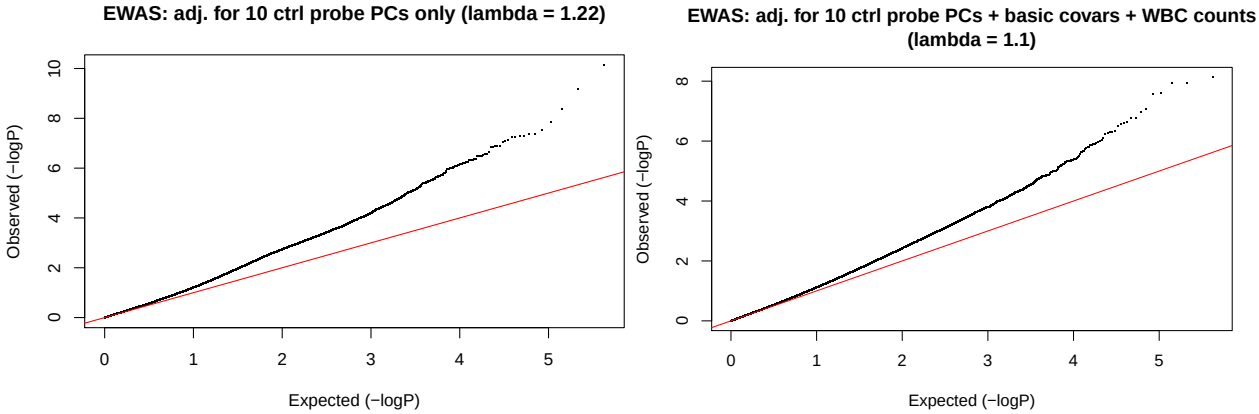
MRS Model Experimentation

Preliminaries

- Currently using two datasets: FHS Offspring and WHI
- 424348 probes from the 450k array passed QC/filtering step
- 2575 individuals from FHS, 304 of whom have an incident event
- 2023 individuals from WHI, 1020 of whom have an incident event

EWAS for incident CVD

Performed in WHI. For reference, Tsai & Bell (2015) estimate that approximately 200 case-control pairs are required to have 80% power to identify a 7% methylation difference.



Using the adjusted results, a gene set enrichment analysis was performed in the following manner:

1. Collect the set of CpGs with FDR<0.2 in EWAS
2. Assign CpGs to genes based on the available Illumina 450k annotation
3. Test for enrichment using the GOseq package, which explicitly accounts for RNA-seq length bias but can be co-opted to account for bias in number of probes per gene (Geeleher 2013)

Table 1: GO-based gene set enrichment testing

term	ontology	numDEInCat	numInCat	over_represented_pvalue	fdr
response to oxygen-containing compound	BP	81	1357	1.00e-07	0.0020941
positive regulation of mesenchymal to ep	BP	4	4	9.40e-06	0.0680629
cellular response to stimulus	BP	245	6254	1.38e-05	0.0680629
response to organic substance	BP	124	2621	1.75e-05	0.0680629
regulation of multicellular organismal p	BP	124	2454	1.77e-05	0.0680629
response to stimulus	BP	279	7490	2.16e-05	0.0680629
response to external stimulus	BP	91	1881	2.32e-05	0.0680629
small molecule biosynthetic process	BP	29	450	2.77e-05	0.0680629
regulation of cell differentiation	BP	82	1397	4.47e-05	0.0680629
regulation of epithelial cell differenti	BP	14	116	4.71e-05	0.0680629
regulation of secretion	BP	41	625	4.86e-05	0.0680629

term	ontology	numDEInCat	numInCat	over_represented_pvalue	fdr
response to chemical	BP	154	3782	5.14e-05	0.0680629
cyclic nucleotide biosynthetic process	BP	15	141	5.20e-05	0.0680629
cyclic purine nucleotide metabolic proce	BP	15	141	5.20e-05	0.0680629
chylomicron	CC	4	12	5.32e-05	0.0680629
response to biotic stimulus	BP	40	762	5.40e-05	0.0680629
response to bacterium	BP	28	477	5.67e-05	0.0680629
positive regulation of epithelial cell d	BP	9	52	5.88e-05	0.0680629
negative regulation of developmental pro	BP	50	758	6.66e-05	0.0719738
positive regulation of biological proces	BP	204	4744	7.08e-05	0.0719738

Initial MRS model selection

Choice of input CpG set

Models below are trained in WHI and the results from testing in FHS are shown. For now, EWAS-related models use *unadjusted* associations (only control-probe PCs). Regressions are adjusted only for control-probe principal components, but no biological covariates.

Table 2: Using direct weights from EWAS

HR_per_SD	p	numberOfCpGsUsed
1.801891	9.95e-28	25
1.869983	2.37e-24	50
1.926658	1.38e-23	100
1.971179	6.96e-24	250
1.895705	1.31e-20	500
1.804639	6.00e-18	1000

Table 3: Using top CpGs from EWAS as input to elastic net

HR_per_SD	p	numberOfCpGsUsed
1.148076	1.37e-04	25
1.134292	5.04e-03	50
1.175283	6.35e-04	100
1.182674	6.52e-08	250
1.209954	3.07e-08	500
1.227412	6.55e-11	1000

Table 4: Using ~1k CpGs near CVD GWAS loci as input to elastic net

HR_per_SD	p
1.097797	0.13

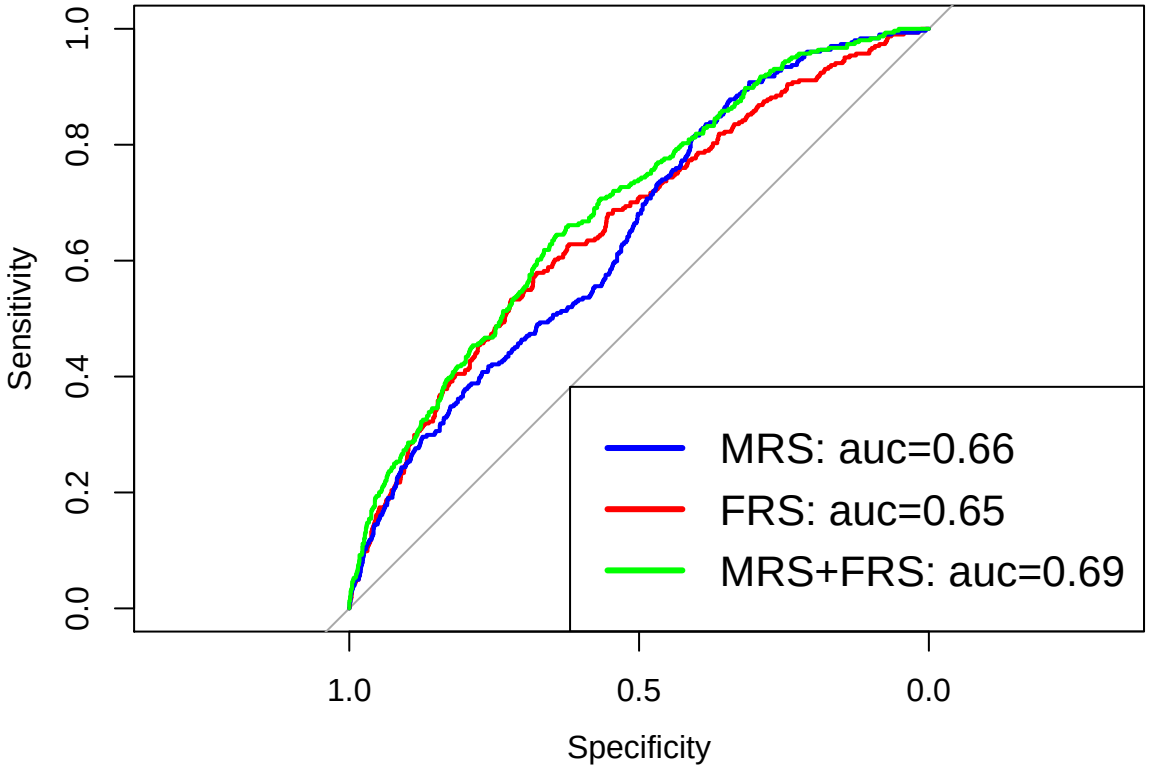
Table 5: Using a naive CpG variance threshold before running elastic net

HR_per_SD	p	varThreshold_percent
1.615744	1.87e-32	0.25
1.716940	3.77e-29	0.50
1.790277	1.55e-27	0.75
1.722101	1.23e-24	0.90

The direct weighted sum from EWAS shows the highest estimated HRs, but the significance is not quite as strong as that from the "naive" elastic net regression. Using these EWAS CpGs as input to a regression seems to be much worse, but one might expect the algorithm to overfit/insufficiently penalize the coefficients since the CpG set was found in the entire training set (WHI) – I have confirmed that manually increasing the penalty parameter can increase the predictive significance to a level greater than the weighted sum. Likewise, the CVD biology-based CpG set over-penalizes the coefficients because it doesn't take into account the higher likelihood that this biologically motivated feature set is a priori more likely to generalize. Using naive variance thresholds before regression shows solid performance, so the 75% threshold (top 25% most variable CpGs) will be used moving forward to compromise between performance and computational requirements.

Evaluation

(Using the 75th quantile variance threshold CpG set elastic net model)



For appropriate comparison to the Framingham Risk Score, the evaluation here is based on a binary classification of whether an incident event did or did not occur. The basic MRS shows a reasonable c-statistic (ROC area under the curve) of 0.66, comparable to that of my implementation of the FRS algorithm. There seems to be a bit of synergy between the two risk scores, as performance increases slightly when using them together. One important caveat is that the FRS is generally estimated to show a c-statistic closer to 0.75, so

it is possible that FHS Offspring is not an optimal test cohort (because cholesterol is not predictive of CVD risk in this population?) or that there is some error in my coding of the FRS.

Are the patterns identified specific to a population subset?

(Using the 75th quantile variance threshold CpG set elastic net model)

Though the MRS above is trained in a female-only cohort, there is no evidence here of any bias of the MRS towards predicting events in females. Based on previous results, there was also reason to believe that prior CVD could be an important confounder – this could manifest as a much better performance of the classifier in those who experienced prior CVD events, but there doesn't seem to be any such effect here.

Table 6: Stratify by sex

HR_per_SD	p	sex
1.844053	4.06e-16	Male
1.718230	3.54e-12	Female

Table 7: Stratify by whether FHS subjects experienced a past CVD event

HR_per_SD	p	pastEvent
1.392058	6.09e-04	Yes
1.760440	5.72e-17	No

Exploration of model variants

To facilitate exploration of binary classifiers, testing from this point on uses logistic regression/odds ratios to evaluate MRS performance rather than Cox regression/hazard ratios. The CpG set used is still the 75th quantile variance threshold set.

Table 8: Test the same 75th quantile MRS from above for binary CVD event outcome

OR_per_SD	p
1.800139	9.05e-22

Table 9: Training and testing using a binary outcome

OR_per_SD	p
1.65512	1.25e-15

Table 10: Using M-values instead of beta values

OR_per_SD	p
1.74153	4.17e-20

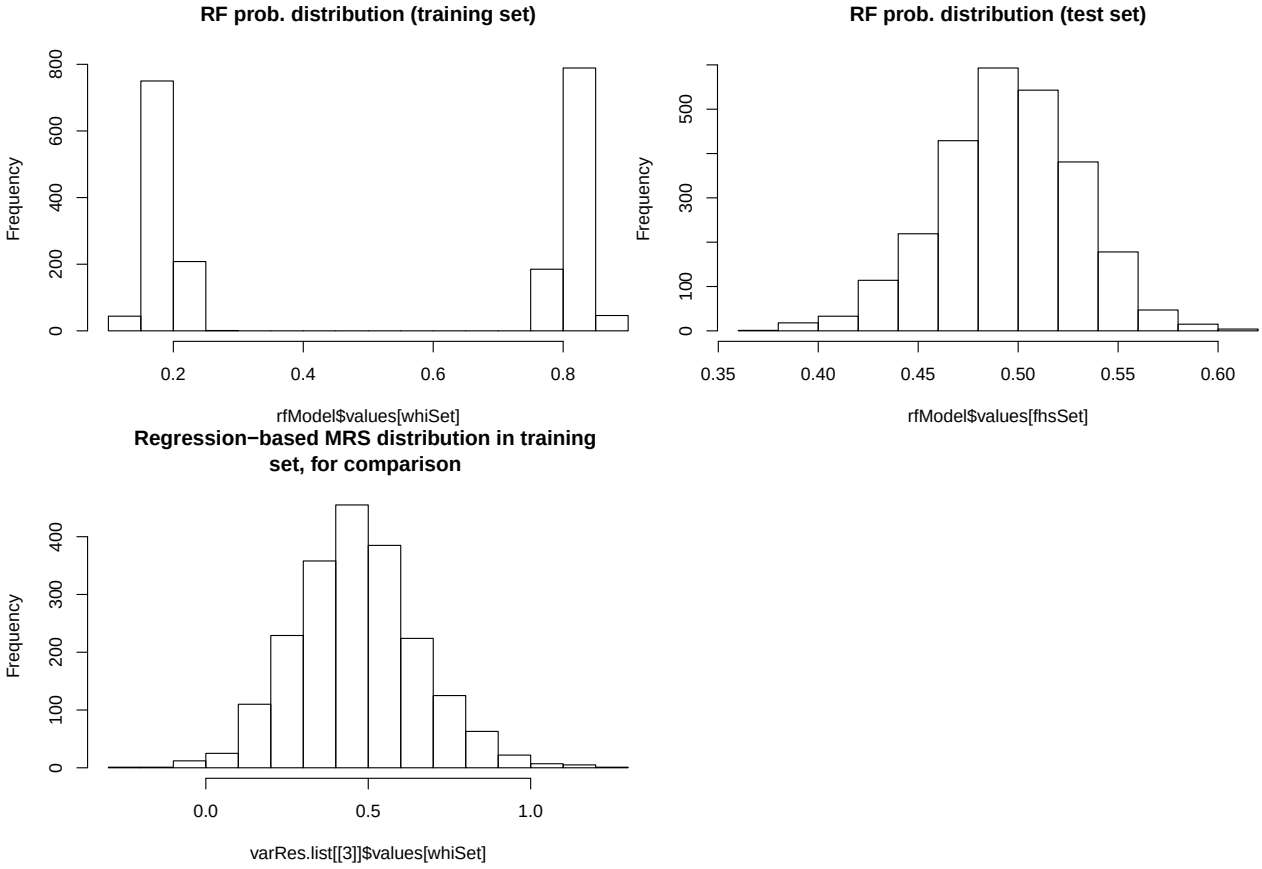
Table 11: Random forest model

OR_per_SD	p
2.118368	0.0464

The first two tables above compare models that are both tested using binary classifiers, but one is trained using a Cox model while the other is trained through a logistic model. As might be expected, a notable increase in performance is achieved by the Cox model, as it is able to take advantage of additional time-to-event information.

Performance of the full training and testing procedure using M-values does not seem to improve performance.

A random forest model trained on the same set of CpGs and a binary event outcome shows a comparably high OR, but the p-value doesn't indicate notable predictive ability – what is going on here? To investigate, here are the distributions of the probabilities output by the random forest algorithm (used here as the MRS).



Clearly, the random forest is overfitting – it is way too good at classifying incident events in the training set. Likely it is identifying WHI-specific population specific or batch effects, especially since I was not able to adjust for control probe PCs in the random forest model training. So, it may be helpful to combine my two datasets before creating a train-test split in order to mitigate these population-specific patterns.

Cross-cohort training

Here, the populations were combined and the full dataset was randomly divided into a 70/30 train/test split, while ensuring proportional numbers of incident events in each set.

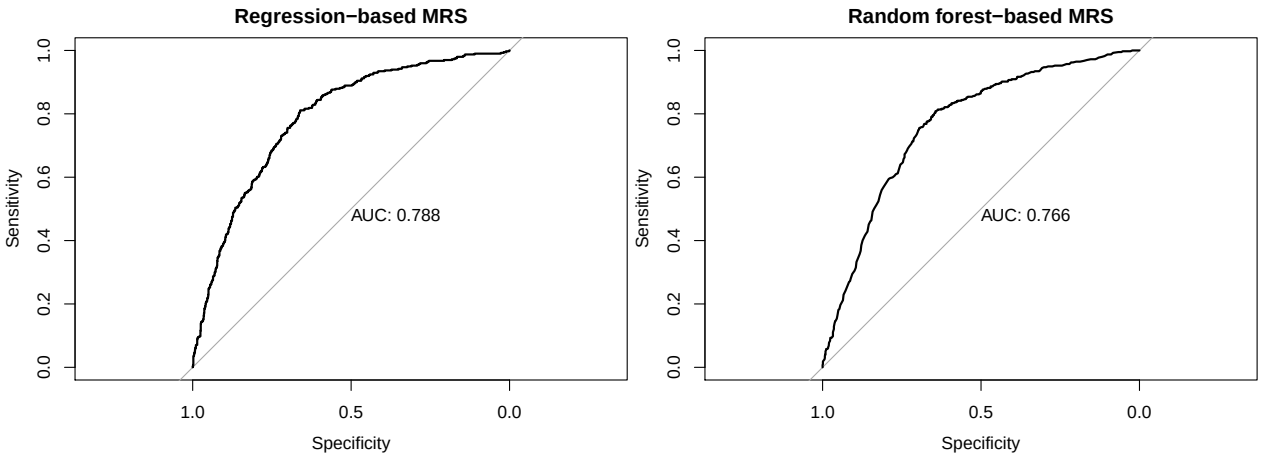
Table 12: Elastic net regression

OR_per_SD	p	model
1.800139	9.05e-22	WHI train, FHS test
3.308466	2.67e-52	Combined 70/30 split

Table 13: Random forest

OR_per_SD	p	model
2.118368	0.0464	WHI train, FHS test
5.049813	1.49306742359257e-45	Combined 70/30 split

Evaluation of cross-cohort trained models



Is this model fitting only dataset-specific patterns?

Try to evaluate this by fitting separate models in WHI and FHS.

Table 14: Train and test in FHS only

OR_per_SD	p
1.59198	2.2e-06

Table 15: train and test in WHI only

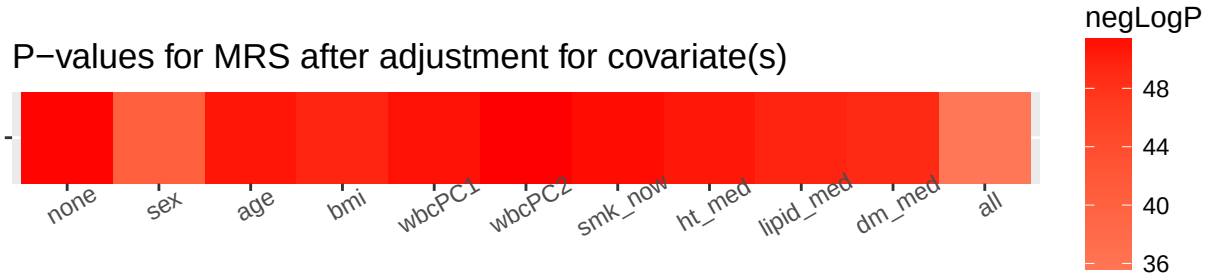
OR_per_SD	p
1.313228	0.0015172

Though ideally there would be a fully independent validation set available, the mediocre within-cohort cross-validation results don't indicate any population-specific effects of a strength that would dominate the above combined regression results.

Biological significance of the MRS

Biological covariates

The intent to this point was to determine the "maximum predictive power" of microarray-based methylation data in predicting CVD events. However, known associations of methylation with sex, age, white blood cell proportions, etc. may explain a substantial fraction of this predictive power.



The AUC of MRS residuals after adjustment for the full set of covariates above is 0.74.

Specific CpG components/weights

Table 16: MRS components CpGs and associated genes

CpG	weight	gene	Literature association
cg21843755	-1.5720291	ZNF638	
cg09741592	-0.9084210	HNRNPA1	smoking
cg09741592	-0.9084210	HNRPA1L-2	smoking
cg18181703	-0.7259223	SOCS3	metabolic syndrome
cg00534468	0.6921193	CMIP	cholesterol efflux capacity
cg22407942	-0.6905793	SNORD93	
cg16867657	0.6281935	ELOVL2	aging (strongly)
cg19693031	-0.5608651	TXNIP	T2D
cg14428733	-0.5209318	HLA-DPB1	
cg26634409	0.4723838	ANP32A	
cg03498895	0.4703612	MYC	
cg00854392	0.4593796	SEPT1	high-fat overfeeding in subQ AT
cg06048973	0.4530961	ACTC1	
cg23287661	-0.4242700	S100A7A	T2D in AT
cg23289079	0.4004365	PRDM6	cancer
cg02113429	0.3976318	OVOL2	
cg25130381	0.3948994	SLC9A1	
cg15925505	-0.3349900	C6orf89	
cg05892167	0.3017754	ROBO1	
cg21917349	0.2898946	APBA2	
cg22004443	0.2680577	PDIA6	
cg22525895	-0.2617878	MIR29B2	
cg18328334	-0.2616762	TNS1	
cg13056927	0.2476976	KBTBD5	
cg14627974	0.2432533	MIR548I4	
cg14627974	0.2432533	CNTNAP2	
cg23997263	0.2424078	FUT9	
cg21740507	0.2324585	IQCA1	

CpG	weight	gene	Literature association
cg24694018	-0.2316633	POLR3GL	
cg23580000	0.2292735	ADCY7	
cg10917602	-0.2286279	HSD3B7	
cg10243348	0.2229748	ITGBL1	
cg25410668	0.2045030	RPA2	
cg23256951	0.2030831	MUC1	
cg21429551	-0.1966234	GARS	
cg07801516	0.1947829	ZNF461	
cg24723140	0.1864230	DPP10	
cg21812670	-0.1696115	SNORD45C	
cg21812670	-0.1696115	RABGGTB	
cg20923885	0.1660029	SEPT9	
cg20098015	-0.1648668	ODF3B	
cg23465749	0.1637560	ARID2	
cg12121643	0.1553623	DGKG	
cg15243034	0.1549155	USP35	
cg13958389	0.1476441	VGLL4	
cg00996053	-0.1411271	SKINTL	
cg08143875	0.1342484	F10	
cg01749249	-0.1273334	TTBK1	
cg12161971	0.1228376	USP35	
cg01243823	-0.1225371	NOD2	
cg09580755	0.1101103	ST6GALNAC3	
cg02996131	-0.1094380	SYNE1	
cg04387347	0.1080093	ZFPM1	
cg27152890	0.1067089	PPP1R13L	
cg22454769	0.1025002	FHL2	
cg24141036	0.0963956	PLEKHA2	
cg09053611	0.0905772	NDUFA5	
cg17293936	0.0872323	DGKG	
cg09090376	0.0862058	DEPDC7	
cg13161901	0.0791287	CDH20	
cg12157673	0.0789443	ZNF578	
cg03757250	0.0700495	TMIGD2	
cg03306615	0.0635137	ASCL2	
cg17351376	0.0612684	CD248	
cg05630111	0.0501493	LASS2	
cg25608547	0.0484148	LAMA2	
cg13606990	0.0445420	ZNF397OS	
cg19584649	0.0434165	ESYT2	
cg04098547	0.0400544	ABCC12	
cg03764818	0.0369207	GALNT9	
cg02634816	-0.0325780	UBOX5	
cg02634816	-0.0325780	FASTKD5	
cg27256066	-0.0317808	IRF8	
cg08122652	-0.0255872	PARP9	
cg08122652	-0.0255872	DTX3L	
cg04208124	0.0233099	AHNAK	
cg07029084	0.0055795	FAM170A	
cg11705975	0.0008842	PRLHR	

. . . the catch, or "Batch effects: 1, Kenny: 0"

It seems that I eventually got bitten by the class imbalance between FHS (mostly controls) and WHI (half cases) – MRS residuals after adjustment for study lose much of their predictive power (AUC = 0.6198295).