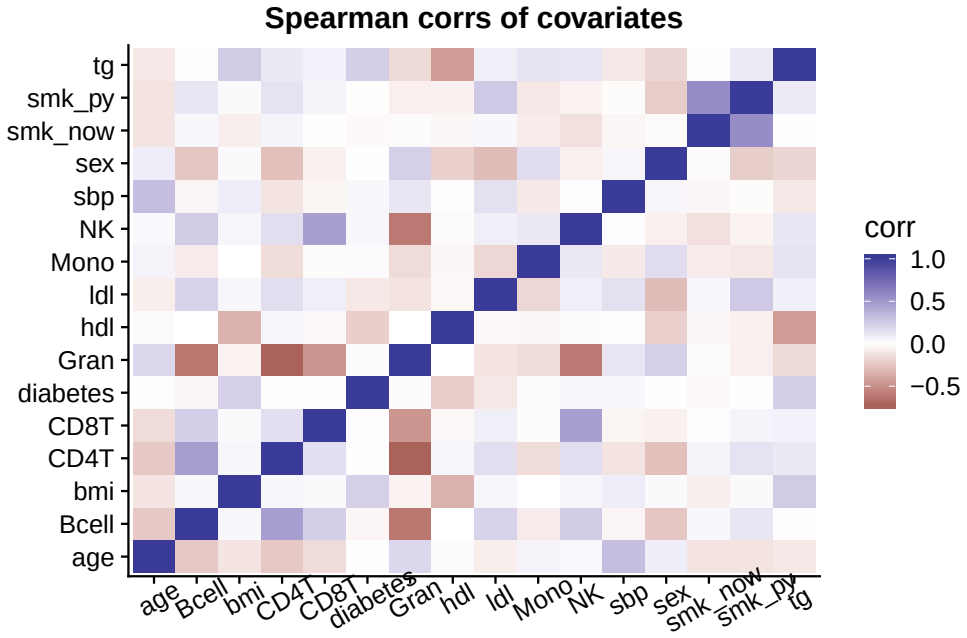
MRS Model Experimentation

1 Goal

Develop a methylation-based risk score to predict incident cardiovascular events (defined here as MI, angina, stroke, or death from CVD). Performance will be assessed in a held-out portion of the FHS dataset and replicated in external datasets. Interactions between the risk score and other measures of cardiovascular risk (biochemistry, genetic) will also be explored.

2 Population characteristics

- WHI: N~2000 women, race breakdown 50/25/25
- FHS: N ~ 2300 whites
  - ~600 from CHD case-control (JHU)
  - Rest from FHS Offspring population (U of Minn.)
- LBC36: N ~ 800 whites



The above covariate plot suggests notable collinearity between (cell counts), (current and pack-years of smoking), and (BMI/TG/HDL). A series of increasingly-adjusted models will be used to assess the performance of the risk score, with a few variables removed :

1. Unadjusted
2. Basic covariates: age + sex + cell counts (CD4T + CD8T + Bcell + NK + Mono, no Gran)
3. Basic covariates plus risk factors (BMI + current smoking + LDL + HDL + SBP)
4. Framingham Risk Score only

Table 1: Results on held-out FHS subset

covariate_set	HR_per_SD	p
unadjusted	1.62	0.000
basic	1.32	0.000
plus_risk_factors	1.30	0.001
FRS_only	1.40	0.000

3 Cross-study learner model development and assessment

3.1 Description

The "stacking" method used here is based on Patil 2018, who suggest using linear combinations of predictions from multiple studies as an alternative to a simple combination of all datasets as was done above. It involves training study-specific predictors, then using predictions from these models as features in a multivariate, penalized Cox regression on the full combined ("stacked") dataset. Normalized coefficients from this regression then act as weights for linear combination of these study-specific predictors into a final risk score.

This method has the added advantage of avoiding the possibility that the regression picks up batch effects due to the highly heterogeneous class balances across studies. Though I adjust for study in the combined model, it is possible that it still picks up batch effect signatures, while the stacking model is a linear combination of models that have only "seen" a single study.

For the following, FHS was divided into two "studies", based on the batch run at JHU (case-control for CHD) and the batch run at U. of Minn. (rest of the Offspring cohort), for a total of four studies. For internal assessment, the U. of Minn. dataset is held out during the stacking procedure, and only included for training of the final model to be used for replication.

3.2 Initial results for FHS-UM holdout

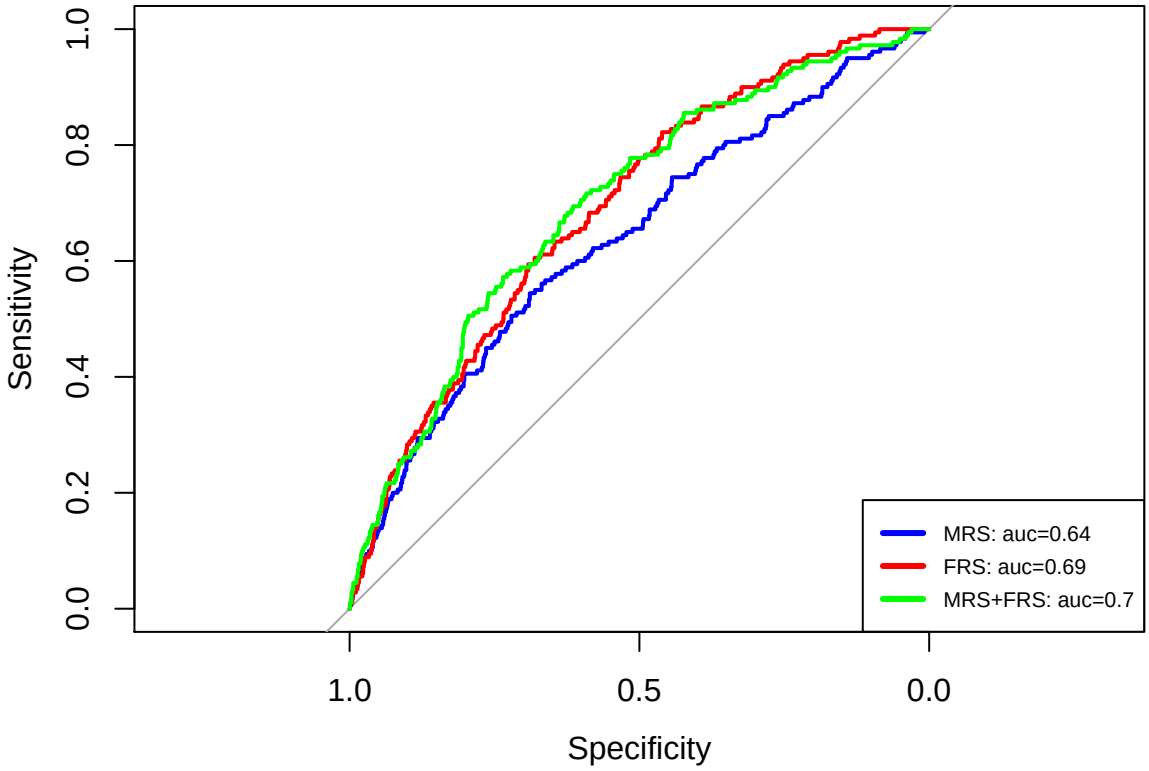
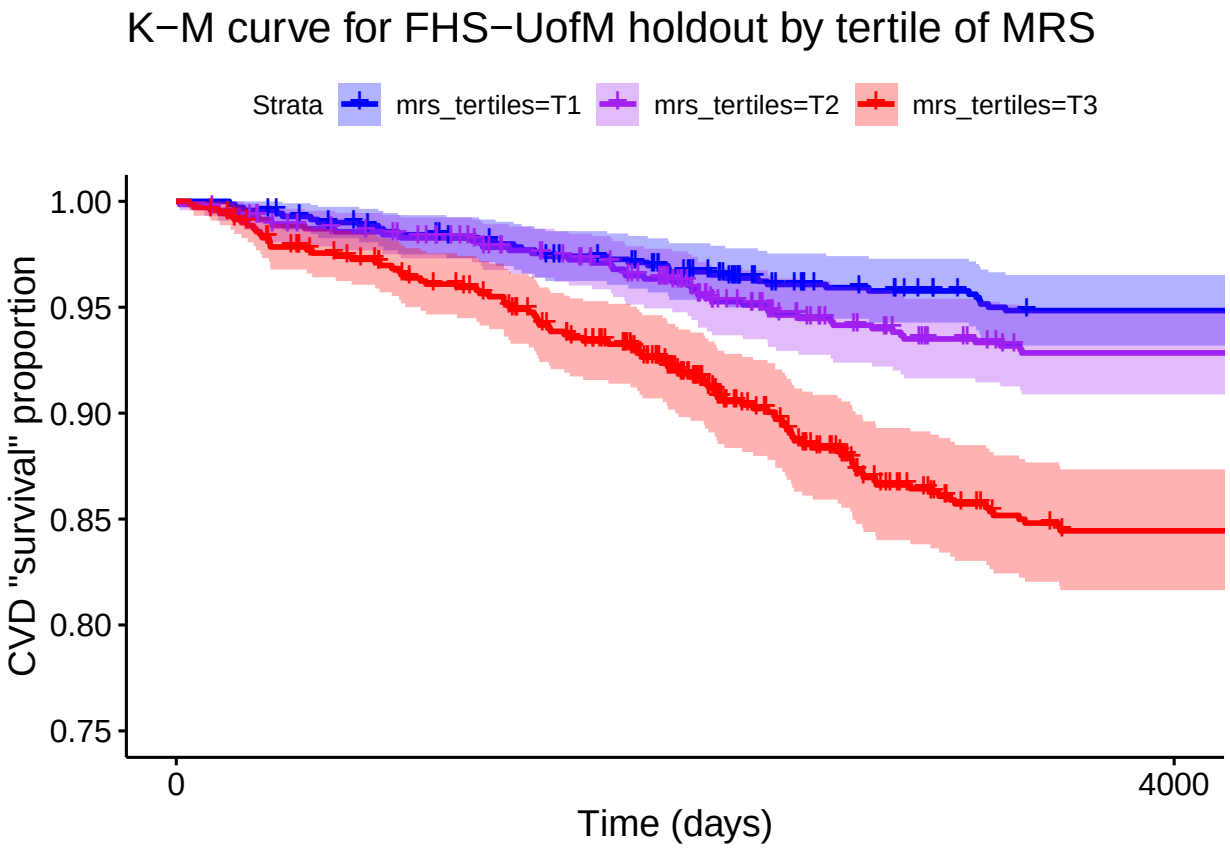


Table 2: Comparison of Cox regression coefficient estimates between the CSL and combined model options.

	HR-per-SD		
covariate_set	combined	combined_no_batch_adjust	CSL
basic	1.21	1.21	1.32
FRS_only	1.39	1.38	1.40
plus_risk_factors	1.12	1.17	1.30
unadjusted	1.51	1.50	1.62



3.3 Comparison with basic combined model

4 Final CSL model

4.1 Construction

. . . details here. . .

Table 3: 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	0.811
LBC36	Samples over multiple visits	0.678
LBC36	Samples over subsequent visits (Wave 1 & 2)	0.675
LBC36	Samples over longer time frame (Wave 1 & 3)	0.629

Table 4: Validation of Framingham Risk Score

study	HR_per_SD	p
whi	1.50	0.000
fhs_JHU	1.42	0.000
fhs_UofM	1.62	0.000
lbc36	0.88	0.042

4.2 Characterization

4.3 MRS stability

5 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. It uses the MRS derived from the U-of-M holdout set, so results corresponding to the U-of-M subset will have lower HRs and the MRS will not be biased towards strong performance.

5.0.1 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.

5.0.2 Genetic risk

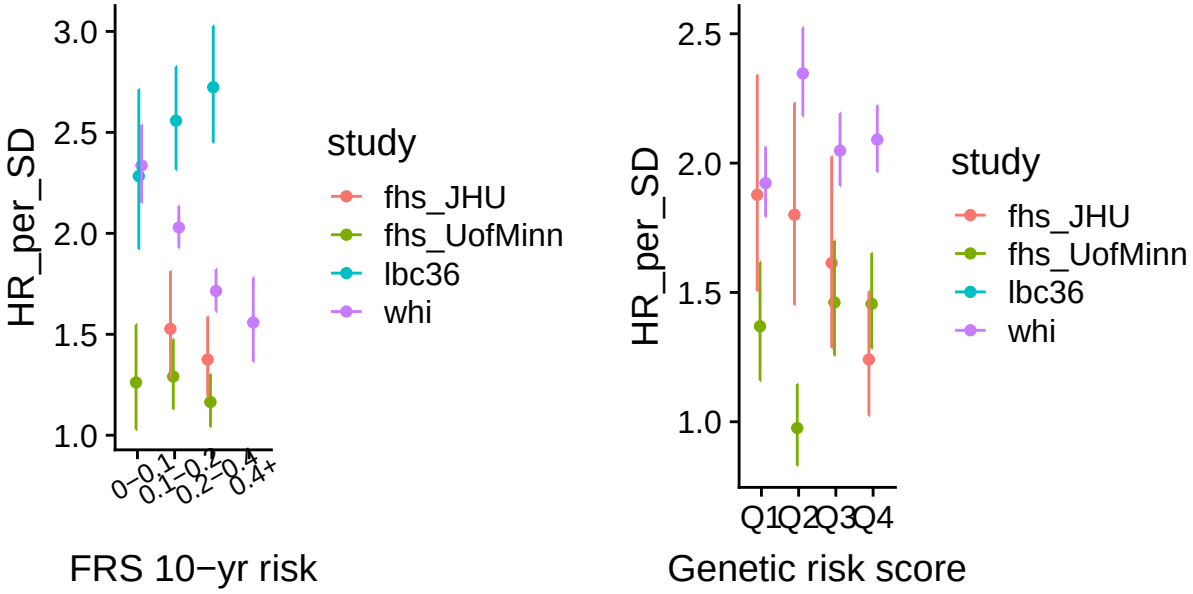
A genetic risk score (GRS) was calculated based on the genome-wide polygenic model used by Khera et al. 2018 for prediction of coronary heart disease (~6M SNPs). This GRS was first tested to confirm its associations in WHI and FHS (both of which had about 80% of the full set of GRS SNPs available after 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.

5.0.3 Stratified plots

Plots show estimated hazard ratio per std. dev. difference in MRS, with standard errors from the Cox regression. Regressions are adjusted for basic covariates (age + cell counts), and study/quantile subsets are excluded if they contain less than 25 events.

Table 5: Validation of genetic risk score

cohort	OR_per_SD	p
WHI	1.26	0.000
FHS_JHU	1.09	0.378
FHS_UofM	1.04	0.558



6 More on CSL method and rationale

6.1 "Batch effects paradox" and the utility of the cross-study learner method

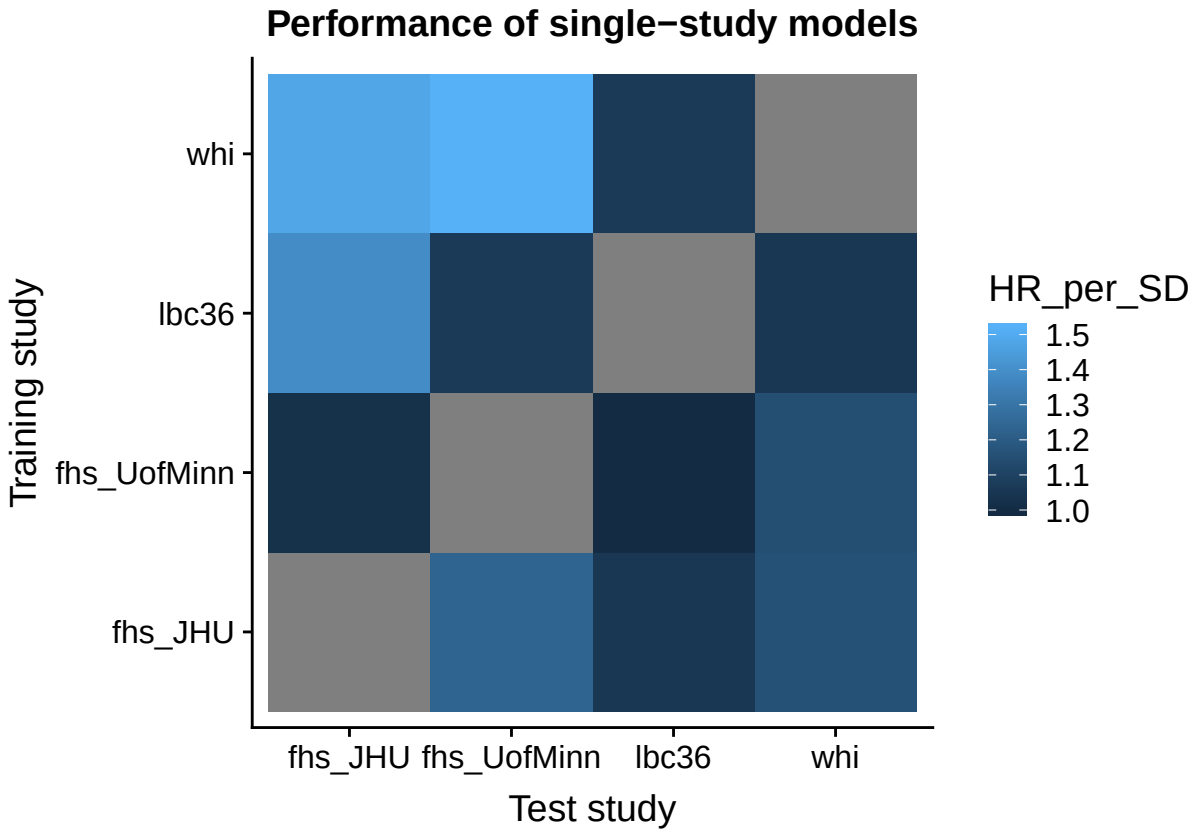
The heterogeneity of the component cohorts here (sex, race, and case/control imbalances) is a strength in some sense by allowing for a potentially more generalizable risk score, but has the major obstacle that "real" methylation differences between cohorts are indistinguishable from batch effects.

One example provides evidence that the systematic study differences are not solely due to batch effects: The single-study model figure above suggests that an LBC36-trained score does not perform strongly/consistently in WHI or FHS. However, when the LBC36 score is tested on WHI and both FHS batches together, it performs much better (HR = 1.471). This should not be due to batch effects, since *the LBC36 score has not "seen" either of the other datasets.* Indeed, it turns out that the LBC36-derived score is strongly higher in the WHI group as a whole, which has a much greater proportion of incident cases. Furthermore, this score is biologically plausible: the top-weighted CpG (cg06500161 near ABCG1) is already known to predict incident cardiometabolic disease based on prior investigations.

This observation provides support for the use of the cross-study learner model, which allows the testing of model performance in multiple cohorts with somewhat less concern that batch effects are confounding the result. In comparison, the "combined" model requires that either batch effects be allowed to confound results, or that study be directly adjusted for, which would remove the type of signal I provide evidence for above.

6.2 Out-of-sample performance – across studies

So, first, how do models trained on each single study perform in the others? While WHI and FHS predict each other well, LBC36 seems more marginally related.



Starting to use the stacking approach to combination of the SSLs, we can look at "leave one study out" performance. Each of the FHS groups can be predicted reasonably well by the others, but there is little discriminative power in WHI or LBC36.

Table 6: Leave-one-study-out performance evaluation

Study	covariate_set	HR_per_SD	p
whi	unadjusted	1.071	0.030
whi	basic	1.125	0.002
whi	plus_risk_factors	1.008	0.845
whi	FRS_only	0.963	0.238
fhs_JHU	unadjusted	1.509	0.000
fhs_JHU	basic	1.451	0.000
fhs_JHU	plus_risk_factors	1.246	0.025
fhs_JHU	FRS_only	1.410	0.000
fhs_UofMinn	unadjusted	1.617	0.000
fhs_UofMinn	basic	1.323	0.000
fhs_UofMinn	plus_risk_factors	1.298	0.001
fhs_UofMinn	FRS_only	1.399	0.000
lbc36	unadjusted	1.074	0.220
lbc36	basic	1.060	0.332
lbc36	plus_risk_factors	1.078	0.251
lbc36	FRS_only	1.106	0.092