American Journal of Preventive Cardiology 4 (2020) 100118

Contents lists available at ScienceDirect

American Journal of Preventive Cardiology

journal homepage: www.journals.elsevier.com/the-american-journal-of-preventive-cardiology

Original Research

County-level phenomapping to identify disparities in cardiovascular outcomes: An unsupervised clustering analysis
Short title: Unsupervised clustering of counties and risk of cardiovascular mortality

Matthew W. Segar [a,1], Shreya Rao [a,1], Ann Marie Navar [a], Erin D. Michos [b], Alana Lewis [c], Adolfo Correa [d], Mario Sims [d], Amit Khera [a], Amy E. Hughes [e], Ambarish Pandey [a,*]

[a] Division of Cardiology, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA
[b] Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[c] Department of Preventive Medicine, Northwestern University, Chicago, IL, USA
[d] Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA
[e] Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA

ARTICLE INFO

Keywords:
Cardiovascular disease
Epidemiology
Risk factors
Machine learning

ABSTRACT

Introduction: Significant heterogeneity in cardiovascular disease (CVD) risk and healthcare resource allocation has been demonstrated in the United States, but optimal methods to capture heterogeneity in county-level characteristics that contribute to CVD mortality differences are unclear. We evaluated the feasibility of unsupervised machine learning (ML)-based phenomapping in identifying subgroups of county-level social and demographic risk factors with differential CVD outcomes.

Methods: We performed a cross-sectional study using county-level data from 2008 to 2018 from the Centers for Disease Control (CDC) WONDER platform and the 2020 Robert Wood Johnson County Health Rankings program. Unsupervised clustering was performed on 46 facets of population characteristics spanning the demographic, health behaviors, socioeconomic, and healthcare access domains. Spatial autocorrelation was assessed using the Moran's I test, and temporal trends in age-adjusted CVD outcomes were evaluated using linear mixed effect models and least square means.

Results: Among 2676 counties, 4 county-level phenogroups were identified (Moran's I p-value <0.001). Phenogroup 1 (N = 924; 24.5%) counties were largely white, suburban households with high income and access to healthcare. Phenogroup 2 counties (N = 451; 16.9%) included predominantly Hispanic residents and below-average prevalence of CVD risk factors. Phenogroup 3 (N = 951; 35.5%) counties included rural, white residents with the lowest levels of access to healthcare. Phenogroup 4 (350; 13.1%) comprised counties with predominantly Black residents, substantial cardiovascular comorbidities, and physical and socioeconomic burdens. Least square means in age-adjusted cardiovascular mortality over time increased in a stepwise fashion from 223 in phenogroup 1 to 317 per 100,000 residents in phenogroup 4.

Conclusions: Unsupervised ML-based clustering on county-level population characteristics can identify unique phenogroups with differential risk of CVD mortality. Phenogroup identification may aid in developing a uniform set of preventive initiatives for clustered counties to address regional differences in CVD mortality.

Substantial improvements in life expectancy and mortality attributable to cardiovascular disease (CVD) have been observed since the early 1980's, driven by reductions in the burden of cardiovascular risk factors and improvements in healthcare [1]. For the first time in decades, however, life expectancy in the United States declined in 2015 and 2016, with rates of CVD on the rise among some groups [2]. Regional variation is apparent in both of these alarming trends. Over 25 years, the Global Burden of Disease 2016 Study found pronounced and persistent

* Corresponding author. Division of Cardiology Department of Internal Medicine University of Texas Southwestern Medical Center 5323 Harry Hines Boulevard, Dallas, TX, 75390, USA.
E-mail address: ambarish.pandey@utsouthwestern.edu (A. Pandey).
[1] Equal Contribution.

https://doi.org/10.1016/j.ajpc.2020.100118
Received 20 September 2020; Received in revised form 21 October 2020; Accepted 23 October 2020

disparities in the burden of CVD across counties [2,3]. Despite declining CVD rates nationally, counties that in 1990 had the lowest rates of CVD and most rapid decline in disease burden continue to demonstrate low rates of CVD today, while those counties experiencing high levels of disease and mortality 25 years ago continue to lag and now report plateauing or increasing rates of CVD [2,4–6].

The highest rates of CVD mortality can be found across the Gulf Coast to West Virginia, concentrated in counties with high rates of cardiac risk factors, inadequate access to healthcare and low socioeconomic status [7–10]. Prior research has focused on the role of these independent risk factors in driving regional disparities, and demonstrates the potential for early intervention in high-risk communities to impact CVD trends. However, targets for multi-level community health interventions to address such disparities are ill-defined, and tools for identifying at-risk communities are lacking [11]. Machine learning (ML) has emerged as a tool uniquely suited to detecting patterns to explain the observed heterogeneity in CVD risk and outcomes [12].

In this study, we hypothesized that unsupervised ML-based clustering could successfully identify clusters of phenotypically similar counties, based on social, demographic, behavioral and health-related risk factors, and that these phenogroups would demonstrate independent associations with observed disparities in CVD outcomes among U.S. counties. Identifying subgroups of counties with overlapping characteristics and potential drivers of CVD outcomes may inform preventive programs and public health policies to address regional differences in CVD mortality.

1. Methods

1.1. Data sources

We analyzed publicly available data from the Robert Wood Johnson Foundation's 2020 County Health Rankings (CHR) database and Centers for Disease Control and Prevention Wide-Ranging Online Data for Epidemiologic Research (CDC WONDER). CHR data includes measures collected between 2016 and 2018 and comprises information on county-level measures of health for more than 3000 U S. counties from multiple sources, including the Behavioral Risk Factor Surveillance System, American Community Survey and National Center for Health Statistics databases [13]. The database includes measures of health outcomes and factors encompassing four domains: (1) health behaviors, (2) clinical care, (3) social and economic factors, and (4) physical environment characteristics (Supplemental Table 1). CHR 2020 data was used as it provides the most contemporary cohort and includes the greatest number of county-level covariates compared to previous years. Counties with >10% missing covariates were excluded. This study was approved with IRB exemption from the University of Texas Southwestern Medical Center, Dallas, Texas.

Outcomes data were obtained from CDC WONDER, an online database of county-level underlying cause of death reported for all U.S. counties from 2008 to 2018 [30]. Our primary outcome of interest was temporal trend in age-adjusted cardiovascular mortality per 100,000 population per year, derived from death certificates of U.S. residents and classified by four-digit International Classification of Diseases 10th Revision (ICD-10) codes [31]. Cardiovascular deaths were defined as those attributed to ischemic heart disease (IHD), hypertensive disease, rheumatic heart disease, myocarditis, heart failure (HF), and arrhythmias under ICD-10 codes I00–I99. Secondary outcomes of interest included age-adjusted mortality for IHD, mortality for HF, and all-cause mortality per 100,000 population.

1.2. Exposures of interest

To identify clinically informative covariates from the CHR database and eliminate redundancy among predictors we evaluated Pearson correlation coefficients among 46 candidate variables, excluding covariates with a correlation coefficient >0.7 (N = 9). Among groups of correlated covariates, we eliminated more downstream risk factors, with the aim of keeping upstream predictors of outcomes in the model. No covariates had a high proportion of missing values (>10%). This process yielded 37 covariates, including measures from all four domains including demographic indicators and markers of income, access to healthcare, high-risk behaviors and physical environment. A list of the included and excluded health factors and their definitions is provided in the Supplement.

1.3. Statistical analysis

Data were scaled and standardized to a mean of 0 and standard deviation of 1. Missing data were imputed using random forest imputation [14]. Phenotypic clusters were defined using unsupervised hierarchical clustering of principal components (HCPC), a process by which data undergo agglomerative hierarchical clustering followed by K-means consolidation [15]. Because extraneous variables can misguide clustering results, principal components analysis was first performed to reduce the dimensionality of the covariate data without losing significant variation among important features. The optimal number of phenogroups was determined based on the absolute loss of within-cluster inertia (a measure of variance). Once phenogroups were identified, spatial autocorrelation was assessed using the Moran's I test [16]. Variable importance of phenogroup characteristics was assessed by a V-test (a measure of the over or underrepresentation of the variable within the phenogroup) with an underlying hypergeometric distribution. Specifically, a positive V-test value indicates the variable overrepresents the phenogroup while a negative number indicates the variable underrepresents the phenogroup. Characteristics across phenogroups were summarized as means (standard deviations) and differences across phenogroups were evaluated using one-way ANOVA.

Stepwise linear mixed effect regression with forward selection based on Akaike Information Criterion was used to identify the 10 strongest predictors of CVD mortality over time. Linear mixed effect modeling was also performed to evaluate the temporal trends in age-adjusted CVD outcomes with phenogroups. Differences in outcomes were displayed as least square means. Models were adjusted for the same 10 covariates identified in the stepwise forward selection step. To assess if phenogroup membership improved prognostic performance above and beyond the individual components, the improvement in the best fit model with versus without phenogroup membership was assessed as previously described [12]. The change in R^2 between the best fit model with and without phenogroup membership was determined by the Davidson-MacKinnon J-test [17]. To assess the temporal stability in the observed differences across phenogroups using the CHR 2020 dataset, sensitivity analyses were performed comparing the county-level characteristics across the phenogroups using data from the CHR 2010 cohort (the earliest available dataset). Analyses were performed using R version 3.6.0 (R Foundation) using the *FactoMineR*, *ape*, and *lmtest* packages [15,18,19]. A two-sided $P < 0.05$ was considered as statistically significant.

2. Results

Of the 3193 counties in the CHR dataset, 517 were excluded due to missing outcomes data for a total of 2676 counties and 37 variables included in the primary analysis. Eigenvalues for the first 24 components were >1 and these components accounted for 91% of the variance in the covariate data. Therefore, we used the first 24 components as input for the HCPC model [15]. The optimal number of phenogroups was 4. Baseline characteristics across county phenogroups are displayed in Table 1. Differences between phenogroups, as determined by the V-test, are summarized in Fig. 1 and Supplemental Table 2. Predictors of phenogroup membership included risk factors such as race, ethnicity, comorbid disease with HIV and diabetes, markers of access to healthcare, socioeconomic determinants such as stability of housing and medium household income, and behavioral risk factors including rates of physical

Table 1
Baseline county-level demographic characteristics, health behaviors, social economic factors, physical environment attributes, and healthcare access measures assessed across clusters.

	Phenogroup 1 (n = 924)	Phenogroup 2 (n = 451)	Phenogroup 3 (n = 951)	Phenogroup 4 (n = 350)
Demographic				
Female, %	50.1 (1.3)	50.2 (2.0)	49.8 (2.2)	50.6 (3.1)
Age 65 and older, %	20.0 (4.6)	15.0 (3.7)	19.8 (3.8)	18.1 (3.3)
Age less than 18 years, %	21.7 (3.1)	23.2 (3.9)	21.7 (2.7)	22.4 (3.0)
Black, %	2.8 (3.9)	9.9 (11.1)	6.6 (8.0)	39.7 (17.0)
American Indian & Alaska Native, %	1.1 (2.0)	2.3 (5.6)	1.9 (4.7)	2.5 (10.1)
Native Hawaiian/Other Pacific Islander, %	0.1 (0.1)	0.3 (1.0)	0.1 (0.2)	0.1 (0.1)
Hispanic, %	5.6 (4.9)	25.6 (21.5)	6.9 (9.1)	6.0 (9.3)
Non-Hispanic White, %	87.8 (7.3)	55.6 (18.9)	82.3 (13.1)	49.8 (15.0)
Rural, %	54.2 (27.6)	18.1 (17.0)	68.5 (23.9)	59.7 (27.2)
Population, n	71,299.7 (94,847.3)	447,188.1 (773,317.9)	41,303.2 (55,985.2)	53,670.4 (126,859.6)
Health Behaviors				
Smokers, %	15.6 (2.0)	15.3 (2.9)	19.4 (2.9)	20.9 (2.8)
Obesity, %	31.5 (4.5)	29.1 (5.0)	34.9 (4.3)	38.1 (4.9)
Diabetes, %	10.5 (2.7)	9.5 (2.6)	14.1 (3.6)	16.1 (4.0)
HIV Prevalence Rate[a]	99.7 (61.4)	269.3 (270.4)	141.0 (90.5)	396.0 (254.6)
Physically Inactive, %	24.6 (4.1)	22.6 (4.7)	30.7 (4.3)	32.8 (5.0)
Excessive Drinking, %	19.6 (2.7)	18.8 (2.7)	16.1 (2.4)	14.1 (2.3)
Vaccinated, %	46.6 (8.6)	44.3 (8.3)	40.6 (8.1)	39.1 (7.5)
Insufficient Sleep, %	30.7 (3.3)	33.2 (3.4)	34.9 (3.1)	38.0 (2.6)
Annual Mammogram, %	45.9 (6.2)	39.0 (7.0)	37.8 (6.0)	38.9 (6.1)
Social Economic Factors				
Median Household Income, $	60645.0 (12272.3)	61837.0 (18201.5)	46857.9 (7704.6)	39491.6 (5958.4)
Income Ratio	4.1 (0.4)	4.7 (0.8)	4.6 (0.6)	5.4 (0.8)
Unemployed, %	3.5 (1.0)	4.1 (1.6)	4.5 (1.3)	5.3 (1.4)
Segregation Index	51.9 (12.1)	47.9 (12.6)	45.2 (14.4)	33.5 (14.2)
Homeowners, %	74.4 (5.7)	62.1 (8.9)	73.9 (5.3)	66.4 (8.2)
Severe Housing Problem, %	12.2 (3.0)	18.6 (4.8)	12.9 (2.5)	16.5 (3.4)
Single-Parent Household, %	27.2 (6.9)	33.1 (8.4)	32.9 (6.2)	49.5 (9.5)
Drive Alone to Work, %	80.4 (4.6)	76.0 (9.5)	82.3 (4.0)	83.2 (4.6)
Limited Access to Healthy Foods, %	6.0 (4.3)	8.3 (5.9)	7.1 (5.7)	10.4 (7.9)
Physical Environment				
Social Association Rate[a]	13.4 (5.5)	9.1 (3.5)	11.1 (4.3)	11.1 (3.9)
Violent Crime Rate[a]	170.9 (106.3)	367.7 (198.2)	240.7 (139.8)	464.6 (256.5)
Fine Particulate Matter, $\mu g/m^3$	8.8 (2.0)	9.1 (2.0)	9.8 (1.4)	10.0 (1.0)
Healthcare Assess				
Uninsured, %	8.5 (3.7)	11.8 (5.7)	12.3 (4.9)	13.9 (3.6)
Preventable Hospitalization Rate	4143.4 (1351.4)	4122.5 (1224.6)	5571.8 (1812.7)	6224.4 (1699.0)
Child Mortality Rate[a]	50.8 (12.0)	51.9 (14.2)	68.3 (15.7)	87.6 (21.9)
Primary Care Physicians Rate[a]	63.3 (33.5)	77.0 (37.0)	39.8 (21.1)	44.5 (24.0)
Dentist Rate[a]	52.9 (22.7)	72.3 (45.1)	33.6 (16.8)	36.7 (23.3)
Mental Health Provider Rate[a]	165.4 (130.6)	293.6 (246.7)	116.1 (142.2)	130.1 (133.9)

Numbers displayed as mean (standard deviation); CV, cardiovascular; HF, heart failure.
[a] Per 100,000 persons

inactivity and smoking. A choropleth map of U.S. by counties grouped by phenogroup is provided in Fig. 2 and demonstrates regional clustering of observed phenogroups (Moran's I = 0.158, p-value<0.001). Phenogroups 1 and 2 represent geographically diverse groups including much of the Western and Northeastern United States. Conversely, phenogroup 3 spans much of the rural Southeast and Midwest and phenogroup 4 is most geographically centered in the urban Southeast and East coast.

Phenogroup 1 (N = 924; 24.5%) included counties with non-Hispanic white residents and highest median household income, percentage of homeownership, access to healthy foods, and greatest access to preventive health measures (Table 1). Phenogroup 2 counties (N = 451; 16.9%) were the most populous with nearly 25% Hispanic occupants, the second highest prevalence of Black residents, and the highest rates of severe housing problems. This phenogroup also had the best access to healthcare and the lowest rates of diabetes, obesity, smoking, and social isolation (Table 1). Phenogroup 3 counties (N = 951; 35.5%) were primarily non-Hispanic white and rural, with the worst access to healthcare, higher rates of physical inactivity, obesity and diabetes, and limited access to healthy food (Table 1). Phenogroup 4 (350; 13.1%) counties consisted of primarily Black residents with the highest rates of single-parent households, child mortality, income inequality, unemployment, HIV prevalence, and traditional CVD risk factors (smoking, obesity, diabetes, and physical inactivity) (Table 1). These counties also had the second lowest rate of healthcare access (by prevalence of primary care physicians) and the highest uninsured resident proportion and preventable hospitalization rate. Similar differences between phenogroups were observed using the limited county-level data from the 2010 CHR (Supplemental Table 3).

In stepwise linear regression analysis, the ten strongest independent predictors of CVD mortality included demographic and socioeconomic characteristics (race, household income, and age), behavioral risk factors (prevalence of smoking and physical inactivity), environmental factors (fine particulate matter) and characteristics of healthcare access and systems (percent with annual mammogram, child mortality rate, and preventable hospitalization rates) (Supplemental Table 4). The age-adjusted mortality rates across phenogroups over time are shown in Fig. 3 and Table 2. Phenogroup 1 counties had the lowest age-adjusted rates of CVD-related mortality with a stepwise increase in average (least square means) age-adjusted mortality rates observed from phenogroup 1 (223 deaths per 100,000 population) to phenogroup 4 (317 deaths) ($p < 0.001$ for trend). A similar trend was also seen in age-adjusted overall mortality with phenogroup 1 having the lowest risk (797 deaths per 100,000 population) and phenogroup 4 having the highest (868 deaths) ($p < 0.001$ for trend). Conversely, while phenogroup 4 still had the highest overall age-adjusted IHD and HF

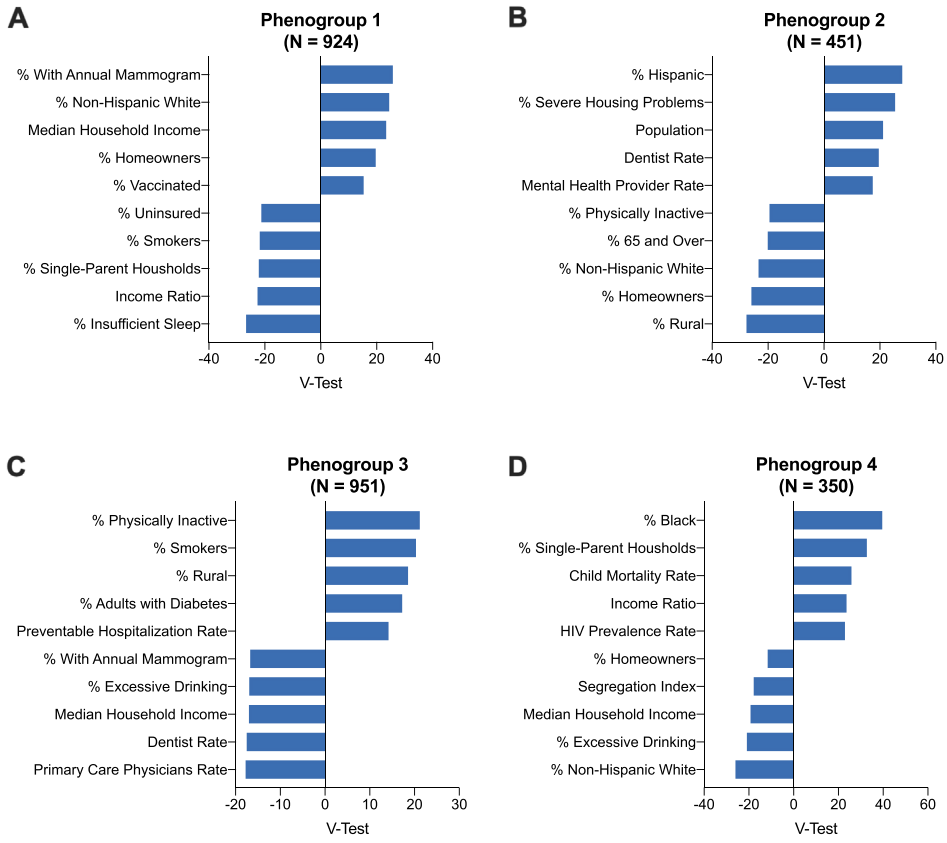

Fig. 1. The top 5 most over and underrepresented variables in each phenogroup as determined by the V-test. A positive value indicates the variable overrepresents the phenogroup while a negative number indicates the variable underrepresents the phenogroup.

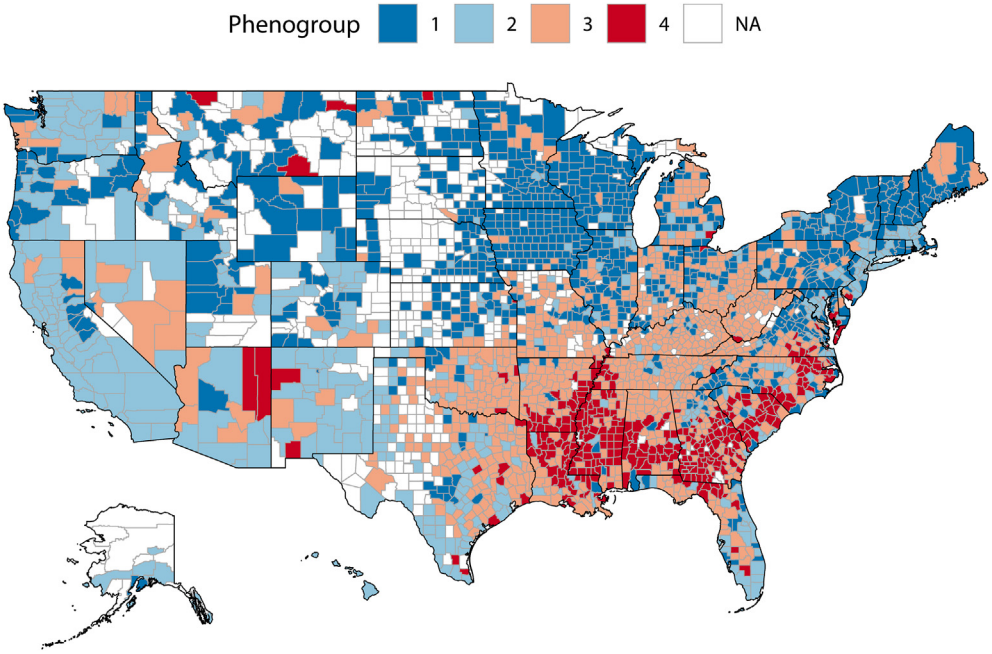

Fig. 2. Choropleth map of United States counties by derived phenogroups.

mortality, phenogroup 2 counties had the lowest overall age-adjusted IHD and HF mortality (102 and 20 deaths per 100,000 population, respectively).

The R^2 for predicting CVD mortality with phenogroups alone was 0.48. The addition of phenogroup membership to the model with the top 10 optimal predictors significantly improved the R^2 for predicting CVD mortality from 0.69 to 0.71 ($p < 0.001$).

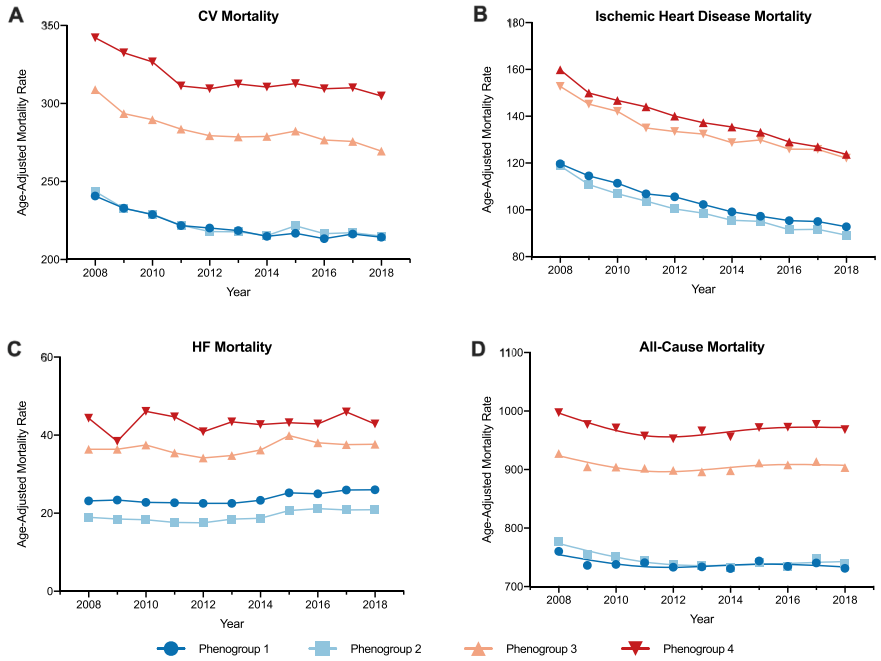

Fig. 3. Overall trends in A) cardiovascular (CV), B) ischemic heart disease, C) heart failure (HF), and D) all-cause mortality from 2008 to 2018 by phenogroups.

Table 2
Least square means age-adjusted mortality rates per 100,000 population between 2008 and 2018 by phenogroup. Differences between phenogroups were determined by linear mixed effect models after adjusting for percent smoking, percent physically inactive, percent Black race, percent with annual mammogram, fine particulate matter, median household income, preventable hospitalization rate, percent non-Hispanic white race, percent aged less than 18 years, and child mortality rate.[a]

	Phenogroup 1 (95% CI)	Phenogroup 2 (95% CI)	Phenogroup 3 (95% CI)	Phenogroup 4 (95% CI)	P-value for difference (ref group: Phenogroup 4)		
					Phenogroup 1	Phenogroup 2	Phenogroup 3
CVD Mortality	222.9 (220.2–225.5)	223.8 (220.0–227.7)	284.0 (281.3–286.6)	317.0 (312.7–321.3)	<0.001	<0.001	0.10
IHD Mortality	106.3 (104.0–108.5)	102.3 (99.1–105.4)	141.1 (126.2–138.9)	136.4 (132.7–140.2)	0.048	<0.001	0.35
HF Mortality	26.2 (24.3–28.0)	20.4 (18.5–22.4)	41.7 (39.7–43.7)	50.7 (47.7–53.8)	0.005	<0.001	0.21
All-Cause Mortality	797.2 (789.4–804.9)	808.0 (801.8–814.2)	852.2 (846.7–857.7)	868.1 (855.5–880.6)	<0.001	<0.001	0.02

[a] CI, confidence interval; CVD, cardiovascular disease; HF, heart failure; IHD, ischemic heart disease

3. Discussion

Our study demonstrated the feasibility of an unsupervised ML clustering strategy in identifying unique county phenogroups with differential CVD, IHD, HF, and all-cause mortality risk. We identified four phenotypically distinct county phenogroups and observed increasing rates of CVD outcomes and mortality over time from the predominantly white, suburban communities in phenogroup 1 to the majority Black, low-income, urban communities of phenogroup 4. While our ML algorithm evaluated the contributions of socio-demographic, behavioral and medical risk factors, phenogroup membership was in large part determined by social factors, including housing stability, household income, access to health care and self-identified race. Phenogroup membership was, moreover, associated with CVD outcomes independent of other county-level predictors, demonstrating the robustness of the observed association.

Despite declining national mortality rates from CVD, significant and persistent regional disparities in CVD outcomes driven by disparities in IHD were demonstrated in the first comprehensive county-level assessment of CVD in the U.S. by Roth et. al [9]. Furthermore, while data from the Behavioral Risk Factor Surveillance Survey [20] previously showed marked disparities in CVD risk factors by region, the contributions of heterogeneity in county-level characteristics toward the risk of mortality and methods for accounting for these differences in developing cohesive health policy interventions are not well understood. Our study represents an important step forward by identifying county phenogroups based on socio-demographic characteristics and demonstrating differences in CVD mortality across phenogroups. The application of unsupervised ML in the study of health disparities has been described previously [21], though no prior study to our knowledge has employed these methods for evaluating CVD disparities in a large, heterogeneous county-level population. Furthermore, inclusion of demographic and social determinants data at the county level—the smallest unit for which aggregate mortality data are available—into a phenogroup analysis provides unique insight into interactions among multiple related CVD mortality risk factors, which can inform the development of health interventions at county and regional levels [9,22]. By combining data on county-level risk factors and CVD outcomes, our study builds on prior observations to further characterize communities experiencing disproportionate CVD burden.

Our autonomous ML algorithm identified four mutually exclusive county phenogroups, defined by differences in racial composition, social determinants, access to healthcare, comorbid disease, and behavioral risk factors. By evaluating the coalescence of markers of socioeconomic distress including housing instability and household income, demographic factors such as race, social environment including rural versus urban differences, and access to healthcare at the county level, we are

able to move beyond prior work that has established racial disparities in order to better understand the effect of structural factors in driving cardiovascular outcomes within communities. On the extremes of the spectrum, phenogroups 1 and 4, demonstrated the lowest and highest rates of CVD, respectively. Focusing on predominantly social markers, this is largely expected. Phenogroup 1 represents high-income, predominantly White counties, with high access to healthcare, while phenogroup 4 offers the opposite extreme: predominantly Black, urban counties, with low household income, high rates of housing instability and low access to care. However, the performance of intermediate phenogroups is less predictable. Phenogroup 2, which represents counties with the second lowest rates of CVD, demonstrates a number of high-risk features, including a large Hispanic population, and high rates of housing instability, but simultaneously exhibits high indices of healthcare access, and low levels of CVD risk factors including diabetes and smoking. This is likely due to phenogroup 2 counties hailing predominantly from regions with large, racially diverse, urban centers that are historically White with a growing Hispanic minority. Thus, though these communities experience housing instability due to increasing gentrification, their urban status provides a higher density of healthcare resources. This is consistent with our observations such that while phenogroup 2 had the highest proportion of Hispanic residents compared to other phenogroups, only 26% of residents in phenogroup 2 were Hispanic compared to 56% of Non-Hispanic White race. "Rural" identification was additionally the strongest predictor of phenogroup non-membership, supporting that phenogroup 2 largely represents urban communities. Similarly, phenogroup 3 communities exhibit a mix of risk features: though predominantly White without housing insecurity issues, these counties demonstrated high rates of traditional CVD risk factors, and low access to healthcare, likely a result of their largely rural locations.

The significant role of race in determining cluster membership highlights the role of race and racism in driving county-level health disparities. Compared with non-Hispanic white adults, Black adults have poor access to health care, higher burden of comorbid conditions, and experience 30% higher risk of CV death [8,23–25]. In our study, self-identified race was a driver of group membership in three of four phenogroups, with high percentages of Black residents defining counties in phenogroup 4 that exhibited the highest CVD mortality rates. The high prevalence of social risk factors evident among counties with higher proportion of Black and Hispanic residents further highlights the role of structural racism—manifested in unstable housing, low income, and inadequate access to care—in contributing to adverse health outcomes. Though traditionally challenging to study, recent evaluations of CVD risk factor prevalence in communities with high foreclosure risk and low home-ownership find consistent results regarding the role of housing instability in driving higher CVD risk as well [26]. Furthermore, the confluence of elevated cardiovascular risk with childhood mortality risk and HIV infection risk in counties with a higher proportion of Black residents demonstrates the impact of social risk on multiple downstream health outcomes. This points to the need for targeted multilayered interventions, policies, and programs to produce health equity. By phenotyping counties based on multiple drivers of CVD risk, our study goes beyond prior work in identifying related and overlapping modifiable social risk factors for adverse CVD outcomes. Future studies are needed to determine the impact of local policies and interventions aimed at housing stability, housing insecurity, residential segregation, and social infrastructure-building on long-term cardiovascular outcomes [27].

Our study can further the work of reducing CVD disparities by informing the development of efficient community health interventions targeted at specific county phenogroups. We demonstrated the value and feasibility of a ML strategy in order to better capture drivers of CVD risk at the regional level. Despite the development in recent years of evidence-based community health interventions, matching appropriate interventions to communities in need has remained a challenge and has limited efforts to scale-up such programs. Our phenogroups identified counties where funding for large-scale public health interventions aimed at reducing physical and financial barriers to preventive care and/or healthier lifestyles may have the largest impacts [28,29].

3.1. Limitations

This study is not without limitations. First, because we conducted a cross-sectional analysis, causal relationships between county characteristics and mortality cannot be drawn. Counties are large sociopolitical (i.e., not tangible/ecological) boundaries that obscure more granular variation in outcomes/covariates (confounding/measurement bias) and influence the health policy that drives outcomes. Analysis in our study is additionally limited to characteristics captured by the CHR, though data presented represents over 97% of U.S. counties. Second, outcomes data was based on ICD codes and not using adjudicated mortality events. Third, this study included county-level data obtained from 2016 to 2018 and outcomes data recorded from 2008 to 2018. While the cardiovascular outcomes occurring in 2008 may not be reflective in the current county characteristics, our analysis showed that differences in county-level mortality rates have been relatively stable between phenogroups. Furthermore, the pattern of differences in county characteristics across the phenogroups in 2010 data (the earliest available CHR) was consistent with that observed in the primary analysis (2016–2018) highlighting the temporal stability of our phenomapping approach. Finally, this is a geographic analysis and our data was limited to the county level. Future studies should examine smaller units of analysis – such as zip codes or census tracts – in order to present a more granular area analysis which may compliment the values we observe at the county level.

4. Conclusions

Unsupervised clustering of demographic and behavior data can identify unique phenogroups with differential risk of CV-related and overall mortality and may aid in designing tailored preventive interventions for similar communities at highest risk for CVD-related adverse events.

5. Funding and disclosures

Dr. Navar has received personal fees from Amgen, AstraZeneca, Janssen Pharmaceuticals, Esperion Therapeutics, Amarin, Sanofi, Regeneron, Novo Nordisk, Novartis, The Medicines Company, New Amsterdam, and Pfizer.

Dr. Hughes is supported by the Texas Health Resources Clinical Scholars Program.

Dr. Pandey is supported by the Texas Health Resources Clinical Scholars Program and has served on the advisory board of Roche Diagnostics.

All other authors report no funding or disclosures.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: **Dr. Navar** has received personal fees from Amgen, AstraZeneca, Janssen Pharmaceuticals, Esperion Therapeutics, Amarin, Sanofi, Regeneron, Novo Nordisk, Novartis, The Medicines Company, New Amsterdam, and Pfizer. **Dr. Hughes** is supported by the Texas Health Resources Clinical Scholars Program. **Dr. Pandey** is supported by the Texas Health Resources Clinical Scholars Program and has served on the advisory board of Roche Diagnostics.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ajpc.2020.100118.

References

[1] Ford ES, Ajani UA, Croft JB, et al. Explaining the decrease in U.S. deaths from coronary disease, 1980-2000. N. Engl. J. Med. 2007;356(23):2388–98.

[2] Global Burden of Cardiovascular Diseases C, Roth GA, Johnson CO, et al. The burden of cardiovascular diseases among US States, 1990-2016. JAMA Cardiol 2018;3(5):375–89.

[3] Roth GA, Johnson C, Abajobir A, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. J. Am. Coll. Cardiol. 2017; 70(1):1–25.

[4] Wallace M, Sharfstein JM, Kaminsky J, Lessler J. Comparison of US county-level public health performance Rankings with county cluster and national Rankings: assessment based on prevalence rates of smoking and obesity and motor vehicle crash death rates. JAMA Netw Open 2019;2(1):e186816.

[5] Rosamond WD. Geographic variation in cardiovascular disease burden: clues and questions. JAMA Cardiol 2018;3(5):366–8.

[6] Fang J, Yang Q, Hong Y, Loustalot F. Status of cardiovascular health among adult Americans in the 50 States and the District of Columbia, 2009. J. Am. Heart Assoc. 2012;1(6):e005371.

[7] Mensah GA, Mokdad AH, Ford ES, Greenlund KJ, Croft JB. State of disparities in cardiovascular health in the United States. Circulation 2005;111(10):1233–41.

[8] Virani SS, Alonso A, Benjamin EJ, et al. Heart disease and stroke statistics-2020 update: a report from the American heart association. Circulation 2020;141(9): e139–596.

[9] Roth GA, Dwyer-Lindgren L, Bertozzi-Villa A, et al. Trends and patterns of geographic variation in cardiovascular mortality among US counties. J. Am. Med. Assoc. 2017;317(19):1976–92.

[10] Anderson TJ, Saman DM, Lipsky MS, Lutfiyya MN. A cross-sectional study on health differences between rural and non-rural U.S. counties using the County Health Rankings. BMC Health Serv. Res. 2015;15:441.

[11] Litaker D, Love TE. Health care resource allocation and individuals' health care needs: examining the degree of fit. Health Pol. 2005;73(2):183–93.

[12] Segar MW, Patel KV, Ayers C, et al. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. Eur. J. Heart Fail. 2020;22(1):148–58. https://doi.org/10.1002/ejhf.1621.

[13] University of Wisconsin Population Health Institute. County health Rankings 2020. https://www.countyhealthrankings.org. Accessed 4/5/2020.

[14] Stekhoven DJ, Buhlmann P. MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–8.

[15] Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis 2008; 25(1):18.

[16] Moran PAP. Notes on continuous stochastic phenomena. Biometrika 1950;37(1/2): 17–23.

[17] Davidson R, MacKinnon J. Several tests for model specification in the presence of alternative hypotheses. Econometrica 1981;49(3):781–93.

[18] Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 2019;35(3):526–8.

[19] Zeileis A, Hothorn T. Diagnostic checking in regression relationships. R. News 2002; 2(3):7–10.

[20] Pilkerton CS, Singh SS, Bias TK, Frisbee SJ. Changes in cardiovascular health in the United States, 2003-2011. J. Am. Heart Assoc. 2015;4(9):e001650.

[21] Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat. Med. 2020;26(1):16–7.

[22] Ezzati M, Friedman AB, Kulkarni SC, Murray CJ. The reversal of fortunes: trends in county mortality and cross-county mortality disparities in the United States. PLoS Med. 2008;5(4):e66.

[23] Alexander M, Grumbach K, Selby J, Brown AF, Washington E. Hospitalization for congestive heart failure. Explaining racial differences. J. Am. Med. Assoc. 1995; 274(13):1037–42.

[24] Lutfiyya MN, McCullough JE, Haller IV, Waring SC, Bianco JA, Lipsky MS. Rurality as a root or fundamental social determinant of health. Dis. Mon. 2012;58(11): 620–8.

[25] Joynt KE, Harris Y, Orav EJ, Jha AK. Quality of care and patient outcomes in critical access rural hospitals. J. Am. Med. Assoc. 2011;306(1):45–52.

[26] Chambers EC, Hanna DB, Hua S, et al. Relationship between area mortgage foreclosures, homeownership, and cardiovascular disease risk factors: the Hispanic Community Health Study/Study of Latinos. BMC Publ. Health 2019;19(1):77.

[27] Sims M, Kershaw KN, Breathett K, et al. Importance of housing and cardiovascular health and well-being: a scientific statement from the American heart association. Circ Cardiovasc Qual Outcomes 2020;13(8):e000089.

[28] Mensah GA, Cooper RS, Siega-Riz AM, et al. Reducing cardiovascular disparities through community-engaged implementation research: a national heart, lung, and blood institute workshop report. Circ. Res. 2018;122(2):213–30.

[29] Record NB, Onion DK, Prior RE, et al. Community-wide cardiovascular disease prevention programs and health outcomes in a rural county, 1970-2010. J. Am. Med. Assoc. 2015;313(2):147–55.

[30] Friede A, Reid JA, Ory HW, CDC WONDER. a comprehensive on-line public health information system of the Centers for Disease Control and Prevention. Am J Public Health 1993;83(9):1289–94. https://doi.org/10.2105/ajph.83.9.1289.

[31] Jacqueline T Vuong, Sophia A Jacob, Kevin M Alexander, Avinainder Singh, Ronglih Liao, Akshay S, Desai, Sharmila Dorbala, Mortality from heart failure and dementia in the United States: CDC WONDER 1999–2016, J Card Fail, 25 (2), 2019, 125-129.