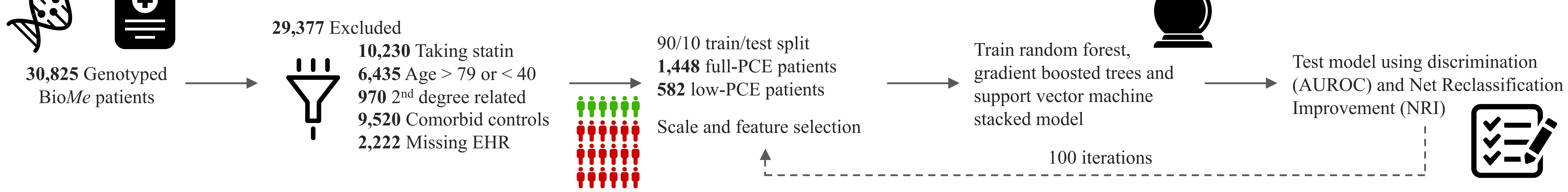
Coronary Risk Estimation Based on Clinical Data in Electronic Health Records

Ben O. Petrazzini, Kumardeep Chaudhary, Carla Márquez-Luna, Iain S. Forrest, Ghislain Rocheleau, Judy Cho, Jagat Narula, Girish Nadkarni and Ron Do

The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Background: The Pooled Cohort Equations[1] (PCE) is used guide cholesterol-lowering medication (statin) by assessing the 10-year risk for a first atherosclerotic event. Limitations of the PCE include overestimation[2,3] and bias[4-6] in certain non-European populations. Recent efforts have evaluated the predictive performance of polygenic risk scores (PRS) for coronary artery disease (CAD)[7]. However, their clinical utility remains unclear[8,9] suggesting a need for alternative approaches. Agnostic feature evaluation of machine learning workflows powered by clinical data in electronic health records (EHR) can identify non-traditional risk factors currently ignored for CAD prognosis. The accumulation of plaque in coronary arteries manifests as a complex trait; thus, additional features could carry novel evidence of CAD susceptibility. Here, we develop a short-term CAD risk prediction tool to test weather EHR clinical features can improve PCE-based CAD risk assessment and evaluated its performance in two large EHR-linked biobanks.

Figure 1. Study workflow.



30,825 Genotyped Bio*Me* patients

29,377 Excluded
10,230 Taking statin
6,435 Age > 79 or < 40
970 2nd degree related
9,520 Comorbid controls
2,222 Missing EHR

90/10 train/test split
1,448 full-PCE patients
582 low-PCE patients

Scale and feature selection

Train random forest, gradient boosted trees and support vector machine stacked model

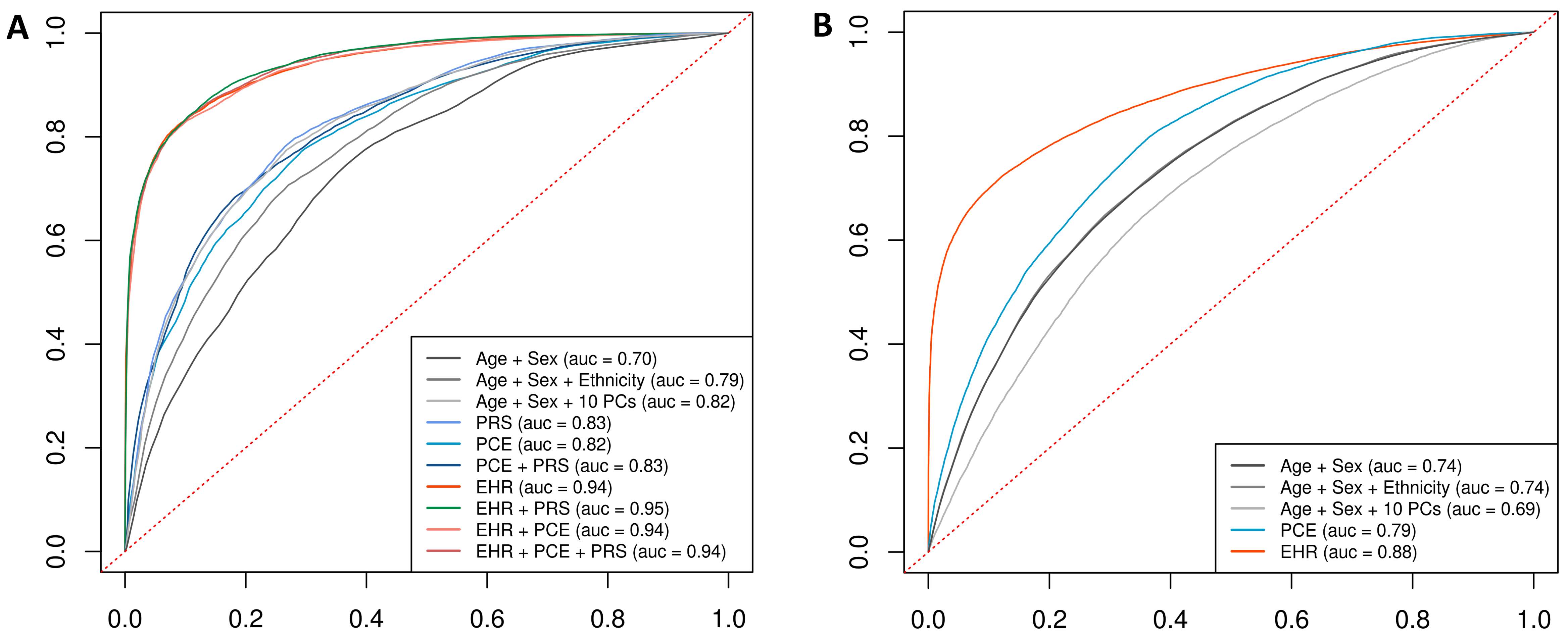
Test model using discrimination (AUROC) and Net Reclassification Improvement (NRI)

100 iterations

Methods: We collected EHR and genetic data from Bio*Me* and the UK Biobank. PCE and PRS were calculated for both datasets. Models integrating combinations of PCE, PRS and EHR data were trained and tested on Bio*Me*, then validated in the UK Biobank. Cases were identified using ICD codes, all EHR one year prior to diagnosis was removed. We trained Random Forest, Gradient Boosted Trees and Support Vector Machine models on a balanced set using 90% of cases, then regressed it's predictions to obtain a unique stacked model. Performance was tested on a balanced 10% dataset blind to scaling and feature selection. This was repeated 100 times to avoid sampling biases (fig. 1). Reported results are the mean area under the receiver operator characteristic (AUROC) curve and Net Reclassification Improvement (NRI) across 100 models. Each model was then validated on a balanced set of UK Biobank participants. Performance was evaluated in parallel on a subpopulation of low-risk individuals (PCE<7.5).

Results: For the EHR alone, discrimination was AUROC=0.94 and positive predictive value (PPV) was 0.88 (fig. 2A and tab. 1A). This was 12% higher than the PCE alone which yielded an AUROC of 0.82 and PPV of 0.74. In low-risk individuals this difference almost doubles (20%), with 0.87 and 0.67 AUROC for the EHR and PCE models, respectively. Similar results were observed on the independent cohort (fig. 2B and tab. 1B). Including the PRS to either model showed no improvement in prediction power (fig. 2). Analyses in the UK Biobank show the EHR model corrects risk overestimation originating from the PCE[2,3] with NRI=28.7% in healthy individuals (tab. 2) and a 36% decrease of false positives in the top 15% of the score (fig. 3). A small number of risk factors are driving the EHR model, with 0.94 AUROC attained using only 10 features; 9 being non-traditional risk factors not used by the PCE such as diagnosed hypertension (I10), depression (F32.9), red blood cell distribution width, basophil, and hemoglobin A1c.

Figure 2. Receiver operator characteristic curves in Bio*Me* (A) and UK Biobank (B). Y and X axes correspond to averaged true positive and false negative rates respectively across 100 iterations.



Conclusions: The EHR score can correct risk overestimation stemming from the PCE. This suggests that the inclusion of non-traditional risk factors can improve one-year risk prediction for CAD over conventional risk assessment tools. Furthermore, the implementation of an EHR score in hospital settings can potentially enable systematic identification of high-risk individuals otherwise undetected by current clinical practices.

Figure 3. Change in prediction score between the PCE and EHR model in UK Biobank individuals.

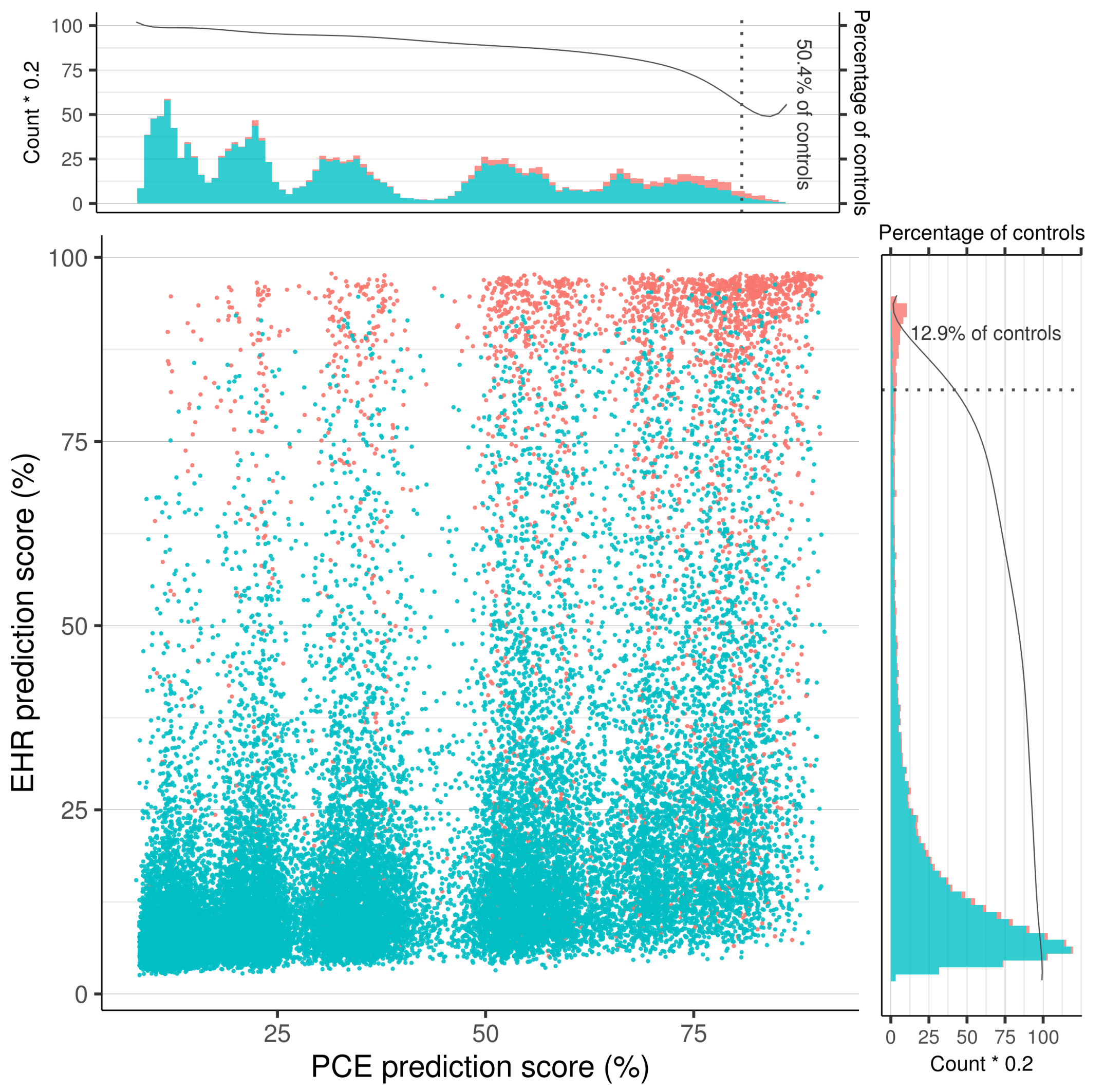


Table 1. Mean performance metrics in Bio*Me* (A) and UK Biobank (B) individuals.

A	**All**			**Low-risk**		
Model	**AUROC**	**PPV**	**NPV**	**AUROC**	**PPV**	**NPV**
Age + Sex	0.70 (0.04)	0.68 (0.04)	0.69 (0.04)	0.64 (0.12)	0.60 (0.17)	0.63 (0.13)
PCE	0.82 (0.04)	0.74 (0.04)	0.73 (0.05)	0.67 (0.13)	0.61 (0.11)	0.63 (0.15)
EHR	**0.94 (0.02)**	**0.88 (0.04)**	**0.85 (0.04)**	**0.87 (0.07)**	**0.81 (0.10)**	**0.78 (0.10)**

B	**All**			**Low-risk**		
Model	**AUROC**	**PPV**	**NPV**	**AUROC**	**PPV**	**NPV**
Age + Sex	0.74 (0.02)	0.67 (0.02)	0.68 (0.02)	0.59 (0.02)	0.57 (0.01)	0.57 (0.03)
PCE	0.79 (0.02)	0.69 (0.02)	0.75 (0.03)	0.69 (0.01)	0.62 (0.02)	0.65 (0.03)
EHR	**0.88 (0.01)**	**0.92 (0.02)**	**0.73 (0.02)**	**0.80 (0.04)**	**0.68 (0.10)**	**0.87 (0.02)**

Table 2. Net reclassification improvement.

	Bio*Me*			**UK Biobank**		
Model	**Overall**	**Cases**	**Control**	**Overall**	**Cases**	**Controls**
Age + Sex	-9.6 (7.8)	-1.8 (7.0)	-7.8 (7.1)	-7.9 (2.8)	-8.9 (4.0)	1.0 (4.1)
EHR	**25.8 (8.5)**	**12.4 (7.5)**	**13.4 (5.6)**	**15.2 (4.1)**	**-13.5 (4.7)**	**28.7 (3.9)**

References:

1. Goff DC, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. Circulation. 2013.
2. Ridker PM and Cook NR. Statins: new American guidelines for prevention of cardiovascular disease. The Lancet. 2013.
3. Kavousi M, et al. Comparison of Application of the ACC/AHA Guidelines […]. JAMA. 2014.
4. Muntner P, et al. Validation of the Atherosclerotic Cardiovascular Disease Pooled Cohort Risk Equations. JAMA. 2014.
5. DeFilippis AP, et al. Risk score overestimation: the impact of individual […]. EHJ. 2016.
6. Rana JS, et al. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Real-World Population. JACC. 2016.
7. Rotter JI and Lin HJ. An Outbreak of Polygenic Scores for Coronary Artery Disease. JACC. 2020.
8. Mosley JD, et al. Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. JAMA. 2020.
9. Elliott J, et al. Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA. 2020.

Scan to access article 🔍