Heterogeneity in the Association Between the Presence of Coronary Artery Calcium and Cardiovascular Events: A Machine Learning Approach in the MESA Study

Expanded Method

The causal forest algorithm to assess the heterogeneity in the association between the presence of coronary artery calcium and cardiovascular events.

Figures S1-S4

1. Application of causal forest model in this study
2. A hypothetical example of a causal tree
3. Honest estimation approach
4. Cross-fitting approach

Figures S5-S9

5. Covariates balance between individuals with CAC=0 and those with CAC>0 before and after the propensity score matching
6. Calibration plot of the causal forest model.
7. Variable importance of the causal forest model.
8. Association between the 10-year ASCVD risk and the estimated increase in the risk of cardiovascular events when CAC>0 using the entire sample before propensity score matching (n=5594).
9. Association between the 10-year ASCVD risk and the estimated increase in the risk of cardiovascular events when CAC>0 among people without statin use.

Tables S1-S4

1. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME) checklist
2. Baseline characteristics of people according to the estimated increase in the risk of cardiovascular events when CAC>0 (vs. CAC=0) among adults with borderline or higher 10-year ASCVD risk (≥5%).
3. Baseline characteristics of people according to the estimated increase in the risk of cardiovascular events when CAC>0 (vs. CAC=0) among adults with low 10-year ASCVD risk (<5%), using the median (3.2% increase) instead of 2.5% as the threshold.
4. Baseline characteristics of people according to the estimated increase in the risk of cardiovascular events when CAC>0 (vs. CAC=0) among adults with borderline or higher 10-year ASCVD risk (≥5%), using the median (5.5% increase) instead of 2.5% as the threshold.

Expanded Method. The causal forest algorithm to assess the heterogeneity in the association between the presence of coronary artery calcium and cardiovascular events.

The causal forest is one of the machine learning-based approaches to estimate how the exposure-outcome associations (or the treatment effects in clinical trials) vary across individuals (Figure S1). It uses an ensemble of "trees" or partitions optimized to detect heterogeneity, and thus estimate the risk due to exposure for a particular individual as a function of observable characteristics. Estimating individual risk of diseases associated with the exposure (i.e., CVD risk when CAC>0 in our study) allows us to identify people who would receive benefit or harm from the exposure of interest (i.e., information gained with CAC measurement in our study).

Figure S1. Application of causal forest model in this study
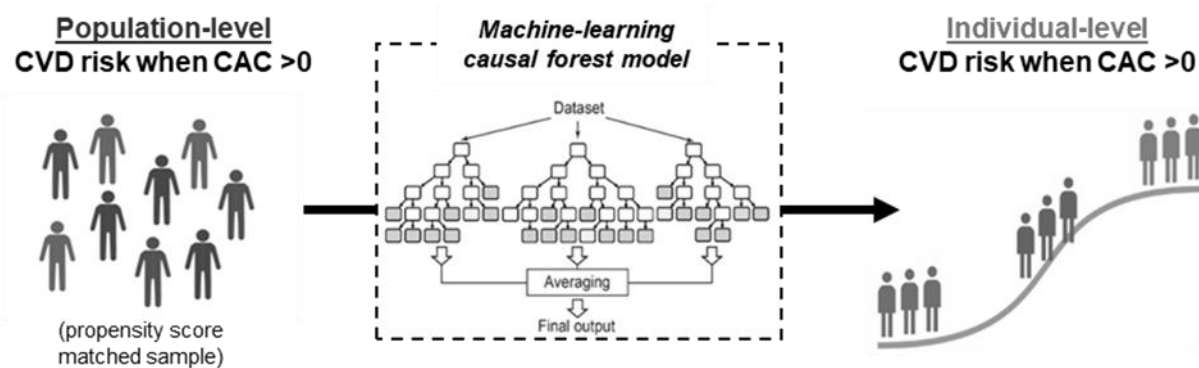
In the present study, we constructed an ensemble of 2,000 causal trees to identify subgroups with different magnitudes of the associations between positive CAC and incident CVD by baseline characteristics (for example, in the right hypothetical example, the estimated increase in CVD risk when CAC>0 was +6.0 pp for female with BMI ≥30, and +2.0 pp for male aged <60 years; Figure S2).

Figure S2. A hypothetical example of a causal tree
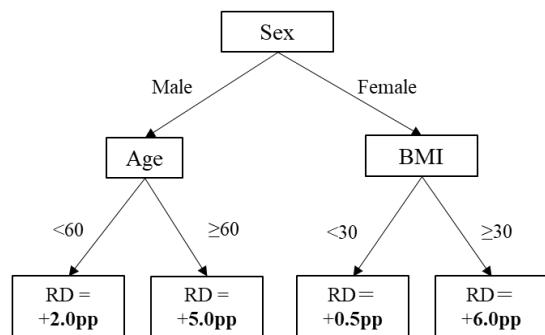
RD, risk difference (increase in CVD risk when CAC>0); pp, percentage point.
This example shows only three variables in the tree, but actual trees are more complex with several variables and splits.
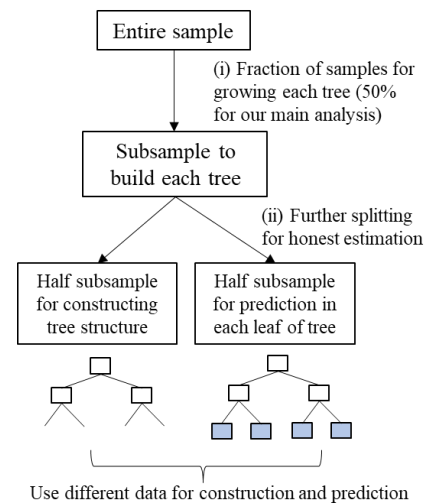
More formally, within the counterfactual framework, the causal forest model allows us to estimate

$$E[Y_{x=1} - Y_{x=0}|C = c]$$

where Y_x denotes potential outcome Y (CVD event) under exposure X (CAC>0 or not) and C denotes a set of baseline characteristics. Causal forest is different from other common machine learning algorithms such as random forest because it assesses the contrast in the average outcome between the exposed and the unexposed individuals ($E[Y_{x=1} - Y_{x=0}|C = c]$) rather than predicting the average outcome itself ($E[Y|C = c]$).


Figure S3. Honest estimation approach

To minimize the risk of overfitting, the following two steps of the "honest" estimation approach were applied to build each tree using observable individual characteristics (Figure S3): i) randomly select the half subsample (or a specific fraction of samples) without replacement from the entire dataset to build each tree algorithm, and ii) further split the fractional subsample into halves and used the first half to construct the tree structure and the second half to make predictions in each leaf of the tree. We also used cross-fitting with 10 folds in which we calculated estimates for each fold (k) based on the causal forest model that was fit with other folds (i.e., other than fold k; Figure S4). This is different from

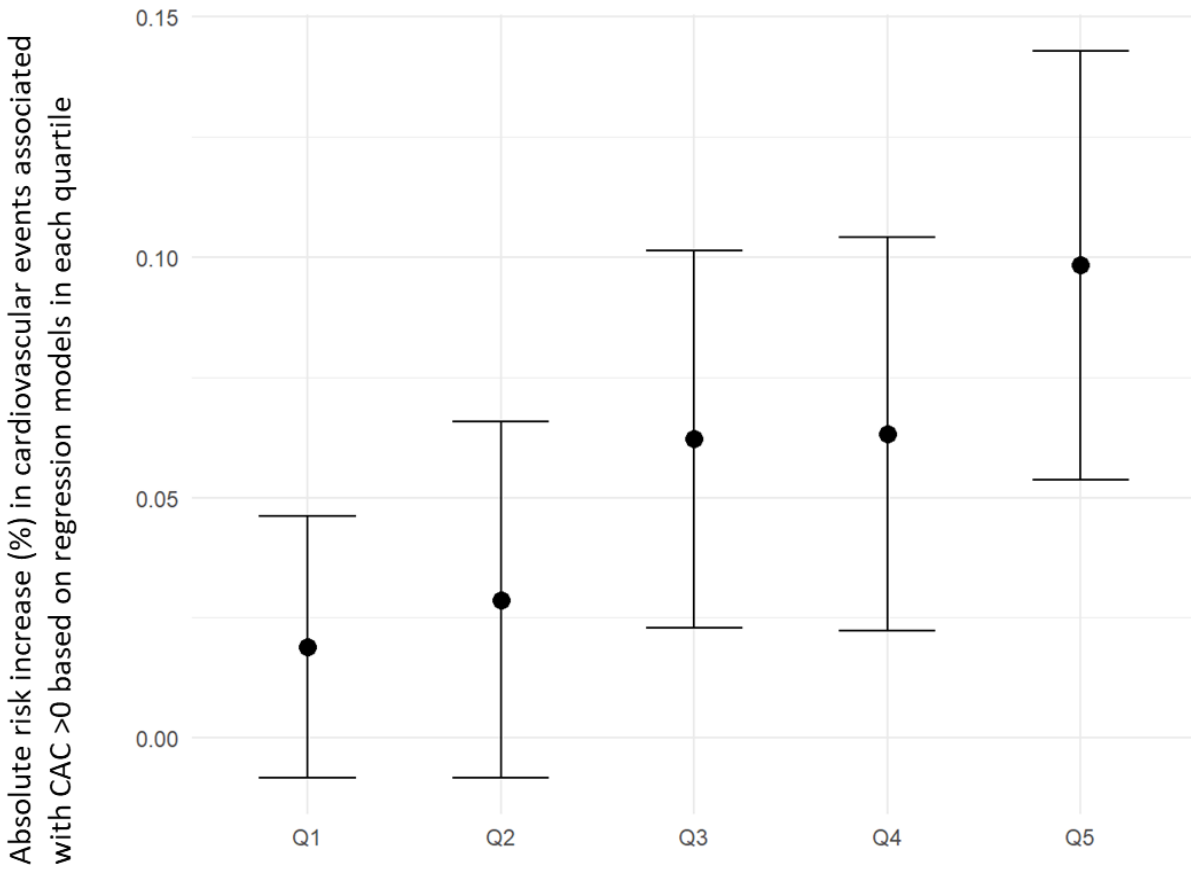**Figure S4**. Cross-fitting approach
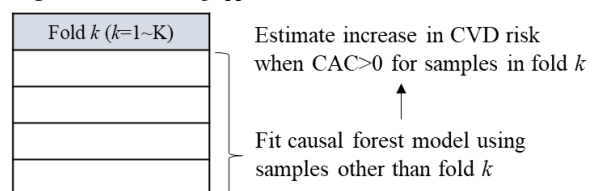
the 10-fold cross-validation approach which was used for tuning the following parameters in the causal forest model; the minimum number of samples a node should contain, the number of variables considered during each split, whether the estimation sample tree should be pruned to avoid empty leaves, maximum imbalance of a split, and penalty for imbalanced splits.

Figure S5. Covariates balance between individuals with CAC=0 and those with CAC>0 before and after the propensity score matching
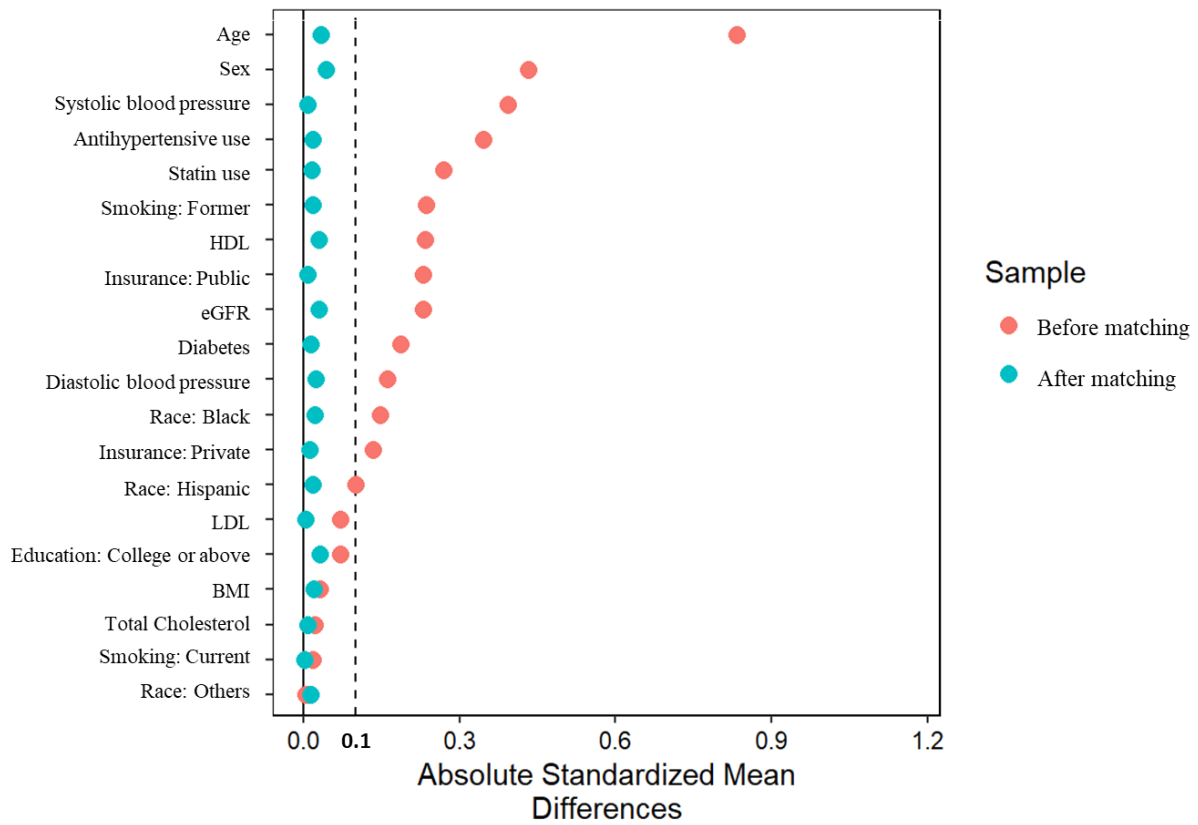
The absolute standardized mean difference of less than 0.1 (dash line) was considered to indicate a negligible imbalance between the 2 groups (CAC=0 vs CAC>0)

Figure S6. Calibration plot of the causal forest model.

Absolute risk increase (%) in cardiovascular events associated with CAC >0 based on regression models in each quartile

Quartiles of absolute risk increase (%) in cardiovascular events associated with CAC >0 based on causal forest model with 10-fold cross-fitting

Figure S7. Variable importance of the causal forest model.

Variable importance
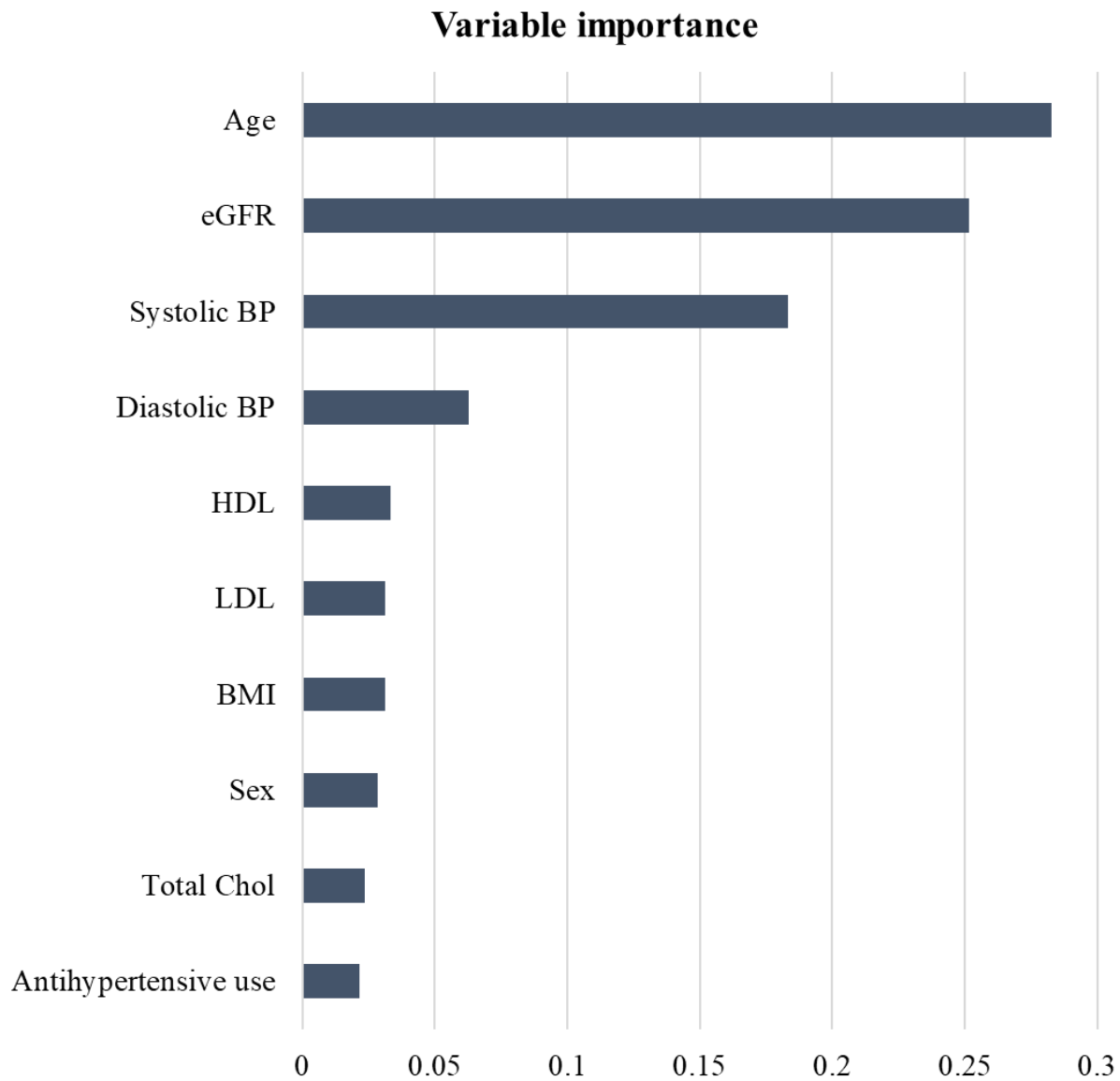
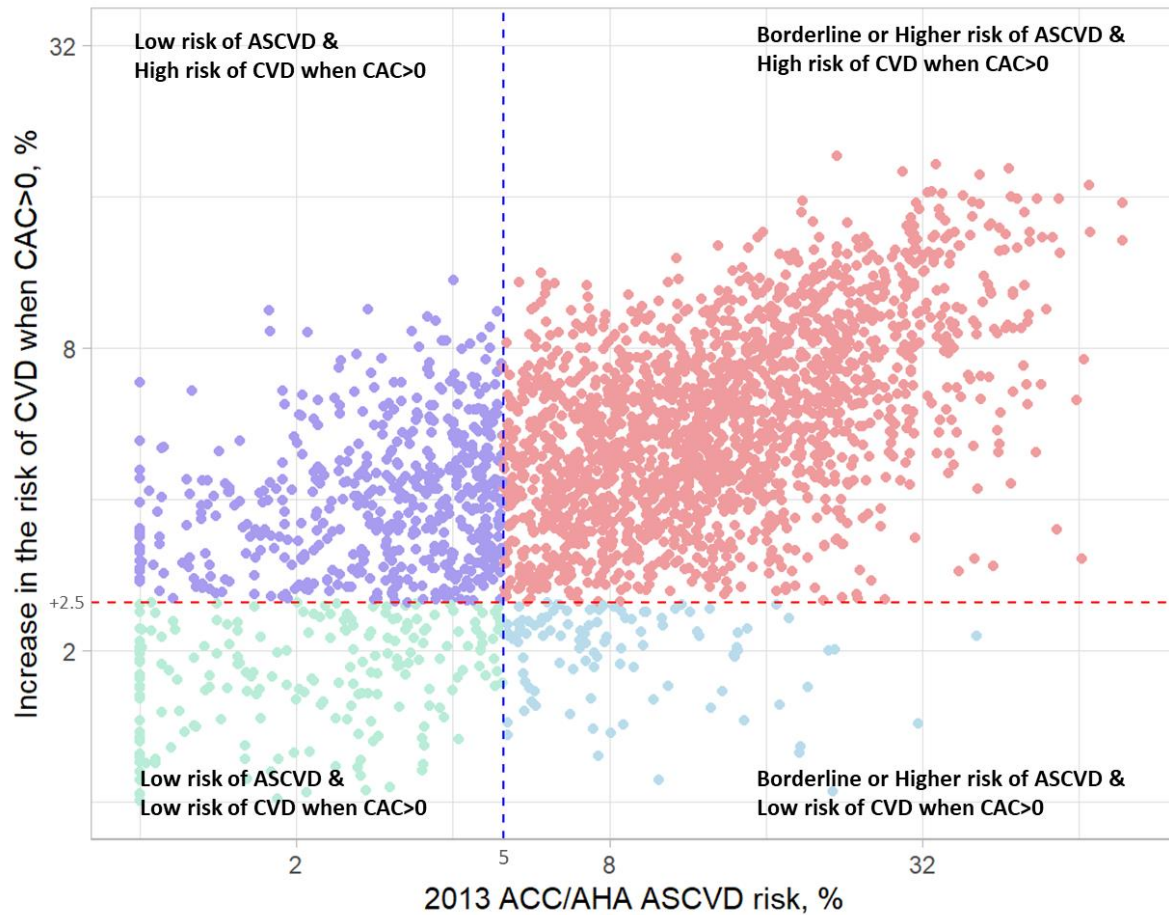The variable importance was calculated by a simple weighted sum of how many times each variable was split at each depth in the causal forest. The top 10 variables are described in this Figure. The results should be carefully interpreted because this rank did not consider the stage of split.

Figure S8. Association between the 10-year ASCVD risk and the estimated increase in the risk of cardiovascular events when CAC>0 using the entire sample before propensity score matching (n=5594).

X-axis shows the 10-year ASCVD risk calculated by the 2013 ACC/AHA pooled cohort equations. Y-axis showed the estimated increase in the risk of cardiovascular events when CAC>0 (calculated by the causal forest model). Spearman's correlation coefficient and Pearson's correlation coefficient between the 10-year ASCVD risk and the estimated increase in the risk of cardiovascular events when CAC>0 were 0.72 (p-value <0.001) and 0.69 (p-value <0.001), respectively.

Figure S9. Association between the 10-year ASCVD risk and the estimated increase in the risk of cardiovascular events when CAC>0 among people without statin use.

X-axis shows the 10-year ASCVD risk calculated by the 2013 ACC/AHA pooled cohort equations. Y-axis showed the estimated increase in the risk of cardiovascular events when CAC>0 (calculated by the causal forest model). Spearman's correlation coefficient and Pearson's correlation coefficient between the 10-year ASCVD risk and the estimated increase in the risk of cardiovascular events when CAC>0 were 0.60 (p-value <0.001) and 0.60 (p-value <0.001), respectively.

Table S1. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME) checklist

1 Designing the Study Plan	
1.1 Describe the need for the application of machine learning to the dataset	Introduction section Paragraph 2
1.2 Describe the objectives of the machine learning analysis	Introduction section Paragraph 3
1.3 Define the study plan	Introduction section Paragraph 3
1.4 Describe the summary statistics of baseline data	Results section Paragraph 1-2
1.5 Describe the overall steps of the machine learning workflow	Results section Paragraph 1-2
2 Data Standardization, Feature Engineering, and Learning	
2.1 Describe how the data were processed in order to make it clean, uniform, and consistent	Statistical analysis (Methods) Paragraph 1
2.2 Describe whether variables were normalized and if so, how this was done	Not applicable for our causal forest algorithm.
2.3 Provide details on the fraction of missing values (if any) and imputation methods	Methods section Paragraph 1
2.4 Describe any feature selection processes applied	Not applicable for our causal forest algorithm.
2.5 Identify and describe the process to handle outliers if any	Not applicable for our causal forest algorithm.
2.6 Describe whether class imbalance existed, and which method was applied to deal with it	Not applicable for our causal forest algorithm.
3 Selection of Machine Learning Models	
3.1 Explicitly define the goal of the analysis e.g., regression, classification, clustering	Identify the conditional average treatment effect
3.2 Identify the proper learning method used (e.g., supervised, reinforcement learning etc.) to address the problem	Statistical analysis (Methods) Paragraph 1
3.3 Provide explicit details on the use of simpler, complex, or ensemble models	Machine Learning Approach (Methods) Paragraph 3
3.4 Provide the comparison of complex models against simpler models if possible	Not applicable for our causal forest algorithm.
3.5 Define ensemble methods, if used	Not applicable for our causal forest algorithm.
3.6 Provide details on whether the model is interpretable	Statistical analysis (Methods) Paragraph 2-3
4 Model Assessment	
4.1 Provide a clear description of data used for training, validation, and testing	Statistical analysis (Methods) Paragraph 1 Cross-fitting approach was employed in this analysis (For each fold k, this procedure fits the causal forest model on observations not included in fold k and predicts the ITEs of the observations in fold k). *American Economic Review*. 2017;107(5):261-65.
4.2 Describe how the model parameters were optimized (e.g., optimization technique, number of model parameters etc.)	Statistical analysis (Methods) Paragraph 1
5 Model Evaluation	
5.1 Provide the metric(s) used to evaluate the performance of the model	Statistical analysis (Methods) Paragraph 1

5.2 Define the prevalence of disease and the choice of the scoring rule used	Results section Paragraph 1-2
5.3 Report any methods used to balance the numbers of subjects in each class	Not applicable for our causal forest algorithm
5.4 Discuss the risk associated to misclassification	Limitation section. Exposure and outcome are less likely to be misclassified in this study.
6 Best Practices for Model Replicability	
6.1 Consider sharing code or scripts on a public repository with appropriate copyright protection steps for further development and non-commercial use	Code available from the corresponding author upon request
6.2 Release a data dictionary with appropriate explanation of the variables	The data and materials that support our findings are available and can be requested at http://www.mesa-nhlbi.org
6.3 Document the version of all software and external libraries used	Statistical analysis (Methods) Paragraph 1
7 Reporting Limitations, Biases and Alternatives	
7.1 Identify and report the relevant model assumptions and findings	Decision tree and random forest (Results)
7.2 If well performing models were tested on a hold-out validation dataset, detail the data of that validation set with the same rigor as that of training dataset (see section 2 above)	Not applicable (cross-fitting approach was used in this paper)

Table S2. Baseline characteristics of people according to the estimated increase in the risk of cardiovascular events when CAC>0 (vs. CAC=0) among adults with borderline or higher 10-year ASCVD risk (≥5%).

Variables[a]	Estimated increase in the risk of cardiovascular events when CAC>0		p-value[b]
	Low (<2.5%)	High (≥2.5%)	
N of participants	135	2293	-
Age	61.5 ± 6.5	65.0 ± 8.1	<0.01
Sex, %			0.48
Male	54.8	51.5	
Female	45.2	48.5	
Race/ethnicity, %			0.50
White	34.8	33.7	
Black	33.3	32.8	
Hispanic	17.1	22.0	
Asian	14.8	11.5	
Education status, %			0.06
Less than college	25.9	39.6	
College or above	74.1	60.4	
Health insurance, %			0.12
Public	23.7	25.7	
Private	63.7	66.9	
Uninsured	12.6	7.4	
Smoking status, %			0.14
Never	44.4	49.2	
Former	30.4	36.5	
Current	25.2	14.3	
BMI, kg/m^2	27.2 ± 6.3	28.6 ± 5.3	<0.01
Systolic blood pressure, mmHg	123.6 ± 19.7	131.5 ± 20.5	<0.01
Diastolic blood pressure, mmHg	71.6 ± 9.1	73.5 ± 10.2	0.05
eGFR, mL/min/1.73 m^2	79.1 ± 14.6	73.7 ± 15.8	<0.01
Total cholesterol, mg/dl	189.5 ± 25.5	195.4 ± 36.0	0.09
HDL cholesterol, mg/dl	54.1 ± 14.9	50.1 ± 14.7	<0.01
LDL cholesterol, mg/dl	111.6 ± 22.4	119.1 ± 32.6	0.02
Diabetes, %	8.2	15.9	0.04
Antihypertensive use, %	28.2	43.8	<0.01
Statin use, %	14.1	15.4	0.69

CVD, cardiovascular diseases; CAC, coronary artery calcium score; BMI, body mass index; eGFR, estimated glomerular filtration rate; HDL, high-density lipoprotein; LDL, low-density lipoprotein.
[a] Mean ± standard deviation was described for continuous variables, otherwise indicated.
[b] *P* values for between-group differences. We used Wilcoxon rank-sum test for continuous variables and Fisher's exact test for categorical variables. P-value was adjusted for multiple comparisons using the Benjamini-Hochberg method.

Table S3. Baseline characteristics of people according to the estimated increase in the risk of cardiovascular events when CAC>0 (vs. CAC=0) among adults with low 10-year ASCVD risk (<5%), using the median (3.2% increase) instead of 2.5% as the threshold.

Variables[a]	Estimated increase in the risk of cardiovascular events when CAC>0		p-value[b]
	Low (<3.2%)	High (≥3.2%)	
N of participants	452	448	-
Age	54.6 ± 5.9	53.7 ± 5.8	0.03
Sex, %			0.32
Male	35.8	39.5	
Female	64.2	60.5	
Race/ethnicity, %			0.02
White	60.6	54.7	
Black	10.2	11.2	
Hispanic	13.3	21.4	
Asian	15.9	12.7	
Education status, %			0.32
Less than college	22.4	25.7	
College or above	77.6	74.3	
Health insurance, %			0.32
Public	5.8	8.5	
Private	84.2	82.1	
Uninsured	10.0	9.4	
Smoking status, %			0.02
Never	49.3	58.3	
Former	42.3	37.1	
Current	8.4	4.7	
BMI, kg/m^2	26.8 ± 5.6	29.4 ± 5.9	<0.01
Systolic blood pressure, mmHg	108.9 ± 12.5	119.4 ± 16.1	<0.01
Diastolic blood pressure, mmHg	66.3 ± 8.1	72.2 ± 10.0	<0.01
eGFR, mL/min/1.73 m^2	75.9 ± 12.3	75.9 ± 23.2	0.99
Total cholesterol, mg/dl	190.5 ± 29.6	202.8 ± 34.8	<0.01
HDL cholesterol, mg/dl	56.8 ± 16.5	50.1 ± 14.6	<0.01
LDL cholesterol, mg/dl	111.6 ± 25.3	125.5 ± 32.1	<0.01
Diabetes, %	1.8	3.1	0.28
Antihypertensive use, %	13.1	22.5	<0.01
Statin use, %	13.1	13.6	0.85

CVD, cardiovascular diseases; CAC, coronary artery calcium score; BMI, body mass index; eGFR, estimated glomerular filtration rate; HDL, high-density lipoprotein; LDL, low-density lipoprotein.
[a] Mean ± standard deviation was described for continuous variables, otherwise indicated.
[b] *P* values for between-group differences. We used Wilcoxon rank-sum test for continuous variables and Fisher's exact test for categorical variables. P-value was adjusted for multiple comparisons using the Benjamini-Hochberg method.

Table S4. Baseline characteristics of people according to the estimated increase in the risk of cardiovascular events when CAC>0 (vs. CAC=0) among adults with borderline or higher 10-year ASCVD risk (≥5%), using the median (5.5% increase) instead of 2.5% as the threshold.

Variables[a]	Estimated increase in the risk of cardiovascular events when CAC>0		p-value[b]
	Low (<5.5%)	High (≥5.5%)	
N of participants	1236	1192	-
Age	62.2 ± 6.8	67.6 ± 8.3	<0.01
Sex, %			0.02
Male	54.3	49.0	
Female	45.7	51.0	
Race/ethnicity, %			0.02
White	31.1	36.5	
Black	33.5	32.1	
Hispanic	23.9	20.0	
Asian	11.6	11.8	
Education status, %			<0.01
Less than college	35.8	41.9	
College or above	64.2	58.1	
Health insurance, %			<0.01
Public	22.8	28.5	
Private	67.6	65.9	
Uninsured	9.6	5.6	
Smoking status, %			<0.01
Never	45.8	52.2	
Former	35.7	36.6	
Current	18.5	11.2	
BMI, kg/m^2	28.8 ± 5.5	28.3 ± 5.2	0.02
Systolic blood pressure, mmHg	125.6 ± 19.1	136.8 ± 20.4	<0.01
Diastolic blood pressure, mmHg	72.0 ± 9.5	74.8 ± 10.5	<0.01
eGFR, mL/min/1.73 m^2	78.0 ± 14.2	70.0 ± 16.3	<0.01
Total cholesterol, mg/dl	192.8 ± 34.5	197.4 ± 36.3	<0.01
HDL cholesterol, mg/dl	50.1 ± 14.3	50.6 ± 15.1	0.37
LDL cholesterol, mg/dl	117.2 ± 31.5	120.2 ± 32.7	0.02
Diabetes, %	15.1	15.9	0.62
Antihypertensive use, %	38.1	47.9	<0.01
Statin use, %	16.2	14.4	0.24

CVD, cardiovascular diseases; CAC, coronary artery calcium score; BMI, body mass index; eGFR, estimated glomerular filtration rate; HDL, high-density lipoprotein; LDL, low-density lipoprotein.
[a] Mean ± standard deviation was described for continuous variables, otherwise indicated.
[b] *P* values for between-group differences. We used Wilcoxon rank-sum test for continuous variables and Fisher's exact test for categorical variables. P-value was adjusted for multiple comparisons using the Benjamini-Hochberg method.