Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Weighted PRS for Ischemic Heart Disease: Candidate-Gene Associations and an Evaluation of Epistasis and Machine-Learning Approaches

Anam Ijaz

anamijaz2021@gmail.com

Queen Mary University of London https://orcid.org/0009-0002-5933-9802

sara aslam

university of the punjab

NA Shabana

University of the Punjab https://orcid.org/0000-0002-6617-1588

Article

Keywords: ischemic heart disease, candidate-gene polymorphisms, polygenic risk score, machine learning, Pakistani population, lipid profile

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10560334/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no duality of interest

Abstract

Ischemic heart disease (IHD) carries a substantial polygenic component, but candidate-gene and polygenic-score evidence remains scarce in South Asians. This Pakistani case–control study evaluated eight candidate variants, their aggregation into conventional and epistasis-weighted polygenic scores, and machine-learning classifiers. 612 participants (306 IHD cases, 306 controls) were genotyped for *ADAMTS7* rs3825807, *ADAMTS13* rs2301612, *APOE* rs769452, *AGT* rs699, *APOB* rs676210, *MMP9* rs3918242, *MTHFR* rs1801133 and *ZC3HC1* rs11556924, with covariate-adjusted single-variant testing. Unweighted and effect-size-weighted polygenic risk scores (PRS) and a three-tier epistasis-weighted candidate-gene PRS (EW-cgPRS) were built, a lipid quantitative-trait scan performed, and four machine-learning classifiers benchmarked. Six of eight variants were nominally associated with IHD; four survived Bonferroni correction across 40 inheritance-model tests: *ADAMTS13* (additive OR 6.85, 95% CI 3.57–13.12), *ADAMTS7* (recessive OR 6.00, 2.69–13.40), *APOE* (dominant OR 4.44, 2.12–9.26) and *MMP9* (recessive OR 14.49, 3.40–61.80). The weighted PRS separated cases from controls (cross-validated AUC 0.861), with a monotonic case-prevalence gradient across strata (top- versus bottom-tertile OR 59.2). Neither epistatic nor functional-annotation weighting improved on the classic score; the interaction tier's apparent superiority reflected covariate access, not epistasis. All four classifiers achieved near-perfect discrimination (AUC 0.98–0.999), traced to lipid-based ascertainment of controls rather than genetic signal. Several candidate variants are associated with IHD in this population, and a simple weighted score captures a reproducible genetic gradient; epistatic or functional enrichment adds nothing. The near-perfect machine-learning performance reflects study design, not clinical utility, and the largest single-variant effects warrant cautious interpretation given the modest sample size, Hardy–Weinberg deviations, and single-population design.

Introduction

IHD is the leading cause of death worldwide, causing 8.99 million deaths and 188.4 million disability-adjusted life-years in 2021 — the largest contributor to non-communicable health loss[1,2]. South Asians develop coronary disease roughly a decade earlier than Europeans (INTERHEART median age at first myocardial infarction: 52 vs 62 years)[3,4]. Pakistan is disproportionately affected: cardiovascular incidence (918 vs 684/100,000) and mortality (358 vs 240/100,000) exceed global averages, and IHD caused ~ 54% of cardiovascular deaths there in 2019[5,6], against a background diabetes prevalence of ~ 31% and hypertension of ~ 37%. Identifying at-risk individuals before a first clinical event is therefore a priority.

IHD arises from the interaction of modifiable environmental exposures with a substantial inherited component. Twin studies place the heritability of fatal coronary artery disease at approximately 40–60%[7], and genome-wide association studies (GWAS) have now implicated more than 240 independent susceptibility loci[8]. Even so, these loci jointly explain only about 15.5% of the liability-scale heritability of the disease[8], leaving a considerable residual genetic burden unaccounted for. This gap is compounded

by ascertainment bias in the evidence base: the overwhelming majority of cardiovascular GWAS and polygenic score derivations have been conducted in European-ancestry cohorts, and the resulting scores are several-fold less accurate when applied to other ancestries because effect sizes, allele frequencies and linkage-disequilibrium structure do not transfer reliably[9]. In Pakistan the issue is further sharpened by a consanguinity rate of approximately 58%, which is likely to concentrate genetic risk in ways that European-derived scores do not capture[5]. Well-characterised candidate-gene studies in under-represented South Asian populations therefore remain informative, both for local risk stratification and for testing the generalisability of established loci.

Atherogenesis offers a coherent framework for candidate-gene selection. The process integrates several partly independent axes: the delivery and clearance of atherogenic lipoproteins; extracellular-matrix remodelling and vascular smooth-muscle-cell proliferation within the developing lesion; thrombosis at the site of plaque erosion or rupture; and vasoactive and metabolic perturbations that amplify endothelial dysfunction. Because these axes act in concert rather than in isolation, a variant panel sampling across them is more likely to capture cumulative susceptibility than one confined to a single pathway. On this rationale the present study examined eight single-nucleotide polymorphisms (SNPs) spanning lipoprotein metabolism (*APOB*, *APOE*), vascular remodelling and plaque stability (*ADAMTS7*, *MMP9*, *ZC3HC1*), haemostasis (*ADAMTS13*), blood-pressure regulation (*AGT*) and homocysteine metabolism (*MTHFR*).

Apolipoprotein B-100, encoded by *APOB*, is the principal structural protein of very-low-density and low-density lipoprotein particles and the ligand recognised by the LDL receptor. The rs676210 variant is a missense substitution (c.8216C > T, p.Pro2739Leu) in this large scaffold protein. It was identified in a genome-wide scan as the only coding variant robustly associated with circulating oxidised LDL concentrations, although in that study it was not itself associated with angiographic coronary disease or myocardial infarction after adjustment[10]. Subsequent work has linked the variant to lipid levels and to the magnitude of LDL-cholesterol lowering achieved with atorvastatin, suggesting a pharmacogenetic as well as an aetiological role[11]. *APOE* occupies an analogous position in remnant-lipoprotein clearance, and the *APOE* region is among the strongest known genetic determinants of circulating lipids. The rs769452 variant is a rare missense substitution (p.Leu46Pro) that is predicted and experimentally shown to destabilise the apoE protein and to impair its lipid-binding properties, although it is currently classified as a variant of uncertain significance and is distinct from the classical ε2/ε3/ε4 haplotype defined by rs429358 and rs7412[12].

A second group of variants acts on the vessel wall rather than on the lipoprotein pool. *ADAMTS7* encodes a secreted metalloproteinase that cleaves thrombospondin-5 and promotes vascular smooth-muscle-cell migration during arterial remodelling; knockout of *Adamts7* reduces atherosclerotic lesion formation in hyperlipidaemic mice[13]. The rs3825807 polymorphism (p.Ser214Pro) lies in the prodomain and affects its cleavage during enzyme maturation. Notably, the direction of effect is opposite to the intuitive expectation: the Pro-encoding G allele is hypomorphic, reducing ADAMTS7 secretion and

smooth-muscle-cell migration, and is associated with *protection* from coronary disease, whereas the Ser-encoding A allele is the risk allele[14,15]. In a Slovenian cohort with type 2 diabetes the AA genotype was associated with myocardial infarction under recessive and codominant models[16]. *MMP9* encodes gelatinase B, which degrades the collagenous fibrous cap of the plaque. The rs3918242 polymorphism (C-1562T) lies in the promoter: the T allele weakens binding of a transcriptional repressor to a regulatory element upstream of the gene, generating a high-activity promoter genotype with elevated basal and cytokine-inducible transcription[17]. An updated meta-analysis of 37 case-control studies comprising 13,035 cases and 11,372 controls reported a significant allelic association with coronary artery disease (odds ratio 1.34, 95% CI 1.20–1.50), with a stronger effect for myocardial infarction in Asian populations[18]. *ZC3HC1* encodes NIPA, an F-box protein that targets cyclin-B1 for ubiquitin-mediated degradation during interphase and thereby regulates mitotic entry. The rs11556924 variant (p.Arg363His) is the lead — and, unusually, the sole — signal at the 7q32.2 coronary locus; the common Arg363 allele confers approximately a 9% increase in disease risk per copy, while the His363 form slows cell proliferation[19,20].

The remaining variants act through haemostatic, vasoactive and metabolic mechanisms. *ADAMTS13*, on chromosome 9q34, encodes a plasma metalloproteinase that cleaves ultra-large von Willebrand factor multimers and thereby restrains platelet-mediated thrombus growth; the von Willebrand factor to ADAMTS13 balance has been repeatedly associated with myocardial infarction and arterial thrombotic risk[21]. The rs2301612 variant (p.Gln448Glu) lies in the cysteine-rich domain involved in substrate recognition and is the leading missense determinant of plasma ADAMTS13 concentration in genome-wide analyses[22]. *AGT* encodes angiotensinogen, the rate-limiting substrate of the renin–angiotensin–aldosterone system. The M235T polymorphism (rs699) substitutes threonine for methionine at residue 235 and is associated with higher plasma angiotensinogen, greater angiotensin II generation and hypertension; a meta-analysis of 16 case-control studies gave an allelic odds ratio for coronary artery disease of 1.38 (95% CI 1.15–1.65)[23], and the variant has been examined specifically in angiographically diagnosed Pakistani patients[24]. *MTHFR* encodes methylenetetrahydrofolate reductase, which commits folate to the remethylation of homocysteine to methionine. The C677T variant (rs1801133; p.Ala222Val) lies near the FAD-binding site and yields a thermolabile enzyme retaining approximately 65% of wild-type activity in CT heterozygotes and 30–35% in TT homozygotes[25,26]. The resulting mild hyperhomocysteinaemia impairs nitric oxide bioavailability and promotes endothelial dysfunction. The clinical significance of this variant is, however, contested: while individual angiographic cohorts report strong associations with disease severity[27], a large meta-analysis of 48,175 cases and 67,961 controls designed to avoid publication bias found essentially no overall effect (TT versus CC odds ratio 1.02, 95% CI 0.98–1.07), with any effect appearing confined to populations of low dietary folate[28]. This folate dependence is itself a clear example of gene–environment interaction and is directly relevant to the analytical framework developed below.

Individually, common variants of this kind confer modest effects, and single-SNP testing is underpowered to support clinical risk stratification. Polygenic risk scores address this by aggregating risk alleles into a single quantitative index, either as an unweighted allele count or weighted by locus-specific effect sizes. Such scores rest on two simplifying assumptions: that each variant contributes strictly additively, and that loci act independently of one another and of the environment. Neither is biologically self-evident. Genes acting within a shared pathway may interact epistatically; variants differ systematically in functional consequence, so that a missense substitution and a promoter variant are not equivalent units of risk; and gene–environment interaction is documented for several of the loci examined here, most clearly for *MTHFR* C677T[28]. Machine-learning classifiers can in principle recover non-additive structure without pre-specification, and applying gradient-boosted models to genotype data has improved polygenic prediction for several cardiometabolic traits[29]. Reported gains have nonetheless been inconsistent and generally modest, and additive scores still outperform machine-learning alternatives in some settings, so any claimed improvement requires formal statistical testing on held-out data rather than a favourable point estimate[30].

Against this background, the present case–control study had three objectives. First, to determine the allele and genotype frequencies of eight candidate SNPs in *ADAMTS7*, *ADAMTS13*, *APOE*, *AGT*, *APOB*, *MMP9*, *MTHFR* and *ZC3HC1* in IHD cases and controls, and to test their association with disease status under additive, dominant, recessive and codominant inheritance models with adjustment for conventional risk factors. Second, to construct unweighted and effect-size-weighted polygenic risk scores from this panel, to evaluate their discriminatory performance and their relationship to the serum lipid profile, and to benchmark them against machine-learning classifiers trained on the same genotypic and clinical information. Third, to develop and formally test an epistasis-weighted candidate-gene polygenic risk score (EW-cgPRS), a tiered framework that successively incorporates inter-SNP epistatic structure, variant-level functional annotation and gene–environment interaction terms, in order to establish whether relaxing the additivity and independence assumptions of the conventional score yields a measurable and statistically supported improvement in risk discrimination in this population.

Materials and Methods

Study design and ethical approval

This was a hospital-based, individually unmatched case–control study conducted between February 2024 and July 2024. Ethical approval was granted by the Research Ethics and Biosafety Committee of the Institute of Microbiology and Molecular Genetics (Approval No. D/645/MMG), which serves as the institutional review board for human-subjects research and oversees biosafety protocols. The study was conducted in accordance with the principles of the Declaration of Helsinki, and written informed consent was obtained from every participant.

A total of 612 unrelated adults of Pakistani origin were recruited consecutively from tertiary-care cardiology units in Lahore: 306 patients with newly diagnosed ischemic heart disease (IHD) and 306

disease-free controls. Cases were confirmed by an on-duty consultant cardiologist on the basis of at least two of the following: characteristic electrocardiographic changes; elevated cardiac biomarkers (troponin T or I); a positive exercise stress test; or angiographic coronary stenosis ≥50%, supported by echocardiographic and clinical assessment. Only patients not yet commenced on lipid-lowering or antihypertensive therapy were enrolled, to avoid treatment effects on the lipid profile. Coexisting diabetes, hypertension, smoking or a positive family history were treated as covariates rather than exclusions. Controls were adults with no history or clinical evidence of cardiovascular disease and a normal resting electrocardiogram, drawn from the same catchment population over the same period; family history was deliberately not an exclusion criterion, as excluding on this basis would render the corresponding case–control comparison circular. Participants in either group were excluded for any other cardiovascular disorder, chronic liver or kidney disease, malignancy, autoimmune disease, chronic infection or pregnancy. All samples were prescreened for HIV, hepatitis B and C and syphilis, and reactive samples were excluded.

Because genotype is fixed at conception, reverse causation does not affect the genetic associations reported here; the clinical and biochemical covariates were measured at a single post-diagnosis time point, and associations involving them are interpreted as cross-sectional throughout.

Sample size and statistical power

The achieved sample was fixed by consecutive recruitment, so the calculation below is presented as a design-sensitivity analysis. For the primary log-additive per-allele test, with 306 cases and 306 controls, a two-sided α of 0.05 and 80% power, the minimum detectable per-allele odds ratio is approximately 1.38 at a minor allele frequency (MAF) of 0.40–0.50, rising to 1.40 at MAF 0.30, 1.46 at 0.20, 1.63 at 0.10 and 1.90 at 0.05. Control MAFs in this panel ranged from 0.036 (*AGT* rs699) to 0.389 (*MTHFR* rs1801133), giving minimum detectable odds ratios of approximately 1.38–1.41 for the four commonest variants and 1.6–2.1 for the four rarest. At the effect sizes reported in large meta-analyses of these loci (allelic odds ratios of approximately 1.34–1.38), power is 67–72% for a common variant but only 22–37% at the low allele frequencies observed here for *MMP9* and *AGT*. The study is therefore well powered only for moderate-to-large effects at common variants, and non-significant single-variant results should be read as inconclusive. No formal power calculation was performed for the aggregate score, machine-learning and EW-cgPRS analyses; their precision is instead characterised empirically.

Clinical data and variable definitions

Age, sex, smoking status, diabetes, hypertension and family history of IHD were recorded by trained staff through structured interviews and review of medical records. Diabetes was defined by WHO criteria as fasting glucose ≥126 mg/dL, two-hour postprandial glucose ≥200 mg/dL, prior physician diagnosis or current antidiabetic use. Hypertension was classified by JNC 7 criteria, prior diagnosis or current antihypertensive use, with blood pressure taken as the mean of two consecutive measurements after five minutes seated at rest; it enters the analysis only through this binary classification. Family history

was defined as a confirmed diagnosis in a first-degree relative before age 55 (men) or 65 (women). Age was analysed as a continuous variable and the remaining five as binary indicators.

Blood sampling and biochemical analysis

After a 10–12 hour overnight fast, 5 mL of venous blood was drawn, allowed to clot for 10–20 minutes, and serum separated by centrifugation at 3000 rpm for five minutes and stored at −20 °C. Triglycerides (TG), total cholesterol (TC), VLDL cholesterol (VLDL-C), LDL cholesterol (LDL-C) and HDL cholesterol (HDL-C) were measured in mg/dL on an automated chemistry analyser using enzymatic colorimetric kits (Spectrum Diagnostics), with LDL-C derived by the Friedewald equation. Reference ranges were TG <200, TC <200, VLDL-C 0–40, LDL-C <130 and HDL-C 35–55 mg/dL. Because LDL-C is derived, the five measures are related by near-linear identities (LDL-C ≈ TC − HDL-C − VLDL-C; VLDL-C ≈ TG/5) and are strongly collinear, which constrained their handling in the regression models.

DNA extraction and genotyping

Genomic DNA was extracted from whole blood by the salting-out method, with quality and quantity assessed by 1% agarose gel electrophoresis and spectrophotometry (A260/A280). All eight SNPs — *ADAMTS7* rs3825807, *ADAMTS13* rs2301612, *APOE* rs769452, *AGT* rs699 (M235T), *APOB* rs676210, *MMP9* rs3918242, *MTHFR* rs1801133 (C677T) and *ZC3HC1* rs11556924 — were genotyped by tetra-primer ARMS-PCR; primer sequences, annealing temperatures and amplicon sizes are given in Supplementary Table S1. Products were resolved on 1.5% or 2% agarose gels stained with ethidium bromide and genotypes assigned from band pattern. Ten percent of samples (n = 62) were randomly regenotyped by Sanger sequencing (Macrogen, Korea), with 100% concordance.

Data curation, genotype coding and single-variant analysis

Analyses were performed in Python 3 (pandas, NumPy, SciPy, statsmodels, scikit-learn, imbalanced-learn, XGBoost, LightGBM, SHAP, NetworkX, Matplotlib, seaborn), two-sided throughout at α = 0.05, with a single fixed random seed (42) applied to every stochastic step so the pipeline is fully reproducible. No missing values or duplicate records were present; imputation routines (k-nearest-neighbour, k = 5, for continuous variables; modal imputation for categorical) were retained in the pipeline but were inoperative on this extract.

Genotypes were recoded as additive allele doses (0, 1, 2), matched by allele set so that allele order did not affect classification. Reference alleles were assigned as the more frequent allele in the sample for six variants. For each variant the direction of association was determined by unadjusted logistic regression and the dose inverted where the odds ratio fell below unity, so that every variant contributed a risk-increasing dose downstream; because this alignment uses the same observations later used for testing, the leakage-controlled estimates are preferred over whole-sample estimates.

Continuous variables are summarised as medians with interquartile ranges and compared by Mann–Whitney U test; categorical variables as counts and percentages, compared by chi-square or Fisher’s

exact test where any expected count fell below five, with Bonferroni and Benjamini–Hochberg corrections applied across this family of eleven comparisons. Hardy–Weinberg equilibrium was tested in controls only by chi-square goodness-of-fit with one degree of freedom; two variants (*ADAMTS7*, *ADAMTS13*) were consistent with equilibrium and six deviated. No variant was excluded on this basis, but a restricted sensitivity analysis was pre-specified. Pairwise linkage disequilibrium was quantified as the squared Pearson correlation (r^2) between allele dosages.

Crude and adjusted odds ratios were estimated by logistic regression. The full adjustment set comprised age, sex, diabetes, hypertension, family history and smoking together with TG, TC, VLDL-C, LDL-C and HDL-C, and was applied identically in every adjusted single-variant model (Tables 3 and 4). Total cholesterol and VLDL-C were subsequently dropped from this set for the machine-learning, polygenic-score and epistasis analyses because they are near-linear functions of TG, LDL-C and HDL-C, leaving a nine-covariate set; the quantitative-trait scan used only the demographic and clinical covariates. Two consequences follow: cases have systematically higher lipid values and the panel is closely related to the clinical basis of case assignment, so adjusted estimates should be read as conditional on the measured lipid profile rather than as total effects; and the collinearity inflates coefficient variance, which was quantified by variance inflation factor. The additive model was pre-specified as primary; dominant, recessive and codominant models were pre-specified secondary analyses reported for completeness and never used to select a best-fitting model. Multiplicity correction (Bonferroni and Benjamini–Hochberg) was applied once, jointly across all 40 inheritance-model tests, rather than within each variant.

Polygenic risk scores and quantitative-trait analysis

An unweighted score was calculated as the sum of the eight risk-aligned doses, and a weighted score as the sum of each dose multiplied by its coefficient from the corresponding covariate-adjusted additive model. At this descriptive stage coefficients were estimated in the complete sample and discrimination summarised by five-fold stratified cross-validated AUC; because the weights derive from all observations this estimate is not fully leakage-free and every comparative claim rests instead on the re-estimated, within-partition results. Participants were stratified into tertiles and quartiles of each score, case proportions calculated per stratum, and trait distributions across tertiles compared by Kruskal–Wallis test.

As a secondary, hypothesis-generating analysis, each lipid trait was regressed by ordinary least squares on each allele dose and on the weighted score. This is the sole analysis not using the full adjustment set: because each lipid measure is in turn the dependent variable and the five are related by the Friedewald identities, cross-adjustment would be mathematically degenerate, so this scan adjusts for the demographic and clinical covariates only. Benjamini–Hochberg correction was applied across all 45 predictor–trait tests, with significance at $q < 0.05$. Given the lipid panel's relationship to case ascertainment, these findings are interpreted as reinforcing rather than independent of the primary associations.

Machine-learning models and the EW-cgPRS framework

The cohort was partitioned 80:20 (489 / 123) with stratification by case status, and the identical partition reused for every subsequent split-based analysis. The continuous covariates — and the polygenic score were used as a feature — were standardised using training-set parameters only. Total cholesterol and VLDL-C were excluded from the covariate set for all machine-learning, polygenic-score and epistasis analyses, leaving a nine-covariate set (age, sex, diabetes, hypertension, family history, smoking, TG, LDL-C and HDL-C). Because the cohort is balanced by design (244 cases versus 245 controls in the training fold), no SMOTE or other class-resampling was applied. Four classifiers were fitted — logistic regression, random forest (400 trees, depth 6), XGBoost and LightGBM (both 300 estimators, depth 4, learning rate 0.05) — under two feature configurations: the eight allele doses plus the nine covariates (17 features), and a single training-derived weighted score plus the same covariates (10 features). Performance was assessed by five-fold stratified cross-validation within the training set and on the hold-out set, reporting AUC with a percentile bootstrap 95% confidence interval (500 resamples for the machine-learning models and 1000 for the polygenic scores), per-fold AUC mean, SD and range, AUPRC, F1, accuracy, precision, recall, Matthews correlation coefficient and Brier score at a 0.5 threshold. Calibration was assessed by quantile-binned reliability diagrams alongside the Brier score, and model behaviour interpreted by SHAP values (tree explainer for ensembles, linear explainer for logistic regression) for every model, feature set and split.

The epistasis-weighted candidate-gene polygenic risk score (EW-cgPRS) comprises three alternative constructions rather than a strictly nested sequence. In Tier I, pairwise mutual information between allele doses was computed from the 3 × 3 genotype contingency table for each of the 28 locus pairs to populate an epistasis adjacency matrix; a unit self-loop was added, the matrix row-normalised by node degree, and each participant’s dosage vector propagated through the network before effect-size weighting. In Tier II, the unpropagated doses were weighted according to the predicted functional consequence of each variant. Missense variants (ADAMTS7, ADAMTS13, APOE, AGT, APOB, MTHFR, and ZC3HC1), which directly alter the encoded amino acid sequence, were assigned a functional weight of 1.00. The upstream regulatory variant in MMP9, which is predicted to influence gene expression rather than protein structure, was assigned a weight of 0.70. These functional annotation weights were multiplied by the corresponding effect sizes and normalised to a mean of one before deriving the weighted genetic risk score. In Tier III, the Tier I propagated dosages scaled by these annotation weights were entered into a single L2-regularised logistic regression (fixed C = 0.1) together with all 72 variant-by-covariate interaction terms and the nine standardised covariates. All effect sizes and mutual-information matrices were re-estimated from scratch within each cross-validation fold and training split and never computed using observations being scored; the externally specified annotation weights were held fixed.

Model comparison and robustness

All models were compared against the conventional weighted score as the common reference on the identical 123 held-out participants, giving twelve comparators: the unweighted score, the three EW-

cgPRS tiers and the eight machine-learning models (four classifiers under each of the two feature configurations). Discrimination was compared by DeLong's test, classification agreement at a 0.5 threshold by McNemar's test (exact where the smallest discordant cell fell below 25) and Cohen's κ. These twelve comparisons were not corrected for multiplicity. A master table reports the full metric panel for each score under two formulations — score-only and score-plus-covariates — across both splits, with ROC curves presented on common axes.

Because single-partition estimates are unstable at this sample size, three generalisation estimates were computed for all four classifiers and all five score types: the full pipeline repeated across 100 independent random 80:20 partitions, with weights and mutual-information matrices re-derived within each; nested cross-validation with five outer folds for performance estimation and three inner folds for hyperparameter selection, taken as the primary generalisation measure (applied to the complete cohort for the classifiers and within the training partition for the score types, so the two are reported separately); and a .632 bootstrap with 1000 resamples combining apparent and out-of-bag estimates. Five sensitivity analyses followed: a 1000-fold label-permutation null for the weighted score; an alternative score weighted by ElasticNet-penalised coefficients (mixing parameter 0.5, penalty selected from a ten-point grid by five-fold cross-validation); an alternative annotation weight derived solely from variant consequence class (missense 1.00, upstream regulatory 0.70, normalised to a mean of one), which carries no circularity with the outcome data; a score restricted to the two variants satisfying Hardy–Weinberg equilibrium in controls; and variance inflation factors across all covariates and allele doses, using thresholds of 5 and 10 for moderate and serious concern. All workflow is given is flow chart Supplementary Figure S1.

Results

Characteristics of the study population

The 612 participants comprised 306 cases and 306 controls. Cases and controls were of comparable age (median 58 years, IQR 51–67 versus 55 years, IQR 52–67; p = 0.86) and did not differ in smoking prevalence (6.2% versus 5.6%; p = 0.86). Cases were more often male (59.8% versus 43.8%; $p = 1.0 \times 10^{-4}$) and carried a substantially higher burden of conventional risk factors: diabetes was present in 43.5% of cases versus 9.2% of controls ($p = 1.3 \times 10^{-21}$), hypertension in 65.0% versus 8.2% ($p = 9.4 \times 10^{-48}$), and a positive family history in 43.5% versus 0.7% ($p = 6.0 \times 10^{-45}$). All four differences remained significant after Bonferroni correction across the eleven comparisons (Table 1).

Cases showed an adverse lipid profile on four of the five measures. Median triglycerides were 255.5 mg/dL (IQR 207.8–400.0) in cases versus 173.3 mg/dL (145.8–197.5) in controls ($p = 4.5 \times 10^{-35}$), LDL-C 122.0 mg/dL (88.8–200.0) versus 64.0 mg/dL (56.0–79.6) ($p = 1.1 \times 10^{-30}$), total cholesterol 214.0 versus 177.8 mg/dL ($p = 1.0 \times 10^{-5}$) and VLDL-C 105.0 versus 77.3 mg/dL ($p = 1.8 \times 10^{-4}$). HDL-C did not differ significantly (87.0 versus 77.7 mg/dL; p = 0.19). Full descriptive statistics are given in Table 1 and the distributions in Supplementary Figure S2.

Genotype distribution, Hardy–Weinberg equilibrium and linkage disequilibrium

Minor allele frequencies in controls ranged from 0.036 for *AGT* rs699 to 0.389 for *MTHFR* rs1801133. Two of the eight variants were consistent with Hardy–Weinberg equilibrium in controls: *ADAMTS7* rs3825807 ($\chi^2 = 0.37$, $p = 0.54$) and *ADAMTS13* rs2301612 ($\chi^2 = 0.11$, $p = 0.75$). The remaining six departed from equilibrium, five of them very strongly — *APOE* rs769452 ($\chi^2 = 177.8$, $p < 10^{-15}$), *ZC3HC1* rs11556924 ($\chi^2 = 121.7$, $p < 10^{-15}$), *APOB* rs676210 ($\chi^2 = 112.8$, $p < 10^{-15}$), *MTHFR* rs1801133 ($\chi^2 = 76.2$, $p < 10^{-15}$) and *MMP9* rs3918242 ($\chi^2 = 32.6$, $p = 1.1 \times 10^{-8}$) — with *AGT* rs699 showing a more modest deviation ($\chi^2 = 7.01$, $p = 0.008$). The direction of departure was heterozygote deficiency for *APOE*, *APOB*, *ZC3HC1*, *MMP9* and *AGT*, and heterozygote excess for *MTHFR* (218 observed versus 145.4 expected heterozygotes). Observed and expected genotype counts are shown in Table 2 and Supplementary Figure S3.

Pairwise linkage disequilibrium between the eight loci was negligible, as expected for variants on different chromosomes: the mean r^2 across the 28 pairs was 0.018 and only one pair exceeded $r^2 = 0.10$ (*AGT*–*MMP9*, $r^2 = 0.157$) (Figure 1a). The loci can therefore be treated as effectively independent for score construction.

Single-variant association with ischemic heart disease

Under the pre-specified additive model with adjustment for age, sex, diabetes, hypertension, family history, smoking and the five lipid measures, six of the eight variants were associated with IHD at $p < 0.05$ before multiplicity correction (Table 3, Table 4, Figure 1b). The strongest signal was at *ADAMTS13* rs2301612, where the risk-aligned per-allele odds ratio was 6.85 (95% CI 3.57–13.12; $p = 6.8 \times 10^{-9}$). *AGT* rs699 (OR 2.71, 1.30–5.66; $p = 0.008$), *MTHFR* rs1801133 (OR 2.71, 1.18–6.22; $p = 0.018$), *MMP9* rs3918242 (OR 2.42, 1.27–4.61; $p = 0.007$), *APOE* rs769452 (OR 2.16, 1.42–3.29; $p = 3.1 \times 10^{-4}$) and *ADAMTS7* rs3825807 (OR 1.89, 1.25–2.86; $p = 0.003$) also reached nominal significance. *APOB* rs676210 (OR 1.29, 0.87–1.91; $p = 0.20$) and *ZC3HC1* rs11556924 (OR 1.22, 0.83–1.81; $p = 0.31$) did not.

In the crude analysis, two variants showed a protective direction of effect and were consequently dose-inverted for all downstream scores: *ADAMTS13* rs2301612 (crude OR 0.10 per minor allele, 0.07–0.15) and *APOB* rs676210 (crude OR 0.66, 0.54–0.80).

After joint Bonferroni correction across all 40 inheritance-model tests, associations at four genes survived at $p_{Bonf} < 0.05$: *ADAMTS13* under the additive ($p_{Bonf} = 2.7 \times 10^{-7}$), dominant ($p_{Bonf} = 5.5 \times 10^{-7}$), recessive and both codominant contrasts; *ADAMTS7* under the recessive (OR 6.00, 2.69–13.40; $p_{Bonf} < 0.0001$) and codominant (minor-versus-normal) models; *APOE* under the additive ($p_{Bonf} = 0.012$) and dominant (OR 4.44, 2.12–9.26; $p_{Bonf} = 0.004$) models; and *MMP9* under the recessive ($p_{Bonf} = 0.012$) and codominant (minor-versus-normal, $p_{Bonf} = 0.024$) models. Under the less conservative Benjamini–Hochberg procedure, 23 of the 40 tests met $q < 0.05$. The secondary inheritance models were fitted for biological interpretation only and were not used to select a preferred model for any variant; several — notably the recessive estimates for *AGT* (OR 24.1, 3.36–172.95) and *ADAMTS13*

(OR 64.0, 6.33–646.50) — rest on very small homozygote counts and have correspondingly uninformative confidence intervals. The complete set of 40 tests is given in Table 4.

Polygenic risk scores

The eight adjusted additive coefficients used as weights ranged from +0.203 (*ZC3HC1*) to +1.924 (*ADAMTS13*). The unweighted score ranged from 0 to 12 across the cohort (mean 4.84, SD 2.03) and was higher in cases than controls (5.99 ± 1.95 versus 3.69 ± 1.35; $p = 2.6 \times 10^{-48}$); the weighted score showed the same pattern (4.481 ± 1.974 versus 2.109 ± 1.074; $p = 2.6 \times 10^{-54}$) (Figure 2a, Supplementary Table S2).

In five-fold cross-validation of a single-predictor model in the complete sample, the unweighted score achieved an AUC of 0.829 and the weighted score 0.861. Case prevalence rose monotonically across weighted-score tertiles, from 15.6% in the lowest to 91.6% in the highest, and across quartiles from 19.6% to 92.8% (Figure 2b, Table 5). The corresponding odds ratio for the top versus bottom tertile was 59.2 (95% CI 32.4–108.2) and for the top versus bottom quartile 52.9 (25.5–110.0). These stratification estimates are calculated in the same sample used to derive the weights and are reported as descriptive.

Across weighted-score tertiles, five of the six continuous traits differed significantly (LDL-C $p = 6.7 \times 10^{-15}$, triglycerides $p = 5.4 \times 10^{-12}$, total cholesterol $p = 6.6 \times 10^{-8}$, VLDL-C $p = 0.047$, HDL-C $p = 0.028$), while age did not ($p = 0.085$).

Quantitative-trait analysis of the lipid panel

Of the 45 predictor–trait tests, 17 survived Benjamini–Hochberg correction at $q < 0.05$ (Figure 3, Supplementary Table S3). The weighted polygenic score was associated with three of the five traits: LDL-C (β = 9.76 mg/dL per score unit, 95% CI 6.19–13.34; $q = 2.7 \times 10^{-6}$), total cholesterol (β = 9.11, 4.45–13.77; $q = 8.8 \times 10^{-4}$) and triglycerides (β = 9.47, 3.94–15.00; $q = 0.004$).

At the individual-variant level, *MMP9* rs3918242 showed the strongest single association, with VLDL-C (β = −35.31 mg/dL per risk allele, −48.20 to −22.42; $q = 2.7 \times 10^{-6}$), followed by *AGT* rs699 with VLDL-C (β = −32.48, −45.26 to −19.69; $q = 1.2 \times 10^{-5}$). *ADAMTS13* rs2301612 was associated with LDL-C (β = 24.50, 13.72–35.28; $q = 1.1 \times 10^{-4}$), triglycerides and total cholesterol; *APOB* rs676210 with triglycerides (β = 25.88, 13.47–38.29; $q = 4.3 \times 10^{-4}$) and VLDL-C; and *APOE* rs769452 with all four of total cholesterol, triglycerides, VLDL-C and LDL-C. *ADAMTS7* rs3825807 and *ZC3HC1* rs11556924 showed no trait association surviving correction.

Two features of these results warrant caution. The negative coefficients for *MMP9* and *AGT* against VLDL-C run opposite to their positive association with case status, and the trait associations as a group are difficult to separate from the case–control contrast itself, since the lipid panel is closely tied to how case status was assigned. This scan is therefore reported as hypothesis-generating and reinforcing of the primary associations rather than as independent evidence.

Machine-learning classification

The 80:20 partition produced training and test sets of 489 and 123 participants. All four classifiers achieved very high discrimination under both feature configurations (Table 6; full per-class metrics in Supplementary Table S4). With the eight allele doses plus the nine-covariate set, hold-out AUC was 0.984 (95% CI 0.968–0.996) for logistic regression, 0.998 for random forest, 0.999 for XGBoost and 0.997 for LightGBM; the corresponding cross-validated training AUCs were 0.975, 0.998, 0.995 and 0.996. Replacing the eight doses with a single training-derived weighted score gave essentially identical results (hold-out AUC 0.983 for logistic regression and 0.994–0.996 for the ensembles), indicating that compressing the panel into one score costs nothing in this setting. On the hold-out set XGBoost misclassified a single participant (accuracy 0.992) and random forest two (0.984), while LightGBM misclassified four (0.967) and logistic regression ten (0.919) The complete confusion matrices for both feature configurations are shown in Supplementary Figures S4 (raw doses plus covariates) and Supplementary Figure S5 (weighted score plus covariates), and the corresponding ROC curves and accuracies in Figures 4 and 5.

These figures are implausibly high for a candidate-gene study and are not interpretable as genuine predictive performance. A decomposition by feature block establishes the source. With the eight allele doses alone, cross-validated AUC was 0.874 (logistic regression) and 0.946 (random forest); with the clinical covariates alone, 0.887 and 0.926. With the lipid measures alone, however, random forest reached 0.999. The machine-learning results are accordingly presented as a demonstration that the covariate set is near-deterministic for case status in this design, and no claim of clinical predictive utility is made from them. Both classifiers were well calibrated on the hold-out set (Supplementary Figure S6). Feature attributions are reported in full: global SHAP importance rankings for the models without and with the polygenic score (Supplementary Figures S7 and S8), the corresponding per-participant beeswarm distributions (Supplementary Figures S9–S12), and a representative single-participant local explanation (Supplementary Figure S13).

The EW-cgPRS framework and formal model comparison

Mean pairwise mutual information across the 28 locus pairs was 0.024 bits (maximum 0.078 bits), indicating negligible detectable epistatic structure among these eight loci in this sample (Supplementary Figure S14). Consistent with this, network propagation produced no meaningful gain: on the hold-out set, EW-cgPRS-I achieved an AUC of 0.857, essentially identical to the 0.858 of the conventional weighted score (DeLong $p = 0.74$), and EW-cgPRS-II, which applies functional annotation weights without propagation, achieved 0.861 (DeLong $p = 0.19$). The unweighted score performed significantly worse than the weighted score (0.805; DeLong $p = 0.049$).

EW-cgPRS-III achieved a hold-out AUC of 0.969, significantly higher than the conventional weighted score (DeLong $p = 0.0049$). This tier differs from the others in incorporating the 72 variant-by-covariate interaction terms and the nine covariates directly, so its advantage reflects the inclusion of the clinical and lipid covariates rather than any epistatic or annotation-based contribution; the same covariate set

drives the machine-learning results. Under the score-plus-covariates formulation, in which every score is given the same covariate access, the five scores converge (hold-out AUC 0.984–0.985 for the four conventional and epistasis tiers, 0.969 for Tier III), confirming that the Tier III advantage in the score-only comparison is attributable to covariate access and not to the epistasis or annotation machinery (Supplementary Figure S15). The condensed comparison appears as Table 6, the full master table as Supplementary Table S5, and the combined ROC curves as Figure 4.

All twelve comparators were tested against the conventional weighted score on the identical 123 held-out participants. Eight of the eight machine-learning models differed significantly (DeLong p ranging from 1.8×10^{-4} to 5.5×10^{-5}), as did EW-cgPRS-III and the unweighted score. Neither EW-cgPRS-I nor EW-cgPRS-II differed significantly. Cohen's κ against the reference classifier ranged from 0.535 (unweighted score) to 0.983 (EW-cgPRS-I). These comparisons were not corrected for multiplicity.

Robustness and sensitivity analyses

Across 100 independent random 80:20 partitions, mean hold-out AUC was 0.854 (SD 0.040; 2.5th–97.5th percentile 0.780–0.916) for the conventional weighted score, 0.855 (0.781–0.920) for EW-cgPRS-I, 0.846 (0.775–0.912) for EW-cgPRS-II, 0.855 (0.780–0.917) for the unweighted score and 0.977 (0.954–0.992) for EW-cgPRS-III; the three ensemble classifiers averaged 0.997–0.998 (SD 0.002) (Figure 6). Nested cross-validation gave closely concordant estimates: 0.852 (SD 0.035) for the weighted score, 0.864 for EW-cgPRS-I, 0.852 for EW-cgPRS-II, 0.855 for the unweighted score and 0.972 for EW-cgPRS-III within the training partition, and 0.977 for logistic regression and 0.998 for each ensemble across the full cohort. The .632 bootstrap produced optimism-corrected estimates of 0.979 for logistic regression and 0.998–0.999 for the three ensembles, with apparent and out-of-bag estimates differing by no more than 0.008 (Figure 7) — indicating that the near-perfect performance is not an artefact of a single favourable partition but a stable property of this feature set (Supplementary Table S6, S7).

The 1000-fold label-permutation test confirmed that the score–outcome association is not attributable to chance: the observed AUC of 0.862 lay far outside the permutation null (mean 0.518, SD 0.015; p = 0.001), the smallest value attainable with 1000 permutations (Figure 8).

ElasticNet penalisation shrank all coefficients except that of *APOB* and *ZC3HC1*, with the largest reductions for *AGT* (0.997 to 0.388) and *MMP9* (0.883 to 0.271), consistent with these being the least stably estimated weights; the resulting score performed marginally better than the unpenalised version (cross-validated AUC 0.871 versus 0.861). Reweighting by variant consequence class left performance unchanged (0.861). Restricting the score to the two variants in Hardy–Weinberg equilibrium reduced the cross-validated AUC from 0.861 to 0.769, indicating that a substantial part of the discrimination of the full score derives from the six loci that deviate from equilibrium and whose genotype calls are therefore less secure. Variance inflation factors were uniformly low, with a maximum of 1.541 (hypertension) across all seventeen covariates and allele doses, excluding multicollinearity as an explanation for the coefficient estimates (Supplementary Table S8).

Discussion

To our knowledge, this is among the most comprehensive integrative candidate-gene investigations of ischemic heart disease (IHD) conducted in a Pakistani population, uniting single-variant association testing, weighted polygenic scoring, an epistasis- and functional-annotation-weighted score, a quantitative-trait scan, and four machine-learning classifiers within a single, leakage-controlled analytical framework. Its principal positive finding is that a compact, eight-variant weighted polygenic risk score (PRS) captured a genuine and reproducible gradient of genetic risk. The score separated cases from controls with a cross-validated area under the curve (AUC) of 0.861, exceeding the unweighted allele-count score (0.829), and case prevalence rose monotonically from 15.6% in the lowest weighted-score tertile to 50.9% in the middle and 91.6% in the highest, corresponding to a top-versus-bottom-tertile odds ratio (OR) of 59.2 (95% CI 32.4–108.2). Critically, this gradient was not an artefact of a single favourable data split: it remained essentially unchanged across 100 independent 80:20 partitions (mean AUC 0.854), under nested cross-validation (0.852) and under a .632 bootstrap (0.853), and a 1000-fold label-permutation test placed the observed AUC (0.862) far outside the null distribution (mean 0.518; $p = 0.001$). The score therefore behaves as a coherent aggregate instrument, though—as developed below—it is a candidate-gene score confined to eight loci and must be read as such, not as a genome-wide PRS.

The single strongest association in our data was at *ADAMTS13* rs2301612, with a risk-aligned per-allele OR of 6.85 (95% CI 3.57–13.12; $p = 6.8 \times 10^{-9}$). *ADAMTS13* encodes the von Willebrand factor–cleaving protease that restrains platelet-mediated thrombus growth, and reduced ADAMTS13 activity has been repeatedly linked to myocardial infarction and arterial thrombotic risk, giving the locus clear biological plausibility[21]. The magnitude of our estimate nonetheless greatly exceeds the published record: in a recent European case–control study, Wrona and colleagues found no association between rs2301612 and coronary artery disease (CAD) or long-term survival, although the same variant was associated with HDL-cholesterol and pro-atherogenic lipid indices and behaved as an expression quantitative-trait locus in vascular tissue[31]. Importantly, and in contrast to five of the other loci examined here, this variant was in Hardy–Weinberg equilibrium among controls ($\chi^2 = 0.11$; $p = 0.75$), so genotyping artefact is an unlikely explanation. The discrepancy is more plausibly attributable to winner's-curse inflation in a modest sample, to the very small homozygote counts that render the recessive and codominant estimates unstable (recessive OR 64.0, 95% CI 6.33–646.5), and to genuine population differences. We therefore treat the size of this effect as provisional and in need of independent, ancestry-matched replication rather than as evidence of a uniquely large ADAMTS13 effect in South Asians.

The remaining Bonferroni-surviving signals align more closely with established data and are mechanistically coherent. *ADAMTS7* rs3825807, a well-replicated CAD locus whose prodomain substitution modulates enzyme maturation and vascular smooth-muscle-cell migration, was associated with IHD under additive (OR 1.89, 1.25–2.86) and, most strongly, recessive models (OR 6.00, 2.69–13.40); this recessive pattern mirrors the AA-genotype association with myocardial infarction reported by Petrovič and colleagues in Slovenian subjects with type 2 diabetes[16]. *APOE* rs769452, which destabilises apoE and impairs its lipid binding, showed additive (OR 2.16, 1.42–3.29) and dominant (OR 4.44, 2.12–

9.26) associations, directionally consistent with meta-analytic evidence that *APOE* variation, and the ε4 allele in particular, raises coronary heart disease risk[32]. *MMP9* rs3918242, the − 1562C > T promoter variant that elevates gelatinase-B transcription and can degrade the fibrous cap, was associated under recessive (OR 14.49, 3.40–61.80) and codominant models—matching the updated meta-analysis of Zhang and colleagues, which reported the strongest matrix-metalloproteinase-9 effect under recessive models and in Asian populations (recessive OR for myocardial infarction 2.29)[18]. Two further variants, *AGT* rs699 (M235T; additive OR 2.71, 1.30–5.66) and *MTHFR* rs1801133 (C677T; additive OR 2.71, 1.18–6.22), reached nominal but not Bonferroni-corrected significance. Both are biologically credible—angiotensinogen is the rate-limiting substrate of the renin–angiotensin system, and the thermolabile MTHFR enzyme raises homocysteine—but our point estimates again exceed the meta-analytic consensus for these loci (allelic OR ≈ 1.38 for *AGT*[23] and ≈ 1.11 for *MTHFR*[33], the latter essentially null in the folate-replete analysis of Clarke and colleagues[28]), and the *AGT* association was recently null in an angiographically defined Pakistani cohort[24]. This upward departure is exactly the inflation expected of statistically significant findings in small candidate-gene studies. Finally, *APOB* rs676210 (OR 1.29, 0.87–1.91) and *ZC3HC1* rs11556924 (OR 1.22, 0.83–1.81) were not associated after adjustment; for *ZC3HC1*, whose common allele confers only an ≈ 9% per-copy increase in risk, a null result at this sample size is unsurprising.

A distinctive and, in our view, instructive feature of this study was the attempt to enrich the additive score with epistatic and functional information through a three-tier epistasis-weighted candidate-gene PRS (EW-cgPRS); in practice, this yielded no advantage. Pairwise mutual information among the eight loci was negligible (mean 0.024 bits; maximum 0.078), indicating essentially no detectable statistical epistasis in this sample. Consistent with this, network propagation of dosage (Tier I, hold-out AUC 0.857) and functional-annotation weighting (Tier II, 0.861) were statistically indistinguishable from the conventional weighted score (0.858; DeLong p = 0.74 and 0.19, respectively). Only the interaction tier (Tier III) appeared superior (0.969; p = 0.0049), and that advantage disappeared once every score was granted equal access to the clinical and lipid covariates (hold-out AUC 0.984–0.985 for all tiers), revealing it as a covariate effect rather than an epistatic one. This is a useful negative result: with only eight loci and a modest sample there is little exploitable interaction structure and added model complexity mainly re-imports covariate information already available to simpler models. It is consistent with formal benchmarking work showing that non-additive and gene–gene interaction terms rarely improve complex-disease prediction beyond well-specified additive models[29,30], and it parallels the finding of Alireza and colleagues (2024) that, in the UK Biobank, machine-learning feature selection could match polygenic scoring with as few as ≈ 50 variants but that neither adding a genome-wide PRS nor increasing model complexity improved on a compact, risk-factor-anchored model[34].

All four classifiers achieved optimal discrimination (hold-out AUC 0.984 for logistic regression and 0.997–0.999 for the ensembles). We interpret this not as evidence of clinically useful prediction but as a cautionary illustration of how study design can inflate machine-learning performance. A decomposition by feature block localised almost the entire signal to the lipid measures: a random forest built from lipids

alone reproduced the near-perfect result (AUC 0.999), whereas genotype-only (0.874–0.946) and clinical-covariate-only (0.887–0.926) models were substantially weaker. Because our controls were ascertained in part on lipid criteria, the lipid panel effectively encodes case status, and the classifiers are exploiting that structural dependence rather than any genetic signal. This phenomenon is well documented as data leakage, a recognised driver of over-optimistic, non-reproducible machine-learning results[35,36], and it is directly illuminated by two of the most relevant external studies. Widen, Lello and colleagues (Genes, 2021) showed that biomarker-only risk scores in the UK Biobank—using no genotype at all—achieved an AUC of ≈ 0.76 for CAD and ≈ 0.95 for diabetes, with the CAD predictor dominated by LDL, apolipoprotein B, total cholesterol and HDL and the diabetes predictor by glycated haemoglobin; in both cases the apparent power flowed from biomarkers that partly define the diagnosis, exactly as the lipid panel does here, whereas their SNP-only heart-attack predictor reached only ≈ 0.59[37]. Likewise, Alireza and colleagues found that genotype-only CAD prediction plateaued near an AUC of 0.55 while models incorporating conventional risk factors exceeded 0.75[34]. Our own repeated-split, nested-cross-validation and .632-bootstrap analyses reconcile these observations: the polygenic score's more modest AUC (≈ 0.85) is stable and reproducible, whereas the ensembles' near-ceiling values are a property of the lipid-defined design. We therefore make no claim of clinical predictive utility from the classifiers and present them primarily to demonstrate, transparently, that the covariate set is near-deterministic for case status in this sample.

The quantitative-trait scan reinforced the primary associations while underscoring their entanglement with the lipid phenotype. Seventeen of forty-five predictor–trait tests survived correction, with the weighted score associated with LDL-cholesterol (β = 9.76 mg/dL per score unit; $q = 2.7 \times 10^{-6}$), total cholesterol (β = 9.11; $q = 8.8 \times 10^{-4}$) and triglycerides (β = 9.47; q = 0.004), and several individual variants —including *ADAMTS13* (LDL β = 24.50), *APOE* (all four of total cholesterol, triglycerides, VLDL and LDL), *APOB* (triglycerides β = 25.88) and, in the opposite direction, *MMP9* and *AGT* against VLDL (β = −35.31 and − 32.48)—reaching significance. The *ADAMTS13*–lipid relationship is notable because it independently echoes the rs2301612–HDL/atherogenic-lipid association reported by Wrona and colleagues[31], lending external support to that specific link. However, because case status was itself defined partly through the lipid panel, and because the negative *MMP9* and *AGT* coefficients run counter to their positive association with disease, these trait associations cannot be cleanly separated from the case–control contrast and are best regarded as hypothesis-generating and corroborative rather than as independent evidence of variant–lipid causality.

These findings should be read against the wider landscape of cardiovascular polygenic prediction. Our eight-locus instrument is categorically different from genome-wide PRS for CAD, which aggregate hundreds to millions of variants, are underpinned by large multi-ancestry discovery efforts, and are the focus of current clinical-translation and guideline activity[38,39]; such scores can identify individuals with monogenic-equivalent risk and refine preventive-therapy targeting[40], but their portability across ancestries remains a central limitation, and scores derived predominantly in European-ancestry cohorts systematically underperform in under-represented populations, including South Asians[9,41,42]. The relative

standing of genetic information as a standalone predictor is illustrated by a MEDINFO 2023 comparison (Studies in Health Technology and Informatics), in which genotypic variables were the weakest single-modality predictor of CAD, carotid-imaging intermediate, and conventional phenotypic risk factors by far the strongest (reported approximate AUCs of 0.56, 0.62 and 0.84, respectively)[43]. Our candidate-gene score is not a substitute for a genome-wide PRS; it is a compact, interpretable, ancestry-specific instrument that may serve as a foundation for South Asian–specific models, in a population bearing one of the highest global burdens of premature coronary disease yet markedly under-represented in genomic databases. The path forward is likely to be biologically informed rather than merely larger: Petrazzini and colleagues (Nature Genetics, 2024) showed that a machine-learning-derived digital CAD phenotype linked to exome-sequencing data recovered 17 genes—14 with prior cardiovascular evidence—demonstrating that precision phenotyping and biologically informed feature selection can improve gene discovery independently of sample size[44]. Our own pathway-driven panel—spanning thrombosis, lipoprotein metabolism, vascular remodelling, extracellular-matrix degradation and blood-pressure regulation—and its identification of *ADAMTS13*, *ADAMTS7*, *APOE* and *MMP9* as leading contributors is in the same spirit.

Several strengths and limitations frame these conclusions. The chief strengths are the breadth of the analytical framework, the deliberate selection of biologically plausible functional variants across complementary atherogenic pathways, and a validation strategy—repeated train–test splitting, nested cross-validation, a .632 bootstrap and label permutation—designed specifically to guard against the optimism and leakage that inflate single-split estimates. The limitations are equally important and, in places, structural. First and most consequential, controls were ascertained partly on lipid criteria, creating a dependence between predictors and outcome that inflates classifier performance and complicates the trait analyses; the machine-learning results in particular should be read as a demonstration of this artefact rather than as clinical prediction. Second, six of eight variants deviated from Hardy–Weinberg equilibrium in controls, raising the possibility of genotyping error or population structure; restricting the score to the two equilibrium loci (*ADAMTS7* and *ADAMTS13*) reduced the cross-validated AUC from 0.861 to 0.769, indicating that a meaningful share of the score's discrimination is carried by loci whose genotype calls are least secure. Third, the sample was modest and drawn from a single population and setting, which limits power, inflates the effect sizes of variants that do reach significance, precludes any assessment of cross-population portability, and was not accompanied by an external replication cohort. Fourth, the score is confined to eight candidate loci and therefore captures only a small fraction of the heritability of IHD. Finally, the design was cross-sectional, so no temporal or causal inference is warranted, intermediate mechanisms (for example homocysteine for *MTHFR*, or ADAMTS13 activity for *ADAMTS13*) were not measured, and the annotation weights used in the epistasis tiers were literature-derived and provisional.

Taken together, the results support a genuine but modest polygenic contribution to IHD risk in this population, captured efficiently by a simple weighted score, and they show that neither epistatic enrichment nor flexible machine learning added interpretable value beyond it once information leakage

was accounted for. Future work should prioritise larger, ancestry-matched and prospectively ascertained South Asian cohorts with independent regenotyping; the incorporation of rare coding variants from exome or whole-genome sequencing; the integration of transcriptomic, proteomic and metabolomic layers; and explainable-artificial-intelligence approaches that keep model behaviour interpretable. Progress in cardiovascular precision medicine for this high-burden, under-represented population is likely to come not from increasingly complex algorithms but from the disciplined integration of biologically informed genomic data, rigorous leakage-aware methodology and ancestry-specific validation.

Conclusion

In this Pakistani case–control study of ischemic heart disease, functional candidate variants in *ADAMTS13*, *ADAMTS7*, *APOE* and *MMP9* were associated with disease after stringent Bonferroni correction, and a compact eight-variant weighted polygenic score aggregated these signals into a reproducible gradient of genetic risk that remained stable across repeated resampling, nested cross-validation and label permutation. Layering epistatic or functional-annotation information onto this score conferred no advantage, indicating that a panel of this size contains little exploitable interaction structure and that added model complexity largely re-imports covariate information already available to simpler additive models. The well calibrated discrimination of the machine-learning classifiers is best understood as a consequence of lipid-based control ascertainment—whereby the lipid panel effectively encodes case status—rather than as evidence of clinical predictive value, and it stands as a cautionary illustration of how information leakage can inflate the apparent performance of flexible models applied to selected case–control data. Several of the individual effect sizes are correspondingly large and likely inflated by winner's curse, and six of the eight loci deviated from Hardy–Weinberg equilibrium, so the associations warrant cautious interpretation. Taken together, the findings support a genuine but modest polygenic contribution to ischemic heart disease in this under-represented, high-burden population—captured efficiently by a simple weighted score that is a compact, interpretable instrument rather than a substitute for a genome-wide polygenic score. Larger, ancestry-matched and prospectively ascertained South Asian cohorts, with independent regenotyping, external replication and leakage-aware study design, will be required before candidate-gene scores or machine-learning classifiers of this kind can inform cardiovascular risk assessment.

Declarations

Ethics approval and consent to participate: The research was approved by the Research Ethics and Biosafety Committee of the University of the Punjab, Pakistan, which follows the principles Declaration of the Helsinki. Written informed consent was obtained from all participants prior to enrollment.

Consent for publication: Not applicable.

Availability of data and materials: The dataset analyzed during this study is available as Supplementary Data S1. The complete analysis code is publicly accessible via the GitHub repository:

https://github.com/anamijaz2021-qmul/Integrative-ML-model-with-genetics-and-lipid-profile-for-Ischemic-Heart-Disease.git

Competing interests: The authors have no competing interests.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: A.I.: Designed the study; collected, curated, and analyzed the data; wrote and proofread the manuscript. **S.A.:** Analyzed data, prepared tables and figures, and proofread the manuscript. **Sh.:** Proofread, reviewed, and approved the draft of the manuscript.

Acknowledgements: Not applicable

References

1. Liang, P. *et al.* Trends and predictive research on the global burden of ischemic heart disease from 1990 to 2021: an analysis of the Global Burden of Disease study 2021. *Frontiers in Public Health* 13, 1569179 (2025).
2. Wang, Y. *et al.* Global trends in the burden of ischemic heart disease based on the global burden of disease study 2021: the role of metabolic risk factors. *BMC Public Health* 25, 310 (2025).
3. Yusuf, S. *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The lancet* 364, 937–952 (2004).
4. Xia, H., Li, J. & Dong, J. Clinical Characteristics and Prognosis of Acute Heart Failure in Patients with Chronic Obstructive Pulmonary Disease. *The American journal of cardiology* (2025).
5. Samad, Z. & Hanif, B. Cardiovascular diseases in Pakistan: imagining a postpandemic, postconflict future. *Circulation* 147, 1261–1263 (2023).
6. Hassan, K. M.*et al.* in *CIRCULATION.* (LIPPINCOTT WILLIAMS & WILKINS TWO COMMERCE SQ, 2001 MARKET ST, PHILADELPHIA …
7. Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nature medicine* 28, 1679–1692 (2022).
8. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nature genetics* 54, 1803–1815 (2022).
9. Martin, A. R. *et al.* Current clinical use of polygenic scores will risk exacerbating health disparities. *Nature genetics* 51, 584 (2019).
10. Mäkelä, K.-M. *et al.* Genome-wide association study pinpoints a new functional apolipoprotein B variant influencing oxidized low-density lipoprotein levels but not cardiovascular events: AtheroRemo Consortium. *Circulation: Cardiovascular Genetics* 6, 73–81 (2013).
11. Abdulfattah, S. Y. & Al-Awadi, S. J. ApoB gene polymorphism (rs676210) and its pharmacogenetics impact on atorvastatin response among Iraqi population with coronary artery disease. *Journal of*

Genetic Engineering and Biotechnology 19, 95 (2021).

12. Wintjens, R. *et al.* Global molecular analysis and APOE mutations in a cohort of autosomal dominant hypercholesterolemia patients in France [S]. *Journal of lipid research* 57, 482–491 (2016).
13. Bauer, R. C. *et al.* Knockout of Adamts7, a novel coronary artery disease locus in humans, reduces atherosclerosis in mice. *Circulation* 131, 1202–1213 (2015).
14. Pu, X. *et al.* ADAMTS7 cleavage and vascular smooth muscle cell migration is affected by a coronary-artery-disease-associated variant. *The American Journal of Human Genetics* 92, 366–374 (2013).
15. Mizoguchi, T. *et al.* Coronary disease association with ADAMTS7 is due to protease activity. *Circulation research* 129, 458–470 (2021).
16. Petrovič, D., Nussdorfer, P. & Petrovič, D. The rs3825807 polymorphism of ADAMTS7 as a potential genetic marker for myocardial infarction in slovenian subjects with type 2 diabetes mellitus. *Genes* 14, 508 (2023).
17. Li, Y.-Y. *et al.* Matrix metalloproteinase-9 gene-1562C > T gene polymorphism and coronary artery disease in the Chinese Han population: a meta-analysis of 5468 subjects. *Frontiers in Physiology* 7, 212 (2016).
18. Zhang, M.-M. *et al.* Association between matrix metalloproteinase 9 C-1562T polymorphism and the risk of coronary artery disease: an update systematic review and meta-analysis. *Oncotarget* 9, 9468 (2017).
19. Jones, P. D. *et al.* The coronary artery disease-associated coding variant in zinc finger C3HC-type containing 1 (ZC3HC1) affects cell cycle regulation. *Journal of Biological Chemistry* 291, 16318–16327 (2016).
20. Linseman, T. *et al.* Functional validation of a common nonsynonymous coding variant in ZC3HC1 associated with protection from coronary artery disease. *Circulation: Cardiovascular Genetics* 10, e001498 (2017).
21. Chion, C. K., Doggen, C. J., Crawley, J. T., Lane, D. A. & Rosendaal, F. R. ADAMTS13 and von Willebrand factor and the risk of myocardial infarction in men. *Blood* 109, 1998–2000 (2007).
22. Ma, Q. *et al.* Genetic variants in ADAMTS13 as well as smoking are major determinants of plasma ADAMTS13 levels. *Blood advances* 1, 1037–1046 (2017).
23. Zhao, H., Zhao, R., Hu, S. & Rong, J. Gene polymorphism associated with angiotensinogen (M235T), endothelial lipase (584C/T) and susceptibility to coronary artery disease: a meta-analysis. *Bioscience Reports* 40, BSR20201414 (2020).
24. Ahmed, S. T., Nasir, M. I., Amir, K. F. & Siddiqui, P. Q. R. Spectrum of AGT (M235T) rs699 and AGTR1 (A1166C) rs5186 gene variants and its association with coronary artery disease in Pakistani patients. *Pakistan Journal of Medical Sciences* 41, 1151 (2025).
25. Klerk, M. *et al.* MTHFR 677C→ T polymorphism and risk of coronary heart disease: a meta-analysis. *Jama* 288, 2023–2031 (2002).

26. Wang, W., Wang, Y., Gong, F., Zhu, W. & Fu, S. MTHFR C677T polymorphism and risk of congenital heart defects: evidence from 29 case-control and TDT studies. *Plos one* 8, e58041 (2013).
27. Bouzidi, N. *et al.* Association of the methylene-tetrahydrofolate reductase gene rs1801133 C677T variant with serum homocysteine levels, and the severity of coronary artery disease. *Scientific reports* 10, 10064 (2020).
28. Clarke, R. *et al.* Homocysteine and coronary heart disease: meta-analysis of MTHFR case-control studies, avoiding publication bias. *PLoS medicine* 9, e1001177 (2012).
29. Elgart, M. *et al.* Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications biology* 5, 856 (2022).
30. Bell, N. Y., Wightman, D. P., de Leeuw, C. & Posthuma, D. Benchmarking non-additive genetic effects on polygenic prediction and machine learning-based approaches. *medRxiv*, 2025.2010. 2010.25337750 (2025).
31. Wrona, J. *et al.* ADAMTS13 Gene Polymorphisms and Coronary Artery Disease Risk, Long-Term Survival, and Risk Factor Profile. *Genes* 17, 508 (2026).
32. Xu Min, X. M. *et al.* Apolipoprotein E gene variants and risk of coronary heart disease: a meta-analysis. (2016).
33. Luo, Z. *et al.* Associations of the MTHFR rs1801133 polymorphism with coronary artery disease and lipid levels: a systematic review and updated meta-analysis. *Lipids in Health and Disease* 17, 191 (2018).
34. Alireza, Z., Maleeha, M., Kaikkonen, M. & Fortino, V. Enhancing prediction accuracy of coronary artery disease through machine learning-driven genomic variant selection. *Journal of Translational Medicine* 22, 356 (2024).
35. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4 (2023).
36. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S. & Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature communications* 15, 1829 (2024).
37. Widen, E., Raben, T. G., Lello, L. & Hsu, S. D. Machine learning prediction of biomarkers from SNPs and of disease risk from biomarkers in the UK Biobank. *Genes* 12, 991 (2021).
38. Klarin, D. & Natarajan, P. Clinical utility of polygenic risk scores for coronary artery disease. *Nature Reviews Cardiology* 19, 291–301 (2022).
39. O'Sullivan, J. W. *et al.* Polygenic risk scores for cardiovascular disease: a scientific statement from the American Heart Association. *Circulation* 146, e93-e118 (2022).
40. Natarajan, P. *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* 135, 2091–2101 (2017).

41. Patel, A. P. *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nature Medicine* 29, 1793–1803 (2023).
42. Busby, G. B. *et al.* Ancestry-specific polygenic risk scores are risk enhancers for clinical cardiovascular disease assessments. *Nature Communications* 14, 7105 (2023).
43. Sng, L. M., Sharma, R., Bagot, S., Bauer, D. C. & Twine, N. A. in *19th World Congress on Medical and Health Informatics, MedInfo 2023*. 1021–1025 (IOS Press).
44. Petrazzini, B. O. *et al.* Exome sequence analysis identifies rare coding variants associated with a machine learning-based marker for coronary artery disease. *Nature genetics* 56, 1412–1419 (2024).

Tables

Tables 1 to 6 are available in the supplementary files section.

Figures

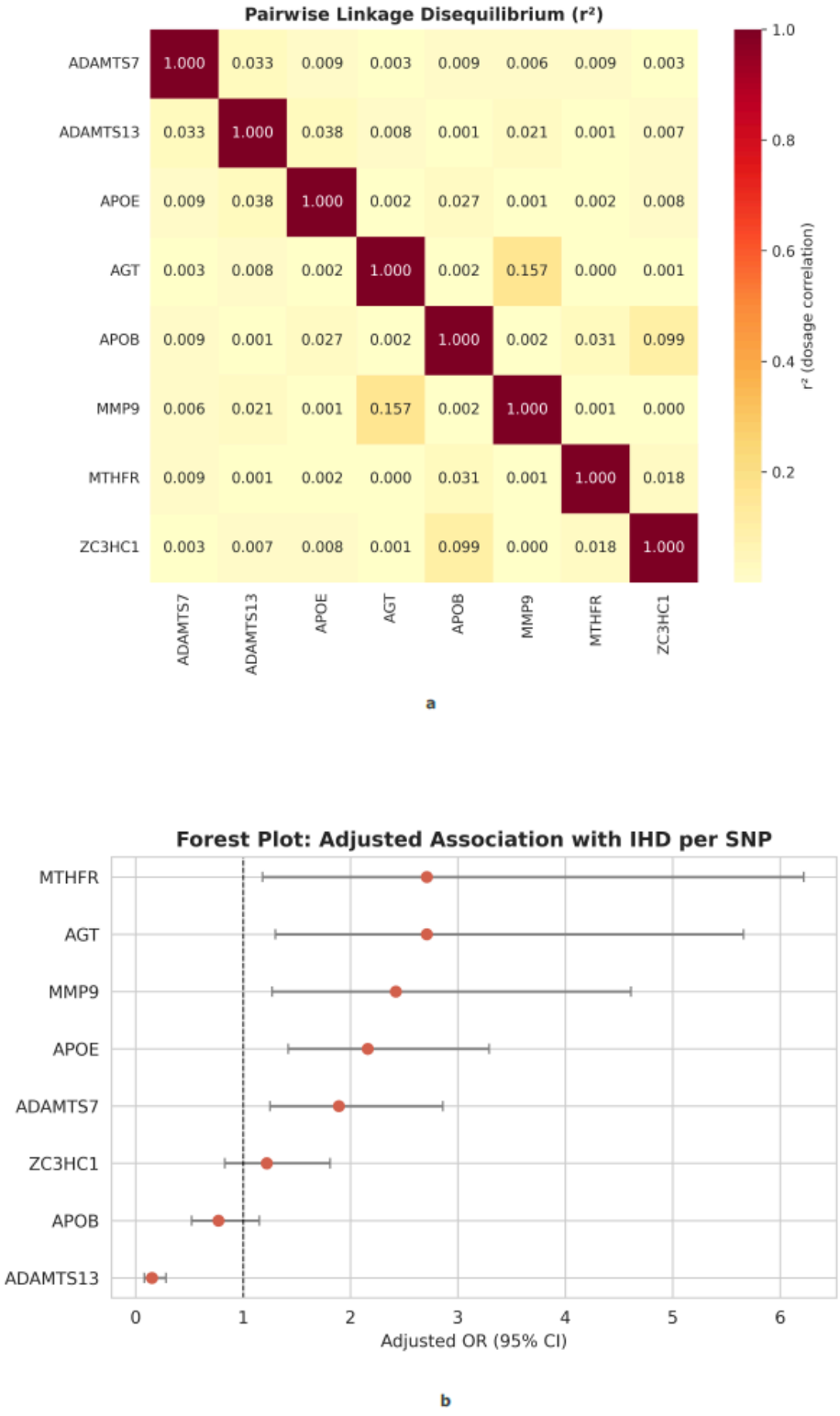
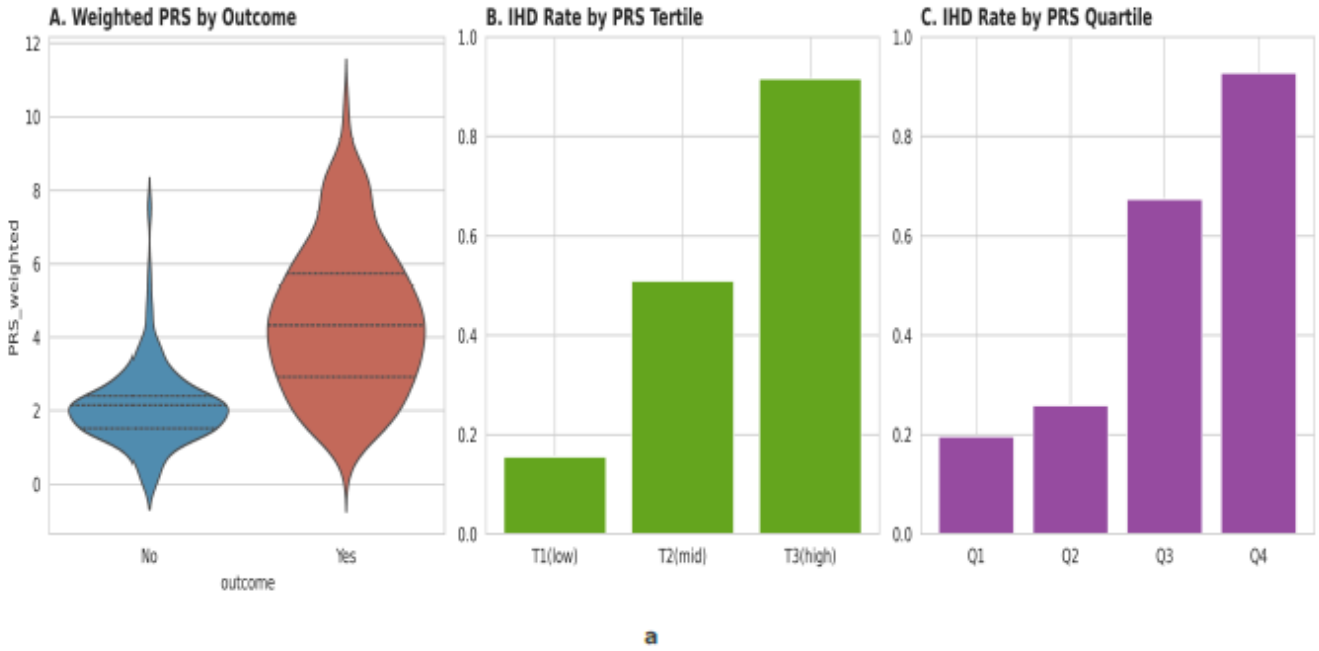
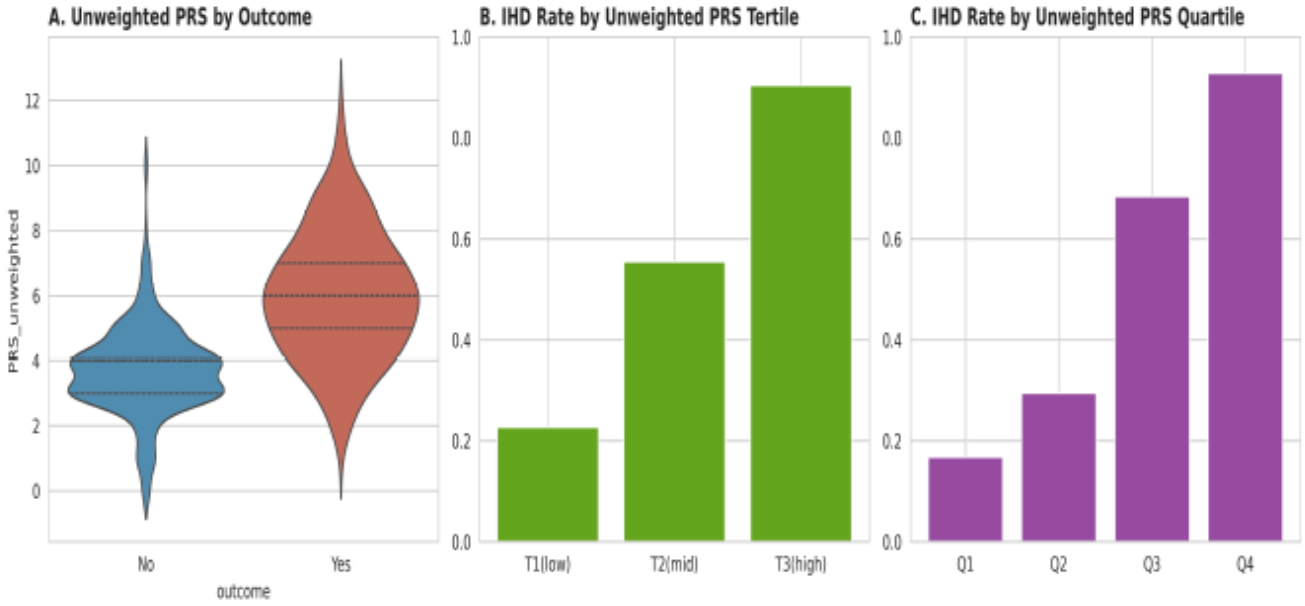


Figure 1

a) Pairwise linkage disequilibrium (r²) among the candidate SNPs confirming independence of markers (all off-diagonal r² < 0.10 except AGT–MMP9 (r² = 0.157); mean r² = 0.018.). b) Forest plot of adjusted ORs



a



b

Figure 2

a) Weighted-PRS stratification, b) Unweighted-PRS stratification

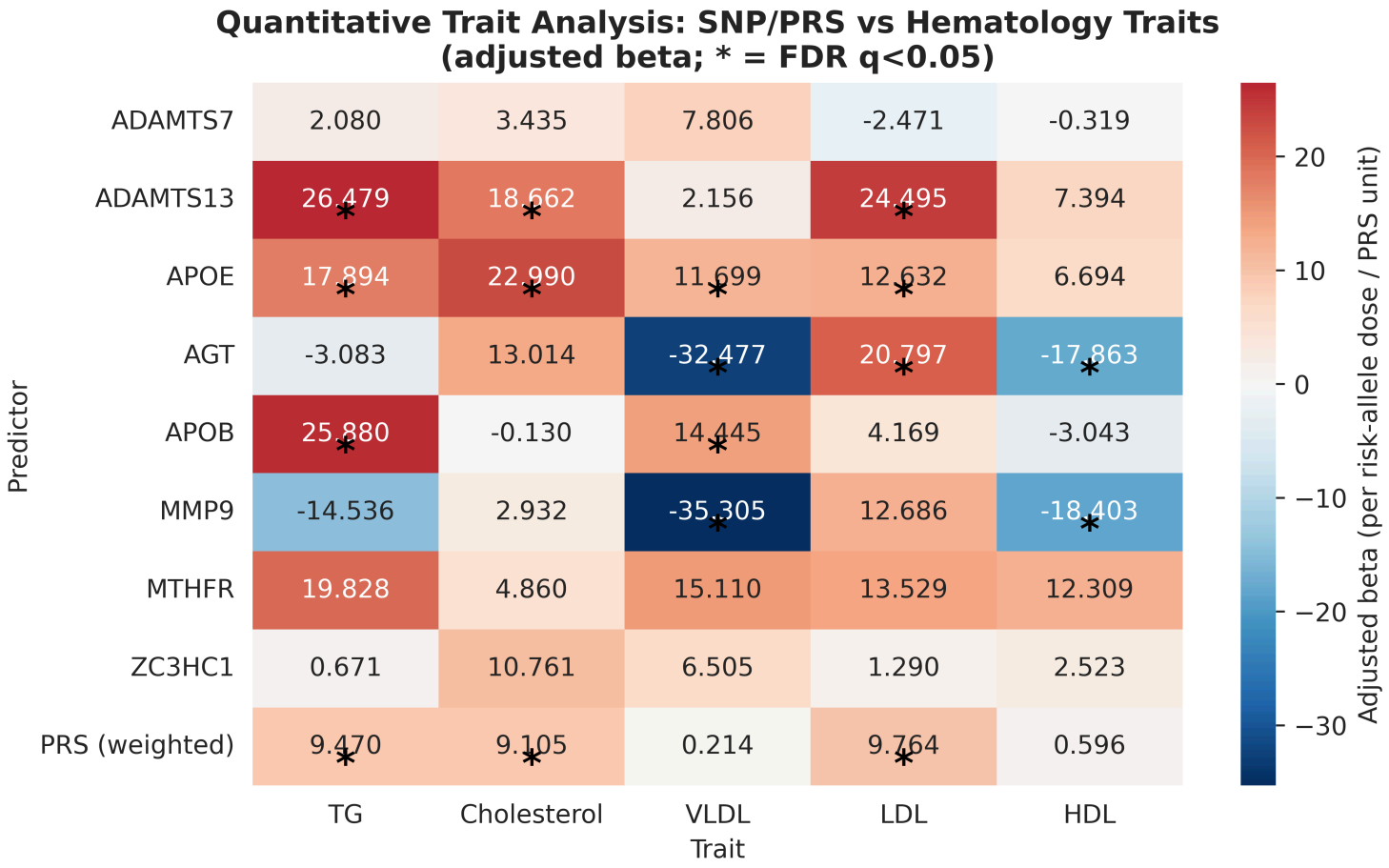


Figure 3

QTL heatmap (SNP/PRS vs hematology traits)

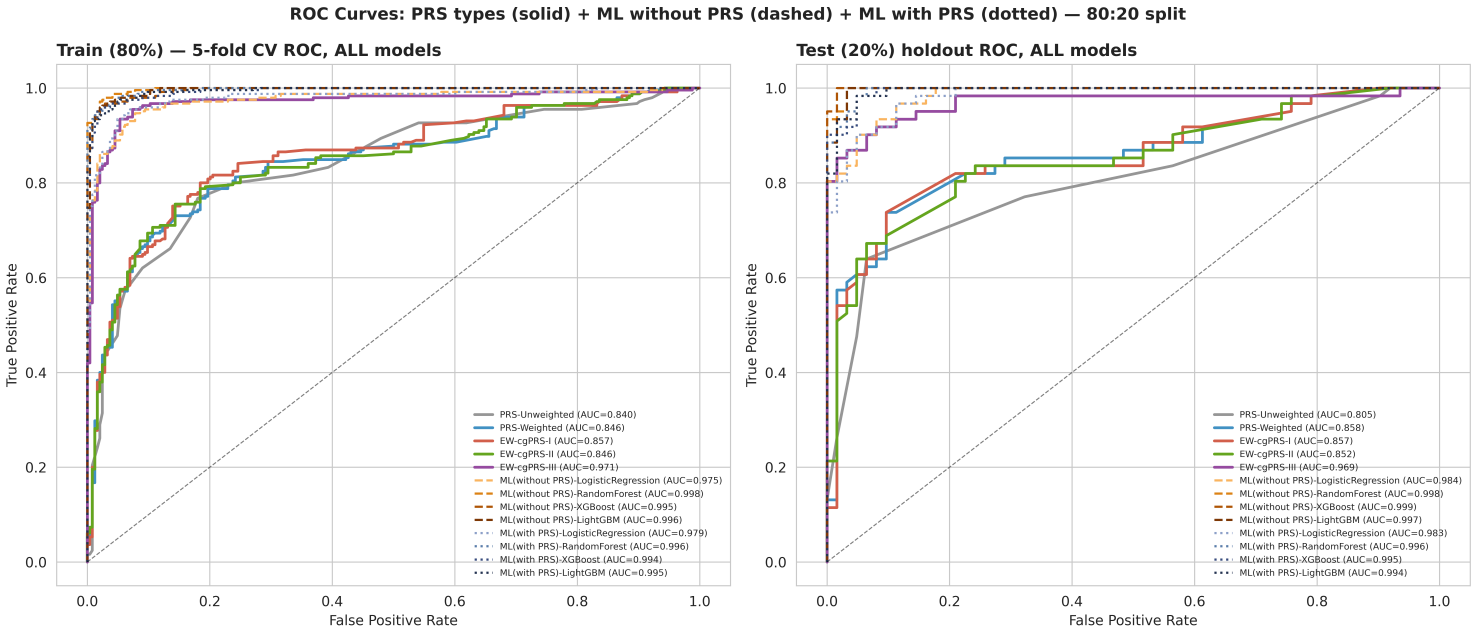


Figure 4

ROC curves of PRS types (solid), ML without PRS (dashes), ML with PRS (dotted) in 80:20 split.

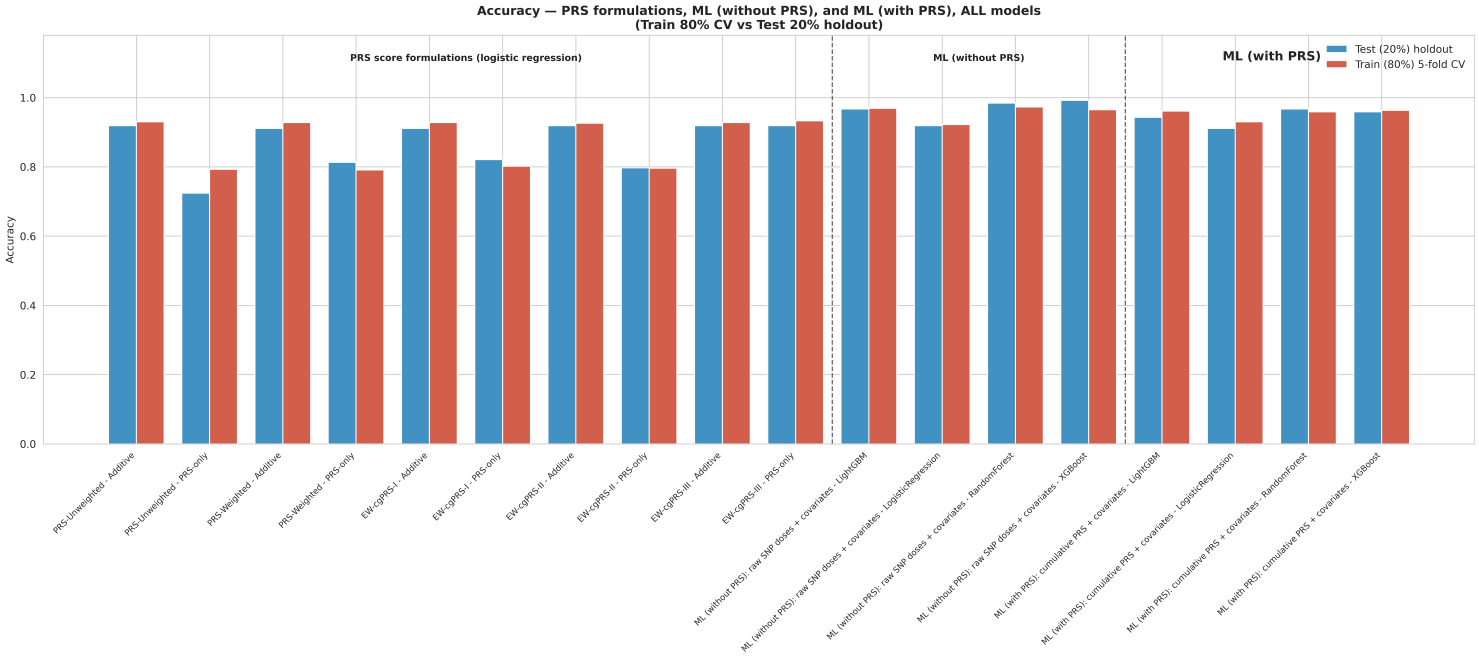


Figure 5

Accuracies of all PRS and ML models in train 80% cross-validation (CV) vs. test 20% holdout set.

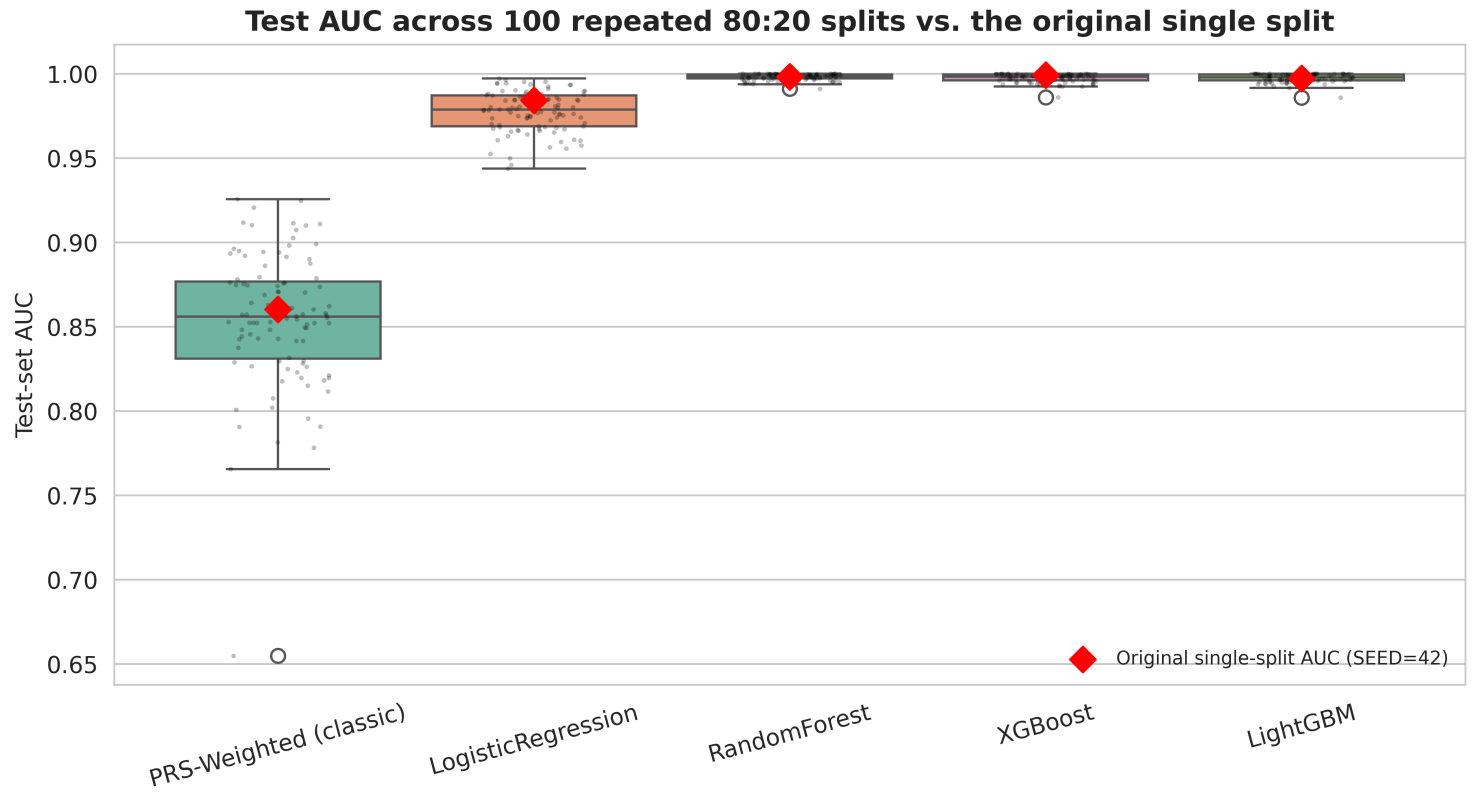


Figure 6

Test AUC across 100 repeated 80:20 splits vs. original single split.

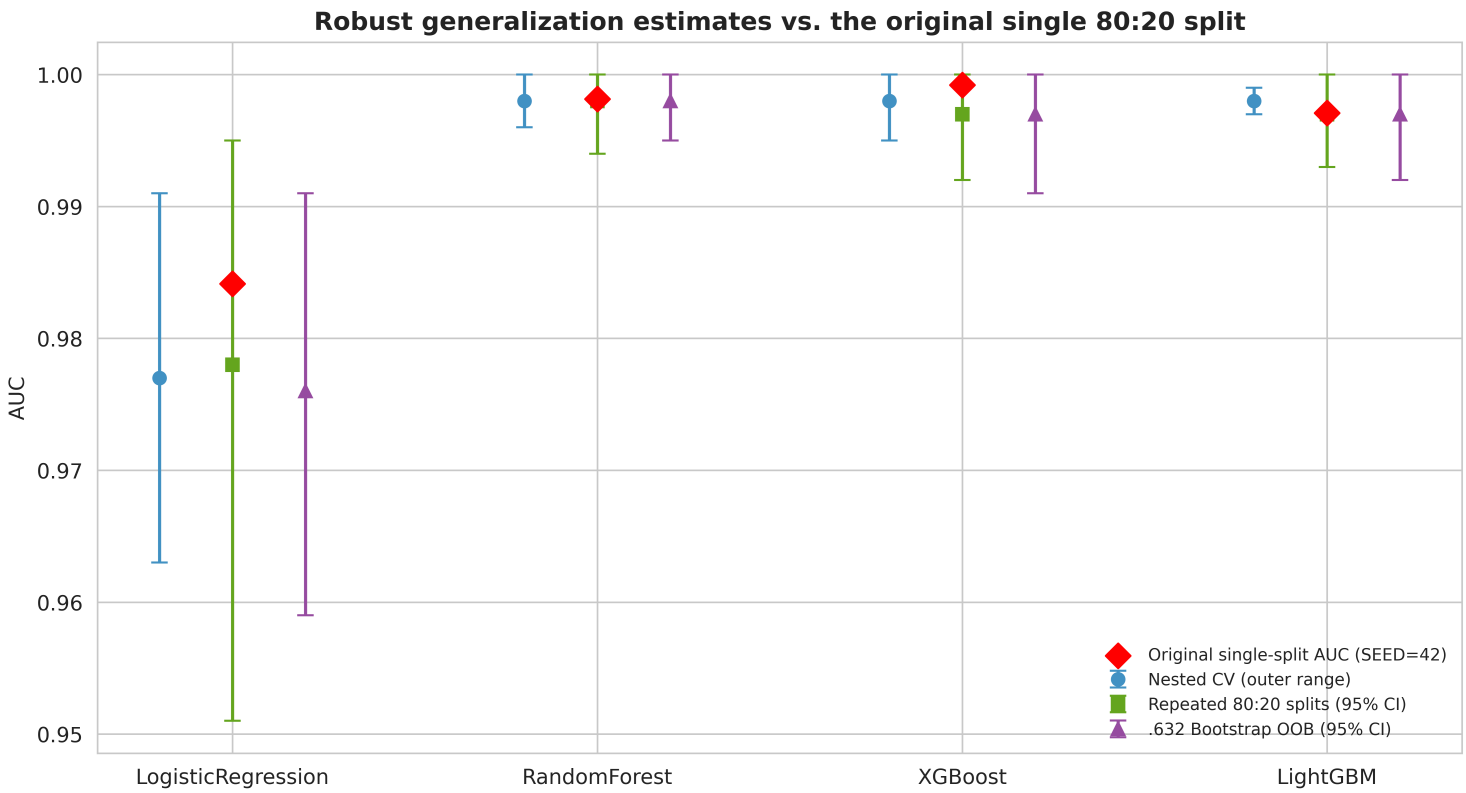


Figure 7

Robust generalization estimates vs. original single 80:20 splits.

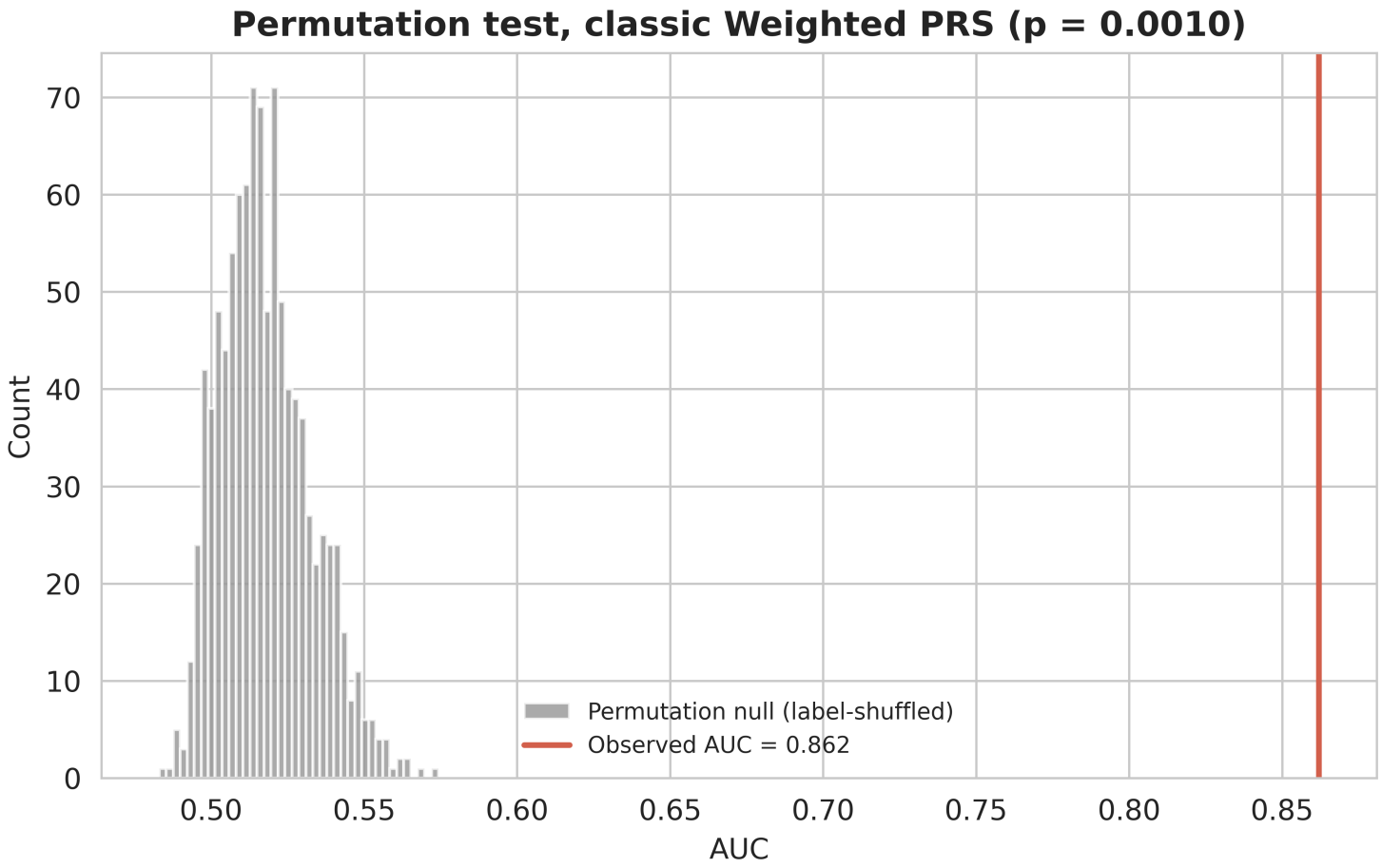


Figure 8

Permutation test, classic weighted PRS (p=0.001).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDataS1.xlsx
- SupplementaryTables.docx
- SupplementaryFigureS1.png
- SupplementaryFigureS2.png
- SupplementaryFigureS3.png
- SupplementaryFigureS4.png
- SupplementaryFigureS5.png
- SupplementaryFigureS6.pdf
- SupplementaryFigureS7.png
- SupplementaryFigureS8.png

- SupplementaryFigureS9.pdf
- SupplementaryFigureS10.pdf
- SupplementaryFigureS11.pdf
- SupplementaryFigureS12.pdf
- SupplementaryFigureS13.pdf
- SupplementaryFigureS14.png
- SupplementaryFigureS15.png
- Tables.docx
- SupplementaryDataS1.docx
- SupplementaryFigureS4.png