Supplemental information

Selection of 51 predictors from 13,782 candidate

multimodal features using machine learning

improves coronary artery disease prediction

Saaket Agrawal, Marcus D.R. Klarqvist, Connor Emdin, Aniruddh P. Patel, Manish D. Paranjpe, Patrick T. Ellinor, Anthony Philippakis, Kenney Ng, Puneet Batra, and Amit V. Khera

TABLE OF CONTENTS

SUPPLEMENTARY TABLES

Table S1 UK Biobank codes used for excluding individuals on the basis of prevalent disease

UK Biobank Field ID	Description	Codings
6150	Vascular/heart problems diagnosed by doctor	1, 2
20002	Non-cancer illness code, self-reported	1088, 1087, 1067, 1074, 1075, 1079, 1076, 1588, 1491, 1583, 1081, 1086, 1082
20004	Operation code	1440, 1102, 1108, 1109, 1105, 1110, 1070, 1095, 1523, 1551, 1550, 1096
40001	Underlying (primary) cause of death: ICD10	I70.0, I70.00, I70.01, I70.2, I70.20, I70.21, I70.8, I70.80, I70.9, I70.90, I73.8, I73.9, I20, I20.0, I20.1, I20.8, I20.9, I21, I21.0, I21.1, I21.2, I21.3, I21.4, I21.9, I22, I22.0, I22.1, I22.8, I22.9, I23, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.8, I24, I24.0, I24.1, I24.8, I24.9, I25.1, I25.2, I25.5, I25.6, I25.8, I25.9
40002	Contributory (secondary) causes of death: ICD10	I70.0, I70.00, I70.01, I70.2, I70.20, I70.21, I70.8, I70.80, I70.9, I70.90, I73.8, I73.9, I20, I20.0, I20.1, I20.8, I20.9, I21, I21.0, I21.1, I21.2, I21.3, I21.4, I21.9, I22, I22.0, I22.1, I22.8, I22.9, I23, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.8, I24, I24.0, I24.1, I24.8, I24.9, I25.1, I25.2, I25.5, I25.6, I25.8, I25.9
41200	Operative procedures - main OPCS4	X09.3, X09.4, X09.5, L21.6, L51.3, L51.6, L51.8, L52.1, L52.2, L54.1, L54.4, L54.8, L59.1, L59.2, L59.3, L59.4, L59.5, L59.6, L59.7, L59.8, L60.1, L60.2, L63.1, L63.5, L63.9, L66.7, K40, K40.1, K40.2, K40.3, K40.4, K40.8, K40.9, K41, K41.1, K41.2, K41.3, K41.4, K41.8, K41.9, K42, K42.1, K42.2, K42.3, K42.4, K42.8, K42.9, K43, K43.1, K43.2, K43.3, K43.4, K43.8, K43.9, K44, K44.1, K44.2, K44.8, K44.9, K45.1, K45.2, K45.3, K45.4, K45.5, K45.6, K45.8, K45.9, K46, K46.1, K46.2, K46.3, K46.4, K46.5, K46.8, K46.9, K49.1, K49.2, K49.3, K49.4, K49.8, K49.9, K50.1, K50.2, K50.4, K75.1, K75.2, K75.3, K75.4, K75.8, K75.9
41202	Diagnoses - main ICD10	I70.0, I70.00, I70.01, I70.2, I70.20, I70.21, I70.8, I70.80, I70.9, I70.90, I73.8, I73.9, I20, I20.0, I20.1, I20.8, I20.9, I21, I21.0, I21.1, I21.2, I21.3, I21.4, I21.9, I22, I22.0, I22.1, I22.8, I22.9, I23, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.8, I24, I24.0, I24.1, I24.8, I24.9, I25.1, I25.2, I25.5, I25.6, I25.8, I25.9
41203	Diagnoses - main ICD9	4400, 4402, 4438, 4439, 410, 4109, 411, 4119, 412, 4129, 413, 4139, 4140, 4148, 4149
41204	Diagnoses - secondary ICD10	I70.0, I70.00, I70.01, I70.2, I70.20, I70.21, I70.8, I70.80, I70.9, I70.90, I73.8, I73.9, I20, I20.0, I20.1, I20.8, I20.9, I21, I21.0, I21.1, I21.2, I21.3, I21.4, I21.9, I22, I22.0, I22.1, I22.8, I22.9, I23, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.8, I24, I24.0, I24.1, I24.8, I24.9, I25.1, I25.2, I25.5, I25.6, I25.8, I25.9
41205	Diagnoses - secondary ICD9	4400, 4402, 4438, 4439, 410, 4109, 411, 4119, 412, 4129, 413, 4139, 4140, 4148, 4149
41210	Operative procedures - secondary OPCS4	X09.3, X09.4, X09.5, L21.6, L51.3, L51.6, L51.8, L52.1, L52.2, L54.1, L54.4, L54.8, L59.1, L59.2, L59.3, L59.4, L59.5, L59.6, L59.7, L59.8, L60.1, L60.2, L63.1, L63.5, L63.9, L66.7, K40, K40.1, K40.2, K40.3, K40.4, K40.8, K40.9, K41, K41.1, K41.2, K41.3, K41.4, K41.8, K41.9, K42, K42.1, K42.2, K42.3, K42.4, K42.8, K42.9, K43, K43.1, K43.2, K43.3, K43.4, K43.8, K43.9, K44, K44.1, K44.2, K44.8, K44.9, K45.1, K45.2, K45.3, K45.4, K45.5, K45.6, K45.8, K45.9, K46, K46.1, K46.2, K46.3, K46.4, K46.5, K46.8, K46.9, K49.1, K49.2, K49.3, K49.4, K49.8, K49.9, K50.1, K50.2, K50.4, K75.1, K75.2, K75.3, K75.4, K75.8, K75.9
41270	Diagnoses - ICD10	I700, I702, I708, I709, I738, I739, I701, I740, I741, I742, I743, I744, I745, I748, I749, I200, I201, I208, I209, I210, I211, I212, I213, I214, I219, I220, I221, I228, I229, I231, I232, I236, I238, I240, I241, I248, I249, I251, I252, I255, I256, I258, I259, I250, I510, I511, I512, R931, T825, T827, T828, T829, Z034, Z951, Z955, Z958, Z959, I110, I130, I132, I420, I421, I422, I425, I428, I429, I500, I501, I509, I426, I427, I430, I431, I438, I517, J81, I600, I601, I602, I603, I604, I605, I606, I607, I608, I609, I610, I611, I612, I613, I614, I615, I616, I618, I619, I630, I631, I632, I633, I634, I635, I636, I638, I639, I64, G450, G451, G458, G459, G461, G462, G463, G464, G465, G466, G467, G468, I629, I650, I651, I652, I653, I658, I659, I660, I662, I663, I664, I668, I669, I672,

		I679, I690, I691, I692, I693, I694, I698, I460, I469, I490, R071, R072, R073, R074, R943
41272	Operative procedures - OPCS4	L216, L518, L541, L544, L548, L591, L592, L593, L594, L595, L596, L597, L598, L601, L602, L631, L635, L639, X094, K471, L161, L162, L201, L204, L205, L206, L208, L211, L212, L213, L218, L251, L253, L261, L262, L263, L265, L266, L267, L291, L293, L294, L295, L303, L311, L314, L351, L353, L371, L378, L381, L382, L383, L384, L388, L391, L392, L393, L395, L398, L412, L419, L421, L422, L423, L428, L431, L432, L433, L435, L438, L439, L451, L452, L463, L464, L468, L471, L472, L474, L503, L505, L506, L508, L511, L512, L513, L514, L515, L516, L519, L521, L522, L528, L531, L532, L542, L549, L581, L582, L583, L584, L585, L599, L603, L604, L608, L621, L622, L632, L633, L682, L711, L761, L762, L763, L765, L766, L768, L769, L895, L896, L899, K401, K402, K403, K404, K408, K409, K411, K412, K413, K414, K422, K424, K431, K433, K439, K441, K442, K448, K449, K451, K452, K453, K454, K455, K456, K458, K459, K461, K462, K463, K464, K469, K491, K492, K493, K494, K498, K499, K501, K502, K504, K751, K752, K753, K754, K758, K759, Y731, Y732
42007	Source of stroke report	0, 1, 2

Table S2 Breakdown of individuals who were missing data for continuous traits in the UK Biobank

UK Biobank Field ID	Field	Participants	Missing	Missing (%)
31	Sex	502,504	0	0.00%
48	Waist circumference	502,504	2,160	0.43%
49	Hip circumference	502,504	2,219	0.44%
50	Standing height	502,504	2,539	0.51%
21001	Body mass index (BMI)	502,504	3,105	0.62%
21002	Weight	502,504	2,774	0.55%
21022	Age at recruitment	502,504	0	0.00%
102	Pulse rate, automated reading	502,504	27,375	5.45%
4079	Diastolic blood pressure, automated reading	502,504	30,113	5.99%
4080	Systolic blood pressure, automated reading	502,504	30,118	5.99%
30000	White blood cell (leukocyte) count	502,504	24,336	4.84%
30010	Red blood cell (erythrocyte) count	502,504	24,331	4.84%
30020	Haemoglobin concentration	502,504	24,331	4.84%
30030	Haematocrit percentage	502,504	24,331	4.84%
30040	Mean corpuscular volume	502,504	24,333	4.84%
30050	Mean corpuscular haemoglobin	502,504	24,334	4.84%
30060	Mean corpuscular haemoglobin concentration	502,504	24,338	4.84%
30070	Red blood cell (erythrocyte) distribution width	502,504	24,333	4.84%
30080	Platelet count	502,504	24,334	4.84%
30090	Platelet crit	502,504	24,338	4.84%
30100	Mean platelet (thrombocyte) volume	502,504	24,339	4.84%
30110	Platelet distribution width	502,504	24,339	4.84%

30120	Lymphocyte count	502,504	25,220	5.02%
30130	Monocyte count	502,504	25,220	5.02%
30140	Neutrophill count	502,504	25,220	5.02%
30150	Eosinophill count	502,504	25,220	5.02%
30160	Basophill count	502,504	25,220	5.02%
30170	Nucleated red blood cell count	502,504	25,231	5.02%
30180	Lymphocyte percentage	502,504	25,214	5.02%
30190	Monocyte percentage	502,504	25,214	5.02%
30200	Neutrophill percentage	502,504	25,214	5.02%
30210	Eosinophill percentage	502,504	25,214	5.02%
30220	Basophill percentage	502,504	25,214	5.02%
30230	Nucleated red blood cell percentage	502,504	25,235	5.02%
30240	Reticulocyte percentage	502,504	32,648	6.50%
30250	Reticulocyte count	502,504	32,648	6.50%
30260	Mean reticulocyte volume	502,504	32,649	6.50%
30270	Mean sphered cell volume	502,504	32,647	6.50%
30280	Immature reticulocyte fraction	502,504	32,649	6.50%
30290	High light scatter reticulocyte percentage	502,504	32,647	6.50%
30300	High light scatter reticulocyte count	502,504	32,648	6.50%
N/A	Black Ethnicity	502,504	97,020	19.31%
N/A	White Ethnicity	502,504	97,020	19.31%
N/A	Normalized polygenic risk score	502,504	97,020	19.31%
N/A	Sotuheast Asian Ethnicity	502,504	97,020	19.31%
N/A	PRS principle component 4	502,504	97,020	19.31%
N/A	PRS principle component 3	502,504	97,020	19.31%
N/A	PRS principle component 2	502,504	97,020	19.31%
N/A	PRS principle component 1	502,504	97,020	19.31%
N/A	East Asian Ethnicity	502,504	97,020	19.31%
30600	Albumin	502,504	72,429	14.41%

30610	Alkaline phosphatase	502,504	32,902	6.55%
30620	Alanine aminotransferase	502,504	33,103	6.59%
30630	Apolipoprotein A	502,504	74,994	14.92%
30640	Apolipoprotein B	502,504	35,298	7.02%
30650	Aspartate aminotransferase	502,504	34,708	6.91%
30670	Urea	502,504	33,233	6.61%
30680	Calcium	502,504	72,572	14.44%
30690	Cholesterol	502,504	32,915	6.55%
30700	Creatinine	502,504	33,147	6.60%
30710	C-reactive protein	502,504	33,936	6.75%
30720	Cystatin C	502,504	32,946	6.56%
30730	Gamma glutamyltransferase	502,504	33,162	6.60%
30740	Glucose	502,504	72,937	14.51%
30750	Glycated haemoglobin (HbA1c)	502,504	36,000	7.16%
30760	HDL cholesterol	502,504	72,633	14.45%
30770	IGF-1	502,504	35,464	7.06%
30780	LDL direct	502,504	33,798	6.73%
30810	Phosphate	502,504	73,253	14.58%
30830	SHBG	502,504	76,658	15.26%
30840	Total bilirubin	502,504	34,944	6.95%
30850	Testosterone	502,504	77,297	15.38%
30860	Total protein	502,504	72,900	14.51%
30870	Triglycerides	502,504	33,290	6.62%
30880	Urate	502,504	33,479	6.66%
30890	Vitamin D	502,504	54,154	10.78%
N/A	Lipoprotein A	502,504	43,618	8.68%
N/A	Weekly American standardized drinks	502,504	0	0.00%
189	Townsend deprivation index at recruitment	502,504	623	0.12%

Table S3 Breakdown of individuals who were missing data from nurse interview questions in the UK Biobank

| Field | UK Biobank Field ID | N | Responses (N) | Prefer not to answer | | Missing |
				N	%	%
Illnesses of father	20107	502,504	487,770	1,102	0.23%	3.25%
Illnesses of mother	20110	502,504	492,908	941	0.19%	2.14%
Illnesses of siblings	20111	502,504	433,903	869	0.20%	16.01%
Medication for cholesterol, blood pressure or diabetes (males); Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones (females)	6177, 6153	502,504	497,788	638	0.13%	1.08%
Smoking status	20116	502,504	501,613	2,057	0.41%	0.59%
Weight change compared with 1 year ago	2306	502,504	501,584	424	0.08%	0.27%
Chest pain or discomfort	2335	502,504	501,581	706	0.14%	0.32%
Vascular/heart problems diagnosed by doctor	6150	502,504	501,575	1,294	0.26%	0.44%
Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor	6152	502,504	501,573	1,210	0.24%	0.43%
Diabetes diagnosed by doctor	2443	502,504	501,573	404	0.08%	0.27%
Average total household income before tax	738	502,504	500,247	49,846	9.96%	10.42%

Table S4 Baseline characteristics in UK Biobank stratified by inclusion status

	Excluded (N = 329,230)	Included (Development + Holdout) (N = 173,274)
Age (years)	57.5 (8.1)	56.2 (8.1)
Men	140,620 (42.7%)	88,502 (51.1%)
Ethnicity Group		
White	306,993 (93.2%)	165,702 (95.6%)
Black	6,662 (2.0%)	2,444 (1.4%)
East Asian	2,878 (0.9%)	1,385 (0.8%)
South Asian	6,008 (1.8%)	2,016 (1.2%)
Other	6,689 (2.0%)	1,727 (1.0%)
Current smoker	34,878 (10.7%)	18,105 (10.4%)
Diabetes	19,645 (6.0%)	8,203 (4.7%)
Cholesterol (mg/dL)	221.1 (47.5)	217.5 (37.7)
HDL-C (mg/dL)	56.3 (15.3)	55.4 (13.9)
LDL-C (mg/dL)	137.9 (35.9)	136.2 (29.2)
SBP (mmHg)	138.1 (18.8)	137.5 (18.4)
Antihypertensive	76,112 (23.1%)	32,601 (18.8%)
CAD Polygenic Risk Score (CAD PRS)	0.02 (1.00)	-0.03 (0.99)
Incident CAD events over median 11-year follow-up	11,382 (3.7%)	5,140 (3.0%)
Predicted 10 year risk (%)		
Framingham risk score (FRS)	6.9 (6.4)	6.9 (6.4)
Pooled cohort equations (PCE)	8.8 (7.9)	8.3 (7.7)
QRISK3-2017 (QRISK3)	11.1 (9.3)	10.0 (8.4)

Table S5 UK Biobank Codes included in Coronary Artery Disease (CAD) Outcome Definition

ICD-10 (41202, 41204)	ICD-10 Cause of Death (40001, 40002)	ICD-9 (41203, 41205)	OPCS-4 (41200, 41210)
I21	I21	410	K40
I22	I22	411	K41
I23	I23	412	K42
I24	I24		K43
I25.2	I25.2		K44
			K45
			K46
			K49
			K50
			K75

Table S9 C-statistics for ML4H$_{EN-COX}$ and clinical risk algorithms in the **development** cohort.

Model (C-statistic)	Entire development (n = 138,619)	Male (n = 70,896)	Female (n = 67,723)	Age < 55 (n = 59,728)	Age >= 55 (n = 78,891)
ML4H$_{EN-COX}$	0.790 (0.783, 0.797)	0.741 (0.733, 0.749)	0.769 (0.753, 0.784)	0.808 (0.793, 0.823)	0.747 (0.739, 0.755)
FRS	0.751 (0.744, 0.758)	0.680 (0.671, 0.689)	0.741 (0.725, 0.757)	0.770 (0.753, 0.784)	0.703 (0.694, 0.711)
PCE	0.750 (0.743, 0.757)	0.680 (0.671, 0.689)	0.739 (0.723, 0.754)	0.773 (0.755, 0.787)	0.700 (0.691, 0.709)
QRISK3	0.756 (0.749, 0.764)	0.688 (0.679, 0.697)	0.739 (0.724, 0.757)	0.775 (0.759, 0.789)	0.708 (0.699, 0.718)

ML4HEN-COX performs similarly in the development cohort as in the holdout cohort (**Table 3**), suggesting that no overfitting has occurred. Bootstrapped 95% confidence intervals indicated in parentheses.

Table S11 C-statistics for XGBoost, SimpleCox51, and SimpleCox20

Model (C-statistic)	Entire holdout (n = 34,655)	Men (n = 17,606)	Women (n = 17,049)	Age < 55 (n = 15,134)	Age >= 55 (n = 19,521)
ML4H$_{EN\text{-}COX}$	0.796 (0.784, 0.809) ref	0.751 (0.735, 0.767) ref	0.780 (0.747,0.811) ref	0.825 (0.799, 0.850) ref	0.755 (0.737, 0.771) ref
XGBoost	0.797 (0.784, 0.810) p = 0.847	0.752 (0.736, 0.770) p = 0.976	0.784 (0.755, 0.813) p = 0.495	0.832 (0.805, 0.855) p = 0.189	0.754 (0.736, 0.770) p = 0.632
SimpleCox51	0.797 (0.784, 0.811) p = 0.350	0.753 (0.736, 0.769) p = 0.637	0.785 (0.753, 0.816) p = 0.041	0.827 (0.800, 0.853) p = 0.383	0.756 (0.739, 0.772) p = 0.802
SimpleCox20	0.794 (0.781, 0.807) p = 0.171	0.749 (0.732, 0.764) p = 0.303	0.777 (0.745, 0.808) p = 0.411	0.820 (0.793, 0.846) p = 0.124	0.753 (0.736, 0.770) p = 0.422

XGBoost, SimpleCox51, and SimpleCox20 have similar discriminatory performance as ML4H$_{EN\text{-}COX}$ across all studied subgroups in the holdout cohort. Bootstrapped 95% confidence intervals indicated in parentheses. P values listed below each C-statistic correspond to DeLong's test comparing each C-statistic to reference (ML4H$_{EN\text{-}COX}$).

Table S12 Categorical reclassification indices for XGBoost, SimpleCox51, and SimpleCox20

Categorical NRI cutoff	Comparator model		
ML4H$_{EN\text{-}COX}$	FRS	PCE	QRISK3
2.5%	6.0% (3.5-8.6%)	6.6% (4.1-9.1%)	5.8% (3.3-8.3%)
5.0%	6.1% (3.1-9.1%)	8.2% (5.1-11.2%)	7.5% (4.6-10.5%)
XGBoost			
2.5%	5.9% (3.3-8.5%)	6.4% (3.8-9.0%)	5.6% (3.1-8.2%)
5.0%	7.9% (5.0-10.9%)	10.0% (6.9-13.1%)	9.3% (6.4-12.2%)
SimpleCox51			
2.5%	6.6% (4.0-9.2%)	7.1% (4.6-9.7%)	6.3% (3.8-8.9%)
5.0%	6.9% (3.9-9.9%)	9.0% (5.9-12.0%)	8.2% (5.2-11.2%)
SimpleCox20			
2.5%	5.8% (3.2-8.4%)	6.3% (3.8-8.9%)	5.5% (3.0-8.1%)
5.0%	6.8% (3.7-9.8%)	8.8% (5.7-11.9%)	8.2% (5.2-11.2%)
ML4H$_{EN\text{-}COX}$ (unknown GPS$_{CAD}$)			
2.5%	2.1% (0.0-4.3%)	2.7% (0.7-4.7%)	1.8% (0.0-3.9%)
5.0%	2.4% (0.0-5.2%)	4.5% (1.9-7.1%)	3.9% (1.4-6.4%)

Note that the ML4H$_{EN\text{-}COX}$ results are identical to those of Table 4 in the main text and are reproduced here for comparison with alternate models XGBoost, SimpleCox51, and SimpleCox20.

An additional sensitivity analysis (ML4H$_{EN\text{-}COX}$ (unknown GPS$_{CAD}$)) is shown to simulate the situation where GPS$_{CAD}$ is unknown and the mean value is assumed.

**FIGURE S1** Correlation plot of the 51 predictors selected by ML4H$_{EN\text{-}COX}$.
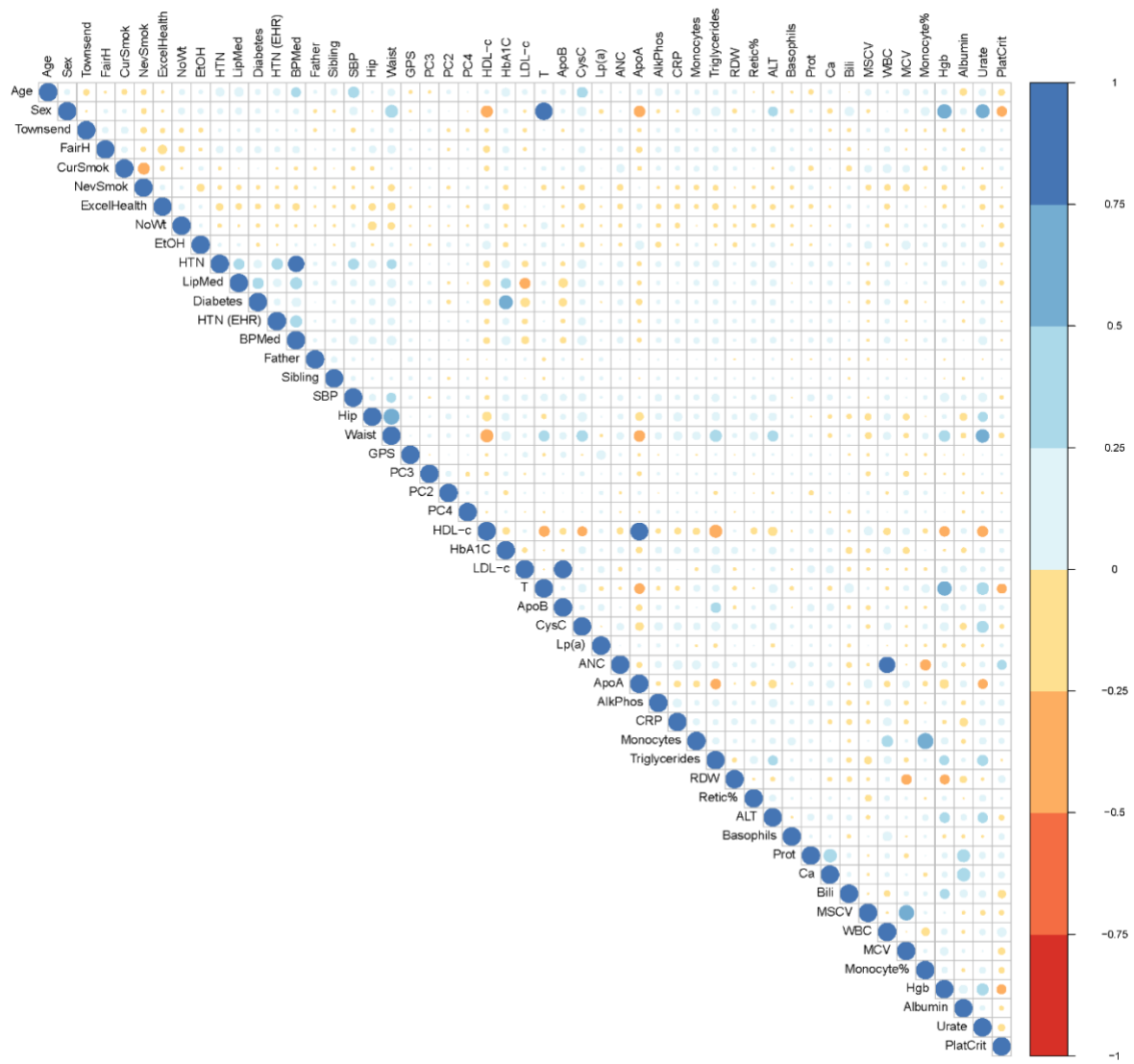
Note that GPS (genome-wide polygenic score for CAD) is not strongly associated with any of the other 50 traits. Five pairs of features had Pearson correlation coefficient magnitude greater than 0.75: Sex/Testosterone, self-reported history of HTN/history of antihypertensive, LDL-c/ApoB, HDL-c/ApoA, Neutrophils/WBC count.

FIGURE S2 Principal components 1 and 2 of genetic ancestry (left), and principal components 3 and 4 (right) plotted for the analyzed cohort colored by self-identified ethnic groups.
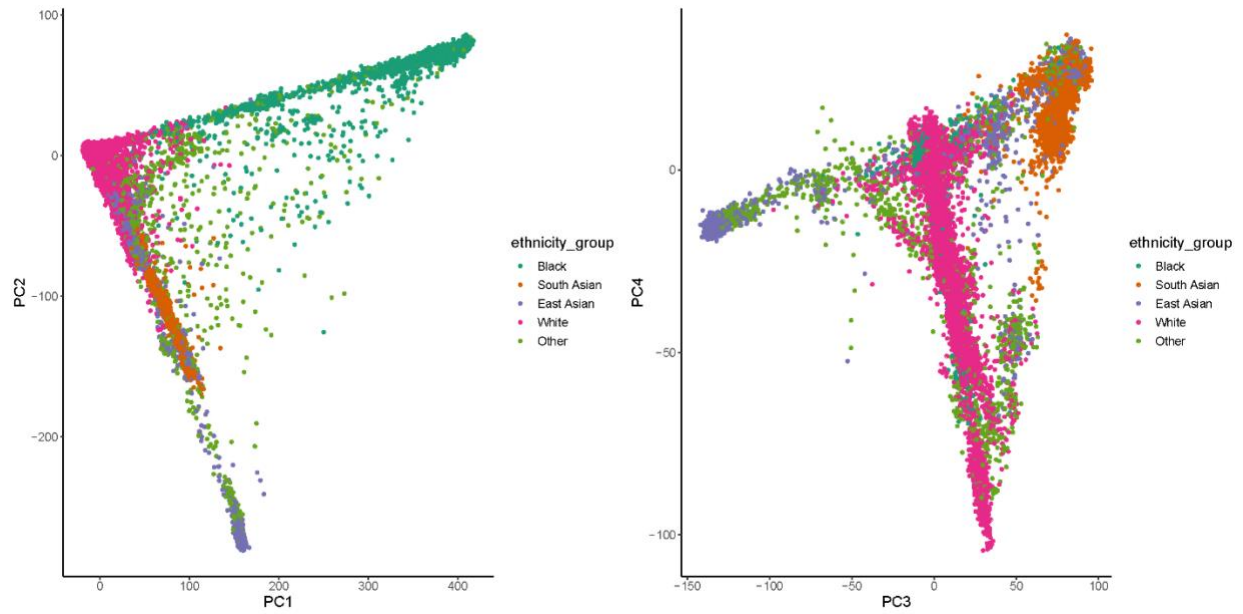

Individuals of South Asian ancestry (orange) are concentrated at the top right of PC3/PC4 plot.

FIGURE S3 Calibration plots for ML4H$_{EN\text{-}COX}$ in the development (left) and holdout (right) cohorts.
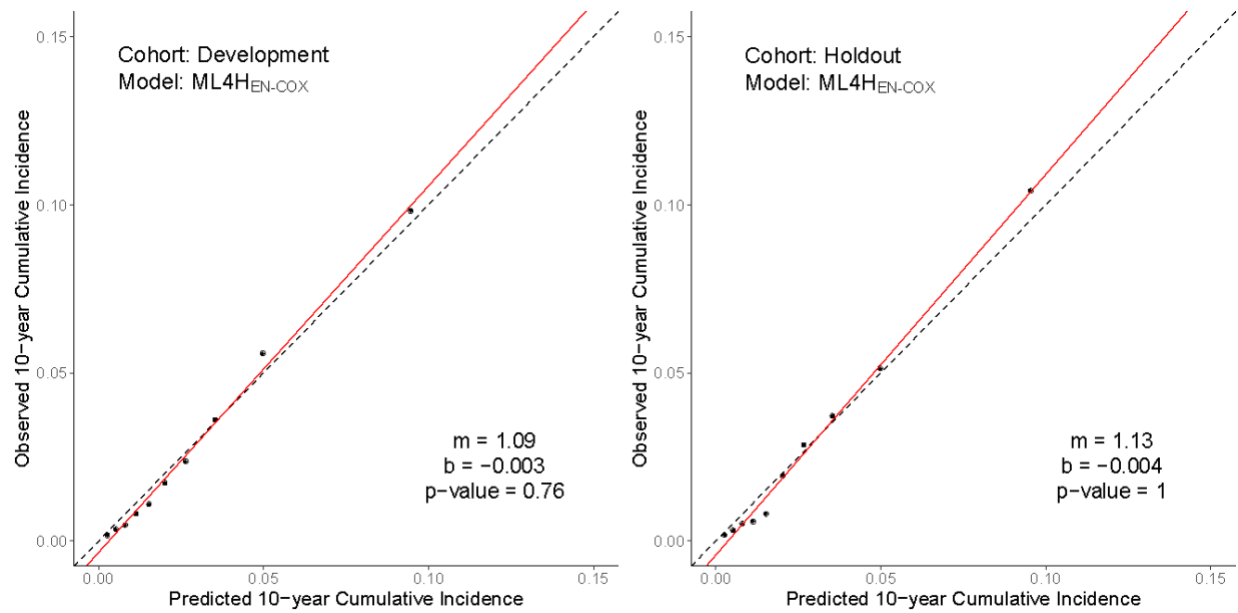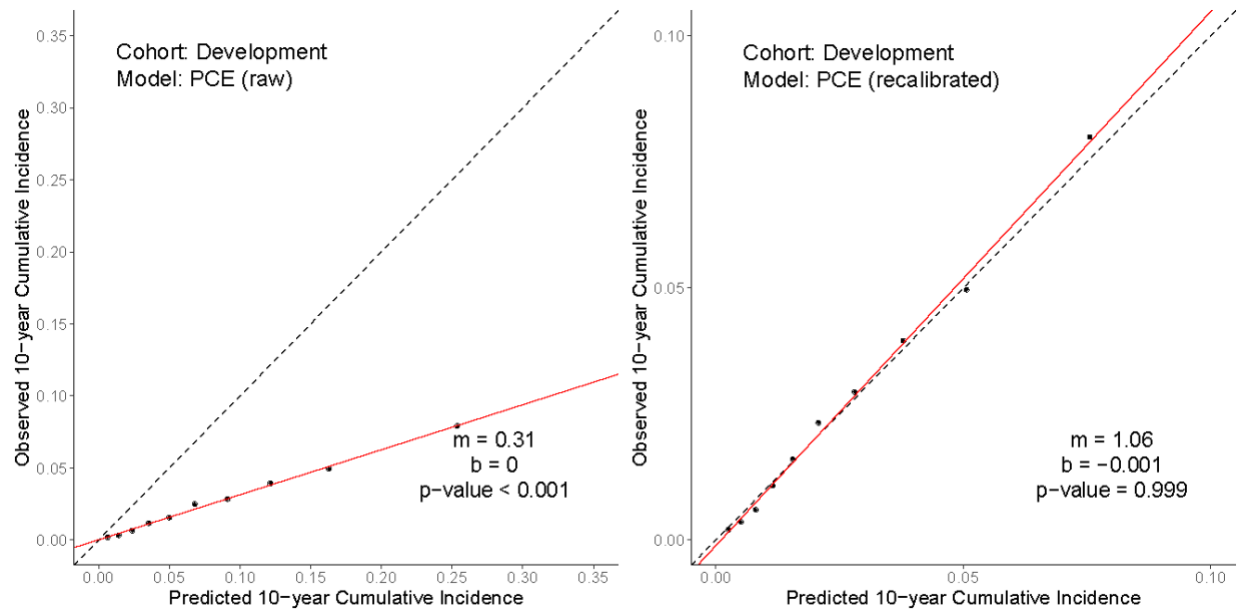

ML4H$_{EN\text{-}COX}$ was well-calibrated in the development and holdout cohorts. Slope (m), y-intercept (b), and Hosmer-Lemeshow significance test (p-value) are displayed.

FIGURE S4 Calibration plots for Framingham risk score (left) and re-calibrated Framingham risk score (right) in the development cohort.
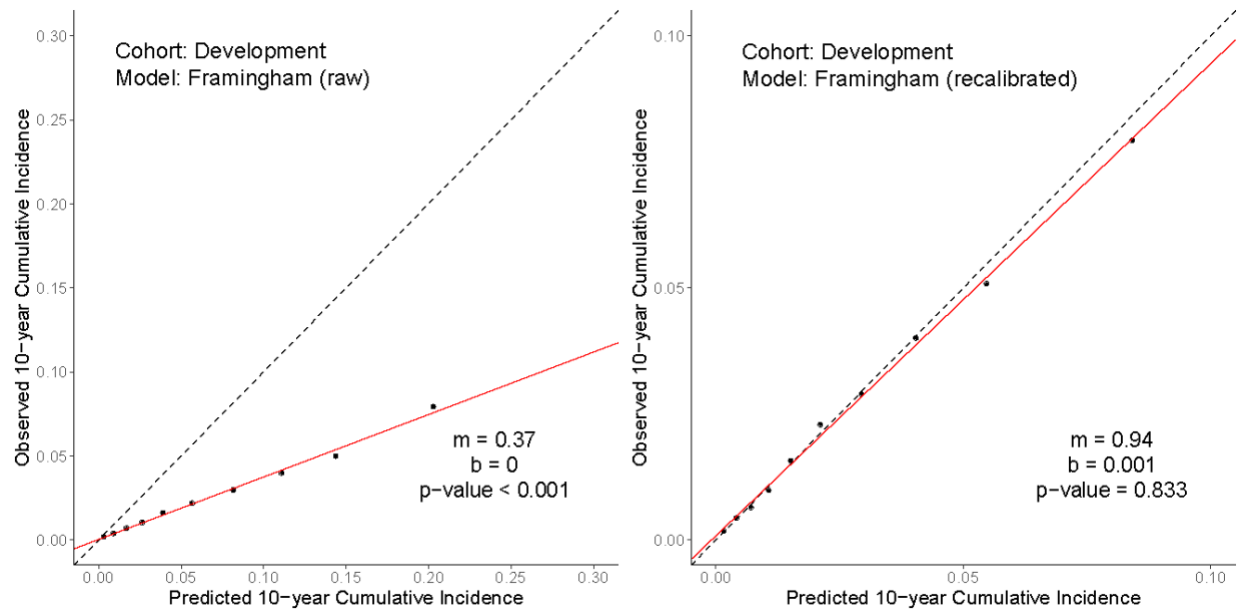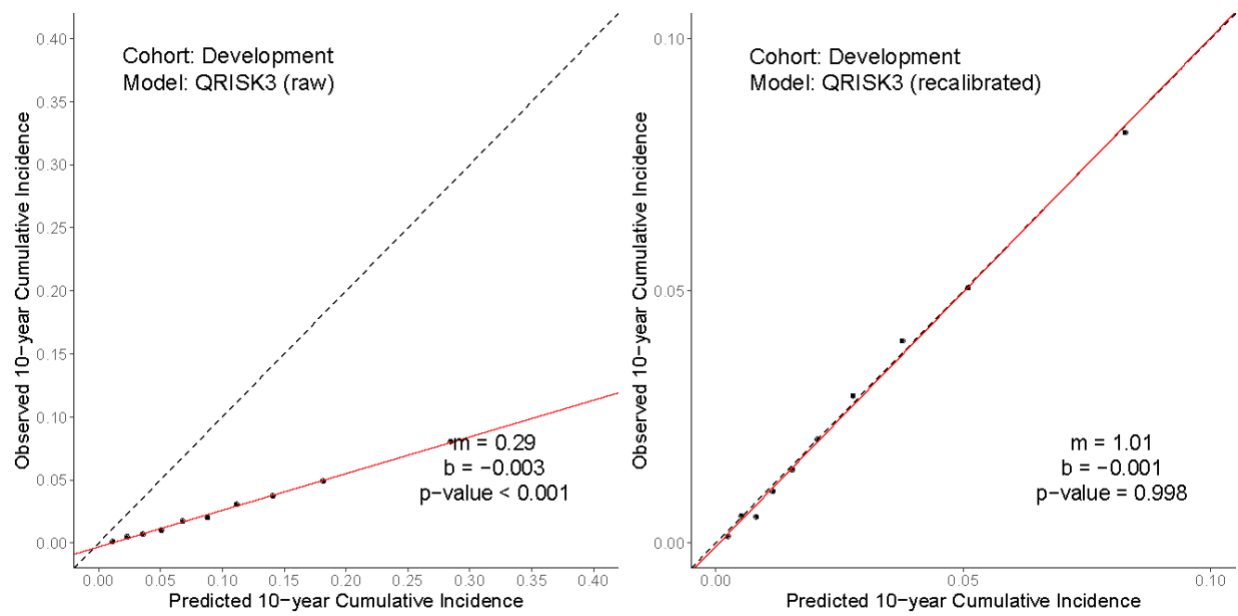

Framingham Risk Score for CAD was successfully recalibrated in the development cohort. Slope (m), y-intercept (b), and Hosmer-Lemeshow significance test (p-value) are displayed.

FIGURE S5 Calibration plots for Pooled Cohort Equations (left) and re-calibrated Pooled Cohort Equations (right) in the development cohort.
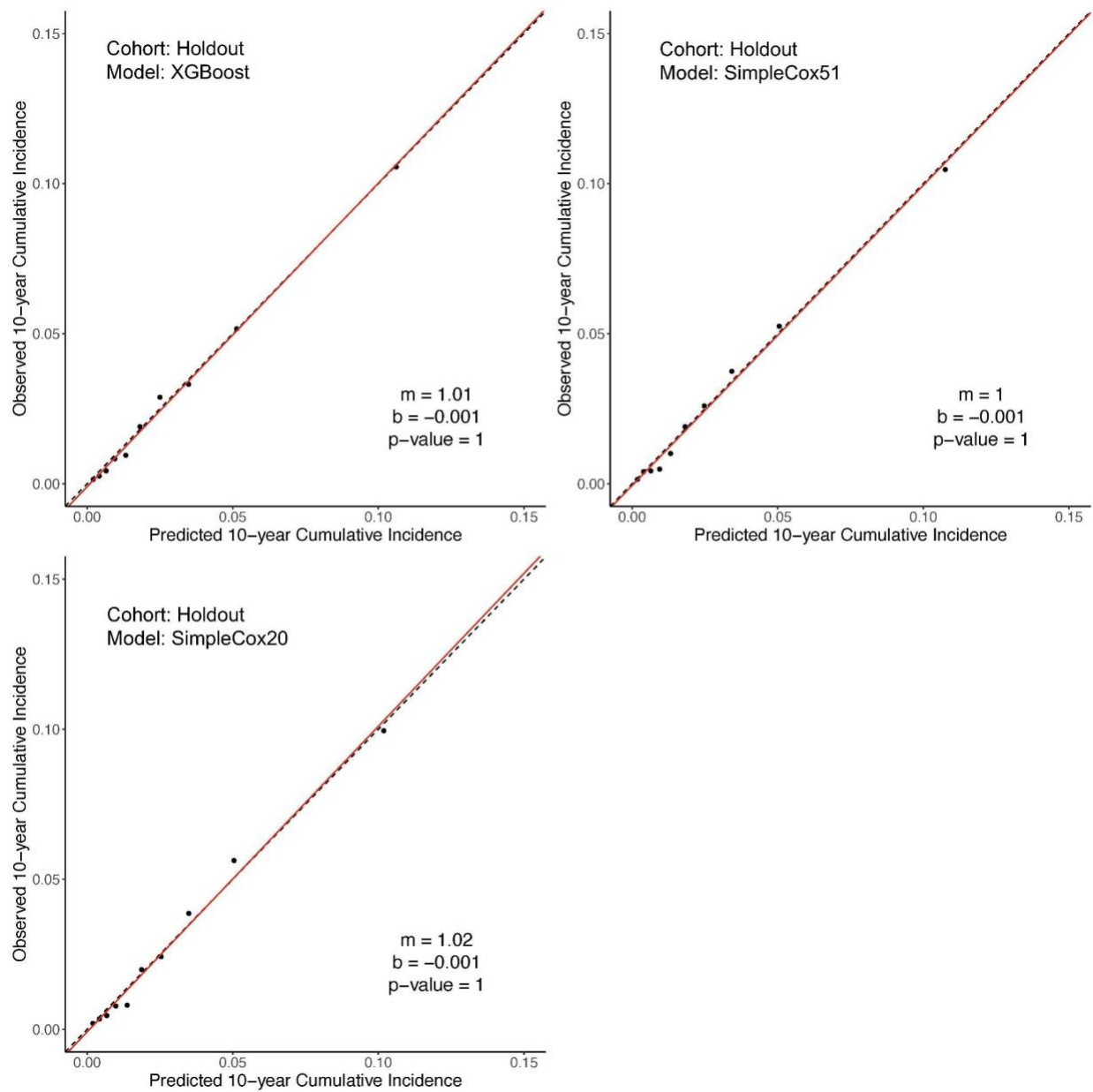

Pooled Cohort Equations predictions were successfully recalibrated in the development cohort. Slope (m), y-intercept (b), and Hosmer-Lemeshow significance test (p-value) are displayed.

FIGURE S6 Calibration plots for QRISK3 (left) and re-calibrated QRISK3 (right) in the development cohort.

QRISK3 was successfully recalibrated in the development cohort. Slope (m), y-intercept (b), and Hosmer-Lemeshow significance test (p-value) are displayed.

FIGURE S7 Calibration plots for XGBoost (top left), SimpleCox51 (top right), and SimpleCox20 (bottom left) in the holdout cohort.

XGBoost, SimpleCox51, and SimpleCox20 models were well-calibrated in the holdout cohort. Slope (m), y-intercept (b), and Hosmer-Lemeshow significance test (p-value) are displayed.

FIGURE S8 Distribution of predicted 10-year risk of CAD for ML4H$_{EN\text{-}COX}$, FRS, PCE, QRISK3, SimpleCox51, and XGBoost in the holdout cohort.