


ORIGINAL RESEARCH

# Unsupervised Learning for Automated Detection of Coronary Artery Disease Subgroups

Alyssa M. Flores , MD; Alejandro Schuler , PhD; Anne Verena Eberhard ; Jeffrey W. Olin , DO; John P. Cooke , MD, PhD; Nicholas J. Leeper , MD; Nigam H. Shah , PhD, MBBS\*; Elsie G. Ross , MD, MSc\*

**BACKGROUND:** The promise of precision population health includes the ability to use robust patient data to tailor prevention and care to specific groups. Advanced analytics may allow for automated detection of clinically informative subgroups that account for clinical, genetic, and environmental variability. This study sought to evaluate whether unsupervised machine learning approaches could interpret heterogeneous and missing clinical data to discover clinically important coronary artery disease subgroups.

**METHODS AND RESULTS:** The Genetic Determinants of Peripheral Arterial Disease study is a prospective cohort that includes individuals with newly diagnosed and/or symptomatic coronary artery disease. We applied generalized low rank modeling and K-means cluster analysis using 155 phenotypic and genetic variables from 1329 participants. Cox proportional hazard models were used to examine associations between clusters and major adverse cardiovascular and cerebrovascular events and all-cause mortality. We then compared performance of risk stratification based on clusters and the American College of Cardiology/American Heart Association pooled cohort equations. Unsupervised analysis identified 4 phenotypically and prognostically distinct clusters. All-cause mortality was highest in cluster 1 (oldest/most comorbid; 26%), whereas major adverse cardiovascular and cerebrovascular event rates were highest in cluster 2 (youngest/multiethnic; 41%). Cluster 4 (middle-aged/healthiest behaviors) experienced more incident major adverse cardiovascular and cerebrovascular events (30%) than cluster 3 (middle-aged/lowest medication adherence; 23%), despite apparently similar risk factor and lifestyle profiles. In comparison with the pooled cohort equations, cluster membership was more informative for risk assessment of myocardial infarction, stroke, and mortality.

**CONCLUSIONS:** Unsupervised clustering identified 4 unique coronary artery disease subgroups with distinct clinical trajectories. Flexible unsupervised machine learning algorithms offer the ability to meaningfully process heterogeneous patient data and provide sharper insights into disease characterization and risk assessment.

**REGISTRATION:** URL: <https://www.clinicaltrials.gov>; Unique identifier: NCT00380185.

**Key Words:** cluster analysis ■ coronary artery disease ■ machine learning ■ phenotype discovery

**A**therosclerotic cardiovascular disease (ASCVD) represents a complex, heterogeneous disorder for which the application of precision health technologies may be of great utility. Although clear risk

factors have been established for disease development, patients with ASCVD exist on a full phenotypic spectrum with varying comorbidities, clinical features, and rates of disease progression.<sup>1</sup> Smoking status,

Correspondence to: Elsie G. Ross, MD, MSc, Division of Vascular Surgery, Stanford University School of Medicine, 780 Welch Road, CJ350C, Palo Alto, CA, 94304. E-mail: [elsie.ross@stanford.edu](mailto:elsie.ross@stanford.edu)

\*N. H. Shah and E. G. Ross contributed equally.

Supplementary Material for this article is available at <https://www.ahajournals.org/doi/suppl/10.1161/JAHA.121.021976>

For Sources of Funding and Disclosures, see page 13.

© 2021 The Authors. Published on behalf of the American Heart Association, Inc., by Wiley. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

JAHA is available at: [www.ahajournals.org/journal/jaha](http://www.ahajournals.org/journal/jaha)

## CLINICAL PERSPECTIVE

### What Is New?

- We use unsupervised machine learning methods to define distinct clinical profiles in a cohort of patients with coronary artery disease from rich data including 155 clinical, sociodemographic, biological, and genetic features.
- By adding a novel component known as “generalized low rank modeling” to our unsupervised learning approach, we were able to combine multiple types of data that are typically siloed to enable phenotype discovery.

### What Are the Clinical Implications?

- Such novel data-driven approaches can be applied to discover cardiovascular disease subclasses using a wide range of heterogeneous data (as is collected in routine clinical practice).
- The discovered patient clusters may be used to tailor care, reveal population health patterns, and provide more refined risk assessment.

## Nonstandard Abbreviations and Acronyms

<b>GLRM</b>	generalized low rank modeling
<b>MACCE</b>	major adverse cardiovascular and cerebrovascular events
<b>PCE</b>	pooled cohort equation

blood pressure, and other traditional risk factors are used to assess an individual’s likelihood for experiencing adverse outcomes and guide treatment strategies.<sup>2</sup> However, conventional risk factors can have different relative contributions to ASCVD across different vascular beds (eg, coronary, carotid, or peripheral arteries).<sup>3–6</sup> Moreover, data beyond these specific variables can drive disease severity and contribute to cardiovascular risk. Thus, assessing the presence or absence of conventional predictors may oversimplify the characterization of a patient with ASCVD. Methods that capture and integrate richer patient characteristics may be important to better understanding prognosis and targeting intensive risk-reduction therapies.

Unsupervised learning algorithms can identify complex interactions within data and may be used to identify unique patient subgroups within a heterogeneous population. This data-driven approach has been extensively applied to high-dimensional data including imaging<sup>7,8</sup> and genomic sequencing<sup>9,10</sup> and has begun to demonstrate promise in electronic health record–based strategies for risk stratification and resource allocation.<sup>11,12</sup> In an effort

to more adequately capture disease heterogeneity and examine subgroups, clustering algorithms have increasingly been performed on clinical data (sometimes called “phenomapping”).<sup>13–16</sup> However, prior phenomapping algorithms have largely relied on the use of 1 type of data, typically continuous variables, to perform clustering. Health care data, though, are known to be heterogeneous, including continuous, categorical, and ordinal data, and are often missing. In our current work, we sought to apply flexible unsupervised learning algorithms to heterogeneous data collected from individuals with ASCVD. We hypothesized that unsupervised learning using rich clinical, sociodemographic, biological, and genetic data would identify distinct subgroups with unique areas to focus care and distinguish clinically significant differences in risk of cardiovascular events and mortality. We then evaluated whether unsupervised clustering improved cardiovascular risk stratification compared with conventional risk assessment based on the 2013 American College of Cardiology/American Heart Association pooled cohort equations (PCEs).<sup>2</sup>

## METHODS

### Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### Study Cohort

The GenePAD (Genetic Determinants of Peripheral Arterial Disease) study is a prospective, multi-institutional study of patients who presented for non-emergent coronary angiography at Stanford Health Care and Mount Sinai Medical Center (ClinicalTrials.gov identifier: NCT00380185). The study enrolled 1789 patients who were aged  $\geq 40$  years and underwent elective angiography to confirm the presence of coronary artery disease (CAD) between April 2004 and July 2008. Patients who had a history of radiation therapy, organ transplant, or chronic infectious diseases were excluded from the original GenePAD cohort. The study collected extensive data at the time of enrollment, including physical, biological, and health-related factors. Individuals with hemodynamically significant CAD were included in our study, defined as having  $\geq 50\%$  stenosis in at least 1 coronary vessel. The left anterior descending, left circumflex (and ramus), and right coronary arteries were the major coronary arteries analyzed. CAD severity based on coronary catheterization was further defined as having 1-vessel, 2-vessel, and triple-vessel or left main disease.

Data used for modeling included nearly all available variables from the GenePAD study (Table S1). In total,

our analysis included 155 variables ranging from socio-demographics; family, medical, and surgical histories; lifestyle and environment factors; angiographic findings; and blood analyses. Fasting blood was collected for measurements of glucose, lipid levels, and select biomarkers<sup>17</sup> as well as single-nucleotide polymorphisms (SNPs) associated with peripheral artery disease (PAD), ankle-brachial index (ABI), and CAD in genome-wide association studies.<sup>18–23</sup> Lifetime physical activity patterns were assessed using the Physical Activity Questionnaire.<sup>24</sup> Patients also completed the Walking Impairment Questionnaire, which consists of 3 categories evaluating walking distance, speed, and stair climbing.<sup>25</sup> Biomarkers were collected from a subsample of participants, including high-sensitivity CRP (C-reactive protein;  $n=459$ ) and cystatin C and  $\beta$ -2 microglobulin ( $n=268$ ). Diabetes was determined by self-reported use of insulin or oral hypoglycemic agents and/or a fasting blood glucose  $>126$  mg/dL. ABIs were measured as previously described using a 5-MHz Doppler ultrasound (Nicolet Elite 5-MHz vascular model 110R Doppler; Nicolet Vascular, Golden, CO).<sup>26</sup> Data on missingness for each variable are reported in Table S2.

All participants were prospectively followed for incident cardiovascular events, hospitalizations, and all-cause mortality. Follow-up data were collected at  $\approx$ 1-year intervals for up to 5 years. The GenePAD study was funded by the National Heart, Lung, and Blood Institute and approved by the Stanford University and Mount Sinai School of Medicine Human Subjects Institutional Review Boards. All participants provided written informed consent.

## Follow-Up and Outcomes

Our primary outcomes were major adverse cardiovascular and cerebrovascular events (MACCE) and all-cause mortality. In the primary outcome analysis, MACCE was defined as a composite of myocardial infarction (MI), stroke, and coronary and/or peripheral revascularization. To compare clustering to the PCEs, MACCE was redefined as MI, stroke, and death to be more consistent with the PCE models. Cardiovascular events, mortality, and cause of mortality were ascertained through medical record review and by contacting the patient or next of kin directly. All mortalities were verified through linkage with the Social Security Death Index. All-cause mortality data were verified through query of the Social Security Death Index as well as phone or postal communication. Follow-up continued through March 2012.

## Unsupervised Learning Based on Generalized Low Rank Modeling

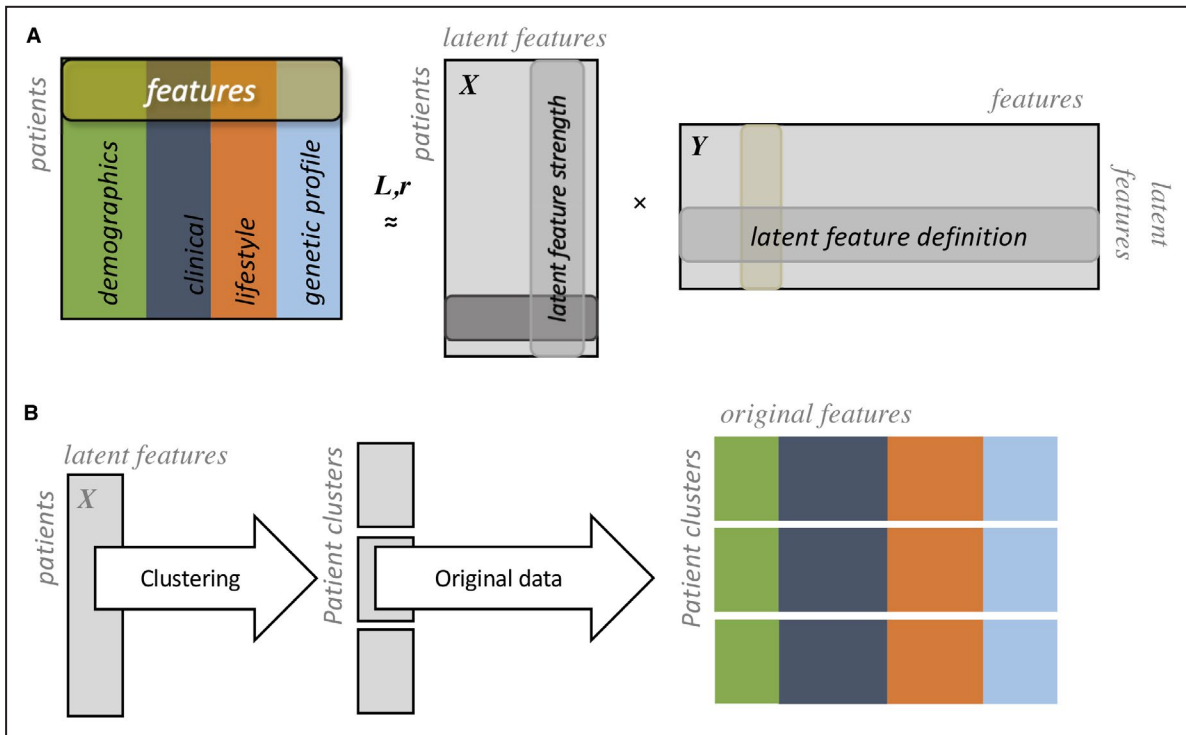
Unsupervised learning algorithms discover underlying patterns in observed data. One approach is to

cluster observations (patients) into self-similar groups. However, most clustering algorithms do not perform well when there are hundreds of variables. Furthermore, most clustering algorithms are built for fully observed numerical data, not the combination of missing values, numbers, categories, and ranks that are common in clinical data sets. Therefore, cluster analysis commonly excludes patients with any missing data variables (resulting in potential bias and deterioration in sample size) and employs data analysis pipelines that separately analyze binary and continuous data.<sup>27</sup> Generalized low rank modeling (GLRM) offers a flexible framework to address these problems.<sup>12,28</sup> GLRM enables simultaneous analysis of high-dimensional data sets with mixed data types and partially missing entries, transforming them into a low-dimensional numerical matrix that is amenable to standard machine learning algorithms. This process is referred to as dimensionality reduction. Dimensionality reduction algorithms posit a small number of unobserved latent features that explain most of the variation in the observed data. The data may then be represented in terms of the latent variables instead (the low-dimensional numeric matrix).

As illustrated in Figure 1A, an original data set is transformed into “latent features” that capture the principal components of variation in the data. This is done by approximating the original data as a product of 2 low-rank matrices  $X$  and  $Y$ . The statistical methodology of GLRM has been described in detail previously.<sup>28</sup> In brief, GLRM is an extension of principal components analysis that enables the ability to add data-type appropriate loss functions (eg, for categorical, ordinal, continuous data) and regularization to approximate a heterogeneous data set and constrain the low-rank representations  $X$  and  $Y$ , respectively. Missing entries are simultaneously imputed in the process of constructing the low-rank matrices.

In our analysis, we applied quadratic loss for continuous features, hinge loss for Boolean features, and ordinal hinge loss for ordinal features.<sup>28</sup> We applied quadratic regularization to the  $X$  and  $Y$  matrices. We tested a range of rank values from 5 to 154 (total features–1) and chose a rank of 50 based on a balance of approximation of the largest drops in training error (Figure S1) while maintaining generalizability (ie, not overfitting our model by choosing rank based on the lowest error). By using a more sparse model with less features, we also optimize the performance of clustering algorithms by reducing noisiness of the data.

The resulting latent features were then used for estimating clusters. Once the clusters were identified, summary statistics and outcomes were evaluated within and across clusters (Figure 1B). We specifically excluded MACCE and mortality event variables from our low-rank modeling preprocessing step. Thus, outcomes were only evaluated after



**Figure 1. Schematic for generalized low rank modeling.**

**A**, Patient data are condensed to fewer dimensions to allow for analysis using unsupervised K-means clustering. The “features” matrix is a high-dimensional data set that includes patient information on demographics and clinical, lifestyle, angiographic, and cardiovascular genetic risk markers. This data set is transformed into a lower dimensional “latent feature” space by approximating the features matrix as the product of 2 matrices, shown as the X (containing each observation) and Y representations (containing the definition for each observation).  $L, r$  indicates the loss function that accounts for the accuracy in the data approximation and regularizes the latent feature representation to prevent overfitting. **B**, After cluster analysis, data are then transformed back to their original form and analyzed to discover subgroup characteristics and compare long-term outcomes across clusters.

cluster assignments. Our GLRM feature matrix was calculated using the GLRM package in Julia (version 0.5).<sup>28</sup>

Based on the low-rank feature matrix, we applied K-means clustering to discover unique groups/clusters. To identify the optimal number of clusters, we first performed validation statistics for stability and internal measures. Stability measures evaluated how much change in a clustering result occurred by removing 1 column of data at a time.<sup>29</sup> The internal validity measures evaluated the degree of connectedness of the clusters (connectivity) and the relative compactness and separation of clusters (silhouette width and Dunn index).<sup>29</sup> The optimal cluster number was chosen based on the majority recommendation in addition to consideration of clinical utility and practicality. The cluster metrics were thus calculated for a range of  $K=2-6$ . K-means clustering was performed using Euclidean distances and the kmeans package in R. Cluster validation metrics were computed using the clValid package. After selecting the number of clusters, we produced cluster plots by applying discriminant

analysis of principal components using the adegenet package in R.<sup>30,31</sup>

## Statistical Analysis

After the identification of distinct clusters, the baseline characteristics were compared across cluster groups. Normality was assessed with the Kolmogorov–Smirnov test. Continuous variables did not follow a normal distribution, and thus the nonparametric Kruskal–Wallis test was used to compare continuous data. Differences in categorical variables were compared using the  $\chi^2$  test or Fisher exact tests, accordingly. Descriptive data are presented as mean and standard deviation for continuous variables or percentages for categorical variables.

In addition, we explored which features were most informative for the cluster identities based on the concept that each K cluster includes individuals based on their distance to the cluster centroid and is defined by a Voronoi cell in the latent space.<sup>32</sup> We therefore obtained a representation of cluster centroids in the original space by multiplying each centroid’s latent

representation by matrix  $Y$ . Results of the features most heavily weighted for each cluster were illustrated as a heatmap using R package `gplots`.

After the exclusion of individuals with missing follow-up dates ( $N=99$  for MACCE,  $N=66$  for all-cause mortality), Kaplan–Meier curves were generated and compared using the log-rank test. To evaluate the relationship between cluster membership and outcomes, hazard ratios (HRs) for MACCE and all-cause mortality were estimated using Cox proportional hazards modeling.

### Comparison of Risk Stratification by Clustering With PCE Modeling

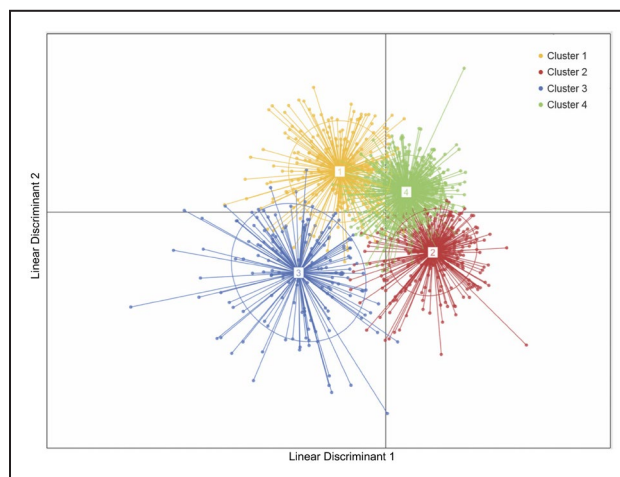
To evaluate how our clustering methodology compared with standard clinical risk stratification, Cox proportional hazards models were also used to estimate the relative hazards of each American College of Cardiology/American Heart Association PCE risk group for MACCE or all-cause mortality.<sup>2</sup> HRs were compared between the PCE model and for a range of cluster numbers ( $N$  clusters recommended from validation statistics $\pm 1$ ). To be consistent with the PCE model, the MACCE composite excluded coronary and lower extremity revascularization events and was redefined to include MI, stroke, and death.

The PCE includes sex, age, race, total cholesterol, high-density lipoprotein cholesterol, diabetes, systolic blood pressure, antihypertensive use, and smoking status. We calculated the 5-year PCE predicted risk of this cohort given that patients were followed for an average of 5 years. To do this, we annualized the 10-year predicted risk and categorized them based on the estimated 5-year risk into the standard PCE groups ( $<2.5\%$  considered low risk,  $2.5\%$ – $4.9\%$  considered intermediate risk, and  $\geq 5\%$  considered high risk), as previously described.<sup>33</sup> Given that our cohort likely had a higher risk of MACCE at baseline, we performed PCE recalibration using the D'Agostino method.<sup>34</sup> In the cluster sensitivity analysis comparing  $N\pm 1$  clusters, we additionally tested a more granular PCE model that categorized the estimated 5-year risks into 4 groups ( $<2.5\%$  considered low risk,  $2.5\%$ – $3.74\%$  as intermediate low risk,  $3.75\%$ – $4.9\%$  as intermediate high risk, and  $\geq 5\%$  considered high risk).<sup>33,35</sup>

All analyses were performed using R version 3.5.2. A  $P$  value  $<0.05$  was considered statistically significant.  $P$  values were adjusted for multiple comparisons using the Benjamini and Hochberg method.

## RESULTS

After excluding individuals without CAD, 1329 participants remained. The overall cohort was 71% men and 54% White and had a mean $\pm$ SD age of  $67\pm 10.5$  years. Aside from the 3 biomarkers that were obtained



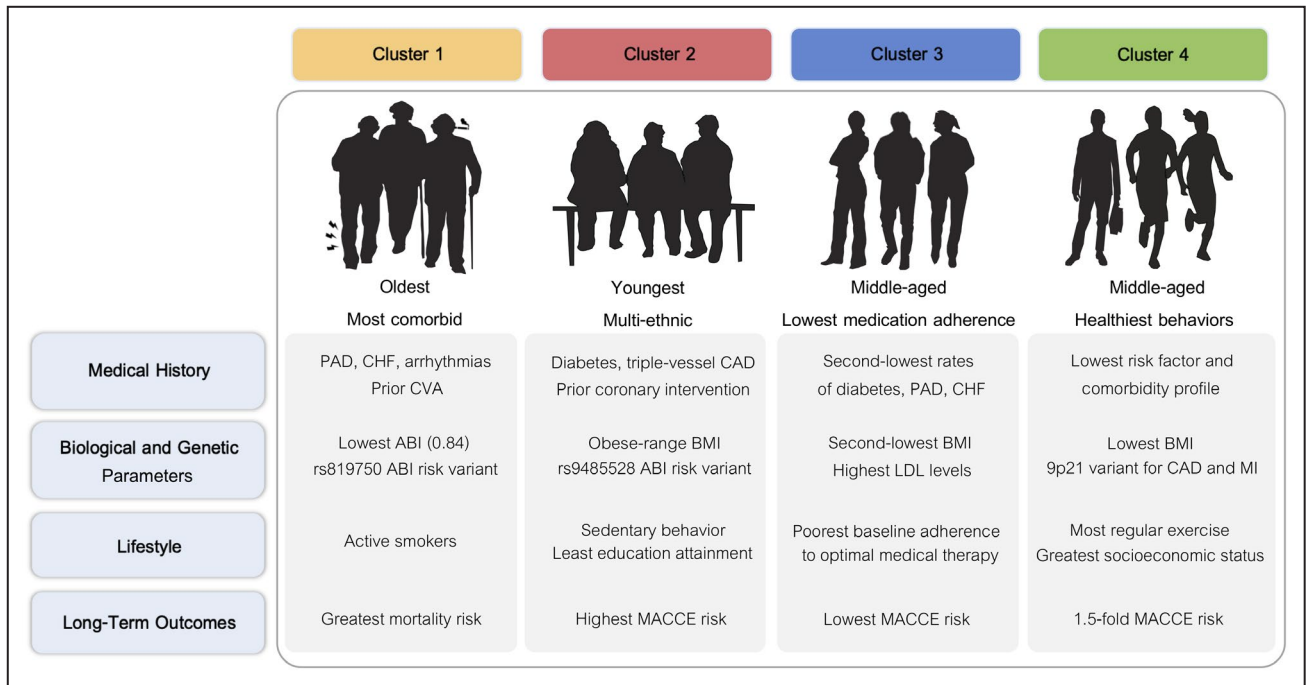
**Figure 2. Distinct subgroups of patients with coronary artery disease identified by unsupervised clustering.**

Plot showing 4 distinct groups of patients identified by K-means clustering. Data are plotted based on the top 20 principal components across the first 2 discriminant functions to form a 2-dimensional plot.

from a subsample of the GenePAD cohort, variables used in the GLRM had a mean of 4.8% missing data (Table S2). Based on this rich data set of demographic, clinical, lifestyle, genetic, and angiographic data, our unsupervised cluster analysis identified 4 distinct subgroups of patients with CAD (Figures 2 and 3). This cluster number was selected based on being most frequently recommended by validation indices (Table S3). The baseline characteristics according to cluster identity are shown in Table 1. Importantly, in addition to cardiovascular risk factors, the identified clusters significantly differed in characteristics that are not always captured by conventional risk assessment.

### Cluster 1: Oldest/Most Comorbid Cluster

Individuals in cluster 1 were the oldest (mean $\pm$ SD,  $70\pm 11$  years) and the most likely to be White (59%) and female (47%). A major feature of this cluster is that patients had the highest rates of several cardiovascular comorbidities, including congestive heart failure (CHF; 15%), cardiac arrhythmias (24%), PAD (26%), and a history of stroke or transient ischemic attack (15%). They also had the lowest mean ABI (0.84) and the second-highest body mass index ( $29.4$  kg/m<sup>2</sup>). In relation to lifestyle behavior, cluster 1 was most likely to be active smokers (13%) and had the greatest pack-year history (26 years) and the second-lowest frequency of weekly exercise (17%). This cluster also reported the greatest parental history of MI or coronary revascularization (25%) and second highest of stroke (14%). Assessment for the presence of select genomic markers revealed that cluster 1 included the highest proportion of carriers of an SNP associated with lower ABI (rs819750,



**Figure 3. Schematic representation of the 4 CAD clusters and their major features.**

ABI indicates ankle-brachial index; BMI, body mass index; CAD, coronary artery disease; CHF, congestive heart failure; CVA, cerebrovascular accident; LDL, low-density lipoprotein; MACCE, major adverse cardiovascular and cerebrovascular events; MI, myocardial infarction; and PAD, peripheral artery disease.

16%) and the second-highest prevalence of an SNP in the chromosome 9p21 region (rs10757269, 51%), a pleiotropic risk variant associated with CAD,<sup>18,22</sup> MI,<sup>19</sup> lower ABI, and PAD.<sup>21,23</sup>

### Cluster 2: Youngest/Multiethnic Cluster

Cluster 2 included the youngest group of patients (64±10 years) composed primarily of racial minorities (71% non-White race). This cluster was the most ethnically diverse (30% Black, 19% Hispanic, 11% South Asian, 5% East Asian), and nearly all (99%) individuals were diabetic. Cluster 2 was the only group with a mean body mass index in the obese range (30.3±6 kg/m<sup>2</sup>) and displayed the most severe pattern of sedentary behavior, as reflected by only 9% who reported engaging in weekly exercise. They also were the least likely to complete education beyond high school. Serological analysis showed that they had the highest levels of serum CRP (8.2 mg/L). In addition, cluster 2 had by far the highest rates of triple-vessel or left main disease (50%), prior MI (29%), prior coronary revascularization (35%), lower extremity amputations (4.9%), and chronic kidney disease (28%). They had the second-highest rates of prior stroke or transient ischemic attack (19%) and reported the greatest parental history of stroke (18%). This group with the second-highest PAD rates (19%) also had the greatest proportion of a genetic variant associated with low ABI (rs9485528, 35%).

However, likely reflecting differences in the genetic architecture of this largely minority ethnic subgroup that were not captured in European genome-wide association studies,<sup>36</sup> cluster 2 showed the lowest frequency of carriers for several of the investigated genetic risk loci for CAD, PAD, and ABI discovered in European populations,<sup>20,23</sup> including rs819750, rs2171209, and the 9p21 risk variant rs10757269.

### Cluster 3: Middle-Aged/Lowest Medication Adherence Cluster

These individuals had lower rates of most comorbidities and cardiovascular risk factors than clusters 1 and 2, such as diabetes (30%), CHF (3.9%), history of PAD (2.8%), prior stroke or transient ischemic attack (2.8%), and hypertension (71%). They had the least severe angiographic CAD, with the highest rates of 1-vessel disease (35%) and the lowest rates of triple-vessel or left main disease (32%). Cluster 3 had the second-highest proportion of men (78%), White patients (56%), and individuals who completed college and graduate-level education (51%). They also had a lower body mass index (28.5 kg/m<sup>2</sup>) and the second-highest rate of weekly exercise (29%), although this group also had the second-highest smoking rate after cluster 1 (12.6%, 18.6 pack years). Despite relatively good health status, overall optimization and adherence to cardiovascular medications was the poorest

**Table 1. Demographic, Socioeconomic, Clinical, and Biological Factors Compared Across Clusters**

	Cluster 1 (oldest/most comorbid), N=271	Cluster 2 (youngest/ multiethnic), N=164	Cluster 3 (middle-aged/ lowest medication adherence), N=316	Cluster 4 (middle- aged/healthiest behaviors), N=578	Adjusted p value
Demographics					
Age, y	70±11	64±10	67±10	66±11	8.7 × 10 <sup>-10</sup>
Male sex	53	57	78	80	8 × 10 <sup>-16</sup>
Race and ethnicity					
White race	59	29	56	59	1.2 × 10 <sup>-10</sup>
Black race	13	30	8	9.5	4.2 × 10 <sup>-11</sup>
South Asian race	4	11	7	7	0.049
East Asian race	4	5	8	8	NS
Hispanic ethnicity	11	19	11	9	0.007
Socioeconomics					
Income					
<\$35 000	17	13	22	16	NS
\$35 000–\$99 000	8	12	18	15	0.005
>\$100 000	8	2	15	21	1.9 × 10 <sup>-10</sup>
Prefer not to answer	66	71	44	47	2.8 × 10 <sup>-12</sup>
Highest education level					
High school or less	45	54	41	31	9.4 × 10 <sup>-8</sup>
College	34	31	29	34	NS
Graduate school	11	7	22	28	8.7 × 10 <sup>-11</sup>
Lifestyle					
Current smoker	13	8.0	12.6	7.3	0.01
Cumulative pack y	26.1±30	16.8±28	18.6±26	15.6±24	7.0 × 10 <sup>-6</sup>
Engages in exercise at least once per wk	17	9.0	29	35	6.4 × 10 <sup>-12</sup>
Physical					
BMI, kg/m <sup>2</sup>	29.4±6	30.3±6	28.5±6	28.4±5	0.0009
Ankle-brachial index	0.84±0.3	0.91±0.3	1.04±0.2	1.05±0.1	8 × 10 <sup>-16</sup>
Systolic blood pressure, mm Hg	141±20	148±21	137±19	137±20	1 × 10 <sup>-5</sup>
Coronary angiography					
One-vessel disease	26	23	35	27	0.006
Two-vessel disease	30	26	29	31	NS
Triple vessel or left main	43	50	32	41	0.003
Clinical history					
Prior MI	0.4	29	2.2	0.7	8 × 10 <sup>-16</sup>
Prior CABG or PCI	1.8	35	1.8	0.2	8 × 10 <sup>-16</sup>
Prior valve surgery	2.9	0	1.3	3.4	0.06
CHF	15	11	3.9	3.3	4.5 × 10 <sup>-10</sup>
Stroke or TIA	15	14	2.8	2.4	3.8 × 10 <sup>-15</sup>
PAD	26	19	2.9	1.6	8 × 10 <sup>-16</sup>
Lower extremity amputation	1.5	5	0	0	2.1 × 10 <sup>-7</sup>
Cardiac arrhythmia	24	15	19	17	NS
Chronic kidney disease	0.4	28	1.9	0.5	3.7 × 10 <sup>-15</sup>
Diabetes	41	99	30	27	8 × 10 <sup>-16</sup>
Biological, mean					
β-2 microglobulin,* µg/mL	3.3±4.5	4.7±7.4	4.2±8.6	3.5±6.9	0.02
Cystatin C,* mg/L	1.0±0.7	1.3±1.1	1.1±1.3	1.0±1.0	0.04

(Continued)

**Table 1. Continued**

	Cluster 1 (oldest/most comorbid), N=271	Cluster 2 (youngest/multiethnic), N=164	Cluster 3 (middle-aged/lowest medication adherence), N=316	Cluster 4 (middle-aged/healthiest behaviors), N=578	Adjusted p value
CRP <sup>†</sup> , mg/L	7.5±18	8.2±25	4.7±19	2.4±5	1.9 × 10 <sup>-7</sup>
Glucose, maximum mg/dL	120±47	188±218	119±65	115±60	8 × 10 <sup>-16</sup>
Total cholesterol, mg/dL	136±38	138±35	146±43	130±33	1 × 10 <sup>-6</sup>
LDL, mg/dL	72±30	73±28	85±50	69±26	9.5 × 10 <sup>-8</sup>
Creatinine, mg/dL	1.3±1.2	2.1±2.0	1.1±0.8	1.1±0.6	1.4 × 10 <sup>-8</sup>
Medications					
Antihypertensive					
Current	90	90	38	96	8 × 10 <sup>-16</sup>
Ever taken	86	93	71	88	1.8 × 10 <sup>-12</sup>
Insulin or hypoglycemic agents					
Current	27	85	11	20	8 × 10 <sup>-16</sup>
Ever taken	32	99	23	20	8 × 10 <sup>-16</sup>
Cholesterol-lowering medications					
Current	78	71	9	99	8 × 10 <sup>-16</sup>
Ever taken	86	86	52	96	8 × 10 <sup>-16</sup>
Aspirin					
Current	72	69	37	84	2.8 × 10 <sup>-7</sup>
Clopidogrel					
Current	48	50	15	47	2.2 × 10 <sup>-12</sup>
Statin					
Current	75	65	7.6	93	8 × 10 <sup>-16</sup>
β-blockers					
Current	59	64	45	68	5.2 × 10 <sup>-7</sup>
Family history (biological mother or father)					
MI, CABG, or PCI (mother)	25	20	17	24	0.01
Stroke (father)	14	18	12	12	0.01
Lower extremity revascularization (father)	1.5	0.6	0.9	0.9	0.004
AAA rupture or repair (father)	0.4	0	0.9	1.6	0.0008
Genetics <sup>‡</sup>					
rs10757269	51	38	40	57	0.01
rs819750	16	7	14	14	0.01
rs94855286	28	35	32	34	0.04
rs2171209	43	35	39	48	NS
rs7100623	24	22	21	29	NS
rs16824978	45	46	37	48	NS
rs7003385	40	34	29	45	NS
rs4659996	46	38	33	45	NS
rs3745274	30	30	38	30	NS
rs290481	28	25	32	28	NS

Values are mean±SD or percentage. AAA indicates abdominal aortic aneurysm; BMI, body mass index; CABG, coronary artery bypass graft; CAD, coronary artery disease; CHF, congestive heart failure; CRP, C-reactive protein; LDL, low-density lipoprotein; MI, myocardial infarction; NS, nonsignificant; PAD, peripheral artery disease; PCI, percutaneous coronary intervention; and TIA, transient ischemic attack.

\*Based on subsample of 268 individuals.

†Based on subsample of 459 individuals.

‡Heterozygous or homozygous carriers.

in cluster 3. For example, individuals in cluster 3 were the least likely to be taking aspirin (37%) or clopidogrel (15%). There were also significant differences

in the relative number of individuals who were prescribed antihypertensives and lipid-lowering medications compared with those who reported maintaining



their current use. This cohort also had the highest average levels of total cholesterol (146 mg/dL) and low-density lipoprotein (85 mg/dL).

### Cluster 4: Middle-Aged/Healthiest Behaviors

Similar to cluster 1, this cluster included a relatively high percentage of White patients (59%). However, they were the most likely to be men (80%) and had the overall best health status of all the clusters. For example, they reported weekly exercise most frequently (35%), had the lowest body mass index (28.4 kg/m<sup>2</sup>), and had the lowest smoking rate of all the clusters (7.3%). Although they had the highest rates of previous valve surgery (3.4%), these individuals had significantly lower rates of most cardiovascular risk factors and other comorbidities, such as prior coronary revascularization (0.2%), CHF (3.3%), PAD (1.6%), stroke or transient ischemic attack history (2.4%), and chronic kidney disease (0.5%). They also most frequently reported a high income (21% reporting >\$100 000 annually) and completed college or graduate-level education (62%). Furthermore, cluster 4 had the lowest levels of total cholesterol (130 mg/dL), low-density lipoprotein (69 mg/dL), and CRP (2.4 mg/L) as well as the highest rates of antiplatelet use (84% taking aspirin). In this cluster, 41% had triple-vessel or left main disease, which was greater than cluster 3. Regarding their familial histories and genetic profiles, cluster 4 included the second-highest parental history of MI or coronary revascularization (24%) and the greatest proportion of carriers (57%) of the chromosome 9p21 cardiovascular risk variant (rs10757269).

### Within Cluster Feature Weight Analysis

Although our unsupervised learning models discovered 4 phenotypically distinct, clinically relevant cohorts, we aimed to gain insight into which features were used most heavily in creating the low-rank model that was subsequently used for clustering. Although direct derivation of feature importance is difficult, we examined the relative importance of features in our models by multiplying cluster centroids by the low-rank matrix  $Y$ . This analysis demonstrated that features most heavily weighted in cluster 1 (oldest/most comorbid) were age, features related to poor walking tolerance (eg, angina, dyspnea, difficulty walking 1 block at average speed), sedentary behavior (eg, daily time spent sitting), cumulative pack years, and the majority of comorbidities including PAD, CHF, valve disease, arrhythmias, and rheumatologic conditions (Figure S2). Cluster 2 (youngest/multiethnic) was most weighted for physical and laboratory measures (eg, height, weight, CRP, fasting glucose, creatinine), diabetes and associated complications (CKD, retinopathy, neuropathy),

medications for diabetes, hypertension, and hyperlipidemia, history of MI or coronary revascularization, and parental cardiovascular history. Cluster 3 (middle-aged/lowest medication adherence) was most driven by aspirin, clopidogrel, and  $\beta$ -blocker use as well as racial minority status. Features most heavily weighted in cluster 4 (middle-aged/healthiest behaviors) were ABI measurements, each of the 3 Walking Impairment category scores (walking distance, speed, stair climbing), and features related to the amount and intensity of physical activity. Clusters 2 and 4 were the clusters that were most heavily influenced by SNPs associated with CAD, PAD, and ABI.

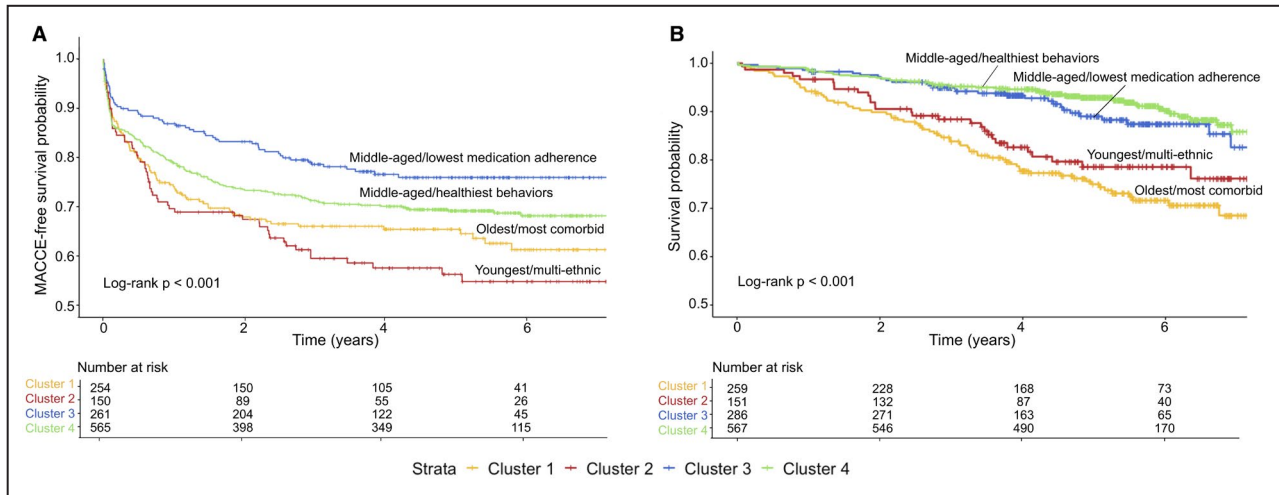
### Association of Cluster Identity With Clinical Outcomes

To assess the ability of cluster analysis to detect phenotypes that are clinically meaningful, we evaluated the association between cluster membership and long-term outcomes. The median duration of follow-up was 5.05 years (interquartile range, 3.8–6.1 years). Figure 4 shows Kaplan–Meier curves for MACCE and all-cause mortality. As illustrated, MACCE occurrence differed significantly across clusters (log-rank  $P$  value=2.0e-04; Figure 4A). Compared with cluster 3 (middle-aged/lowest medication adherence) with the lowest risk of MACCE, the MACCE risk was greatest in cluster 2 (youngest/multiethnic diabetics; HR, 2.2; 95% CI, 1.5–3.1), followed by cluster 1 (oldest/most comorbid; HR, 1.8; 95% CI, 1.3–2.5) and cluster 4 (middle-aged/healthiest behavior; HR, 1.5; 95% CI, 1.1–2.0). Although the individual rates of MI and stroke were similar across clusters, the occurrence of coronary and peripheral revascularization significantly differed (Table 2). The oldest/most comorbid cluster (cluster 1) had the highest rates of incident peripheral revascularization (5.1%). Rates of coronary revascularization were significantly higher among the youngest/multiethnic diabetic cluster 2 (33%), followed by the oldest/most comorbid cluster 1 (28%) and middle-aged group with healthiest behaviors in cluster 4 (27%).

With regard to all-cause mortality (Figure 4B), compared with cluster 4, the oldest/most comorbid cluster 1 was at the greatest risk of death (HR, 3.3; 95% CI, 2.3–4.8), followed by the youngest/multiethnic cluster 2 (HR, 2.6; 95% CI, 1.7–4.1; log-rank  $P$  value=2.0e-04). Of note, although cluster 4 had greater rates of MACCE compared with cluster 3, their mortality rates did not differ.

### Unsupervised Clustering Enhances Risk Stratification Compared With Standard Risk Prediction

Lastly, to explore how an unsupervised learning framework compares with traditional risk assessment, we compared the estimated hazards of a more



**Figure 4. Long-term outcomes of the 4 coronary artery disease clusters.** Kaplan–Meier curves showing (A) MACCE\* and (B) all-cause mortality. \*Primary MACCE composite included myocardial infarction, stroke, coronary revascularization, and peripheral revascularization. MACCE indicates major adverse cardiovascular and cerebrovascular events.

restrictive MACCE definition and all-cause mortality determined by cluster membership and PCE. As shown in Figure 5, classification based on the conventional PCE did not result in significant discrimination of MACCE risk between high-risk (HR, 1.1; 95% CI, 0.8–1.6), intermediate-risk (HR, 0.82; 95%, CI 0.5–1.4), or low-risk individuals in this cohort (Table S4). Similarly, the granular PCE model did not discriminate MACCE risk across 4 groups (Table S5). This is in contrast to the cluster models, which each demonstrated informative discrimination. Of the 4 clusters from our main unsupervised analysis, there were 3 distinct groups with significantly different MACCE risks (cluster 1 versus cluster 2 versus cluster 3/4; Figure 5, Table S6). The cluster sensitivity analysis including 5 clusters provided the most refined risk characterization with 4 unique MACCE risk groups (clusters 1/2 versus cluster 3 versus cluster 4 versus cluster 5); however, the clusters showed poorer between-cluster separation (Table S7). The 3-cluster model included 2 significantly different high versus low MACCE risk groups (clusters 1/3 versus cluster 2; Table S8).

In evaluating all-cause mortality alone, PCE classification similarly did not differentiate mortality risk in both the standard and granular PCE models (Tables S9 and 10). The only comparison that approached significance was with the high-risk PCE group (HR, 1.5; 95% CI, 0.98–2.3;  $P=0.06$ ). Cluster affiliation distinguished 3 unique mortality risk groups in the main 4 cluster models (cluster 1 versus cluster 2 versus clusters 3/4) as well as in the 5-cluster model. The 3-cluster model distinguished 2 distinct mortality risk groups (Tables S11 through S13).

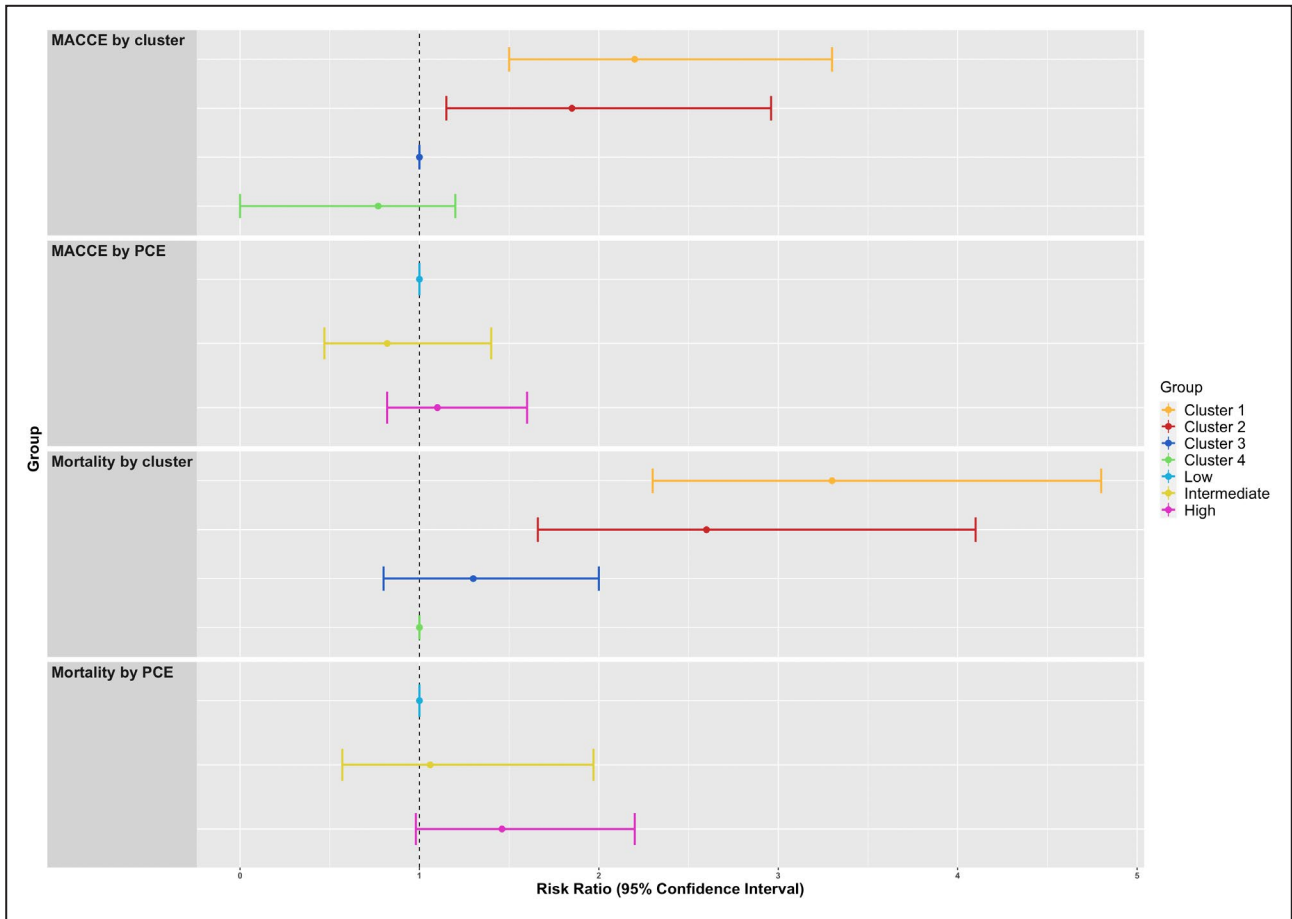
## DISCUSSION

In this longitudinal analysis of individuals with ASCVD, we applied contemporary unsupervised machine learning methods to discover subgroups with distinct sociodemographic, clinical, biological, and genetic characteristics. We demonstrate that unsupervised learning algorithms may be used to handle clinical data with heterogeneous and missing entries and generate

**Table 2. Crude Outcome Rates, Stratified by Clusters**

	Cluster 1 (oldest/most comorbid), N=254	Cluster 2 (youngest/multiethnic), N=150	Cluster 3 (middle-aged/lowest medication adherence), N=261	Cluster 4 (middle-aged/healthiest behaviors), N=565	Adjusted P value
MACCE	34.6	41.3	22.6	30.4	$1 \times 10^{-4}$
Myocardial infarction	2	2	2.7	1.6	NS
Stroke	2	3.3	1.1	1.6	NS
Coronary revascularization	27.6	33.3	19	26.5	0.0008
Peripheral revascularization	5.1	4	1.1	1.2	0.002
All-cause mortality	26.3	20	11	9	$9.2 \times 10^{-10}$

Primary MACCE was defined as a composite of myocardial infarction, stroke, coronary and/or peripheral revascularization. MACCE indicates major adverse cardiovascular and cerebrovascular events; and NS, nonsignificant.



**Figure 5. Comparison of clustering to PCE risk groups for prediction of MACCE\* and all-cause mortality.** \*PCE-consistent MACCE included myocardial infarction, stroke, and death. MACCE indicates major adverse cardiovascular and cerebrovascular events; and PCE, pooled cohort equations.

clinically important subclassifications with varying risks for long-term cardiovascular events. Furthermore, compared with traditional methods of risk assessment that employ a handful of known risk factors, our results highlight the potential of more refined cardiovascular risk stratification based on machine learning-based classification algorithms.

We took advantage of the extensive individual-level data in our ASCVD cohort to identify patient subgroups and obtain a composite view of their cross-group variation. What is progressive about our approach is that we did not provide specifications about how to partition the data based on data type or our expertise, but instead used agnostic approaches to process highly heterogeneous data types (including missing entries) and successfully identified 4 phenotypically and prognostically distinct patient groups. As clinical, biological, and imaging data continue to grow at an exponential pace and are widely deposited into electronic health records, clustering frameworks using low-rank modeling techniques may be increasingly important as they are particularly suited for processing heterogeneous patient data.

The resulting clusters may be used to further improve understanding of unique cohorts. For example, clusters 1 and 2 had characteristics that intuitively make sense to the cardiovascular practitioner—such as an older group of individuals with extensive histories of smoking and significant cardiovascular comorbidities including PAD and CHF.<sup>37</sup> However, our unsupervised learning algorithm also helped distinguish 2 groups that appeared otherwise similar. Although clusters 3 and 4 had apparently similar cardiovascular health (as reflected by conventional clinical and lifestyle risk factors), cluster 3 was defined by very low rates of adherence to preventive therapies, which might explain why this “healthy appearing” cluster had the highest cholesterol levels of all subgroups. Furthermore, in assessing within-cluster feature weighting, the most prominent features used for low-rank modeling in cluster 4 was family history of ASCVD events and the presence of the widely replicated 9p21 locus for CAD and MI.<sup>18,19,22</sup> Given these characteristics, it is interesting to note that cluster 4 had higher rates of MACCE, a prognosis that may be contextualized with relatively higher rates of triple-vessel/left main disease and a nearly 2-fold

increase in incident rates of coronary revascularization. It is possible that the elevated long-term MACCE risk in cluster 4 could be attributed to a greater prevalence of individuals in cluster 3 who had already been revascularized before study enrollment (1.8% versus 0.2%). However, despite healthier behaviors, knowledge that a group such as cluster 4 carries higher genetic risk can contribute importantly to targeting more aggressive primary and secondary prevention strategies.<sup>38–41</sup>

As exemplified by the discovered clusters of patients with CAD, a major advantage to unsupervised learning is that it can help reveal population health patterns and at the same time assist in developing tailored care strategies to specific groups. For example, recent advances in cardiovascular medicine include the demonstrated benefit of proprotein convertase subtilisin/kexin type 9 serine protease (PCSK9) inhibitors<sup>42</sup> and rivaroxaban,<sup>43</sup> however, broad use of these therapies is limited by cost and an associated increase in bleeding risk, respectively.<sup>44</sup> Our cluster analysis identified an “unfavorable phenotype” of younger patients who had a high prevalence of PAD and prior MI and mean low-density lipoprotein above target goals and were at the greatest long-term risk for cardiovascular events (cluster 2). Particularly in patients with prior MACCE history or concomitant arrhythmias, the optimal management strategy may include subsidization of therapies such as PCSK9 inhibitors or intensified antithrombotic regimens such as anticoagulant therapy. It is also important to note that this cluster was largely composed of minorities with lower socioeconomic status compared with other subgroups. Such a finding could reflect the need to better address social determinants of health that contribute to adverse outcomes in this group of individuals, such as improving access to health care and health-enhancing resources. Improving outcomes for cluster 2 may thus require different interventions for outreach and education. This is an especially important point as we found that many of these patients were prescribed appropriate medical therapy for their comorbidities yet still had poorer outcomes.

Lastly, compared with the American College of Cardiology/American Heart Association PCEs, we found that unsupervised clustering yields improved characterization of cardiovascular risk in a diverse and heterogeneous cohort of patients with ASCVD. Unlike the conventional PCE subgroups, clusters had statistically distinct risks for MI, stroke, and death. Cluster affiliations continued to be more informative than PCE scores in cluster sensitivity analyses and in comparison with a more granular PCE model with 4 risk categories, which suggests that the enhanced cluster performance was not primarily attributed to having more subgroups. Taken together, these results suggest that unsupervised clustering may be used to support integration of multiple types of patient data to better capture differing

trajectories of disease risk. Although the application of machine learning in medicine has yet to fully materialize in clinical use, our data support their ability to identify clinically important ASCVD strata. Indeed, implementing these data-driven methods may enable automated clinical scoring systems and generation of meaningful clinical insights from data already captured throughout the health care system.

## Study Limitations

Our findings should be interpreted in the context of several limitations. First, the generalizability of the current study may be limited by inclusion of patients from 2 tertiary academic centers where study enrollment was conditioned on need for angiography. Thus our participant group is likely higher risk given that angiography is generally performed in individuals with a high probability of having hemodynamically significant disease. Our findings could have been biased by variables that influenced being selected for the study and were also associated with experiencing MACCE (eg, Berkson's bias).<sup>45</sup> These factors may include health care access, symptom severity, and symptom recognition and could have induced distorted associations between variables that may explain the unexpected finding of elevated MACCE risk in cluster 4 compared with cluster 3. Furthermore, the phenotypic differences in the cohort were likely influenced by population structure, including social strata and genetic variation related to geographic distribution. Thus, our clustering algorithm should be evaluated in broader populations and in data sets of differing structures to explore reliability (eg, with more bias or missingness than a carefully maintained clinical trial registry). Evaluation in larger cohorts is also important to validate the within-cluster phenomenon observed in our study, including characteristics that appear to increase or attenuate risk. Similarly, we acknowledge that patient groupings and thus defining features may differ depending on available clinical variables. Lastly, the model included a select list of SNPs used as markers of systemic atherosclerosis (including PAD) and therefore provided a limited assessment of the cohort's genetic risk of cardiovascular disease and events. Because of the complex nature of ASCVD, combining deep phenotypic information with a broader reflection of genetic risk, such as through a polygenic risk score, may provide more powerful risk estimation.<sup>38</sup> It is also noteworthy that in the multiethnic cluster 2, the allele frequency was low for many of the included genome-wide association study variants. This observation may reflect the limited transferability of variants derived from European genome-wide association studies<sup>36</sup> and highlight the importance of ongoing genetic discovery in non-European populations to permit broad and equitable implementation of genetic risk prediction tools.

## CONCLUSIONS

By applying contemporary unsupervised learning techniques to ASCVD, we identify 4 groups of patients that differ across a wide range of health variables and subsequent risk of adverse outcomes. We show that flexible clustering analysis of heterogeneous data (including mixed and missing values) is feasible and is able to identify prognostically distinct clusters that partition at greater resolution than groups formed solely based on standard risk factors. Our results confirm the heterogeneity of ASCVD and highlight the possibility that flexible and unbiased machine learning algorithms can be used to identify important subpopulations based on data that are collected in clinical practice.

## ARTICLE INFORMATION

Received April 9, 2021; accepted October 14, 2021.

### Affiliations

Division of Vascular Surgery, Department of Surgery, Stanford University School of Medicine, Stanford, CA (A.M.F., A.V.E., N.J.L., E.G.R.); Center for Biomedical Informatics Research, Stanford University, Stanford, CA (A.S., N.H.S., E.G.R.); Zena and Michael A. Wiener Cardiovascular Institute, Marie-Josée and Henry R. Kravis Center for Cardiovascular Health, Icahn School of Medicine at Mount Sinai, New York, NY (J.W.O.); Department of Cardiovascular Sciences, Houston Methodist Research Institute, Houston, TX (J.P.C.); Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA (N.J.L.); and Stanford Cardiovascular Institute, Stanford, CA (N.J.L., E.G.R.).

### Acknowledgments

The corresponding author had full access to all the data in the study and takes responsibility for its integrity and the data analysis.

### Sources of Funding

This study was supported by American Heart Association EIA34770065, Dallas, TX (Dr Leeper); National Institutes of Health/National Heart, Lung, and Blood Institute R35HL144475, Bethesda, MD (Dr Leeper); National Institutes of Health/National Library of Medicine R01LM011369-06, Bethesda, MD (Dr Shah); Deutsche Herzstiftung (A.V. Eberhard); the 2018-2019 Society of University Surgeons Junior Faculty Award, Los Angeles, CA (Dr Ross); and National Institutes of Health/National Heart, Lung, and Blood Institute 1K01HL148639-01, Bethesda, MD (Dr Ross). The GenePAD study was supported by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute K12HL087746 and National Institutes of Health/National Heart, Lung, and Blood Institute R01HL075774, Bethesda, MD (Dr Cooke).

### Disclosures

None.

### Supplementary Material

Tables S1–S13  
Figures S1–S2

## REFERENCES

- Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J III. Factors of risk in the development of coronary heart disease—six year follow-up experience. The Framingham Study. *Ann Intern Med.* 1961;55:33–50. doi: 10.7326/0003-4819-55-1-33
- Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association task force on practice guidelines. *Circulation.* 2014;129:S49–S73. doi: 10.1161/01.cir.0000437741.48606.98
- Stoeckenbroek RM, Boekholdt SM, Luben R, Hovingh GK, Zwiderman AH, Wareham NJ, Khaw KT, Peters RJ. Heterogeneous impact of classic atherosclerotic risk factors on different arterial territories: the EPIC-Norfolk prospective population study. *Eur Heart J.* 2016;37:880–889. doi: 10.1093/eurheartj/ehv630
- Price JF, Mowbray PI, Lee AJ, Rumley A, Lowe GD, Fowkes FG. Relationship between smoking and cardiovascular risk factors in the development of peripheral arterial disease and coronary artery disease: Edinburgh artery study. *Eur Heart J.* 1999;20:344–353.
- Ding N, Sang Y, Chen J, Ballew SH, Kalbaugh CA, Salameh MJ, Blaha MJ, Allison M, Heiss G, Selvin E, et al. Cigarette smoking, smoking cessation, and long-term risk of 3 major atherosclerotic diseases. *J Am Coll Cardiol.* 2019;74:498–507.
- Levin MG, Klarin D, Assimes TL, Freiberg MS, Ingelsson E, Lynch J, Natarajan P, O'Donnell C, Rader DJ, Tsao PS, et al. Genetics of smoking and risk of atherosclerotic cardiovascular diseases: a Mendelian randomization study. *JAMA Netw Open.* 2021;4:e2034461. doi: 10.1001/jamanetworkopen.2020.34461
- Shah RV, Yeri AS, Murthy VL, Massaro JM, D'Agostino R, Freedman JE, Long MT, Fox CS, Das S, Benjamin EJ, et al. Association of multiorgan computed tomographic phenomaps with adverse cardiovascular health outcomes: the Framingham Heart Study. *JAMA Cardiol.* 2017;2:1236–1246. doi: 10.1001/jamacardio.2017.3145
- Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, Marwick TH. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol.* 2019;73:1317–1335. doi: 10.1016/j.jacc.2018.12.054
- Levine J, Simonds E, Bendall S, Davis K, Amir el AD, Tadmor M, Litvin O, Fienberg H, Jager A, Zunder E, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell.* 2015;162:184–197. doi: 10.1016/j.cell.2015.05.047
- Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, Zhou A, Eyob H, Balakrishnan S, Wang C-Y, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature.* 2015;526:131–135. doi: 10.1038/nature15260
- Grant RW, McCloskey J, Hatfield M, Uratsu C, Ralston JD, Bayliss E, Kennedy CJ. Use of latent class analysis and k-means clustering to identify complex patient profiles. *JAMA Netw Open.* 2020;3:e2029068. doi: 10.1001/jamanetworkopen.2020.29068
- Schuler A, Liu V, Wan J, Callahan A, Udell M, Stark DE, Shah NH. Discovering patient phenotypes using generalized low rank models. *Pac Symp Biocomput.* 2016;21:144–155.
- Ahmad T, Pencina MJ, Schulte PJ, O'Brien E, Whellan DJ, Pina IL, Kitzman DW, Lee KL, O'Connor CM, Felker GM. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J Am Coll Cardiol.* 2014;64:1765–1774.
- Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiane M, Bonow RO, Huang CC, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation.* 2015;131:269–279. doi: 10.1161/CIRCULATIONAHA.114.010637
- Inohara T, Shrader P, Pieper K, Blanco RG, Thomas L, Singer DE, Freeman JV, Allen LA, Fonarow GC, Gersh B, et al. Association of atrial fibrillation clinical phenotypes with treatment patterns and outcomes: a multicenter registry study. *JAMA Cardiol.* 2018;3:54–63. doi: 10.1001/jamacardio.2017.4665
- Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, Dahlstrom U, O'Connor CM, Felker GM, Desai NR. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *J Am Heart Assoc.* 2018;7:e008081. doi: 10.1161/JAHA.117.008081
- Nead KT, Zhou MJ, Caceres RD, Sharp SJ, Wehner MR, Olin JW, Cooke JP, Leeper NJ. Usefulness of the addition of beta-2-microglobulin, cystatin C and C-reactive protein to an established risk factors model to improve mortality risk prediction in patients undergoing coronary angiography. *Am J Cardiol.* 2013;111:851–856. doi: 10.1016/j.amjcard.2012.11.055
- Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447:661–678.
- Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, et al.

- A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*. 2007;316:1491–1493. doi: 10.1126/science.1142842
20. Wassel CL, Lamina C, Nambi V, Coassin S, Mukamal KJ, Ganesh SK, Jacobs DR, Franceschini N, Papanicolaou GJ, Gibson Q, et al. Genetic determinants of the ankle-brachial index: a meta-analysis of a cardiovascular candidate gene 50K SNP panel in the candidate gene association resource (CARE) consortium. *Atherosclerosis*. 2012;222:138–147. doi: 10.1016/j.atherosclerosis.2012.01.039
  21. Cluett C, McDermott MM, Guralnik J, Ferrucci L, Bandinelli S, Miljkovic I, Zmuda JM, Li R, Tranah G, Harris T, et al. The 9p21 myocardial infarction risk allele increases risk of peripheral artery disease in older people. *Circ Cardiovasc Gene*. 2009;2:347–353. doi: 10.1161/CIRCGENETICS.108.825935
  22. McPherson R, Pertsemlidis A, Kavavlar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, et al. A common allele on chromosome 9 associated with coronary heart disease. *Science*. 2007;316:1488–1491. doi: 10.1126/science.1142447
  23. Murabito JM, White CC, Kavousi M, Sun YV, Feitosa MF, Nambi V, Lamina C, Schillert A, Coassin S, Bis JC, et al. Association between chromosome 9p21 variants and the ankle-brachial index identified by a meta-analysis of 21 genome-wide association studies. *Circ Cardiovasc Gene*. 2012;5:100–112. doi: 10.1161/CIRCGENETICS.111.961292
  24. Wilson AM, Sadrzadeh-Rafie AH, Myers J, Assimes T, Nead KT, Higgins M, Gabriel A, Olin J, Cooke JP. Low lifetime recreational activity is a risk factor for peripheral arterial disease. *J Vasc Surg*. 2011;54:427–432, 432 e421–424. doi: 10.1016/j.jvs.2011.02.052
  25. Nead KT, Zhou M, Diaz Caceres R, Olin JW, Cooke JP, Leeper NJ. Walking impairment questionnaire improves mortality risk prediction models in a high-risk cohort independent of peripheral arterial disease status. *Circ Cardiovasc Qual Outcomes*. 2013;6:255–261. doi: 10.1161/CIRCOUTCOMES.111.000070
  26. Nead KT, Cooke JP, Olin JW, Leeper NJ. Alternative ankle-brachial index method identifies additional at-risk individuals. *J Am Coll Cardiol*. 2013;62:553–559. doi: 10.1016/j.jacc.2013.04.061
  27. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: an Introduction to Cluster Analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.; 1990.
  28. Udell M, Horn C, Zadeh R, Boyd S. Generalized low rank models. arXiv preprint arXiv:1410.0342. 2014. Available at: <https://arxiv.org/abs/1410.0342>
  29. Dalton L, Ballarin V, Brun M. Clustering algorithms: on learning, validation, performance, and applications to genomics. *Curr Genomics*. 2009;10:430–445.
  30. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. 2010;11:94. doi: 10.1186/1471-2156-11-94
  31. Jombart T. Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 2008;24:1403–1405. doi: 10.1093/bioinformatics/btn129
  32. Okabe A, Boots B, Sugihara K, Chiu SN. *Spatial Tessellations: Concepts and applications of Voronoi diagrams*. 2nd ed; 2000.
  33. Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, Ballantyne CM, Go AS. Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population. *J Am Coll Cardiol*. 2016;67:2118–2130.
  34. D'Agostino RB, Grundy S, Sullivan LM, Wilson P, Group CHDRP. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*. 2001;286:180–187. doi: 10.1001/jama.286.2.180
  35. Muntner P, Colantonio LD, Cushman M, Goff DC Jr, Howard G, Howard VJ, Kissela B, Levitan EB, Lloyd-Jones DM, Safford MM. Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations. *JAMA*. 2014;311:1406–1415. doi: 10.1001/jama.2014.2630
  36. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell*. 2019;177:26–31. doi: 10.1016/j.cell.2019.02.048
  37. Lu JT, Creager MA. The relationship of cigarette smoking to peripheral arterial disease. *Rev Cardiovasc Med*. 2004;5:189–193.
  38. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50:1219–1224. doi: 10.1038/s41588-018-0183-z
  39. Natarajan P, Young R, Stitzel NO, Padmanabhan S, Baber U, Mehran R, Sartori S, Fuster V, Reilly DF, Butterworth A, et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*. 2017;135:2091–2101. doi: 10.1161/CIRCULATION.116.024436
  40. Mega JL, Stitzel NO, Smith JG, Chasman DI, Caulfield MJ, Devlin JJ, Nordio F, Hyde CL, Cannon CP, Sacks FM, et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet*. 2015;385:2264–2271. doi: 10.1016/S0140-6736(14)61730-X
  41. Kullo IJ, Jouni H, Austin EE, Brown S-A, Krusselbrink TM, Isseh IN, Haddad RA, Marroush TS, Shameer K, Olson JE, et al. Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (the MI-GENES clinical trial). *Circulation*. 2016;133:1181–1188. doi: 10.1161/CIRCULATION.115.020109
  42. Sabatine MS. PCSK9 inhibitors: clinical evidence and implementation. *Nat Rev Cardiol*. 2019;16:155–165. doi: 10.1038/s41569-018-0107-8
  43. Eikelboom JW, Connolly SJ, Bosch J, Dagenais GR, Hart RG, Shestakovska O, Diaz R, Alings M, Lonn EM, Anand SS, et al. Rivaroxaban with or without aspirin in stable cardiovascular disease. *N Engl J Med*. 2017;377:1319–1330. doi: 10.1056/NEJMoa1709118
  44. Hlatky MA, Kazi DS. PCSK9 inhibitors: economics and policy. *J Am Coll Cardiol*. 2017;70:2677–2687. doi: 10.1016/j.jacc.2017.10.001
  45. Munafo MR, Tilling K, Taylor AE, Evans DM, Davey SG. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol*. 2018;47:226–235. doi: 10.1093/ije/dyx206

# **SUPPLEMENTAL MATERIAL**

**Table S1. Original variables used in the generalized low rank model approximation.**

<b>Categories</b>	<b>Variables</b>
Demographics	Age, sex, self-reported race/ethnicity
Physical measures	Height, weight, blood pressure, heart rate, BMI, ABI
Laboratory results	Total Cholesterol, LDL, HDL, non-HDL, total cholesterol to HDL ratio, LDL to HDL level, serum glucose Creatinine, glomerular filtration rate Biomarkers: C-Reactive Protein, Cystatin C, $\beta$ -2 microglobulin
Imaging	Coronary angiography findings (graded by one-vessel, two-vessel, or triple-vessel/left main disease)
Medical history	CAD, MI, PAD, carotid stenosis, stroke, CHF, arrhythmias, AAA Diabetes, CKD, diabetic neuropathy, menopause
Surgical history	Coronary revascularization (coronary bypass or percutaneous coronary intervention), abdominal aortic aneurysm repair, lower extremity amputation, valve surgery
Reported symptoms	Angina, shortness of breath, claudication, joint pain
Physical activity assessment	Walking Impairment Questionnaire (walking distance, walking speed, and stair climbing) Physical activity questionnaire (lifetime physical activity patterns)



Family history	History of cardiovascular diseases in parents including MI and coronary revascularization, stroke, lower extremity revascularization, and AAA rupture or surgery
Medications	Anti-platelet therapy: Aspirin, clopidogrel Lipid-lowering therapy: statins, other Anti-hypertensive: ACE inhibitor, ARB, diuretics, other Beta-blockers, insulin or hypoglycemic agents Polypharmacy (total number of medications taking at baseline)
Genetic variants associated with CAD, PAD, or lower ABI	rs290481, rs819750, rs7100623, rs7003385, rs94855286, rs4659996, rs3745274, rs2171209, rs16824978, rs10757269
Social factors	Ever married, living situation, total education, current income, employment status
Lifestyle behavior	Smoking: Ever smoked, current smoker, cumulative pack years Alcohol: weekly alcohol consumption, alcohol consumption pattern over lifetime

AAA, abdominal aortic aneurysm. ABI, ankle-brachial index. ACE, angiotensin converting enzyme. ARB, angiotensin II receptor blocker. BMI, body mass index. CAD, coronary artery disease. CHF, congestive heart failure. CKD, chronic kidney disease. HDL, high-density lipoprotein. LDL, low-density lipoprotein. MI, myocardial infarction.

**Table S2. Missingness for variables used in the generalized low rank model approximation.**

<b>Variable</b>	<b>% Missing</b>
<b>Demographics</b>	
Age,	0.45%
Sex	0.07%
Race/ethnicity	0.15%
<b>Socioeconomics</b>	
Income	1.2%
Highest education level	0.38%
<b>Lifestyle</b>	
Current smoker	1.3%
Cumulative pack years	4.9%
Engages in exercise at least once per week	7.3%
<b>Clinical</b>	
BMI	0.98%
Ankle-Brachial Index	0.83%
Systolic blood pressure	0.07%
CAD angiography grading	1.9%
<b>Clinical History</b>	
Prior MI	1.2%
Prior CABG or PCI	1.3%
Prior valve surgery	1.3%

CHF	1.3%
Stroke or TIA	0%
PAD	0.7%
Lower extremity amputation	0.98%
Cardiac arrhythmia	1.3%
Chronic kidney disease	1.2%
Diabetes	2.6%
<b>Biological</b>	
$\beta$ -2 microglobulin*	79%
Cystatin C*	79%
CRP <sup>†</sup>	65%
Glucose, maximum	2.1%
Total Cholesterol	1.4%
LDL	2.6%
Creatinine	0.98%
<b>Medications</b>	
Anti-hypertensive	0.15%
Insulin or hypoglycemic agents	0.15%
Cholesterol-lowering meds	0.15%
Aspirin	12.3%
Clopidogrel	12.3%
Statin	12.3%

β-blockers	12.3%
<b>Genetics</b>	
rs10757269	12%
rs819750	11%
rs94855286	11%
rs2171209	11%
rs7100623	13%
rs16824978	12%
rs7003385	11%
rs4659996	11%
rs3745274	13%
rs290481	11%
<p>*Based on subsample of 268 individuals; †Based on subsample of 459 individuals; BMI, body mass index. CABG, coronary artery bypass graft. CAD, coronary artery disease. CHF, congestive heart failure. CRP, C-reactive protein. LDL, low-density lipoprotein. MI, myocardial infarction. NS, non-significant <i>P</i> value. PAD, peripheral arterial disease. PCI, percutaneous coronary intervention. TIA, transient ischemic attack.</p>	

**Table S3. Cluster validation statistics.**

<b>Measure</b>	<b>Result</b>	<b>Cluster number recommendation</b>
<b>Stability measures</b>		
Average distance	2.6	6
Figure of Merit	0.27	6
Average proportion of non-overlap	0	4
Average distance between means	0.0	4
<b>Internal measures</b>		
Dunn index	0.77	4
Connectivity	2.9	2
Silhouette width	0.65	2
Results of cluster validation statistics on generalized linear model approximations. Cluster number was chosen based on the majority recommendation.		

**Table S4. Hazard ratios for risk of PCE-consistent MACCE composite by Pooled Cohort**

**Equations group:** Standard PCE model (<2.5% considered low risk, 2.5-4.9% considered intermediate risk,  $\geq$ 5% considered high risk)

MACCE risk by Pooled Cohort Equations group			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Low Risk	Ref	-	-
Intermediate Risk	0.82	0.5-1.4	<i>NS</i>
High Risk	1.1	0.8-1.6	<i>NS</i>

**Table S5. Hazard ratios for risk of PCE-consistent MACCE composite by Pooled Cohort**

**Equations group:** Granular PCE model (<2.5% considered low risk, 2.5-3.74% as intermediate low risk, 3.75-4.9% as intermediate high risk, and  $\geq 5\%$  considered high risk)

MACCE risk by Pooled Cohort Equations group			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Low Risk	Ref	-	-
Intermediate Low	0.89	0.44-1.8	<i>NS</i>
Intermediate High	0.74	0.33-1.6	<i>NS</i>
High Risk	1.1	0.82-1.6	<i>NS</i>

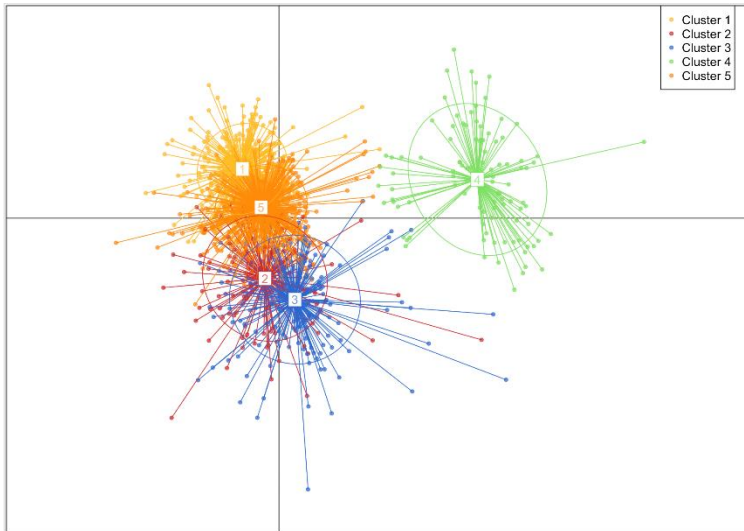
**Table S6. Hazard ratios for risk of PCE-consistent MACCE composite by Clusters:**

Main four cluster model

MACCE risk by 4 clusters			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Cluster 1	2.2	1.5-3.3	<b>0.004</b>
Cluster 2	1.9	1.2-2.9	<b>0.03</b>
Cluster 3	Ref	-	-
Cluster 4	0.77	0-1.2	<i>NS</i>

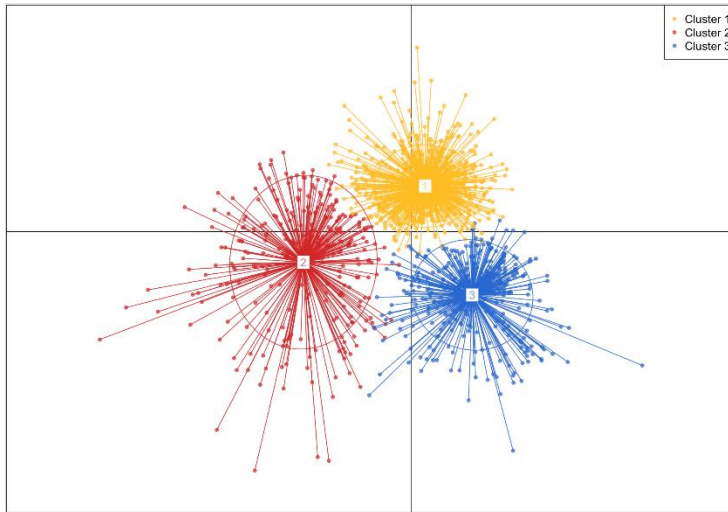


**Table S7. Cluster sensitivity analysis including 5 clusters**



MACCE risk by Cluster 5 (sensitivity analysis)			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Cluster 1	Ref		
Cluster 2	1.3	0.86-1.9	0.2
Cluster 3	1.7	1.1-2.6	<b>0.02</b>
Cluster 4	2.6	1.8-3.8	<b>1.5e-7</b>
Cluster 5	2.8	1.8-4.3	<b>6.6e-6</b>

**Table S8. Cluster sensitivity analysis including 3 clusters**



MACCE risk by 3 clusters			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Cluster 1	Ref	-	-
Cluster 2	2.4	1.8-3.3	<b>1.6e-8</b>
Cluster 3	1.2	0.86-1.8	<i>NS</i>

**Table S9. Hazard ratios for all-cause mortality risk by Clusters or Pooled Cohort**

**Equations group:** Standard PCE model (<2.5% considered low risk, 2.5-4.9% considered intermediate risk,  $\geq$ 5% considered high risk)

All-cause mortality risk by Pooled Cohort Equations group			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Low Risk	Ref	-	-
Intermediate Risk	1.1	0.57-1.9	<i>NS</i>
High Risk	1.5	0.98-2.2	0.06

**Table S10. Hazard ratios for all-cause mortality risk by Clusters or Pooled Cohort**

**Equations group:** Granular PCE model (<2.5% considered low risk, 2.5-3.74% as intermediate low risk, 3.75-4.9% as intermediate high risk, and  $\geq 5\%$  considered high risk)

All-cause mortality risk by Pooled Cohort Equations group			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Low Risk	Ref	-	-
Intermediate Low	1.2	0.55-2.4	<i>NS</i>
Intermediate High	0.9	0.40-2.2	<i>NS</i>
High Risk	1.5	0.98-2.2	0.06

**Table S11. Hazard ratios for risk of all-cause mortality by Clusters: Main four cluster model**

All-cause mortality risk by Cluster 4 (main analysis)			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Cluster 1	3.3	2.3-4.8	<b>1.3e-09</b>
Cluster 2	2.6	1.7-4.1	<b>1.6e-04</b>
Cluster 3	1.3	0.8-2.0	<i>NS</i>
Cluster 4	Ref	-	-

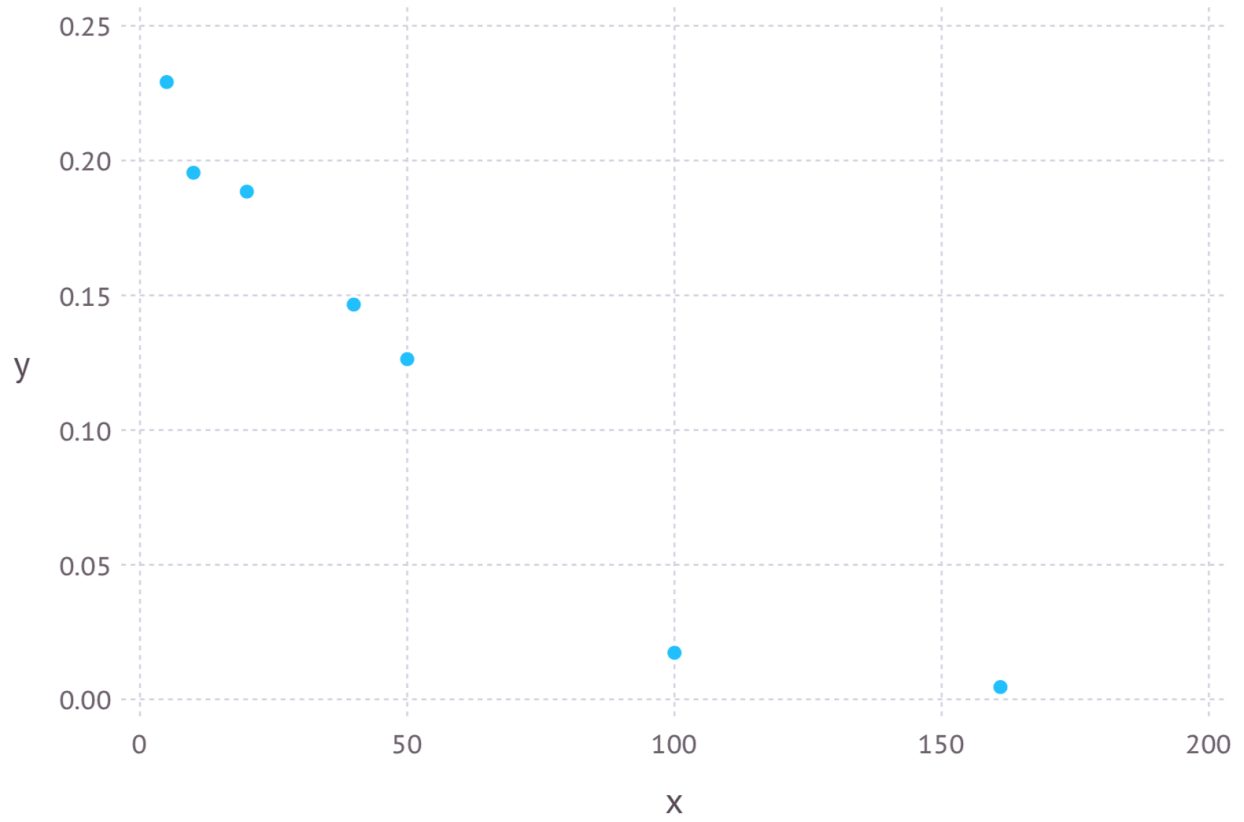
**Table S12. Hazard ratios for risk of all-cause mortality by Clusters: Five cluster model**

All-cause mortality risk by Cluster 5 (sensitivity analysis)			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Cluster 1	Ref	-	-
Cluster 2	1.4	0.80-2.5	0.2
Cluster 3	2.5	1.5-4	<b>0.002</b>
Cluster 4	2.1	1.2-3.7	<b>0.008</b>
Cluster 5	0.75	0.47-1.2	0.2

**Table S13. Hazard ratios for risk of all-cause mortality by Clusters: Three cluster model**

All-cause mortality risk by Cluster 3 (sensitivity analysis)			
	Hazard Ratio	95% CI	Adjusted <i>P</i> value
Cluster 1	Ref	-	-
Cluster 2	2.8	2.0-3.9	<b>1.2e-9</b>
Cluster 3	1.2	0.8-1.9	0.27

**Figure S1. Plot of rank vs. training error used to select the rank for the GLRM.** Model rank was set to 50 to maximize the respective decrease in training error, while preventing the risk of overfitting at higher rank. x-axis: rank. y-axis: training error.





**Figure S2. Heatmap of feature weight in the low rank model used for clustering.** Darker colors represent features most heavily weighted and used to build a centroid for each cluster.

