MRS Model Experimentation

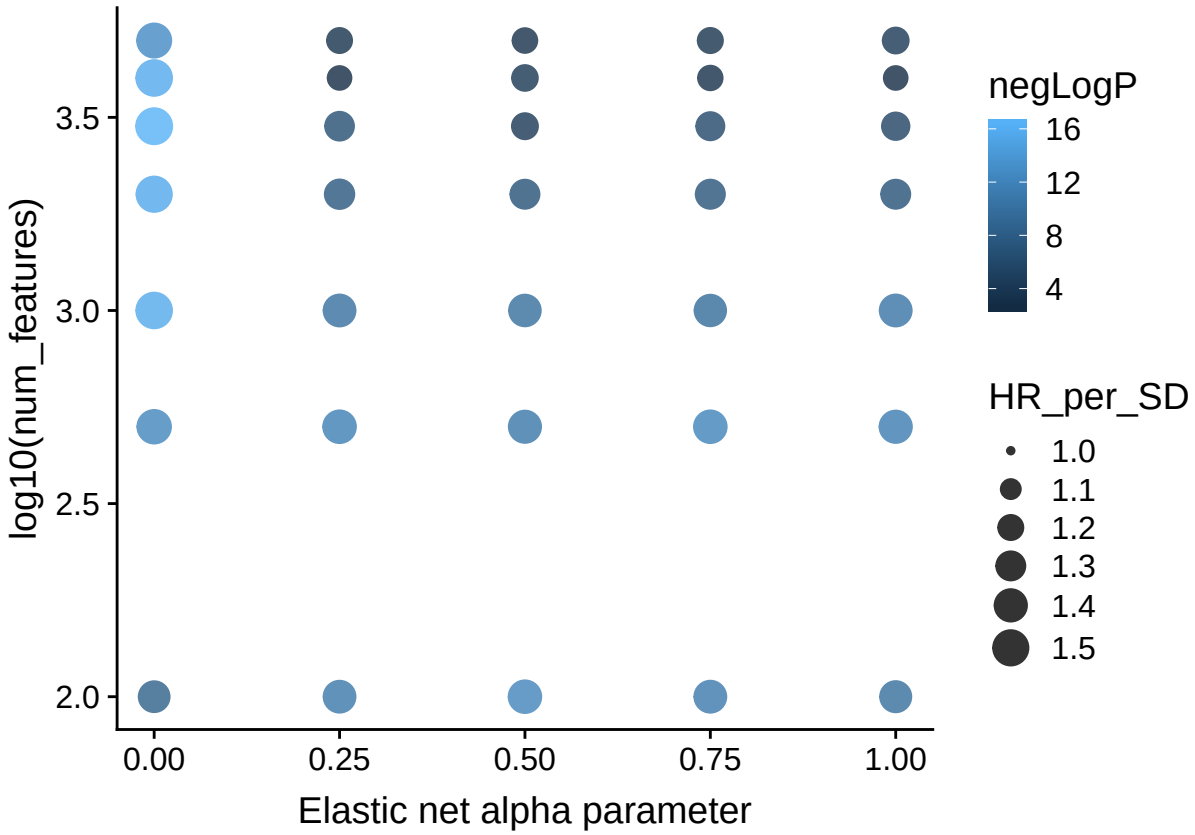
0.1 Methylation risk score construction

Basic setup:

- Combine WHI/FHS/LBC36 and split into 50% training, 25% validation, and 25% testing
- Test marginal associations of each CpG with incident CVD (training set)
- Grid search over parameters for elastic net (# top CpGs, alpha) (training set, evaluate in validation set)
- Train tuned model in training + validation set and evaluate performance in test set

0.1.1 Initial EWAS-like screening to get marginal associations

0.1.2 Tuning



A grid search suggests that performance is optimized with an alpha parameter of zero (i.e. ridge regression) and use of the top 3000 CpG sites based on marginal associations with incident CVD. Using these parameters, a model was trained in the combined training + validation sets and tested on the fully held-out portion of the dataset.

Table 1: Performance of the MRS in the held-out test set.

Covariate_set	HR_per_SD	p
none	1.46	0.000
age_sex	1.42	0.000
plus_cell_counts	1.45	0.000
plus_bmi	1.41	0.000
plus_diabetes	1.28	0.000
plus_smoking	1.31	0.000
plus_lipids_sbp	1.21	0.009
full_model	1.15	0.082
FRS_only	1.35	0.000

0.1.3 Testing

0.2 Final MRS construction and characterization

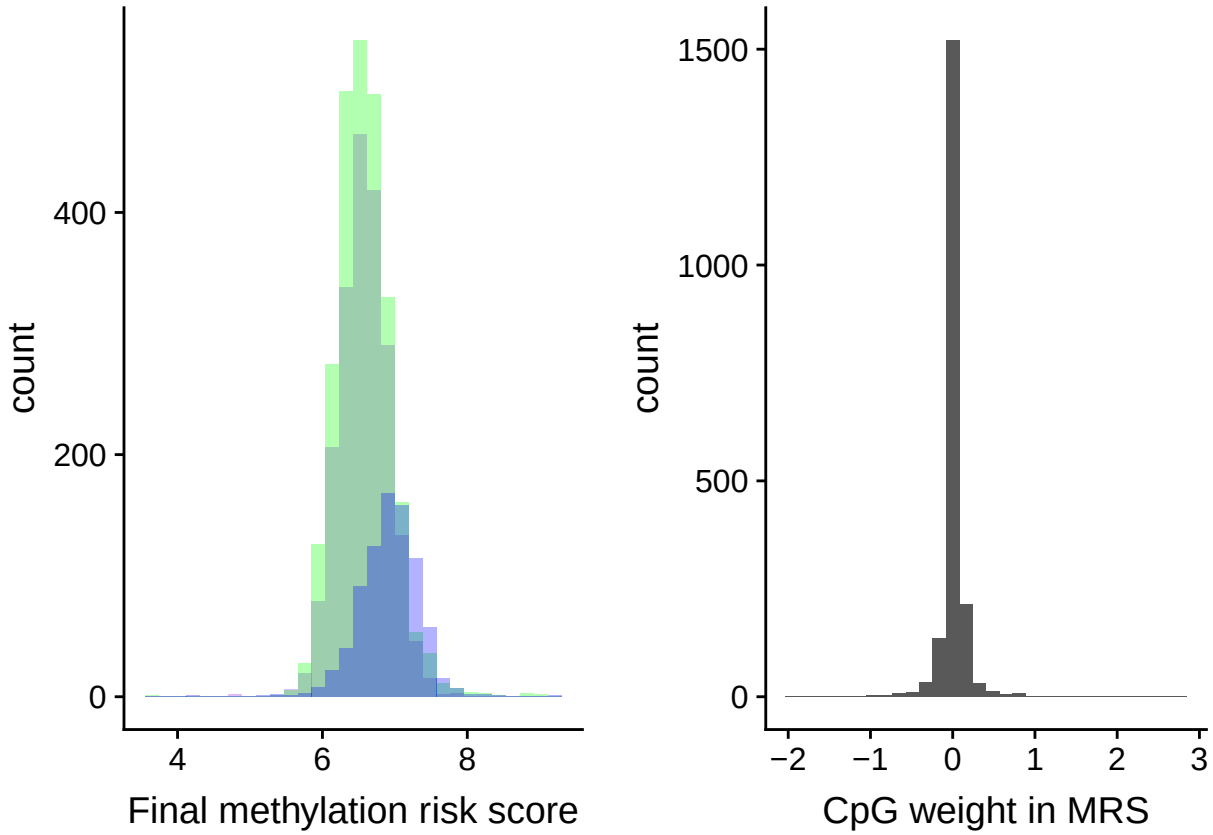


Table 2: MRS stability as evaluated by using multiple within-subject measurements. Generic ICC heuristics for reference: 0-0.5 = poor, 0.5-0.75 = moderate, 0.75 - 0.9 = good, 0.9-1 = excellent.

Cohort	Group_type	ICC
FHS	Duplicates	0.875
LBC36	Samples over multiple visits	0.521

Table 3: Validation of Framingham Risk Score

study	HR_per_SD	p
whi	1.50	0.000
fhs	1.67	0.000
lbc36	0.88	0.042

0.2.1 MRS stability

0.3 Risk score interactions with demographic and risk-based attributes

The following set of results is based on models adjusting for age, sex (when not the stratifying factor), and cell counts.

0.3.1 Demographics

- Sex (noting that the training set was considerably skewed towards females)
- Race
- Age

0.3.2 Traditional risk

Framingham Risk Score (2008 generalized CVD version) was used to calculate cardiovascular risk. Diabetes was defined as either blood sugar medication use or measured fasting glucose > 125 mg/dL.

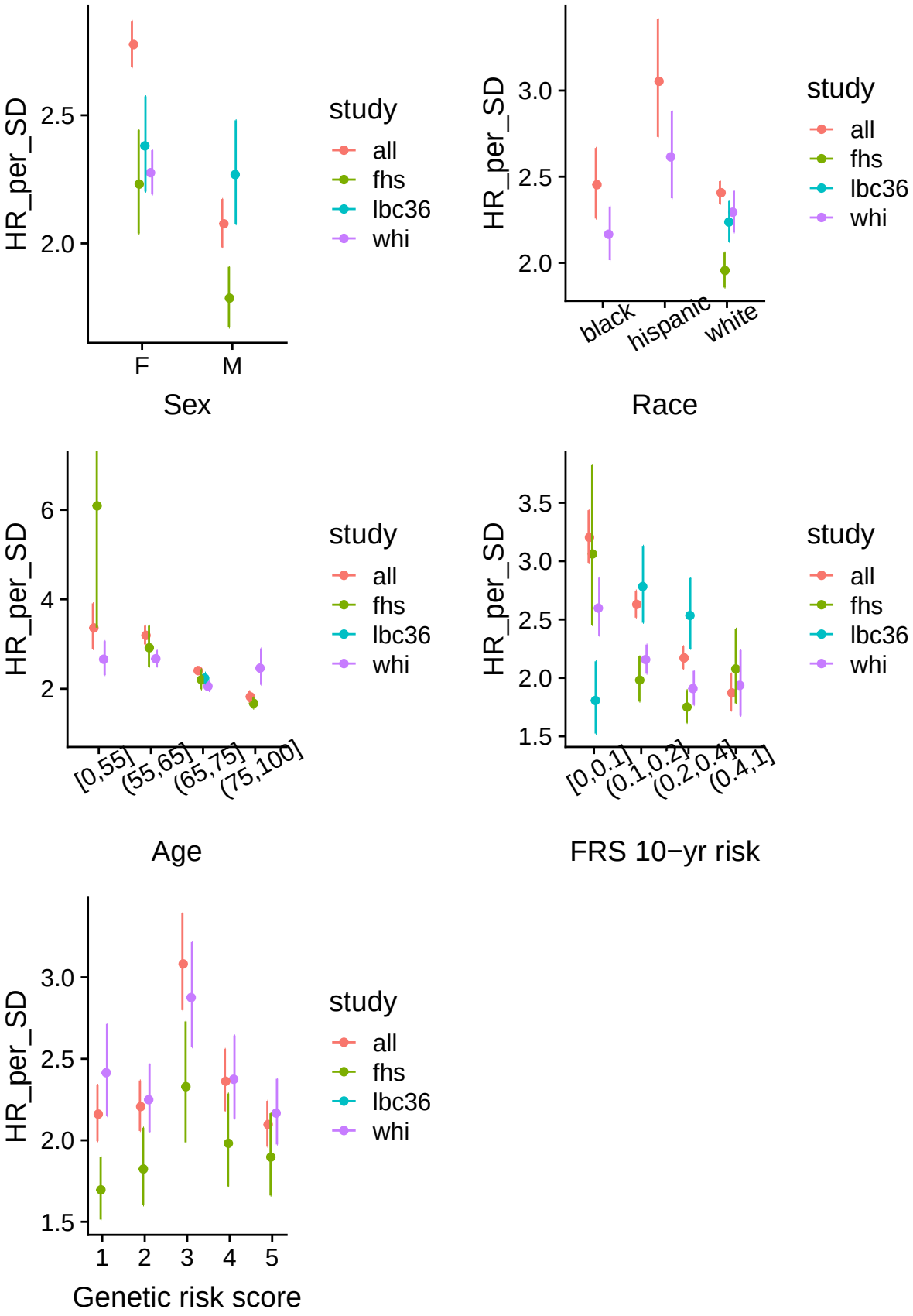
0.3.3 Genetic risk

A genetic risk score (GRS) was calculated based on the genome-wide polygenic model used by Khera et al. 2018 for prediction of coronary heart disease (~6M SNPs). This GRS was first tested to confirm its associations in WHI and FHS (both of which had about 80% of the full set of GRS SNPs available after imputation and QC; table in Supplementary Info). While all CVD cases are incident in WHI, past and incident events were merged into a single binary variable for FHS in order to test the GRS.

Table 4: Validation of genetic risk score

cohort	OR_per_SD	p
WHI	1.21	0.000
FHS	1.05	0.368

0.3.4 Stratified plots



0.4 Replication in REGICOR

- Age & sex-balanced but not incident and MI only
- Look for sex, age, and FRS interactions here

0.5 Replication in KORA/InCHIANTI

- Look for sex, age, FRS, and genetic interactions here