3-7-2023

Social determinants, cardiovascular disease, and health care cost: A nationwide study in the United States using machine learning

Feinuo Sun

Jie Yao

Shichao Du

Feng Qian

Allison A Appleton

See next page for additional authors

Authors

Feinuo Sun, Jie Yao, Shichao Du, Feng Qian, Allison A Appleton, Cui Tao, Hua Xu, Lei Liu, Qi Dai, Brian T Joyce, Drew R Nannini, Lifang Hou, and Kai Zhang

Journal of the American Heart Association

ORIGINAL RESEARCH

Social Determinants, Cardiovascular Disease, and Health Care Cost: A Nationwide Study in the United States Using Machine Learning

Feinuo Sun, PhD; Jie Yao, MS; Shichao Du, BA; Feng Qian, PhD; Allison A. Appleton, ScD; Cui Tao, PhD; Hua Xu, PhD; Lei Liu, PhD; Qi Dai, PhD, MD; Brian T. Joyce, PhD; Drew R. Nannini, PhD; Lifang Hou, PhD, MD; Kai Zhang, PhD

BACKGROUND: Existing studies on cardiovascular diseases (CVDs) often focus on individual-level behavioral risk factors, but research examining social determinants is limited. This study applies a novel machine learning approach to identify the key predictors of county-level care costs and prevalence of CVDs (including atrial fibrillation, acute myocardial infarction, congestive heart failure, and ischemic heart disease).

METHODS AND RESULTS: We applied the extreme gradient boosting machine learning approach to a total of 3137 counties. Data are from the *Interactive Atlas of Heart Disease and Stroke* and a variety of national data sets. We found that although demographic composition (eg, percentages of Black people and older adults) and risk factors (eg, smoking and physical inactivity) are among the most important predictors for inpatient care costs and CVD prevalence, contextual factors such as social vulnerability and racial and ethnic segregation are particularly important for the total and outpatient care costs. Poverty and income inequality are the major contributors to the total care costs for counties that are in nonmetro areas or have high segregation or social vulnerability levels. Racial and ethnic segregation is particularly important in shaping the total care costs for counties with low poverty rates or social vulnerability level. Demographic composition, education, and social vulnerability are consistently important across different scenarios.

CONCLUSIONS: The findings highlight the differences in predictors for different types of CVD cost outcomes and the importance of social determinants. Interventions directed toward areas that have been economically and socially marginalized may aid in reducing the impact of CVDs.

Key Words: cardiovascular disease ■ health care costs ■ machine learning ■ racial and ethnic segregation ■ social determinants of health

Cardiovascular disease (CVD) is the leading cause of death in the United States. It accounted for ≈875 000 deaths in 2019, and the average annual estimated direct and indirect economic cost was $378.0 billion in 2017 to 2018.[1] Cardiovascular outcomes vary by geographical location across the United States. For example, counties with high CVD mortality rates cluster in southeastern Oklahoma along the Mississippi River Valley to eastern Kentucky, whereas counties with low CVD mortality rates are found in the Southwest, Northeast, and southern Florida.[2] Understanding the determinants of population-based

Correspondence to: Kai Zhang, PhD, Department of Environmental Health Sciences, School of Public Health, University at Albany, State University of New York, Albany, NY 12144. Email: kzhang9@albany.edu
Lifang Hou, PhD, MD, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611. Email: l-hou@northwestern.edu

For Sources of Funding and Disclosures, see page 15.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- This study is the first that examines the role of a wide spectrum of social determinants in relation to health care costs for cardiovascular diseases in the United States using a machine learning approach.
- The findings highlight the importance of contextual factors including social vulnerability, poverty and income inequality, and racial and ethnic segregation in shaping the health care costs for cardiovascular diseases.

What Are the Clinical Implications?

- Identifying major social determinants can help prioritize and allocate resources, design, and implement prevention and interventions to reduce the health care costs for cardiovascular diseases.

Nonstandard Abbreviations and Acronyms

SVI	Social Vulnerability Index
XGBoost	extreme gradient boost

CVD outcomes, such as the prevalence of CVD and its related health care costs, is crucial to addressing this enormous public health problem.

Previous research has documented the determinants of CVD risk on human, behavioral, psychological, and social factors.[3] Older adults and racial and ethnic minority groups, especially Black people, are particularly suffering from a higher prevalence of CVD.[4,5] Lifestyle factors, such as tobacco smoking, consumption of alcohol and high cholesterol foods, and poor physical activity, exhibit greater risk of CVD.[3,6–8] Stress, anger, anxiety, and depression can increase CVD risk.[3,7,8] According to the commonly used social determinants of health framework, socioeconomic status (SES) and contextual characteristics, including health care system, neighborhood environments, and community and social contexts, also shape population health outcomes.[9] Thus, social determinants of CVD include SES and aspects of the social environments. Education, income, and employment status are highly related to risk factors for CVD and cardiovascular outcomes, with less education, income, and unemployment associated with CVD and CVD mortality.[6,10,11] Environmental factors (ie, neighborhood or community characteristics) also play important roles in determining individuals' CVD risk. In general, better economic conditions and residential environments, higher availability and accessibility of health care services, and more social support and networks can improve cardiovascular health and thus lower risk of CVD.[3,6,12,13] Racial and ethnic residential segregation, as a neighborhood characteristic, has been found to be associated with poverty and unemployment rate, poor housing conditions, and limited health care services, leading to increased CVD risk.[14]

Although these previous studies identified factors associated with CVD, 2 limitations need to be addressed. First, most studies evaluate CVD risk but rarely examine the importance of social determinants, such as racial and ethnic segregation, in shaping the CVD-related care costs nationwide. Nevertheless, the health care costs for CVD could vary substantially across geographical units because of the variations in the demand and supply of health care services. On the demand side, the different prevalence of CVD (which is determined by demographic composition, SES, risk behaviors, and relative health conditions) leads to different health care costs for CVD. The use of health care services also varies across different population groups, which is associated with health care costs. For example, racial and ethnic minority groups, uninsured, and unemployed individuals have limited access to medical care because of barriers such as language, perceived racial biases, and low economic capacity.[6,10] On the supply side, health care costs are largely determined by the availability of health care resources, such as the number of hospitals and health care providers. Neighborhood characteristics (eg, residential segregation) could affect health care costs bidirectionally. For example, not only highly segregated communities have limited availability of health care providers, but racial and ethnic minority residents in those areas also have lower use of health care services compared with White residents.[15–17]

The second limitation is the methodological perspective. Few studies leverage artificial intelligence approaches, like machine learning, to examine the determinants of CVD health care costs at the county level. Research has indicated that if properly designed and applied, artificial intelligence could enhance health equity in heart failure care by improving diagnosis and identifying patient groups at risk for racially diverse populations.[18] In recent years, machine learning tools have been increasingly adopted in clinical research, given their superior performance over traditional statistical tools.[19,20] Machine learning has shown significant promise over regression approaches in predicting CVD risk, incidence, and outcomes because of its improved flexibility and fewer assumptions.[21] Several studies[22–24] have used machine learning to identify key predictors of heart disease and stroke based on neighborhood-level data and highlight the older population structure and risk factors like obesity and inactivity in shaping cardiovascular health. This approach has also been applied to understand the determinants of health care costs, such as end-of-life care costs and costs for breast cancer.[25–28]

Filling these knowledge gaps, our article is among the first to examine the determinants of county-level CVD health care costs in the United States using the extreme gradient boost (XGBoost) machine learning approach. We focus on the county level because counties have departments of health that are equipped with public health personnel and have the capacity to report and monitor real-world cardiovascular prevalence and costs. County-level analysis is also widely used in ecological health research. We aim to examine the relative importance of a wide spectrum of determinants of the total, inpatient, and outpatient CVD health care costs as well as CVD prevalence based on the total sample and a series of stratified samples. The findings have important policy implications for controlling the health care costs from CVD.

METHODS

Data

Data used in this study are publicly available and described in this section. The complete data set for analysis is available upon request. Data sources include the *Interactive Atlas of Heart Disease and Stroke* provided by the Centers for Disease Control and Prevention,[29] a geographic data set assembled from a variety of national data sources such as the Centers for Medicare and Medicaid Services Chronic Conditions Data Warehouse, the Deaths National Vital Statistics System, the Health Resources and Services Administration Area Health Resources Files, and the US Census Bureau. The most recent data are from the year 2017. We also used the American Community Survey 2015 to 2019, 5-year estimates,[30] PolicyMap,[31] and Centers for Disease Control and Prevention Social Vulnerability Index (SVI)[32] as supplementary sources for variables that are not available in the *Interactive Atlas*. The total sample contains 3137 US counties with complete health care data available. This study is exempt from institutional review board review because we use publicly available county-level data.

The dependent variables include the prevalence of diagnosed CVD among Medicare beneficiaries and 3 different costs (ie, total, inpatient, and outpatient costs) of care per capita for those beneficiaries diagnosed with CVD in 2017, which are the average costs incurred by Medicare beneficiaries with diagnosed CVD in a given county. Inpatient and outpatient costs are differentiated by overnight stay in a hospital. Total care costs include inpatient, outpatient, post–acute care (eg, skilled nursing facility care and home health), hospice, physician, testing, and imaging costs. Data are from the Centers for Medicare and Medicaid Services Chronic Conditions Data Warehouse.[29] Here the beneficiaries are limited to Medicare fee-for-service beneficiaries with both Medicare Part A and Part B coverage and who were aged ≥65 years. Based on the *International Classification of Diseases, Tenth Revision, Clinical Modification* (*ICD-10-CM*),[33,34] CVD is defined including atrial fibrillation (*ICD-10-CM* I48.0, I48.2, I48.91), acute myocardial infarction (*ICD-10-CM* I21.0–I21.4, I22), congestive heart failure (*ICD-10-CM* I09.81, I11.0, I13.0, I13.2, I50.1–I50.4, I50.9), and ischemic heart disease (*ICD-10-CM* I20, I21.0–I21.4, I22, I24, I25.1–I25.2, I25.42, I25.5–I25.9). Medicare payments are standardized to address geographic differences in payment rates for individual services, and the cost data are age-standardized based on the 2000 US Standard Population.

We consider 21 independent variables with regard to demographic composition, risk factors, as well as SES and contextual factors from major domains of the social determinants of health framework that have previously been associated with CVD.[3–6,14] Demographic composition refers to the percentage of individuals aged ≥65 years, men, non-Hispanic Black people, non-Hispanic Asian people, Hispanic people, and individuals aged ≥15 years who are married. Risk factors include the percentage of current smokers and the prevalence of high cholesterol among individuals aged ≥18 years, as well as the prevalence of diagnosed diabetes, obesity, and leisure-time physical inactivity among individuals aged ≥20 years.

Factors from the social determinants of health framework include education, economic conditions, racial and ethnic segregation, SVI, and rural–urban status. Education is measured by the percentage of individuals aged ≥25 years who did not complete a high school diploma. Economic conditions refer to poverty rate and income inequality, measured by the percentage of individuals living in poverty and the Gini index at the county level, or the degree of inequality in income distribution in a region that ranges from 0 to 1, with higher values indicating more inequality.[35] The level of racial and ethnic segregation in a county is indicated by the Theil's H index from the Census and PolicyMap,[31] which measures the extent to which racial and ethnic groups are evenly distributed in a county as compared with a larger area. This index ranges from 0 to 1, with higher values meaning being more segregated. Data for SVI are provided by the Centers for Disease Control and Prevention.[32] SVI ranks counties based on 4 themes, SES, household composition and disability, racial and ethnic minority status and language, and housing type and transportation, and also ranges from 0 to 1 with higher values meaning more social vulnerability. Rural–urban status is measured by the

National Center for Health Statistics rural–urban classification scheme.[36] Counties are divided into large central metro, large fringe metro, medium/small metro, and nonmetro counties. Large central metro is defined as counties in metropolitan statistical areas with a population of ≥1 million that contain the entire population or have the entire population contained in the largest principal city, or contain at least 250 000 inhabitants of any principal city. Large fringe metro refers to counties in metropolitan statistical areas with a population of ≥1 million but are not considered as large central metro, whereas medium/small metro refers to other counties in metropolitan statistical areas. All other counties that are not in metropolitan statistical areas are nonmetro. The Table in the Results section shows the data sources for each variable.

Statistical Analysis

We applied the XGBoost machine learning approach to determine each determinant's importance in predicting the 4 outcomes. The approach has been used in studies about cardiovascular health outcomes and health care costs for multiple diseases,[22–26] but rarely in CVD health care costs. The machine learning approach has several advantages over traditional statistical methods. Compared with traditional statistical modeling, it has fewer assumptions and higher flexibility and thus is able to take into account the complex nonlinear relationships and interactions between a large number of determinants and the outcome.[21,37,38] Supervised machine learning is able to perform predictions from observations of a set of features that are linked to the target outcome based on available data.[21,37] XGBoost

Table. Summary Statistics and Data Sources for Outcome Variables and Independent Variables of 3137 US Counties

Variables	Mean	SD	Minimum	Maximum	Data sources
Outcome variables					
Total care costs per capita, $	19 672.06	2786.31	8212.00	36 976.00	CMS, 2017
Inpatient care costs per capita, $	6386.98	1123.41	1694.00	21 005.00	CMS, 2017
Outpatient care costs per capita, $	4372.06	1556.68	1253.00	15 047.00	CMS, 2017
CVD prevalence, %	35.78	5.54	18.00	55.20	CMS, 2017
Independent variables					
Age 65+ y, %	18.79	4.65	3.20	56.70	ACS, 2015–2019
Men, %	50.09	2.35	42.81	72.72	ACS, 2015–2019
Non-Hispanic Black people, %	8.91	14.49	0	87.20	ACS, 2015–2019
Non-Hispanic Asian people, %	1.38	2.82	0	42.60	ACS, 2015–2019
Hispanic people, %	9.21	13.63	0	99.10	ACS, 2015–2019
Married, %	53.16	6.63	20.26	82.48	ACS, 2015–2019
Current smoker, %	20.13	4.02	6.90	45.70	PLACES, 2018
High cholesterol prevalence, %	37.13	3.47	21.20	46.00	PLACES, 2017
Diabetes prevalence, %	10.49	3.53	2.20	28.70	CDC, 2017
Obesity prevalence, %	33.45	5.90	11.00	58.90	CDC, 2017
Physical inactivity prevalence, %	25.36	5.78	8.80	49.80	CDC, 2017
Less than high school, %	13.03	6.17	1.10	46.70	ACS, 2015–2019
Poverty rate	14.46	5.80	2.70	47.70	Census Bureau, 2018
Gini index	0.45	0.04	0.30	0.71	ACS, 2015–2019
Racial and ethnic segregation index	0.40	0.11	0.11	0.82	Census and PolicyMap, 2010
Social Vulnerability Index	0.50	0.29	0	1	CDC, 2018
Large central metro, %	2.17				NCHS, 2013
Large fringe metro, %	11.74				
Medium/small metro, %	23.25				
Nonmetro, %	62.84				

ACS indicates American Community Survey; CDC, Centers for Disease Control and Prevention; CMS, Centers for Medicare and Medicaid Services; CVD, cardiovascular disease; and NCHS, National Center for Health Statistics.

is a tree-based ensemble machine learning method, an improved algorithm over gradient boosting.[38,39] Unlike common gradient boosting that uses mean square error to split decision trees, XGBoost applies similarity score and gain to determine the nodes of decision trees. Regularization is another advantage of XGBoost that can prevent overfitting in the single regression tree model. With the application of these optimization techniques, XGBoost can achieve faster training speed and better predictive performance than regular tree-based models.[22,39,40]

XGBoost is used to model the complex relationship between the determinants and outcomes and obtain the rank order of importance of input features (ie, influential variables).[38,39] The model building in our analysis was performed in R 4.1.0. The package used was xgboost (version 1.6.0.1).[41] In the XGBoost model, we chose 1000 as the maximum number of iterations. Step size shrinkage was set to 0.01 to avoid overfitting. A depth of 3 was selected for the tree. Hyperparameters were defined in the model by checking the training error and test error of the model.

Through the XGBoost approach, we first obtained the ranking of all 21 independent variables on the 4 different dependent variables, respectively. Next, we stratified all counties into different groups according to their contextual characteristics including rural–urban status, poverty rate, racial and ethnic segregation index, and SVI. Then, we ran XGBoost for each stratified data set separately and compared the results of the ranking of determinants on total care costs obtained by the XGBoost approach. We compared the groups of large central metro counties, large fringe metro counties, medium/small metro counties, and nonmetro counties as well as different groups of counties according to poverty rate, racial and ethnic segregation, and SVI. Counties a with high poverty rate, high segregation level, or high SVI are compared with counties that have low values on these variables. For example, counties with a low poverty rate are those where poverty rates are under the lower quartile (25%), whereas counties with a high poverty rate are those where poverty rates are above the upper quartile (75%).

RESULTS

The Table shows the summary statistics of all variables for all counties as well as their data sources. The average prevalence of CVD in all counties is 35.78%. The average total care cost per capita of CVD is $19 672.06, whereas the average inpatient and outpatient care costs are $6386.98 and $4372.06, respectively. The ranges for 3 types of costs are all large, with total costs ranging from $8212 to $36 976, inpatient costs from $1694 to $21 005, and outpatient costs from $1253 to $15 047. For contextual factors, the average poverty rate is 14.46%, and the average values of the Gini index, racial and ethnic segregation, and SVI are 0.45, 0.40, and 0.50, respectively. The breakdown for large central metro, large fringe metro, medium/small metro, and nonmetro counties are 2.17%, 11.74%, 23.25%, and 62.84%, respectively.

Figure 1 shows the spatial distributions of the 4 outcomes. Counties with the lowest total care costs for CVD are in Idaho, Colorado, New Mexico, Northern California, and Oregon, whereas counties with the highest total care costs are clustered in Texas and Southern California. Inpatient care costs for CVD show great spatial heterogeneity. Some counties in southern Nevada and southern California have the highest inpatient costs, whereas Idaho, Utah, and Iowa have the lowest. For outpatient costs, counties in the Northwest and Midwest including Washington, Idaho, Montana, North Dakota, South Dakota, Nebraska, Kansas, Minnesota, Iowa, and Maine have the highest costs, whereas counties in southern California and the South have the lowest. In contrast, counties in the South suffer from the highest CVD prevalence.

Figure 2 shows the XGBoost ranking results of the models for the 4 outcomes based on the entire sample. SVI, the percentage of people with less than a high school education, poverty rate, the percentage of older adults (ie, aged ≥65 years), and physical inactivity are the top 5 ranked predictors of total care costs. For inpatient costs, the percentages of married individuals, current smokers, non-Hispanic Black people, older adults, and obesity are the top 5 predictors. For outpatient costs, the percentage of non-Hispanic Black people, the nonmetro indicator, racial and ethnic segregation index, the percentages of people with less than a high school education, and Hispanic people are the top 5 predictors. For the prevalence of CVD, the percentage of current smokers, physical inactivity, the percentage of non-Hispanic Black people, high cholesterol, and the percentage of those with less than a high school education are the top 5 predictors. Additional analysis shows that the results do not change if we use other tree-based models, such as CatBoost models.

Examination of the independent variables associated with the 4 outcomes exhibited some consistent associations. Demographically, the percentage of non-Hispanic Black people ranks highly for 3 outcomes (third for inpatient costs, first for outpatient costs, and third for prevalence of CVD). For total care costs, it ranks sixth. The percentage of Hispanic people ranks in fifth place for outpatient costs but does not rank highly for the other 3 outcomes. The percentage of older adults ranks fourth for both total and inpatient costs. The percentage of married individuals is particularly important (ranking in first place) for inpatient costs. Lifestyle factors are consistently

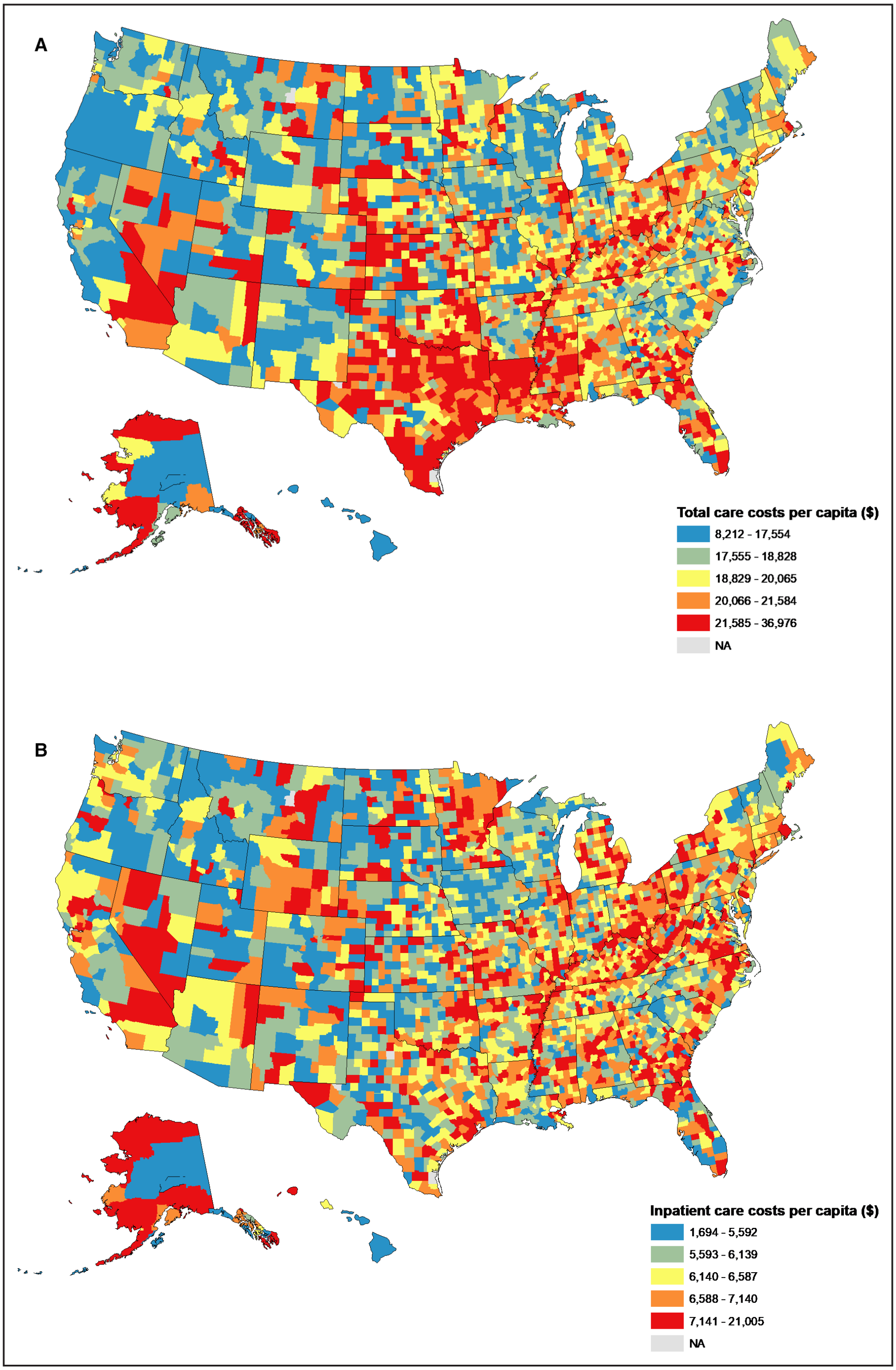

Figure 1. Four CVD outcomes of 3137 US counties, 2017.
A, Total care costs per capita. **B**, Inpatient care costs per capita. **C**, Outpatient care costs per capita. **D**, Prevalence of CVD. (Data source: *Interactive Atlas of Heart Disease and Stroke*). CVD indicates cardiovascular disease. NA indicates insufficient data.

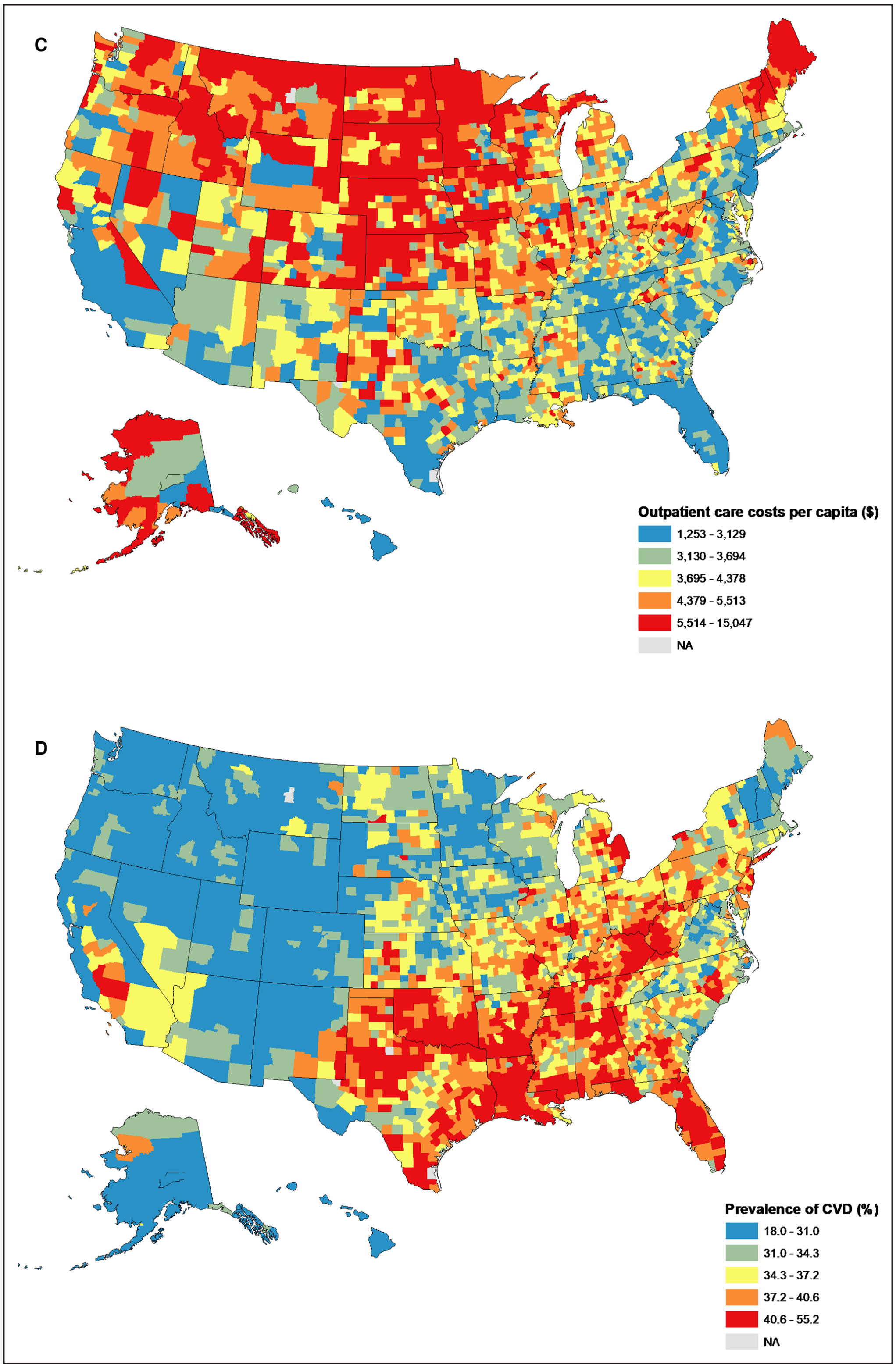

Figure 1. **Continued**

associated with the prevalence of CVD and inpatient costs. The percentage of smokers ranks first and second for the prevalence of CVD and inpatient costs, respectively. Obesity ranks fifth for inpatient costs. Physical inactivity and high cholesterol are among the top 5 determinants of CVD prevalence. Contextual factors are not consistently high ranking for the prevalence of CVD and inpatient costs, but they are major

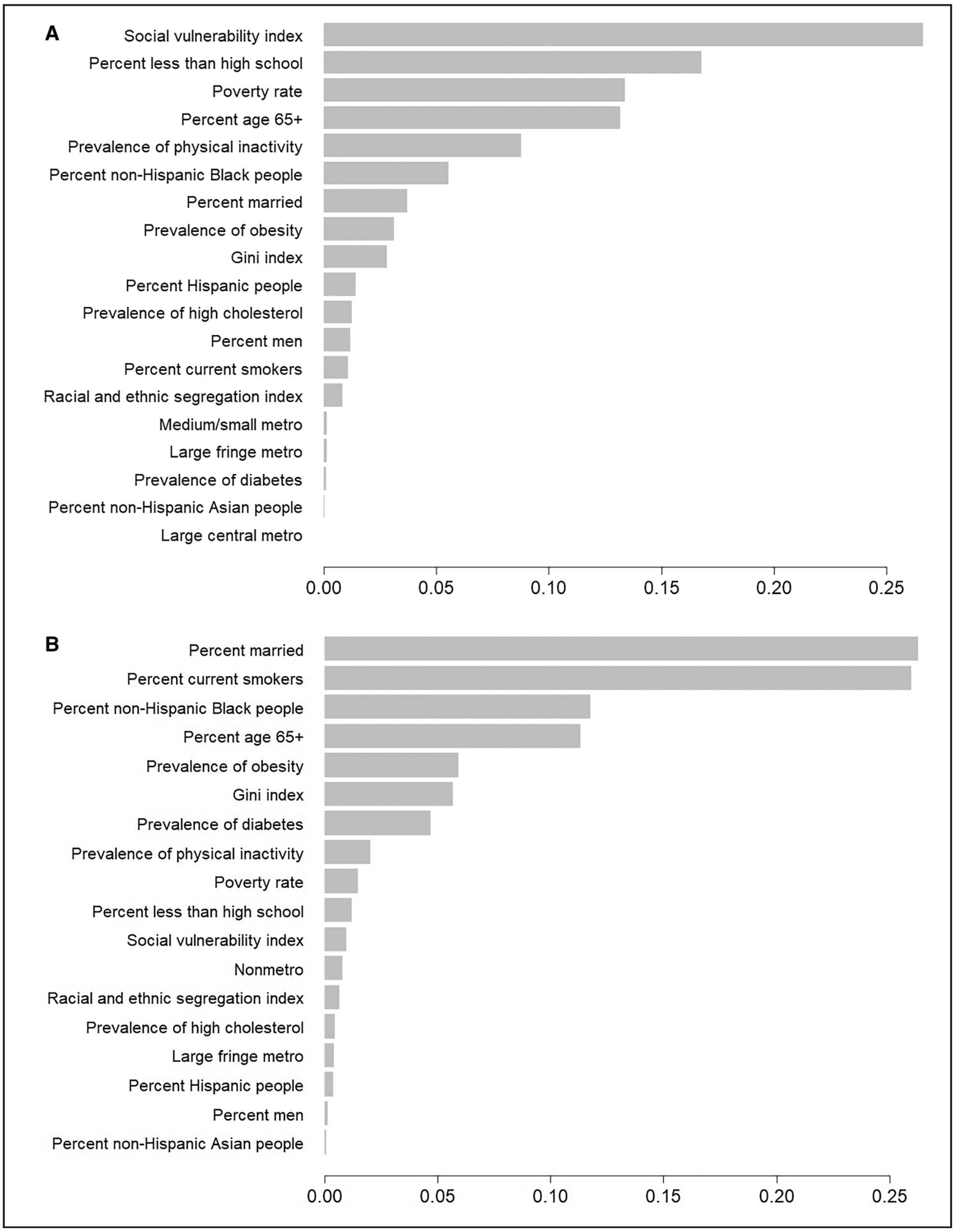

Figure 2. Extreme gradient boosting ranking results for 4 outcomes of 3137 US counties, 2017.
A, Total care costs per capita. **B**, Inpatient care costs per capita. **C**, Outpatient care costs per capita. **D**, Prevalence of CVD. CVD indicates cardiovascular disease.

contributors of total care costs and outpatient costs, ranking higher than risk factors.

Figure 3 displays the ranking results for total care costs based on large central metro, large fringe metro, medium/small metro, and nonmetro counties. For total care costs of CVD in large central metro counties, the top 5 predictors are the percentage of smokers, the Gini index, SVI, and the percentages of people with less than a high school education and non-Hispanic Black people. For large fringe metro counties, SVI, the percentages of older adults and Hispanic people, physical inactivity, and the percentage of current smokers are the top 5 important determinants of total care costs for CVD. Total care costs for CVD in medium/small metro counties rank SVI, physical inactivity, the percentages of current smokers and people who are married, and poverty rate highly in the top 5 places. For nonmetro counties, the percentages of people with less than a high school education and older adults, economic conditions (ie, poverty rate and the Gini

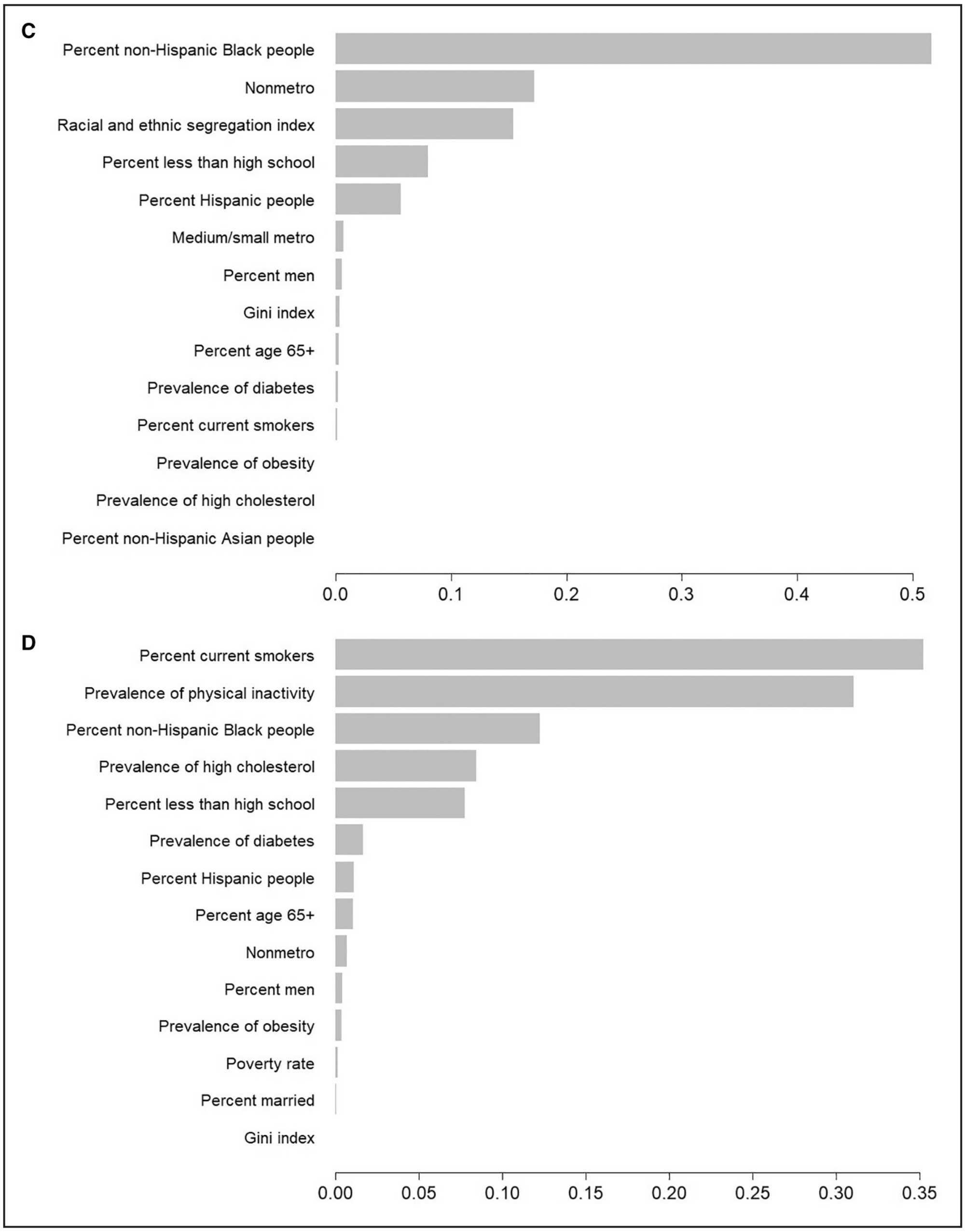

Figure 2. **Continued**

index), and physical inactivity are the top 5 determinants of total care costs for CVD.

The comparison between counties with low and high values for poverty rate, racial and ethnic segregation, and SVI is presented in Figure 4. Demographics, such as age, sex, marital status, and education rank high for total care costs for CVD both in counties with high and low poverty rates. For counties with low poverty rates, the top 2 rankings are percentages of men and people with less than a high school education. For counties with high poverty rates, the percentages of smokers, older adults, and people with less than a high school education rank in the first 3 places. Contextual factors are also important. For counties with low poverty rates, racial and ethnic segregation ranks fourth, whereas SVI ranks fifth for counties with high poverty rates.

There are also differences in the ranking of determinants across counties with low and high racial and ethnic segregation. The top 5 predictors of total costs in counties with low segregation are physical inactivity, the percentages of people with less than a high school education, non-Hispanic Black people, current smokers, and men. For total costs in counties with high segregation, the top 5 predictors are the percentages of married

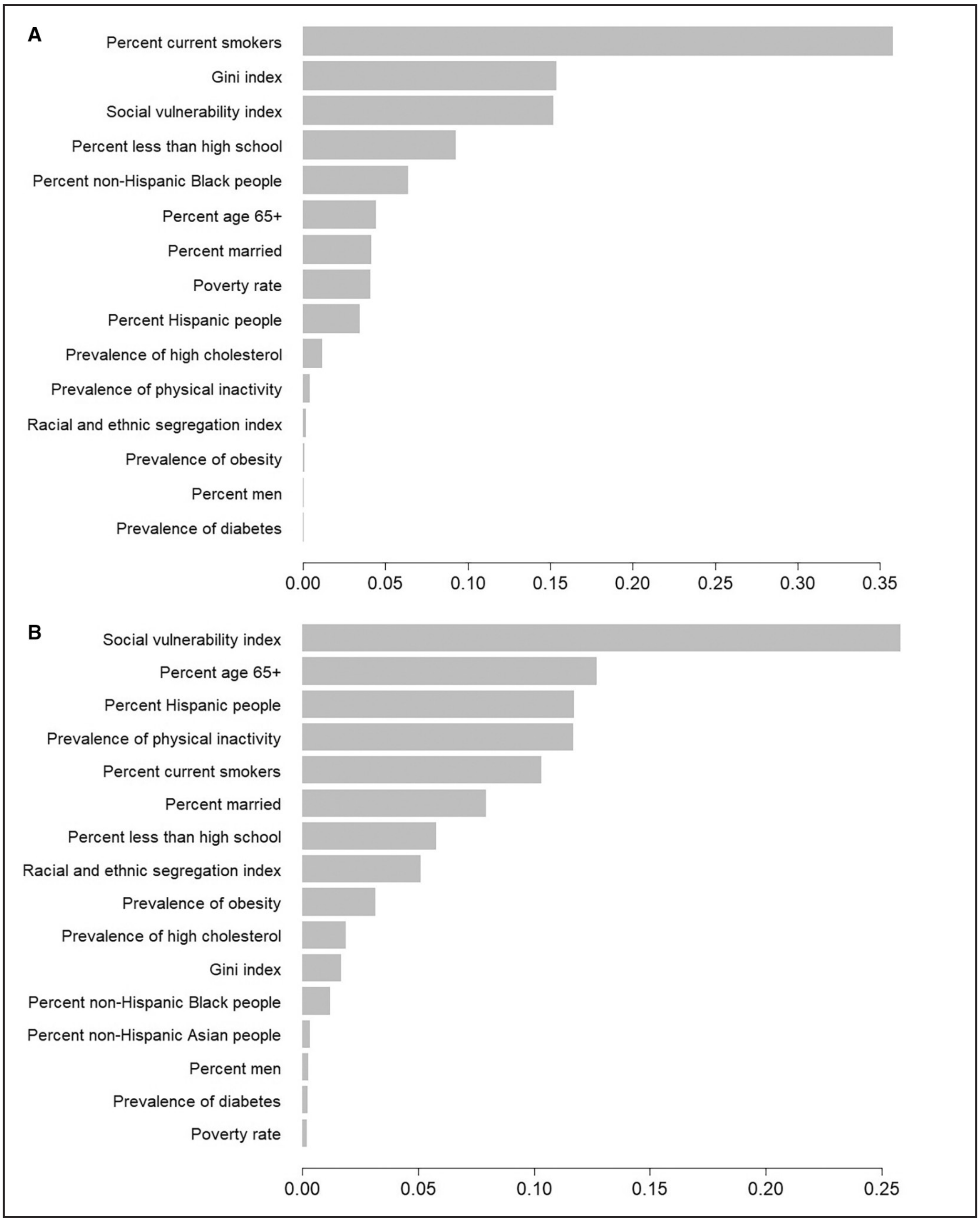

Figure 3. Extreme gradient boosting ranking results for total care costs per capita of 3137 US counties based on the National Center for Health Statistics urban–rural classification scheme, 2017.

A, Large central metro. **B**, Large fringe metro. **C**, Medium/small metro. **D**, Nonmetro.

individuals and people with less than a high school education, SVI, poverty rate, and obesity. Thus, risk factors (physical inactivity and smoking) and racial composition (the percentage of non-Hispanic Black people) rank higher for total costs in counties with low segregation, whereas contextual factors, including SVI and poverty rate, rank higher in counties with high segregation.

For counties with low SVI, racial and ethnic segregation, the percentages of married individuals and men, SVI, and the Gini index are the top 5 determinants of care costs for CVD. The Gini index, the percentages of older adults and Hispanic people, poverty rate, and the percentage of married individuals rank in the top 5 places for counties with high SVI.

DISCUSSION

Using a novel machine learning approach, this study aimed to identify the important predictors of county-level CVD prevalence and related care cost outcomes.

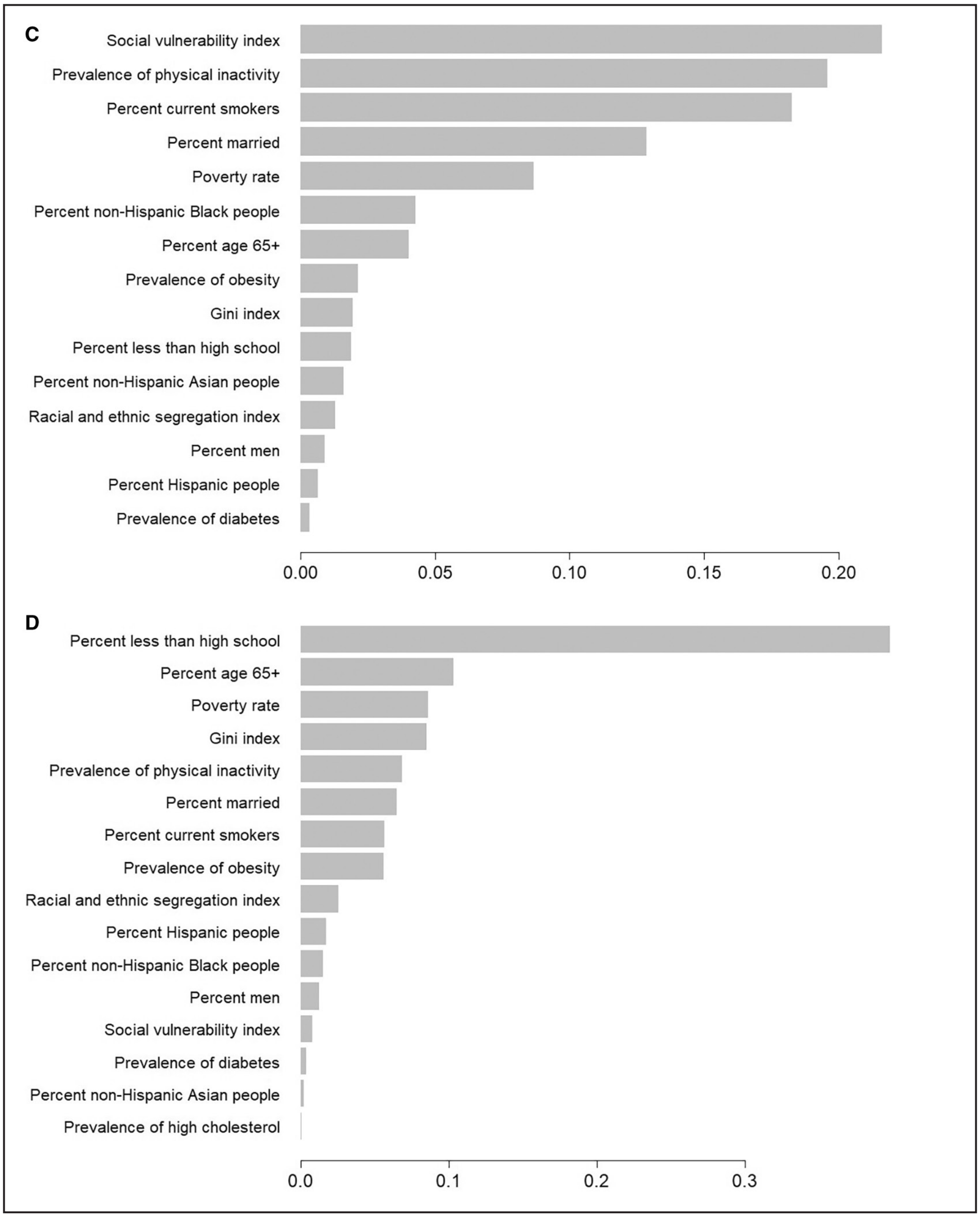

Figure 3. **Continued**

This is among the first studies to examine the role of social determinants nationwide in shaping health care costs. The differential spatial patterning of total, inpatient, and outpatient care costs for CVD and CVD prevalence suggests that it is worth examining CVD prevalence and different types of CVD costs separately. Our results do show that the ranking of determinants varies across CVD outcomes and counties.

First, the prevalence of CVD and inpatient costs are determined by the proportion of groups with a higher risk of CVD, whereas outpatient costs are impacted more by contextual characteristics of the living environments. This difference can be explained in that inpatient costs are probably driven by severe CVD, such as heart attack, whereas outpatient costs are likely affected by asymptomatic CVD, such as asymptomatic hypertension. Therefore, inpatient costs for CVD are determined by the percentages of groups at high risk but less affected by accessibility, whereas for asymptomatic CVD, accessibility (which is determined by contextual environments) becomes a major issue. Moreover, the finding is consistent with previous literature that finds that non-Hispanic Black people, older adults, individuals with lower education level, smokers, and individuals who have health issues and medical conditions are disproportionately affected by CVD.[3–6] It

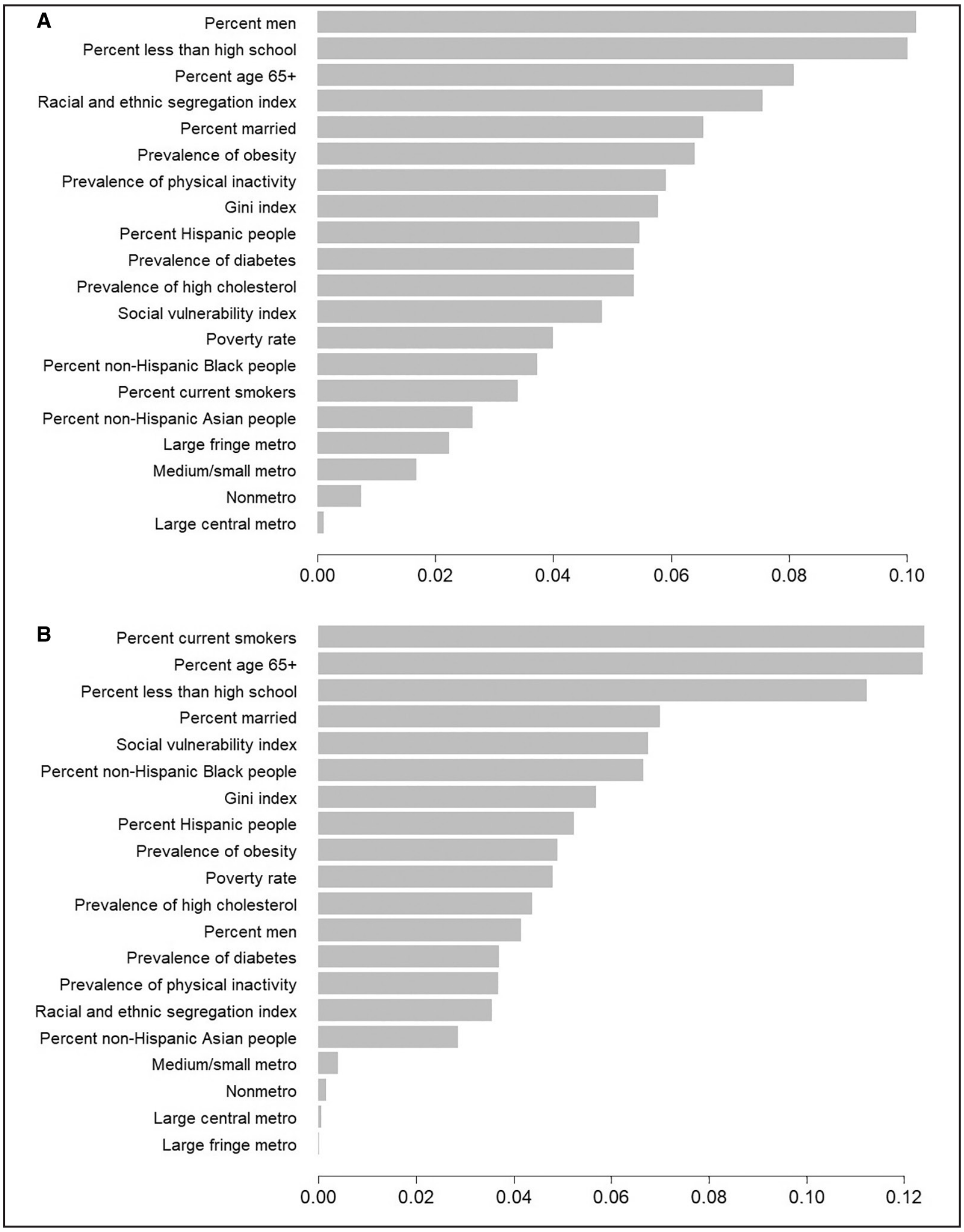

Figure 4. Extreme gradient boosting ranking results for total care costs per capita of 3137 US counties based on low and high values of 3 social determinants, 2017.
A and **B**, Low and high poverty rate. **C** and **D**, Low and high racial and ethnic segregation. **E** and **F**, Low and high Social Vulnerability Index.

also underlines the importance of social environments. Economic conditions, racial and ethnic segregation, and social vulnerability are all highly related to care costs for CVD. One possible explanation is that aspects of social environments may impact an individual's health care use and thus impact care costs. On one hand, individuals living in underresourced communities may have worse health conditions and thus greater demand for health services. Previous research using local data finds that poverty and social vulnerability are strongly associated with health care underuse.[42,43] On the other hand, unfavorable contexts may hinder access to health care services. For example, racial and ethnic minority groups that live in communities of color are less likely to have physician visits.[16] Thus, racial and ethnic segregation may negatively affect health care service use. Similarly, individuals in rural areas may also use fewer health services than their urban counterparts because of lower availability of services and greater distances to health care services.[44]

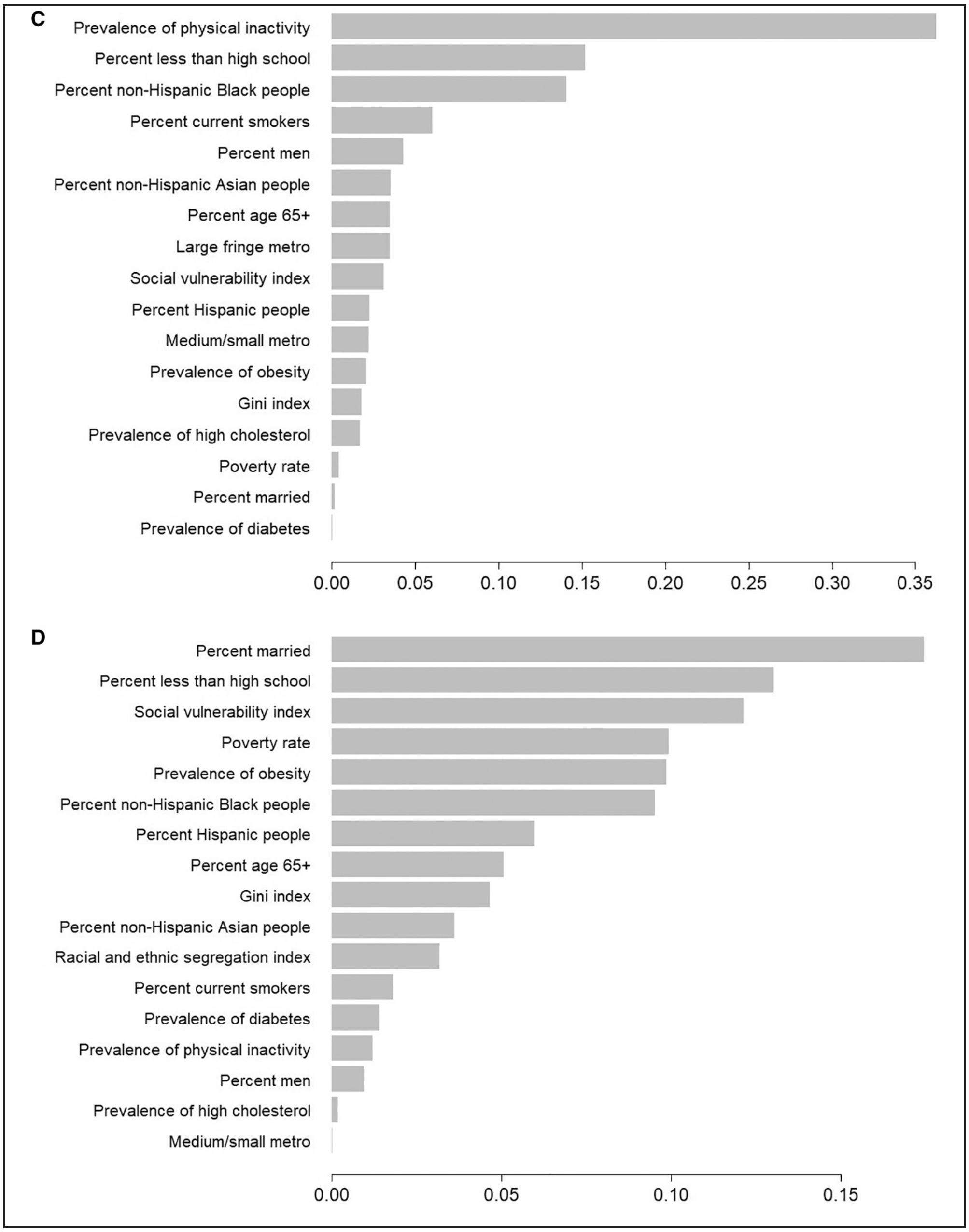

Figure 4. **Continued**

Second, we find that the determinants of total care costs for CVD in metro areas or counties having low poverty, segregation level, or social vulnerability level differ from counties experiencing disadvantages, though demographic composition variables, education and SVI are important for both. For counties experiencing disadvantages, poverty and income inequality contribute more to total care costs for CVD. There is only 1 exception, namely that smoking is the most important predictor for total care costs in counties with high poverty rate. Nonetheless, with respect to total care costs for CVD, counties with unfavorable contextual characteristics may be impacted more by economic distress and inequality than their more advantaged counterparts. For example, the Gini index and poverty rate are important contributors to CVD costs for non-metro counties and counties with high SVI, whereas poverty rate ranks high for counties with high segregation. Populations that have been marginalized, such as racial and ethnic minority groups and individuals living in poverty, may live in unfavorable health environments and have limited availability of health care resources and longer travel times to health services. This explanation is supported by evidence that in highly

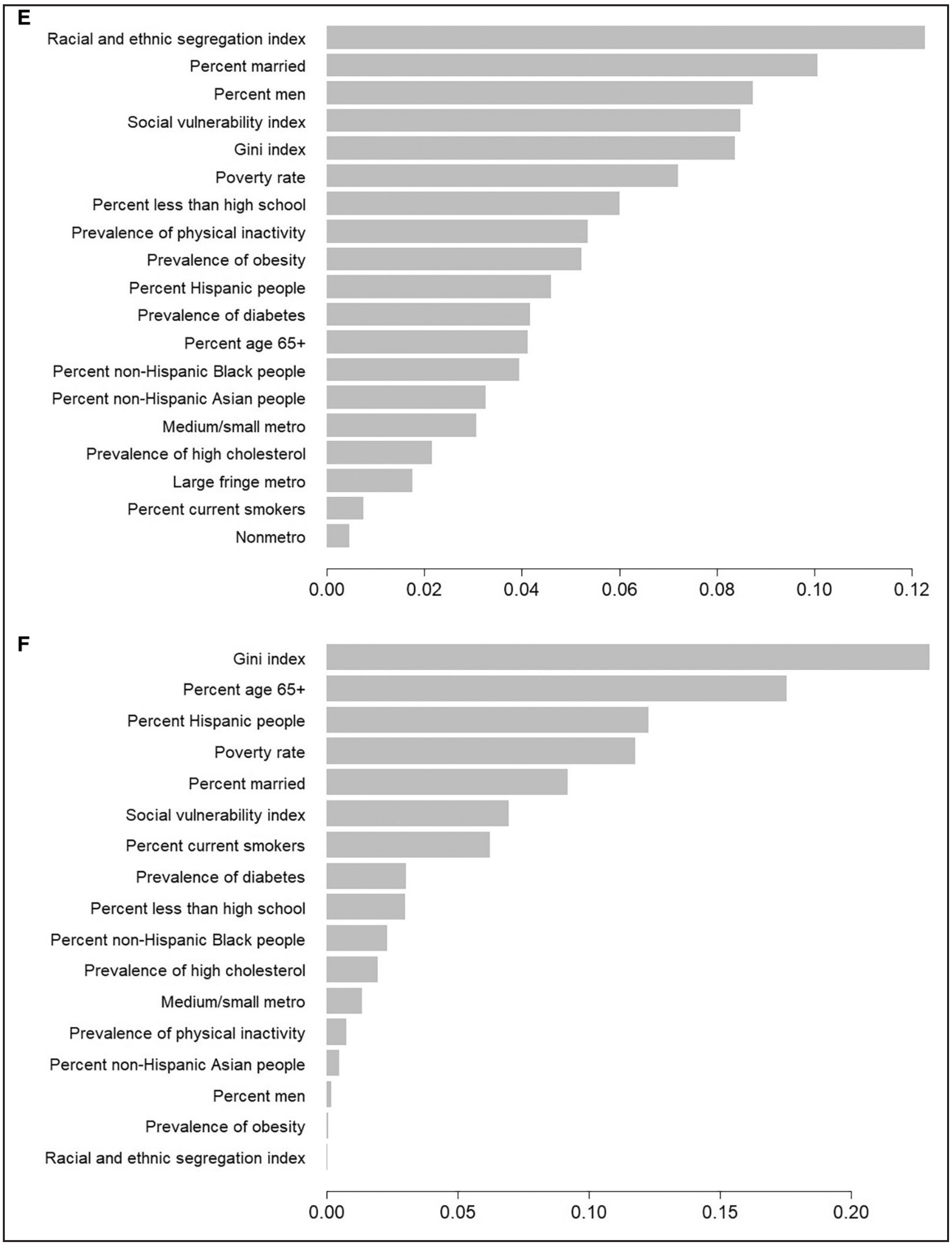

Figure 4. **Continued**

segregated areas, the accessibility of health providers and the use of health care services among racial and ethnic minority groups are more limited, potentially explaining observed worse health outcomes in these populations.[16,17,45,46]

In comparison, among relatively advantaged counties, such as large central, large fringe, and medium/small metro counties and counties with low segregation, smoking and physical inactivity are important risk factors. This finding suggests health behaviors and lifestyles of individuals are more important for more advantaged neighborhoods in decreasing CVD risk. In addition, for counties with low poverty or low SVI, racial and ethnic segregation is among the top predictors, indicating that even for those less disadvantaged counties, such segregation can largely impact CVD cost outcomes.

Advancing previous research that mainly focuses on individual health care costs,[25,47,48] our findings particularly highlight the important role of social determinants on CVD care costs, as well as the interactions between these determinants. On the one hand, economically deprived populations may suffer more in areas that have been marginalized, but on the other hand, the concentration of groups at high risk of CVD and segregation level are particularly important in shaping CVD costs in more advantaged areas. To control the costs of CVD, our study suggests the need to differentiate types of costs and target the diverse needs and disadvantages of different counties. Particularly, to improve the cardiovascular health of populations that are disproportionally affected, such as people of color, we may increase awareness of equity by education and training for patients, payers, and regulators; set up data and research strategies to monitor cardiovascular health equity; and emphasize the importance of diversity and inclusion in any trials, studies, or evaluations.

The study is subject to several limitations. First, the XGBoost approach only provides the ranking of independent variables in determining outcomes, not the direction of impacts. Nevertheless, these findings advance our understanding of the relationships between possible influential factors and CVD outcomes, such as prevalence and costs. Second, the analysis is cross-sectional and thus does not capture the dynamics of CVD prevalence and care costs. Future work should consider incorporating a longitudinal perspective. Third, our analysis is at the county level and thus is inevitably subject to the ecological fallacy and the modifiable areal unit problem. Research at other levels (eg, zip code) is needed to verify our findings at different levels. In addition, our study is only for the United States, but the findings may not be generalized to other countries. Relative importance of the factors considered here for CVD may be different in nations with different health care systems. Lastly, there is no nationwide data set available for use to validate our model-based findings. Further studies are needed in this regard.

Nonetheless, there are several strengths of this study. First, it is among the first to use a machine learning approach on comprehensive measures of CVD outcomes including the prevalence of CVD and 3 different health care costs. Second, it uses the most recent high-quality national data at the county level. Third, it examines a wide spectrum of variables with respect to not only demographic composition and risk factors, but also contextual factors in the social determinants of health domain, which include education, economic conditions, rural–urban status, racial and ethnic segregation, and social vulnerability. Our findings highlight the importance of these social contextual determinants as well as their interactions in shaping CVD care costs and call for more policy attention to economically and socially underresourced (eg, highly segregated) areas and residents experiencing disproportionate impact (eg, racial and ethnic minority groups and people experiencing poverty) living in these areas.

In conclusion, this study applies the rapidly developed machine learning approach to examine the relative importance of a variety of determinants in shaping CVD prevalence and care costs. We observed that predictors of CVD outcomes differ not only by types of outcomes but also by geographical location. In addition to demographic composition and risk factors (ie, medical conditions and health risk behaviors), these results highlight contextual factors, such as poverty rate, income inequality, social vulnerability, and racial and ethnic segregation in determining CVD outcomes. This study enhances our understanding of the geographic variations in CVD outcomes and contributes to controlling the development of CVD by highlighting the underresourced areas and informing the distribution of community resources. Moreover, this research suggests that policy focused on reducing CVD prevalence and health care costs should consider county-level variability in not only traditional CVD risk markers, like clinical conditions and health risk behaviors, but also the regional importance of contextual factors and social determinants of health. Findings from our study may help direct health policy implications in terms of prioritizing tasks of CVD prevention and cost control in different counties.

ARTICLE INFORMATION

Received August 22, 2022; accepted January 17, 2023.

Affiliations

Global Aging and Community Initiative, Mount Saint Vincent University, Halifax, Nova Scotia, Canada (F.S.); Department of Epidemiology and Biostatistics, School of Public Health (J.Y., A.A.A.), Department of Sociology (S.D.), and Department of Health Policy, Management and Behavior, School of Public Health (F.Q.), University at Albany, State University of New York, Albany, NY; School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX (C.T., H.X.); Division of Biostatistics, Washington University in St. Louis, St. Louis, MO (L.L.); Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, School of Medicine, Vanderbilt University, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN (Q.D.); Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL (B.T.J., D.R.N., L.H.); and Department of Environmental Health Sciences, School of Public Health, University at Albany, State University of New York, Albany, NY (K.Z.).

Sources of Funding

K.Z. and L.H. were partially funded by the American Heart Association (grant 19TPA34830085). K.Z., L.L., and L.H. were partially funded by the National Institute on Aging of the National Institutes of Health under award number R01AG081244. C.T. was partially funded by the American Heart Association (grant 19GPSGC35180031).

Disclosures

Dr Qian reports stock in Coinbase and options in AiDANT Intelligent Technology (Canada) and AUM Biosciences (Singapore). The remaining authors have no disclosures to report.

REFERENCES

1. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, Boehme AK, Buxton AE, Carson AP, Commodore-Mensah Y, et al. Heart disease and stroke statistics-2022 update: a report from the American Heart Association. *Circulation*. 2022;145:e153–e639. doi: 10.1161/CIR.0000000000001052

2. Roth GA, Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, Morozoff C, Naghavi M, Mokdad AH, Murray CJL. Trends and patterns of geographic variation in cardiovascular mortality among US counties, 1980-2014. *JAMA*. 2017;317:1976–1992. doi: 10.1001/jama.2017.4150
3. Kumar S. Cardiovascular disease and its determinants: public health issue. *J Clin Med Ther*. 2016;1:1–10.
4. Kleipool EE, Hoogendijk EO, Trappenburg MC, Handoko ML, Huisman M, Peters MJ, Muller M. Frailty in older adults with cardiovascular disease: cause, effect or both? *Aging Dis*. 2018;9:489–497. doi: 10.14336/AD.2017.1125
5. Rooks RN, Simonsick EM, Klesges LM, Newman AB, Ayonayon HN, Harris TB. Racial disparities in health care access and cardiovascular disease indicators in Black and White older adults in the Health ABC Study. *J Aging Health*. 2008;20:599–614. doi: 10.1177/0898264308321023
6. Havranek EP, Mujahid MS, Barr DA, Blair IV, Cohen MS, Cruz-Flores S, Davey-Smith G, Dennison-Himmelfarb CR, Lauer MS, Lockwood DW, et al. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the American Heart Association. *Circulation*. 2015;132:873–898. doi: 10.1161/CIR.0000000000000228
7. Tindle H, Davis E, Kuller L. Attitudes and cardiovascular disease. *Maturitas*. 2010;67:108–113. doi: 10.1016/j.maturitas.2010.04.020
8. Frasure-Smith N, Lesperance F. Depression and cardiac risk: present status and future directions. *Postgrad Med J*. 2010;86:193–196. doi: 10.1136/hrt.2009.186957
9. Artiga S, Hinton E. Beyond health care: the role of social determinants in promoting health and health equity. *Health*. 2019;20:1–13.
10. Meneton P, Kesse-Guyot E, Mejean C, Fezeu L, Galan P, Hercberg S, Menard J. Unemployment is associated with high cardiovascular event rate and increased all-cause mortality in middle-aged socially privileged individuals. *Int Arch Occup Environ Health*. 2015;88:707–716. doi: 10.1007/s00420-014-0997-7
11. Mensah GA, Mokdad AH, Ford ES, Greenlund KJ, Croft JB. State of disparities in cardiovascular health in the United States. *Circulation*. 2005;111:1233–1241. doi: 10.1161/01.CIR.0000158136.76824.04
12. Diez Roux AV. Residential environments and cardiovascular risk. *J Urban Health*. 2003;80:569–589. doi: 10.1093/jurban/jtg065
13. Schultz WM, Kelli HM, Lisko JC, Varghese T, Shen J, Sandesara P, Quyyumi AA, Taylor HA, Gulati M, Harold JG, et al. Socioeconomic status and cardiovascular outcomes: challenges and interventions. *Circulation*. 2018;137:2166–2178. doi: 10.1161/CIRCULATIONAHA.117.029652
14. Kershaw KN, Albrecht SS. Racial/ethnic residential segregation and cardiovascular disease risk. *Curr Cardiovasc Risk Rep*. 2015;9:9. doi: 10.1007/s12170-015-0436-7
15. Ko M, Needleman J, Derose KP, Laugesen MJ, Ponce NA. Residential segregation and the survival of U.S. urban public hospitals. *Med Care Res Rev*. 2014;71:243–260. doi: 10.1177/1077558713515079
16. Gaskin DJ, Dinwiddie GY, Chan KS, McCleary R. Residential segregation and disparities in health care services utilization. *Med Care Res Rev*. 2012;69:158–175. doi: 10.1177/1077558711420263
17. Gaskin DJ, Dinwiddie GY, Chan KS, McCleary RR. Residential segregation and the availability of primary care physicians. *Health Serv Res*. 2012;47:2353–2376. doi: 10.1111/j.1475-6773.2012.01417.x
18. Johnson AE, Brewer LC, Echols MR, Mazimba S, Shah RU, Breathett K. Utilizing artificial intelligence to enhance health equity among patients with heart failure. *Heart Fail Clin*. 2022;18:259–273. doi: 10.1016/j.hfc.2021.11.001
19. Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ Digit Med*. 2021;4:154. doi: 10.1038/s41746-021-00524-2
20. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, Floridi L. Clinical applications of machine learning algorithms: beyond the black box. *BMJ*. 2019;364:l886. doi: 10.1136/bmj.l886
21. Zhao Y, Wood EP, Mirin N, Cook SH, Chunara R. Social determinants in machine learning cardiovascular disease prediction models: a systematic review. *Am J Prev Med*. 2021;61:596–605. doi: 10.1016/j.amepre.2021.04.016
22. Hu L, Liu B, Ji J, Li Y. Tree-based machine learning to identify and understand major determinants for stroke at the neighborhood level. *J Am Heart Assoc*. 2020;9:e016745. doi: 10.1161/JAHA.120.016745
23. Hu L, Liu B, Li Y. Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: a Bayesian machine learning approach. *Prev Med*. 2020;141:106240. doi: 10.1016/j.ypmed.2020.106240
24. Li Y, Liu SH, Niu L, Liu B. Unhealthy behaviors, prevention measures, and neighborhood cardiovascular health: a machine learning approach. *J Public Health Manag Pract*. 2019;25:E25–E28. doi: 10.1097/PHH.0000000000000817
25. Li L, Hu L, Ji J, McKendrick K, Moreno J, Kelley AS, Mazumdar M, Aldridge M. Determinants of total end-of-life healthcare costs of medicare beneficiaries: a quantile regression forests analysis. *J Gerontol A Biol Sci Med Sci*. 2022;77:1065–1071. doi: 10.1093/gerona/glab176
26. Hu L, Li L, Ji J, Sanderson M. Identifying and understanding determinants of high healthcare costs for breast cancer: a quantile regression machine learning approach. *BMC Health Serv Res*. 2020;20:1066. doi: 10.1186/s12913-020-05936-6
27. Rakshit P, Zaballa O, Perez A, Gomez-Inhiesto E, Acaiturri-Ayesta MT, Lozano JA. A machine learning approach to predict healthcare cost of breast cancer patients. *Sci Rep*. 2021;11:12441. doi: 10.1038/s41598-021-91580-x
28. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. *Biomed Eng Online*. 2018;17:131. doi: 10.1186/s12938-018-0568-3
29. Interactive Atlas of Heart Disease and Stroke. Centers for Disease Control and Prevention. Accessed December 15, 2021. http://nccd.cdc.gov/DHDSPAtlas
30. American Community Survey 5-Year Data (2015-2019). United States Census Bureau. Accessed June 15, 2021. https://www.census.gov/data/developers/data-sets/acs-5year.html
31. Census and PolicyMap: Racial and Ethnic Segregation (2010). PolicyMap. Accessed June 15, 2021. https://www.policymap.com
32. CDC/ATSDR Social Vulnerability Index (2018). Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. Accessed June 15, 2021. https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html
33. 2022 ICD-10-CM Diagnosis Code Z13.6. ICD10Data. Accessed April 15, 2022. https://www.icd10data.com/ICD10CM/Codes
34. CCW Condition Algorithms (rev. 07/2016). CMS Chronic Conditions Data Warehouse. Accessed April 15, 2022. https://aspe.hhs.gov/sites/default/files/private/pdf/252376/AppendixA.pdf
35. Gini Index. US Census Bureau. Accessed April 15, 2022. https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/gini-index.html
36. Ingram DD, Franco SJ. 2013 NCHS urban-rural classification scheme for counties. *Vital Health Stat 2*. 2014:1–73.
37. Hatton CM, Paton LW, McMillan D, Cussens J, Gilbody S, Tiffin PA. Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. *J Affect Disord*. 2019;246:857–860. doi: 10.1016/j.jad.2018.12.095
38. Qiu Z, Wang J, Yu B, Liao L, Li J. Identification of passive solar design determinants in office building envelopes in hot and humid climates using data mining techniques. *Build Environ*. 2021;196:107566. doi: 10.1016/j.buildenv.2020.107566
39. Makridis CA, Zhao DY, Bejan CA, Alterovitz G. Leveraging machine learning to characterize the role of socio-economic determinants on physical health and well-being among veterans. *Comput Biol Med*. 2021;133:104354. doi: 10.1016/j.compbiomed.2021.104354
40. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794.
41. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, et al. xgboost: Extreme Gradient Boosting; 2022. Accessed February 10, 2023. https://github.com/dmlc/xgboost
42. Cooper RA, Cooper MA, McGinley EL, Fan X, Rosenthal JT. Poverty, wealth, and health care utilization: a geographic assessment. *J Urban Health*. 2012;89:828–847. doi: 10.1007/s11524-012-9689-3
43. Rickless DS, Wilt GE, Sharpe JD, Molinari N, Stephens W, TT LB. Social vulnerability and access of local medical care during Hurricane Harvey: a spatial analysis. *Disaster Med Public Health Prep*. 2021;17:e12. doi: 10.1017/dmp.2020.421
44. Andersen R, Newman JF. Societal and individual determinants of medical care utilization in the United States. *Milbank Q*. 2005;83:1–28. doi: 10.1111/j.1468-0009.2005.00428.x
45. Yang TC, Zhao Y, Song Q. Residential segregation and racial disparities in self-rated health: how do dimensions of residential segregation matter? *Soc Sci Res*. 2017;61:29–42. doi: 10.1016/j.ssresearch.2016.06.011

46. Anderson KF. Racial residential segregation and the distribution of health-related organizations in urban neighborhoods. *Soc Probl*. 2017;64:256–276. doi: 10.1093/socpro/spw058

47. Slabaugh SL, Curtis BH, Clore G, Fu H, Schuster DP. Factors associated with increased healthcare costs in Medicare Advantage patients with type 2 diabetes enrolled in a large representative health insurance plan in the US. *J Med Econ*. 2015;18:106–112. doi: 10.3111/13696998.2014.979292

48. Reschovsky JD, Hadley J, Saiontz-Martinez CB, Boukus ER. Following the money: factors associated with the cost of treating high-cost medicare beneficiaries. *Health Serv Res*. 2011;46:997–1021. doi: 10.1111/j.1475-6773.2011.01242.x