Supplemental Material

DATA S1: IMBALANCED DATA: NEATER

The MESA data are severely imbalanced in terms of outcomes, that is, the size of the class with events (i.e., minority class) is much smaller than the size of the class without events (i.e., majority class), and thus the decision boundary for ML methods would be severely biased and could result in poor performance. To cope with this skewed class distribution issue, we selected the filtering of over-sampled data using non-cooperative game theory (NEATER) algorithm[2], which is an oversampling data augmentation algorithm that employs cooperative game theory to generate artificial data of the minority class. Non-cooperative game theory[3] addresses the interaction between individual rational decision makers, where all the data are players and the goal is to uniformly and consistently label all of the synthetic data created by any oversampling technique. Unlike other over-sampling approaches, NEATER does not automatically consider synthetic data as part of the minority class. Instead, it keeps synthetic samples unlabeled, at first. These samples then participate in a non-cooperative game to determine their most likely class membership, minority or majority. All the synthetic data that end up belonging to the minority class are kept, and the rest are eliminated.

A detailed description of the main steps of the NEATER implementation in this work can be summarized as follows. First, an oversampled method is used, such as the Synthetic Minority Over-Sampling Technique (SMOTE)[4] to generate synthetic data. The use of SMOTE is justified by the fact that it creates samples that are closely related to the minority class, which causes the classifier to create larger decision regions. Then, both the original and synthetically generated data are considered as players, and the possible class memberships are considered strategies available to all game players. Note that, only the synthetic data play to determine their class membership. There are two types of players: I_c, which denotes players that already belong to a

class, and I_u, which denotes unlabeled players or synthetic samples. Each I_u player interacts with a number of its neighbors I_ϕ, one neighbor at a time. Also, each player can choose among two available strategies $S_i = \{m, M\}$ with a probability of 0.5, where m stands for minority and M for majority. A mixed strategy x_i (i.e., combination of strategies from which one is randomly chosen with specified probability) for player i is the probability distribution over his set of strategies S_i. Then, for each player i, its k, where $k = 5$, nearest neighbors are computed and for each player interacting with each of its k neighbors, the utility functions are computed as follows:

$$u_i(x) = \sum_{j \in I_\phi \cap I_u} (x_i^T A_{ij} x_j) + \sum_{d=1}^{2} \sum_{j \in I_\phi \cap I_{c|d}} (x_i^T A_{ij} e_j^d) ,$$

where $d = 1$ is playing the minority class and $d = 2$ is the majority class, $e_j^d \in S_i$ is an extreme mixed strategy with $e_j^1 = (1,0)$ and $e_j^2 = (0,1)$, and A_{ij} is the partial payoff matrix between two players i and j. The set $I_{c|d}$ is the set of players who always play their d^{th} strategy. After that, the average payoff in the whole population is computed:

$$u(x) = x_i^T A_{ij} x_j .$$

Then, iteratively, discrete-time replicator dynamic is applied to study the evolution of the minority strategy probability:

$$x_i^m(t+1) = \frac{\alpha + u_i(e_i^m)}{\alpha + u_i\big(x(t)\big)} x_i^m,$$

if a maximum number of iterations is reached, the process stops, otherwise, t is increased by one and the average payoff for the next player is computed. Finally, for each player in I_u, the class membership with the highest probability is assigned.

An example of the number of the synthetic data of the minority class generated by NEATER and their characteristics for the "Male White Race" MESA subgroup can be seen in **TABLE S6**.

DATA S2: TWO-FOLD CROSS VALIDATION

To ensure and increase the model's robustness and ability to generalize under unknown samples, we employed two-fold cross validation to randomly split the original dataset into two equally sized halves, a training set to train the model, and a test set to evaluate it. This type of cross validation has been widely used in the machine learning literature for predicting high-risk individuals.[5-8] To ensure that the random split of the dataset will always result in having positive and negative examples in both training and testing sets, we employed the following procedure. First, we randomly shuffle the sub-cohort of samples with CVD events into two parts (50% of the positive samples for training and the remaining 50% for testing). Then, the remaining sub-cohort of negative samples is also randomly split into two halves, and the corresponding training and testing subsets are fused so that each of them will contain positive and negative examples. The training and testing sets are independent and do not overlap with each other. At this point, we train our model on subset A and evaluate on subset B, and next we reverse the order (i.e., train on subset B and evaluate on subset A). This process is repeated 10 times, so that statistical reliability of the evaluation process may be ensured[9-11], with each of the different subsets used exactly once as the validation data, and the results are averaged over all the examined configurations. Note that at each iteration the training and evaluation processes start from scratch so that there is no memory of any the previous learned model, and thus biased results are avoided. One of the main reasons for using two-fold cross validation is that the MESA data are extremely imbalanced and there is not enough data of the positive class; furthermore, by repeating the random split multiple times, we are able to train on more positive examples. A fair way to evaluate the model is to split the dataset into two halves and train on as many positive examples as possible, since it is a powerful general technique, when the data are sparse.

DATA S3: SUPPORT VECTOR MACHINE

Support Vector Machine (SVM)[1] is a discriminative classifier, which is designed for supervised learning. The learning model is given a training set of examples (or inputs), belonging to two classes, with associated class labels (or output values). The examples are in form of attribute vectors and the SVM finds the optimal maximum-margin hyper-plane, which separates the input data. Although there exist multiple hyper-planes that offer a solution to the problem, a hyper-plane may be a bad solution if it lies too close to the points, as it is noise-sensitive and may not generalize well. Thus, SVMs aim at finding the hyper-plane that gives the largest minimum distance to the training examples. In other words, given a set of N training examples that consists of pairs of feature vectors x_i with $i = 1, \dots, N$, that denote the pattern to be classified, along with their corresponding class labels y_i, where $x \in \mathbb{R}^d$, with d being the number of features for each sample (i.e., age, sex, ethnicity, total cholesterol, HDL cholesterol, systolic blood pressure, hypertension, diabetes, and smoking status) and $y \in \{-1, +1\}$, where label "-1" corresponds to subjects without an event and label "+1" corresponds to subjects with an event.

The problem is defined as constructing the decision function that correctly classifies an input pattern that is not the training set. The SVM determines the decision hyper-plane between the two classes, the positive class y_1 (i.e., subjects with an event) and the negative class y_2 (i.e., subjects without an event), which is obtained by the solution of the following optimization problem:

$$\underset{w,b,\xi}{\text{minimize}} \left\{ \frac{1}{2} \left|\left|w\right|\right|^2 + C \sum_{i=1}^{N} \xi_i \right\},$$

$$\text{subject to: } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad i = 1, \dots, N$$

where w is a is a normal vector perpendicular to the hyper-plane, $\left|\left|w\right|\right|^2$ indicates the size of the

margin, C is a positive constant that reflects the influence of margin errors, b determines the offset of the hyper-plane from the origin along the normal vector w, and ξ_i are the slack variables, which measure the degree of misclassification of the datum x_i. In our implementation, the kernel "trick" is used with a function $\phi(x_i)$ that maps the data into a higher dimensional space, where various separating planes would be evaluated and ultimately a hyper-plane can be found.

The minimization process is a problem of Lagrangian optimization that can be solved by transforming to the dual form and using Lagrange multipliers to obtain the weight vector w and the bias b of the optimal hyper-plane as follows:

$$\underset{a}{\text{minimize}}\ R(a) = \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} a_i\,,$$

$$\text{subject to: } \sum_{i-1}^{N} y_i a_i = 0, \qquad 0 \le a_i \le C$$

For each testing sample, the kernel matrix K between each of the training samples and the respective testing sample is computed. Thus, the decision function $f(x)$ is given by:

$$f(x) = \ \text{sgn}\left(\sum_{i=1}^{N} y_i a_i K(x_i, x) + b\right),$$

where the terms a_i, with $i = 1, \ldots, N$ constitute a dual representation for the weight vector w in terms of the training set, such as:

$$w = \sum_{i=1}^{N} a_i y_i x_i\,.$$

Moreover, in our experiments, we used as kernel function the radial basis function (RFB) kernel, which is defined as:

$$K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right), \qquad \gamma > 0$$

To estimate the value of the training parameters, for each of the 16 ML-based models (i.e., eight ML-based models for "Hard CVD" events and eight models for "All CVD" events), we used two-fold cross-validation by setting the values of parameter C to 2^k, with $k \in \{-5, \ldots, 15\}$ and the values of the kernel coefficient γ were set to 2^k, with $k \in \{-15, \ldots, 3\}$.

For visualization purposes, we projected the high-dimensional feature space into a 3D feature space using the Principal Component Analysis (PCA). Because of the high dimensionality of the input training data, the decision hyper-plane between the class samples with an event and the class samples without an event is transformed into a hyper-surface. An example of the 3D hyper-surface for the MESA male group for classifying "Hard CVD" and "All CVD" events can be seen in the **FIGURE S1**.

TABLE S1. MESA Cohort Baseline Characteristics of Study Population and Subgroups of Interest. Continuous variables are expressed as mean ± standard deviation. Categorical variables are presented as absolute numbers and frequencies. *The ACC/AHA Risk Calculator does not use these variables; therefore, they were not included in the Machine Learning CVD predictive models. †High sensitivity C-reactive protein (hsCRP) is also expressed as a geometric mean with 90% confidence interval since this variable is not normally distributed.

	Non-Statin Users							**Statin Users**
	All (N = 5,415)	**Hard CVD (N = 381)**	**All CVD (N = 775)**	**ACC/AHA < 9.75% 13yr risk (N = 3,092)**	**ACC/AHA ≥ 9.75% 13yr risk (N = 2,323)**	**ML: Low Risk (13yr) (N = 4,844)**	**ML: High Risk (13yr) (N = 571)**	**Lipid Lowering Medication (N = 1,044)**
Age, y	60.6 ± 9.7	65.5 ± 9.2	65.5 ± 9.0	54.9 ± 6.9	68.2 ± 7.2	59.9 ± 9.6	66.0 ± 8.6	65.0 ± 8.3
Male, n%	2,563 (47.3%)	222 (58.3%)	477 (61.6%)	1,119 (36.2%)	1,445 (62.2%)	2,204 (45.5%)	359 (62.9%)	497 (47.6%)
Female, n%	2,852 (52.7%)	159 (41.7%)	298 (38.5%)	1,973 (63.8%)	878 (37.8%)	2,640 (54.5%)	212 (37.1%)	547 (52.4%)
Ethnicity, n%								
White	2,028 (37.5%)	145 (38.0%)	322 (41.5%)	1,224 (39.6%)	804 (34.6%)	1,806 (37.3%)	222 (38.9%)	456 (43.7%)
Asian	663 (12.2%)	27 (7.1%)	52 (6.7%)	405 (13.1%)	258 (11.1%)	602 (12.4%)	61 (10.7%)	104 (10.0%)
African American	1,484 (27.4%)	107 (28.1%)	223 (28.8%)	717 (23.2%)	767 (33.0%)	1,334 (27.5%)	150 (26.2%)	296 (28.3%)
Hispanic	1,240 (22.9%)	102 (26.8%)	178 (23.0%)	746 (24.1%)	494 (21.3%)	1,102 (22.8%)	138 (24.2%)	188 (18.0%)
Total Cholesterol, mg/dL	196.6 ± 35.5	197.6 ± 33.8	195.9 ± 36.1	196.2 ± 34.7	197.1 ± 36.5	196.7 ± 35.9	195.8 ± 31.3	182.9 ± 35.3

HDL Cholesterol, mg/dL	51.0 ± 14.9	47.7 ± 14.0	47.8 ± 13.7	52.7 ± 15.0	48.7 ± 14.5	51.5 ± 15.1	46.8 ± 12.9	50.3 ± 13.9
Systolic Blood Pressure, mm Hg	125.3 ± 21.0	135.7 ± 22.2	134.5 ± 21.9	116.7 ± 16.5	136.8 ± 21.0	124.2 ± 20.8	134.9 ± 20.3	129.2 ± 21.5
Hypertension, n%	1,724 (31.8%)	173 (45.4%)	364 (47.0%)	578 (18.7%)	1,146 (49.3%)	1,468 (30.3%)	256 (44.8%)	627 (60.1%)
Diabetes, n%	505 (9.3%)	69 (8.1%)	147 (19.0%)	99 (3.2%)	406 (17.5%)	451 (9.3%)	54 (9.5%)	224 (21.5%)
Smoking, n%								
Current Smoking	765 (14.1%)	79 (20.7%)	145 (18.7%)	352 (11.4%)	413 (17.8%)	663 (13.7%)	102 (17.9%)	104 (10.0%)
Prior Smoking	1,938 (35.8%)	134 (35.2%)	314 (40.5%)	1,036 (33.5%)	902 (38.8%)	1,724 (35.6%)	214 (37.4%)	427 (40.9%)
Never	2,712 (50.1%)	168 (44.1%)	316 (40.8%)	1,704 (55.1%)	1,008 (43.4%)	2,457 (50.7%)	255 (44.7%)	513 (49.1%)
*Family History Heart Attack, n%	2,082 (38.5%)	184 (48.3%)	370 (47.7%)	1,158 (37.5%)	923 (39.7%)	1,830 (37.8%)	252 (44.1%)	511 (48.9%)
*Coronary Artery Calcification, Agatston	118.1 ± 370.0	284.8 ± 557.3	355.4 ± 713.2	36.3 ± 155.7	227 ±515.9	103.6 ± 344.6	242.0 ± 524.7	246.0 ±556.3
*†hsCRP, mg/L	3.9 ± 6.0 1.96 (1.88 - 2.03)	4.4 ± 6.3 2.29 (2.01 - 2.56)	4.6 ± 7.0 2.37 (2.17 - 2 .56)	3.7 ± 5.4 1.82 (1.72 - 1.92)	4.2 ± 6.6 2.16 (2.05 - 2.26)	3.9 ± 5.9 1.98 (1.90 - 2.05)	3.6 ± 6.0 1.79 (1.57 - 2.02)	3.3 ± 5.0 1.76 (1.61 - 1.92)

TABLE S2. Risk Calculator Comparison, when Excluding Statin Users from the Analysis: Sensitivity-Specificity-Other Performance Metrics.

Event	Model	Sn (95% CI)	p-value	Sp (95% CI)	p-value	FN	FP	TP	TN	Acc (95% CI)	p-value	NRI (95% CI)	p-value
Male													
Hard CVD	ACC/AHA Risk Calculator	0.84 ± 0.1 (0.78 - 0.88)	--	0.46 ± 0.1 (0.44 - 0.48)	--	36	1,259	186	1,082	0.50 ± 0.1 (0.48 - 0.51)	--	--	--
	ML Risk Calculator	0.89 ± 0.1 (0.84 0.93)	≤0.001	0.93 ± 0.1 (0.92 - 0.94)	≤0.001	24	161	198	2,180	0.93 ± 0.1 (0.92 - 0.94)	≤0.001	0.52 (0.50 - 0.54)	≤0.001
All CVD	ACC/AHA Risk Calculator	0.77 ± 0.1 (0.72 - 0.80)	--	0.53 ± 0.1 (0.50 - 0.55)	--	112	988	365	1,098	0.57 ± 0.1 (0.55 - 0.59)	--	--	--
	ML Risk Calculator	0.97 ± 0.1 (0.95 - 0.99)	≤0.001	0.83 ± 0.1 (0.81 - 0.84)	≤0.001	13	358	464	1,728	0.86 ± 0.1 (0.84 - 0.87)	≤0.001	0.50 (0.48 - 0.52)	≤0.001
Female													
Hard CVD	ACC/AHA Risk Calculator	0.61 ± 0.1 (0.53 - 0.67)	--	0.71 ± 0.1 (0.69 - 0.73)	--	62	781	97	1,912	0.70 ± 0.1 (0.69 - 0.72)	--	--	--
	ML Risk Calculator	0.79 ± 0.1 (0.72 - 0.85)	≤0.001	0.97 ± 0.1 (0.96 - 0.98)	≤0.001	33	86	126	2,607	0.96 ± 0.1 (0.95 - 0.97)	≤0.001	0.44 (0.42 - 0.46)	≤0.001
All CVD	ACC/AHA Risk Calculator	0.54 ± 0.1 (0.48 - 0.60)	--	0.76 ± 0.1 (0.74 - 0.78)	--	137	608	161	1,946	0.74 ± 0.1 (0.72 - 0.75)	--	--	--
	ML Risk Calculator	0.92 ± 0.1 (0.88 - 0.94)	≤0.001	0.92 ± 0.1 (0.90 - 0.93)	≤0.001	25	217	273	2,337	0.92 ± 0.1 (0.90 - 0.93)	≤0.001	0.54 (0.52 - 0.56)	≤0.001
All													
Hard CVD	ACC/AHA Risk Calculator	0.74 ± 0.1 (0.70 - 0.79)	--	0.60 ± 0.1 (0.58 - 0.61)	--	98	2,040	283	2,994	0.60 ± 0.1 (0.59 - 0.62)	--	--	--
	ML Risk Calculator	0.85 ± 0.1 (0.81 - 0.88)	≤0.001	0.95 ± 0.1 (0.94 - 0.96)	≤0.001	57	247	324	4,787	0.94 ± 0.1 (0.93- 0.95)	≤0.001	0.46 (0.45 - 0.47)	≤0.001
All CVD	ACC/AHA Risk Calculator	0.73 ± 0.1 (0.70 - 0.76)	--	0.62 ± 0.1 (0.61 - 0.64)	--	204	1,752	571	2,888	0.64 ± 0.1 (0.63 - 0.65)	--	--	--

	ML Risk Calculator	0.95 ± 0.1 (0.93 - 0.97)	≤0.001	0.88 ± 0.1 (0.86 - 0.89)	≤0.001	38	575	737	4,065	0.89 ± 0.1 (0.88 - 0.90)	≤0.001	0.48 (0.47 - 0.49)	≤0.001

TABLE S3. FLEMENGHO Cohort Baseline Characteristics of Study Population and Subgroups of Interest Including the Statin Users in the Study Population. Continuous variables are expressed as mean ± standard deviation. Categorical variables are presented as absolute numbers and frequencies.

	All (N = 1,348)	**Hard CVD (N = 265)**	**ACC/AHA < 9.75% 13yr risk (N = 844)**	**ACC/AHA ≥ 9.75% 13yr risk (N = 504)**	**ML: Low Risk (13yr) (N = 1,008)**	**ML: High Risk (13yr) (N = 340)**
Male, n%	672 (49.9%)	155 (58.5%)	324 (38.4%)	348 (69.1%)	446 (44.2%)	226 (66.5%)
Female, n%	676 (50.1%)	110 (41.5%)	520 (61.6%)	156 (30.9%)	562 (55.8%)	114 (33.5%)
Age, y	56.9 ± 9.5	61.3 ± 9.4	52.1 ± 6.5	65.0 ± 8.1	55.3 ± 9.1	61.6 ±8.8
Total Cholesterol, mg/dL	232.4 ± 46.8	238.5 ± 48.4	227.2 ± 41.3	237.6 ± 45.3	230.4 ± 45.9	238.4 ± 49.1
HDL Cholesterol, mg/dL	54.5 ± 16.9	51.9 ± 17.2	58.0 ± 15.6	48.2 ± 15.0	56.4 ± 17.1	48.9 ± 15.1
Systolic Blood Pressure, mm Hg	132.2 ± 18.0	137.5 ± 20.2	126.4 ± 14.7	141.8 ± 18.8	129.8 ± 16.4	139.2 ± 20.6
Hypertension, n%	305 (22.6%)	78 (29.4%)	116 (13.7%)	189 (37.5%)	201 (19.9%)	104 (30.6%)
Diabetes, n%	56 (4.2%)	16 (6.0%)	16 (1.9%)	40 (7.9%)	38 (3.8%)	18 (5.3%)
Smoking, n%						
Current Smoking	357 (26.5%)	90 (34.0%)	168 (19.9%)	189 (37.5%)	268 (26.6%)	89 (26.2%)
Prior Smoking	440 (32.6%)	80 (31.2%)	285 (33.8%)	155 (30.8%)	319 (31.6%)	121 (35.6%)
Never	551 (40.9%)	95 (35.8%)	391 (46.3%)	160 (31.7%)	421 (41.8%)	130 (38.2%)

TABLE S4. Risk Calculator Comparison between Models Trained and Tested on FLEMENGHO Cohort: Sensitivity-Specificity-Other Performance Metrics.

Model	Sn (95% CI)	p-value	Sp (95% CI)	p-value	FN	FP	TP	TN	Acc (95% CI)	p-value	NRI (95% CI)	p-value
Male												
ACC/AHA Risk Calculator	0.74 ± 0.1 (0.66 - 0.80)	--	0.55 ± 0.1 (0.50 - 0.59)	--	41	234	114	283	0.59 ± 0.1 (0.55 - 0.63)	--	--	--
ML Risk Calculator	0.85 ± 0.1 (0.79 - 0.90)	≤0.001	0.99 ± 0.1 (0.98 - 1.00)	≤0.001	23	5	132	512	0.96 ± 0.1 (0.94 - 0.97)	≤0.001	0.55 (0.51 - 0.59)	≤0.001
Female												
ACC/AHA Risk Calculator	0.48 ± 0.1 (0.39 - 0.58)	--	0.82 ± 0.1 (0.78 - 0.85)	--	57	103	53	463	0.76 ± 0.1 (0.73 - 0.79)	--	--	--
ML Risk Calculator	0.71 ± 0.1 (0.61 - 0.79)	≤0.001	0.97 ± 0.1 (0.95 - 0.98)	≤0.001	32	16	78	550	0.93 ± 0.1 (0.91 - 0.95)	≤0.001	0.38 (0.34 - 0.41)	≤0.001
All												
ACC/AHA Risk Calculator	0.63 ± 0.1 (0.57 - 0.69)	--	0.69 ± 0.1 (0.66 - 0.72)	--	98	337	167	746	0.68 ± 0.1 (0.65 - 0.70)	--	--	--
ML Risk Calculator	0.79 ± 0.1 (0.74 - 0.84)	≤0.001	0.98 ± 0.1 (0.97 - 0.99)	≤0.001	55	21	210	1,062	0.94 ± 0.1 (0.93 - 0.95)	≤0.001	0.45 (0.42 - 0.48)	≤0.001

TABLE S5. Risk Calculator Comparison between Models Trained on "White Race" FLEMENGHO Cohort and Tested on "White Race" MESA Cohort.

Model	Sn (95% CI)	p-value	Sp (95% CI)	p-value	FN	FP	TP	TN	Acc (95% CI)	p-value	NRI (95% CI)	p-value
Male												
ACC/AHA Risk Calculator	0.85 ± 0.1 (0.77 - 0.91)	--	0.45 ± 0.1 (0.42 - 0.48)	--	16	602	91	488	0.48 ± 0.1 (0.46 - 0.51)	--	--	--
ML Risk Calculator	0.86 ± 0.1 (0.78 - 0.92)	≤0.001	0.78 ± 0.1 (0.76 - 0.81)	≤0.001	15	236	92	854	0.79 ± 0.1 (0.77 - 0.81)	≤0.001	0.34 (0.31 - 0.37)	≤0.001
Female												
ACC/AHA Risk Calculator	0.58 ± 0.1 (0.49 - 0.68)	--	0.72 ± 0.1 (0.70 - 0.75)	--	34	336	46	871	0.71 ± 0.1 (0.69 - 0.74)	--	--	--
ML Risk Calculator	0.78 ± 0.1 (0.67 - 0.86)	≤0.001	0.79 ± 0.1 (0.77 - 0.81)	≤0.001	18	253	62	954	0.79 ± 0.1 (0.77 - 0.81)	≤0.001	0.27 (0.25 - 0.30)	≤0.001
All												
ACC/AHA Risk Calculator	0.73 ± 0.1 (0.66 - 0.79)	--	0.59 ± 0.1 (0.57 - 0.61)	--	50	938	137	1,359	0.60 ± 0.1 (0.58 - 0.62)	--	--	--
ML Risk Calculator	0.82 ± 0.1 (0.76 - 0.87)	≤0.001	0.79 ± 0.1 (0.77 - 0.80)	≤0.001	33	489	154	1,808	0.79 ± 0.1 (0.77 - 0.81)	≤0.001	0.29 (0.27 - 0.31)	≤0.001

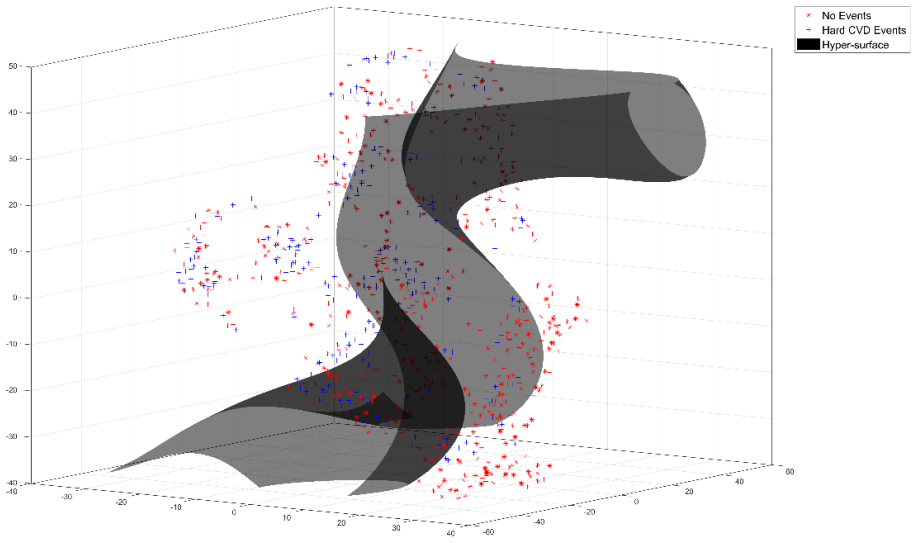
TABLE S6. Characteristics of Synthetic Data Generated by NEATER for "Male White Race" MESA Cohort and Subgroups of Interest. Continuous variables are expressed as mean ± standard deviation. Categorical variables are presented as absolute numbers and frequencies.

	Synthetic Data Generated by NEATER (N = 824)	**Synthetic Data Kept by NEATER (N = 467)**	**p-value***	**Synthetic Data Discarded by NEATER (N = 357)**	**p-value†**	**Majority Data (N = 1,090)**
Age, y	65.5 ± 8.1	66.2 ± 7.7	0.015	64.7 ± 8.5	0.019	61.6 ± 9.6
Total Cholesterol, n%	191.6 ± 30.9	192.6 ± 29.5	0.378	190.5 ± 32.8	0.369	189.4 ± 34.9
HDL, mg/dL	42.1 ± 10.7	41.3 ± 10.9	0.012	43.1 ± 10.2	0.010	45.4 ± 12.0
SBP, mg/dL	132.8 ± 19.2	135.7 ± 19.9	0.002	129.9 ± 17.7	0.001	122.8 ± 17.9
Hypertension, n%	333 (40.4%)	190 (40.7%)	0.486	143 (40.1%)	0.464	347 (31.8%)
Diabetes, n%	159 (19.3%)	86 (18.4%)	0.923	73 (20.4%)	0.855	62 (5.7%)
Smoking, n%			0.834		0.738	
Current Smoking	120 (14.6%)	68 (14.6%)		52 (14.6%)		117 (10.7%)
Prior Smoking	416 (50.5%)	239 (51.1%)		177 (49.6%)		536 (49.2%)
Never Smoking	288 (34.9%)	160 (34.3%)		128 (35.8%)		437 (40.1%)

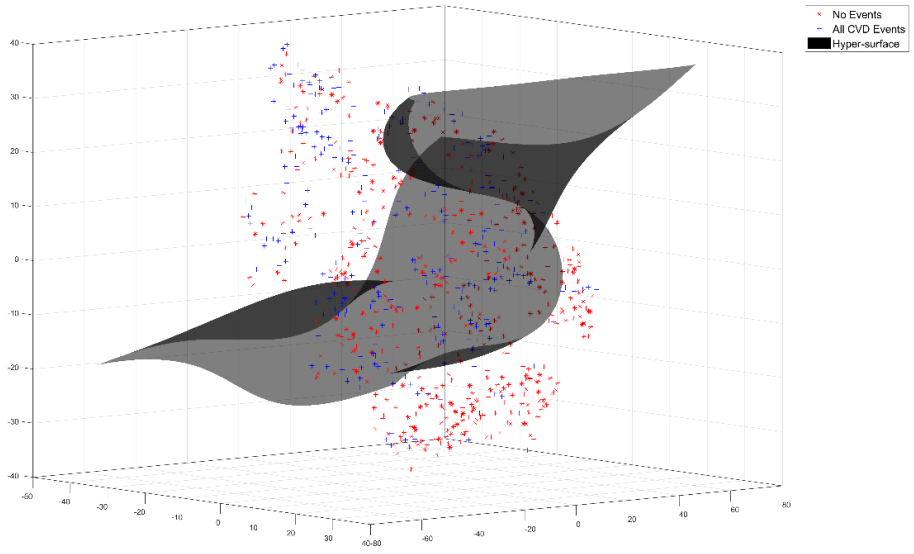
* Interaction between all synthetic data and synthetic data kept by NEATER using multivariate ANOVA

† Interaction between all synthetic data and synthetic data discarded by NEATER using multivariate ANOVA

FIGURE S1. SVM separating hyper-surface for male-group in MESA cohort for classifying (a) Hard CVD events and (b) All CVD events.

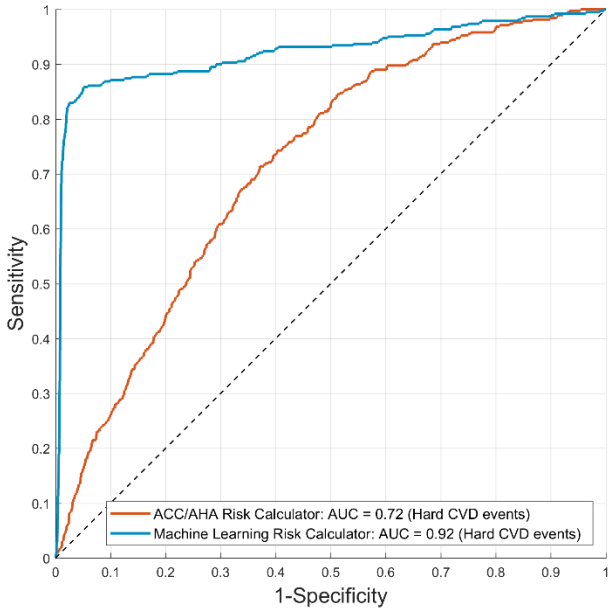


(a)

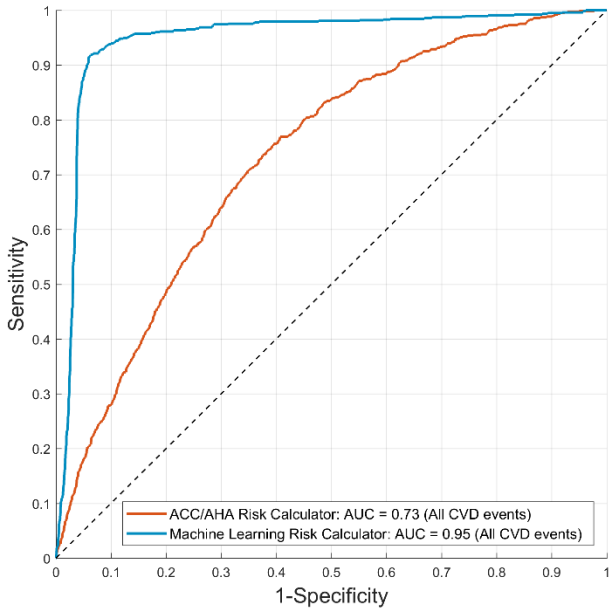


(b)

FIGURE S2. ROC curves for prediction of (a) Hard CVD events and (b) All CVD events, excluding the statin users, comparing the ML Risk Calculator (blue) with the ACC/AHA Risk Calculator (red). AUC: Area under the curve.

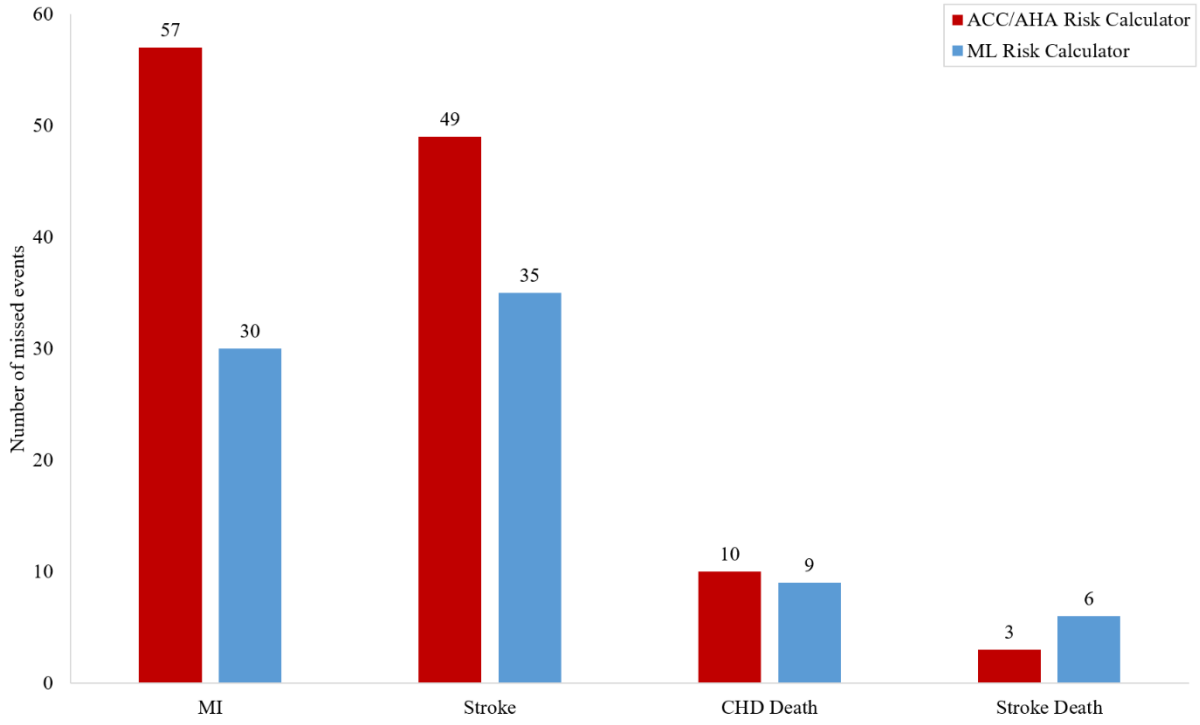


(a)

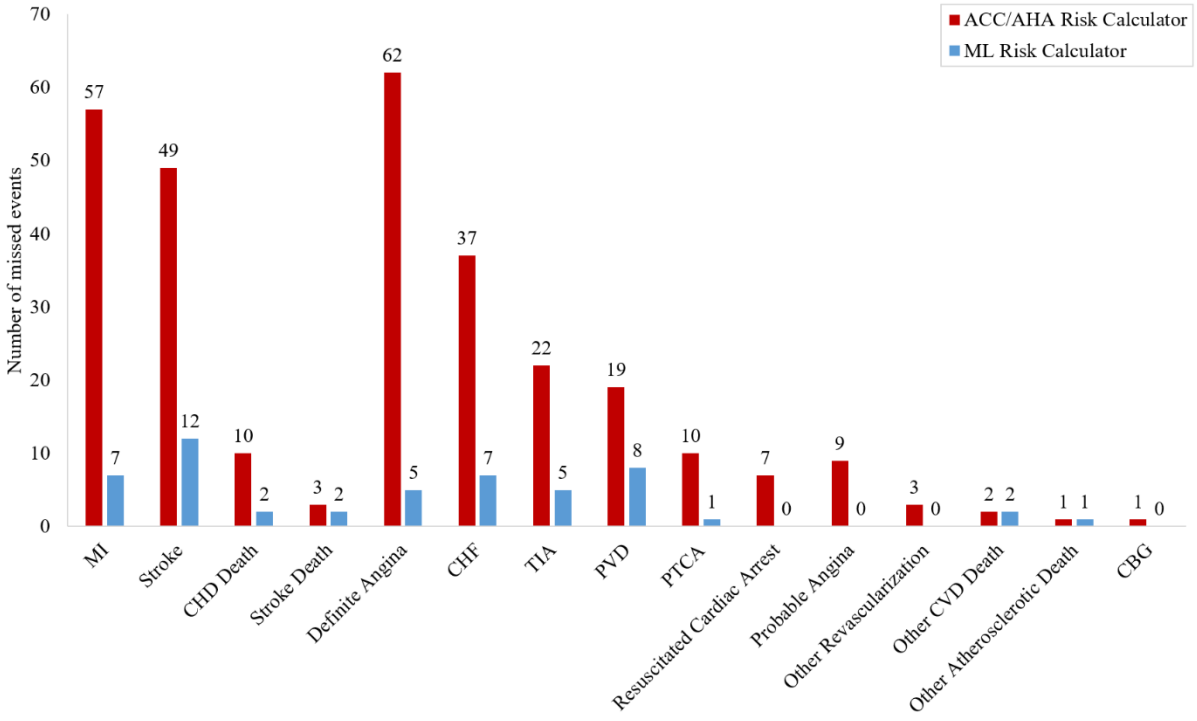

Sensitivity
1-Specificity
ACC/AHA Risk Calculator: AUC = 0.73 (All CVD events)
Machine Learning Risk Calculator: AUC = 0.95 (All CVD events)

(b)

FIGURE S3. Breakdown of the missed (a) Hard CVD events and (b) All CVD events comparing the ML Risk Calculator (blue) with the ACC/AHA Risk Calculator (red). MI: myocardial infarction; CHD: coronary heart disease; CVD: cardiovascular disease; CHF: congestive heart failures; PVD: peripheral vascular diseases; PTCA: percutaneous transluminal coronary angioplasties; CBG: coronary bypass grafts; TIA: transient ischemic attacks.

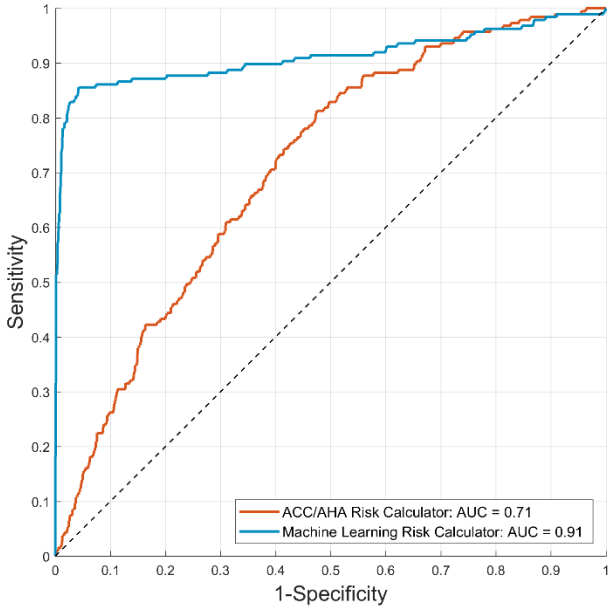


(a)

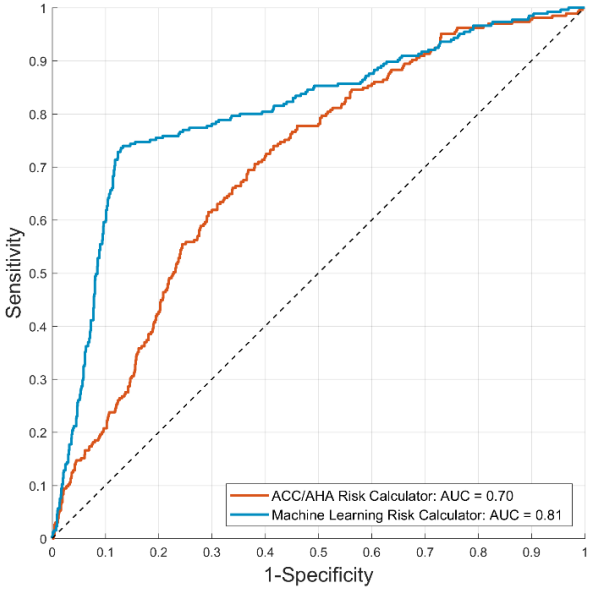

ACC/AHA Risk Calculator
ML Risk Calculator
Number of missed events
70
60
50
40
30
20
10
0
57
7
49
12
10
2
3
2
62
5
37
7
22
5
19
8
10
1
7
0
9
0
3
0
2
2
1
1
1
0
MI
Stroke
CHD Death
Stroke Death
Definite Angina
CHF
TIA
PVD
PTCA
Resuscitated Cardiac Arrest
Probable Angina
Other Revascularization
Other CVD Death
Other Atherosclerotic Death
CBG

(b)

FIGURE S4. ROC curves for prediction of Hard CVD events (a) when training and testing on "White Race" MESA cohort, and (b) when training on "White Race" MESA cohort and testing on FLEMENGHO cohort comparing the ML Risk Calculator (blue) with the ACC/AHA Risk Calculator (red). AUC: Area under the curve.



(a)



(b)

Supplemental References:

1. Chang C-C, Lin C-J. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. 2011;2:1-27.

2. Almogahed BA, Kakadiaris IA. Neater: Filtering of over-sampled data using non-cooperative game theory. *Proc. International Conference of Pattern Recognition*. 2014:1371--1376.

3. Nash J. Non-cooperative games. *Annals of Mathematics*. 1951;54:286-295.

4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* 2002;16:321-357.

5. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12:e0174944.

6. Isler Y, Narin A, Ozer M. Comparison of the effects of cross-validation methods on determining performances of classifiers used in diagnosing congestive heart failure. *Meas Sci Rev*. 2015;15:196-201.

7. Wiens AD, Inan OT. Accelerometer body sensor network improves systolic time interval assessment with wearable ballistocardiography. *Proc IEEE Eng Med Biol Soc*. 2015;2015:1833-1836.

8. Alcaraz R, Martínez A, Rieta JJ. Role of the p-wave high frequency energy and duration as noninvasive cardiovascular predictors of paroxysmal atrial fibrillation. *Comput Methods Programs Biomed*. 2015;119:110-119.

9. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 1998;10:1895-1923.

10. Arlot S, Celisse A. Approximate statistical tests for comparing supervised classification learning algorithms. *Statist Surv*. 2010;4:40-79.

11. Burman P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*. 1989;76:503-514.