter 6** to construct integrated scores. The

performance of the PRSs in the testing set was also compared with the best-performing external PRS (Patel *et al.*[14]) in **Chapter 4**.

5.3 Results

5.3.1 PRS training

The CHD GWAS of MCPS participants did not identify any novel genetic loci. Similarly, the GWAS meta-analysis did not reveal additional novel associations when compared with previously published CHD GWAS studies. The detailed results of the admixed-American CHD GWAS meta-analysis (including MCPS data) will be presented in a separate publication. The following sections present the results of PRS training using P+T, LDpred2[1] and PRS-CSx[2].

5.3.1.1 Pruning and Thresholding PRS

Figure 5.2 and 5.3 compare the overall performance of PRSs generated from each hyper-parameter set and GWAS combination with the P+T method, averaged across the 10 validation folds. PRSs generated using GWAS based solely on admixed-American participants showed relatively lower performance compared to multi-ancestry PRSs, both in terms of mean AUC and OR per SD. For multi-ancestry PRS, except when r^2=1, PRS performance generally peaked at lower p-value thresholds. Generally, as r^2 increased, more SNPs were retained, and better performance was observed. When r^2 was set to 1, no SNPs were removed due to high LD, as pairwise correlation cannot exceed 1. In this case, SNPs were filtered only based on the p-value threshold. Under this condition, PRS performance did not improve significantly by including more SNPs (setting a higher p-value cut-off). Using input GWAS that included admixed-American and European ancestries (UKB+CC4D+HIS+MCPS), with hyper-parameters r^2= 0.9 and p-value threshold=0.0055, achieved the highest performing PRS both in mean AUC (0.5556) and OR per SD (1.21).

5.3.1.2 LDpred2 PRS

As mentioned in **Section 5.2.6.2**, some hyper-parameter combinations may fail to converge in the LDpred2[1] algorithm and not produce SNP posterior effect sizes (i.e., weights). Therefore, to ensure a fair comparison between PRSs, only the hyper-parameter sets that converged in all 10 folds of the 10-fold CV were put forward for further evaluation. The majority of the non-convergence could be attributed to h^2. **Table 5.5** shows the number of hyper-parameter sets for which all 10 folds reached convergence across the 20 h^2 levels. In total, each level of h^2 was expected to produce 432 PRSs (108 fraction of causal marker thresholds × 2 sparsity settings × 2 effective sample size calculation methods). Non-convergence was associated with the ancestry included in the input GWAS. For PRSs generated with multi-ancestry GWAS, the number of hyper-parameter sets that reached algorithm convergence in all 10 CV folds decreased rapidly as the h^2 level increased. From h^2=0.09 onwards, none of the hyper-parameter sets with multi-ancestry input GWAS reached convergence in all 10 folds. On the other hand, PRSs generated using GWAS based solely on admixed-American ancestry showed a stable convergence rate as the h^2 level increased.

Figure 5.4 and Figure 5.5 display the performance of LDpred2 PRSs across different hyper-parameter set values and GWAS combinations. Since no hyper-parameter sets reached convergence after h^2=0.09 for multi-ancestry GWAS inputs, the figures only show the PRS performance up to h^2=0.1. For PRS generated using genetic information solely from admixed-Americans, the AUC and OR per SD performance were very similar regardless of input parameters, and were generally lower than multi-ancestry PRSs. For LDpred-inf PRSs generated using multi-ancestry GWAS, effective sample sizes calculated using the GWAS meta-analysis-specific method (Eq.5.6) outperformed those calculated using the standard approach (Eq.5.5). However, overall, LDpred-inf PRSs showed lower mean AUC and OR per SD compared to LDpred2 PRSs generated using user-defined hyper-parameters. PRSs generated by incorporating admixed-American (including external meta-analysis), European, and East-Asian ancestry infor-

mation demonstrated the highest performance in both mean AUC and OR per SD. However, the associated optimal hyper-parameter sets differed between the two performance metrics. PRS generated with hyper-parameter sets with h^2=0.02, p=0.38, sparsity=True, and using standard effective sample size calculation achieved the highest mean AUC, 0.5623. In contrast, PRS generated with hyper-parameter sets with h^2=0.04, p=0.23 and the same sparsity and effective sample size calculation method achieved the highest mean OR per SD, 1.26.

5.3.1.3 PRS-CSx PRS

PRS-CSx outputs ancestry-specific PRS for each input ancestry, as well as a multi-ancestry PRS, which is generated using SNP weights meta-analysed from the ancestry-specific weights. **Figure 5.6 and Figure 5.7** compare the performance of PRS-CSx PRS across hyper-parameter values and ancestry-specific GWAS combinations inputs. Multi-ancestry PRSs generally outperformed ancestry-specific PRSs regardless of hyper-parameter values (including the 'auto' option), both in terms of mean AUC and OR per SD. After 'borrowing' information from other ancestry-specific GWAS through the ancestry-shared hyper-parameter ϕ, the admixed-American ancestry-specific PRS outperformed ancestry-specific PRSs of other ancestries. In contrast, the East-Asian ancestry-specific PRSs demonstrated the lowest performance. Using ancestry-specific GWAS from admixed-Americans, Europeans, and East-Asians, the PRS generated with ϕ=1x10^{-4} demonstrated the strongest performance in both mean AUC (0.5663) and OR per SD (1.27).

Table 5.6 and Table 5.7 present the mean AUC and OR per SD for the best-performing PRS hyper-parameter sets of the three PRS methods, along with their corresponding optimal GWAS combinations. Overall, the best performing PRS was generated using PRS-CSx, which outperformed PRSs from other methods in both performance metrics. The best performing hyper-parameter set was the same for PRSs generated using P+T and PRS-CSx based on the two performance metrics, while it was slightly different for PRS generated using LDpred2. There-

fore, four PRS hyper-parameter sets (and three GWAS combinations) were taken forward to the testing stage.

5.3.2 PRS testing

All four MCPS-informed candidate PRSs were associated positively and log-linearly with the odds of CHD (see **Figure 5.8**). Those in the top 20% group for CHD genetic risk had 2.01 to 2.23 times higher odds of CHD compared to the reference (lowest) group. **Figure 5.9** compares the association strengths of the four MCPS-informed candidate PRSs and the Patel *et al.* PRS[14], the best-performing external PRS in **Chapter 4**, with CHD in the testing set. The PRS generated by PRS-CSx (included 1,180,546 SNPs) demonstrated the strongest associations, of all candidate PRSs, with CHD risk. With no other adjustments, a 1SD increase in the score was associated with a 34% increase in the odds of CHD (OR per SD=1.34, 95%CI 1.26-1.42). In contrast, the PRS generated using the P+T method showed the weakest association, with a 1SD increase associated with a 27% increase in the odds of CHD (OR per SD=1.27, 1.19-1.34). The strength of the associations remained robust after additional adjustment for age, sex, and genetic PCs, and persisted even with further adjustment for established CHD risk factors. However, only the PRS generated using PRS-CSx demonstrated an association strength comparable to that of the Patel *et al.* PRS[14] (OR per SD=1.34, 1.26-1.43).

In terms of predictive ability, the model AUCs for all five evaluated PRSs were similar across the different model adjustment approaches. The AUC was approximately 0.57-0.58 for models with no adjustments, 0.71 for the partially adjusted models, and 0.75 for the fully adjusted models. Only the PRS generated using PRS-CSx slightly outperformed the model AUC of the Patel et al. PRS[14], with the difference observable only at the third decimal place. The PRS-CSx PRS was retained as the novel MCPS-informed PRS for score integration in **Chapter 6**.

5.4 Discussion

5.4.1 Summary of findings

In this chapter, through the evaluation of thousands of combinations of PRS methods, hyper-parameter sets and GWAS meta-analyses, a novel CHD PRS was developed. CHD candidate PRSs were developed using P+T[64], LDpred2[1], and PRS-CSx[2] and trained through a 10-fold cross-validation approach to identify the optimal hyper-parameter sets and GWAS combinations for each method. During the training stage, multi-ancestry PRSs outperformed ancestry-specific PRSs across all methods. The best-performing P+T PRS incorporated genetic ancestry information from European and admixed-American populations. In contrast, the best-performing PRSs derived from LDpred2[1] or PRS-CSx[2] used genetic information from admixed-American, European, and East-Asian ancestries. All the PRSs carried over to the testing stage were associated positively and log-linearly with CHD risk at ages 35-79. The novel multi-ancestry CHD PRS developed using the PRS-CSx[2] method (with 1,180,546 SNPs), demonstrated superior performance compared to candidate PRSs developed by the other two methods. This score leveraged the genetic information across three ancestries, admixed American, European, and East-Asian, drawing from the GWAS data and included 159,172 CHD cases and 806,532 controls. Notably, this PRS included the largest number of admixed-American participants to date (240,791), eight times larger than the next most representative PRS[14]. The score was both trained and tested exclusively among individuals of admixed-American ancestry in the MCPS cohort. Although the PRS-CSx[2] PRS did not significantly improve CHD predictability, its performance was comparable to the best-performing PRS evaluated previously in **Chapter 4** (i.e., Patel *et al*. PRS[14]) while including over 92,000 fewer SNPs. Due to the absence of available admixed-American or indigenous-American cohorts with adequate sample size, the novel MCPS-informed PRS derived in this chapter could not be evaluated externally.

5.4.2 Limitations of PRS construction methods developed for single ancestry populations

Both P+T and LDpred2[1] were originally developed for PRS construction within a single ancestry population and did not account for population heterogeneity, ancestry-specific LD patterns, or allele frequency differences when they were designed[2,75]. Although previous studies using these two methods with multi-ancestry GWAS inputs reported improved performance over ancestry-specific PRSs[10,13,76,77], their limitations when applied to multi-ancestry GWAS inputs were observed in our analysis.

For PRSs generated using the P+T method, incorporating GWAS data from European-descent populations improved performance. In contrast, adding genetic information from East-Asian ancestries did not lead to further improvements in association strength or predictive ability. This may be attributed to a mismatch in LD patterns between the ancestries represented in the input GWAS samples and those in the validation set[2,75]. As mentioned in **Section 5.2.6.1**, pairwise correlations between SNPs were estimated using genetic data from the validation set[64], which consisted exclusively of admixed-Americans with a high degree of indigenous American ancestry[16]. These correlation estimates were then used to prune SNPs in the input GWAS, which included individuals from multiple ancestries. This mismatch in ancestry composition between the validation dataset and the input GWAS may have led to both over-pruning and under-pruning of SNPs. PRSs incorporating European GWAS were more predictive, possibly due to shared genetic ancestry stemming from historical colonisation as introduced in **Chapter 1**[78].

The LDpred2[1] algorithm experienced non-convergence more frequently when using multi-ancestry GWAS inputs, whereas this issue was not observed with admixed-American-only GWAS. As the heritability prior input value increased, the number of models that successfully converged dropped sharply, an effect that was particularly evident for GWAS including all three ancestries (over half of the input hyper-parameter sets did not converge when the heritability prior h^2 reached 0.05). Similar to the P+T method, the LD matrix used for SNP weight computation in LDpred2

was also estimated using genetic data from individuals in the validation set[1]. The mismatch in LD structure between the ancestries represented in the input GWAS and the validation set would likely cause the MCMC algorithm to fail to converge, as the algorithm may encounter conflicting signals between the SNP covariance estimated from the LD matrix (estimated from the Mexican-only validation dataset) and the covariance implied by the multi-ancestry GWAS meta-analysis z-scores. As the heritability prior increases, a higher proportion of SNPs are assumed to have stronger non-zero (causal) effects. These conflicts then become more pronounced, placing greater computational burden on the MCMC process and hence increasing the likelihood of non-convergence. However, despite these limitations, both methods showed that multi-ancestry PRS outperformed admixed-American-only PRS. This is likely due to the fact that multi-ancestry GWAS studies normally have greater statistical power, thereby enabling more precise effect sizes to be estimated for true causal variants shared across ancestries. Moreover, pooling genetic information from other ancestries could also complement effect size estimates for SNPs that are rare in admixed-American populations. These benefits possibly offset the limitations in PRS construction algorithms.

Although this analysis included over 240,000 admixed-American participants for CHD PRS construction, the largest sample ever for this ancestry, the number was still significantly smaller than that of the European ancestry group included, which was approximately twice as large. Future efforts should focus on building larger-scale cohorts of admixed-American populations to enable more powerful GWAS, thereby allowing more accurate estimation of genetic predisposition to CHD within admixed-Americans.

5.4.3 Multi-ancestry PRS construction method

PRS-CSx[2] was developed specifically for constructing multi-ancestry PRS to improve disease risk prediction in underrepresented ancestries. It models ancestry-specific LD and allele frequencies, and borrows information from well-powered ancestries (e.g., European) to improve

estimates in under-powered ancestries. The method has been shown to outperform other single-ancestry approaches (e.g., PRS-CS[9] and LDpred2[1]) when applied to diverse ancestries[77,79]. Among candidate PRSs generated using PRS-CSx[2], while the multi-ancestry PRS demonstrated the best overall performance, admixed-American-specific PRSs ranked second (when using HIS-meta for admixed-American-specific GWAS input) and outperformed those developed for other ancestries when applied to the MCPS population. Although current admixed-American GWAS still have much lower sample sizes, the results suggest that, with adequately powered datasets, admixed-American-specific PRSs are likely to offer the most accurate risk predictions for this population[2] (**Figure 5.6 and Figure 5.6**).

5.4.4 Comparing the MCPS-informed PRS with external CHD PRS

The MCPS-informed PRS developed using PRS-CSx[2] outperformed PRSs developed from other methods and showed comparable performance to the Patel *et al.* PRS[14], which was the best-performing external CHD PRS in **Chapter 4**. The Patel *et al.* PRS[14] is a multi-ancestry, multi-trait PRS developed using a two-stage approach (details provided in **Chapter 4**). The score incorporates genetic information from CHD-associated traits and risk factors across five ancestries (admixed-Americans, African, East and South-Asian and European). Ancestry-specific PRS of each CHD-related trait were first combined and then trait-specific PRS were combined into the final PRS. Due to the complex structure involved in its construction, replicating the Patel *et al.* PRS[14] for use in another disease context would be challenging. Not only does the score require ancestry-specific GWAS for CHD but it also relies on ancestry-specific GWAS for CHD-related traits. For diseases with less common risk factors, such trait-specific GWAS may be limited or unavailable, potentially reducing the predictive performance of the resulting PRS. For example, metabolic risk factors have been found to increasingly promote cancer[80,81], but GWAS for some of the metabolic risk factors such as high fasting plasma glucose remain limited in scale and power[82]. Moreover, the selection of ancestry-specific scores to include in each trait-specific PRS was based on the Akaike Information Criterion (AIC) from logistic regressions. As a result,

for less common traits, non-European ancestry-specific PRSs that are informative but statistically underpowered may be excluded, limiting the applicability and generalisability of the PRS in under-represented populations.

The PRS generated by PRS-CSx[2] reached comparable performance to the Patel *et al*. PRS[14] both in terms of association strength and discriminatory ability, despite being constructed using GWAS data for only a single trait, CHD. This could be partly explained by the fact that this PRS includes approximately eight times more individuals of admixed-American ancestry (MCPS + admixed-American GWAS meta-analysis) during its construction than the Patel *et al.* PRS[14], allowing CHD genetic predisposition for this population to be more precisely estimated. Moreover, the PRS-CSx[2] method allows underpowered ancestry-specific GWAS to be integrated with well-powered GWAS from other ancestries, enabling power gains for under-represented groups. This ensures that underpowered ancestries can still contribute meaningfully to the final PRS construction, rather than being excluded. The PRS hyper-parameter tuning method employed in this chapter is also reproducible, as it requires only ancestry-specific GWAS for the disease of interest and is relatively straightforward to implement.

5.5 Conclusion

To conclude, this chapter evaluated three PRS algorithms (P+T[64], LDpred2[1] and PRS-CSx[2]) across thousands of hyper-parameter combinations using a 10-fold cross-validation approach, to develop a novel CHD PRS with the largest inclusion of admixed-American participants to date. PRS methods designed for single-ancestry (P+T[64] and LDpred2[1]) demonstrated limitations when applied to multi-ancestry GWAS meta-analysis inputs. The novel MCPS-informed CHD PRS was generated using PRS-CSx[2], which incorporated genetic information from over 965,000 individuals from admixed-American, European, and East-Asian and included 1,180,546 SNPs. Internal testing of the PRS demonstrated comparable performance to the best-performing external PRS in **Chapter 4**. However, the MCPS-informed PRS leveraged genetic information from all input

ancestry-specific GWAS, regardless of their statistical power, and employed a simpler construction process. Consequently, this approach can be easily replicated for other disease outcomes.

In both this chapter and **Chapter 4**, it is evident that CHD PRSs explain only a fraction of the total variability in CHD risk. Vascular risk factors substantially improved model discrimination when included as covariates, and they predicted CHD risk independently of the PRS. Therefore, in the next chapter, PRSs will be integrated with clinical risk scores to assess whether this combination can further enhance CHD risk prediction.

Table 5.1: Cross validation data split case control proportion

Fold	GWAS set		Validation set	
	Case (%)	Control (%)	Case (%)	Control (%)
1	3723 (3.88 %)	92186 (96.12 %)	426 (4.00 %)	10231 (96.00 %)
2	3679 (3.84 %)	92231 (96.16 %)	470 (4.41 %)	10186 (95.59 %)
3	3741 (3.90 %)	92168 (96.10 %)	408 (3.83 %)	10249 (96.17 %)
4	3736 (3.90 %)	92174 (96.10 %)	413 (3.88 %)	10243 (96.12 %)
5	3784 (3.95 %)	92125 (96.05 %)	365 (3.42 %)	10292 (96.58 %)
6	3727 (3.89 %)	92182 (96.11 %)	422 (3.96 %)	10235 (96.04 %)
7	3728 (3.89 %)	92181 (96.11 %)	421 (3.95 %)	10236 (96.05 %)
8	3723 (3.88 %)	92187 (96.12 %)	426 (4.00 %)	10230 (96.00 %)
9	3736 (3.90 %)	92173 (96.10 %)	413 (3.88 %)	10244 (96.12 %)
10	3764 (3.92 %)	92146 (96.08 %)	385 (3.61 %)	10271 (96.39 %)

Table 5.2: Characteristics of external GWAS used in GWAS summary statistics

Name	Genome build	Ancestry	CHD case/control (Sample size)
UK Biobank (UKB [25])	GRCh37	European-dominant	34,541/261,984 (296,525)
CARDIoGRAMplusC4D Consortium (CC4D[27])	GRCh37	European-dominant	60,801/123,504 (184,305)
Japanese CHD GWAS (JAP[37])	GRCh37	East-Asian	25,892/142,336 (168,228)
China Kadoorie Biobank (CKB[41])	GRCh38	East-Asian	13,748/62,107 (75,855)
Admixed-American GWAS meta-analysis (HIS)	GRCh38	Admixed-American	20,450/124,432 (144,882)

Table 5.3: Summary of GWAS meta-analysis performed

Name	Included GWAS	Included ancestry	ten-fold CV	CHD case/control (sample size)
HIS+MCPS (HIS-meta)	1. MCPS GWAS 2. Admixed-American GWAS Meta-analysis	Admixed-American-only	Yes	24,190 / 216,601 (240,791)
UKB+CC4D+HIS+MCPS	1. UKB CHD GWAS 2. CC4D GWAS meta-analysis 3. MCPS GWAS 4. Admixed-American GWAS Meta-analysis	Admixed-Americans and European	Yes	119,532 / 602,089 (721,621)
UKB+CKB+CC4D+JAP+MCPS	1. UKB CHD GWAS 2. CKB CHD GWAS 3. CC4D GWAS meta-analysis 4. Japanese CHD GWAS 5. MCPS GWAS	Admixed-Americans, European and East-Asian	Yes	138,722 / 682,100 (820,822)
UKB+CKB+CC4D+JAP+HIS+MCPS	1. UKB CHD GWAS 2. CKB CHD GWAS 3. CC4D GWAS meta-analysis 4. Japanese CHD GWAS 5. MCPS GWAS 6. Admixed-American GWAS Meta-analysis	Admixed-Americans, European and East-Asian	Yes	159,172 / 806,532 (965,704)
EUR-meta	1. UKB CHD GWAS 2. CC4D GWAS meta-analysis	European-only	No	95,342 / 385,488 (480,830)
EAS-meta	1. CKB CHD GWAS 2. Japanese CHD GWAS	East-Asian only	No	39,640 / 204,443 (244,083)

The MCPS GWAS sample size is 95,909. While the number of cases and controls varies across each fold, the summary table uses 3,740 cases and 92,169 controls as representative values.

Table 5.4: Summary of hyperparameters selected for tuning for each PRS method

Method name	GWAS input (ancestry included)	Hyperparameter sets
Pruning and Thresholding	1. HIS+MCPS 2. UKB+CC4D+HIS+MCPS 3. UKB+CKB+CC4D+JAP+MCPS 4. UKB+CKB+CC4D+JAP+HIS+MCPS	**r^2**: 0.2 to 1 in the interval of 0.1 (9 values)
		p: a total of 266 values, comprising: • Discrete values: $5x10^{-8}$, $5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, 1 (5 values) • Continuous range: $5x10^{-4}$ to 0.05 in the interval of $5x10^{-4}$ (100 values) • Continuous range: 0.1-0.9 in the interval of 0.005 (161 values)
LDpred2[1]	1. HIS+MCPS 2. UKB+CC4D+HIS+MCPS 3. UKB+CKB+CC4D+JAP+MCPS 4. UKB+CKB+CC4D+JAP+HIS+MCPS	**h^2**: 0.01 to 0.2 in the interval of 0.01 (20 values)
		p: a total of 108 values, comprising: • Discrete values: $5x10^{-10}$, $5x10^{-9}$, $5x10^{-8}$, $5x10^{-7}$, $5x10^{-6}$, $5x10^{-5}$, $5x10^{-4}$, $5x10^{-3}$ (8 values) • Continuous range:0.01 to 1 in the interval of 0.01 (100 values)
		Sparsity: True or False (2 values)
		Effective sample size calculation: 2 methods
		LDpred-inf
PRS-CSx[2]	1. HIS-meta+EUR-meta (Admixed-American, and European) 2. MCPS+EUR-meta+EAS-meta (Admixed-American, European and East-Asian) 3. HIS-meta+EUR-meta+EAS-meta (Admixed-American, European and East-Asian)	**ϕ:** $1x10^{-6}$, $1x10^{-4}$, 0.01, 1.
		PRS-CSx-auto

Table 5.5: LDpred2 convergence model counts

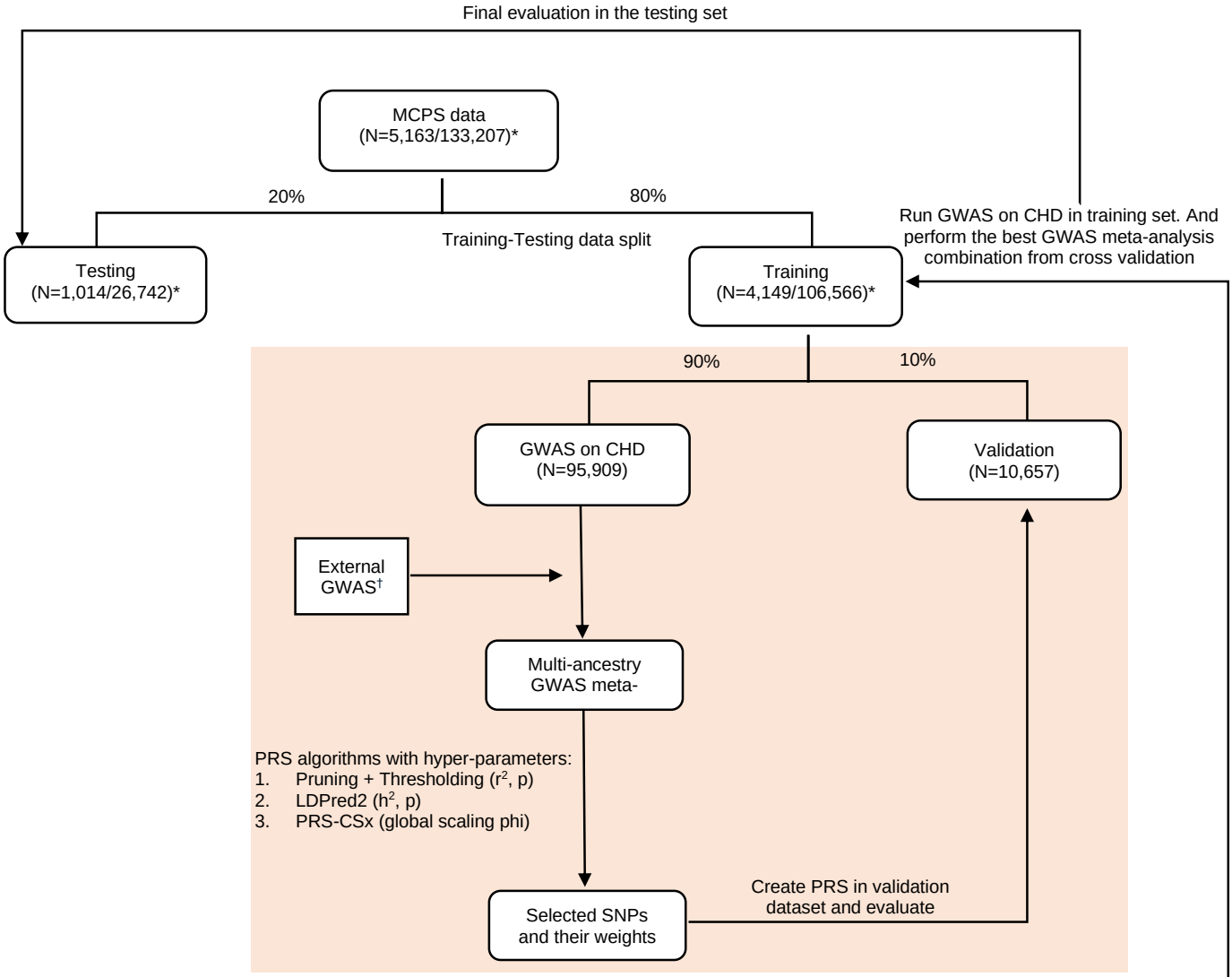
h^2	Converged model counts for each GWAS input			
	UKB+CKB+CC4D+JAP+MCPS	UKB+CC4D+HIS+MCPS	HIS+MCPS	UKB+CKB+CC4D+JAP+HIS+MCPS
0.01	400	404	432	400
0.02	388	396	431	388
0.03	368	391	431	369
0.04	316	381	432	315
0.05	181	358	432	182
0.06	32	310	428	36
0.07	0	208	428	0
0.08	0	59	429	0
0.09	0	4	427	0
0.1	0	0	424	0
0.11	0	0	428	0
0.12	0	0	425	0
0.13	0	0	425	0
0.14	0	0	425	0
0.15	0	0	424	0
0.16	0	0	423	0
0.17	0	0	425	0
0.18	0	0	423	0
0.19	0	0	421	0
0.2	0	0	422	0

Table 5.6: Mean AUC comparison of the best performing PRS of each method

Method	GWAS	Parameters	Mean AUC
Pruning and Thresholding	UKB+CC4D+HIS+MCPS Ancestry included: Admixed-American and European	r^2= 0.9 p=0.0055	0.5556
LDpred2	UKB+CKB+CC4D+JAP+HIS+MCPS Ancestry included: Admixed-American, East Asian and European	h^2=0.02 p=0.38 sparsity=True Normal effective sample calculation	0.5623
PRS-CSx	HIS-meta+EUR-meta+EAS-meta Ancestry included: Admixed-American, European and East-Asian	ϕ =1e-04 Ancestry=Multi	0.5663

Table 5.7: Mean OR per SD comparison of the best performing PRS of each method

Method	GWAS	Parameters	Mean OR per SD
Pruning and Thresholding	UKB+CC4D+HIS+MCPS Ancestry included: Admixed-American and European	r^2= 0.9 p=0.0055	1.21
LDpred2	UKB+CKB+CC4D+JAP+HIS+MCPS Ancestry included: Admixed-American, East Asian and European	h^2=0.04 p=0.23 sparsity=True Normal effective sample calculation	1.26
PRS-CSx	HIS-meta+EUR-meta+EAS-meta Ancestry included: Admixed-American, European and East-Asian	ϕ =1e-04 Ancestry=Multi	1.27



Model AUC				
	Parameter 1	Parameter 2	Parameter 3	…
AUC 1				
AUC 2				
AUC 3				
AUC 4				
AUC 5				
AUC 6				
AUC 7				
AUC 8				
AUC 9				
AUC 10				
Mean AUC	**0.74**	**0.75**	**0.73**	

Figure 5.1: PRS construction flowchart

*The numbers presented are in the case/control format. †The analysis used five external CHD GWAS summary statistics of European (UKB, CC4D), East-Asian (CKB, JAP) and admixed-American (HIS) ancestries (see Table 5.2).

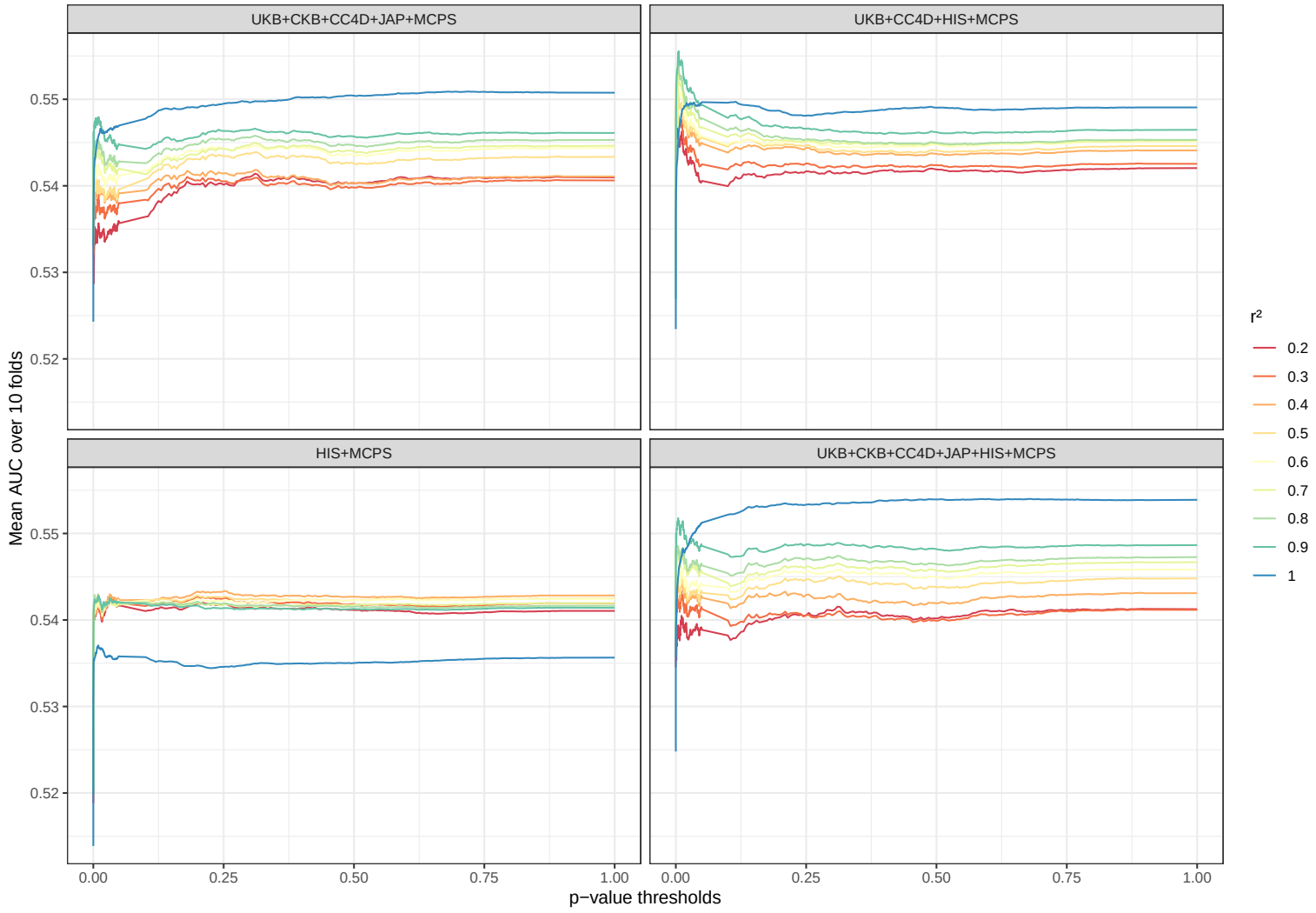


Figure 5.2: Mean AUC comparison of P+T algorithms parameters

During the 10-fold CV, the candidate P+T PRSs were evaluated in regression models without any additional adjustments. The mean AUC was calculated by averaging the model AUC for each PRS across the 10 folds.

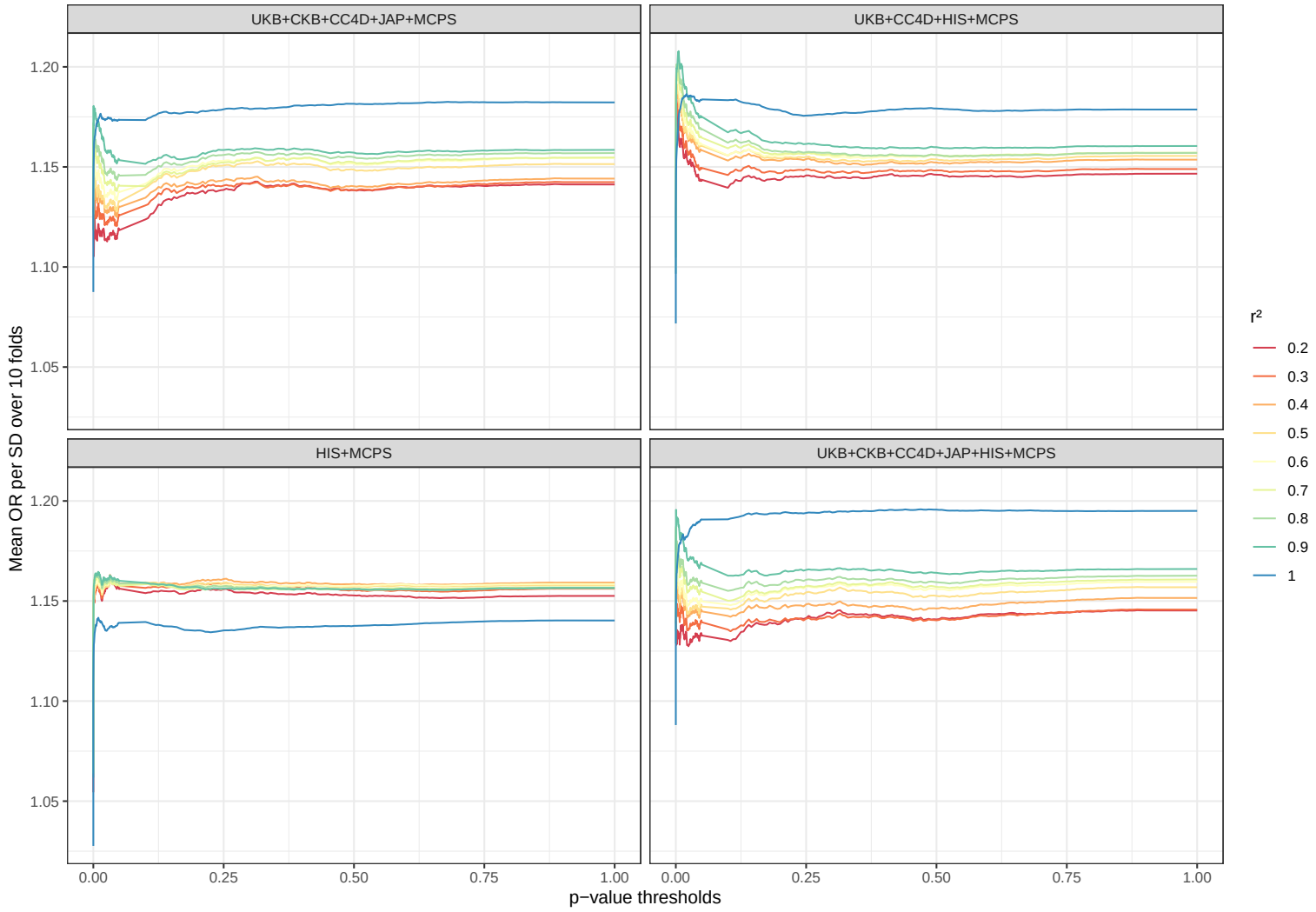


Figure 5.3: Mean OR per SD comparison of P+T algorithms parameters

Analysis as for Figure 5.2. The mean OR per SD was calculated by averaging the original PRS regression coefficients (β) across the 10 folds, and then taking the exponential of that mean.

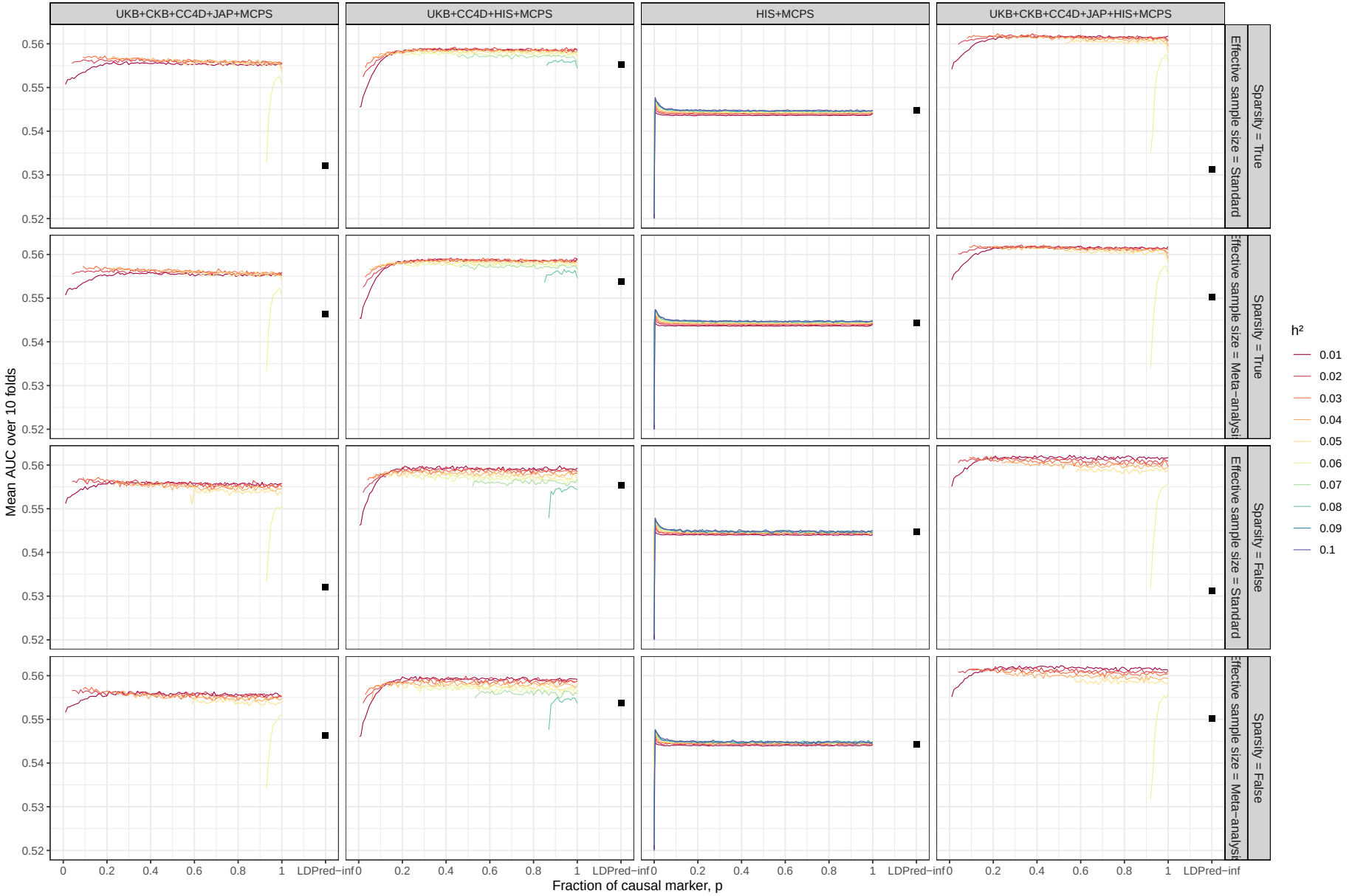


Figure 5.4: Mean AUC comparison of LDpred2 algorithms parameters

Analysis as for Figure 5.2.

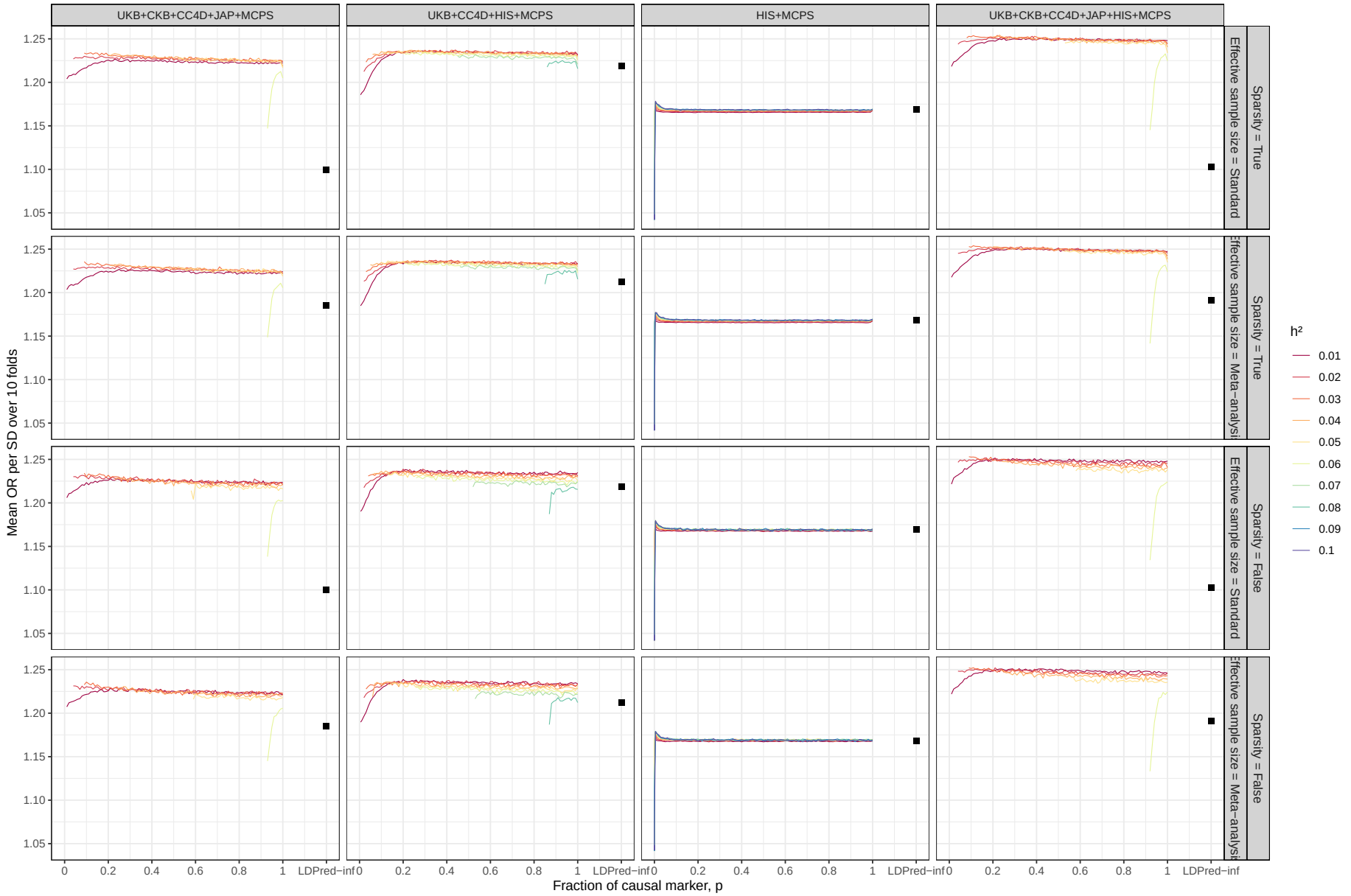


Figure 5.5: Mean OR per SD comparison of LDpred2 algorithms parameters

Analysis as for Figure 5.3.

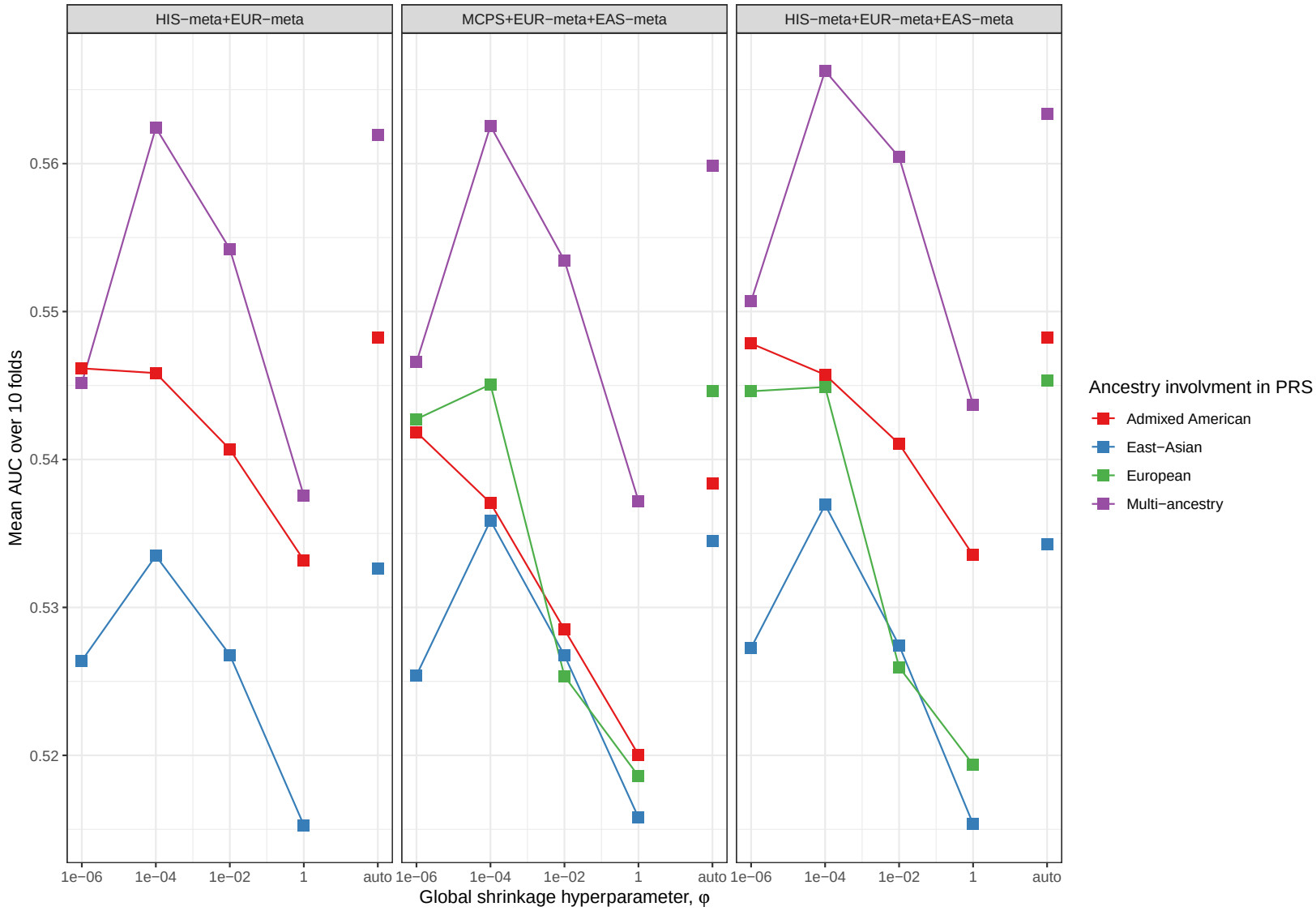


Figure 5.6: Mean AUC comparison of PRS-CSx algorithms parameters

Analysis as for Figure 5.2.

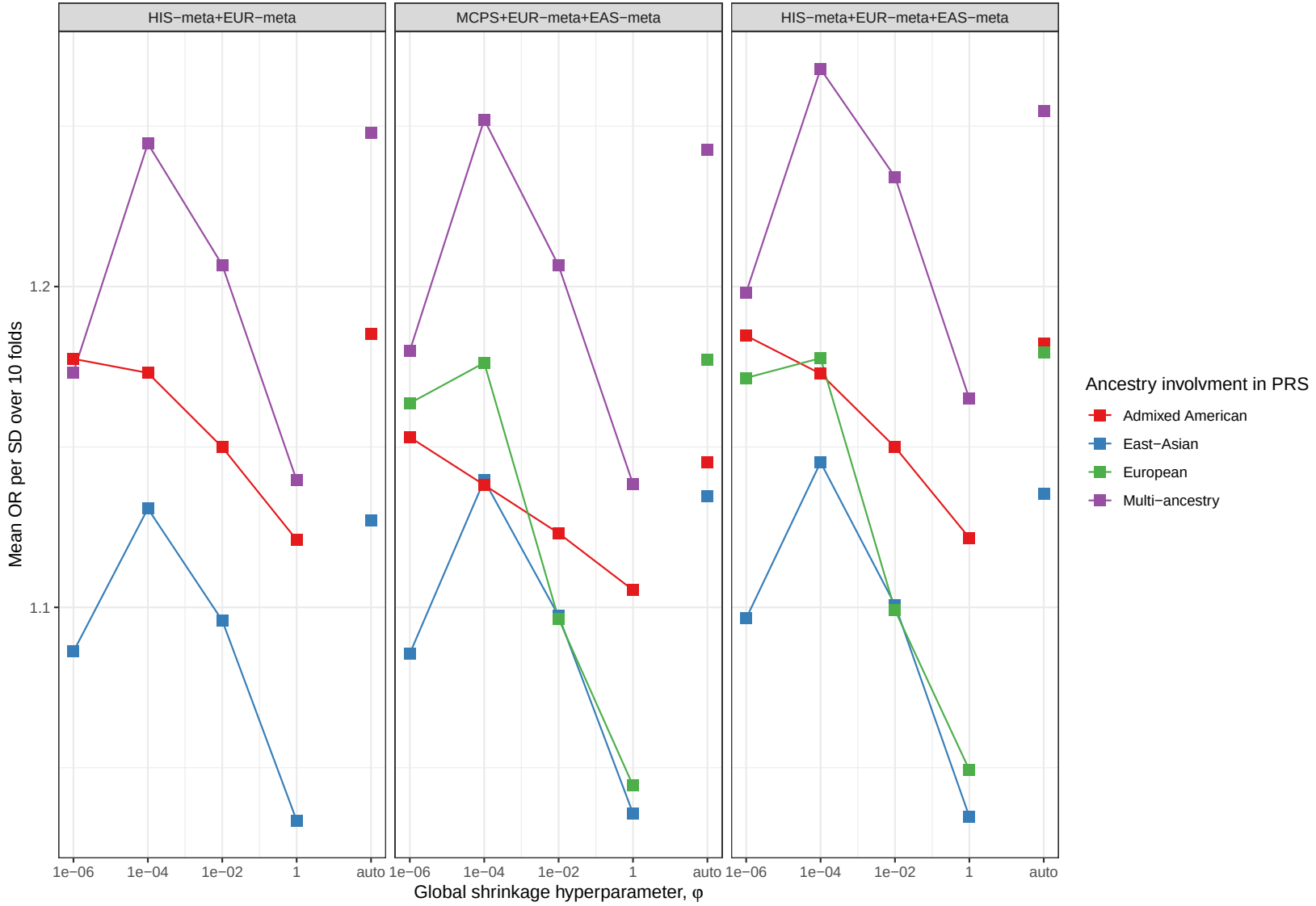


Figure 5.7: Mean OR per SD comparison of PRS-CSx algorithms parameters

Analysis as for Figure 5.3.

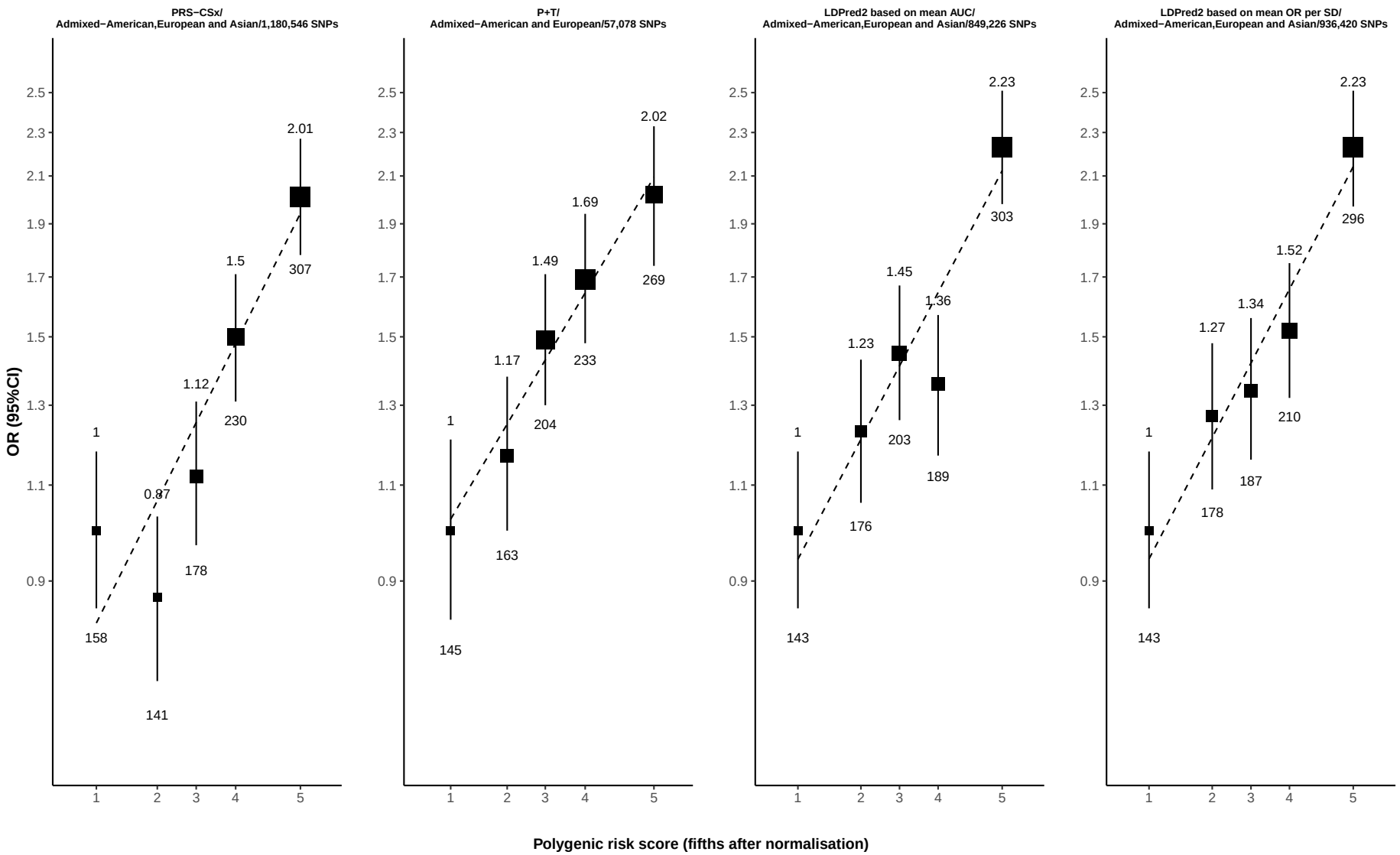


Figure 5.8: Odds of premature CHD by fifth of each MCPS-informed PRS, in the testing set

Analysis as for Figure 4.3 in Chapter 4,using the 20% data that withheld from PRS training.

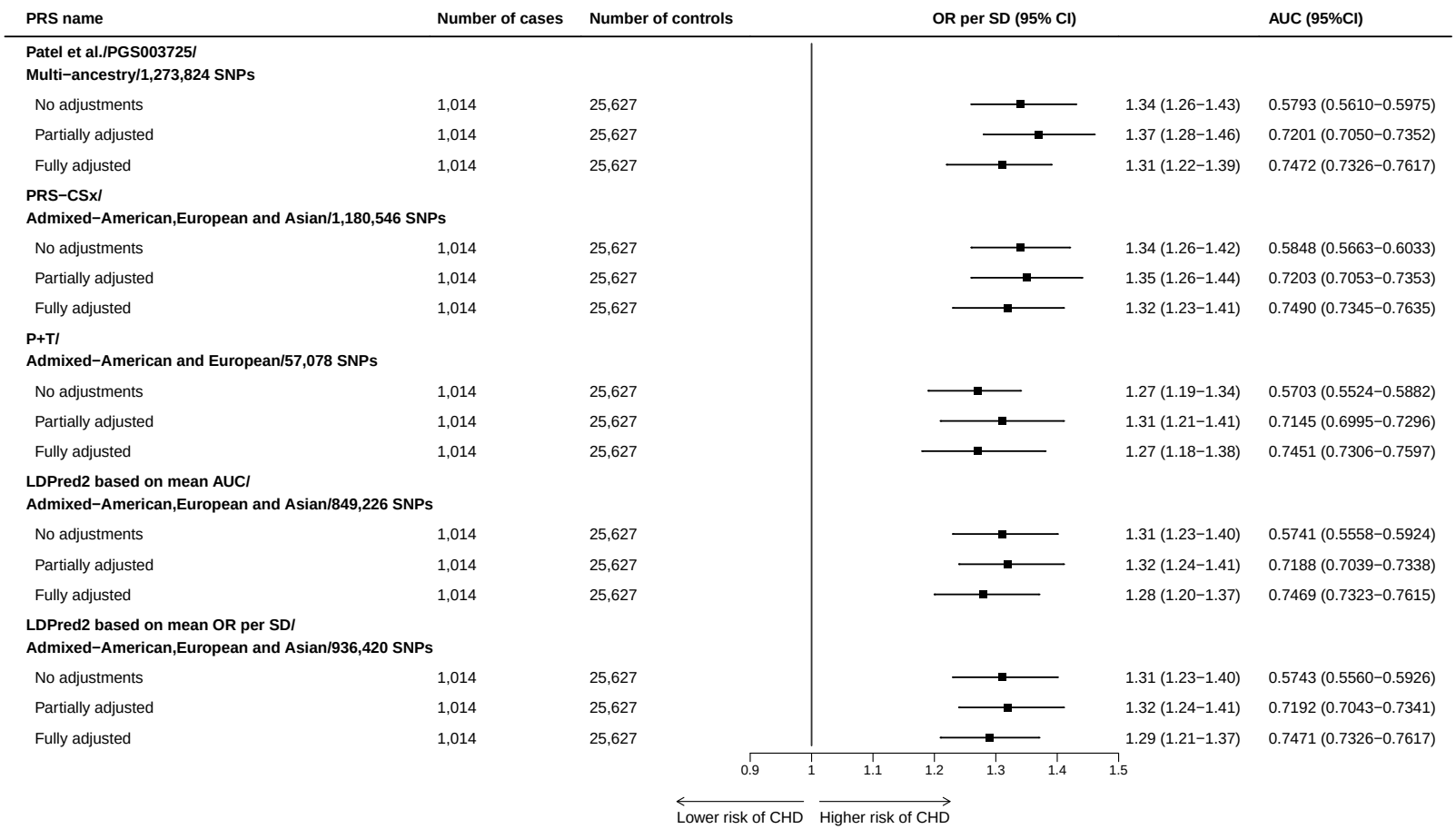


Figure 5.9: Odds of premature CHD per 1SD increase in each PRS, in the testing set

The analysis was conducted as described for 4.4 in Chapter 4, for partially and fully adjusted models. For models without adjustments, analyses were performed between each PRS and CHD without any additional covariates, using the 20% of data withheld from PRS training.

5.6 References

1. Privé F, Arbel J, and Vilhjálmsson BJ. “LDpred2: better, faster, stronger”. *Bioinformatics* 2021;36(22-23):pp. 5424–5431.
2. Ruan Y, Lin Y.-F, Feng Y.-CA, Chen C.-Y, Lam M, Guo Z, et al. “Improving polygenic prediction in ancestrally diverse populations”. *Nature Genetics* 2022;54(5):pp. 573–580.
3. Mills MC and Rahal C. “The GWAS Diversity Monitor tracks diversity by disease in real time”. *Nature Genetics* 2020;52(3):pp. 242–243.
4. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, et al. “Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups”. *American Journal of Human Genetics* 2020;106(5):pp. 707–716.
5. Zaccardi F, Timmins IR, Goldney J, Dudbridge F, Dempsey PC, Davies MJ, et al. “Self-reported walking pace, polygenic risk scores and risk of coronary artery disease in UK biobank”. *Nutrition, Metabolism and Cardiovascular Diseases* 2022;32(11):pp. 2630–2637.
6. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. *Nat Genet* 2018;50(9):pp. 1219–1224.
7. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. “Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 636–645.
8. Mak TSH, Porsch RM, Choi SW, Zhou X, and Sham PC. “Polygenic scores via penalized regression on summary statistics”. *Genetic Epidemiology* 2017;41(6):pp. 469–480.
9. Ge T, Chen CY, Ni Y, Feng YA, and Smoller JW. “Polygenic prediction via Bayesian regression and continuous shrinkage priors”. *Nature communications* 2019;10(1):p. 1776.
10. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. “Large-scale genome-wide association study of coronary artery disease in genetically diverse populations”. *Nature Medicine* 2022;28(8):pp. 1679–1692.
11. Oni-Orisan A, Haldar T, Cayabyab MAS, Ranatunga DK, Hoffmann TJ, Iribarren C, et al. “Polygenic Risk Score and Statin Relative Risk Reduction for Primary Prevention of Myocardial Infarction in a Real-World Population”. *Clinical Pharmacology & Therapeutics* 2022;112(5):pp. 1070–1078.
12. Saad M, El-Menyar A, Kunji K, Ullah E, Al Suwaidi J, and Kullo IJ. “Validation of Polygenic Risk Scores for Coronary Heart Disease in a Middle Eastern Cohort Using Whole Genome Sequencing”. *Circulation* 2022;Genomic and precision medicine.e003712.
13. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, et al. “Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease”. *Nature Genetics* 2020;52(11):pp. 1169–1177.
14. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. *Nature Medicine* 2023;29(7):pp. 1793–1803.
15. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. “Million Veteran Program: A mega-biobank to study genetic influences on health and disease”. *J Clin Epidemiol* 2016;70:pp. 214–23.
16. Ziyatdinov A, Torres J, Alegre-Díaz J, Backman J, Mbatchou J, Turner M, et al. “Genotyping, sequencing and analysis of 140,000 adults from Mexico City”. *Nature* 2023;622(7984):pp. 784–793.
17. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. “Next-generation genotype imputation service and methods”. *Nat Genet* 2016;48(10):pp. 1284–1287.
18. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. *Nature Genetics* 2021;53(7):pp. 1097–1103.

19. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. “Genetic analyses of diverse populations improves discovery for complex traits”. *Nature* 2019;570(7762):pp. 514–518.

20. Hoerl AE and Kennard RW. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. *Technometrics* 2000;42(1):pp. 80–86.

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

22. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. “A reference panel of 64,976 haplotypes for genotype imputation”. *Nature Genetics* 2016;48(10):pp. 1279–1283.

23. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. “Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel”. *Nature Communications* 2015;6(1):p. 8111.

24. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. “A map of human genome variation from population-scale sequencing”. *Nature* 2010;467(7319):pp. 1061–73.

25. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. “The UK Biobank resource with deep phenotyping and genomic data”. *Nature* 2018;562(7726):pp. 203–209.

26. van der Harst P and Verweij N. “Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease”. *Circ Res* 2018;122(3):pp. 433–443.

27. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. *Nat Genet* 2015;47(3):pp. 284–90.

28. Nikpay M, Goel A, Won H.-H, Hall LM, Willenborg C, Kanoni S, et al. “A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease”. *Nature genetics* 2015;47(10):pp. 1121–1130.

29. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. “Large-scale association analysis identifies new risk loci for coronary artery disease”. *Nature Genetics* 2013;45(1):pp. 25–33.

30. Mägi R and Morris AP. “GWAMA: software for genome-wide association meta-analysis”. *BMC Bioinformatics* 2010;11:p. 288.

31. Cochran WG. “The Combination of Estimates from Different Experiments”. *Biometrics* 1954;10(1):pp. 101–129.

32. Higgins JP and Thompson SG. “Quantifying heterogeneity in a meta-analysis”. *Stat Med* 2002;21(11):pp. 1539–58.

33. Kuriyama S, Yaegashi N, Nagami F, Arai T, Kawaguchi Y, Osumi N, et al. “The Tohoku Medical Megabank Project: Design and Mission”. *J Epidemiol* 2016;26(9):pp. 493–511.

34. Tsugane S and Sawada N. “The JPHC study: design and some findings on the typical Japanese diet”. *Japanese journal of clinical oncology* 2014;44(9):pp. 777–782.

35. Watanabe S, Tsugane S, Sobue T, Konishi M, and Baba S. “Study design and organization of the JPHC study. Japan Public Health Center-based Prospective Study on Cancer and Cardiovascular Diseases”. *J Epidemiol* 2001;11(6 Suppl):S3–7.

36. Hamajima N. “The Japan Multi-Institutional Collaborative Cohort Study (J-MICC Study) to detect gene-environment interactions for cancer”. *Asian Pac J Cancer Prev* 2007;8(2):pp. 317–23.

37. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. “Overview of the BioBank Japan Project: Study design and profile”. *J Epidemiol* 2017;27(3s):S2–s8.

38. Ishigaki K, Akiyama M, Kanai M, Takahashi A, Kawakami E, Sugishita H, et al. “Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases”. *Nature Genetics* 2020;52(7):pp. 669–679.

39. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. “A global reference for human genetic variation”. *Nature* 2015;526(7571):pp. 68–74.

40. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. “Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies”. *Nature Genetics* 2018;50(9):pp. 1335–1341.

41. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, et al. “China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up”. *Int J Epidemiol* 2011;40(6):pp. 1652–66.

42. Walters RG, Millwood IY, Lin K, Schmidt Valle D, McDonnell P, Hacker A, et al. “Genotyping and population characteristics of the China Kadoorie Biobank”. *Cell Genom* 2023;3(8):p. 100361.

43. Willer CJ, Li Y, and Abecasis GR. “METAL: fast and efficient meta-analysis of genomewide association scans”. *Bioinformatics* 2010;26(17):pp. 2190–1.

44. Tu Y., Chittoor G., Justice A.and Wang Z., Pereira A., Frankel E., Below J., et al. *Genetic Underpinnings of Coronary Heart Disease in Hispanic/Latino Populations*. Conference Paper. 2024.

45. Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, Kaplan RC, Daviglus ML, Giachello AL, et al. “Design and implementation of the Hispanic Community Health Study/Study of Latinos”. *Ann Epidemiol* 2010;20(8):pp. 629–41.

46. Lavange LM, Kalsbeek WD, Sorlie PD, Avilés-Santa LM, Kaplan RC, Barnhart J, et al. “Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos”. *Ann Epidemiol* 2010;20(8):pp. 642–9.

47. Lin D.-Y, Tao R, Kalsbeek WD, Zeng D, Gonzalez F, Fernández-Rhodes L, et al. “Genetic Association Analysis under Complex Survey Sampling: The Hispanic Community Health Study/Study of Latinos”. *The American Journal of Human Genetics* 2014;95(6):pp. 675–688.

48. Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, et al. “The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study”. *American Journal of Epidemiology* 2011;174(7):pp. 849–859.

49. The Women’s Health Initiative Study Group . “Design of the Women’s Health Initiative Clinical Trial and Observational Study”. *Controlled Clinical Trials* 1998;19(1):pp. 61–109.

50. Women’s Health Initiative . *Genetic and Omic Data in WHI*. Web Page. 2021. `https://www.whi.org/md/gwas`.

51. National Center for Biotechnology Information. *The BioMe Biobank at Mount Sinai (Study Accession: phs001644.v3.p2)*. Web Page.

52. Kolonel LN, Henderson BE, Hankin JH, Nomura AM, Wilkens LR, Pike MC, et al. “A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics”. *Am J Epidemiol* 2000;151(4):pp. 346–57.

53. Kolonel LN, Altshuler D, and Henderson BE. “The multiethnic cohort study: exploring genes, lifestyle and cancer risk”. *Nature Reviews Cancer* 2004;4(7):pp. 519–527.

54. The All of Us Research Program Investigators . “The “All of Us” Research Program”. *New England Journal of Medicine* 2019;381(7):pp. 668–676.

55. Ramirez AH, Gebo KA, and Harris PA. “Progress With the All of Us Research Program: Opening Access for Researchers”. *JAMA* 2021;325(24):pp. 2441–2442.

56. Bick AG, Metcalf GA, Mayo KR, Lichtenstein L, Rura S, Carroll RJ, et al. “Genomic data in the All of Us Research Program”. *Nature* 2024;627(8003):pp. 340–346.

57. Pulley J, Clayton E, Bernard GR, Roden DM, and Masys DR. “Principles of Human Subjects Protections Applied in an Opt-Out, De-identified Biobank”. *Clinical and Translational Science* 2010;3(1):pp. 42–48.

58. Snyder Bill. *Nashville Biosciences and Illumina announce agreement to establish preeminent clinico-genomic resource for life sciences research & development*. Press Release. 2022. `https://news.vumc.org/2022/01/10/nashville-biosciences-and-illumina-announce-agreement-to-establish-preeminent-clinico-genomic-resource-for-life-sciences-research-development/`.

59. Fisher-Hoch SP, Rentfro AR, Salinas JJ, Pérez A, Brown HS, Reininger BM, et al. “Socioeconomic status and prevalence of obesity and diabetes in a Mexican American community, Cameron County, Texas, 2004-2007”. *Prev Chronic Dis* 2010;7(3):A53.

60. Maldonado BL, Piqué DG, Kaplan RC, Claw KG, and Gignoux CR. “Genetic risk prediction in Hispanics/Latinos: milestones, challenges, and social-ethical considerations”. *J Community Genet* 2023;14(6):pp. 543–553.

61. Sabotta CM, Kwan S.-Y, Petty LE, Below JE, Joon A, Wei P, et al. “Genetic variants associated with circulating liver injury markers in Mexican Americans, a population at risk for non-alcoholic fatty liver disease”. *Frontiers in Genetics* 2022;Volume 13 - 2022.

62. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. “The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research”. *Genet Med* 2016;18(9):pp. 906–13.

63. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. “The UCSC Genome Browser Database: update 2006”. *Nucleic Acids Research* 2006;34(suppl_1):pp. D590–D598.

64. Choi SW and O’Reilly PF. “PRSice-2: Polygenic Risk Score software for biobank-scale data”. *GigaScience* 2019;8(7).

65. Privé F, Vilhjálmsson BJ, Aschard H, and Blum MGB. “Making the Most of Clumping and Thresholding for Polygenic Scores”. *Am J Hum Genet* 2019;105(6):pp. 1213–1221.

66. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores”. *The American Journal of Human Genetics* 2015;97(4):pp. 576–592.

67. Bulik-Sullivan BK, Loh P.-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. *Nature Genetics* 2015;47(3):pp. 291–295.

68. Korner-Nievergelt Fränzi, Roth Tobias, Felten Stefanie von, Guélat Jérôme, Almasi Bettina, and Korner-Nievergelt Pius. “Chapter 12 - Markov Chain Monte Carlo Simulation”. In: *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and STAN*. Ed. by Fränzi Korner-Nievergelt, Tobias Roth, Stefanie von Felten, Jérôme Guélat, Bettina Almasi, and Pius Korner-Nievergelt. Boston: Academic Press, 2015, pp. 197–212. doi: https://doi.org/10.1016/B978-0-12-801370-0.00012-5. https://www.sciencedirect.com/science/article/pii/B9780128013700000125.

69. Privé F, Aschard H, Ziyatdinov A, and Blum MGB. “Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr”. *Bioinformatics* 2018;34(16):pp. 2781–2787.

70. Privé F, Albiñana C, Arbel J, Pasaniuc B, and Vilhjálmsson BJ. “Inferring disease architecture and predictive ability with LDpred2-auto”. *The American Journal of Human Genetics* 2023;110(12):pp. 2042–2055.

71. Tamlander M, Mars N, Pirinen M, FinnGen, Widen E, and Ripatti S. “Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes”. *Communications Biology* 2022;5(1):p. 158.

72. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. “Software for Computing and Annotating Genomic Ranges”. *PLOS Computational Biology* 2013;9(8):e1003118.

73. Pagès Hervé. *SNPlocs.Hsapiens.dbSNP155.GRCh38: Human SNP locations and alleles extracted from dbSNP Build 155 and placed on the GRCh38/hg38 assembly*. Report. Bioconductor, 2023. https://bioconductor.org/packages/SNPlocs.Hsapiens.dbSNP155.GRCh38.

74. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. *GigaScience* 2015;4(1).

75. Lewis CM and Vassos E. “Polygenic risk scores: from research tools to clinical instruments”. *Genome Medicine* 2020;12(1):p. 44.

76. Wang Y, Kanai M, Tan T, Kamariza M, Tsuo K, Yuan K, et al. “Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology”. *Cell Genomics* 2023;3(10):p. 100408.

77. Gunn S, Wang X, Posner DC, Cho K, Huffman JE, Gaziano M, et al. “Comparison of methods for building polygenic scores for diverse populations”. *Human Genetics and Genomics Advances* 2025;6(1).

78. “Socioeconomic indices, demography and population structure”. In: *The Evolution and Genetics of Latin American Populations*. Ed. by Francisco M. Salzano and Maria C. Bortolini. Cambridge Studies in Biological and Evolutionary Anthropology. Cambridge: Cambridge University Press, 2001, pp. 55–102. doi: DOI:10.1017/CBO9780511666100.004. https://www.cambridge.org/core/product/99516DDC896C6D7E0A0C12D333A19858.

79. Ge T, Irvin MR, Patki A, Srinivasasainagendra V, Lin Y.-F, Tiwari HK, et al. “Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations”. *Genome Medicine* 2022;14(1):p. 70.

80. Tran KB, Lang JJ, Compton K, Xu R, Acheson AR, Henrikson HJ, et al. “The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019”. *The Lancet* 2022;400(10352):pp. 563–591.

81. Safiri S, Nejadghaderi SA, Karamzad N, Kaufman JS, Carson-Chahhoud K, Bragazzi NL, et al. “Global, Regional and National Burden of Cancers Attributable to High Fasting Plasma Glucose in 204 Countries and Territories, 1990-2019”. *Front Endocrinol (Lausanne)* 2022;13:p. 879890.

82. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. “The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource”. *Nucleic Acids Research* 2022;51(D1):pp. D977–D985.

Chapter 6: Integration of polygenic and clinical risk scores to improve overall prediction of coronary heart disease risk

Summary

Although polygenic risk scores (PRSs) have been shown to be effective for coronary heart disease (CHD) risk stratification, they explain only a small fraction of the overall variability in CHD risk in adults. Several studies have demonstrated the potential improvement in disease risk stratification beyond a PRS or clinical risk score alone when integrating them together into one integrated risk-prediction model[1–5]. However, previous studies on integrated risk scores have been conducted mainly in populations of European ancestry. In this chapter, the best-performing PRS for predicting CHD risk that was previously evaluated in **Chapter 4** was integrated with three guideline-recommended clinical risk scores, SCORE2[6] with two variations (diabetes-recalibrated and standard) and the Pooled Cohort Equation (PCE)[7], and were assessed for their joint performance in CHD risk prediction and stratification among 133,207 participants from the Mexico City Prospective Study (MCPS) aged 35-79 at recruitment. All three scores (standardised) associated positively with CHD risk in the MCPS study population, with the average strength of association, odds ratio (OR) for CHD per standard deviation (SD) score, being 1.62 (95% confidence interval [CI], 1.59-1.65), 1.57 (1.54-1.60) and 1.59 (1.57-1.62), respectively. When considered in combination, the integrated (genetic x clinical) risk score resulted in slightly greater CHD risk separation between participants in the top versus bottom 20% of the predicted risk distribution than was observed when using the clinical risk score alone. For example, the observed OR for CHD among individuals in the top 20% of the SCORE2-diabetes-recalibrated risk strata increased from 12.92 to 13.66 after integration of the Patel *et al.*[8] PRS. Moreover, based on a 7.5% clinical-risk threshold for identifying participants at high risk of CHD, this integrated score significantly improved CHD case reclassification compared to the SCORE2-diabetes-recalibrated

alone. The categorical net reclassification index (NRI), with a 7.5% risk threshold, was 1.50% (95% CI, 0.73%-2.27%), while the continuous NRI was 17.50% (14.75%-20.24%). Sensitivity analysis using the novel MCPS-informed PRS generated in **Chapter 5** confirmed the robustness of the main findings.

6.1 Background and aims

Despite being a highly heritable disease, only a small fraction of the total variability in CHD risk can be explained by aggregating currently identified genetic variants into CHD PRSs. This may be in part due to limitations of the current methods for the construction of the scores, which typically combine single nucleotide polymorphism (SNP) effects linearly, without accounting for their joint effects during GWAS or the PRS construction. As shown in **Chapter 5**, PRSs alone could only achieve a model discrimination (area under the receiver operator curve, AUC) of 0.58, which is consistent with findings from several previous studies[9–12]. However, CHD risk is not driven solely by genetics. Environmental, biological, and lifestyle factors also contribute significantly to CHD susceptibility, as detailed in **Chapter 1**. Clinical risk scores such as the Framingham Risk Score (FRS)[13], SCORE2[6], PCE[7], QRISK[14,15] and ASSIGN[16] are designed to estimate the 10-year individual cardiovascular (CVD) risk, by combining these non-genetic risk factors together. They can also be applied for CHD risk prediction, as common CHD observational risk factors are shared with other CVDs. On their own, these clinical risk scores achieved discrimination rates (Harrell's C-index) of over 0.70 for CHD risk prediction[2,17,18]. One potential limitation of these scores and their utility in other populations may arise from the fact that they were all originally derived and calibrated using populations of European ancestry. Although recently recalibrated versions, such as SCORE2-ASIA, have been shown to perform well among Asians[19], external validation studies in other non-European, non-Asian populations are largely missing.

Chapter 4 demonstrated that including genetic information with conventional CHD clinical risk factors, such as sex, age, smoking status, and blood pressure, slightly improved model predic-

tive performance. Moreover, PRSs were shown to independently predict CHD risk beyond these clinical risk factors. Hence, combining CHD PRS with clinical risk factors into an integrated score may further enhance overall CHD risk prediction. Several studies have investigated the potential benefits of such integrated scores (either for CVD or CHD prediction), each using different methods of integration, and have consistently reported improvements in disease risk stratification and reclassification[1–5]. However, similar to genetic studies on CHD, previous research on integrated risk scores has focused primarily on populations of European ancestry, with no studies to date assessing their performance in populations of admixed-American ancestry from Latin America. Using data from MCPS, the aim of this chapter is to first assess the performance of three selected clinical risk scores, two variations of SCORE2[6] and PCE[7], for predicting CHD events among Mexican adults, and second, to integrate these clinical risk scores with the best-performing CHD PRS evaluated, the Patel *et al.*[8] PRS, in **Chapter 4**. The integrated scores are then evaluated for their performance in CHD risk prediction and reclassification, compared to the original clinical scores. The main analysis is subsequently repeated in the 20% subset of MCPS data used for PRS testing in **Chapter 5**, by alternatively incorporating the novel MCPS-informed PRS (MCPS-PRS) instead of the Patel *et al.* PRS[8] into the integrated score to further assess its utility.

6.2 Methods

The analyses conducted in this chapter used the dataset from the MCPS. **Chapter 3** described the MCPS study design, overall methods, analysis population, and CHD outcome definitions in detail. Briefly, all MCPS participants aged between 35-79 at recruitment and with genotype data available were included in the main analysis. The outcome assessed in this chapter is CHD events, defined as self-reported CHD at baseline or death before age 80, with CHD mentioned anywhere on the death certificate.

6.2.1 Clinical risk score

Based on the availability and practicality of the specific risk variable requirements, three clinical risk scores, two variations of SCORE2[6] and PCE[7], were selected to estimate CHD risk among participants in MCPS, as all the required variables were available in the dataset. These scores were computed for all participants in MCPS, using the R package RiskScorescvd-0.2.0[20]. Although other CVD risk scores, such as the WHO CVD risk chart[21], are available, they were not included in this analysis because SCORE2 and PCE are more widely adopted in clinical practice and therefore more suitable for comparative evaluation.

If an input variable was missing during the computation of clinical risk scores, it was imputed following the method described in **Chapter 3**.

6.2.1.1 SCORE2 for CHD risk prediction

SCORE2[6] is a clinical risk score developed in 2021 to predict the 10-year risk of first-onset fatal and non-fatal CVD among diabetes-free individuals aged 40 to 69 years. The score was developed using data from 677,684 participants across 45 cohorts in 13 countries. The variables included in the SCORE2[6] risk-prediction model include well-established CHD risk factors, specifically, age, sex, smoking status, systolic blood pressure (SBP), total cholesterol (total-C), and high-density lipoprotein cholesterol (HDL-C)[22–24]. In addition, based on the country-specific data included in the development of SCORE2, four risk regions (low to high) were defined according to the WHO standardised CVD mortality rates per 100,000 people. These risk regions were used to further calibrate SCORE2 according to the regional CVD risk burden. The score was subsequently externally validated in 1,133,181 individuals from 15 countries across Europe, using survival analysis, and demonstrated good model discrimination (Harrell's C-index), ranging from 0.67 to 0.81[6].

Follow-up efforts aimed to model SCORE2[6] for older adults, leading to the development of SCORE2-OP[25], targeting individuals aged 70 years and older. Using a similar methodology

to that of SCORE2 above, SCORE2-OP[25] involved 28,503 individuals without prior CVD at recruitment included during the score derivation. External validation involved 338,615 individuals from six different cohorts, and SCORE2-OP[25] demonstrated a C-index range from 0.63 to 0.67. For the analysis presented in this chapter, SCORE2 and SCORE2-OP were computed for participants aged below and above 70 years, respectively (see *Table S4 to Table S7* for coefficient values and computation method). For simplicity, both SCORE2 and SCORE2-OP are referred to collectively as SCORE2 throughout the analysis and in subsequent sections.

SCORE2 ranges from 0 to 100, with higher scores indicating a higher risk of CVD. For instance, a SCORE2 value of 10 corresponds to a 10% estimated absolute risk of developing CVD within the next 10 years.

Diabetes mellitus status information was initially included in the modelling during the development of SCORE2 for calibration purposes, but it was excluded as a risk predictor, as SCORE2 was intended for individuals free from diabetes[6]. SCORE2-Diabetes[26] was subsequently developed to specifically estimate 10-year CVD risk among individuals with diabetes. However, the age of diagnosis for diabetes is required for SCORE2-Diabetes[26] computation, and this information was not available reliably in the MCPS dataset. Therefore, SCORE2-Diabetes[26] could not be calculated for participants with diabetes in MCPS. Although diabetes status was not included as a risk predictor for SCORE2 computation, its effect was still accounted for in our analysis by incorporating the excluded diabetes coefficient during the SCORE2 calculation for MCPS participants. This modified version was referred to as SCORE2-diabetes-recalibrated in this chapter. For comparison, the standard SCORE2, calculated without adjustments for diabetes, was also computed and referred to as SCORE2-standard. Therefore, two variations of SCORE2 were evaluated in this chapter, SCORE2-diabetes-recalibrated and SCORE2-standard.

Based on the standardised CVD mortality rate of 150 per 100,000 population in Mexico in 2021[27], the country falls within the ”moderate risk” category as defined by the SCORE2 risk region classifications[6].

6.2.1.2 Pooled cohort equation for CHD risk prediction

The Pooled Cohort Equation (PCE)[7] was developed by the American College of Cardiology and the American Heart Association (ACC/AHA) in 2013 to estimate the 10-year risk of atherosclerotic cardiovascular disease (ASCVD), an outcome that includes myocardial infarction (MI), CHD death, and stroke. The equation was derived from five cohorts of American individuals of African and European ancestry aged 40 to 79 years. The score incorporates sex, age, ancestry (European or African), total-C, HDL-C, diabetes status, smoking status, SBP, and whether an individual with hypertension takes treatment. The PCE has been widely used in previous studies among populations of European ancestry and has demonstrated high model discrimination (Harrell's C-index), ranging from 0.70 to 0.80 for CVD risk prediction[17,28–30]. PCE ranges from 0 to 1, with higher values corresponding to a greater risk of ASCVD. For example, a PCE score of 0.1 indicates a 10% estimated 10-year risk of developing ASCVD.

The PCE algorithm employs valid input ranges for total-C (130 to 320 mg/dL), HDL-C (20 to 120 mg/dL), and SBP (90 to 200 mmHg) and is particularly sensitive to out-of-range values. Values outside these ranges may result in extrapolation and potentially inaccurate risk estimates. Therefore, after missing value imputation was performed (following the method outlined in **Chapter 3**), a second imputation step was applied for total-C, HDL-C, and SBP values falling outside the valid PCE input ranges. Values below the lower bound were imputed to the lower bound, and values above the upper bound were imputed to the upper bound. Approximately 17% of the data underwent the second imputation step, with the majority of cases resulting from total-C and HDL-C values falling below the lower bound. Without this additional imputation, the PCE algorithm would have estimated a risk score of 1 (i.e., 100% CHD risk) for a few MCPS participants due to extremely low HDL-C values. For the ancestry input, since all participants in the MCPS are of admixed-American ancestry, which is not directly represented among the ancestry categories accepted by the original PCE model, we applied the coefficients designated for European ancestry during the PCE score computation, as recommended by the current PCE guidelines (see

Table S8 for coefficient values)[7,31]. To enable comparison between PCE and SCORE2, the PCE estimated for each participant was multiplied by 100 to align it with the scale range of SCORE2.

6.2.2 Polygenic risk score for CHD

Details of the MCPS genetic data are provided in **Chapter 3**. The CHD PRS selected to be integrated with the clinical risk scores (described above) was the Patel *et al.* PRS[8]. Details of this score and its recreation for MCPS participants are described in **Chapter 4**. Briefly, the Patel *et al.* PRS[8] is a multi-ancestry, multi-trait score that leveraged genetic information across five ancestries (South-Asian, East-Asian, African, admixed-American and European) for CHD and its related traits and risk factors during its construction. The score was derived using LDpred2[32] and included 1,296,172 SNPs. The Patel *et al.* PRS[8] demonstrated the strongest association and predictability with CHD risk in **Chapter 4**.

To avoid overfitting, the novel MCPS-PRS developed in **Chapter 5** was generated using the 20% subset of MCPS data that was withheld for PRS testing, and was used in this chapter for sensitivity analysis. Details of the score generation are described in **Chapter 5**. Briefly, this PRS was constructed using PRS-CSx[33], incorporating genetic information from European, East Asian, and admixed-American ancestries for CHD, and included 1,180,546 SNPs. The PRS demonstrated performance comparable to that of the Patel *et al.* PRS[8].

6.2.3 Integration of PRS with clinical risk scores

Both SCORE2[6] and PCE[7] estimate absolute CHD risk, meaning they directly quantify the disease risk of an individual. In contrast, a PRS provides a relative risk estimate, which is only meaningful when interpreted in the context of a reference population. Consequently, integrated risk scores were generated for each MCPS participant *i* using the following equation:

$$\text{Integrated score}_i = \text{Clinical risk score}_i \times (\text{RR}_{\text{PRS}})^{\text{PRS}_i} \quad (6.1)$$

$$RR_{PRS} \approx \frac{OR_{PRS}}{1 - p + (p \times OR_{PRS})} \quad (6.2)$$

Where we define p as the risk of CHD in the study population. Based on the prevalence of fatal or non-fatal CHD in MCPS, 5,163 cases out of 133,207 individuals, we set p=0.04. OR_{PRS} represents the average association strength between the included PRS and CHD, estimated from a logistic regression model where the PRS was the sole predictor and CHD was the outcome (OR per SD=1.27 for the Patel *et al.* PRS[8]). RR_{PRS} adjusted the OR based on CHD prevalence of 0.04 in MCPS and was 1.26 per SD higher PRS.

In this thesis, every PRS was standardised (see **Chapter 3**) to mean of 0 and SD of 1. A PRS_i of value 0 therefore means that this individual is not at higher or lower genetic risk relative to others in the cohort. Based on our formula, $(RR_{PRS})^{PRS_i}$ in this case will be 1, because PRS_i will equal 0, meaning the PRS does not alter the risk estimated by the clinical risk score. If PRS_i is greater than 0, $(RR_{PRS})^{PRS_i}$ will be greater than 1 and if PRS_i is less than 0, $(RR_{PRS})^{PRS_i}$ will be between 0 and 1. Since PRS follows a standard normal distribution, only 2% of individuals will have scores over 2 and resulting in an $(RR_{PRS})^{PRS_i}$ of 1.58. Therefore, only in the case of SCORE2 or PCE exceeding 63 could the integrated score potentially exceed 100 (i.e., over 100% estimated CHD risk, and in such an extreme case a value of 100 was assumed). One participant with a SCORE2-diabetes-recalibrated × PRS integrated score and 20 participants with a PCE × PRS integrated score exceeded 100. All of them had hypertension, were aged over 60, and the majority of them were men.

Using the above formulae, each constructed integrated score enhanced the clinical risk score by incorporating genetic information and translating the relative risk estimated by the PRS into the framework of a clinical risk score, while largely preserving the original absolute risk interpretation. Overall, the integrated scores ranged from 0 to 100.

6.2.4 Statistical analysis

6.2.4.1 Primary analysis

Overall, three clinical risk scores were evaluated: two variations of SCORE2 (SCORE2-diabetes-recalibrated and SCORE2-standard) and PCE. Their performance in CHD risk prediction was assessed using logistic regression models in which each clinical risk score served as the sole predictor, and CHD as the outcome. No additional covariates were included, as clinical risk scores already include the main CHD risk factors, so adjusting for them would artificially reduce the apparent relevance of the risk score to CHD risk. Each clinical risk score was categorised into five equally sized groups and treated as a categorical variable in analyses against CHD, with the lowest group used as reference group. Subsequently, each score was treated as a continuous variable to assess its average strength of association with CHD risk. To allow meaningful comparison later in the analysis, each clinical score was standardised, to mean of 0 and SD of 1, when treated as a continuous variable. The AUC was used to estimate model discrimination.

The integrated scores (clinical risk score x PRS) initially followed the same analytical approach as the clinical risk score. To evaluate the risk reclassification performance of each integrated score, the categorical Net Reclassification Index (NRI) was computed by comparing its risk-group assignments to those based on the clinical risk score alone, using a 7.5% 'high-risk' cut-off[34,35]. The 7.5% threshold represents the age-specific very-high-risk boundary recommended by the European Society of Cardiology (ESC) guidelines for initiation of preventive treatment, based on the SCORE2 estimates for individuals younger than 50 years (a group that comprised over 50% of the MCPS participants at the time of recruitment)[36]. For consistency, a 7.5% CHD risk threshold was also adopted for the PCE, as this has been widely used in the previous literature[17,37,38]. Continuous NRI was also calculated to measure the overall upward and downward shifts in predicted CHD risk for all cases and controls, independent of risk thresholds. In addition, the integrated discrimination improvement (IDI) was calculated to quantify improvement in

the discriminative ability of the integrated score compared to the clinical score, by measuring the difference in mean predicted risk improvement between cases and controls[34,35].

6.2.4.2 Subgroup-specific analysis

To assess difference in clinical risk by sex and diabetes status, the main analyses were repeated in their respective subgroups. For the sex-specific analysis, each clinical score was treated as a categorical variable with five equally sized groups, defined separately for women and men. Women were assigned labels from 1-5, and men were assigned labels from 6-10. The lowest group for women (group 1) was used as reference group. A similar approach was used for the baseline diabetes status-specific analysis. However, only 24,796 MCPS participants self-reported being diagnosed with diabetes at baseline. Categorising them into five groups would have resulted in the lowest group having no CHD cases. To address this, each clinical score was categorised into four equally sized groups, defined separately for participants with and without diabetes at baseline. Participants without diabetes were assigned labels from 1-4, and participants with diabetes were assigned labels from 5-8.

To investigate the potential effect modification of the clinical risk score (considered as five equally sized groups) in genetic predisposition to CHD (i.e., to see if there was an interaction between the clinical and polygenic risk scores), a logistic regression was performed with the Patel *et al.* PRS[8] as the sole predictor and CHD as outcome, stratified by clinical risk score levels, with a test for trend across the resulting odds ratios performed.

6.2.4.3 Sensitivity analyses

To evaluate the performance of the novel MCPS-PRS within an integrated score, the main analyses described above were repeated, integrating the MCPS-PRS with clinical risk scores among the 20% subset of MCPS data that had been retained for PRS testing (see **Chapter 5**), thereby avoiding overfitting. Due to the much smaller sample size for this analysis, clinical risk scores were treated only as continuous variables and were not categorised.

Since both SCORE2[6,25] and PCE[7] were developed to estimate the 10-year risk of CVD, the analysis of the clinical risk scores was repeated using Cox regression with CHD death within the first 10 years of follow-up as the outcome, restricted to participants in the MCPS cohort without a history of CHD at baseline. To assess the risk reclassification performance of the integrated risk scores against the clinical score over the first 10 years of follow-up, we computed the prospective (time-dependent) NRI[39] alongside the standard (binary) NRI.

Additional analyses were conducted to compare the categorical NRI, sensitivity, and specificity of the integrated risk score relative to the clinical score at various risk thresholds.

6.3 Results

Baseline characteristics of the MCPS participants included in the main analysis, as well as the breakdown of fatal and non-fatal CHD cases, are described in **Chapter 4**. Briefly, after the exclusion of participants without genetic information, with unreliable mortality linkage or aged above 80 years at recruitment, 133,207 eligible participants aged 35-79 years at baseline remained for inclusion in the main analysis. Of these, 1,901 (1.4%) participants self-reported a history of medically diagnosed CHD at baseline, which included myocardial infarction (MI) and angina. As of September 30, 2022, a total of 3,479 CHD deaths occurring before age 80 years had accrued, and were listed as the primary or contributing cause of mortality on the death certificate. Among those who died from CHD, 217 individuals had also reported CHD at baseline, resulting in a total of 5,163 participants meeting the criteria for CHD.

Figure 6.1 presents the distributions of the three clinical risk scores evaluated, along with their input components, in MCPS. All three clinical risk scores were positively skewed, and the valid input ranges set by the PCE suggested that total cholesterol levels were generally lower in MCPS than in other reference populations. Overall, the addition of PRS slightly broadened the distribution of predicted risks (see left panels of **Figure 6.2 to Figure 6.4**), suggesting that including genetic information into clinical risk scores may enhance their ability to distinguish between individuals

with similar clinical profiles but different genetic predisposition to CHD. The Bland-Altman plots (right panels of **Figure 6.2 to Figure 6.4**) indicate that inclusion of PRS does not systematically inflate or deflate the predicted clinical risk.

6.3.1 Association between clinical risk score and CHD risk

For the three CHD clinical risk scores selected for analysis, a higher estimated level of clinical risk was associated positively with higher odds of having CHD in the MCPS population (**Figure 6.5** and **Figure 6.6**). When each clinical risk score was treated as a categorical variable, the top 20% of clinical risk had 10.29 to 12.92 higher odds of CHD compared to the lowest (reference) 20% group. The standard SCORE2, which was not calibrated for diabetes, conferred lower odds of CHD in the top 20% of clinical risk compared to the reference group (OR=10.29), whereas SCORE2-diabetes-calibrated yielded higher odds for subsequent CHD (OR=12.92). When each clinical risk score was treated as a continuous variable in an analysis without additional covariate adjustments, all three scores showed similar strengths of association with CHD. SCORE2-diabetes-recalibrated demonstrated the strongest strength of association with CHD in MCPS, with an OR per SD of 1.62 (95%CI, 1.59-1.65), followed by PCE (OR per SD=1.59,1.57-1.62) and SCORE2-standard (OR per SD=1.57, 1.54-1.60). In terms of discrimination, both SCORE2-diabetes-recalibrated and PCE achieved a model AUC of 0.73, while SCORE2-standard showed a slightly lower AUC of 0.71.

When restricting the analysis to CHD mortality in the first 10 years of follow-up, the overall shape of the association between clinical risk scores and CHD death risk remained largely consistent. However, the risk conferred at the top 20% of the risk strata, relative to the reference group, was notably higher compared to the main analysis (**Figure 6.7**). The average strength of the association between clinical risk scores and subsequent CHD death remained comparable across the three scores, but all of them were higher than in the main analysis. The HRs per SD were 1.91 (95%CI, 1.88-1.95) for SCORE2-diabetes-recalibrated, 1.90 (1.86-1.94) for SCORE2-standard

and 1.97 (1.93-2.02) for PCE (**Figure 6.8**). Model discrimination measured by Harrell's C-index was 0.84 for SCORE2-diabetes-recalibrated, 0.82 for SCORE2-standard and 0.83 for PCE.

6.3.2 Association between integrated score and CHD risk

Each clinical risk score was then combined with the Patel *et al.* PRS[8] into integrated risk score. Integrating genetic information related to CHD predisposition with traditional clinical risk scores resulted in higher estimated risks for individuals in the top 20% of the combined score distribution, compared to those identified as high-risk using conventional CHD factors alone (**Figure 6.9 to Figure 6.11**). For SCORE2-diabetes-recalibrated, the OR for individuals in the top 20% risk stratum was 12.92 when compared to the lowest 20% reference group, which increased to 13.66 when it was integrated with the Patel *et al.* PRS[8]. A similar level of improvement was observed for SCORE2-standard and PCE, where the OR increased from 10.29 to 10.96 and from 10.77 to 11.98, respectively.

When each integrated score was treated as a continuous variable, they showed largely no improvement in the strength of association with CHD risk (**Figure 6.12**). The OR per SD of each integrated score remained largely similar compared to the clinical risk scores they were designed to enhance. Model AUC was also largely unchanged, except for a slight improvement observed for both SCORE2 integrated scores, irrespective of diabetes calibration. Compared to their respective original SCORE2 versions, the integrated scores increased model AUC from 0.73 to 0.74 (diabetes-recalibrated) and from 0.71 to 0.72 (standard) (**Figure 6.12**).

In terms of improvement in risk reclassification, both SCORE2 integrated scores showed significant categorical NRIs when compared to their corresponding SCORE2 alone, using a CHD risk threshold of 7.5%. When comparing the integrated score, SCORE2-diabetes-recalibrated x Patel *et al.* PRS[8], to SCORE2-diabetes-recalibrated alone, 245 of 2,164 MCPS participants (11%) originally incorrectly classified as CHD-free were correctly reclassified to case and 159 of 2,999 (5.3%) MCPS participants originally correctly classified as CHD cases were incorrectly

reclassified to controls (**Table 6.1**). Therefore, with the integrated score, an additional 86 CHD cases were reclassified correctly (of all 5,163 cases), leading to a categorical case NRI of 1.67% (95% CI, 0.90%-2.43%) (**Table 6.2**). For controls, 3,564 of 95,689 (3.7%) that were originally correctly classified controls were incorrectly reclassified to cases, and 3,349 of 32,355 (10%) originally incorrectly classified cases were correctly reclassified as CHD-free by the combined scores (**Table 6.1**). Consequently, 215 more controls were incorrectly classified, resulting in a slight negative control NRI of -0.17% (-0.30%- -0.04%) (**Table 6.2**). Together, the overall categorical NRI combining CHD cases and controls was therefore 1.50 (0.73-2.27).

In terms of overall improvement in the scoring for CHD cases versus controls, the continuous NRI, which looks at whether the integrated score predicted cases with higher scores and controls with lower scores compared to SCORE2-diabetes-recalibrated, was positive for both groups (**Table 6.2**). The continuous NRI for cases was 16.52% (95%CI, 13.83%-19.21%) and for controls was 0.97% (0.43%-1.52%), which led to an overall continuous NRI of 17.50 (14.75-20.24). Inclusion of the Patel *et al.* PRS^{8} to the clinical scores resulted in a 0.78% improvement in discriminative ability (i.e., better separation between cases and controls) compared to SCORE2-diabetes-recalibrated alone, as indicated by the IDI (95%CI, 0.69%-0.87%).

The results were largely similar when comparing the integrated score SCORE2-standard x Patel *et al.* PRS^{8} to SCORE2-standard. In this comparison, 307 cases were reclassified correctly and 174 cases were reclassified incorrectly to controls. For controls, 3,138 of them were reclassified correctly and 3,532 incorrectly to cases (**Table 6.3**). Therefore, for categorical NRI, reclassification was improved by 2.58% for cases (95%CI, 1.75%-3.41%) and undermined by 0.31% for controls (-0.43%- -0.18%), resulting in an overall categorical NRI of 2.27 (1.43-3.11) (**Table 6.4**). Due to the same computation procedure as with SCORE2-diabetes-recalibrated, continuous NRI was the same for the SCORE2-standard integrated score. IDI was estimated to be 0.57% (95%CI, 0.50%-0.64%).

Comparing the integrated score PCE x Patel *et al.* PRS^{8} to PCE alone, resulted in slightly differ-

ent reclassification results using the recommended 7.5% risk threshold. Among CHD cases, 155 were correctly reclassified to cases and 124 incorrectly reclassified to controls. Among controls, 2,459 were correctly reclassified and 2,979 were incorrectly reclassified to CHD cases (**Table 6.5**). The large degree of incorrect reclassification of controls to cases and the limited improvement in correct reclassification of CHD cases, resulted in an insignificant improvement in the overall reclassification. The categorical NRI was 0.19 (95%CI, -0.45-0.84), resulting from a case NRI of 0.60% (95%CI, -0.03%-1.23%) and control NRI of -0.41% (-0.52%- -0.29%) (**Table 6.6**). However, when the clinically-informative CHD risk threshold of 7.5% was not conditioned on, the integrated score resulted in a significant continuous NRI of 16.65 (95%CI, 13.92-19.38). This improvement was primarily driven by a 16.56% increase in case NRI (95% CI, 13.88%-19.24%), despite an insignificant improvement in control NRI of 0.09% (95% CI, -0.44% to 0.62%). The integrated score also showed an improvement in discriminative ability compared to PCE alone, as indicated by an IDI of 1.18% (95% CI, 1.04%-1.33%).

6.3.3 Difference in clinical risk by sex and baseline diabetes status

The impact of male sex on CHD risk in the MCPS population was reasonably well-captured by the clinical risk scores. For, at each *predicted* level of risk based on sex and the other clinical risk factors, the *observed* odds ratios were broadly similar for men and women, although there was some variation, particularly for those in the top 20% of the risk distribution (**Figure 6.13**). By contrast, the impact of diabetes on CHD risk in the MCPS population was less well captured by the clinical risk scores. For, at each *predicted* level of risk based on diabetes and other clinical risk factors (i.e., for SCORE2-diabetes-recalibrated and PCE), the *observed* odds ratios remained notably higher for those with than without diabetes (**Figure 6.14**). For the SCORE2-standard risk score, which does not take diabetes into account in its prediction of risk, the greatly increased CHD risks for those with diabetes were clearly demonstrated.

6.3.4 Genetic predisposition to CHD risk across clinical risk score strata

For the Patel *et al.* PRS[8], a slight U-shaped pattern was observed in the association between PRS and CHD risk at each level of clinical risk. Compared to the first stratum, the association was slightly weaker in the second risk stratum but became progressively stronger across the higher risk strata (three to five) for all three clinical risk scores evaluated (**Figure 6.15 to Figure 6.17**). A test for trend (i.e., interaction) was significant for all three scores, with p-values less than 0.001.

6.3.5 Sensitivity analysis

The MCPS participants used in the sensitivity analysis presented in this chapter comprised only 20% of the total sample as the remaining subset was used during the derivation and cross validation of the MCPS-PRS within the cohort. The characteristics of the participants included in the sensitivity analysis are shown in **Table 6.7** and indicate largely similar baseline traits to those in the full analysis cohort (**Table 4.2** in **Chapter 4**). Similar to the main analysis results, integrating the MCPS-PRS with the clinical risk scores showed largely no improvement in the average strength of association with subsequent CHD risk compared to models with the clinical risk score alone. Model discrimination improved slightly, from 0.73 to 0.74 for SCORE2-diabetes-recalibrated, from 0.70 to 0.72 for SCORE2-standard, and from 0.72 to 0.73 for PCE when compared to their respective MCPS-PRS integrated scores (**Figure 6.18**).

Similar to the main analysis (summarised in **Section 6.3.2** above), both SCORE2 integrated scores demonstrated positive categorical NRI values when compared to the clinical risk score they were derived from. The categorical NRI was 2.90 (95% CI, 0.95-4.86) when comparing SCORE2-diabetes-recalibrated x MCPS-PRS to SCORE2-diabetes-recalibrated, and 3.49 (95% CI, 1.40-5.57) when comparing SCORE2-standard x MCPS-PRS to SCORE2-standard (**Table 6.8** to **Table 6.9**). Continuous NRI was also strongly positive, at 27.17 (95%CI, 21.09-33.26). The IDI remained positive as 1.07% (0.83%-1.32%) for the diabetes-recalibrated SCORE2 integrated

score and 0.80% (0.61%-0.99%) for the standard SCORE2 integrated scores, respectively.

When comparing PCE × MCPS-PRS compared to PCE alone, the categorical NRI was not statistically significant, being 1.59 (95% CI: -0.04-3.23) (**Table 6.10**), whereas the continuous NRI was strongly positive at 26.11 (20.04-32.18). The PCE integrated score achieved an IDI of 1.64% (1.24%-2.03%).

When the analysis was restricted to the first 10 years of follow up, all integrated scores showed non-significant categorical NRI when compared to their corresponding clinical risk scores. However, both the continuous NRI and prospective NRI were strongly positive, suggesting that including genetic inforation into clinical risk scores improved overall discrimination (see **Table 6.11** to **Table 6.13**) .

Table 6.14 to **Table 6.15** present the categorical NRI comparing clinical and integrated risk scores, and the sensitivity and specificity of each score, across five risk thresholds (i.e., 5%, 7.5%, 10%, 12.5% and 15%). As the risk threshold for CHD increases, the overall NRI also increases, primarily driven by a greater improvement in sensitivity(i.e., more cases were reclassified correctly). This was due to the initially low sensitivity of the clinical risk scores, combined with an imbalance between low case numbers and high control numbers, which made improvements in sensitivity more pronounced.

6.4 Discussion

6.4.1 Summary of findings

In this chapter, three guideline-recommended conventional clinical risk scores (two variations of SCORE2[6] and PCE[7]) for CVD were evaluated for their predictability of CHD, both before and after integration with CHD PRS, to assess whether their combined relevance could improve CHD reclassification in a large study of Mexicans. Overall, all three clinical risk scores showed reasonable transferability, with positive associations observed between each score and CHD risk in this

Mexican population. Individuals classified as being at higher clinical risk also exhibited stronger genetic predisposition to CHD, further supporting the utility of clinical risk scores. The average strength of the associations with CHD risk across these clinical scores was largely similar, however, PCE and SCORE2-diabetes recalibrated demonstrated better model discrimination. In addition, all three clinical scores demonstrated stronger discriminative power for CHD events occurring within the first 10 years of follow-up, the time frame for which these scores were originally developed. The results for the integrated genetic-clinical scores were consistent when replacing the externally-developed Patel *et al.* PRS[8] with the novel MCPS-PRS developed in this thesis. While the clinical risk scores effectively captured the impact of sex on CHD risk, they did not adequately account for diabetes status.

Integrating PRS with each CHD clinical risk score did not improve the average strength of association compared to the clinical risk score alone, and only slightly improved the model discrimination. However, both SCORE2 (diabetes-recalibrated and standard) × Patel *et al.* PRS[8] integrated scores significantly improved case reclassification at the 7.5% risk threshold, correctly identifying more cases than the original SCORE2 from which they were derived. Moreover, after incorporating genetic information into the clinical risk scores, CHD risk classification improved in the expected direction: more cases had their predicted risk increased, and more controls had their predicted risk decreased.

6.4.2 Performance of external non-admixed-American based clinical risk score in MCPS

Although both SCORE2 and PCE were derived from non-admixed-American populations, they demonstrated strong performance in MCPS, with discrimination over 0.73. However, at 7.5% CHD risk threshold, the sensitivity for CHD cases were low for SCORE2 and PCE in this admixed-American populations (see **Table 6.15**). The standard SCORE2 score misclassified over half of the true CHD cases as controls at a 7.5% risk threshold, while the diabetes-recalibrated SCORE2 and PCE misclassified 35-40% of true cases as controls (lower than the model dis-

crimination). Lowering the risk threshold would improve sensitivity but would reduce specificity (see **Table 6.15**). This performance gap is likely due to both the lack of incidence data and the lack of calibration of these scores for admixed-American populations. Previous evidence suggested that SCORE2 recalibrated for Asians demonstrated improvement in CVD identification[19]. Therefore, recalibration of the clinical risk scores evaluated is needed using admixed-American-specific CHD rates to improve their utility and transferability beyond the source populations and to achieve more accurate and balanced CHD risk estimation in populations of admixed-American ancestry[40].

6.4.3 Combining genetic and clinical risk predictors to enhance CHD prediction in clinical settings

The integration of polygenic risk scores (PRS) with conventional clinical risk models represents a growing opportunity to personalise CHD risk prediction. While most studies have included CHD PRS with conventional clinical predictors as separate covariates in risk-prediction models[8,17,18,28,29,41,42], only a few studies have developed a CHD integrated score in populations of European ancestry[2,5]. Yet, all these CHD integrated scores demonstrated significant improvements in risk stratification and reclassification, suggesting meaningful potential for clinical translation[2,5]. This chapter contributes to that evidence by presenting, for the first time, a CHD integrated risk score tailored for a large admixed-American population. The SCORE2-diabetes-recalibrated x Patel *et al.* PRS[8] significantly improved risk classification, with a categorical NRI of 1.5, and modestly improved risk prediction compared to clinical scores alone. Moreover, the improvement became more evident, with categorical NRI of 2.9, when the external PRS was replaced by the novel MCPS-PRS in the integrated score, although the confidence interval was wider. These improvements, although modest in absolute terms, are clinically meaningful as they reflect the ability to more accurately classify individuals who will go on to develop CHD, enabling earlier and potentially more effective preventive interventions.

Moreover, the approach used in this thesis for combining genetic and conventional risk predictors adds the individual genetic predisposition to CHD in a manner that can be readily integrated into clinical calculators used in clinical settings. Conventionally, a measure of genetic liability to CHD (i.e., PRS) represents a relative risk for CHD, and it is difficult to interpret without a specific population context. As demonstrated in this chapter, the absolute risk feature of the conventional CHD clinical risk scores (otherwise used routinely in practice) was preserved in the score integrated with PRS information, thus allowing clinicians to use a familiar prediction tool and to directly apply established CHD risk thresholds (now adjusted for polygenic liability by a constant term embedded within the risk-calculator), in order to determine whether initiation, intensification or reduction in CHD preventive treatments is needed. This approach ensures the simplicity and practicality of the integrated score developed in this thesis for routine clinical use.

With a similar aim, one clinical trial has evaluated the use of an integrated score incorporating a CVD PRS and QRISK2[14] in real-world clinical settings, aiming to assess the acceptance and feasibility of such a tool[43]. The study found that the integrated score had a high acceptance rate among both clinicians and patients, and was associated with changes in CVD prevention strategies in over 2% of the participants. This suggests that even modest improvements in risk prediction could meaningfully impact the clinical outcomes if applied at a population scale. Furthermore, this highlights the general applicability and feasibility of a genetically-enhanced risk score in clinical settings.

Finally, the findings also underscore the value of increased representation of admixed-Americans during PRS construction. The integrated score with the MCPS-informed PRS, which included the largest sample of admixed-Americans to date during derivation and was tested within an admixed-American-only cohort, outperformed the integrated score with the Patel *et al.* PRS[8], which, although also multi-ancestry, included only one-eighth as many admixed-Americans during derivation compared to the MCPS-informed PRS. This reinforces the urgent need for improved ancestral diversity in genetic studies. Without better representation of under-represented

populations, the clinical utility of PRS will remain limited and inequitable.

6.4.4 Potential impact of the implementation of a genetically-enhanced CHD risk score at national level in Mexico

In 2020, the total population of Mexico was about 126 million with about 50 million aged 35-79 years (based on estimates from the Mexican National Institute of Statistics and Geography (INEGI)[44]). Using the categorical NRI estimated from the comparison between the integrated SCORE2-diabetes-recalibrated × Patel *et al.* PRS[8] and SCORE2-diabetes-recalibrated alone as an example (**Table 6.2**), the genetically-enhanced integrated score resulted in the lowest misclassification rate for controls. As the overall NRI simply combines the case and control components and may therefore be misleading, the case and control NRIs were evaluated separately to assess the impact of the integrated risk score. Specifically, the NRI was 1.67% for cases and -0.17% for controls. Assuming a CHD case rate of 3.5% (based on the age-standardised CHD prevalence reported previously[45]) extrapolated to the whole country, there would be approximately 1,764,000 individuals with CHD among the entire adult population of 50 million. Applying the integrated score to this population over a ten-year period, an estimated 29,458 additional CHD cases would be correctly reclassified while approximately 82,680 CHD-free individuals would be misclassified as having CHD. Despite more CHD-free individuals being misclassified than true CHD cases correctly reclassified, the integrated score still offers a net benefit in risk stratification by enabling earlier identification of thousands of additional CHD cases, individuals who might otherwise be missed by clinical risk scores alone. With a well-calibrated clinical risk score and a PRS that captures predisposition to CHD risk among admixed-American populations, the integrated score represents a better practical and interpretable tool for personalised CHD prevention in Mexicans in particular and other admixed-American populations in general.

In Mexico, there remain evident inequalities in access to health insurance services between formal and informal sector workers[46,47]. Although covered by health insurance systems, many

individuals still have to spend out-of-pocket health payments, hence the early and more accurate identification of high-risk individuals (and prevention) becomes even more critical. By improving risk stratification and enabling targeted preventive strategies, this genetically-enhanced integrated score has the potential to reduce the burden of late-stage CHD and its associated healthcare costs. Therefore, more resources should be made available to achieve universal health care and cover these additional CHD screening needs.

6.5 Conclusion

In summary, this chapter first evaluated the performance of three guideline-recommended clinical risk scores (SCORE2-diabetes recalibrated, SCORE2-standard, and PCE) within the MCPS cohort. These clinical scores were then integrated with a CHD PRS to assess improvement in risk prediction. The findings indicated that clinical risk scores not calibrated for admixed-American populations tend to underestimate CHD risk in these populations. Genetically-enhanced integrated scores significantly improved CHD risk classification and effectively translated the relative risk conferred by PRS into a format suitable for clinical application. If extrapolated to the whole Mexican population, the genetically-enhanced SCORE2-diabetes-recalibrated integrated score would reclassify approximately 30,000 CHD cases correctly among adults of aged 35-79 years. Currently there is not a consensus on the methods recommended for constructing genetically-enhanced clinical risk scores. Therefore, future research should focus on developing and validating alternative approaches that better integrate polygenic and clinical risk scores to maximise a more reliable prediction of CHD risk, especially among admixed-Americans, and to inform personalised management of CHD risk.

Table 6.1: Reclassification of cases and controls based on 7.5% risk threshold SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated × Patel *et al.* PRS

SCORE2-diabetes-recalibrated risk threshold,%	SCORE2-diabetes-recalibrated x The Patel *et al.* PRS risk threshold,%		
	<7.5%	≥7.5%	%Reclassified*
Cases			
< 7.5 %	1919	245	11
≥ 7.5 %	159	2840	5.3
Controls			
< 7.5 %	92125	3564	3.7
≥ 7.5 %	3349	29006	10

*%Reclassified calculates the percentage of individuals whose risk category (<7.5% vs ≥7.5%) has changed when using the new integrated risk score was applied.

Table 6.2: NRI and IDI comparing SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated × Patel *et al.* PRS

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	1.50 (0.73 - 2.27)	1.67 (0.90 - 2.43)	-0.17 (-0.30 - -0.04)			
Continuous†	17.50 (14.75 - 20.24)	16.52 (13.83 -19.21)	0.97 (0.43 - 1.52)	0.78 (0.69 - 0.87)	0.95	-0.17

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)

*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.

†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.

‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the combined changes in mean predicted risk for cases and control.

Table 6.3: Reclassification of cases and controls based on 7.5% risk threshold SCORE2-standard with SCORE2-standard × Patel *et al.* PRS

SCORE2-standard risk threshold,%	SCORE2-standard x Patel *et al.* PRS risk threshold,%		
	<7.5%	≥7.5%	%Reclassified*
Cases			
< 7.5 %	2522	307	11
≥ 7.5 %	174	2160	7.5
Controls			
< 7.5 %	99358	3532	3.4
≥ 7.5 %	3138	22016	12

*%Reclassified calculates the percentage of individuals whose risk category (<7.5% vs ≥7.5%) has changed when using the new integrated risk score was applied.

Table 6.4: NRI and IDI comparing SCORE2-standard with SCORE2-standard × Patel *et al.* PRS

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	2.27 (1.43 - 3.11)	2.58 (1.75 - 3.41)	-0.31 (-0.43 - -0.18)			
Continuous†	17.50 (14.75 - 20.24)	16.52 (13.83 - 19.21)	0.97 (0.43 - 1.52)	0.57 (0.50- 0.64)	0.70	-0.13

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)

*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.

†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.

‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the combined changes in mean predicted risk for cases and control.

Table 6.5: Reclassification of cases and controls based on 7.5% risk threshold pooled cohort equation with pooled cohort equation × Patel *et al.* PRS

Pooled cohort equation risk threshold,%	Pooled cohort equation x Patel *et al.* PRS risk threshold,%		
	<7.5%	≥7.5%	%Reclassified*
Cases			
< 7.5 %	1723	155	8.3
≥ 7.5 %	124	3161	3.8
Controls			
< 7.5 %	88082	2979	3.3
≥ 7.5 %	2459	34524	6.6

*%Reclassified calculates the percentage of individuals whose risk category (<7.5% vs ≥7.5%) has changed when using the new integrated risk score was applied.

Table 6.6: NRI and IDI comparing pooled cohort equation with pooled cohort equation × Patel *et al.* PRS

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	0.19 (-0.45 - 0.84)	0.60 (-0.03 - 1.23)	-0.41 (-0.52 - -0.29)			
Continuous†	16.65 (13.92 - 19.38)	16.56 (13.88 - 19.24)	0.09 (-0.44 - 0.62)	1.18 (1.04-1.33)	1.41	-0.22

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)

*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.

†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.

‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the combined changes in mean predicted risk for cases and control.

Table 6.7: Baseline characteristics of 26,641 MCPS participants included in the sensitivity analysis of MCPS-PRS

	Men n=8,721 (33%)	Women n=17,920 (67%)	All n=26,641
Age, years	51.9 (12.0)	50.7 (11.7)	51.1 (11.8)
Resident of Coyoacán	3,640 (42%)	6,570 (37%)	10,210 (38%)
Ancestry admixture percentage			
Indigenous American	66.6 (17.9)	67.0 (17.9)	66.9 (17.9)
African	3.4 (2.6)	3.5 (2.9)	3.5 (2.8)
East Asian	1.4 (2.0)	1.4 (1.7)	1.4 (1.8)
European	28.6 (16.3)	28.1 (16.1)	28.2 (16.2)
Highest attained educational level			
University/high school	2,101 (24%)	2,092 (12%)	4,193 (16%)
Middle school	2,309 (26%)	4,381 (24%)	6,690 (25%)
Elementary	3,565 (41%)	8,971 (50%)	12,536 (47%)
Other	743 (9%)	2,464 (14%)	3,207 (12%)
Missing	3 (0%)	12 (0%)	15 (0%)
Smoking status			
Never	1,790 (21%)	11,045 (62%)	12,835 (48%)
Former	2,657 (30%)	2,580 (14%)	5,237 (20%)
Current	4,266 (49%)	4,281 (24%)	8,547 (32%)
Missing	8 (0%)	14 (0%)	22 (0%)
Alcohol intake			
Never	538 (6%)	4,683 (26%)	5,221 (20%)
Former	1,552 (18%)	2,092 (12%)	3,644 (14%)
Current	6,626 (76%)	11,140 (62%)	17,766 (67%)
Missing	5 (0%)	5 (0%)	10 (0%)
Physical measures			
SBP, mmHg	128.6 (15.9)	126.5 (16.8)	127.2 (16.5)
DBP, mmHg	84.5 (9.9)	82.4 (10.1)	83.1 (10.1)
BMI, kg/m²	28.0 (4.4)	29.7 (5.3)	29.2 (5.1)
Waist-to-hip Ratio	0.95 (0.07)	0.88 (0.07)	0.90 (0.08)
Laboratory measurements			
HDL-C, mmol/L	0.93 (0.19)	1.03 (0.22)	1.00 (0.21)
LDL-C, mmol/L	2.40 (0.80)	2.49 (0.80)	2.46 (0.80)
Triglycerides, mmol/L	1.65 (0.68)	1.55 (0.65)	1.58 (0.66)
HbA1c, %	6.11 (1.74)	6.10 (1.72)	6.10 (1.73)
eGFR, ml/min/1.73m²§	101.1 (15.9)	102.0 (16.3)	101.7 (16.2)
Prior disease*			
Coronary heart disease	170 (2%)	216 (1%)	386 (1%)
Stroke	87 (1%)	170 (1%)	257 (1%)
Cancer	50 (1%)	259 (1%)	309 (1%)
Diabetes†	1,695 (19%)	3,251 (18%)	4,946 (19%)
Other‡	478 (5%)	1,794 (10%)	2,272 (9%)

Numbers are n (%) or mean (SD). SBP=systolic blood pressure, DBP=diastolic blood pressure, BMI=body mass index, HDL-C=high density lipoprotein cholesterol, LDL-C=low density lipoprotein cholesterol, HbA1c=glycosylated haemoglobin A1c

*Self-reported previous diagnoses unless otherwise stated.

†Self-reported previously-diagnosed diabetes or glycosylated haemoglobin ≥6.5%.

‡Other diseases include self-reported emphysema, chronic kidney disease, peptic ulcer, liver cirrhosis, and peripheral arterial disease.

§Calculated using the 2021 CKD-EPI equation, based on NMR-measured creatinine levels.

Table 6.8: NRI and IDI comparing SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated x MCPS-PRS among the 20% of MCPS participants used for PRS testing

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	2.90 (0.95 - 4.86)	3.25 (1.32 - 5.19)	-0.35 (-0.66 - -0.04)			
Continuous†	27.17 (21.09 - 33.26)	24.85 (18.89 - 30.81)	2.32 (1.10 - 3.55)	1.07 (0.83 - 1.32)	1.30	-0.23

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)
*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.
†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.
‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the improvement in sensitivity and specificity.

Table 6.9: NRI and IDI comparing SCORE2-standard with SCORE2-standard x MCPS-PRS among the 20% of MCPS participants used for PRS testing

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	3.49 (1.40 - 5.57)	4.14 (2.08 - 6.21)	-0.66 (-0.96 - -0.35)			
Continuous†	27.17 (21.09 - 33.26)	24.85 (18.89 - 30.81)	2.32 (1.10 - 3.55)	0.80 (0.61 - 0.99)	0.99	-0.19

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)
*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.
†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.
‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the improvement in sensitivity and specificity.

Table 6.10: NRI and IDI pooled cohort equation with Pooled cohort equation x MCPS-PRS among the 20% of MCPS participants used for PRS testing

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	1.59 (-0.04 - 3.23)	2.37 (0.76 - 3.98)	-0.78 (-1.06 - -0.50)			
Continuous†	26.11 (20.04 - 32.18)	24.56 (18.61 - 30.51)	1.56 (0.37 - 2.75)	1.64 (1.24 - 2.03)	1.93	-0.30

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)
*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.
†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.
‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the improvement in sensitivity and specificity.

Table 6.11: Net reclassification index comparing SCORE2-diabetes-recalibrated with SCORE2-diabetes-recalibrated x Patel *et al.* PRS in the first 10 years of follow up

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	0.25 (-0.83 - 1.33)	0.46 (-0.62 - 1.53)	-0.20 (-0.33 - -0.08)			
Continuous†	16.27 (11.88 - 20.67)	15.47 (11.11 - 19.84)	0.80 (0.26 - 1.35)	0.83 (0.65 - 1.02)	1.01	-0.17
Prospective Continuous§	16.33 (12.13 - 19.50)	15.52 (11.08 - 18.78)	0.81 (0.29 - 1.27)			

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)
*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.
†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.
§The prospective NRI looks at all upward or downward shifts in predicted risk over a specified follow-up period. It calculaties the proportion of cases by that time whose predicted risk went up and down, and the proportion of controls whose predicted risk went up and down, while accounting for censoring.
‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the combined changes in mean predicted risk for cases and control.

Table 6.12: Net reclassification index comparing SCORE2-standard with PCE x Patel *et al.* PRS in the first 10 years of follow up

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	-0.06 (-1.29 - 1.17)	0.30 (-0.92 - 1.53)	-0.37 (-0.49 - -0.24)			
Continuous†	16.27 (11.88 - 20.67)	15.47 (11.11 - 19.84)	0.80 (0.26 - 1.35)	0.58 (0.44 - 0.72)	0.71	-0.13
Prospective Continuous§	16.33 (12.68 - 21.45)	15.52 (12.20 - 20.32)	0.81 (0.11 - 1.35)			

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)
*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.
†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.
§The prospective NRI looks at all upward or downward shifts in predicted risk over a specified follow-up period. It calculaties the proportion of cases by that time whose predicted risk went up and down, and the proportion of controls whose predicted risk went up and down, while accounting for censoring.
‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the combined changes in mean predicted risk for cases and control.

Table 6.13: Net reclassification index comparing SCORE2-diabetes-recalibrated with PCE x Patel *et al.* PRS in the first 10 years of follow up

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)	IDI,% (95%CI)‡	Change in mean predicted risk for cases,%	Change in mean predicted risk for controls,%
Categorical*	0.76 (-0.05 - 1.57)	1.17 (0.37 - 1.97)	-0.40 (-0.52 - -0.29)			
Continuous†	15.35 (10.97 - 19.74)	15.42 (11.07 - 19.78)	-0.07 (-0.60 - 0.46)	1.29 (0.99 - 1.59)	1.52	-0.23
Prospective Continuous§	15.42 (11.29 - 19.87)	15.48 (11.34 - 19.74)	-0.06 (-0.67 - 0.39)			

Net reclassification index (NRI) = NRI_{case} + $NRI_{control}$ =($P_{up|case} - P_{down|case}$) + ($P_{down|control} - P_{up|control}$)
*The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.
†The continuous NRI looks at all upward or downward shifts in predicted risk. It calculates the proportion cases had their predicted risk go up (from clinical risk score to the new integrated score) and down, and the proportion of controls had their predicted risk go up and down.
§The prospective NRI looks at all upward or downward shifts in predicted risk over a specified follow-up period. It calculaties the proportion of cases by that time whose predicted risk went up and down, and the proportion of controls whose predicted risk went up and down, while accounting for censoring.
‡Integrated discrimination improvement (IDI) measures how well the new integrated score discriminates cases and controls comparing to clinical risk score. It calculates the combined changes in mean predicted risk for cases and control.

Table 6.14: Categorical NRI comparing each clinical risk score with their corresponding integrated risk score across different risk thresholds

	Overall NRI, (95%CI)	Case NRI,% (95%CI)	Control NRI,% (95%CI)
SCORE2-diabetes-recalibrated x Patel et al. PRS vs SCORE2-diabetes-recalibrated			
5% risk threshold	-0.08 (-0.77 - 0.62)	-0.10 (-0.77 - 0.58)	0.02 (-0.11 - 0.16)
7.5% risk threshold*	1.50 (0.73 - 2.27)	1.67 (0.90 - 2.43)	-0.17 (-0.30 - -0.04)
10% risk threshold	2.53 (1.67 - 3.40)	3.02 (2.17 - 3.88)	-0.49 (-0.61 - -0.37)
12.5% risk threshold	3.33 (2.47 - 4.20)	3.89 (3.03 - 4.75)	-0.56 (-0.67 - -0.45)
15% risk threshold	3.86 (3.03 - 4.70)	4.55 (3.72 - 5.38)	-0.69 (-0.79 - -0.59)
SCORE2-standard x Patel et al. PRS vs SCORE2-standard			
5% risk threshold	1.15 (0.35 - 1.94)	1.36 (0.57 - 2.14)	-0.21 (-0.35 - -0.07)
7.5% risk threshold*	2.27 (1.43 - 3.11)	2.58 (1.75 - 3.41)	-0.31 (-0.43 - -0.18)
10% risk threshold	3.51 (2.63 - 4.38)	4.01 (3.14 - 4.88)	-0.50 (-0.62 - -0.39)
12.5% risk threshold	3.21 (2.39 - 4.04)	3.64 (2.82 - 4.46)	-0.43 (-0.53 - -0.33)
15% risk threshold	2.84 (2.11 - 3.57)	3.37 (2.64 - 4.10)	-0.53 (-0.61 - -0.44)
Pooled cohort equation x Patel et al. PRS vs Pooled cohort equation			
5% risk threshold	0.91 (0.32 - 1.50)	-1.53 (-2.10 - -0.96)	2.44 (2.31 - 2.56)
7.5% risk threshold*	0.19 (-0.45 - 0.84)	0.60 (-0.03 - 1.23)	-0.41 (-0.52 - -0.29)
10% risk threshold	0.93 (0.24 - 1.62)	0.23 (-0.45 - 0.91)	0.70 (0.59 - 0.80)
12.5% risk threshold	1.75 (1.03 - 2.48)	2.11 (1.39 - 2.83)	-0.36 (-0.46 - -0.26)
15% risk threshold	2.04 (1.31 - 2.76)	1.70 (0.99 - 2.42)	0.33 (0.23 - 0.43)

Net reclassification index (NRI) = $NRI_{case} + NRI_{control} = (P_{up|case} - P_{down|case}) + (P_{down|control} - P_{up|control})$
This table shows the categorical NRI comparing each clinical risk score with their corresponding integrated risk score, across different risk thresholds. The categorical NRI calculates the proportion of actual reclassification (based on the reclassification table) in cases and in controls between clinical risk score and the new integrated score.
*Risk threshold used in the primary analysis.

Table 6.15: Sensitivity and specificity of risk scores across different risk thresholds

Risk Threshold	5%		7.5%		10%		12.5%		15%	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
SCORE2-diabetes-recalibrated	0.747	0.615	0.581	0.747	0.430	0.831	0.306	0.886	0.210	0.923
x Patel et al./PGS003725/ Multi-ancestry/1,273,824 SNPs	0.746	0.615	0.598	0.746	0.460	0.826	0.345	0.880	0.255	0.916
SCORE2-standard	0.652	0.67	0.452	0.804	0.288	0.880	0.179	0.925	0.107	0.955
x Patel et al./PGS003725/ Multi-ancestry/1,273,824 SNPs	0.666	0.668	0.478	0.800	0.328	0.875	0.215	0.921	0.141	0.950
Pooled cohort equation	0.764	0.578	0.636	0.711	0.558	0.764	0.468	0.819	0.407	0.846
x Patel et al./PGS003725/ Multi-ancestry/1,273,824 SNPs	0.749	0.603	0.642	0.707	0.560	0.771	0.489	0.816	0.424	0.849

Sensitivity measures the proportion of true cases correctly identified by the risk score as having the disease, while specificity measures the proportion of controls correctly identified as disease-free.

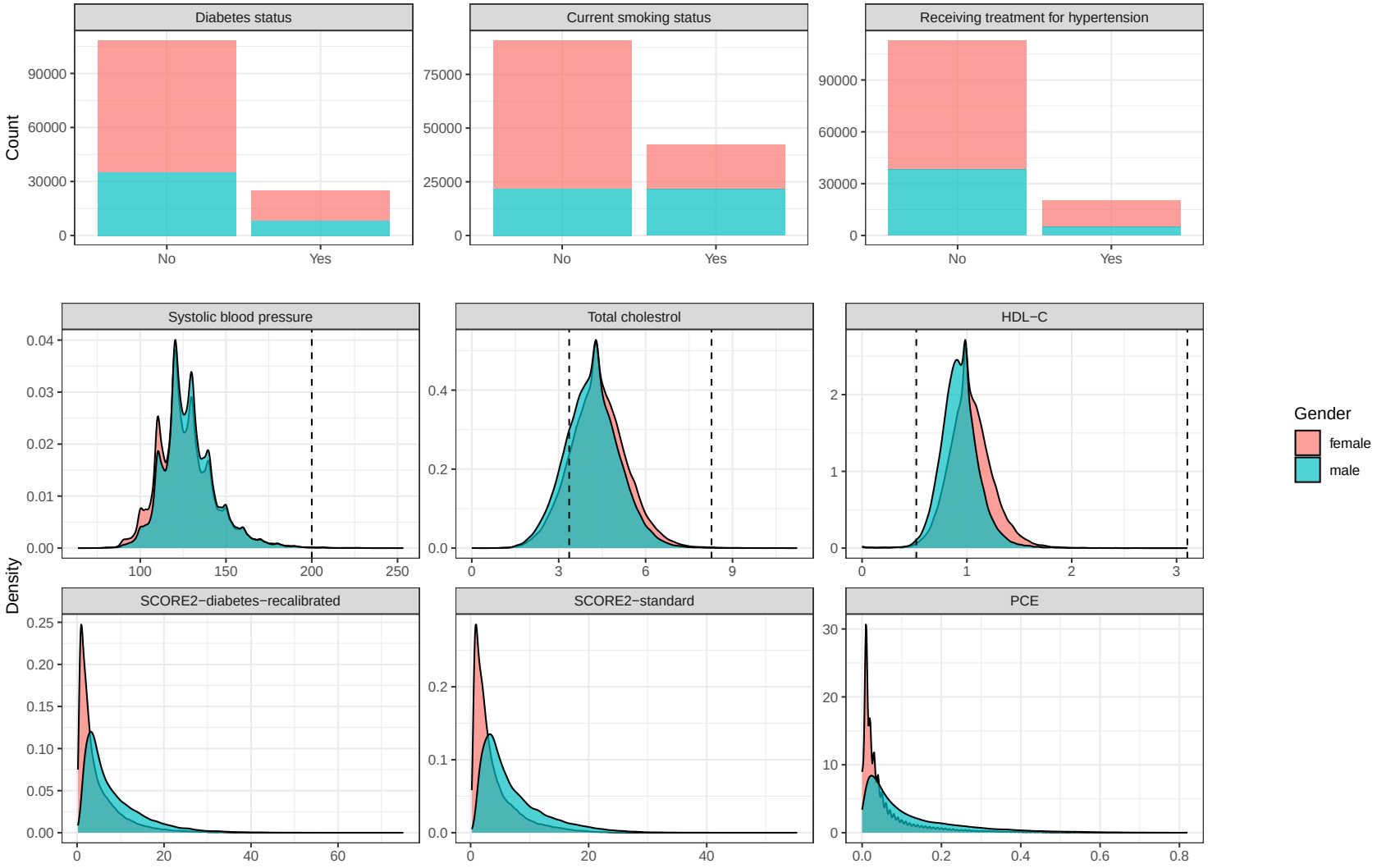


Figure 6.1: Distribution of clinical risk scores and their components in MCPS

The SCORE2-standard risk prediction model includes age, sex, current smoking status, SBP, total cholesterol and HDL-C. SCORE2-diabetes-recalibrated additionally includes diabetes status. The PCE risk prediction model includes age, sex, total cholesterol, HDL-C, diabetes status, current smoking status, SBP, and whether an individual with hypertension takes treatment. The dashed lines represent the valid input ranges for PCE.

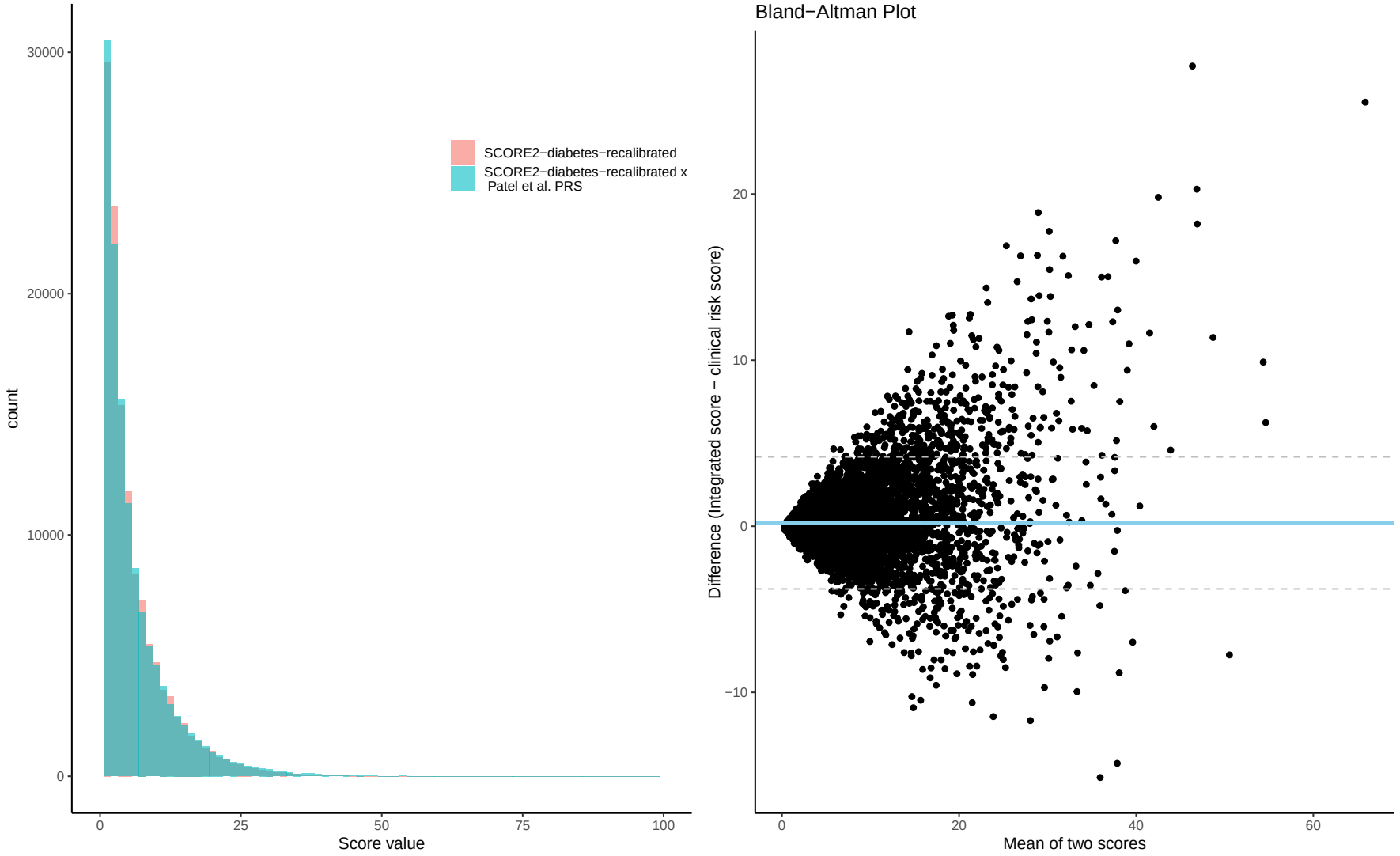


Figure 6.2: Distribution of SCORE2-diabetes-recalibrated and SCORE2-diabetes-recalibrated x Patel *et al.* PRS (left), and agreement between the two scores (right).

The left panel shows the distribution of SCORE2-diabetes-recalibrated and the integrated score combining SCORE2 and the Patel *et al.* PRS. The right panel shows a Bland-Altman plot comparing the two scores, with the mean difference (solid blue line) and its associated 95% confidence interval (dashed grey lines).

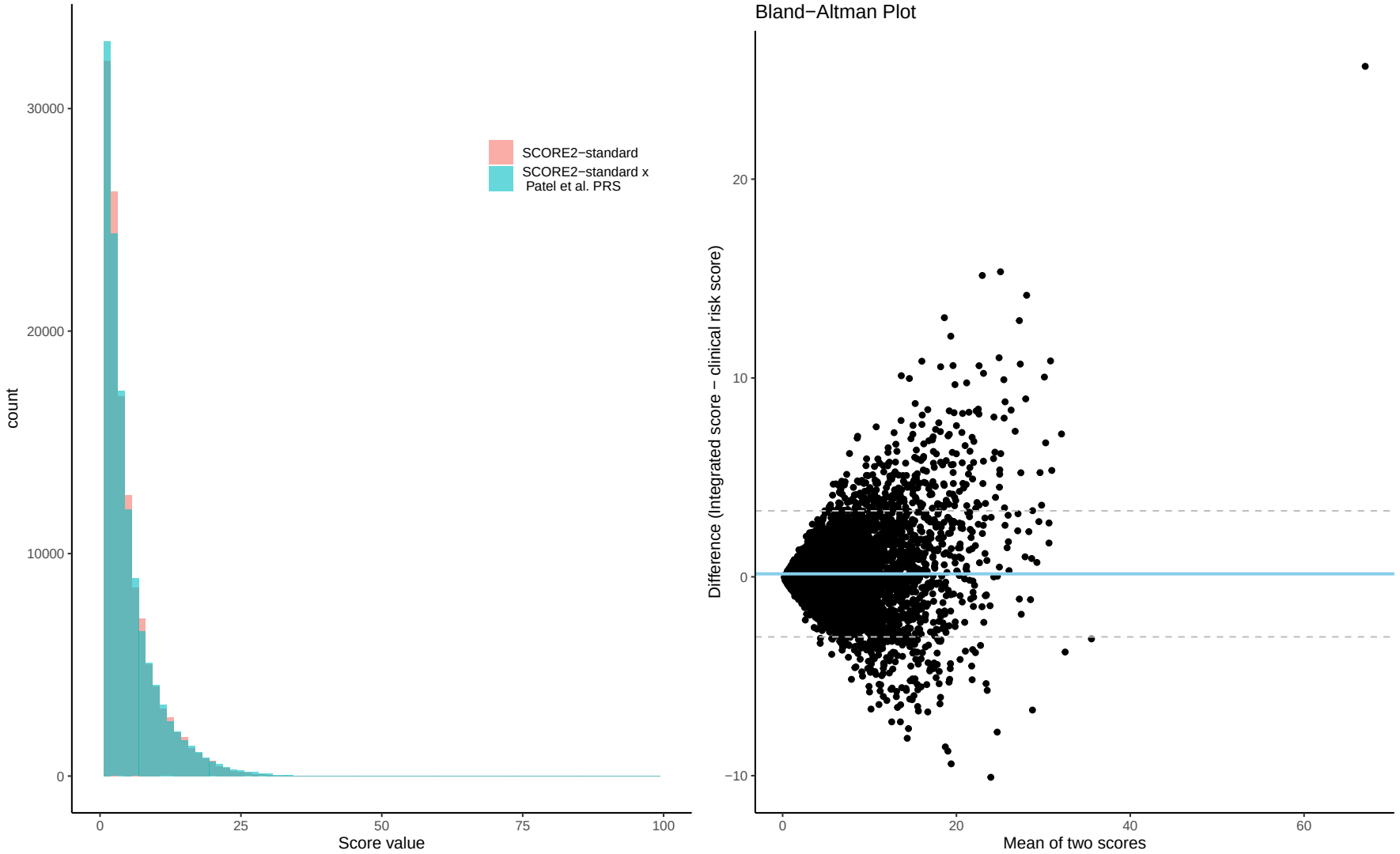


Figure 6.3: Distribution of SCORE2-standard and SCORE2-standard x Patel *et al.* PRS (left), and agreement between the two scores (right).

Analysis as per Figure 6.2

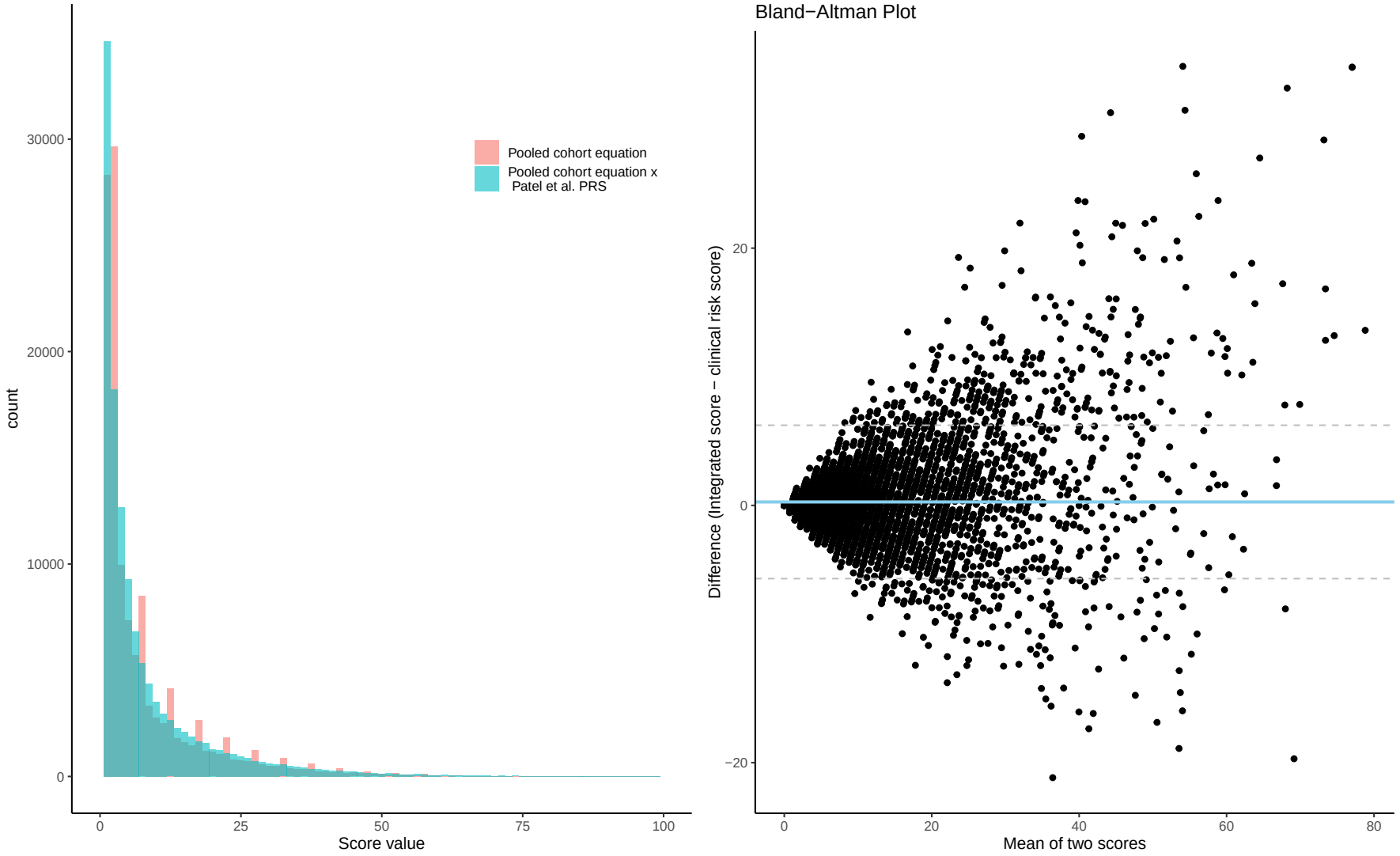


Figure 6.4: Distribution of PCE and PCE x Patel *et al.* PRS (left), and agreement between the two scores (right).

Analysis as per Figure 6.2

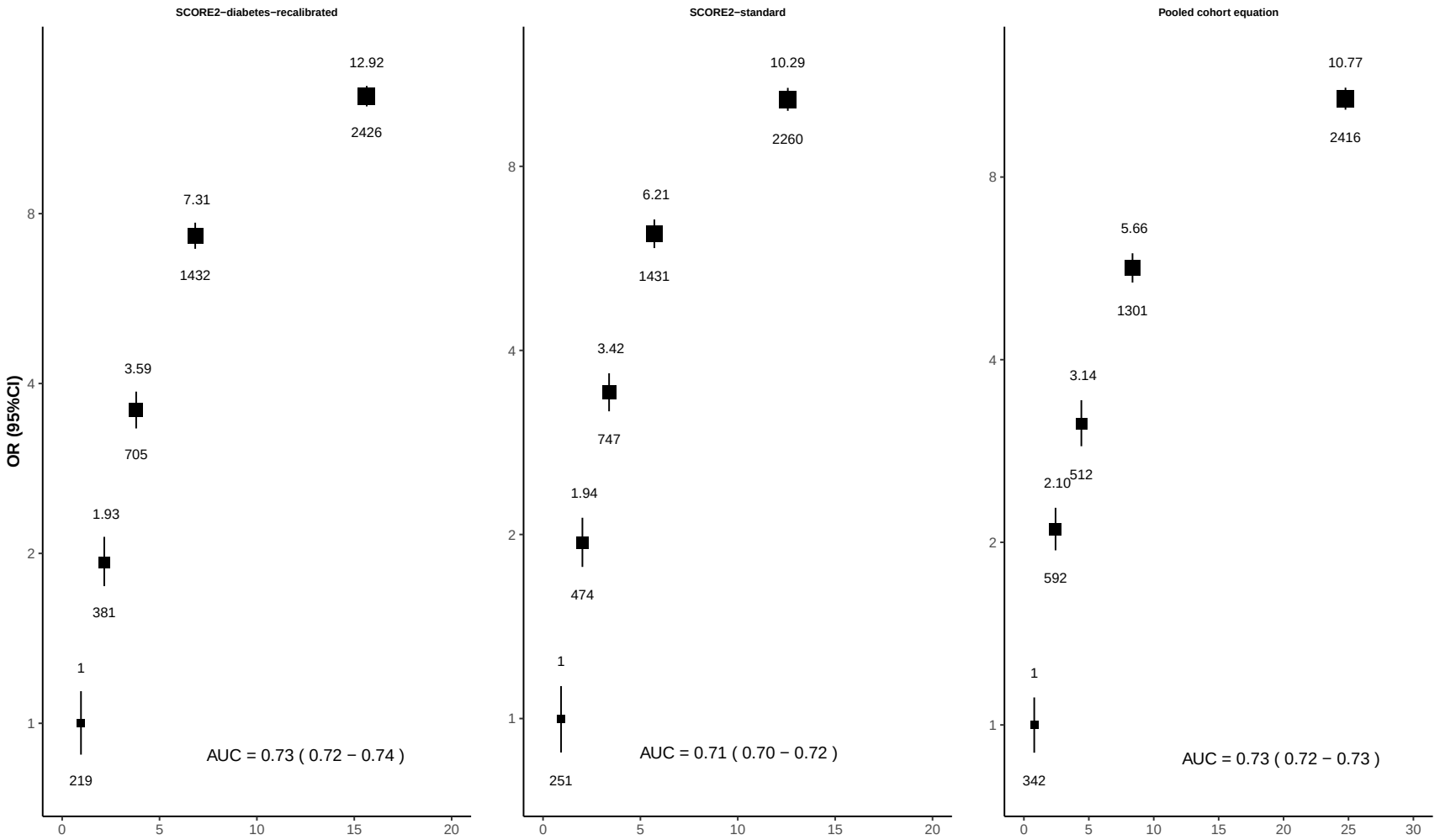


Figure 6.5: Shape of association between each clinical risk score and CHD

Each clinical risk score was split into five equally-sized group. Analyses had no other adjustments. Each group is plotted against the mean of the clinical risk score. The vertical lines through each point represent 95% confidence intervals and are shown for each category (including the reference category with RR=1.0). The area of each square is inversely proportional to the square of the standard error of the log odds ratio (i.e., it is proportional to the amount of statistical information). ORs are shown above each point and the number of CHD cases below each point.

Clinical score (standardised)	Number of cases	Number of controls	OR per SD (95% CI)	AUC (95%CI)
SCORE2−diabetes−recalibrated	5,163	128,044	1.62 (1.59−1.65)	0.73 (0.72−0.74)
SCORE2−standard	5,163	128,044	1.57 (1.54−1.60)	0.71 (0.70−0.72)
Pooled cohort equation	5,163	128,044	1.59 (1.57−1.62)	0.73 (0.72−0.73)

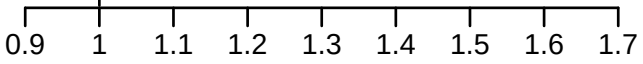


Figure 6.6: Average strength of association between each standardised clinical risk score and CHD

ORs are obtained from univariable logistic regression models with each clinical risk score as sole predictor and no other adjustments.

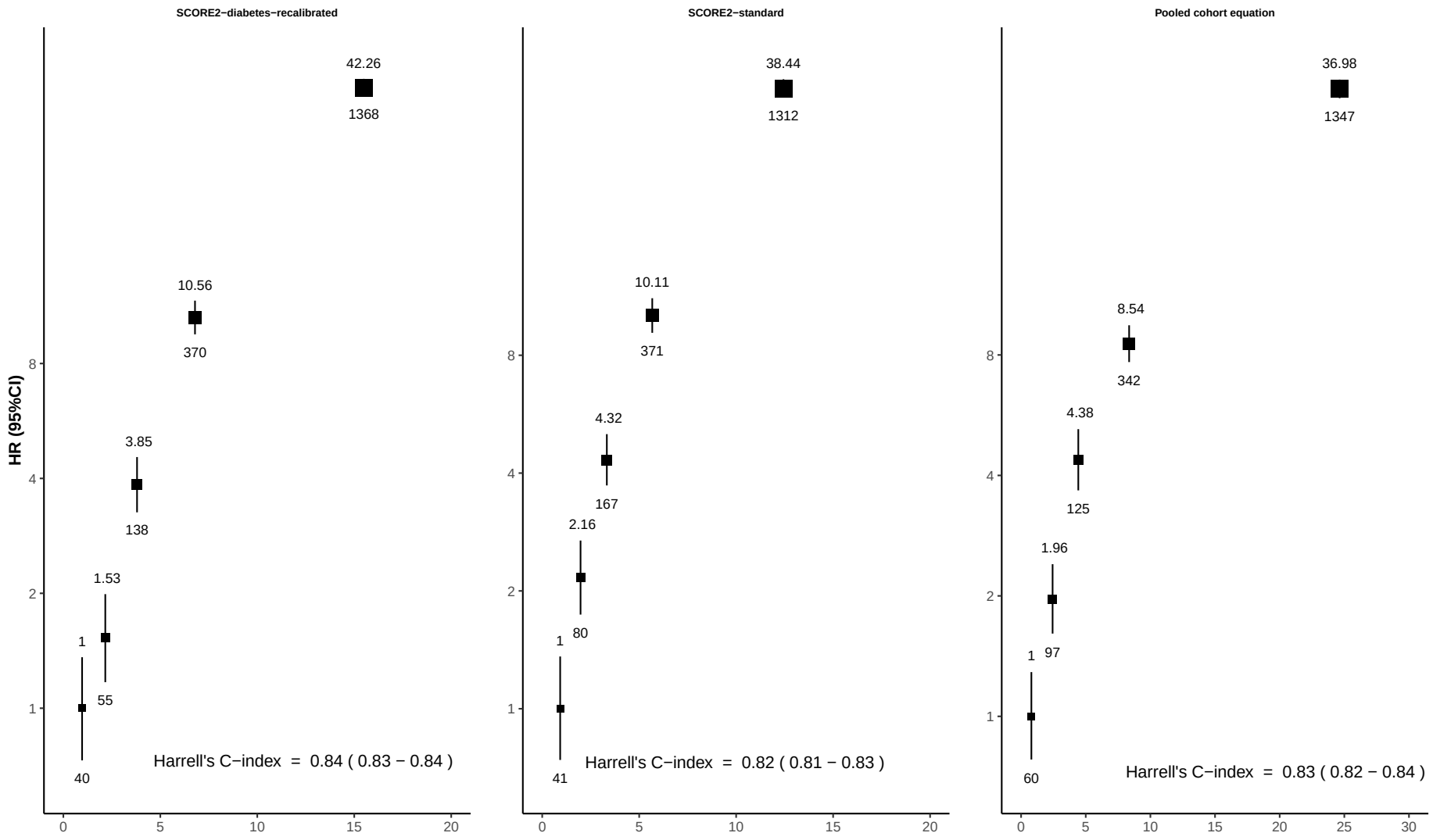


Figure 6.7: Shape of association between each clinical risk score and CHD death in the first 10 years of follow-up

Analyses as per Figure 6.5, using Cox models to estimate the associations between clinical risk scores and fatal CHD, with follow-up censored at 10 years.

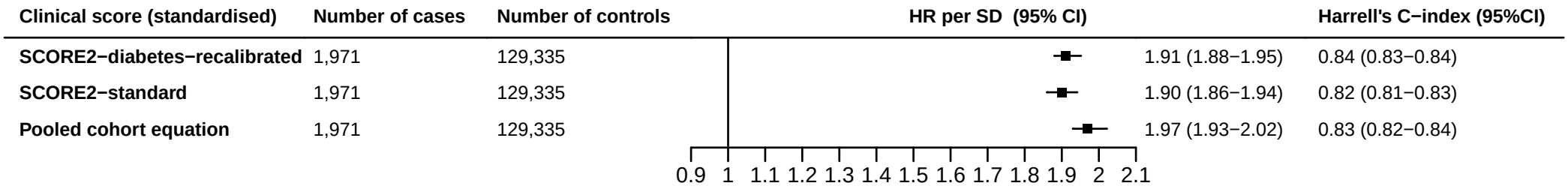


Figure 6.8: Average strength of association between each standardised clinical risk score and CHD death in the first 10 years of follow-up

HRs were derived from univariable Cox regression models including each clinical risk score as the only predictor fatal CHD as the outcome, with follow-up censored at 10 years.

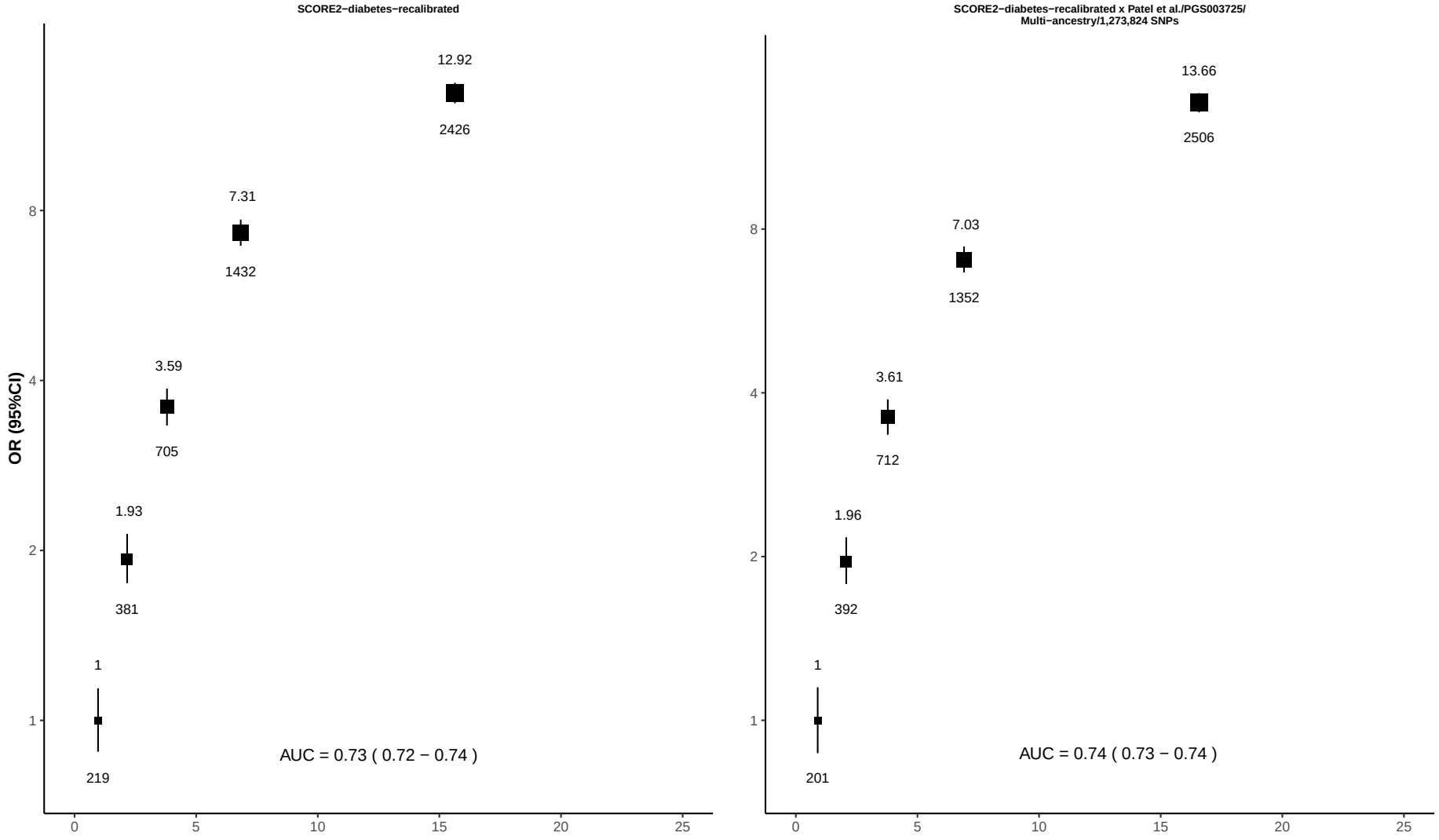


Figure 6.9: Shape of association between SCORE2-diabetes-recalibrated integrated score and CHD

Analyses as for Figure 6.5

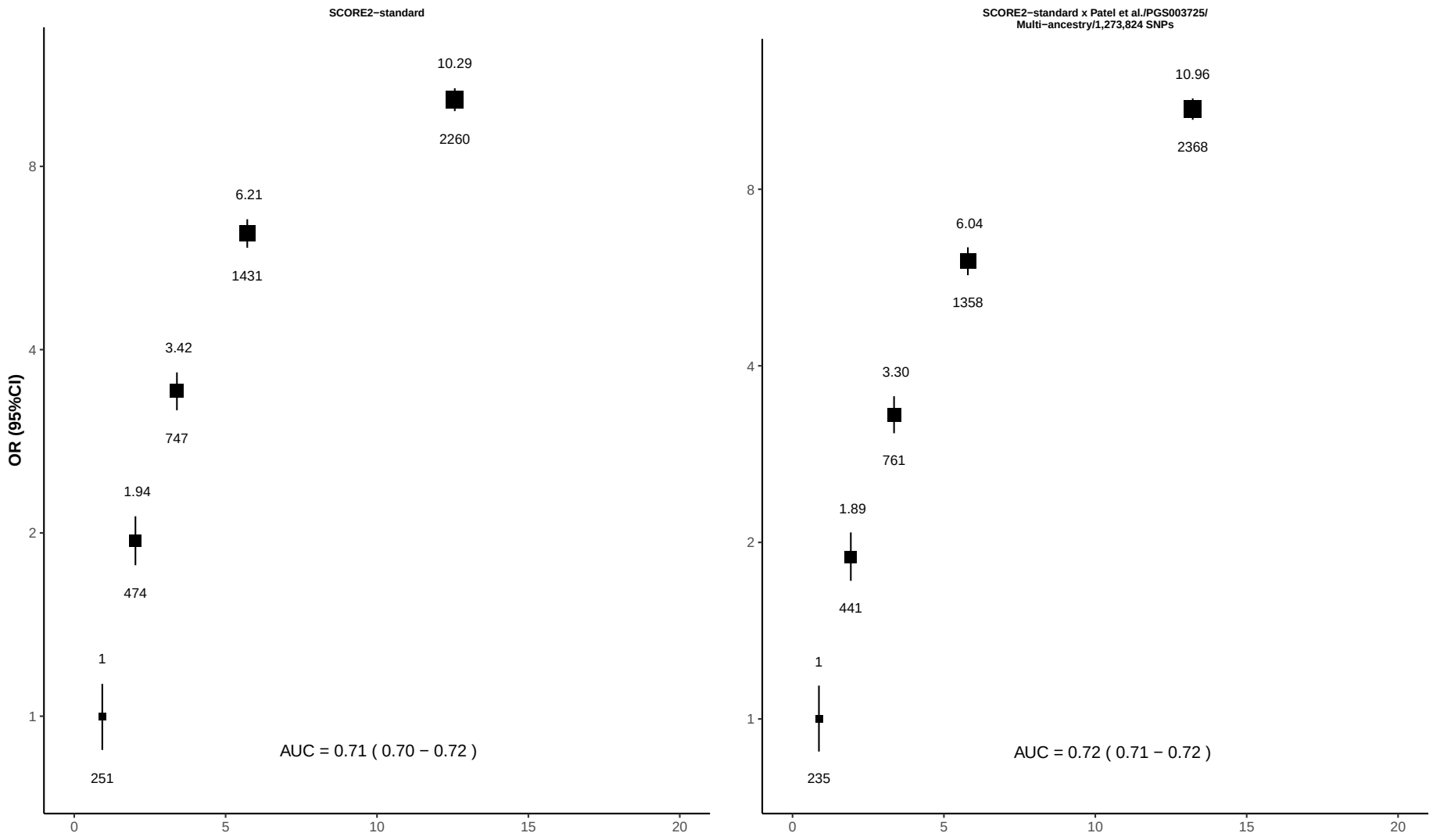


Figure 6.10: Shape of association between SCORE2-standard integrated score and CHD

Analyses as for Figure 6.5

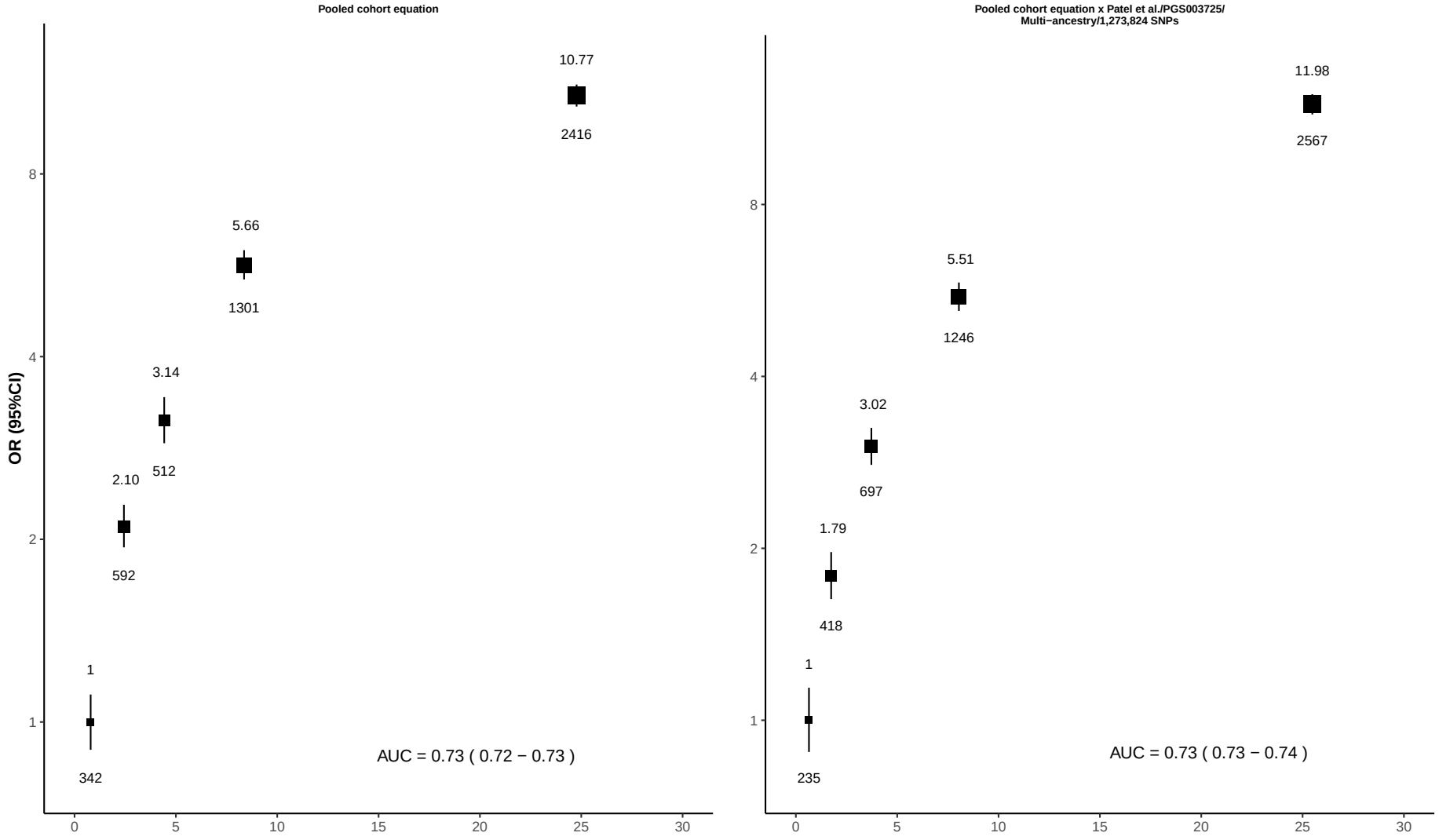


Figure 6.11: Shape of association between PCE and CHD

Analyses as for Figure 6.5

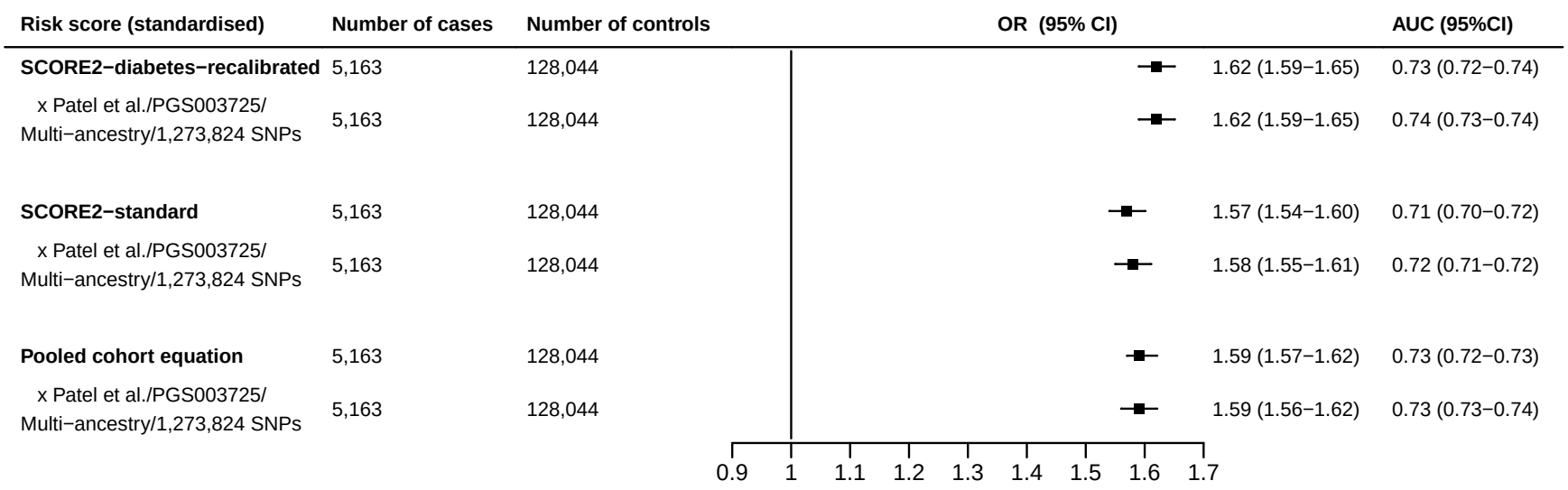


Figure 6.12: Average strength of association between each standardised integrated score and CHD

Analyses as for Figure 6.6

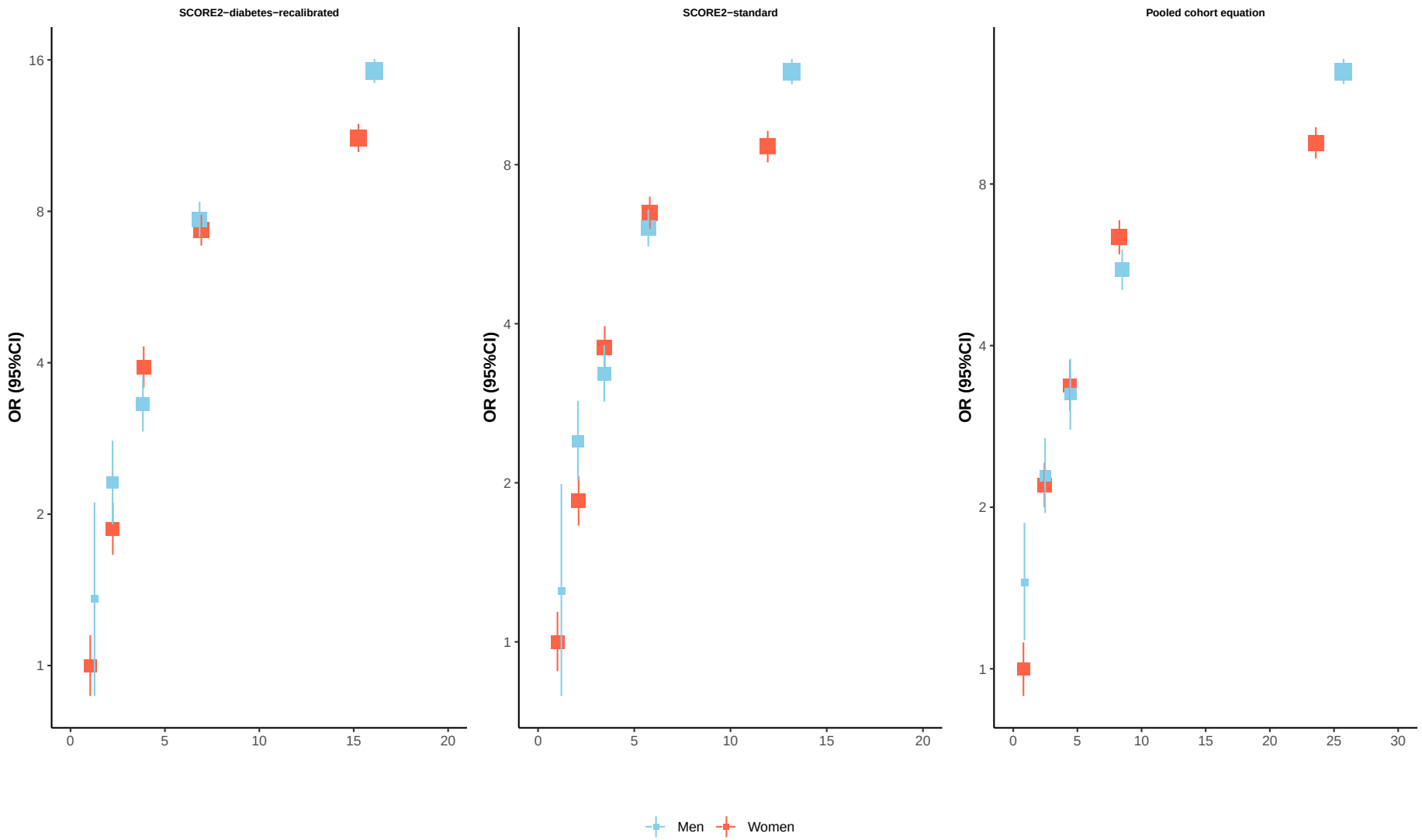


Figure 6.13: Shape of association between each clinical risk score and CHD, by sex

Each clinical risk score was into ten equally sized groups, with groups 1–5 representing women and groups 6–10 representing men. The analysis followed the same approach as in Figure 6.5, using women in group 1 (the lowest 20%) as the reference group.

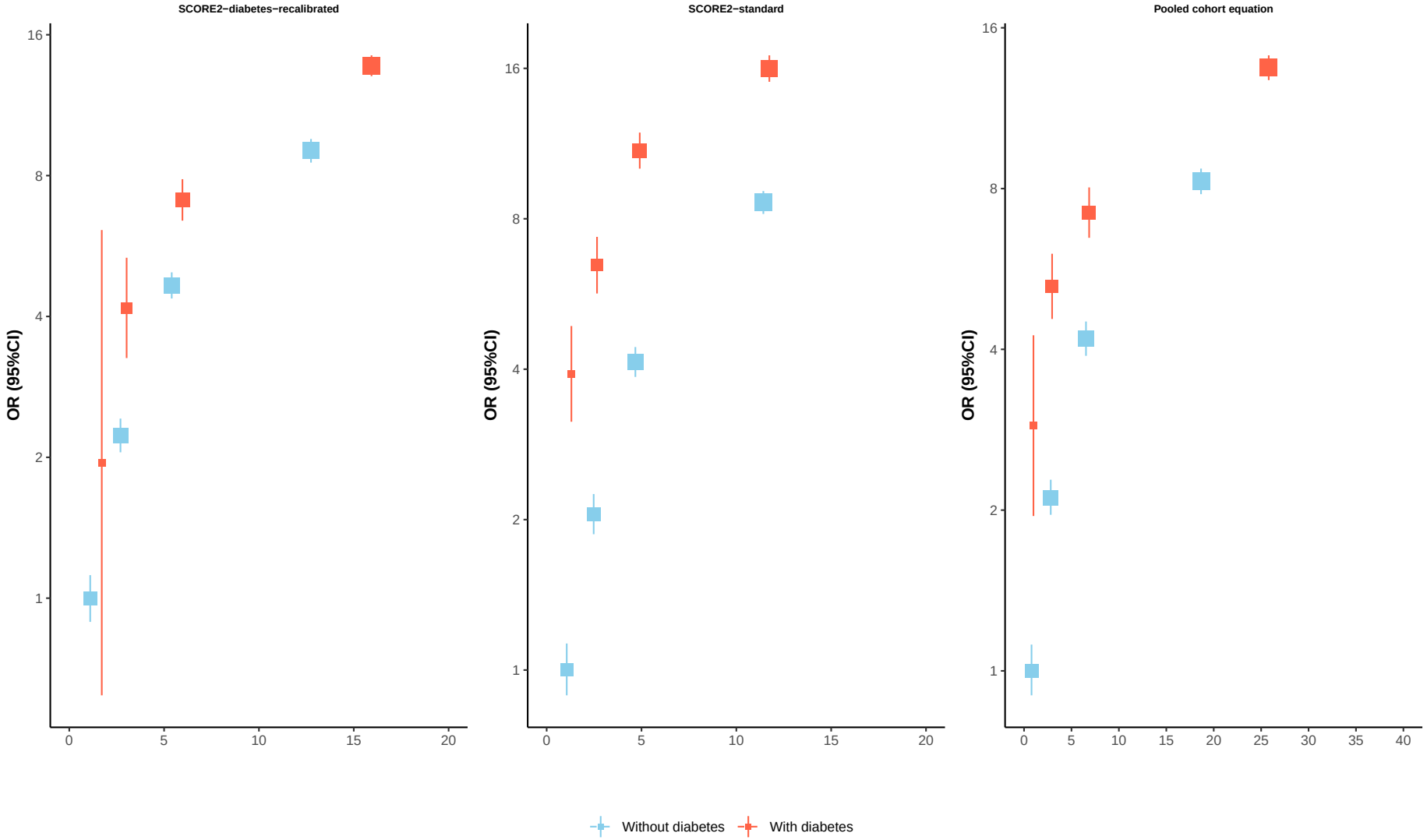


Figure 6.14: Shape of association between each clinical risk score and CHD, by baseline diabetes status

Each clinical risk score was divided into 8 equally sized groups, with groups 1-4 representing those without diabetes and groups 6-10 representing those with diabetes. The analysis followed the same approach as in Figure 6.5, using those without diabetes in group 1 (the lowest 20%) as the reference group.

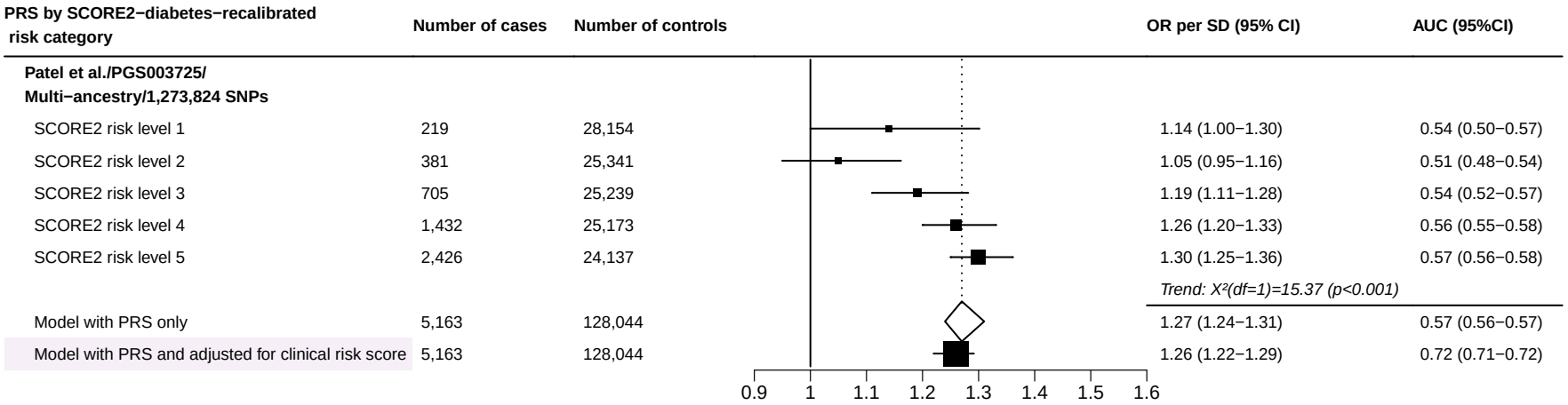


Figure 6.15: Odds of premature CHD per 1 SD increase in PRS, by SCORE2-diabetes-recalibrated risk strata

Subgroup-specific effects are from the model with PRS as predictor and CHD as outcome with no other adjustments.

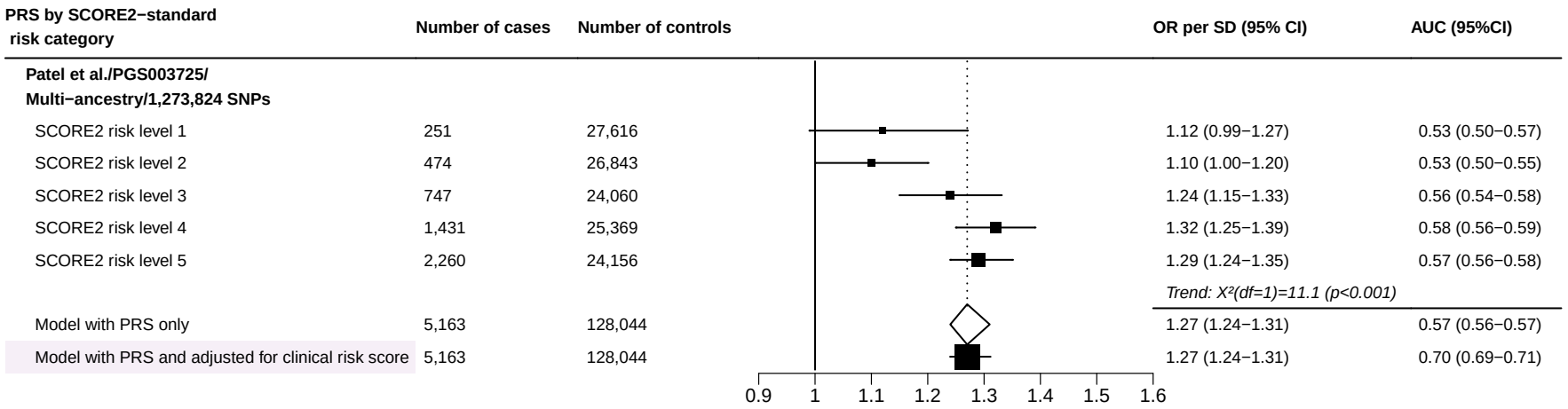


Figure 6.16: Odds of premature CHD per 1 SD increase in PRS, by SCORE2-standard risk strata

Analyses as for Figure 6.15

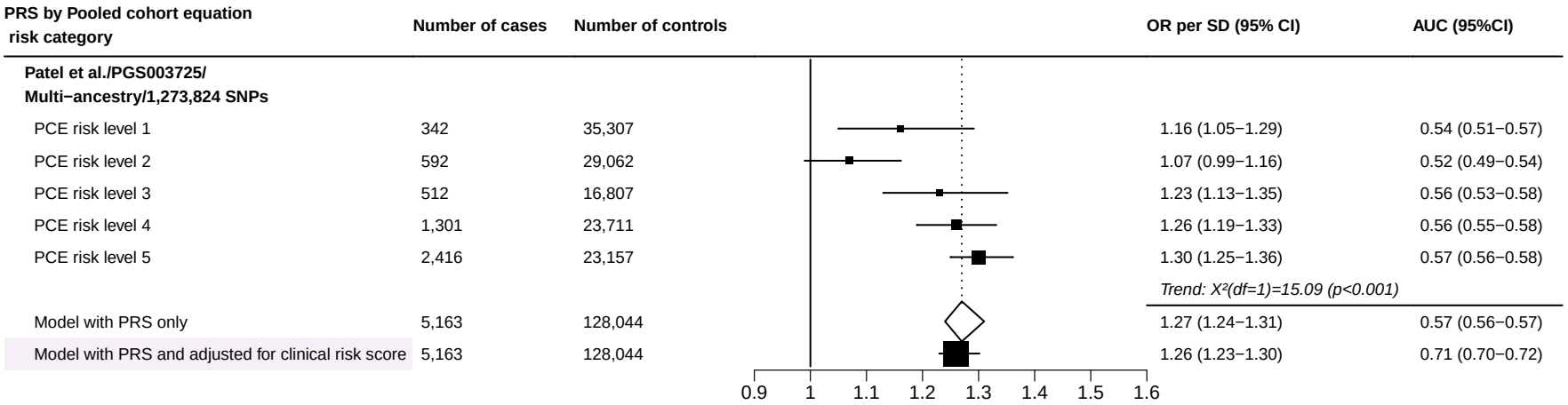


Figure 6.17: Odds of premature CHD per 1 SD increase in PRS, by PCE risk strata

Analyses as for Figure 6.15

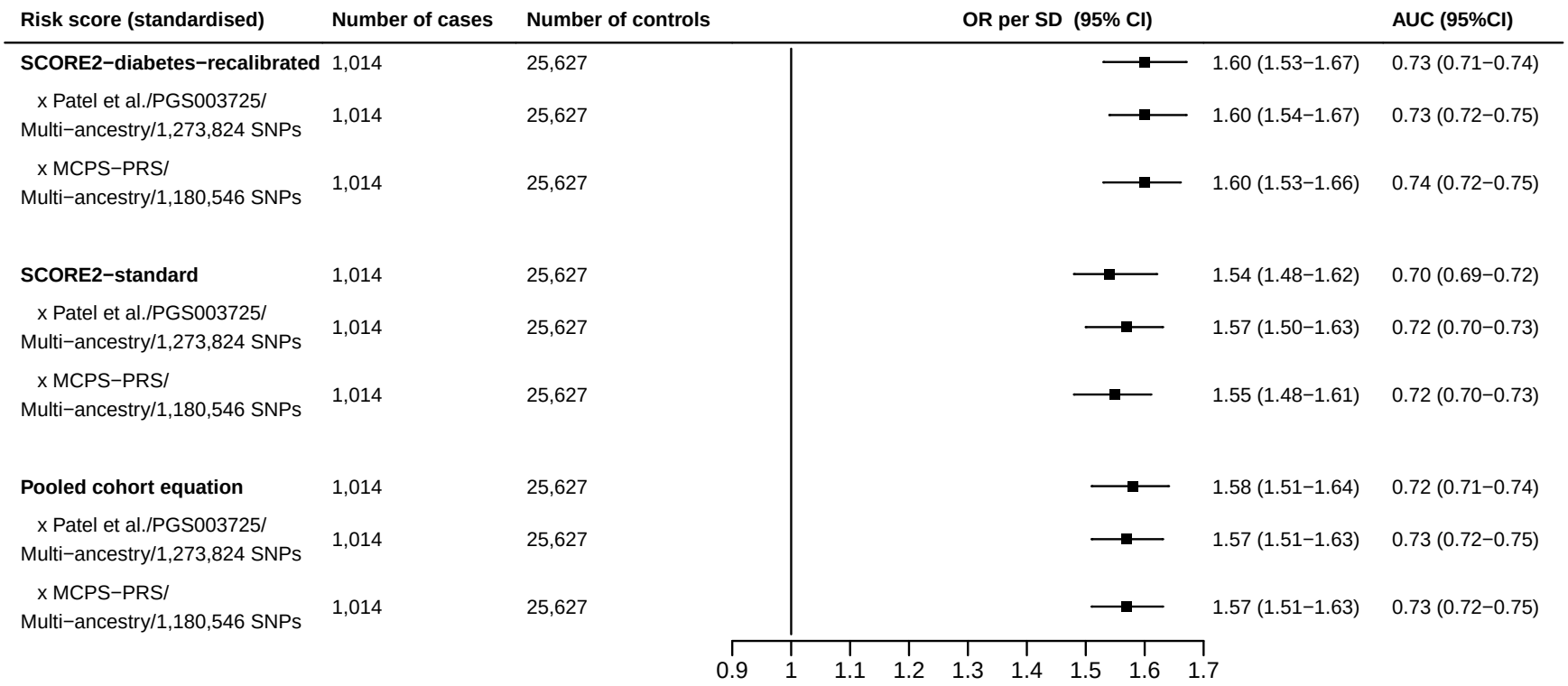


Figure 6.18: Average strength of association between each standardised integrated score and CHD, among 20% subset of the MCPS participants that retained for PRS testing

Analyses as for Figure 6.12

6.6 References

1. Li L, Pang S, Starnecker F, Mueller-Myhsok B, and Schunkert H. “Integration of a polygenic score into guideline-recommended prediction of cardiovascular disease”. *Eur Heart J* 2024.
2. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, et al. “Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction”. *Circ Genom Precis Med* 2021;14(2):e003304.
3. Samani NJ, Beeston E, Greengrass C, Riveros-McKay F, Debiec R, Lawday D, et al. “Polygenic risk score adds to a clinical risk score in the prediction of cardiovascular disease in a clinical setting”. *European Heart Journal* 2024;45(34):pp. 3152–3160.
4. Weale ME, Riveros-Mckay F, Selzam S, Seth P, Moore R, Tarran WA, et al. “Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries”. *American Journal of Cardiology* 2021;148:pp. 157–164.
5. Ramírez J, Duijvenboden S van, Young WJ, Tinker A, Lambiase PD, Orini M, et al. “Prediction of Coronary Artery Disease and Major Adverse Cardiovascular Events Using Clinical and Genetic Risk Scores for Cardiovascular Risk Factors”. *Circ Genom Precis Med* 2022;15(5):e003441.
6. SCORE2 working group ESC Cardiovascular risk collaboration. “SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe”. *European Heart Journal* 2021;42(25):pp. 2439–2454.
7. Goff D. C. J, Lloyd-Jones DM, Bennett G, Coady S, D’Agostino R. B. S, Gibbons R, et al. “2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines”. *J Am Coll Cardiol* 2014;63(25 Pt B):pp. 2935–2959.
8. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. *Nature Medicine* 2023;29(7):pp. 1793–1803.
9. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. “Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention”. *J Am Coll Cardiol* 2018;72(16):pp. 1883–1893.
10. Zhang H, Zeng Y, Yang H, Hu Y, Chen W, Ying Z, et al. “Familial factors, diet, and risk of cardiovascular disease: a cohort analysis of the UK Biobank”. *The American journal of clinical nutrition* 2021;114(5):pp. 1837–1846.
11. Morris RW, Cooper JA, Shah T, Wong A, Drenos F, Engmann J, et al. “Marginal role for 53 common genetic variants in cardiovascular disease prediction”. *Heart* 2016;102(20):pp. 1640–7.
12. Isgut M, Sun J, Quyyumi AA, and Gibson G. “Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later”. *Genome Medicine* 2021;13(1):p. 13.
13. D’Agostino R. B. S, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. “General cardiovascular risk profile for use in primary care: the Framingham Heart Study”. *Circulation* 2008;117(6):pp. 743–53.
14. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. “Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2”. *BMJ* 2008;336(7659):pp. 1475–1482.
15. Hippisley-Cox J, Coupland C, and Brindle P. “Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study”. *BMJ* 2017;357:j2099.
16. Woodward M, Brindle P, and Tunstall-Pedoe H. “Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC)”. *Heart* 2007;93(2):pp. 172–176.
17. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. “Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 636–645.

18. King A, Wu L, Deng HW, Shen H, and Wu C. “Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease”. *BMC Medicine* 2022;20(1):p. 385.

19. SCORE2 Asia-Pacific writing group, Hageman SHJ, Huang Z, Lee H, Kaptoge S, Dorresteijn JAN, et al. “Risk prediction of cardiovascular disease in the Asia-Pacific region: the SCORE2 Asia-Pacific model”. *European Heart Journal* 2024;46(8):pp. 702–715.

20. Perez-Vicencio Daniel, Doudesis Dimitrios, Thurston Alexander J. F., and Selva Jeremy. *RiskScorescvd: Cardiovascular Risk Scores Calculator*. Computer Program. 2025. `https://github.com/dvicencio/RiskScorescvd`.

21. Kaptoge S, Pennells L, De Bacquer D, Cooney MT, Kavousi M, Stevens G, et al. “World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions”. *The Lancet Global Health* 2019;7(10):e1332–e1345.

22. Hajar R. “Risk Factors for Coronary Artery Disease: Historical Perspectives”. *Heart Views* 2017;18(3):pp. 109–114.

23. Fuchs FD and Whelton PK. “High Blood Pressure and Cardiovascular Disease”. *Hypertension* 2020;75(2):pp. 285–292.

24. Grant PJ, Cosentino F, and Marx N. “Diabetes and coronary artery disease: not just a risk factor”. *Heart* 2020;106(17):pp. 1357–1364.

25. SCORE2-OP working group ESC Cardiovascular risk collaboration . “SCORE2-OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions”. *European Heart Journal* 2021;42(25):pp. 2455–2467.

26. SCORE2-Diabetes Working Group and the ESC Cardiovascular Risk Collaboration. “SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe”. *European Heart Journal* 2023;44(28):pp. 2544–2556.

27. Ferrari AJ, Santomauro DF, Aali A, Abate YH, Abbafati C, Abbastabar H, et al. “Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021”. *The Lancet* 2024;403(10440):pp. 2133–2161.

28. de La Harpe R, Thorball CW, Redin C, Fournier S, Müller O, Strambo D, et al. “Combining European and U.S. risk prediction models with polygenic risk scores to refine cardiovascular prevention: the CoLaus|PsyCoLaus Study”. *Eur J Prev Cardiol* 2023;30(7):pp. 561–571.

29. Mosley JD, Gupta DK, Tan J, Yao J, Wells QS, Shaffer CM, et al. “Predictive Accuracy of a Polygenic Risk Score Compared with a Clinical Risk Score for Incident Coronary Heart Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 627–635.

30. Pérez de Isla L, Alonso R, Mata N, Fernández-Pérez C, Muñiz O, Díaz-Díaz JL, et al. “Predicting Cardiovascular Events in Familial Hypercholesterolemia: The SAFEHEART Registry (Spanish Familial Hypercholesterolemia Cohort Study)”. *Circulation* 2017;135(22):pp. 2133–2144.

31. Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. “2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines”. *Circulation* 2019;139(25):e1082–e1143.

32. Privé F, Arbel J, and Vilhjálmsson BJ. “LDpred2: better, faster, stronger”. *Bioinformatics* 2021;36(22-23):pp. 5424–5431.

33. Ruan Y, Lin Y.-F, Feng Y.-CA, Chen C.-Y, Lam M, Guo Z, et al. “Improving polygenic prediction in ancestrally diverse populations”. *Nature Genetics* 2022;54(5):pp. 573–580.

34. Cook NR and Ridker PM. “Advances in Measuring the Effect of Individual Predictors of Cardiovascular Risk: The Role of Reclassification Measures”. *Annals of Internal Medicine* 2009;150(11):pp. 795–802.

35. Pencina MJ, D’Agostino R. B. S, D’Agostino R. B. J, and Vasan RS. “Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond”. *Stat Med* 2008;27(2):157–72, discussion 207–12.

36. Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. “2021 ESC Guidelines on cardiovascular disease prevention in clinical practice: Developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies With the special contribution of the European Association of Preventive Cardiology (EAPC)”. *European Heart Journal* 2021;42(34):pp. 3227–3337.

37. Murphy BS, Hershey MS, Huang S, Nam Y, Post WS, McClelland RL, et al. “PREVENT Risk Score vs the Pooled Cohort Equations in MESA”. *JACC: Advances* 2025;4(6, Part 1):p. 101825.

38. Dhaliwal JS, Gaonkar M, Patel N, Shetty NS, Li P, Vekariya N, et al. “Differences in Statin Eligibility With the Use of Predicting Risk of Cardiovascular Disease EVENTs Versus Pooled Cohort Equations in the UK Biobank”. *American Journal of Cardiology* 2025;241:pp. 43–51.

39. Pencina MJ, D’Agostino Sr RB, and Steyerberg EW. “Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers”. *Statistics in Medicine* 2011;30(1):pp. 11–21.

40. Kuehn BM. “Better Risk Assessment Tools Needed for Hispanic or Latino Patients”. *Circulation* 2019;139(18):pp. 2186–2187.

41. Lu X, Liu Z, Cui Q, Liu F, Li J, Niu X, et al. “A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study”. *European Heart Journal* 2022;43(18):pp. 1702–1711.

42. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, et al. “Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses”. *PLoS Medicine* 2021;18(1) (no pagination).

43. Fuat A, Adlen E, Monane M, Coll R, Groves S, Little E, et al. “A polygenic risk score added to a QRISK®2 cardiovascular disease risk calculator demonstrated robust clinical acceptance and clinical utility in the primary care setting”. *European Journal of Preventive Cardiology* 2024;31(6):pp. 716–722.

44. National Institute of Statistics and Geography. *Demographic and Social Information*. `https://en.www.inegi.org.mx/programas/ccpv/2020/`. 2024.

45. Stark B, Johnson C, and Roth GA. “Global prevalence of coronary artery disease: an update from the global burden of disease study”. *Journal of the American College of Cardiology* 2024;83(13_Supplement):pp. 2320–2320.

46. Meneses Navarro S, Pelcastre-Villafuerte BE, Becerril-Montekio V, and Serván-Mori E. “Overcoming the health systems’ segmentation to achieve universal health coverage in Mexico”. *The International Journal of Health Planning and Management* 2022;37(6):pp. 3357–3364.

47. Garcia-Diaz R. “Effective access to health care in Mexico”. *BMC Health Services Research* 2022;22(1):p. 1027.

Chapter 7: Discussion

7.1 Thesis overview

Coronary heart disease (CHD) is a major global cause of death and has always been a focus of public health targets for countries around the world. Genetics plays a significant role in explaining inter-individual risk differences for CHD and, by aggregating the effects of many genes into a single summary measure, polygenic risk scores (PRSs) can serve as useful tools to identify those at higher risk. However, non-Europeans such as admixed-Americans are highly under-represented in the field of genetic studies. The aim of this thesis was therefore to improve CHD risk prediction among a Mexican population with the help of a CHD PRS.

A literature review on existing PRSs built for CHD risk prediction in diverse ancestries was first conducted in **Chapter 2**. The majority of the identified literature focused exclusively on populations of European ancestry, with no studies of admixed-American populations. However, previously published reports indicated that multi-ancestry PRSs may have greater transferability than the Euro-centric scores and could potentially improve CHD risk predictions among non-European populations. Building on the findings from the literature review, this thesis first evaluated the transferability of eight previously-published CHD PRSs identified from the literature review to predict CHD risk among 130,000 adults from the Mexico City Prospective Study (MCPS) (**Chapter 4**). In the second stage of the thesis, a novel multi-ancestry CHD PRS was developed by leveraging genetic data from the MCPS cohort, and external cohorts of European, East-Asian and admixed-American ancestries. The novel PRS included eight times more admixed-American individuals than any admixed-American samples included in previously-published PRSs, and the predictive performance of the novel MCPS-based score was comparable to that of the best-performing external CHD PRS evaluated in **Chapter 4** (**Chapter 5**). Finally, to further improve

CHD risk prediction and stratification among admixed-Americans, the analyses conducted in this thesis integrated a CHD PRS with three guideline-recommended clinical risk scores (two variations of SCORE2[1,2] and the Pooled Cohort Equation [PCE][3]) that are routinely used to assess CHD risk in clinical settings and assessed the ability of integrated scores to improve CHD risk classification (**Chapter 6**). In each chapter, results were summarised and discussed in light of the existing evidence. This chapter provides an overall summary of the key findings from the analyses described above, and discusses potential implications for prevention, clinical transferability and assessment of CHD risk in admixed-American populations. In addition, this final chapter explores potential areas for future research informed by the present work.

7.2 The relevance of previously-published PRSs to CHD risk in Mexicans

7.2.1 Summary of the main findings

Eight previously-published CHD candidate PRSs identified from literature review (comprising 44 to over 6 million SNPs)[4–11] were evaluated for their association strengths and prediction against CHD among 133,207 MCPS participants aged 35-79 at recruitment. Overall, 5,163 MCPS participants had a history of CHD at recruitment or died before age 80 with CHD mentioned anywhere on the death certificate.

All eight external PRSs associated positively and log-linearly with the risk of CHD, with the scores based on more abundant SNPs and advanced computing methods displaying stronger associations than the more economical PRSs. In particular, the PRS developed by Patel *et al.*[8], which incorporated genetic information across five ancestries (i.e., South and East-Asian, African, Hispanics and European) and included over one million SNPs, demonstrated the strongest association with CHD risk in the MCPS population, with an odds ratio (OR) per standard deviation (SD) of 1.29 (95% confidence interval [CI], 1.35-1.33), when adjusted for age, sex and the first seven genetic principal components (PCs) for population structure. By comparison, the Tada *et al.*[4] PRS, which included only 44 SNPs, showed the weakest association with CHD risk in MCPS,

with an OR per SD of 1.05 (1.03-1.08). Further accounting for conventional CHD risk factors (i.e., blood pressure, highest education attainment, smoking status, baseline diabetes status and waist-to-hip ratio [WHR]) did not materially alter the strengths of the associations between each PRS and CHD risk. Based on a logistic model including just age, sex and the first seven genetic PCs (and without PRS), the model AUC was 0.716 (95% CI, 0.709-0.722). Upon adding conventional CHD risk factors, the AUC improved to 0.745 (0.739-0.751). Including a PRS into these models led to marginal further improvements in model discrimination. For instance, including the Patel *et al.*[8] PRS into these models increased the model AUC to 0.724 (0.717-0.730) and 0.749 (0.743-0.755), respectively.

A striking finding from these initial analyses was the big sex differences found with genetic predisposition to CHD among the Mexican study population. Six out of the eight external PRSs (with thousands of SNPs included) demonstrated significantly stronger associations with CHD risk in men compared to women. After accounting for age and the first seven genetic PCs, the Patel *et al.*[8] PRS showed the strongest association with CHD risk, with an OR per SD of 1.37 (95%CI, 1.32-1.43) in men and 1.23 (1.18-1.28) in women. These findings remained largely consistent after further accounting for conventional CHD risk factors.

Additional investigation of potential effect-modification by different levels of age, educational attainment, body fat distribution, blood pressure, smoking status, diabetes status at recruitment, and indigenous ancestry proportion, were largely consistent with the main findings. When restricting the analysis to unrelated participants, approximately 60% of the participants were retained and 2,947 had CHD (based on the primary CHD definition). The results remained consistent with the main analysis results.

The overall associations for the combined CHD outcomes described above were largely consistent across the specific subtypes of CHD disease-components, specifically, with myocardial infarction, angina or death with CHD as primary cause of death. In addition, sex heterogeneity was apparent for all CHD alternative definitions for the Patel *et al.*[8] PRS, the PRS displayed the

strongest association with CHD in the main analysis, and the most significant for self-reported CHD at baseline.

7.2.2 Finding in context

7.2.2.1 Sex heterogeneity

The sex heterogeneity in genetic predisposition to CHD risk reported in **Chapter 4** aligns with previous findings from European populations, where men consistently exhibited higher genetic risk for CHD than women[8,11,12]. Huang *et al.*[12] further analysed sex differences in CHD-specific mediating factors (i.e., blood pressure, blood lipids and body mass index) and found that only blood pressure showed significant sex heterogeneity. This difference may be partly attributed to a novel genetic locus *21q22.11*, which demonstrated a sex-specific association with CHD risk and was also linked to pulse pressure. In addition, a review on genetic predisposition to CHD risk suggested that the weaker association in women may be attributed to sex-biases in clinical definitions of CHD outcomes (e.g., progression, presentation, age at onset), which have historically been based on male-dominated data[13]. This thesis also found that the sex differences were the most pronounced for self-reported CHD at baseline, raising the possibility that differential awareness of disease status may also contribute to the observed weaker association to CHD risk in women.

7.2.2.2 Genetic predisposition to CHD risk and conventional CHD risk factors

The analyses showed that genetic risk of CHD, as captured by the eight evaluated PRSs, was largely independent of conventional CHD risk factors (e.g., blood pressure, smoking status). This is consistent with current evidence suggesting that PRSs capturing underlying genetic susceptibility complement established risk factors in predicting CHD[11,14–17]. This complementarity in relevance further highlights the need and benefits of integrating genetic and conventional CHD risk factors to enhance CHD risk prediction, for each bear a multiplicative effect on CHD[14,18,19]. It also draws attention to the fact that continuing to address preventable conventional risk factors,

by better controlling their levels towards lower risk, remains a 'must do' at any level of inherited genetic predisposition for CHD. While this approach cannot curb the genetic liability itself, it can meaningfully lower the overall risk, which would otherwise be amplified by modifiable, observable CHD risk factors acting in combination with genetic susceptibility.

7.2.2.3 Multi-ancestry CHD PRS for CHD risk prediction in populations of non-European ancestry

PRS derived from European populations remained positive but showed diminished performance in Mexicans, which aligns with previous studies that applied European-derived PRSs to non-European cohorts[7,20–22]. In an ideal situation, the best-performing PRS in an admixed-American population should be developed using large-scale admixed-American GWAS that could best-capture the genetic risk factors specific to this population. However, due to lack of large admixed-American cohorts, there has been no PRS developed using only populations of admixed-American ancestry. Multi-ancestry PRSs that incorporate genetic information from multiple ancestries have demonstrated superior performance compared to single ancestry PRSs derived from large European cohorts in previous studies among non-European populations[7–9], a pattern that was also observed in the analysis of the MCPS cohort. This suggests that leveraging genetic information from diverse ancestries could improve generalisability of a PRS, as causal variants are likely shared across ancestries. However, the number of admixed-American individuals included in the eight evaluated external PRSs was low. Incorporating a larger representation of admixed-American individuals in the PRS development may further enhance the accuracy of polygenic risk prediction for CHD in this population.

Findings from **Chapter 4** underscore the need for a CHD PRS that better represents admixed-American genetic information to improve CHD risk prediction in this population. These results also highlight the importance of integrating PRS with conventional clinical risk factors to enhance the overall accuracy of CHD risk prediction models.

7.3 Developing a novel CHD PRS leveraging genetic data from MCPS

7.3.1 Summary of the main findings

To address the gap identified in **Chapter 4**, a novel CHD PRS was developed in **Chapter 5**. As a central aim of this thesis, this effort focused on developing a CHD PRS capable of more precisely estimating CHD risk in Latin American populations. By leveraging genetic data from MCPS and external GWAS summary statistics from individuals of European, East-Asian and admixed-American ancestry, a novel MCPS-informed CHD PRS was developed. The score was trained using a 10-fold cross validation (CV) approach and evaluated three methods (Pruning and thresholding [P+T][23], LDpred2[24] and PRS-CSx[25]). This MCPS-informed PRS incorporated genetic information from the largest number of admixed-Americans individuals to date, eight times more than the next most representative PRSs for this population[7,8], which included around 30,000 admixed-American samples from Million Veterans Program (MVP)[26] during PRS construction. External validation of the novel PRS was not possible due to absence of available replication cohorts.

7.3.1.1 PRS training results

The P+T method[23] involved tuning 2,394 combinations of two hyper-parameters, the r^2 and the p-value thresholds. PRS performance generally improved with higher r^2 values and peaked at lower p-value thresholds. The best-performing P+T candidate PRS, based on both a mean AUC of 0.5556 and a mean OR per SD of 1.21, included genetic information from admixed-American and European ancestries, with hyper-parameters r^2=0.9 and p-value threshold as 0.0055.

For PRSs developed using LDpred2[24,27], four hyper-parameters were tuned across 8,640 combinations to identify the optimal hyper-parameter combination: heritability (h^2), fraction of causal markers (p), sparsity and the method for calculating effective sample size. As h^2 increased, the frequency of non-convergence in the LDpred2 model also increased, resulting in the failure to generate PRSs for certain hyper-parameter combinations. Non-convergence was associated

with ancestries included in the input GWAS since, as the number of included ancestries increased, the proportion of hyper-parameter sets that reached convergence declined more rapidly with increasing h^2. The GWAS input that incorporated genetic data from European, East Asian, and admixed-American ancestries achieved the highest performance during training. Based on a mean AUC of 0.5623, the best-performing hyper-parameters for LDpred2 were h^2=0.02, p=0.38, sparsity=True and standard effective sample size calculation method. Based on a mean OR per SD of 1.26, the best performing hyper-parameters were h^2=0.04 and p=0.23, with the same sparsity setting and sample size calculation method as for the best-performing candidate PRS based on mean AUC.

For PRSs derived using PRS-CSx[25], five values of a single hyper-parameter, global scaling ϕ, were used for tuning. Multi-ancestry candidate PRSs constructed using meta-analysed multi-ancestry SNP effect estimates outperformed ancestry-specific candidate PRSs across all tested hyper-parameter values. The best-performing PRS-CSx[25] candidate PRS, based both on a mean AUC of 0.5663 and a mean OR per SD of 1.27, was generated with the hyper-parameter ϕ=1x10^{-4} and incorporated genetic information from European, East Asian, and admixed-American ancestries.

7.3.1.2 PRS testing results

Four MCPS-informed candidate PRSs were selected for testing: one from the P+T method[23], one from PRS-CSx[25], and two from LDpred2[24], each representing the best-performing hyper-parameter configurations identified during training. All four candidate PRSs associated positively and log-linearly with the odds of CHD when applied to the testing samples (20% of MCPS data) that were withheld during PRS training. The besting-performing PRS, in terms of both AUC and OR per SD, was the PRS-CSx[25] PRS, which included 1,10,546 SNPs and incorporated genetic information of three ancestries. When evaluated in the testing set using logistic regression with no other covariate adjustments, this PRS showed a strong association with CHD risk, with an

OR per SD of 1.34 (95% CI, 1.26-1.42). The strength of the association remained consistent after adjusting for age, sex and the first seven genetic PCs, and further after adjustment for conventional CHD risk factors. Without other covariate adjustments, the PRS generated using PRS-CSx[25] achieved marginally higher model AUC of 0.5848 (0.5663-0.6033) than candidate PRSs developed by other methods and hence was retained as the novel MCPS-informed PRS. This MCPS-informed PRS also demonstrated comparable performance with the best-performing external PRS (Patel *et al.*[8]) evaluated in **Chapter 4** (OR per SD=1.34, 1.26-1.43; AUC=0.5793, 0.5610-0.5975) and had a slightly higher AUC. However, after further adjustment for conventional CHD risk factors, the improvement in model AUC by this novel MCPS-informed PRS was only observable at the third decimal place.

7.3.2 Findings in context

7.3.2.1 PRS construction methods comparison

PRS construction methods that originally developed for a single ancestry population (i.e., P+T and LDpred2[24]) do not account for population heterogeneity or differences in linkage disequilibrium (LD) patterns and allele frequencies across ancestries[25,28]. These limitations became evident when these PRS construction methods were applied to multi-ancestry GWAS inputs. In the P+T method, using GWAS summary statistics that differ from the validation set (e.g., admixed-American individuals) can lead to inappropriate pruning of SNPs. This mismatch may cause over or under-pruning due to differing linkage disequilibrium (LD) patterns, thereby reducing PRS performance. Similarly, for the LDpred2[24] method, the algorithm exhibited a higher rate of non-convergence as more diverse ancestries were included in the input GWAS summary statistics. The discrepancy between the LD patterns in the validation set and those estimated from the input GWAS likely caused the Markov Chain Monte Carlo (MCMC) sampling of the method to fail to converge, due to conflicting signals between the two sources. As the input heritability h^2 increased, more SNPs were assumed to have non-zero effect sizes, the conflicts intensified, leading to a higher rate of non-convergence. However, for both methods, PRSs de-

rived from multi-ancestry GWAS still outperformed those based on single-ancestry data, likely due to greater statistical power, which may help refine effect size estimates of SNPs that are rare in admixed-American populations.

PRS-CSx[25] is a PRS construction method designed for multi-ancestry applications. The method is specifically built to incorporate genetic information from diverse ancestries using ancestry-specific LD reference for each ancestry-specific GWAS, thereby eliminating the issues of ancestry mismatch. For underpowered ancestry-specific GWAS, the method can borrow information from well-powered GWAS (e.g., those based on European populations). In this thesis, multi-ancestry PRSs generated using this method were shown to outperform PRSs derived from single-ancestry PRS methods, consistent with previous evidence[29,30]. Moreover, the ancestry-specific PRS output by this method showed that with adequate power (i.e., larger sample size), PRSs derived using admixed-American GWAS would offer the highest accuracy in CHD risk prediction in MCPS.

7.3.2.2 Comparing the novel MCPS-informed PRS with external CHD PRS

The novel MCPS-informed PRS demonstrated comparable performance to the Patel *et al.* PRS[8], which demonstrated the strongest performance for CHD prediction in **Chapter 4**. However, the Patel *et al.* PRS[8] was constructed using a complex two-stage process that required not only CHD GWAS from different ancestries but also GWAS of CHD risk factors. This method may have limited reproducibility to other complex diseases, for which GWAS of key risk factors are scarce. This limitation could potentially reduce the predictive performance of PRSs generated using the Patel *et al.*[8] method. Moreover, this method favours well-powered GWAS and may exclude ancestry-specific GWAS that are informative but significantly underpowered, limiting the generalisability of the resulting PRS.

In contrast, the MCPS-informed PRS achieved similar performance with only CHD GWAS, supported by the inclusion of eight times more individuals of admixed-American ancestry, making it more representative of this population. Moreover, the PRS-CSx[25] method does not discard un-

derpowered ancestry-specific GWAS, enabling a more equitable and comprehensive integration of diverse genetic information. Additionally, the hyper-parameter tuning approach used here is more reproducible and adaptable to other complex diseases for future research.

This work represents a significant advancement over previous genetic studies on CHD in admixed-American populations, all of which relied on data from the MVP, a cohort composed of over 90% men and generally healthier individuals, given that all participants are military veterans[7,8,20]. The novel MCPS-informed PRS developed in this thesis used a more sex balanced and regionally relevant cohort that better represents the general population. The findings in **Chapter 5** contribute to the growing efforts to develop multi-ancestry CHD PRSs that are more representative of admixed-American ancestry to improve CHD risk prediction in this population. Moreover, with the availability of more prospective cohorts of admixed-American ancestry providing greater statistical power, it may become feasible to develop a CHD PRS based solely on data from this population in the future.

7.4 Clinical and genetically-integrated risk scores and risk for CHD in Mexicans

7.4.1 Summary of the main findings

In both the analyses of **Chapter 4 and 5**, it was evident that PRSs and conventional CHD risk factors are independent of each other and, therefore, have a multiplicative impact on CHD risk prediction (i.e., their odds ratios multiply together). (Alternatively, their log odds ratios add together.) The final analyses presented in this thesis attempted to clarify the added relevance of genetic information to commonly available routine clinical tools for CHD risk assessment. Three guideline-recommended clinical risk scores for CHD risk assessment (SCORE2-standard[1], SCORE2-diabetes-recalibrated and PCE[3]) in adults from the general population, were evaluated on their own as well as when integrated with the Patel *et al.*[8] PRS to assess improvement in CHD risk classification in Mexican adults from MCPS. The three clinical risk scores were all strongly positively associated with CHD risk. When treated as standardised continuous variables, all three scores

showed similar strengths of association with the odds of CHD in unadjusted logistic regression models, with the OR per SD ranging from 1.57 to 1.62. Clinical scores calibrated for diabetes (SCORE2-diabetes-recalibrated and PCE) achieved higher model discrimination with AUC as 0.73 compared to the SCORE2-standard (AUC=0.71), which was not calibrated for diabetes.

The clinical risk scores largely captured the impact of sex on CHD risk, showing by similar observed ORs for men and women in the sex-specific analysis. However, the impact of diabetes was less well-captured, the participants with diabetes exhibited significantly higher ORs than those without diabetes. When assessing genetic predisposition to CHD risk (as inferred by CHD PRS) stratified by levels of observed clinical risk measured, a significantly stronger genetic effect was observed among individuals with higher clinical risk.

7.4.1.1 The relevance of the integrated genetically-enhanced clinical risk score to CHD risk in Mexicans

When each clinical score was integrated with the Patel *et al.*[8] PRS, the combined score improved risk estimation for individuals in the top 20% risk stratum compared to those in the lowest group. Although each integrated score demonstrated a similar association strength compared to the clinical scores that they derived from, it marginally improved model AUC. In terms of risk reclassification of individuals with different underlying risk-profiles, both SCORE2-diabetes-recalibrated x PRS and SCORE-standard x PRS integrated scores showed positive and significant categorical Net Reclassification Index (NRI) relative to the original clinical score they derived from, as 1.50 (95%CI, 0.73-2.27) and 2.27 (1.43-3.11), respectively, when adopting a 7.5% risk threshold. In contrast, the PCE integrated score showed a non-significant categorical NRI of 0.19 (95%CI, -0.45-0.84). On the continuous scale, without the 7.5% risk threshold, all three integrated scores demonstrated strongly positive continuous NRIs and significant integrated discrimination indexes (IDIs), indicating an overall improvement in the ability to differentiate individuals with and without CHD.

The analyses were then repeated by integrating the novel MCPS-informed PRS developed in **Chapter 5** with the clinical risk scores (as opposed to external pre-existing PRSs) using the 20% of MCPS data reserved for PRS testing, and the findings were consistent. For risk reclassification, the SCORE2 x MCPS-PRS integrated scores demonstrated stronger categorical and continuous NRI values when compared to SCORE2 alone, although with wider confidence intervals. The categorical NRI was 2.90 (95% CI, 0.95-4.86) for SCORE2-diabetes calibrated integrated score, 3.49 (95% CI, 1.40-5.57) for SCORE2-standard integrated score and 1.59 (-0.04-3.23) for PCE integrated score. Similarly, the continuous NRIs and IDIs were strongly significant for all three MCPS-PRS integrated scores.

7.4.2 Findings in context

7.4.2.1 Performance of non-admixed-American derived clinical risk score in MCPS

All three evaluated clinical risk scores demonstrated strong CHD risk prediction performance in MCPS. However, low sensitivity for CHD cases was observed for both SCORE2 variations and PCE. Over half of the CHD cases were misclassified as controls by SCORE2-standard, while 35-40% of the cases were misclassified by diabetes calibrated SCORE2 and PCE. This misclassification may be partially attributed to the fact that the score was not calibrated for admixed-American population and the lack of incidence CHD data. Previous studies have reported that SCORE2 recalibrated for populations of Asian ancestry correctly identified more cases than the original uncalibrated version[31].

7.4.2.2 Combining genetic and clinical risk factors to enhance CHD risk prediction

Currently, there has been no standard way of integrating clinical risk scores with PRSs. However, studies on integrated scores all reported positive reclassification results when comparing their integrated score with clinical risk score[14,18]. This study further confirmed the utility of integrated score in CHD risk prediction among admixed-Americans, with NRI being 1.5 when integrating diabetes-recalibrated SCORE2 and the Patel *et al.*[8] PRS. This highlights that combining genetic

and clinical information (obtained from SCORE2) together could lead to more accurate CHD classification of individuals. This improvement in CHD case classification increased to an NRI of 2.9 when using MCPS-informed PRS in the integrated score, although with a wider confidence interval due to smaller evaluation sample size. The finding reinforces the need for improved admixed-American representation during PRS construction. With this, and a clinical risk score that is well calibrated for the admixed-American population, the improvement in CHD risk classification could be even greater.

The score integration approach that this thesis adopted ensures the genetic predisposition to CHD is combined with a clinical risk score using an algorithm that is easily implementable in clinical practice, and could be used in a similar fashion to the guideline-approved tools clinicians are practiced at and familiar with, while still preserving the absolute risk feature that clinical risk score have. In this way, the integrated score remains interpretable to patients in clinical settings. A recent study on the feasibility of clinical risk scores in clinics showed that the use of a CVD integrated score received a high acceptance rate and significantly changed the CVD prevention strategies for over 2% of the participants, highlighting the applicability of a genetically-enhanced integrated score in clinical settings.

7.4.2.3 Estimated impact at the national level in Mexico

Based on the total population of Mexico in 2020 (126,014,024) and around 40% between ages 35-79[32], around 50 million people fall within the CHD target age group. Assuming a CHD prevalence of 3.5% (based on the age-standardised CHD prevalence reported previously[33]) and using the NRI estimated for SCORE2-diabetes recalibrated x Patel *et al.*[8] PRS, which is the integrated score with the lowest misclassified rate for controls, the integrated score would correctly reclassify approximately 30,000 CHD cases compared with the clinical risk score alone based on the case NRI of 1.67%. Therefore, with a well-calibrated clinical risk score and a CHD PRS that better reflects admixed-American genetic information, an integrated score has the potential of becoming

a practical tool for the admixed-American population in CHD prevention in the future.

Although these improvements were modest in absolute terms, they are clinically meaningful. For public health at a population level, the integrated score may not change the overall demand drastically as while some people will be newly identified as higher risk, others will shift to lower risk. The planning of preventive and treatment service would balance out on this exchange. However, at the individual level, the impact would be significant. People who are at much higher 'true' risk (but underestimated by clinical risk scores alone) may now be identified more accurately, allowing for timely and potentially life-saving interventions and reducing risk of fatal CHD in the future. In addition, technology advancements have significantly reduced the cost of genotyping, eliminating it as a barrier to implementing such integrated scores in clinical settings.

7.5 Strengths and limitations

7.5.1 Strengths

The first major strength of this study is that it was conducted using one of the largest blood-based prospective cohorts of admixed-American ancestry, a population that has been understudied in previous genetic studies. MCPS contains more than four times as many individuals of admixed-American ancestry as the largest previous PRS evaluation studies[7,8], and 20 years of follow-up for CHD (or other fatal events) enabling well-powered analysis to be carried out when evaluating previously-published CHD PRS. In addition, as mentioned in **Chapter 5**, by combining the genetic information of this large sample with other external admixed-American CHD GWAS meta-analyses, this thesis constructed a novel MCPS-informed PRS with greater representation of admixed-American ancestry, involving eight times more individuals than the next most admixed-American representative PRS[8]. The genetic data collected by MCPS are of high quality and are well-imputed[34], allowing external CHD PRSs to be recreated for MCPS participants with high variant matching rates. This also enables the novel CHD PRS constructed using MCPS genetic data to be easily reproduced and externally validated by other independent cohorts in the

future. Another strength is the careful selection of eight PRSs for evaluation in **Chapter 4**. As explained in **Chapter 2**, the selection process tried to cover a wide range of PRSs with different construction methodology and ancestry involvement in order to evaluate the factors contribute to the better performance of PRSs in admixed-American population. This step provided important guidelines for subsequent PRS construction in **Chapter 5**.

The third and most novel aspect of the thesis is the well-structured approach used to construct the novel MCPS-informed PRS in **Chapter 5**. Firstly, a ten-fold cross-validation (CV) strategy for PRS hyper-parameter tuning was used to avoid overfitting and reduce bias that is likely caused by a single split. Moreover, ten-fold CV allows all the samples in the training data set to be used both for training and evaluating PRS, resulting in no data waste. Secondly, the high degree of relatedness among MCPS participants was most carefully accounted for by using the RE-GENIE[35] platform for conducting GWAS, a method that has shown efficiency for samples with high relatedness and population structure. Thirdly, three PRS construction methods (P+T, LD-pred2[24]and PRS-CSx[25]) were systematically assessed and compared for their performance, via ten-fold CV. Each method was tuned on a wide range of hyper-parameter sets and also GWAS ancestry combinations, with the optimal set taken forward to the testing stage. This strategy ensured the novel MCPS-informed CHD PRS is robust to overfitting while being optimised for CHD predictability. Moreover, the approach is reproducible and can be extended to PRS development for other complex diseases.

The final strength of the thesis is the method integrating a CHD PRS and a clinical risk score described in **Chapter 6**. A CHD PRS is a relative risk and only interpretable when comparing it to a reference population. The integrated score retains the absolute risk feature (0-100) as a clinical risk score, allowing a straightforward interpretation in a clinical setting and enabling clinicians to make threshold-related treatment decisions.

7.5.2 Limitations

This thesis has several limitations that need to be acknowledged. The first limitation is the lack of incident CHD cases in MCPS. In addition, although the study has followed its participants for over 20 years, accruing a reasonable number of CHD mortality cases, the overall CHD case rate remained relatively low (3.9%) comparing to other PRS studies evaluated in **Chapter 4**, which ranged from 4.6% to 20%[7–9,11]. Both of these factors could result in attenuation of the estimated effect sizes of external PRSs on CHD when they are evaluated in this population, and may also affect PRS construction, as the full disease burden of the population was not captured. MCPS has recently conducted further health resurveys to collect non-fatal disease on a subset of participants. Although these data were not available for the current thesis, such information could improve statistical power and provide a more comprehensive assessment of CHD risk in future studies.

Another limitation of the thesis is that participants of MCPS were recruited from only two regions of Mexico City, and do not fully reflect the genetic diversity of Mexico or Latin America more widely. However, previous studies have shown that non-representative prospective cohorts can still provide reliable evidence about the association between the disease and its risk factors that are generalisable[36]. Therefore, the findings from **Chapter 4** are likely to be at least broadly generalisable to the population of Mexico and also individuals of admixed-American ancestry. Moreover, genetic information of admixed-Americans in the US and populations of European and East-Asian ancestries were leveraged during PRS construction in **Chapter 5**, this multi-ancestry approach further improving the generalisability of the novel MCPS-informed PRS developed in this thesis.

The third limitation is the absence of an external admixed-American cohort to validate the novel MCPS-informed PRS in an independent population. Therefore, assessing the external validity of this PRS is currently not possible. Future validation will be essential once other admixed-

American cohorts of sufficient sample size become available.

Although other CVD risk scores, such as the WHO CVD risk chart[37], are available, they were not included in this thesis for analysis. This is because SCORE2 and PCE are more widely adopted and therefore more appropriate for comparative evaluation, and the majority of variables they include are shared across most CVD risk scores. Nonetheless, the exclusion of less commonly used scores may limit the generalisability of findings to settings where alternative clinical risk models are preferred.

The final limitation is that the NMR-based high density lipoprotein cholestrol (HDL-C) and total cholesterol measurements were relatively low in many cases. Around 17% of participants had a value lower than the lower bound sets by PCE. A lower HDL-C increases risk estimated by SCORE2[1,2] and PCE[3] while a lower total cholesterol reduces the risk estimated by both clinical risk scores. In addition, due to the use of manual sphygmomanometers at recruitment, baseline blood pressure exhibited strong digit preference, with values often ending in 5 or 0[38]. HDL-C, total cholesterol and systolic blood pressure were all inputs for calculating SCORE2[1,2] and PCE[3]. The low NMR measurements values and rounding issue in blood pressure would reduce the precision of risk estimated by these clinical risk scores.

7.6 Implication for public health

The research outcome in **Chapter 6** has major implications for public health. In recent years, Mexico has adopted several strategies for CVD prevention (CHD is the most common subtype of CVD), such as the HEARTS initiative promoted by the World Health Organisation (WHO)[39,40], which focuses on controlling blood pressure, smoking habits, obesity and screening for diabetes. However, as with most national guidelines on CHD prevention of other countries, Mexican public policies have not incorporated clinically-relevant genetic information into CVD or CHD risk assessment in clinics. As mentioned in **Chapter 6**, the use of an integrated score could modestly but significantly improve risk classification risk compared to clinical score alone by 1.5% for

individuals in MCPS (with previous studies among European populations also reporting similar findings)[18,41]. Extrapolating these estimates to the whole of Mexico suggests that approximately 30,000 more CHD cases could be identified among adults age 35 to 79. This corresponds to approximately one additional CHD case correctly detected for every 60 individuals who carry CHD within this age range in Mexico but may not yet be identified. In Mexico, many individuals still have to spend out-of-pocket health payments[42,43], hence the early and more accurate identification of high-risk individuals becomes even more critical. These findings highlight the public health values of including such integrated risk tool in national CHD screening process to improve early detection of CHD cases and hence prevention.

7.7 Future research

The thesis also underscores several areas that warrant further research. Firstly, sex stratified analysis was performed in **Chapter 4** and significant sex heterogeneity in genetic predisposition to CHD was observed. The relative risks associated with PRSs in women were significantly lower than those in men, likely reflecting inherent sex-biases in genetic research[13,44]. Therefore, more efforts should be made to improve the representation of women in CHD genetic research (i.e., GWAS and PRS studies) to allow accurate risk estimation in both sexes.

Secondly, as shown in **Chapter 5**, most advanced PRS construction tools were developed from large European-only cohorts and hence only aimed for a population of single ancestry[24,45]. For non-European cohorts, the majority have smaller sample sizes compared to those of European populations. Therefore, those non-European studies often combine their genetic information with European GWAS to boost their sample size to obtain more precise SNP effect size estimates for PRS construction. Although a few emerging methods (e.g., PolyPred[46], PROSPER[47], SBayesRC[48]) also address genetic prediction across multiple ancestries, PRS-CSx[25] is currently the only method explicitly designed for multi-ancestry PRS, as it jointly models GWAS summary statistics and ancestry-specific LD. In contrast, single-ancestry PRS methods often encounter

issues when applied directly to multi-ancestry GWAS results. Future research on PRS construction methodology should account for the need of multi-ancestry PRS to improve disease risk prediction in understudied populations.

In addition, although MCPS is one of the largest admixed-American focused prospective studies, it is smaller than major cohorts of other ancestries such as UK Biobank[49] and China Kadoorie Biobank[50]. More effort should therefore be directed to building larger admixed-American focused prospective cohorts (or meta-analyses of cohorts across Latin America) to further improve statistical power of genetic research for such ancestries and, ultimately, contributing to the construction of an admixed-American-only CHD PRS.

Finally, as mentioned in **Chapter 6**, low sensitivity for CHD cases were observed for both SCORE2[1] and PCE[3]. This highlights the need for a finely tuned, admixed-American-specific CHD risk predictor that better captures clinical risk in this population. Moreover, to assess the clinical utility of the genetically-enhanced integrated score, future research could consider piloting its use in a small cohort to evaluate its effectiveness, ease of implementation, and suitability in real-world settings. Further research is also needed to determine the optimal way of combining conventional and genetic risk factors of CHD to maximise the predictive potential of both risk scores in CHD risk assessment.

7.8 Conclusion

In this large population of admixed-Americans, previously-published PRSs demonstrated reasonable transferability and were independent from conventional risk factors in predicting CHD risk. Sex heterogeneity was observed for six out of the eight evaluated PRSs. Multi-ancestry PRSs generally outperformed European-only PRSs in terms of association strengths.

A novel MCPS-informed PRS was constructed incorporating genetic information across admixed-American, European and East-Asian ancestries. This novel PRS included 1,10,546 SNPs and was constructed using PRS-CSx method[25]. During internal testing, the novel PRS demonstrated

comparable performance with the best-performing external PRS. This PRS was constructed using only CHD GWAS and included the greatest number of admixed-Americans to date.

Clinical risk scores were assessed for its performance in MCPS alone and when integrated with CHD PRS. All three evaluated clinical risk scores that did not calibrate for admixed-American population underestimated CHD risk in this cohort. When combining clinical risk score with PRS, the integrated scores only improved model discrimination slightly comparing to clinical risk scores alone. However, a significant improvement in risk classification was seen with for SCORE2[1] integrated score. When extrapolating the estimates to Mexico, such integrated score could reclassify correctly around 30,000 more CHD cases.

In summary, the findings from this thesis demonstrated the potential of CHD PRSs to enhance CHD risk prediction in a Mexican population when used alongside established clinical risk factors, while also highlighting the importance of improving representation of admixed-American individuals in PRS development and the opportunities and challenges of integrating genetic risk into CHD prevention.

7.9 References

1. SCORE2 working group ESC Cardiovascular risk collaboration. “SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe”. *European Heart Journal* 2021;42(25):pp. 2439–2454.
2. SCORE2-OP working group ESC Cardiovascular risk collaboration . “SCORE2-OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions”. *European Heart Journal* 2021;42(25):pp. 2455–2467.
3. Goff D. C. J, Lloyd-Jones DM, Bennett G, Coady S, D’Agostino R. B. S, Gibbons R, et al. “2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines”. *J Am Coll Cardiol* 2014;63(25 Pt B):pp. 2935–2959.
4. Tada H, Melander O, Louie JZ, Catanese JJ, Rowland CM, Devlin JJ, et al. “Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history”. *Eur Heart J* 2016;37(6):pp. 561–7.
5. Oni-Orisan A, Haldar T, Cayabyab MAS, Ranatunga DK, Hoffmann TJ, Iribarren C, et al. “Polygenic Risk Score and Statin Relative Risk Reduction for Primary Prevention of Myocardial Infarction in a Real-World Population”. *Clinical Pharmacology & Therapeutics* 2022;112(5):pp. 1070–1078.
6. Tamlander M, Mars N, Pirinen M, FinnGen, Widen E, and Ripatti S. “Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes”. *Communications Biology* 2022;5(1):p. 158.
7. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. “Large-scale genome-wide association study of coronary artery disease in genetically diverse populations”. *Nature Medicine* 2022;28(8):pp. 1679–1692.
8. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. *Nature Medicine* 2023;29(7):pp. 1793–1803.
9. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, et al. “Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease”. *Nature Genetics* 2020;52(11):pp. 1169–1177.
10. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. *Nat Genet* 2018;50(9):pp. 1219–1224.
11. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. “Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention”. *J Am Coll Cardiol* 2018;72(16):pp. 1883–1893.
12. Huang Y, Hui Q, Gwinn M, Hu YJ, Quyyumi AA, Vaccarino V, et al. “Sexual Differences in Genetic Predisposition of Coronary Artery Disease”. *Circulation. Genomic and Precision Medicine* 2021;14(1):e003147.
13. Byars SG and Inouye M. “Genome-Wide Association Studies and Risk Scores for Coronary Artery Disease: Sex Biases”. *Adv Exp Med Biol* 2018;1065:pp. 627–642.
14. Li L, Pang S, Starnecker F, Mueller-Myhsok B, and Schunkert H. “Integration of a polygenic score into guideline-recommended prediction of cardiovascular disease”. *Eur Heart J* 2024.
15. Lambert SA, Abraham G, and Inouye M. “Towards clinical utility of polygenic risk scores”. *Hum Mol Genet* 2019;28(R2):R133–r142.
16. Xiang R, Kelemen M, Xu Y, Harris LW, Parkinson H, Inouye M, et al. “Recent advances in polygenic scores: translation, equitability, methods and FAIR tools”. *Genome Medicine* 2024;16(1):p. 33.
17. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, et al. “Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses”. *PLoS Medicine* 2021;18(1) (no pagination).

18. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, et al. “Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction”. *Circ Genom Precis Med* 2021;14(2):e003304.

19. de La Harpe R, Thorball CW, Redin C, Fournier S, Müller O, Strambo D, et al. “Combining European and U.S. risk prediction models with polygenic risk scores to refine cardiovascular prevention: the CoLaus|PsyCoLaus Study”. *Eur J Prev Cardiol* 2023;30(7):pp. 561–571.

20. Vassy JL, Posner DC, Ho YL, Gagnon DR, Galloway A, Tanukonda V, et al. “Cardiovascular Disease Risk Assessment Using Traditional Risk Factors and Polygenic Risk Scores in the Million Veteran Program”. *JAMA Cardiol* 2023;8(6):pp. 564–574.

21. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, et al. “Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups”. *American Journal of Human Genetics* 2020;106(5):pp. 707–716.

22. Aragam KG, Dobbyn A, Judy R, Chaffin M, Chaudhary K, Hindy G, et al. “Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease”. *Journal of the American College of Cardiology* 2020;75(22):pp. 2769–2780.

23. Choi SW and O’Reilly PF. “PRSice-2: Polygenic Risk Score software for biobank-scale data”. *GigaScience* 2019;8(7).

24. Privé F, Arbel J, and Vilhjálmsson BJ. “LDpred2: better, faster, stronger”. *Bioinformatics* 2021;36(22-23):pp. 5424–5431.

25. Ruan Y, Lin Y.-F, Feng Y.-CA, Chen C.-Y, Lam M, Guo Z, et al. “Improving polygenic prediction in ancestrally diverse populations”. *Nature Genetics* 2022;54(5):pp. 573–580.

26. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. “Million Veteran Program: A mega-biobank to study genetic influences on health and disease”. *J Clin Epidemiol* 2016;70:pp. 214–23.

27. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores”. *The American Journal of Human Genetics* 2015;97(4):pp. 576–592.

28. Lewis CM and Vassos E. “Polygenic risk scores: from research tools to clinical instruments”. *Genome Medicine* 2020;12(1):p. 44.

29. Gunn S, Wang X, Posner DC, Cho K, Huffman JE, Gaziano M, et al. “Comparison of methods for building polygenic scores for diverse populations”. *Human Genetics and Genomics Advances* 2025;6(1).

30. Ge T, Irvin MR, Patki A, Srinivasasainagendra V, Lin Y.-F, Tiwari HK, et al. “Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations”. *Genome Medicine* 2022;14(1):p. 70.

31. SCORE2 Asia-Pacific writing group, Hageman SHJ, Huang Z, Lee H, Kaptoge S, Dorresteijn JAN, et al. “Risk prediction of cardiovascular disease in the Asia-Pacific region: the SCORE2 Asia-Pacific model”. *European Heart Journal* 2024;46(8):pp. 702–715.

32. National Institute of Statistics and Geography. *Demographic and Social Information*. `https://en.www.inegi.org.mx/programas/ccpv/2020/`. 2024.

33. Stark B, Johnson C, and Roth GA. “Global prevalence of coronary artery disease: an update from the global burden of disease study”. *Journal of the American College of Cardiology* 2024;83(13_Supplement):pp. 2320–2320.

34. Ziyatdinov A, Torres J, Alegre-Díaz J, Backman J, Mbatchou J, Turner M, et al. “Genotyping, sequencing and analysis of 140,000 adults from Mexico City”. *Nature* 2023;622(7984):pp. 784–793.

35. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. *Nature Genetics* 2021;53(7):pp. 1097–1103.

36. Rothman KJ, Gallacher JE, and Hatch EE. “Why representativeness should be avoided”. *Int J Epidemiol* 2013;42(4):pp. 1012–4.

37. Kaptoge S, Pennells L, De Bacquer D, Cooney MT, Kavousi M, Stevens G, et al. “World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions”. *The Lancet Global Health* 2019;7(10):e1332–e1345.

38. Tapia-Conyer R, Kuri-Morales P, Alegre-Díaz J, Whitlock G, Emberson J, Clark S, et al. “Cohort profile: the Mexico City Prospective Study”. *Int J Epidemiol* 2006;35(2):pp. 243–9.

39. Pan American Health Organization. “HEARTS in the Americas: Evaluation framework for continuous quality improvement in primary care centers” 2025:62 p.

40. Ordunez P, Campbell NRC, Giraldo Arcila GP, Angell SY, Lombardi C, Brettler JW, et al. “HEARTS in the Americas: innovations for improving hypertension and cardiovascular disease risk management in primary care”. *Rev Panam Salud Publica* 2022;46:e96.

41. Ramírez J, Duijvenboden S van, Young WJ, Tinker A, Lambiase PD, Orini M, et al. “Prediction of Coronary Artery Disease and Major Adverse Cardiovascular Events Using Clinical and Genetic Risk Scores for Cardiovascular Risk Factors”. *Circ Genom Precis Med* 2022;15(5):e003441.

42. Meneses Navarro S, Pelcastre-Villafuerte BE, Becerril-Montekio V, and Serván-Mori E. “Overcoming the health systems’ segmentation to achieve universal health coverage in Mexico”. *The International Journal of Health Planning and Management* 2022;37(6):pp. 3357–3364.

43. Garcia-Diaz R. “Effective access to health care in Mexico”. *BMC Health Services Research* 2022;22(1):p. 1027.

44. Sakkers TR, Mokry M, Civelek M, Erdmann J, Pasterkamp G, Diez Benavente E, et al. “Sex differences in the genetic and molecular mechanisms of coronary artery disease”. *Atherosclerosis* 2023;384:p. 117279.

45. Mak TSH, Porsch RM, Choi SW, Zhou X, and Sham PC. “Polygenic scores via penalized regression on summary statistics”. *Genetic Epidemiology* 2017;41(6):pp. 469–480.

46. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera AV, et al. “Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores”. *Nat Genet* 2022;54(4):pp. 450–458.

47. Zhang J, Zhan J, Jin J, Ma C, Zhao R, O’Connell J, et al. “An ensemble penalized regression method for multi-ancestry polygenic risk prediction”. *Nature Communications* 2024;15(1):p. 3238.

48. Zheng Z, Liu S, Sidorenko J, Wang Y, Lin T, Yengo L, et al. “Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries”. *Nature Genetics* 2024;56(5):pp. 767–777.

49. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. “The UK Biobank resource with deep phenotyping and genomic data”. *Nature* 2018;562(7726):pp. 203–209.

50. Walters RG, Millwood IY, Lin K, Schmidt Valle D, McDonnell P, Hacker A, et al. “Genotyping and population characteristics of the China Kadoorie Biobank”. *Cell Genom* 2023;3(8):p. 100361.

Appendix

Table S1: Identified papers that constructed a new PRS for CHD risk.

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
UK Biobank **,(Inouye et al.,2018[1])**	Mainly European	22242 / 482629 , (4.61 %)	mean 6.2	NA , NA	Incident CHD (Including prevalent (n=9729) case at baseline) Outcome definition: fatal or non-fatal MI, PTCA, CABG	A meta-score based on the weighted average of 3 different standardised PRSs that created using CC4D GWAS summary statistics	Nikpay et al., Deloukas et al	Logistic, Cox	Logistic :sex, age, genotyping array, 10 PCs Cox: 1) sex, genotyping array, 10 PCs 2)+ smoking, diabetes, family history of heart disease, BMI, hypertension, high cholesterol (only among incident cases)	HR per SD -- all CHD cases: 1.71 (1.68-1.73) incident CHD model 1: 1.58 (1.55 to 1.61), model 2: 1.48(1.45 to 1.51), Q5 vs Q1 : 4.17 (3.97- 4.38)	AUC-- Logistic: 0.79 (+2.8% over model without the score) Cox: incident model 1: 0.623 (0.615-0.630), model 2: 0.696 (0.688-0.703)
UK Biobank **,(Tikkanen et al.,2018[2])**	Mainly European	8518 / 468095 , (1.82 %)	median 6.1	NA , NA	Incident CHD Outcome definition: MI, PTGA and CABG	Used coefficients of 62 SNPs as weights from CC4D GWAS	Nikpay et al	Cox	1.Age, sex, ethnicity, genotype array,10 PCs, UK assessment centre 2.+ diabetes, smoking, SBP, BMI, lipid medication, IPAQ , grip strength	1.HR: High tertile vs low: 1.77(1.67-1.87), high 5% vs bottom 5%: 2.74 (2.38-3.17) 2.HR: High tertile vs low: 1.73 (1.64-1.83), high 5% vs bottom 5%: 2.67 (2.31-3.08)	
UK Biobank **,(Ntalla et al.,2019[3])**	Mainly European	21051 / 425196 , (4.95 %)	prevalent	NA , NA	Prevalent CHD Outcome definition: MI and its complications, Coronary failure/insufficiency, PTCA, CABG, triple heart bypass	Used coefficients of 300 SNPs from UKBB GWAS (5% FDR) as weights	Nelson et al	Logistic	age, sex, 5 PCs, genotyping array	OR per SD: 1.65(1.62-1.65), Q5 vs Q2-4 OR: 2.05 (1.99-2.12), Q1 vs Q2-4: 0.53 (0.51-0.56)	
UK Biobank **,(Patel et al.,2022[4])**	Mainly European	30455 / 408470 , (7.46 %)	prevalent	NA , NA	Prevalent CHD Outcome definition: Angina, MI and its complications, Coronary failure/ insufficiency, chronic IHD, PCI, CABG	Used coefficients of CC4D GWAS as weights; included 182 significant SNPs derived from clumping	Nikpay et al	Logistic	age, sex	OR per SD:1.34 (1.32-1.35)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
UK Biobank ,**(Howe et al.,2020[5])**	Mainly European	2848 / 393108 , (0.72 %)	NA	NA , NA	Incident MI and CHD death (n=921) Outcome definition: MI , CHD death(angina, MI and its complications, Coronary failure/ insufficiency, chronic IHD, sudden death)	Used coefficients of CC4D for 182 SNPs as weights, derived from clumping and also passed 5e-6 significance threshold	Nikpay et al	Logistic	age, sex	OR per SD: MI: 1.34 (1.29-1.38), CHD death: 1.31 (1.23-1.40), CHD death /MI: 1.33 (1.29-1.38)	
UK Biobank ,**(Yuan et al.,2022[6])**	Mainly European	6732 / 384093 , (1.75 %)	median 11.91	NA , NA	Incident MI Outcome definition: MI and its complications	Used coefficients of CC4D as weights for 68 SNPs identified to be associated with CHD in previous studies	Nikpay et al	Cox (HDL/LDL groups and PRS interaction)	race, age, TDI, gender, smoking history, alcohol use, education level, BG, BMI, SBP, DBP, TG	HDL-C/LDL-C<0.4: intermediate PRS (Q2-4) vs Q1 HR: 1.29(1.18-1.41), high (Q5) vs Q1 HR: 1.79 (1.62-1.98)	
UK Biobank ,**(Fan et al.,2020[7])**	Mainly European	4381 / 357246 , (1.23 %)	median 8.5	NA , NA	Incident CHD Outcome definition: Angina, MI and its complications, Coronary failure/ insufficiency, chronic IHD	Included 74 genome-wide significant SNPs, using UKBB GWAS SNP coefficients as weights	Nelson et al	Cox (sleep pattern and PRS interaction)	age, sex, TDI, ethnicity, total physical activity level, smoking status, alcohol consumption, family history of heart diseases, BMI, prevalent hypertension, prevalent diabetes	Healthy sleep: intermediate PRS (Q2-4) vs Q1 HR:1.44 (1.27-1.63), high (Q5) vs Q1 OR: 2.13 (1.85-2.44)	
UK Biobank ,**(Zhang et al.,2021[8])**	Mainly European	Not available / 349462 , (NA %)	median 11.2	NA , NA	Incident CVD and CHD Outcome definition: CHD: MI and angina; CVD: CHD, cerebrovascular disease, emboli/thrombosis, heart failure, hypertensive	Used coefficients of CC4D as weights and LDPred2 method	Nikpay et al	Logistic (initial validation for CVD only), Cox	Logistic: None, Cox: Sex, age group, ethnicity, employment status, smoking status, BMI, alcohol intake, physical activity, vitamin and mineral supplement, tea intake, change in weight in the past	Logistic (OR per SD): 1.54 (1.50-1.59), Cox (for CVD): High PRS always had higher HR in all categories of food intake on incident CHD and CVD.	AUC -- Logistic: 0.54

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
					disease, arrhythmia/conduction disorder				year, family history of CVD		
UK Biobank **,(Huang, Y et al.,2021[9])**	Mainly European	20205 / 317509 , (6.36 %)	median 6.1	NA , NA	Incident and Prevalent CHD Outcome definition: MI and its complications, Coronary failure/ insufficiency, chronic IHD, replacement, CABG, PTCA	Used GWAS coefficients of CC4D as weights, constructed a PRS (CHD-GRS) using 161 CHD related SNPs and a genome-wide PRS using LDPred (CHD-LDPred)	Nikpay et al	Logistic, Cox (incident only, (9847/307151).	age, population stratification, smoking status, alcohol consumption, BMI, history of diabetes or hypertension, cholesterol medication use, education, TDI	OR per SD: CHD-GRS: 1.41 (1.38 -1.44), CHD-LDpred: 1.44 (1.41-1.47) HR per SD: CHD-GRS: 1.34 (1.31-1.36), CHD-LDPred: 1.36 (1.33-1.39)	
UK Biobank **,(Carter et al.,2022[10])**	Mainly European	14481 / 317055 , (4.57 %)	NA	NA , NA	Incident CHD Outcome definition: Angina, MI and its complications, Coronary failure/ insufficiency, chronic IHD	Using GWAS coefficients of CC4D as weights, constructed 3 PRSs based on 3 significance thresholds cut-off on the GWAS: genome-wide 5e-8 (main analysis), 0.05 and 0.5	Nikpay et al	Logistic	age, sex, 40PCs	OR per SD: genome-wide PRS: 1.22 (1.20-1.24), 0.05 PRS:1.23 (1.21-1.25), 0.5 PRS: 1.22 (1.20-1.24)	
UK Biobank **,(Daghlas et al.,2019[11])**	Mainly European	3513 / 310917 , (1.13 %)	median 7.04	NA , NA	Incident MI Outcome definition: fatal and non-fatal MI	Included 68 genome-wide significant SNPs for CHD from prior CHD GWAS as weights.	Nikpay et al , Deloukas et al., Howson et al., Webb et al., Stitziel et al	Cox	genotyping array, 10 PCs, age, sex, ethnicity, smoking status, frequency of alcohol consumption, history of heart disease in family, marital status, education, income, TDI, employment status, physical activity, television watching, grip	HR per SD: 1.31 (1,27-1.35); in quartile, Q4 vs Q1 HR:1.91(1.74 to 2.10)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
									strength, BMI, waist-hip ratio, history of seeing provider for mental health, snoring, use of sleep medications, self-reported or medical record derived sleep apnea, and self-reported insomnia, probable type 2 diabetes, hypertension, use of blood pressure lowering medication, history of high cholesterol, use of cholesterol lowering medication, use of aspirin		
1. FinnGen 2. UK Biobank (external validation) **,(Tamlander et al.,2022[12])**	Mainly European	33628 / 309154 , (10.88 %)	median 15.3	NA , 18698/343672	Prevalent and incident CHD Outcome definition: MI, Coronary angioplasty, CABG	Developed a genome-wide PRS using PRS-CS, using 1k genome European sample as LD reference panel and CC4D GWAS coefficients as weights	Nikpay et al	Logistic	year of birth, sex, 10 PCs, batch, genotyping array (FinnGen only)	OR per SD in FinnGen: prevalent: 1.59 (1.57-1.62), incident: 1.44 (1.41-1.47), incident + prevalent: 1.56 (1.53-1.58) OR per SD in UK Biobank: prevalent: 1.77 (1.73-1.80), incident: 1.61 (1.57-1.65), incident + prevalent: 1.72 (1.70-1.75)	AUC-- In FinnGen: prevalent: 0.869 (0.867–0.871), incident: 0.913 (0.911–0.916), incident + prevalent: 0.871 (0.869–0.873) In UK Biobank: prevalent: 0.811 (0.808–0.815), incident: 0.756 (0.751–0.761), incident + prevalent: 0.792 (0.789–0.795)
UK Biobank **,(Huang, Y et al. 2,2022[13])**	Mainly European	9847 / 307147 , (3.21 %)		NA , NA	Incident CHD Outcome definition: MI and its complications, Coronary failure/ insufficiency, chronic IHD, replacement, CABG, PTCA	Used 161 SNPs (identified by Van der Harst 2018) with CC4D GWAS coefficients as weights	Nikpay et al	Cox	age, sex, alcohol consumption, education, history of hypertension, history of diabetes, cholesterol lowering medication use, BMI, TDI, 10 PCs	HR per SD: 1.34 (1.31-1.36), Q2-4 vs Q1 HR: 1.44 (1.35-1.53), Q5 vs Q1: 2.20 (2.06-2.35)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
UK Biobank **,(Heianza et al.,2020[14])**	Mainly European	1162 / 152975 , (0.76 %)	max 5	NA , NA	Incident MI Outcome definition: MI: fatal or non-fatal MI	Included 89 SNPs based on previous studies and used their coefficients as weights	Schunkert et al., Nikpay et al., Clarke et al., Klarin et al., Howson et al., Webb et al	Cox	1. age, sex, ethnicity, 5PCs 2. age, sex, ethnicity, education, TDI, smoking habit, multivitamin use, total energy intake, alcohol consumption, physical activity 3. 2+BMI, hypertension, dyslipidemia, T2D	Low vs High (based on median PRS) for low adherence to healthful diet HR: model 1 : 0.81 (0.67-0.98), model 2:0.81 (0.67, 0.98), model 3:0.82 (0.68, 0.99)	
Biobank Japan **,(Koyama et al.,2020[15])**	South Asian (Japanese)	23003 / 151405 , (15.19 %)	prevalent, median 7.7 (death)	16823 (validation, in 10-fold CV)/10999 (testing) , NA	Prevalent CHD and CHD mortality (464/49230) Outcome definition: Stable or unstable angina, MI	The paper selected the best performing score based on R2 and AUC from 875 PRSs derived from different combinations of GWASs weighting (UK Biobank, CC4D and Biobank Japan) and methods (P+T, LDPred and metaGRS)	GWAS meta-analysis of Biobank Japan (conducted on the training set), Nikpay et al., Van der Harst et al	Logistic (in test cohort), Cox (CHD mortality)	Logistic: age, sex ; Cox: adjusted for sex, age, age2, 10 PCs	OR per SD: 1.840 (1.744–1.943), top 10%. vs rest:2.649 (2.295–3.046); HR per SD: 1.216 (1.109-1.333)	AUC -- Logistic:0.674 (0.661–0.687)
UK Biobank **,(Verweij et al.,2017[16])**	Mainly European	10898 / 143936 , (7.57 %)	prevalent	NA , NA	Prevalent CHD, CHD death (723) Outcome definition: MI and its complications, Coronary failure/ insufficiency, Chronic IHD, replacement, CABG, PTCA	Included 71 genome-wide significant SNPs associated with CHD, using CC4D GWAS coefficients as weights	Nikpay et al	Logistic (CHD), Cox (CHD mortality)	age, gender, 15 PCs, genotyping chip	Logistic (OR):2.21(2.11-2.32) 2. Cox (HR): 1.75 (1.48-2.08)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
FinnGen (subset FINRISK for downstream:1209/20165) **,(Mars et al.,2020[17])**	Mainly European	20179 / 135300 , (14.91 %)	max 46	NA , NA	Incident CHD Outcome definition: MI and its complications, Angina pectoris, Other coronary atherosclerosis, CABG, Coronary angioplasty	Used GWAS UKBB SAIGE coefficients as weights and LDPred to account for LD	Zhou et al	Cox (main), Logistic	survey collection year (in FINRISK), genotyping array, 10 PCs, sex	FinnGen (HR per SD): 1.31(1.29-1.33) FINRISK(OR per SD): incident: 1.31 (1.25-1.38), prevalent+incident: 1.38 (1.32-1.44), HR per SD: 1.27 (1.22-1.32)	AUC -- FINRISK model adjusted for sex, age: 0.832 (0.828-0.836), adjusted for ASCVD: 0.820 (0.816–0.824) NRI -- FINRISK model adjusted for ASCVD: 1.1 (-0.1-2.2)
UK Biobank **,(Khera et al.,2018[18])**	Mainly European	3963 / 120280 , (3.29 %)	prevalent	8676/288978 , NA	Prevalent CHD Outcome definition: MI or coronary revascularization (CABG or coronary angioplasty with or without stenting)	This paper selected the best performing PRS out of 31 different candidate PRSs (LDPred or P+T) based on AUC, using CC4D GWAS coefficients as weight	Nikpay et al	Logistic	age, sex, genotyping array, 4 PCs.	training (OR per SD): 1.72 (1.67-1.78); testing (OR): top 20% vs rest: 2.55 (2.43-2.67), top 10% vs rest: 2.89 (2.74-3.05),top 5% vs rest: 3.34 (3.12-3.58), top 1% vs rest: 4.83 (4.25-5.46), top 0.5% vs rest:5.17 (4.34-6.12)	AUC -- training:0.806 (0.800-0.813) testing: 0.81 (0.81-0.81)
UK Biobank, Million veteran program, Gene&Health (**Patel et al., 2023[19]**)	Multi-ancestry	4,412/112237(3.78%)	prevalent	2140/14293 (Hispanic)2917 1/95296(European)4831/2826 5(African), 853/16021(South Asian)	Prevalent CHD Outcome definition: angina, MI, coronary bypass or revascularisation (CABG, PTCA), coronary thrombolysis, chronic IHD, acute MI discharge	For each of 9 CHD trait, a multi-ancestry PRS was created by summing over ancestry specific PRS. The CHD PRS summed over all multi-ancestry PRS of all traits	Genes & Health, FinnGen, Million Veteran Program, Biobank Japan and CC4D excluding UK Biobank samples	Logistic	Age, sex, genotyping array, 10 genetic PCs	OR per SD: Hispanic: 1.61 (1.53-1.70),European: 1.72 (1.69-1.75),South Asian: 1.83 (1.69-1.99), African: 1.25 (1.21-1.29) Other PRS in Hispanic: Koyama et al: 1.49 (1.41-1.57), Wang,M et al 1.34 (1.27-1.41), Tamlander et al: 1.39 (1.32-1.46), Inouye et al: 1.38 (1.31-1.45), Khera et al: 1.33 (1.27-1.40), Tada et al: 1.15(1.09-1.20)	
UK Biobank **,(Ye et al.,2021[20])**	Mainly European	4746 / 92928 , (5.11 %)		3467/176238 , NA	Incident CHD Outcome definition: fatal and non-fatal MI, CABG, coronary angioplasty with	Constructed 130 PRSs using 4 different PRS methods and threshold, and	Nikpay et al	Cox	age, sex, 4 PCs, years of education, TDI, self-reported annual household income	testing (HR) with intermediate lifestyle and PRS 40-60% as reference: 60-80%: 1.25 (1.11-1.40), 80-90%: 1.54 (1.35-1.75), 90-	AUC -- AnnoPred: training: 0.643 (0.635-0.651) testing:0.643 (0.637–0.648)

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
					or without stenting	selected the best performing PRS based on AUC in the tuning dataset (Annopred)				95%: 1.98 (1.70-2.31), 95-99%: 2.27 (1.93-2.66), >99%: 4.23 (3.39–5.28)	
Million veteran program (MVP) ,(**Tcheandjieu et al.,2022[21])**	European, Hispanic (574/6314), and African American (1552/17072)	6158 / 67738 , (9.09 %)	prevalent	95151/292438(white), 17202/76709 (black), 6378/30648(hispanic) , NA	Incident CHD and Prevalent CHD Outcome definition: discharge diagnosis of acute MI, coronary revascularization , or any two of the encounter (acute or old MI, coronary syndrome, angina, chronic IHD)	4 existing PRS , 1 new PRS created using P+T on MVP GWAS meta-analysis coefficients, with randomly selected MVP participants as LD reference panel	GWAS meta-analysis (White) MVP+Van der Harst et al+Nikpay et al ; GWAS meta-analysis(Hispanic and Black) MVP +Biobank Japan)	Logistic	batch, age, sex, 10 genetic PCs	OR per SD: prevalent CHD, meta GRS: White: 1.24 (1.21-1.27), Black: 1.16 (1.10-1.23), Hispanic: 1.18 (1.07-1.30) population specific P+T PRS: White 1.35 (1.31-1.38), Black: 1.21 (1.15-.28), Hispanic: 1.43 (1.27-1.61)	
UK Biobank **,(Riveros-Mckay et al.,2021[22])**	Mainly European	NA / 60000 , (NA %)	max 10	incidence: 186451, prevalent+incidence: 248703 , NA	Incident CHD and Prevalent CHD Outcome definition: MI, coronary angioplasty or CABG	Applied LDpred firstly on causal variants (not only for CHD) posterior OR and then on to rest of the variants available in the genomic plc repository. The resulted PRS contains more than 3.5M SNPs	Nikpay et al., training set GWAS	Cox	none	HR per SD: Incident: : 1.62 (1.57-1.67); prevalent and incident: 1.73 (1.70-1.75)	Harell's C index -- Incident: 0.633 (0.625-0.641), incident and prevalent: 0.662 (0.658-0.665), outperformed Elliot et al., Khera et al and Inouye et al
1.ARIC 2. WGHS 3.MDCS, 4 Bioimage	Mainly European	5103 / 51425 , (9.92 %)	median 18.8,2	NA , NA	Incident CHD Outcome definition: MI, coronary	Used coefficients of 50 SNPs identified in	Deloukas et al , Erdmann et al.,	Cox, combined cohort results using	age, sex, self-reported education level, 5 PCs (not available in MDCS), initial trial	HR: Q5 vs Q1:1.91 (1.75 to 2.09), Q4 vs Q1: 1.50 (1.36-1.64), Q3 vs Q1:	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
study (used for sensitivity and small sample size, not included here) **,(Khera et al. 2,2016[23])**			0.5,19.4		revascularization, death from coronary causes	various previous studies as weights	Schunkert et al	RE meta-analysis	randomisation results (WGHS only)	1.30 (1.18-1.42), Q2 vs Q1:1.22 (1.11-1.34)	
GERA (white population) **,(Iribarren et al.,2016[24])**	Mainly European	1864 / 51294 , (3.63 %)	mean 5.9	NA , NA	Incident CHD Outcome definition: Unstable and stable angina, acute MI, coronary revascularisation procedure, death due to angina, MI or hypertensive heart disease	4 PRSs were constructed using coefficients from CC4D GWAS as weights, each contains 8 (GRS_8), 8 (GRS_12), 32 (GRS_36) and 47 variants (GRS_51)	Deloukas et al	Cox	model a: just GRS model b: age, sex, total cholesterol, high density lipoprotein cholesterol, SBO, DBP, smoking status, diabetes mellitus model c: + family history of heart disease model d: +BMI, antihypertensives, lipid-lowering drugs, alcohol consumption	HR per SD model a :GRS_8: 1.21(1.15-1.26), GRS_12:1.20(1.15-1.25), GRS_36:1.23(1.18-1.29), GRS_51:1.25(1.20-1.31) model d: GRS_8: 1.21 (1.15-1.26), GRS_12:1.20 (1.15-1.26), GRS_36:1.23 (1.17-1.28), GRS_51:1.23 (1.17-1.28)	Harell's C index -- model (FRS+GRS): GRS_8: 0.700, GRS_12:0.699, GRS_36:0.700, GRS_51:0.701
1.MDCS 2. JUPITER (RCT) 2.ASCOT **,(Mega et al.,2015[25])**	Mainly European	NA / 34926 , (NA %)		NA , NA	Incident CHD Outcome definition: in JUPITER, ASCOT: coronary heart death, MI, unstable angina, In MDCS: fatal or non-fatal MI, CABG, PCI	Included 27 SNPs that were significantly associated with CHD in previous GWAS, weighted by their coefficients	Schunkert et al	Cox, combined cohort results using RE meta-analysis	age, sex, diabetes status, smoking, race (if applicable), family history of CHD, HDL-C, LDL-C, hyertension.	HR per SD: 1.21 (95% CI 1.17-1.26); HR: Q2-4 vs Q1:1·31 (1.19-1.45), Q5 vs Q1: 1.72 (1.53–1.92)	
UK Biobank **,(Elliot et al.,2020[26])**	Mainly European	15947 / 31894 , (50 %)	testing: median 8	6272/352660 , NA	Incident CHD Outcome definition: MI, coronary angioplasty, CABG	Constructed PRSs using two different methods (P+T, lassosum) and CC4D GWAS coefficients as weights, and selected the best PRS based on AUC	Nikpay et al	Logistic (tuning), Cox (main)	Logistic: genotyping arrary, 10 PCs Cox: 1)None, 2)age, sex, 3)PCE	HR per SD (model 1): 1.32 (1.30-1.34)	AUC-- Tuning: 0.63 (0.62-0.64), Harell's C -- testing model 1): 0.61 (0.60-0.62), model 2)0.76 (0.75-0.76), model 3)0.78 (0.77-0.79) NRI -- testing model 3): 0.040 (0.031-0.049)

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
1. FINRISK,Health 2000, MDC-CC, MPP , COROGENE (no CHD cases) 2. FINRISK, Health 2000, MDC-CC **,(Ripatti et al.,2010[27])**	Mainly European	5093 / 30725 , (16.58 %)	median 10.7	NA , NA	Incident CHD (1264/26901) and Prevalent CHD, sensitivity: CVD and MI Outcome definition: CHD: MI, unstable angina pectoris, coronary revascularisation (CABG, PTCA); CVD: CHD and ischaemic stroke	Included 13 SNPs identified before 2009 to be associated with MI or CHD and used their GWAS coefficients as weights	Helgadottir et al., Samani et al., Kathiresan et al., Erdmann et al., Gudbjartsson et al., Clarke et al., McPherson et al	Cox (incident) and Logistic (prevalent), combine cohort results using FE meta-analysis	Logistic: age and sex Cox: sex, LDL-C, HDL-C, smoking, BMI, SBP, DBP, blood pressure treatment, diabetes	OR: Q5 vs Q1: 1.63 (1.24-2.15), Q4 vs Q1: 1.27 (0.95-1.70), Q3 vs Q1: 1.20 (0.90-1.59), Q2 vs Q1: 1.15 (0.86-1.53) HR:Q5 vs Q1: 1.66 (1.35-2.04), Q4 vs Q1: 1.39 (1.12-1.72), Q3 vs Q1: 1.17 (0.94-1.46), Q2 vs Q1: 1.00 (0.80–1.25)	AUC -- Cox: 0.872 NRI -- Cox: 0.022 (0.010-0.055)
FINRISK 1992, 1997, 2002, Health 2000 **,(Tikkanen et al. 2,2013[28])**	Mainly European	1093 / 24124 , (4.53 %)	median 12	NA , NA	Incident CHD Outcome definition: MI, unstable angina pectoris, coronary revascularization CABG or PTCA), death due to CHD	Included 28 SNPs identified previously to be associated with CHD or MI, and used their GWAS coefficients as weights	Kathiresan et al., Erdmann et al., Schunkert et al., Peden et al	Cox, combine cohort results using FE meta-analysis	sex, TC, HDL-C, BMI, SBP, blood pressure treatment, current smoking status, diabetes mellitus	HR per SD: 1.27 (1.20-1.35); HR vs middle 20%: top 20% :1.71 (1.42-2.06), top 10%: 2.07 (1.68-2.56), top 5%: 2.12 (1.62-2.77)	AUC -- 0.849, model (+family history and PRS vs model with only adjustments): 0.005 (0.004-0.006)
MDCS **,(Tada et al.,2016[29])**	Mainly European	2213 / 23595 , (9.38 %)	median 14.4	NA , NA	Incident CHD Outcome definition: fatal or nonfatal MI, CVD death, cardiovascular revascularizations (CABG, PCI)	Tested one PRS created by Mega et al (27 SNPs); created a new PRS with 23 SNPs identified in CC4D and C4D GWAS, using their coefficients as weights	Deloukas et al., Pedan et al	Cox	age, sex, SBP, hypertension treatment, smoking status, apoB, apoA-I, prevalent diabetes	HR per SD: GRS-27: 1.20 (1.15-1.25), GRS-50: 1.23(1.18-1.28)	AUC -- (based on events occurring in the first 10 years of follow-up)GRS-27: 0.748, GRS-50: 0.749
GERA - only included statin nonuser info in this table **,(Oni-Orisan et al.,2022[30])**	Multi-ethnic (White (1185/21824), Black (28/540),	1185 / 21284 , (5.57 %)	median 8.2	NA , NA	Prevalent MI Outcome definition: fatal or non-fatal MI	Used 164 SNPs identified as associated with CHD in a review, used	Erdmann review 2018	Cox	sex, age, hypertension, diabetes, smoking status	HR per SD: White: 1.59 (1.42-1.78) ; Black: 0.77 (0.32-1.85); East-Asian: 2.05 (95% CI, 1.27-3.31); Latinx:1.87 (1.19–2.95)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
	Latinx (71/1538)and East Asian (61/1489)					coefficients in reviewed GWASs as weights					
UK Biobank **,(King et al.,2022[31])**	Mainly European	9499 / 18998 , (50 %)	testing median 12.33	7036/272307 , NA	Prevalent CHD Outcome definition: MI, CABG, coronary angioplasty	PRS was constructed using meta-analysed GWAS and 7 different methods with the best performing PRSs selected based on AUC to construct An integrated PRS	Nikpay et al., Kurki et al., Ishigaki et al	Logistic (for tuning), Cox (for testing)	Logistic: PCE,5 PCs, genotype array, age, gender, SBP, smoking status, cholesterol, HDL-C, diabetes Cox: PCE, age, gender, SBP, smoking status, cholesterol, HDL-C, diabetes	testing (HR): 1.77 (1.745–1.796)	AUC -- tuning: 0.641 (0.635–0.648), testing : without PCE :0.640 (0.634–0.646), PCE adjusted: 0.753 (0.748-0.758) NRI -- testing: 0.093 (0.08-0.104)
1. ARIC 2. FHS 3. CARDIA 4. MESA **,(Di Narzo et al.,2021[32])**	Mainly European	2279 / 17603 , (12.95 %)	prevalent	NA , NA	prevalent MI Outcome definition: fatal or non-fatal MI	PRSs was constructed based on 8 different significance threshold cut-off, using CC4D GWAS coefficients as weight	Nikpay et al	Logistic, combined cohort results using RE meta-analysis	None	PRS CHD strongly associated with MI, regardless of the GWAS p-value to derive PRS	
1. Gene and Health (G&H) 2. eMerge (comparison) **,(Huang,Q et al.,2022[33])**	1. South Asian (British Pakistani and Bangeladeshi) 2. European Ancestry	996 / 17348 , (5.74 %)	max 10	6815/32816 , NA	Prevalent and Incident CHD Outcome definition: MI or coronary revascularization (CABG, coronary angioplasty with or without stenting)	Firstly tested 2 existing PRSs and then created PRSs using 3 different methods: clumping and thresholding , meta-PGSs proposed by Marquez-Luna et al and PRS-CSx	Nelson et al	Logistic, Cox	Logistic: age, age2, sex, 10 PCs, QRISK3 Cox (420/8112): age at recruitment, sex, QRISK3		AUC -- Logistic: best existing PRS in G&H (Wang,M et al): 0.009 (0.006-0.012), in eMERGE (Inouye et al): 0.022 (0.019-0.024), P+T: G&H: 0.006(0.001-0.015), eMERGE: 0.018(0.011-0.026), meta-PGS:0.009(0.0068-0.0120), PGS-CSx: 0.011(0.008-0.014) Cox: 0.853(0.838–0.867)

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
											NRI -- Cox: 3.9%(0.9–7.0%)
German data combined from 6 imputed datasets **,(Gola et al. 2,2020[34])**	Mainly European	7736 / 14510 , (53.31 %)	prevalent	1000 , NA	Prevalent CHD Outcome definition: GerMIFSI-II: MI before the age of 60 and at least one 1st-degree relative with CHD, GerMIFSIII:MI between ages 26 and 74; GerMIFSIV:CHD diagnosed before the age of 65 (men) or 70 (women); LURIC: >50% angiographic confirmation of vascular obstruction in at least one coronary vessel	The coefficients of GWAS conducted on the training dataset was used as weight for the PRS	GWAS performed in training set for sex)	Logistic	No		AUC -- 0.922(0.905-0.94)
UCLEB Consortium (7 prospective studies) **,(Morris et al.,2016[35])**	Mainly European	1444 / 11851 , (12.18 %)	max 10	NA , NA	Incident CVD Outcome definition: CVD includes CHD(non-fatal MI, any revascularisation procedure (CABG, angioplasty)) and stroke (non fatal ischaemic, haemorrhagic stroke, exclude transient ischaemic attacks)	Included 53 SNPs , weighted by CC4D GWAS coefficients	Deloukas et al	Logistic, combine results from each dataset using FE meta-analysis	none	OR per SD: 1.09 (1.03-1.15); include prevalent: 1.17 (1.10-1.25)	AUC -- 0.524 (0.504-0.541); include QRISK2: 0.623 (0.608-0.639) NRI -- include QRISK2 (10% risk cut-off): 1.18% (−0.23 to 2.60%), 20%: 0.68% (−1.16-2.52%)

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
1. German data combined from 6 imputed datasets 2. UK Biobank 3. Estonian Biobank **,(Gola et al. 1,2020[36])**	Mainly European	365 / 10000 , (3.65 %)	prevalent	431814 (UKB), 27048 (EB), 5581 (German) , NA	Prevalent CHD Outcome definition: German: MI, acute coronary syndrome, angina pectoris or coronary stenosis>50%, UKB: MI, CABG, coronary angioplasty with or without stenting, Estonian: MI and its complications	Tested two existing PRSs by Khera et al (2018) and Inouye et al (2018). New PRSs was constructed using PRSice based on the training dataset with different hyper-parameters.	Training datasets	Logistic	NA		AUC -- "Khera in GERMAN: 0.6699 (0.6557–0.6840), in EB: 0.5617 (0.5402–0.5833), in UKB: 0.6374 (0.6335–0.6412) Inouye in GERMAN: 0.5015 (0.4830–0.5140), EB: 0.6597 (0.6405–0.6789), UKB: 0.6377 (0.6339–0.6416) Trained PRS on combined in EB: 0.6112 (0.5919–0.6305), UKB: 0.5988 (0.5949–0.6027)"
1.ARIC (main), 2. Rotterdam(2068), 3. Framingham Offspring Studies(2339) **,(Brautbar et al.,2012[37])**	Mainly European	1110 / 8542 , (12.99 %)	max 18, 10-year event rate	NA , Rotterdam: 2068, Framingham: 2339	Incident CHD Outcome definition: ARIC: MI, a definite CHD death, or coronary revascularization . Rotterdam: fatal or nonfatal MI, CABP, PTCA. Framingham: nonfatal MI, death due to CHD.	Included 13 SNPs identified in previous literature, their coefficients in the identified GWAS were used as weights	Samani et al.,Kathiresan et al , Erdmann et al., Bare et al., Shiffman et al	Cox	age, sex, smoking, diabetes, SBP, antihypertensive medication use, TC, HDL-C	HR: ARIC: 2.30 (1.87-2.83), Rotterdam: 2.05 (1.50-2.70), Framingham: 1.12 (1.10-1.15)	AUC -- Increment: ARIC: 0.009 (0.006-0.011), Rotterdam: 0.006, Framingham: 0.011 NRI -- ARIC: 7.3% (3.6-4.5%), Rotterdam: 3.6%, Framingham: 4.5%
ARIC **,(Goldstein et al.,2014[38])**	Mainly European	620 / 8491 , (7.3 %)	max 10	NA , NA	Incident CHD Outcome definition: non-fatal or fatal MI, coronary revascularization procedure	Constructed 3 GRSs with CC4D GWAS coefficients as weights, one contains 50 SNPs influenced CHD risk, one with 33 non-risk factors SNPs and one	Deloukas et al	log-linear	age, sex	RR:full GRS: 1.31(1.22-1.41, Non-risk factors: HR 1.29(1.20-1.39), risk factor SNPs: 1.11(1.03-1.20)	AUC -- full GRS: 70.4%, non-risk factors: 70.1%, risk factors 69.2%

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
						with 17 risk-factors SNPs					
ARIC **,(Goldstein et al. 2,2015[39])**	Mainly European	620 / 8491 , (7.3 %)	max 10	NA , NA	Incident CHD Outcome definition: non-fatal or fatal MI, coronary revascularization procedure	The paper constructed 256 PRSs, using coefficients of CARDIOGRAM as weights and 4 different thresholding choices (p-value, LD, imputation quality and allele coding)	Schunkert et al	log-linear	age, gender, blood pressure, TC, HDL-C, smoking status, diabetes	Optimal PRS (p<0.001)RR per SD:1.277(1.186, 1.375)	AUC -- Optimal PRS (p<0.001): 0.783(0.766-0.800)
SMART **,(Tragante et al.,2013[40])**	Multi-ethnic	3788 / 8446 , (44.85 %)	prevalent	NA , NA	Prevalent CHD/MI Outcome definition: MI: At least two of the following: 1. Chest pain for at least 20 minutes, not disappearing after administration of nitrates. 2. ST-elevation >1 mm in two following leads or a left bundle branch block on the ECG 3. CK elevation of at least two times the normal value of CK and a MB-fraction >5% of the total CK.	Used 30 SNPs identified in previous GWASs to be associated with CHD, weighted by their GWAS coefficients.	Deloukas et al , Schunkert et al , Erdmann et al., Wang et al., Aoki et al	Logistic	age, sex, smoking, diabetes, body, BMI, hypertension, LDL-C	OR: Q4 vs Q1: 1.89 (1.51-2.37), Q3 vs Q1: 1.74 (1.39-2.19), Q2 vs Q1 HR: 1.42 (1.14-1.78)	
1. UK Biobank (South Asian ethnicity) 2.BRAVE study (testing)	South Asian	398 / 7244 , (5.49 %)	prevalent	491 , NA	Prevalent CHD Outcome definition: UK Biobank: MI, coronary	Tested on 8 different LDPred parameter settings with	Nikpay et al	Logistic	age, sex, 5PCs	OR per SD:train: 1.58 (1.42-1.76), test: 1.60 (1.32-1.94)	AUC -- train:0.8045, test: 0.66 NRI -- train: 0.3804, test: 0.3546

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
,(Wang,M et al. 2,2020[41])					revascularization (CABG, coronary angioplasty with or without stenting) BRAVE: MI	CC4D GWAS coefficients as weights, the best performing PRS was determined based on AUC					
Qatar Cardiovascular Biorepository (cases) and QBB (control) **,(Saad et al.,2022[42])**	Middle Eastern	1014 / 7023 , (14.44 %)	prevalent	NA , NA	Prevalent CHD Outcome definition: MI, unstable angina, PCI (from published protocol)	The paper created an Ensemble PRS by summing 5 published PRSs. Also tested those 5 PRSs individually	Nikpay et al., Deloukas et al., Van der Harst et al	Logistic	age, sex, BMI, 20 PCs	OR per SD: Koyama et al: 1.81(1.66–1.98) Wang, M et al: 1.53(1.42–1.64), Inouye et al: 1.54(1.43–1.66), Gola et al: 1.66(1.51–1.82), EnsemblePRS: 1.8 (1.66–1.96)	AUC -- Koyama et al: 0.667(0.649–0.685) Wang, M et al :0.683(0.665–0.701), Inouye et al: 0.686(0.667–0.704), Gola et al: 0.645(0.627–0.663), ensemble PRS: 0.702 (0.684-0.720), elliot et al dropped due to underperformance
Framingham Heart Study (FHS) **,(Powell et al.,2021[43])**	Mainly European	1091 / 7017 , (15.55 %)	median 12.0	NA , NA	Incident CVD Outcome definition: CHD (coronary death, MI, coronary insufficiency and angina), cerebrovascular events (including ischaemic stroke, haemorrhagic stroke and transient ischaemic attack), peripheral artery disease (intermittent claudication), heart failure	Used 63 SNPs previously identified to be associated with CVD, weighted by CC4D coefficients	Nikpay et al	Cox	education, age, sex, SBP, HDL-C, LDL-C, smoking, diabetes, BMI	HR per SD: unadjusted: 1.15 (1.08-1.22), adjusted: 1.13 (1.07-1.20)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
Rotterdam Study **,(De Vries et al.,2015[44])**	Mainly European	1449 / 5899 , (24.56 %)	mean 12.8	NA , NA	Prevalent CHD and Incident CHD (incident 964) Outcome definition: MI, myocardial revascularisation , CHD mortality	Constructed 3 PRSs based on 152 CHD related SNPs, using CC4D GWAS coefficients as weights: one included 49 genome-significant SNPs; one included 103 5% FDR SNPs; one included all 152 SNPs	Deloukas et al	Cox, logistic	1: age and sex, 2. +total and HDL cholesterol, systolic BP, prevalent T2B, antihypertensive medication, lipid-lowering medication and current smoking. 3. + family history of MI	model 3: OR per SD: GRS all:1.32(1.20-1.47) GRS gws: 1.24(1.12-1.37), GRS fdr:1.23(1.11-1.37) HR per SD: GRS all:1.13(1.06-1.20), GRS gws: 1.11(1.05-1.19), GRS-fdr: 1.07 (1.01-1.14)	AUC -- Prevalent: model 3: GRS-all: 0.009 (0.003-0.015), GRS-gws: 0.005 (0.000-0.010), GRS-fdr: 0.005 (0.000-0.010) Incident: model 3: GRS-all: 0.003 (-0.001-0.007), GRS-gws: 0.002 (-0.001-0.006), GRS-fdr: 0.002 (-0.001-0.004) NRI -- Prevalent: GRS-all: 0.004 (-0.034-0.041), GRS-gws: 0.041 (0.002-0.081), GRS-fdr:0.034 (-0.002-0.069) Incident: GRS-all: 0.017 (- 0.025-0.058), GRS-gws: 0.016 (-0.019-0.040), GRS-fdr:0.007 (-0.026-0.040)
1. WTCCC-CHD, MIGREN-Harps (tuning) 2.FINRISK, Framingham heart study (evaluation) **,(Abraham et al.,2016[45])**	Mainly European	2414 / 5883 , (41.03 %)	prevalent	NA , 1344/16082	Prevalent (tuning) and Incident CHD (evaluation) Outcome definition: WTCCC: MI, coronary revascularization (CABG, PTCA); MIGen: MI; FINRISK: MI, coronary revascularisation , death from CHD; Framingham heart study: MI, CHD death, angina pectoris,	PRSs were constructed using LD pruning and CC4D coefficients as weights. The best performing PRS was determined based on AUC. Also evaluated 3 existing PRSs	Deloukas et al	Logistic (tuning); Cox (evaluating); combined cohort results using FE meta-analysis	Logistic: age, sex, first 5 PCs; Cox: sex, geographic location, cohorts	Tuning: OR per SD: 1.70 (1.61-1.80); testing: HR per SD: 1.66 (1.55–1.78)	AUC -- Tuning: 0.64 (0.63-0.66) Harell's C index -- Testing (FRS+GRS) compare to FRS: 0.016 (0.01–0.02), ACC/AHA13+GRS compare to ACC/AHA13: 0.015 (0.009–0.02) NRI -- testing categorical NRI: FRS: 0.1 (0.055-0.145), ACC/AHA13:0.104 (0.059-0.149)

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing and External validation sample size	Outcome and Outcome definition	PRS construction	GWAS source/ Author	Regression Model	Adjustments	Results	
										OR/HR/RR (95% CI)	Discrimination
					coronary insufficiency						
Partners Healthcare Biobank (1/3 tuning and 2/3 testing) **,(Ge et al.,2019[46])**	Mainly European	920 / 5417 , (16.98 %)	prevalent	1839/10834 , NA	Prevalent CHD Outcome definition: algorithmically defined CHD	Used a newly proposed method PRS-CS and PRS-CS-auto, with CC4D GWAS coefficients as input. Also computed PRSs using to P+T, LD Pred, LdPred-inf and PRS unadjusted for comparison	Nikpay et al	Logistic	age, sex, 10 PCs	OR (top 10% vs rest): PRS-CS: 2.3447, PRS-CS-auto:2.0122, LDPred: 2.2918, P+T:1.9172, LDPred-inf:2.0157, PRS-unadjust: 1.8637	AUC -- PRS-CS: 0.5882, PRS-CS-auto:0.5782, LDPred: 0.5893, P+T:0.5703, LDPred-inf:0.5642, PRS-unadjust: 0.5573
UK Biobank **,(Manikpurage et al.,2021[47])**	Mainly European	219 / 5000 , (4.38 %)	median 11.1 (training), 11.04 (validation)	32475/403422 , NA	Prevalent CHD, prevalent and incident MI (n=14827, 7746) Outcome definition: ICD-10: MI, other acute CHD, atherosclerotic / chronic ischemic heart disease, CABG, coronary angioplasty, with or without stenting	The PRSs were constructed using LDPred and summary statistics of CC4D GWAS	Nikpay et al	Logistic (prevalent CHD, MI), cox(incident MI)	age, sex, 10 PCs	CHD OR per SD :training: 1.48 (1.29–1.70), validation: 1.56 (1.54–1.58) MI: OR per SD : 1.63 (1.60-1.65), HR(per SD):1.53 (1.49–1.56)	AUC -- validation: CHD: 0.766, prevalent MI: 0.772 Harell's C index -- incident MI: 0.729 (0.724-0.735) NRI -- incident MI (2%): 0.0452 (0.0333-0.0573)

MI: Myocardial Infarction, CHD: Coronary Artery Disease, CVD: Cardiovascular Disease, BMI: Body Mass Index, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, OR: Odds Ratio, RR: Risk Ratio, HR: Hazard Ratio, CI: Confidence interval, PRS: Polygenic Risk Score, PC: Principal Components, CC4D: CARDIoGRAMplusC4D consortium, GWAS: Genome-wide Association Study, AUC: Area under the ROC curve, NRI: Net reclassification Index, P+T: Pruning and Thresholding, SD: standard deviation, TC: Total cholesterol, HDL-C: High Density Lipoprotein Cholesterol, LDL-C: Low Density Lipoprotein Cholesterol, TDI: Townsend Deprivation Index, IPAQ: The international Physical Activity Questionnaire, CABG: Coronary Artery bypass graft, PTCA: Percutaneous Transluminal coronary angioplasty, PCI: Percutaneous Coronary Intervention, SNP: Single Nucleotide Polymorphisms, BG: Blood Glucose, TG: Triglycerides, PCE: pooled cohort equation, SD: Standard deviation, FRS: Framingham Risk Score, MDCS: Malmo Diet and Cancer Study, FE: Fixed effect, RE: Random effect, ASCVD: atherosclerotic cardiovascular disease

Table S2: Identified papers that applied existing PRSs for CHD risk estimate in their studies.

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing sample size	Outcome and Outcome definition	PRS used	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
UK Biobank, (**Goodman et al., 2022[48]**)	Mainly European	26699 / 476851 , (5.6 %)	prevalent	NA	Prevalent CHD Outcome definition: MI or Heart attack	Khera et al.	Logistic regression	mutual interaction of age and BMI by sex (via tensor-product splines), smoking and its interaction with sex, self-reported white race, 5 PCs, genotype platform, asthma, COPD	OR per SD: 1.695(1.672-1.718)	
UK Biobank, (**Isgut et al., 2021[49]**)	Mainly European	35564 / 396991 , (8.96 %)	prevalent	3952/44110	Prevalent CHD and MI (n=15930) Outcome definition: CHD: MI and its complications, PTCA, CABG, Triple heart bypass, angina, other acute CHD, chronic CHD; MI: MI and its complications	FDR202, 1.7M score (by inouye et al.), 46K (Abraham et al.), Khera et al. .	Logistic regressions with each PRS alone and all 4 together (10-fold CV)	1. None (for univariate PRS) 2.+sex, the first 4 PCs, 3. +T2B, family history of heart disease, age, SBP, BMI, Cholesterol, TG, LDL, smoking status		AUC -- 1. CHD: FDR202: 0.57, GRS46K:0.57, 1.7M: 0.58, and 6M: 0.60, all 4 together: 0.61 MI: FDR202: 0.60, GRS46K:0.59, 1.7M: 0.60, and 6M: 0.63, all 4 together: 0.64 2. All combined -- CHD: 0.67, MI: 0.73 3. All combined -- CHD: 0.81, MI: 0.84
UK Biobank, (**Zaccardi et al., 2022[50]**)	Mainly European	10862 / 380693 , (2.85 %)	median 11.9	NA	Incident CHD Outcome definition: MI and its complications, other acute CHD, CABG, PTCA	Mega et al. (PGS-10), Tada et al. (PGS-11), Abrham et al. (PGS-12), Khera et al. (PGS-13), Inouye et al. (PGS-18), paquette et al. (PGS-19), Natarajan et al. (PGS-57), Morieri et al. (PGS-58), Hajek et al. (PGS-59)	Roystob-Parmar-Lambert parametric survival model	age, sex, TDI, SBP, LDL-C, smoking status, history of diabetes, family history of MI	PGS-13 were most associated with CHD and were mainly analysed. Women: Avg vs Low (walking) 1.16 (1.04-1.30), Brisk vs Low (walking): 1.15 (1.01-1.30) Men: Avg vs Low (walking) 1.08 (1.01-1.16), Brisk vs Low (walking): 1.13 (1.05-1.22)	Harell's C -- Women: 0.801 (0.793 - 0.808), increment 0.038 (0.032-0.043), men 0.732 (0.728-0.737), increment 0.045 (0.041-0.049)
UK Biobank, (**Kim, Y et al., 2022[51]**)	Mainly European	9185 / 373026 , (2.46 %)	median 12.6	NA	Incident CHD Outcome definition: ICD-9: MI and its complications, other acute CHD	Ntalla et al.	Cox regression	sex, genotyping array type, 10 PCs	HR: high vs low tertile: 2.04 (1.94–2.15), medium vs low tertile: 1.43 (1.36–1.52)	
UK Biobank, (**Sun, L et al., 2021[52]**)	Mainly European	3333 / 306654 , (1.09 %)	median 8.1	NA	Incident CHD (main outcome CVD (5608)) Outcome definition:	For CHD, used Inouye et al. meta score of 1.7M SNPs	Cox regression	1. age, sex 2. study centre, sex, age at baseline, smoking status, history of diabetes, SBP, TC,	1. HR per SD: CHD: 1.56 (1.51-1.62), CVD: 1.36 (1.32-1.40)	Harell's C increment--2. CHD: 0.0211 (0.0167-0.0255), CVD: 0.0107 (0.0081-0.0132)

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing sample size	Outcome and Outcome definition	PRS used	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
					CHD: MI and its complications, fatal acute CHD and other chronic CHD; CVD: CHD, stroke, PTCA, CABG			HDL-C (with two PRSs on CVD)		continuous NRI-- CHD: 0.3157 (0.2778-0.3536), CVD: 0.2068 (0.1785-0.2351)
UK Biobank, (**Ramirez et al., 2022[53]**)	Mainly European	6186 / 189787 , (3.26 %)	prevalent	6185/189787 (for evaluation)	Prevalent CHD Outcome definition: MI and its complication, PTCA, CABG, triple heart bypass	25 genetic risk scores (GRSs) were derived, a GRS for CHD, for CHD, PRS was constructed using Khera et al. (2018)	Logistic regression (main predictor is QRISK3)	genetic array, first 10 PCs	OR per SD: no adjustments: 1.559 (1.519-1.598), with adjustments: 1.559 (1.519-1.600.), with adjustment (+QRISK3): 1.537 (1.498-1.579)	NRI: with adjustment (+QRISK3 vs no QRISK3): 0.077 (0.075 - 0.079)
UK Biobank, (**Livingstone et al., 2021[54]**)	Mainly European	1141 / 77004 , (1.48 %)	mean 7.8	NA	Incident MI and CVD mortality (364) Outcome definition: CVD mortality: mortality of MI, ischaemic stroke, haemorrhage stroke ; MI: MI and its complication	Ntalla et al. (2019)	Cox regression	age, sex, TDI, smoking status, physical activity, medication use, family history of CVD, energy intake, 8 PCs, genotyping batch, diet quality score	HR: MI: 1.33(1.25-1.41), CVD mortality: 1.08 (0.98-1.20)	
UK Biobank, (**Livingstone et al., 2021[55]**)	Mainly European	1140 / 76958 , (1.48 %)	mean 7.8	NA	Incident MI and CVD mortality (364) Outcome definition: CVD mortality: mortality of MI, ischaemic stroke, haemorrhage stroke ; MI: MI and its complication	Ntalla et al. (2019)	Cox regression	age, sex, TDI, 8 PCs, genotyping batch, lifescore	HR: MI: 1.35(1.27-1.43), CVD mortality: 1.11(1.00-1.23)	
HUNT, (**Rostami et al., 2020[56]**)	Mainly European	2609 / 61465 , (4.24 %)	mean 17.7	NA	Incident MI Outcome definition: MI and its complications, coronary artery interventions	The paper tried 6 published PRSs, which included as few as 6 SNPs and as many as 157 SNPs for CHD from (Irribarren et al., Theriault et al. and Beaney et al.)	Cox regression	sex	HR per SD: GRS6: 1.18(1.13-1.22), GRS10:1.17(1.12-1.21), GRS157:1.23(1.19-1.28), GRS25: 1.16 (1.11-1.20), GRS18: 1.20 (1.16-1.25), GRS21: 1.20 (1.16-1.25)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing sample size	Outcome and Outcome definition	PRS used	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
1. Partners Healthcare Biobank, 2.Penn Medicine Biobank, 3. Mount Sinai BioMe Biobank,(**Aragam et al., 2020[57])**	African (6.3%), AdMixed American (5.9%), East Asian (1.2%), European (85.8%), South Asian (0.8%)	11020 / 47108 , (23.39 %)	prevalent	NA	Prevalent CHD Outcome definition: MI, other acute forms of CHD, other chronic CHD, coronary bypass, coronary revascularisation (CABG, PTCA), coronary thrombolysis	Khera et al., mean centred and normalised by genetic ancestry within each cohort	Logistic regression (for each cohort), combined results from 3 cohorts using FE meta-analysis	age, sex, up to ten PCs of ancestry. The Partners Biobank and Penn Medicine Biobank additionally adjusted for genotyping array.	OR per SD (overall): 1.42 (1.38-1.46), top 5% vs rest: 2.28 (2.04-2.53), top 20% vs rest: 1.92 (1.80-2.04), OR per SD for admixed American (n=7048): 1.50 (1.44 - 1.57)	AUC -- Partners Biobank 0.59, Penn Medicine Biobank: 0.60, BioMe: 0.61, for admixed American (n=7048): 0.63
eMERGE network, (**Dikilitas et al., 2020[58])**	European ancestery (2221/45,645), African ancestry (311/7,597, Hispanic Ethnicity(120/2,493)	2221 / 45645 , (4.87 %)	median 11.7 (EA), 9.2(AA), 10.4(HE)	NA	Incident (primary) and prevalent CHD Outcome definition: MI, Coronary Revascularisation (PCI, CABG)	2 restricted PRS: 1. Tikkanen et al. and 2. Tada et al., 2 genome-wide PRSs 3. Inouye et al. and 4. Khera er al	Cox regression Logistic regression including all prevalent (EA=5887, AA=527, HE=299) and incident CHD	Cox: sex, eMERGE site, 5 PCs Logistic: sex, eMERGE site, 5 PCs, age at first EHR record, duration of HER	HR per SD EA: 1. 1.18 (1.13–1.23), 2. 1.20 (1.15–1.25), 3.1.53 (1.46–1.60), 4. 1.50 (1.43–1.56); AA:1. 1.11 (0.99–1.24), 2. 1.05 (0.94–1.17), 3. 1.27 (1.13–1.43), 4. 1.19 (1.07–1.33); HE: 1. 1.14 (0.94–1.37), 2. 1.13 (0.93–1.36), 3. 1.53 (1.23–1.90), 4. 1.16 (0.96–1.41) OR per SD: EA: 1. 1.24 (1.21–1.28), 2. 1.28 (1.25–1.32), 3.1.73 (1.68–1.78), 4. 1.66 (1.62–1.71); AA:1. 1.07 (0.99–1.16), 2. 1.05 (0.98–1.14), 3. 1.40 (1.30–1.52), 4. 1.30 (1.21–1.41); HE: 1. 1.27 (1.12–1.42), 2. 1.20 (1.06–1.35), 3. 1.93 (1.67–2.22), 4. 1.42 (1.25–1.61)	AUC -- Cox: EA: 1.0.697, 2.0.698, 3. 0.719, 4. 0.719; AA: 1. 0.652, 2. 0.649, 3. 0.663, 4. 0.656; HE: 1. 0.655, 2. 0.654, 3. 0.683, 4. 0.659 Logistic EA: 1.0.748, 2.0.750, 3. 0.772, 4. 0.770; AA: 1. 0.763, 2. 0.763, 3. 0.775, 4. 0.771; HE: 1. 0.771, 2. 0.769, 3. 0.794, 4. 0.776
1.China-PAR 2.China MUCA-1998 3.InterASIA,	East Asian (Chinese)	1373 / 41149 , (3.34 %)	median 13.01	NA	Incident CHD Outcome definition: unstable angina, acute MI, CHD death	Lu et al. (2022)	Cox regression	1. age, sex 2. +geographical region, urbanisation, education level, smoking, drinking, work-related physical activity	HR: 1. Q2-4 vs Q1: 1.56 (1.32-1.84), Q5 vs Q1:2.81 (2.35, 3.35); 3. Q2-4 vs Q1: 1.40 (1.18, 1.66), Q5	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing sample size	Outcome and Outcome definition	PRS used	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
(**Liang et al., 2022[59]**)								3.+BMI,SBP, serum glucose, TC	vs Q1:2.33 (1.94, 2.79)	
1. MDCS 2. UK Biobank (replication), (**Hindy et al., 2020[60]**)	Mainly European	4122 / 28556 , (14.43 %)	median 21.3, UK Biobank median 8.1	7708/325003	Incident CHD Outcome definition: MDCS: fatal or nonfatal MI, CABG, PCI, CHD death, UK Biobank: MI, CABG, coronary angioplasty with or without stenting	Khera et al., ancestry adjusted	Cox regression	1.sex, age 2.+ 10 PCs, SBP, DBP, apolipoprotein A, apolipoprotein B, TC, HDL-C, LDL-C, BMI, smoking status, diabetes, family history of CHD	HR per SD: MDCS: 1. 1.45 (1.40–1.49) 2. 1.45 (1.34–1.56) UKB: 1. 1.53 (1.49–1.56) 2. 1.46 (1.42–1.49)	Harrell's C -- MDCS: 1. 0.759(0724-0.794), added PCE: 0.802(0.763-0.841); UKB: 1. 0.756(0.750-0.762); added PCE: 0.768(0.760-0.776) NRI -- adding PRS to PCE: MDCS (10-year): 0.165 (0.076-0.182); UKB (8-year): 0.085 (0.071-0.098)
FINRISK, (**Martikainen et al., 2021[61]**)	Mainly European	2063 / 26203 , (7.87 %)	mean 13.8	NA	Incident CHD Outcome definition: MI, unstable angina, CABG, PCI, death (MI and its complications, angina, other chronic CHD, other acute CHD, cardiac arrest, sudden cause unknown, unattended)	Tested Abraham et al., validated using Tada et al..	Cox regression	1. sex, region of residence, calendar year, study batch, 10 PCs 2. +smoking, alcohol use, body mass index, HDL-C, TC, blood pressure, diabetes (+ education as main predictor)	HR Abraham 1. Q2 vs Q1: 1.29(1.11-1.49), Q3 vs Q1: 1.57(1.36-1.81), Q4 vs Q1: 2.26(1.97-2.59) ; 2. Q2 vs Q1: 1.24(1.08-1.44), Q3 vs Q1: 1.50(1.30-1.73), Q4 vs Q1: 2.12(1.84-2.43) Tada:1. Q2 vs Q1: 1.14(1.00-1.30), Q3 vs Q1: 1.17(1.03-1.33), Q4 vs Q1: 1.55(1.38-1.76) ; 2. Q2 vs Q1: 1.14(1.01-1.30), Q3 vs Q1: 1.16(1.02-1.32), Q4 vs Q1: 1.53(1.35-1.73)	
MDCS, (**Hindy et al. 2, 2018[62]**)	Mainly European	3217 / 24443 , (13.16 %)	median 19.4	NA	Incident CHD Outcome definition: fatal or nonfatal MI, CABG, PCI, death due to MI or chronic CHD	Tada et al.	Logistic regression (the assumption of proportionality was violated in Cox regression)	1. age, sex, education, total energy intake, leisure time physical activity, alcohol consumption 2. age, sex, ApoB , Apo AI, SBP, antihypertensive medication, diabetes mellitus at baseline, family history of MI	1. OR per tertile: 1.31 (1.25–1.37)	AUC -- 2. never smoker: 0.757(0.742-0772), former smoker: 0.749(0.735-0.763), current smoker: 0.744 (0.728-0.759) NRI -- 2. 0.18

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing sample size	Outcome and Outcome definition	PRS used	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
MDCS, (**Fritz et al., 2017[63]**)	Mainly European	2213 / 23595 , (9.38 %)	median 14.4	NA	Incident CHD Outcome definition: coronary revascularization, fatal or nonfatal MI, CHD death	Tada et al.	cox regression	age at baseline, sex, smoking status	HR: Q5 vs rest: 1.53 (1.39–1.68), Q5 vs Q1: 2.01 (1.76–2.30)	
MDCS, (**Svensson et al., 2017[64]**)	Mainly European	1938 / 18559 , (10.44 %)	mean 14.6 (CHD), 15.1 (CVD death)	NA	Incident CHD, CVD deaths (1071) Outcome definition: CHD: fatal or non-fatal MI, CABG, PCI, CVD: CHD, acute rheumatic fever, chronic rheumatic heart disease, hypertensive disease, disases of pulmonary circulation, cerebrovascular disease, disases of arteries, arterioles and capillaries, other heart diseases	Tada et al.	Cox regression	1. gender and age 2. +education, SEI, smoking, drinking, diabetes status, BMI, hypertension, use of lipid-lowering medication.	HR: CHD: 2. HR: Q2 vs Q1 1.24 (1.08–1.42); Q3 vs Q1: 1.43 (1.25–1.64); Q4 vs Q1: 1.72 (1.51–1.96) CVD death: Q2 vs Q1 1.00 (0.83–1.19); Q3 vs Q1: 1.26 (1.06–1.50); Q4 vs Q1: 1.29 (1.08–1.53)	
GERA, (**Iribarren et al. 2, 2018[65]**)	1. African (95/2089) 2. Latinos(316/4349) 3. East Asians(39/4804)	450 / 11242 , (4 %)	median 8.7	NA	Incident CHD Outcome definition: Hospital primary discharge diagnoses of MI, angina (stable or unstable), coronary revascularization procedures (CABG, PCI), death due to angina, MI, hypertensive heart disease	Used 2 PRSs from Iribarren et al. (GRS_12, GRS_51)	Cox regression, combine subset results using FE meta-analysis	a) 6 PCs b)+age, gender, TC, HDL-C, SBP, DBP, smoking status and diabetes c)+family history of heart disease d)education level, BMI, antihypertensives, lipid lowering drugs, alcohol consumption	HR per SD: GRS_12 a) 1.15(1.04-1.26), b)1.17(1.06-1.28), c) 1.17(1.06-1.28) d)1.15(1.04-1.27); GRS_51: a) 1.17(1.06-1.29), b) 1.18 (1.07-1.30), c) 1.18 (1.07-1.30), d) 1.16(1.05-1.28)	Harell's C -- only adjusted for FRS: GRS_12: 0.725; GRS_51: 0.723 NRI -- include FRS: GRS_12: 0.06 (0.02–0.11), GRS_51: 0.03 (−0.00–0.06)
1. MHI phase 1-2 2. CARTaGENE, (**Wunnemann et al., 2019[66]**)	Mainly European	3639 / 11021 , (33.02 %)	prevalent	NA	Prevalent CHD Outcome definition: MI, CABG, PCI	Khera et al., Inouye et al.	Logistic regresion in each of the three populario n and combine results using	age, sex, 4 PCs, statin use (MHI); age, sex,4 PCs, recruitment centre (CARTaGENE)	OR per SD: Khera et al.: 1.61 (1.51–1.73), Inouye et al.: 1.69 (1.58-1.81)	AUC -- Khera: MH1: 0.72 (0.70-0.74), MH2: 0.89 (0.88-0.91), CARTaGENE: 0.84 (0.81-0.87) Inouye: MH1: 0.72 (0.70-0.75), MH2: 0.89 (0.88-0.91),

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing sample size	Outcome and Outcome definition	PRS used	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
							FE meta-analysis.			CARTaGENE: 0.84 (0.81-0.87)
ARIC, (**Hasbani et al., 2022[67]**)	Mainly European (1725/8372) with a small proportion of Black ethnicity (427/2314)	2152 / 10686 , (20.14 %)	median 26.4	NA	Incident CHD Outcome definition: hospitalised MI, cardiac revascularisation, CHD death (known nonatherosclerotic or noncardiac atherosclerotic causes, chronic CHD, coronary insufficiency, angina,)	Used Khera et al., a residual PRS was then created after adjusting for the first 11 principal components for ancestry.	Non parametric survival model	family history of CHD, sex, study center	culmulative incidence function: Q1: 19.5% (17.1%–22.1%), Q2-4: 27.3% (25.1%–29.6%), Q5:36.9 (34.0%–40.0%),	
Canadian Longitudinal Study on Aging (CLSA), (**Menniti et al., 2021[68]**)	Mainly European	1286 / 9892 , (13 %)	prevalent	NA	Prevalent CHD Outcome definition: MI, angina, heart failure	Used 39 out of 50 SNPs of the scores created by Khera et al. 2	Logistic regression	age, 5 PCs, sex, education level, province at recruitment, total household income, smoking status, urban/rural classification, immigration status	OR: 1.06 (1.04-1.07)	
Health and Retirement study (HRS), (**Hamad et al., 2022[69]**)	White (3084/7,522) and Black (407/1,198)	3084 / 7522 , (41 %)	prevalent	NA	Prevalent CVD Outcome definition: self-reported	Used HRS constructed PRS (release 3) which constructed separatly for white and black ethnicity	Ordinary least squares linear regression, separate for white and black ethnicity	gender, foreign-born status, birth year, census region of residence, 10 PCs		r-squred -- white: 6.0 (5.0-7.0), black: 2.5 (0.8-4.2), The variance explained by PRS is limited, especially in black ethnicty
1.ARIC (7480 (visit 1), 4847 (visit 4)) 2.MESA (n=2390), (**Mosley et al., 2020[70]**)	Mainly European	1005 / 7480 , (13.44 %)	prevalent	NA	Prevalent CHD+ incident CHD Outcome definition: ARIC: MI, heart or arterial surgery, CABG, or angioplasty; MESA : MI, resuscitated cardiac arrest, definite or probable angina if followed by a revascularization, CHD death	Khera et al.	Logistic regression	age, sex, 5 PCs	OR per SD for ARIC: 1.89(1.75-2.03), top 20% vs rest: 2.89 (2.49-3.36), top 10% vs rest: 3.19 (2.64-3.84), top 5% vs rest: 4.14 (3.25-5.26)	

Study population, (Author/Year)	Ethnicity	Study Population Sample Size, Case (%)	Follow-up time (yr)	Testing sample size	Outcome and Outcome definition	PRS used	Regression Model	Adjustments	Results	
									OR/HR/RR (95% CI)	Discrimination
1. Framingham Offspring Study . 2. ARIC , (**Khan et al., 2022[71]**)	Mainly European	1073 / 5740 , (18.69 %)	max 35	NA	Incident CHD Outcome definition: FOS: fatal or nonfatal MI, death resulting from CHD, new onset angina, coronary insufficiency; ARIC: MI, a definite CHD death, an unrecognized MI defined by ARIC ECG readings, or coronary revascularization	Khera et al.	Cox regression. 3 age group analysed separately	age, sex	HR per SD(late-middle life) unadjusted:1.21 (1.14, 1.28), adjusted: 1.22 (1.15, 1.30)	AUC -- unadjusted: 0.555 (0.537, 0.572), adjusted: 0.655 (0.639, 0.671)

MI: Myocardial Infarction, CHD: Coronary Artery Disease, CVD: Cardiovascular Disease, BMI: Body Mass Index, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, OR: Odds Ratio, RR: Risk Ratio, HR: Hazard Ratio, CI: Confidence interval, PRS: Polygenic Risk Score, PC: Principal Components, CC4D: CARDIoGRAMplusC4D consortium, GWAS: Genome-wide Association Study, AUC: Area under the ROC curve, NRI: Net reclassification Index, P+T: Pruning and Thresholding, SD: standard deviation, TC: Total cholesterol, HDL-C: High Density Lipoprotein Cholesterol, LDL-C: Low Density Lipoprotein Cholesterol, TDI: Townsend Deprivation Index, IPAQ: The international Physical Activity Questionnaire, CABG: Coronary Artery bypass graft, PTCA: Percutaneous Transluminal coronary angioplasty, PCI: Percutaneous Coronary Intervention, SNP: Single Nucleotide Polymorphisms, BG: Blood Glucose, TG: Triglycerides, PCE: pooled cohort equation, SD: Standard deviation, FRS: Framingham Risk Score, MDCS: Malmo Diet and Cancer Study, FE: Fixed effect, RE: Random effect, ASCVD: atherosclerotic cardiovascular disease, COPD: Chronic Obstructive Pulmonary Disease, T2B: Type 2 Diabetes, EHR: Electronic Health Record

Table S3: Identified GWAS studies that sourced the development or application of the selected external PRSs for CHD in the identified studies.

Author (Cohort)/Year of publish	Imputation	Datasets included	Ethnicity	Discovery set sample size	Replication set sample size	Outcome definition	No. of Variants tested	Model	Adjustments
Tcheandjieu et al. (Hispanic)/2022[21]	1000 Genomes phase 3 version 5	Million veteran program, Meta-analysis: Nikpay et al., Van der Harst et al., Biobank Japan	European, Hispanic and African American	118,731/399,795 (MVP); **Hispanic:** 6378/30648		Discharge diagnosis of acute MI, coronary revascularization, or any two of the encounter (acute or old MI, coronary syndrome, angina, chronic IHD)	over million	GWAS stratified for 3 ancestries. 2 IVW Meta-analyses: 1) MVP white participants, the CC4D 1KG study and the UK Biobank CHD study; 2) MVP Black participants, MVP Hispanic participants and Biobank Japan.	sex, first 10 PCs
Huang Q et al. (South Asian)/2022[33]	GenomeAsia pilot reference panel	Genes & Health (G&H)	British Pakistani and Bangladeshi individuals	1,110/22,008		MI or coronary revascularization in either primary and secondary care data	336,133	GWAS conducted using SAIGE	age, age2, sex and the first 20PCs
Kurki et al. (FinnGen)/2022[72]	population specific imputation reference panel of 3,775 high-coverage (25-30x) whole-genome sequenced (WGS) Finns	FinnGen, replication in Estonian biobank and UK biobank	Mainly European	11,622/218,792 (FinnGen)	2,413/136,724	MI and its complications	16,387,711	Conducted pheWAS, replicated in Estonian Biobank and UK Biobank	
Koyama et al. (BBJ)/2020[15]	Reference panel created using a small sample of sequence data (n=2,504)	Biobank Japan, Meta-analysis: Nikpay et al., Van der Harst et al	South Asian	25,892/168,228 (BBJ)		stable and unstable angina, and MI	19,707,525	GWAS in BBJ, then meta-analysed with Nikpay et al and UKBB (Van der Harst) GWAS using MANTRA algorithm	sex, age, age-squared, top 10 PCs
Ishigaki et al. (Japan Biobank)/2020[73]	1KG Phase3	BBJ, replication: OACIS,NCGG	Japanese	29,319/212,453 (BBJ)	2,855/18,066	stable and unstable angina, and MI	8,712,794 autosomal variants and 207,198 X chromosome variants	conducted a GWAS by employing a GLMM using SAIGE	sex, age, first 5 PCs
Van de Harst et al. (UKBB)/2018[74]	HRC v1.1. panel	UK Biobank, replication in CC4D	Mainly European	34,541/296,525 (UKB)	88,192/ 250,736(CC4D)	heart attack/MI, coronary angioplasty +/- stent, CABG and triple heart bypass	7,947,838	GWAS in UKB, replicated in CC4D. Then performed a meta-analysis of the GWAS of UK Biobank and replication dataset.	age, gender, the first 30 PCs, genotyping array (UK Biobank versus UK BiLEVE)
Zhou et al. (UKBB SAIGE)/2018[75]	HRC reference panel	UK Biobank	Mainly European	31,355/408,961		Myocardial infarction (MI), coronary artery bypass grafting, or coronary artery angioplasty, angina	28 million	Conducted GWAS using SAIGE method	sex, birth year, the first 4 PCs
Nelson et al. (UKBB FDR)/2017[76]	1KG panel	UK Biobank, meta-analysis with Nikpay et al	Mainly European	71,602/332,477	4,412/8,322	HARD CHD : fatal or nonfatal MI, PTCA or coronary artery CABG. SOFT CHD (primary	9,149,595	UKB GWAS meta-analysed with Nikpay et al and Exome chip meta-analysis using fixed effect IVW, then performed FDR analysis	genotyping array (UK Biobank versus UK BiLEVE), the first five PCs

						outcome): HARD CHD, chronic IHD and angina.			
Webb et al./2017[77]		20 discovery studies, 8 replication studies	Mainly European, 2 studies in the replication studies were of south asian ancestery	42,335/120,575	30,533/73,063	Varied across studies, Including stable and unstable angina, MI, coronary revascularisation,	29,383	Logistic regression in each studies and fixed-effect meta-analysis. Replication performed for variants with suggestive association.	10 PCs
Howson et al./2017[78]	No imputation	Primary GWAS consist: 4 studies contain EUR ancestry, 2 studies of SA ancestry, 7 studies of EA ancestry, 8 studies of AA ancetsry. Meta-analysis: CC4D (Nikpay et al)	Primary cohort: European (52%), South Asian (23%), East Asian (17%) and African American (8%) ancestries	29,976/56,309 (primary GWAS); 88,192/250,736 (meta-analysis)			79,070	GWAS for the primary cohort using GEMMA. Fixed effect IVW to combine study results from primary GWAS and CC4D	5 PCs
Klarin et al./2017[79]	Combined 1KG/UK10K reference panel	UK Biobank, meta-analysis with Nikpay et al	Mainly European	4,831/120,286 (UKB)	Stage 2: 42,355/120,595 , Stage 3: 60,801/ 184,305	Myocardial infarction (MI), coronary artery bypass grafting, or coronary artery angioplasty	around 9 million	3 stage sequential analysis, stage 1: GWAS in UK Biobank, stage 2: took forward 2,190 nominal significant SNPs for meta-analysis with CARDIOGRAM exome consortia. Stage 3:took forward 387,174 variants that reached nominal significance in stage 1 (and not available in stage 2) for meta-analysis with Nikpay et al.	age, gender, chip array
Stitziel et al. (MI genetics and CARDIOGRAM Exome Consortia)/2016[80]		20 discovery studies, 8 replication studies	Mainly European, 2 studies in the replication studies were of south asian ancestery	42,335/120,575	30,533/73,063	Varied across studies. Includes, stable and unstable angina, MI, coronary revascularization,	54,003 coding-sequence exom	GWAS in each studies and fixed-effect IVW meta-analysis. Replication performed for variants with suggestive association.	10 PCs
Nickpay et al. (CARDIOGRMplusC4D 1000 genome)/2015[81]	1KG phase 1 version	48 studies	(77%) European ancestry; 13% south Asian (India and Pakistan); 6% east Asian (China and Korea), smaller samples of Hispanic and African Americans	60,801/184,305; **Hispanic:** 758/4,095		MI, acute coronary syndrome, chronic stable angina or coro- nary stenosis > 50%	8.6 million SNPs and 836k indels	GWAS in each studies and then fixed-effect IVW meta-analysis	

Deloukas et al. (CARDIOGRMplusC4D consortium)/2013[82]	Metabochip array only (60% of samples) or HapMap	34 additional studies (stage-2) in addition to 14 in Schunkert et al (stage-1): 48 studies, replication in 4 studies	Mainly European and a small proportion of South Asian	63,746/194,427	3,630/15,613	MI, acute coronary syndrome, chronic stable angina	79,138	Fixed effect IVW meta-analysis in stage 2 datasets: The combination of evidence across stage 1 and stage 2 meta-analysis results was performed using Fisher's combined P-values method. Fixed effect IVW in stage 3 replication analysis.	sex, age
Schunkert et al. (CARDIoGRAM GWAS)/2011[83]	HapMap 2	14 GWAS studies of European descent in the discovery and more than 10 in the replication	Mainly European	22,233/86,995	32,584-136,416 for significant SNPs in discovery set	CHD and MI	around 2.3 million	GWAS in each studies and then fixed effect IVW meta-analysis	sex,age, genotyping array
Peden et al. (C4D Consortium)/2011[84]	Whole genome sequencing	4 studies in the discovery and 10 in the replication	Mainly European and South Asians	15,420/30,482	21,408/40,593	CHD and MI	574,919	GWAS in each studies and then fixed-effect meta-analysis	
Wang et al./2011[85]	No imputation	GeneID	Chinese	stage 1 discovery: 230/460, stage 2 discovery validation: 572/1,008	Stage 3: 2,668/6,585	coronary angiography, coronary artery bypass graft, percutaneous coronary intervention and/or myocardial infarction	around 350,000	3 stage GWAS, stage 1 GWAS discovery (Pearson's 2×2 or 2×3 contingency test), stage 2 validation of significant association (Multivariate logistic regression analysis), stage 3 replication	age, gender, smoking, hypertension, diabetes, lipid levels
Akoi et al./2011[86]	No imputation	BBJ, OACIS	Japanese	stage 1 discovery: 194/1,733, stage 2 discovery: 1.394/2,819, stage 3 discovery: 1,500/2,856, 5722	2,283/5,722	MI	210,785	3 stage GWAS, stage 1 -3 are discovery, each with more stringent p-value cutoff and SNPs passed cuttoff will be brought to the next stage. Stage 4 is replication.	
Erdmann et al./2009[87]	No imputation	stage 1: GerMIFS II, stage 2:WTCCC CHD study4, MIGen/IATVB stage 3:KORA-MI, ECTIM, UK-MI, PopGen, LURIC, LMD, Angio-GER, Atherogene	Mainly European	stage1: 1,222/2,520	stage 2: 5,768/13,425, stage 3: 12,417/24,828	MI and Angiographically validated CHD	567,119	stage 1 is exploratory GWAS, stage 2 silico-replication study, stage 3:Prospectively planned pooled replication study. Meta-Analysis of all 3 stages (IVW, fixed effect and random effect)	
Clarke et al./2009[88]	No imputation	Recruited for the study from 4 countries, replication datasets from 3 independent population	Mainly European	discovery: 3,145/6,497	4,846/9,440	MI or symptomatic acute coronary syndrome (SACS) before the age of 66 years	34,399	Logisic regression, replication analysis: fixed effect IVW to meta-analysed 3 cohorts	country of origin

Kathiresan et al., Myocardial Infarction Genetics Consortium/2009[89]	HapMap CEU	stage 1: MIGen, stage 2: WTCCC 3, German MI Family Study I3, PennCATH, MedStar, stage 3: MI Gene Study/Dortmund Health Study, Verona Heart Study29, Mid-America Heart Institute Study30, Irish Family Study31, German MI Family Study II, and INTERHEART32	Mainly European	stage 1 :2,967/6,042, stage 2: 3,942/7,884,	5,469/10,938	MI	around 2.5 million	3 stages: the first stage conducted GWAS, 1441 SNPs were taken into stage 2 for silico analysis, 33 SNPs for replication	age, gender, study site, stage2 and 3: age, gender
Gudbjartsson et al./2009	No imputation	Iceland patients, six sample sets of European ancestry	Mainly European	2,625/38,875	3,925/10,921	MI	15	Examined the association of the 15 SNPs identified through the blood-eosinophil-counts scan in chronic obstructive pulmonary disease (COPD) and MI using Icelandic GWA scan data and replicated in European dataset	
Helgadottir et al./2007[90]	No imputation	Iceland patients	Mainly European	1,607/8,335	665/4,198		305,953	first conducted GWAS in 8,335 individuals, then further explored 3 SNPs that had p-values closer to genome-wide significance	
Samani et al./2007[91]	No imputation	Discovery: WTCCC, replication: German MI family study	Mainly European	1,926/4,864	875/2,519	MI, coronary artery bypass grafting, coronary artery angioplasty	377,857	GWAS in WTCCC, then replicated the strong associations in the German MI study.	
McPherson et al./2007[92]	No imputation	OHS 1-3, ARIC, CCHS, DHS,	Mainly European	322/634,	replication 1: 311/637, replication 2: 1,347/10,401, validation: 2,326/12,735	Myocardial infarction (MI), coronary artery bypass grafting, or coronary artery angioplasty	72,864	Conducted discovery GWAS and then replicated separately in two cohorts, Validated 2 significant SNPs in additional cohorts (CCHS, DHS and OHS-3).	
Shiffman et al./2006[93]	No imputation	CSC, USCF	Mainly European	study 1: 1200/1,462, study 2: 434/938	1200, 434, 187/621	early-onset MI	11,647	Tested for associations in study 1 using multiple logistic regression and then in study 2 to reduce false positive. Then replicated for early-onset MI in study 3.	replication: age, smoking, diabetes, dyslipidaemia, hypertension, BMI

MI: Myocardial Infarction, CHD: Coronary Artery Disease, CVD: Cardiovascular Disease, BMI: Body Mass Index, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, OR: Odds Ratio, RR: Risk Ratio, HR: Hazard Ratio, CI: Confidence interval, PRS: Polygenic Risk Score, PC: Principal Components, CC4D: CARDIoGRAMplusC4D consortium, GWAS: Genome-wide Association Study, AUC: Area under the ROC curve, NRI: Net reclassification Index, P+T: Pruning and Thresholding, SD: standard deviation, TC: Total cholesterol, HDL-C: High Density Lipoprotein Cholesterol, LDL-C: Low Density Lipoprotein Cholesterol, TDI: Townsend Deprivation Index, IPAQ: The international Physical Activity Questionnaire, CABG: Coronary Artery bypass graft, PTCA: Percutaneous Transluminal coronary angioplasty, PCI: Percutaneous Coronary Intervention, SNP: Single Nucleotide Polymorphisms, BG: Blood Glucose, TG: Triglycerides, PCE: pooled cohort equation, SD: Standard deviation, FRS: Framingham Risk Score, MDCS: Malmo Diet and Cancer Study, FE: Fixed effect, RE: Random effect, ASCVD: atherosclerotic cardiovascular disease

Table S4: Equation parameter of SCORE2 (before age 70) for estimation of uncalibrated 10-year risk of CVD

Risk factor (units)	Transformation equation†	Sex coefficient	
		Male	Female
Age (yrs)	cage = (age - 60)/5	0.3742	0.4648
Smoking (current vs. other)	Current = 1, other = 0	0.6012	0.7744
Systolic blood pressure (SBP, mm Hg)	csbp = (sbp - 120)/20	0.2777	0.3131
Diabetes* (yes vs. no)	Yes = 1, no = 0	0.6457	0.8096
Total cholesterol (mmol/L)	ctchol = (tchol - 6)/1	0.1458	0.1002
HDL cholesterol (mmol/L)	chdl = (hdl - 1.3)/0.5	-0.2698	-0.2606
Smoking x age interaction	cage x smoking	-0.0755	-0.1088
SBP x age interaction	cage x csbp	-0.0255	-0.0277
Total cholesterol x age interaction	cage x ctchol	-0.0281	-0.0226
HDL cholesterol x age interaction	cage x chdl	0.0426	0.0613
Diabetes* x age interaction	cage x diabetes	-0.0983	-0.1272
Baseline survival		0.9605	0.9776

*Excluded in SCORE2-standard computation

† Transformation equation used to transform each risk factor (or risk factor combination). For instance, cage is the transformation for age, which can be calculated by using the equation cage=(age-60)/5 and no transformation is required for smoking.

Linear predictor of SCORE2 =∑ (risk factor transformation x sex coefficient)

10-year risk estimation (un-calibrated) = 1-sex specific baseline survival $^{\exp(\text{linear predictor})}$ (see **Table S5** for calibration factor)

Table S5: Region specific scaling factor of SCORE2 (before age 70)

	Male		Female	
Risk region	**Scale1**	**Scale2**	**Scale1**	**Scale2**
Low risk region	-0.5699	0.7476	-0.7380	0.7019
Moderate risk region	-0.1565	0.8009	-0.3143	0.7701
High risk region	0.3207	0.9360	0.5710	0.9369
Very high risk region	0.5836	0.8294	0.9412	0.8329

Calibration of risk estimate according to region specific scaling factors

Calibrated 10-year risk = 1-exp(-exp(scale1 + scale2 x ln(-ln(1-un-calibrated 10-yr risk))

Table S6: Equation parameter of SCORE2-OP (above age 70) for estimation of uncalibrated 10-year risk of CVD

Risk factor (units)	Transformation equation†	Sex coefficient	
		Male	Female
Age (yrs)	cage = (age - 73)	0.0634	0.0789
Smoking (current vs. other)	Current = 1, other = 0	0.3524	0.4921
Systolic blood pressure (SBP, mm Hg)	csbp = (sbp - 150)	0.0094	0.0102
Diabetes* (yes vs. no)	Yes = 1, no = 0	0.4245	0.6010
Total cholesterol (mmol/L)	ctchol = (tchol - 6)	0.0850	0.0605
HDL cholesterol (mmol/L)	chdl = (hdl - 1.4)	-0.3564	-0.3040
Smoking x age interaction	cage x smoking	-0.0247	-0.0255
SBP x age interaction	cage x csbp	-0.0005	-0.0004
Total cholesterol x age interaction	cage x ctchol	0.0073	-0.0009
HDL cholesterol x age interaction	cage x chdl	0.0091	0.0154
Diabetes* x age interaction	cage x diabetes	-0.0174	-0.0107
Sex-specific mean predictor		0.0929	0.229
Baseline survival		0.7576	0.8082

*Excluded in SCORE2-standard computation

† Transformation equation used to transform each risk factor (or risk factor combination). For instance, cage is the transformation for age, which can be calculated using the equation cage=(age-73) and no transformation is required for smoking.

Linear predictor of SCORE2-OP = (∑risk factor transformation x coefficient) – sex-specific mean predictor

10-year risk estimation (un-calibrated) = 1-sex specific baseline survival $^{\exp(\text{linear predictor})}$ (see **Table S7** for calibration factor)

Table S7: Region specific scaling factor of SCORE2-OP (above age 70)

	Male		Female	
Risk region	**Scale1**	**Scale2**	**Scale1**	**Scale2**
Low risk region	-0.34	1.19	-0.52	1.01
Moderate risk region	0.01	1.25	-0.1	1.1
High risk region	0.08	1.15	0.38	1.09
Very high risk region	0.05	0.7	0.38	0.69

Calibration of risk estimate according to region specific scaling factors

Calibrated 10-year risk = 1-exp(-exp(scale1 + scale2 x ln(-ln(1-un-calibrated 10-yr risk))

Table S8: Equation Parameters of the Pooled Cohort Equations for Estimation of 10-Year Risk

	Women			Men		
	Coefficient	**Individual Example Value**	**Coefficient × Value**	**Coefficient**	**Individual Example Value**	**Coefficient × Value†**
(Example: 55 years of age with total cholesterol 213 mg/dL, HDL–C 50 mg/dL, untreated systolic BP 120 mm Hg, nonsmoker, and without diabetes)						
Ln Age (y)	–29.799	4.01	–119.41	12.344	4.01	49.47
Ln Age, Squared	4.884	16.06	78.44	N/A	N/A	N/A
Ln Total Cholesterol (mg/dL)	13.540	5.36	72.59	11.853	5.36	63.55
Ln Age×Ln Total Cholesterol	–3.114	21.48	–66.91	–2.664	21.48	–57.24
Ln HDL–C (mg/dL)	–13.578	3.91	–53.12	–7.990	3.91	–31.26
Ln Age×Ln HDL–C	3.149	15.68	49.37	1.769	15.68	27.73
Log Treated Systolic BP (mm Hg)	2.019	–	–	1.797	–	–
Log Untreated Systolic BP (mm Hg)	1.957	4.79	9.37	1.764	4.79	8.45
Current Smoker (1=Yes, 0=No)	7.574	0	0	7.837	0	0
Log Age×Current Smoker	–1.665	0	0	–1.795	0	0
Diabetes (1=Yes, 0=No)	0.661	0	0	0.658	0	0
Individual Sum			–29.67			60.69
Mean (Coefficient× Value)			–29.18			61.18
Baseline Survival			0.9665			0.9144
Estimated 10-Y Risk for hard ASCVD			2.1%			5.3%

Linear predictor= =∑ coefficient x value - mean (coefficient x value)

10-year risk =1- sex specific baseline survival $^{\exp(\text{linear predictor})}$

8.1 References

1. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. “Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention”. *J Am Coll Cardiol* 2018;72(16):pp. 1883–1893.
2. Tikkanen E, Gustafsson S, and Ingelsson E. “Associations of Fitness, Physical Activity, Strength, and Genetic Risk With Cardiovascular Disease: Longitudinal Analyses in the UK Biobank Study”. *Circulation* 2018;137(24):pp. 2583–2591.
3. Ntalla I, Kanoni S, Zeng L, Giannakopoulou O, Danesh J, Watkins H, et al. “Genetic Risk Score for Coronary Disease Identifies Predispositions to Cardiovascular and Noncardiovascular Diseases”. *Journal of the American College of Cardiology* 2019;73(23):pp. 2932–2942.
4. Patel RS, Denaxas S, Howe LJ, Eggo RM, Shah AD, Allen NE, et al. “Reproducible disease phenotyping at scale: Example of coronary artery disease in UK Biobank”. *PLoS One* 2022;17(4):e0264828.
5. Howe LJ, Dudbridge F, Schmidt AF, Finan C, Denaxas S, Asselbergs FW, et al. “Polygenic risk scores for coronary artery disease and subsequent event risk amongst established cases”. *Human Molecular Genetics* 2020;29(8):pp. 1388–1395.
6. Yuan S, Huang X, Ma W, Yang R, Xu F, Han D, et al. “Associations of HDL-C/LDL-C with myocardial infarction, all-cause mortality, haemorrhagic stroke and ischaemic stroke: A longitudinal study based on 384 093 participants from the UK Biobank”. *Stroke and Vascular Neurology* 2022;(no pagination).
7. Fan M, Sun D, Zhou T, Heianza Y, Lv J, Li L, et al. “Sleep patterns, genetic susceptibility, and incident cardiovascular disease: a prospective study of 385 292 UK biobank participants”. *Eur Heart J* 2020;41(11):pp. 1182–1189.
8. Zhang H, Zeng Y, Yang H, Hu Y, Chen W, Ying Z, et al. “Familial factors, diet, and risk of cardiovascular disease: a cohort analysis of the UK Biobank”. *The American journal of clinical nutrition* 2021;114(5):pp. 1837–1846.
9. Huang Y, Hui Q, Gwinn M, Hu YJ, Quyyumi AA, Vaccarino V, et al. “Sexual Differences in Genetic Predisposition of Coronary Artery Disease”. *Circulation. Genomic and Precision Medicine* 2021;14(1):e003147.
10. Carter AR, Harrison S, Gill D, Davey Smith G, Taylor AE, Howe LD, et al. “Educational attainment as a modifier for the effect of polygenic scores for cardiovascular risk factors: Cross-sectional and prospective analysis of UK Biobank”. *International Journal of Epidemiology* 2022;51(3):pp. 885–897.
11. Daghlas I, Dashti HS, Lane J, Aragam KG, Rutter MK, Saxena R, et al. “Sleep Duration and Myocardial Infarction”. *Journal of the American College of Cardiology* 2019;74(10):pp. 1304–1314.
12. Tamlander M, Mars N, Pirinen M, FinnGen, Widen E, and Ripatti S. “Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes”. *Communications Biology* 2022;5(1):p. 158.
13. Huang Y, Hui Q, Gwinn M, Hu YJ, Quyyumi AA, Vaccarino V, et al. “Interaction between genetics and smoking in determining risk of coronary artery diseases”. *Genetic Epidemiology* 2022;46(3-4):pp. 199–212.
14. Heianza Y, Zhou T, Sun D, Hu FB, Manson JE, and Qi L. “Genetic susceptibility, plant-based dietary patterns, and risk of cardiovascular disease”. *American Journal of Clinical Nutrition* 2020;112(1):pp. 220–228.
15. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, et al. “Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease”. *Nature Genetics* 2020;52(11):pp. 1169–1177.
16. Verweij N, Eppinga RN, Hagemeijer Y, and van der Harst P. “Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure”. *Scientific Reports* 2017;7(1):p. 2761.

17. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, et al. “Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers”. *Nature Medicine* 2020;26(4):pp. 549–557.

18. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. *Nat Genet* 2018;50(9):pp. 1219–1224.

19. Patel AP, Wang M, Ruan Y, Koyama S, Clarke SL, Yang X, et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. *Nature Medicine* 2023;29(7):pp. 1793–1803.

20. Ye Y, Chen X, Han J, Jiang W, Natarajan P, and Zhao H. “Interactions Between Enhanced Polygenic Risk Scores and Lifestyle for Cardiovascular Disease, Diabetes, and Lipid Levels”. *Circulation. Genomic and Precision Medicine* 2021;14(1):e003128.

21. Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, et al. “Large-scale genome-wide association study of coronary artery disease in genetically diverse populations”. *Nature Medicine* 2022;28(8):pp. 1679–1692.

22. Riveros-Mckay F, Weale ME, Moore R, Selzam S, Krapohl E, Sivley RM, et al. “Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction”. *Circ Genom Precis Med* 2021;14(2):e003304.

23. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. “Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease”. *New England Journal of Medicine* 2016;375(24):pp. 2349–2358.

24. Iribarren C, Lu M, Jorgenson E, Martinez M, Lluis-Ganella C, Subirana I, et al. “Clinical Utility of Multimarker Genetic Risk Scores for Prediction of Incident Coronary Heart Disease: A Cohort Study Among Over 51 000 Individuals of European Ancestry”. *Circulation. Cardiovascular Genetics* 2016;9(6):pp. 531–540.

25. Mega JL, Stitziel NO, Smith JG, Chasman DI, Caulfield M, Devlin JJ, et al. “Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials”. *Lancet* 2015;385(9984):pp. 2264–2271.

26. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. “Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 636–645.

27. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. “A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses”. *Lancet* 2010;376(9750):pp. 1393–400.

28. Tikkanen E, Havulinna AS, Palotie A, Salomaa V, and Ripatti S. “Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease”. *Arteriosclerosis, Thrombosis and Vascular Biology* 2013;33(9):pp. 2261–2266.

29. Tada H, Melander O, Louie JZ, Catanese JJ, Rowland CM, Devlin JJ, et al. “Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history”. *Eur Heart J* 2016;37(6):pp. 561–7.

30. Oni-Orisan A, Haldar T, Cayabyab MAS, Ranatunga DK, Hoffmann TJ, Iribarren C, et al. “Polygenic Risk Score and Statin Relative Risk Reduction for Primary Prevention of Myocardial Infarction in a Real-World Population”. *Clinical Pharmacology & Therapeutics* 2022;112(5):pp. 1070–1078.

31. King A, Wu L, Deng HW, Shen H, and Wu C. “Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease”. *BMC Medicine* 2022;20(1):p. 385.

32. Di Narzo A, Frades I, Crane HM, Crane PK, Hulot JS, Kasarskis A, et al. “Meta-analysis of sample-level dbGaP data reveals novel shared genetic link between body height and Crohn’s disease”. *Human Genetics* 2021;140(6):pp. 865–877.

33. Huang QQ, Sallah N, Dunca D, Trivedi B, Hunt KA, Hodgson S, et al. “Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals”. *Nature communications* 2022;13(1):p. 4664.

34. Gola D, Erdmann J, Müller-Myhsok B, Schunkert H, and König IR. “Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status”. *Genet Epidemiol* 2020;44(2):pp. 125–138.

35. Morris RW, Cooper JA, Shah T, Wong A, Drenos F, Engmann J, et al. “Marginal role for 53 common genetic variants in cardiovascular disease prediction”. *Heart* 2016;102(20):pp. 1640–7.

36. Gola D, Erdmann J, Läll K, Mägi R, Müller-Myhsok B, Schunkert H, et al. “Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease”. *Circ Genom Precis Med* 2020;13(6):e002932.

37. Brautbar A, Pompeii LA, Dehghan A, Ngwa JS, Nambi V, Virani SS, et al. “A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies”. *Atherosclerosis* 2012;223(2):pp. 421–426.

38. Goldstein BA, Knowles JW, Salfati E, Ioannidis JP, and Assimes TL. “Corrigendum: Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example”. *Front Genet* 2015;6:p. 231.

39. Goldstein BA, Yang L, Salfati E, and Assimes TL. “Contemporary Considerations for Constructing a Genetic Risk Score: An Empirical Approach”. *Genetic Epidemiology* 2015;39(6):pp. 439–45.

40. Tragante V, Doevendans PA, Nathoe HM, Graaf Y van der, Spiering W, Algra A, et al. “The impact of susceptibility loci for coronary artery disease on other vascular domains and recurrence risk”. *Eur Heart J* 2013;34(37):pp. 2896–904.

41. Wang M, Menon R, Mishra S, Patel AP, Chaffin M, Tanneeru D, et al. “Validation of a Genome-Wide Polygenic Score for Coronary Artery Disease in South Asians”. *J Am Coll Cardiol* 2020;76(6):pp. 703–714.

42. Saad M, El-Menyar A, Kunji K, Ullah E, Al Suwaidi J, and Kullo IJ. “Validation of Polygenic Risk Scores for Coronary Heart Disease in a Middle Eastern Cohort Using Whole Genome Sequencing”. *Circulation* 2022;Genomic and precision medicine.e003712.

43. Powell KL, Stephens SR, and Stephens AS. “Cardiovascular risk factor mediation of the effects of education and Genetic Risk Score on cardiovascular disease: a prospective observational cohort study of the Framingham Heart Study”. *BMJ Open* 2021;11(1):e045210.

44. De Vries PS, Kavousi M, Ligthart S, Uitterlinden AG, Hofman A, Franco OH, et al. “Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: The Rotterdam Study”. *International Journal of Epidemiology* 2015;44(2):pp. 682–688.

45. Abraham G, Havulinna AS, Bhalala OG, Byars SG, De Livera AM, Yetukuri L, et al. “Genomic prediction of coronary heart disease”. *Eur Heart J* 2016;37(43):pp. 3267–3278.

46. Ge T, Chen CY, Ni Y, Feng YA, and Smoller JW. “Polygenic prediction via Bayesian regression and continuous shrinkage priors”. *Nature communications* 2019;10(1):p. 1776.

47. Manikpurage HD, Eslami A, Perrot N, Li Z, Couture C, Mathieu P, et al. “Polygenic Risk Score for Coronary Artery Disease Improves the Prediction of Early-Onset Myocardial Infarction and Mortality in Men”. *Circ Genom Precis Med* 2021;14(6):e003452.

48. Goodman MO, Cade BE, Shah NA, Huang T, Dashti HS, Saxena R, et al. “Pathway-Specific Polygenic Risk Scores Identify Obstructive Sleep Apnea-Related Pathways Differentially Moderating Genetic Susceptibility to Coronary Artery Disease”. *Circulation. Genomic and Precision Medicine* 2022;15(5):e003535.

49. Isgut M, Sun J, Quyyumi AA, and Gibson G. “Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later”. *Genome Medicine* 2021;13(1):p. 13.

50. Zaccardi F, Timmins IR, Goldney J, Dudbridge F, Dempsey PC, Davies MJ, et al. “Self-reported walking pace, polygenic risk scores and risk of coronary artery disease in UK biobank”. *Nutrition, Metabolism and Cardiovascular Diseases* 2022;32(11):pp. 2630–2637.

51. Kim Y, Yeung SLA, Sharp SJ, Wang M, Jang H, Luo S, et al. “Genetic susceptibility, screen-based sedentary activities and incidence of coronary heart disease”. *BMC Medicine* 2022;20(1):p. 188.

52. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, et al. “Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses”. *PLoS Medicine* 2021;18(1) (no pagination).

53. Ramírez J, Duijvenboden S van, Young WJ, Tinker A, Lambiase PD, Orini M, et al. “Prediction of Coronary Artery Disease and Major Adverse Cardiovascular Events Using Clinical and Genetic Risk Scores for Cardiovascular Risk Factors”. *Circ Genom Precis Med* 2022;15(5):e003441.

54. Livingstone KM, Abbott G, Bowe SJ, Ward J, Milte C, and McNaughton SA. “Diet quality indices, genetic risk and risk of cardiovascular disease and mortality: a longitudinal analysis of 77 004 UK Biobank participants”. *BMJ Open* 2021;11(4):e045362.

55. Livingstone KM, Abbott G, Ward J, and Bowe SJ. “Unhealthy Lifestyle, Genetics and Risk of Cardiovascular Disease and Mortality in 76,958 Individuals from the UK Biobank Cohort Study”. *Nutrients* 2021;13(12).

56. Rostami S, Hoff M, Dalen H, Hveem K, and Videm V. “Genetic risk score associations for myocardial infarction are comparable in persons with and without rheumatoid arthritis: the population-based HUNT study”. *Scientific reports* 2020;10(1):p. 20416.

57. Aragam KG, Dobbyn A, Judy R, Chaffin M, Chaudhary K, Hindy G, et al. “Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease”. *Journal of the American College of Cardiology* 2020;75(22):pp. 2769–2780.

58. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, et al. “Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups”. *American Journal of Human Genetics* 2020;106(5):pp. 707–716.

59. Liang F, Liu F, Li J, Huang K, Yang X, Chen S, et al. “Genetic risk modifies the effect of long-term fine particulate matter exposure on coronary artery disease”. *Environment International* 2022;170 (no pagination).

60. Hindy G, Aragam KG, Ng K, Chaffin M, Lotta LA, Baras A, et al. “Genome-Wide Polygenic Score, Clinical Risk Factors, and Long-Term Trajectories of Coronary Artery Disease”. *Arteriosclerosis, Thrombosis and Vascular Biology* 2020;40(11):pp. 2738–2746.

61. Martikainen P, Korhonen K, Jelenkovic A, Lahtinen H, Havulinna A, Ripatti S, et al. “Joint association between education and polygenic risk score for incident coronary heart disease events: a longitudinal population-based study of 26 203 men and women”. *Journal of epidemiology and community health.* 2021;06.

62. Hindy G, Wiberg F, Almgren P, Melander O, and Orho-Melander M. “Polygenic Risk Score for Coronary Heart Disease Modifies the Elevated Risk by Cigarette Smoking for Disease Incidence”. *Circulation: Genomic and Precision Medicine* 2018;11(1):E001856.

63. Fritz J, Shiffman D, Melander O, Tada H, and Ulmer H. “Metabolic Mediators of the Effects of Family History and Genetic Risk Score on Coronary Heart Disease-Findings From the Malmo Diet and Cancer Study”. *Journal of the American Heart Association* 2017;6(3):p. 20.

64. Svensson T, Kitlinski M, Engström G, and Melander O. “A genetic risk score for CAD, psychological stress, and their interaction as predictors of CAD, fatal MI, non-fatal MI and cardiovascular death”. *PLoS One* 2017;12(4):e0176029.

65. Iribarren C, Lu M, Jorgenson E, Martínez M, Lluis-Ganella C, Subirana I, et al. “Weighted Multi-marker Genetic Risk Scores for Incident Coronary Heart Disease among Individuals of African, Latino and East-Asian Ancestry”. *Sci Rep* 2018;8(1):p. 6853.

66. Wünnemann F, Sin Lo K, Langford-Avelar A, Busseuil D, Dubé MP, Tardif JC, et al. “Validation of Genome-Wide Polygenic Risk Scores for Coronary Artery Disease in French Canadians”. *Circ Genom Precis Med* 2019;12(6):e002481.

67. Hasbani NR, Ligthart S, Brown MR, Heath AS, Bebo A, Ashley KE, et al. “American Heart Association’s Life’s Simple 7: Lifestyle Recommendations, Polygenic Risk, and Lifetime Risk of Coronary Heart Disease”. *Circulation* 2022;145(11):pp. 808–818.

68. Menniti G, Paquet C, Han HY, Dube L, and Nielsen DE. “Multiscale Risk Factors of Cardiovascular Disease: CLSA Analysis of Genetic and Psychosocial Factors”. *Frontiers in Cardiovascular Medicine* 2021;8:p. 599671.

69. Hamad R, Glymour MM, Calmasini C, Nguyen TT, Walter S, and Rehkopf DH. “Explaining the Variance in Cardiovascular Disease Risk Factors: A Comparison of Demographic, Socioeconomic, and Genetic Predictors”. *Epidemiology* 2022;33(1):pp. 25–33.

70. Mosley JD, Gupta DK, Tan J, Yao J, Wells QS, Shaffer CM, et al. “Predictive Accuracy of a Polygenic Risk Score Compared with a Clinical Risk Score for Incident Coronary Heart Disease”. *JAMA - Journal of the American Medical Association* 2020;323(7):pp. 627–635.

71. Khan SS, Page C, Wojdyla DM, Schwartz YY, Greenland P, and Pencina MJ. “Predictive Utility of a Validated Polygenic Risk Score for Long-Term Risk of Coronary Heart Disease in Young and Middle-Aged Adults”. *Circulation* 2022;146(8):pp. 587–596.

72. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner K, et al. “FinnGen: Unique genetic insights from combining isolated population and national health register data”. *medRxiv* 2022:p. 2022.03.03.22271360.

73. Ishigaki K, Akiyama M, Kanai M, Takahashi A, Kawakami E, Sugishita H, et al. “Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases”. *Nature Genetics* 2020;52(7):pp. 669–679.

74. van der Harst P and Verweij N. “Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease”. *Circ Res* 2018;122(3):pp. 433–443.

75. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. “Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies”. *Nature Genetics* 2018;50(9):pp. 1335–1341.

76. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, et al. “Association analyses based on false discovery rate implicate new loci for coronary artery disease”. *Nature Genetics* 2017;49(9):pp. 1385–1391.

77. Webb TR, Erdmann J, Stirrups KE, Stitziel NO, Masca NG, Jansen H, et al. “Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease”. *J Am Coll Cardiol* 2017;69(7):pp. 823–836.

78. Howson JMM, Zhao W, Barnes DR, Ho W.-K, Young R, Paul DS, et al. “Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms”. *Nature Genetics* 2017;49(7):pp. 1113–1119.

79. Klarin D, Zhu QM, Emdin CA, Chaffin M, Horner S, McMillan BJ, et al. “Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease”. *Nature Genetics* 2017;49(9):pp. 1392–1397.

80. Stitziel NO, Stirrups KE, Masca NG, Erdmann J, Ferrario PG, König IR, et al. “Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease”. *N Engl J Med* 2016;374(12):pp. 1134–44.

81. Nikpay M, Goel A, Won H.-H, Hall LM, Willenborg C, Kanoni S, et al. “A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease”. *Nature genetics* 2015;47(10):pp. 1121–1130.

82. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. “Large-scale association analysis identifies new risk loci for coronary artery disease”. *Nature Genetics* 2013;45(1):pp. 25–33.

83. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. “Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease”. *Nat Genet* 2011;43(4):pp. 333–8.

84. Coronary Artery Disease (C4D) Genetics Consortium. “A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease”. *Nat Genet* 2011;43(4):pp. 339–44.

85. Wang F, Xu CQ, He Q, Cai JP, Li XC, Wang D, et al. “Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population”. *Nat Genet* 2011;43(4):pp. 345–9.

86. Aoki A, Ozaki K, Sato H, Takahashi A, Kubo M, Sakata Y, et al. “SNPs on chromosome 5p15.3 associated with myocardial infarction in Japanese population”. *J Hum Genet* 2011;56(1):pp. 47–51.

87. Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, Hall AS, et al. “New susceptibility locus for coronary artery disease on chromosome 3q22.3”. *Nat Genet* 2009;41(3):pp. 280–2.

88. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, et al. “Genetic variants associated with Lp(a) lipoprotein level and coronary disease”. *N Engl J Med* 2009;361(26):pp. 2518–28.

89. Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, et al. “Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants”. *Nat Genet* 2009;41(3):pp. 334–41.

90. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. “A common variant on chromosome 9p21 affects the risk of myocardial infarction”. *Science* 2007;316(5830):pp. 1491–3.

91. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. “Genomewide Association Analysis of Coronary Artery Disease”. *New England Journal of Medicine* 2007;357(5):pp. 443–453.

92. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. “A common allele on chromosome 9 associated with coronary heart disease”. *Science* 2007;316(5830):pp. 1488–91.

93. Shiffman D, Rowland CM, Louie JZ, Luke MM, Bare LA, Bolonick JI, et al. “Gene variants of VAMP8 and HNRPUL1 are associated with early-onset myocardial infarction”. *Arterioscler Thromb Vasc Biol* 2006;26(7):pp. 1613–8.