NORTHWESTERN UNIVERSITY

Complex Exposures to Social Determinants of Health through Young Adulthood and Associations with Mid-life Cardiovascular Health and Events: The Coronary Artery Risk Development in Young Adults (CARDIA) Study

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Health Sciences Integrated Program

By

Lindsay Zimmerman

EVANSTON, ILLINOIS

September 2021

ABSTRACT

In the U.S., approximately 840,000 Americans die from cardiovascular disease (CVD) each year, and it is the leading cause of morbidity and mortality worldwide. The prevalence of CVD is on the rise and widespread disparities in CVD exist across economic, racial, and ethnic groups. In order to address the rising prevalence of CVD and persistent disparities, there has been a shift in focus to strategies promoting cardiovascular health (CVH) across the life course. CVH is a broader and more positive construct beyond the absence of CVD, and allows for clinical and public health strategies focused on disease prevention and health promotion, rather than solely on treatment once CVD develops. Despite this recent focus on improving CVH, widespread disparities still exist, and social determinants of health (SDOH) appear to be important contributors to these continued disparities. The World Health Organization (WHO) defines SDOH as the “structural determinants and conditions in which people are born, grow, live, work, and age.” There has been limited work studying how longitudinal exposures of SDOH change over time, perform in the prediction of CVH, and are associated with later-life CVH and CVD events.

The primary objectives of this dissertation are to identify patterns of SDOH exposure over time by generating data-driven SDOH clusters using a novel machine learning method, and determine whether the addition of SDOH variables allowed for better prediction of an individual’s CVH status and were associated with mid-life CVH (a critical milestone in CVH maintenance) and CVD events. The primary hypothesis for this study is that a diverse set of SDOH from young adulthood through middle age will be

predictive of mid-life CVH and will be associated with mid-life CVH and CVD events, independent of baseline CVH and other covariates.

This dissertation begins with a general introduction highlighting the burden of CVD and the current evidence and methods used to link SDOH and CVD. I then present the three chapters included in this dissertation, all of which leveraged comprehensive longitudinal data from the Coronary Artery Risk Development in Young Adults (CARDIA) study cohort. Chapter 2 includes a detailed description of the methods used to identify frequent time-dependent SDOH patterns and generate the novel SDOH clusters in a well-phenotyped long-standing community-based study. This chapter demonstrates that the clusters generated improved the prediction of mid-life CVH. Chapter 3 presents a more detailed description of time-dependent individual- and neighborhood-level SDOH exposure patterns and shows an association between clusters and mid-life CVH, overall and by self-identified race groups. Chapter 4 evaluates whether the SDOH clusters are associated with mid-life CVD events before and after adjustment for mid-life CVH and subclinical CVD. The clusters are associated with mid-life CVD events, but not after adjustment for mid-life CVH and subclinical CVD.

The three chapters support our original hypothesis. First, we can use a novel machine learning method to identify time-dependent SDOH patterns from young adulthood to middle age and create novel SDOH clusters of those patterns that provide insight into the complex inter-relationships of SDOH. Additionally, the clusters were predictive of mid-life CVH and associated with mid-life CVH and CVD events. Further refinement and validation of the clusters is necessary. The findings from this dissertation may be used to inform programs looking to develop targeted, timely, and multi-component interventions to address SDOH and improve CVH in young adults, with the potential to improve population CVH and reduce disparities.

ACKNOWLEDGEMENTS

I would not be where I am today without the love and support of my family and friends. They have been a constant source of positivity, laughter, and inspiration.

This dissertation was made possible by the unwavering support of my committee, who shared their expertise in cardiovascular health and data science and were so giving of their time. Dr. Don Lloyd-Jones, my chair, taught me what it means to be a true mentor, encouraging and supportive leader, and that revision *is* the research process—not just part of it. Dr. Kiarri Kershaw, Dr. Yuan Luo, and Dr. David Rehkopf taught me *how* to be an interdisciplinary researcher, contribute to the scientific community, and understand the broader implications of my work. I will be forever thankful to have had this incredible team of mentors as my primary thought partners throughout this process.

My time during the PhD was supported by three doctoral training and dissertation grants. My second and third years were funded by the National Library of Medicine pre-doctoral (T32) training program in Biomedical Data Driven Discovery (T32LM012203). My dissertation work was supported by the National Heart, Lung, and Blood Institute Pre-doctoral (F31) Ruth L. Kirschstein National Research Service Award (F31HL151178), as well as the Northwestern University Graduate School Presidential Fellowship. And thank you to the Coronary Artery Risk Development in Young Adults (CARDIA) Study participants for donating their time, specimens, and data to this study and to the CARDIA investigators for the peer-review of our proposals and manuscripts.

And finally, I am grateful for the encouragement and feedback from the HSIP leadership and my academic community during my time at Northwestern.

ABBREVIATIONS

AUC = Area-Under-the-Curve

CAC = Coronary Artery Calcification

CARDIA = Coronary Artery Risk Development in Young Adults Study

CES-D = Center for Epidemiologic Studies Depression Scale

CI = Confidence Interval

CVD = Cardiovascular Disease

CVH = Cardiovascular Health

DBP = Diastolic Blood Pressure

HR = Hazard Ratio

HS = High School

LVMI = Left Ventricular Mass Index

MV = Multivariable

NMF = Non-Negative Matrix Factorization

SANMF = Subgraph Augmented Non-Negative Matrix Factorization

SBP = Systolic Blood Pressure

SES = Socioeconomic Status

SD = Standard Deviation

SDOH = Social Determinants of Health

TSEI = Total-Based Socioeconomic Index

Y = Exam Year

To my parents and lifelong supporters, Nancy and Jeff.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION 15

1.1 BACKGROUND 15

1.2 OBJECTIVES AND OUTLINE 20

1.3 CONCEPTUAL FRAMEWORK 22

1.4 INNOVATION 22

CHAPTER TWO: APPLICATION OF A NOVEL SEQUENTIAL PATTERN MINING METHOD TO STUDY LONGITUDINAL SOCIAL DETERMINANTS OF HEALTH: THE CORONARY ARTERY RISK DEVELOPMENT IN YOUNG ADULTS (CARDIA) STUDY 26

2.1 ABSTRACT 26

2.2 INTRODUCTION 28

2.3 OBJECTIVES 29

2.4 METHODS 31

2.4.1 Dataset, Predictors, and Outcome 31

2.4.2 Overall Workflow 33

2.4.3 Graph Construction 33

2.4.4 Identification of Time-Dependent SDOH Patterns with Frequent Subgraph Mining 34

2.4.5 Creation of Clusters with Sparse Non-negative Matrix Factorization 36

2.4.6 Characterization of SDOH Clusters 37

2.4.7 Prediction of Cardiovascular Health 38

2.5 RESULTS 40

2.5.1 SDOH Clusters 40

2.5.2 Prediction of CVH 41

2.5.3 Important Clusters and Subgraphs 41

2.6 Discussion 43

2.7 Conclusion 47

CHAPTER THREE: NOVEL TIME-DEPENDENT CLUSTERS OF SOCIAL DETERMINANTS THROUGH YOUNG ADULTHOOD AND ASSOCIATIONS WITH MID-LIFE CARDIOVASCULAR HEALTH: THE CARDIA STUDY 48

3.1 Abstract 48

3.2 Introduction 51

3.3 Objectives 52

3.4 Methods 53

3.4.1 Study Design and Population 53

3.4.2 Primary Exposures 53

3.4.3 Primary Outcome 55

3.4.4 Statistical Analysis 56

3.5 Results 58

3.5.1 Study Sample 58

3.5.2 Characterization of Clusters 58

3.5.3 Associations of SDOH Clusters with Mid-Life CVH 59

3.6 Discussion 61

3.7 Conclusions 64

3.8 Supplement 66

CHAPTER FOUR: TIME-DEPENDENT CLUSTERS OF SOCIAL DETERMINANTS OF HEALTH IN YOUNG ADULTHOOD AND MID-LIFE CARDIOVASCULAR DISEASE EVENTS: THE CORONARY ARTERY RISK DEVELOPMENT IN YOUNG ADULTS (CARDIA) STUDY 91

4.1 Abstract 91

4.2 Introduction 93

4.3 Objectives 94

4.4 Methods 95

4.4.1 Study Design and Population 95

4.4.2 Exposures 95

4.4.3 Outcomes 96

4.4.4 Statistical Analysis 98

4.5 Results 100

4.5.1 Study Sample 100

4.5.2 Characterization of SDOH Clusters 100

4.5.3 Primary Analyses of SDOH-CVD Event Associations 101

4.5.4 Secondary Analyses Adjusting for Mid-life CVH and Subclinical CVD 103

4.6 Discussion 104

4.7 Conclusions 106

4.8 Supplement 108

CHAPTER FIVE: CONCLUSION AND FUTURE DIRECTIONS 110

5.1 Principal Findings 110

5.2 Methodological Approach 111

5.2.1 Strengths and Limitations 112

5.3 Implications 114

5.3.1 Intersectionality of SDOH exposures 115

5.3.2 SDOH as causal factors for CVH and CVD 115

5.3.3 Support for a primordial prevention strategy 117

5.4 Future Research 117

5.5 Summary 121

TABLES AND FIGURES 122

.... 135

.... 138

REFERENCES 158

LIST OF TABLES AND FIGURES

TABLE 1: VARIABLE DESCRIPTIONS FOR EACH OF THE 48 INDIVIDUAL- AND NEIGHBORHOOD-LEVEL SOCIAL DETERMINANT OF HEALTH (SDOH) AND PSYCHOSOCIAL VARIABLES ASSESSED IN CARDIA PARTICIPANTS 122

TABLE 2: SUMMARY OF DEMOGRAPHIC, EDUCATION, CARDIOVASCULAR HEALTH, AND SELECT SDOH CHARACTERISTICS FOR TRAIN AND TEST SETS AMONG CARDIA STUDY PARTICIPANTS 130

TABLE 3: CHARACTERIZATION OF THE FIVE SDOH CLUSTERS WITH THEIR TOP 10 TIME-DEPENDENT SDOH PATTERNS FROM NMF 132

TABLE 4: PREDICTIVE PERFORMANCE (DISCRIMINATION) OF SELECTED MODELS FOR MID-LIFE CARDIOVASCULAR HEALTH BY PREDICTOR GROUP 134

TABLE 5: SUMMARY OF DEMOGRAPHIC AND EDUCATION MEASURES OVERALL AND BY RACE 135

TABLE 6: MULTIVARIABLE-ADJUSTED LOGISTIC REGRESSION ANALYSIS OF ASSOCIATIONS OF SDOH CLUSTERS THROUGH YOUNG ADULTHOOD WITH POOR CVH STATUS IN MIDDLE AGE AMONG ALL PARTICIPANTS.............. 136

TABLE 7: MULTIVARIABLE-ADJUSTED LOGISTIC REGRESSION ANALYSIS OF ASSOCIATIONS OF SDOH CLUSTERS (CREATED USING ALL PARTICIPANTS) THROUGH YOUNG ADULTHOOD WITH POOR CVH STATUS IN MIDDLE AGE AMONG BLACK AND WHITE PARTICIPANTS... 138

TABLE 8: SUMMARY OF DEMOGRAPHIC, EDUCATION, CVD RISK FACTOR, AND SOCIAL DETERMINANTS OF HEALTH MEASURES OVERALL AND BY COHORT 139

TABLE 9: SUMMARY OF CVD EVENTS, SUBCLINICAL CVD, AND CVH MEASURES BY COHORT 141

TABLE 10: HAZARD RATIOS FOR CVD EVENTS, UNADJUSTED AND ADJUSTED FOR AGE AT BASELINE, SEX, RACE, AND EDUCATION 142

TABLE 11: HAZARD RATIOS FOR CVD EVENTS, ADJUSTED FOR BASELINE CVH SCORE, AGE AT BASELINE, SEX, RACE, AND EDUCATION 143

TABLE 12: RACE-STRATIFIED ANALYSES PARTIALLY AND FULLY ADJUSTED BY COVARIATES 144

TABLE 13: SECONDARY ANALYSES FOR SDOH-CVD EVENT ASSOCIATIONS ADJUSTING FOR MID-LIFE CVH AND SUBCLINICAL CVD 145

FIGURE 1: SOCIAL DETERMINANTS OF HEALTH – KEY DOMAINS AND ISSUE.. 147

FIGURE 2: SDOH AND CVH CONCEPTUAL FRAMEWORK 148

FIGURE 3: SUBGRAPH AUGMENTED NON-NEGATIVE MATRIX FACTORIZATION (SANMF) WORKFLOW 149

FIGURE 4: SUBGRAPH AUGMENTED NON-NEGATIVE MATRIX FACTORIZATION MODEL 150

FIGURE 5: VARIABLE IMPORTANCE PLOT FROM THE BASE + CVH + SDOH CLUSTERS LOGISTIC REGRESSION MODEL 151

FIGURE 6: VARIABLE IMPORTANCE PLOT FROM THE BASE + CVH + SUBGRAPHS LASSO REGRESSION MODEL....... 152

FIGURE 7: SUBGRAPH-AUGMENTED NON-NEGATIVE MATRIX FACTORIZATION (SANMF) METHODOLOGY TO CREATE AND CHARACTERIZE SOCIAL DETERMINANTS OF HEALTH (SDOH) CLUSTERS 153

FIGURE 8: CHARACTERIZATION OF THE FIVE SOCIAL DETERMINANTS OF HEALTH (SDOH) CLUSTERS CREATED AMONG ALL PARTICIPANTS. 154

FIGURE 9: CHARACTERIZATION OF SDOH CLUSTERS PRESENTING PATTERNS WITH THE TOP TEN HIGHEST MEMBERSHIP COEFFICIENTS 155

FIGURE 10: CVD EVENT RATES PER 1,000 PERSON-YEARS BY RACE AND SEX SUBGROUPS AND SDOH CLUSTER.... 157

CHAPTER ONE: INTRODUCTION

1.1 Background

Cardiovascular disease is a critical public health problem with widespread health disparities.

Cardiovascular disease (CVD) is the leading cause of death for Americans, accounting for approximately 840,000 deaths per year.[1] The prevalence of CVD is on the rise, and by 2030, 40.5% of the population will have some form of CVD, up from 36.9% in 2010.[2] There remain widespread health disparities in CVD burden across economic, racial, ethnic, and geographic groups both nationally and worldwide.[3,4] For example, non-Hispanic (NH) Black individuals bear a disproportionate burden, with substantially and persistently higher prevalence of CVD and higher CVD death rates compared with NH white individuals.[1] Men also have higher CVD prevalence and death rates than women.[1] The disparities in CVD burden can be largely explained by differences in modifiable, and not genetically pre-determined, risk factors and behaviors.[3,5–8]

The improvement of cardiovascular health has been a recent focus for the prevention of CVD, although major racial disparities exist and are not well understood.

In order to address the rising number of Americans affected by CVD and the persistent disparities affecting marginalized populations in the country, there must be a focus on public health and preventive strategies to address CVD, beyond focusing solely on what can be done clinically at a hospital or doctor’s office once CVD develops.

To spark this change, the American Heart Association (AHA) in 2010 set a Strategic Impact Goal "By 2020, to improve the cardiovascular health of all Americans by 20% while reducing deaths from cardiovascular diseases and stroke by 20%."[9] **Cardiovascular health (CVH)**, as defined by Lloyd-Jones (committee chair) et al. for the AHA, is determined by seven key health factors and behaviors including: total cholesterol, blood pressure, fasting plasma glucose, body mass index, smoking status, physical activity, and healthy diet score.[9] These component metrics are each classified as poor, intermediate, or ideal based on clinically relevant cutpoints, and a composite CVH score can be used to represent overall CVH in an individual or in populations. CVH is a broader, more positive construct beyond just the absence of CVD and can be measured for all individuals including those in younger age groups.[9] Among U.S. adults, only 13% have ideal levels of 5 CVH component metrics, 5% have ideal levels of 6 metrics, and <1% have ideal levels of all 7 metrics.[1] More favorable mid-life CVH status is associated with many positive cross-sectional and longitudinal outcomes including total mortality, CVD-related mortality and non-fatal events, incident cancer, cognition, depression, quality of life, healthy longevity, compression of morbidity, and healthcare charges, among many others.[9–21]

Despite the recent focus on improving CVH, major racial disparities still remain and are not well understood.[22] NH Black and Hispanic individuals generally have fewer ideal levels of the CVH metrics than those who self-identify as being of White or other races; presence of greater than or equal to 4 metrics at ideal levels is most common among

Asian individuals (48%), followed by white (38%), Hispanic (34%), and Black individuals (30%), and others (24%).[1]

Social determinants of health, made up of five key domains and measured on both individual and neighborhood levels, may account for CVH disparities.

Social determinants of health (SDOH) may account for some of the persistent CVH disparities. The World Health Organization (WHO) defines SDOH as the "structural determinants and conditions in which people are born, grow, live, work, and age" that affect health, functioning, and quality of life.[23] There are five key domains of SDOH including economic stability, neighborhood and built environment, education, social and community context, and health and health care.[24] The SDOH domains and their associated issues are pictured in Figure 1. SDOH can be measured at both an individual and neighborhood level. Individual-level SDOH information comes from direct patient report through surveys, cohort studies, and physician interviews. Neighborhood-level SDOH provide a picture of the geographic area where an individual lives. Data from the U.S. Census Bureau and other public datasets are rich resources for neighborhood-based SDOH information at various geographic areas.[25,26]

Data and research are sparse regarding associations of complex SDOH constructs with CVH.

The National Heart, Lung, and Blood Institute (NHLBI)'s strategic vision and the AHA's recent scientific statement highlight the importance of understanding SDOH as a determinant of CVH.[27,28] Associations between a limited set of SDOH, most often from the economic stability and education domains, and some of the individual CVH component metrics have been described, cross-sectionally and longitudinally.[27,29,30] For example, it is well established that those with low education have poorer levels of CVH metrics, including higher rates of smoking, high blood pressure, and high total cholesterol, compared with well-educated groups.[7,31] Individuals with several social risk factors including low family income, low education, minority race, and single-living status, had lower odds of having 5-7 versus 0 ideal CVH components cross-sectionally.[32] Additionally, racial differences in the modifiable CVH behaviors, like smoking, physical activity, and diet, may be primarily explained by socioeconomic factors.[8] In one study using the CARDIA cohort, socioeconomic status mediated 48.9%–70.1% of the association between race and the CVH health behavior score; psychosocial factors mediated 20.3%–30.0% of the association and neighborhood factors mediated 22.1%–41.4%.[8] Low socioeconomic status throughout life is also associated with poor levels of the CVH component metrics.[33] Early studies demonstrate a link between overall CVH and neighborhood environment, including access to favorable food stores, physical activity resources, walking/physical activity environment, and neighborhood socioeconomic status.[30,34] A clearer understanding of the associations between complex SDOH constructs and CVH may provide helpful information for programs looking to address SDOH and ultimately improve CVH. There is already some evidence that addressing SDOH can improve CVH, as shown by decreased CVD mortality following Medicaid expansion[35] and decreased diastolic blood pressure following disbursement of funds from the Earned Income Tax Credit.[36] **Despite this existing work, data are sparse regarding the associations of the full spectrum of SDOH with overall CVH.**

Current methodological and statistical approaches limit our understanding of how the complex interplay of SDOH exposures may be associated with CVH cross-sectionally and longitudinally.

There are also limitations to the current statistical approaches used to study SDOH and the CVH component metrics. Investigators have traditionally examined socioeconomic status (SES) variables or indices as exposures at one time point in regression-based models. Composite SES indices combine multiple SES variables into a multidimensional concept using principal component analysis.[37–39] These indices simplify analysis and reduce multicollinearity between the SES exposures, but are difficult to interpret because they are unit-less values and the association between the outcome and distinct SES variables can no longer be interpreted. In regard to timing of the exposures, SES variables are often assessed at the same time as the CVH component metrics. If timing of the SES exposures is considered, there are three typical life course study designs: 1) SES exposures measured in early life with the CVH outcome occurring later in the life course, 2) changes in a limited set of SES variables measured over a short time period (often two time points) with the CVH outcome occurring later in life, or 3) summary index of SES events over the life course with the CVH outcome occurring later in life.[33] All three life course designs do not account for many SDOH variables and their patterns of change over a longer time period. Both the cross-sectional and longitudinal methods used previously may be imperfect in their ability to account for the complex relationships between SDOH variables, in addition to the complex relationships between many SDOH exposures over time and overall CVH and its component metrics. For example, social exposures like lower income and lower levels of education are often associated with exposure to poor social networks and social support.[40] Current methods are unable to account for the co-occurrence and collinearity of these variables and their joint relationship with health outcomes over time. Thus, **current methodological approaches used to incorporate SDOH in statistical models limit our understanding of the magnitude and timing of their impact on CVH and CVD.**

1.2 Objectives and Outline

It is unclear how individual and neighborhood-level SDOH, across all five domains, may be associated with each other and may change over time. No single variable or domain fully captures an individual's SDOH, therefore new, and easily interpretable, methods for studying the effect of SDOH over time are necessary.[27,41] In addition, it is unknown how a set of SDOH may perform in the prediction of CVH. An understanding of how multiple SDOH are associated with CVH and CVD events may assist researchers and public health professionals in identifying new targets for intervention on SDOH to improve population-level CVH. **I therefore propose to address these knowledge and methodological gaps and to examine how a**

broader array of SDOH perform in predictive models for CVH, and are associated with mid-life CVH and CVD events over time. The *primary objectives* of this dissertation are to identify patterns of SDOH exposure over time, create SDOH clusters using a novel machine learning method, and determine whether the addition of SDOH variables allows for better prediction of an individual's CVH status and are associated with mid-life CVH (a critical milestone in CVH maintenance) and CVD events. The *primary hypothesis* for this study is that a diverse set of SDOH up to age 45 will be associated with and will improve prediction of CVH, subclinical CVD, and CVD events at mid-life, independent of baseline CVH and other covariates.

To pursue these overall objectives, this dissertation begins with a general introduction highlighting the current disparities in CVD, the need to focus on prevention-based strategies, the current evidence linking SDOH and CVD, and the traditional methods used to study SDOH and CVD. I also provide a conceptual model and highlight the innovative components of this work. I then lead into the three chapters with distinct objectives. In Chapter 2, I present a detailed description of the methods used to identify the frequent time-dependent SDOH patterns and generate the novel SDOH clusters. I also explore whether the SDOH clusters improve the prediction of mid-life CVH using multiple regression-based and machine learning methods. Chapter 3 contains a more detailed description of time-dependent individual- and neighborhood-level SDOH exposure patterns and evaluates whether there is an association between the clusters and mid-life CVH. In Chapter 4, I then evaluate whether the SDOH clusters are associated with mid-life CVD events before and after adjustment for mid-life subclinical

CVD and CVH, in order to understand potential pathways that may link upstream SDOH and downstream health outcomes.

1.3 Conceptual Framework

This dissertation is primarily focused on identifying patterns of SDOH exposures from young adulthood to middle age, generating clusters of the patterns, and then examining whether the clusters are predictive of mid-life CVH, and associated with mid-life CVH and CVD events. Figure 2 depicts the conceptual framework for the pathways linking SDOH, CVH, subclinical CVD, and CVD events, as modified from the WHO Commission on Social Determinants of Health Conceptual Framework.[42] This framework highlights the importance of the relationship between upstream SDOH and CVH and health disparities; the social effects are directly felt by vulnerable populations through stress, knowledge, and time.[43–45] Structural racism is a major SDOH, driver of other SDOH, and a root cause of disease—including CVD.[46–48]

1.4 Innovation

Novel methodology to create and understand impact of SDOH patterns and clusters

Studies of SDOH and their associations with various health outcomes to date have been largely cross-sectional or examined change in a limited set of SDOH over a short period. In this study, we examined patterns of SDOH exposure from all five domains over time and created clusters of these trends. This approach has been used for other health-related topics, such as grouping trends in physiologic variables among patients in the ICU[49] and creating phenotypes of multiple organ dysfunction in children[50], but it has not been used to study SDOH, CVH, and CVD events. We simultaneously incorporated SDOH from all five SDOH domains, which is different from existing approaches that often examine SDOH individually or as a limited set. Additionally, CARDIA is a unique dataset in that it allowed us to examine SDOH across all domains on both an individual and neighborhood level for the same participant. This provided a broader understanding of the SDOH experienced by each participant individually and in the larger environment where he/she lives. By understanding longitudinal SDOH patterns and their clusters at individual and neighborhood levels, we identified directions for potential future work focused on developing timely and multi-component social, public health, and healthcare interventions and policies.

Application of machine learning and simultaneous incorporation of measures from all five SDOH domains to study CVH

Machine learning algorithms have not been widely applied to problems within social science research. This dissertation applied two newer supervised machine learning classifiers and regression models. We were able to compare whether longitudinal SDOH patterns and their corresponding subgroups offer improved prediction of CVH, while also comparing more traditional statistical models with machine learning algorithms as methods for prediction. Additionally, a limited set of SDOH variables have been used to study the seven CVH component metrics individually; however, the incorporation of variables representing all five SDOH domains as predictors for overall

CVH is novel. This provided us with a better understanding of which exposure patterns may be associated with and predictive of CVH, and at what time during young adulthood. Findings from this study may help researchers, clinicians, and policymakers better predict patients and populations at risk for low CVH during middle age.

Identification of SDOH trends that may be associated with and differ by race

Previous work has identified differences in SDOH cross-sectionally and also disparities in overall CVH and CVH behaviors. In this study, we identified longitudinal SDOH exposure patterns and clusters of these trends. We also stratified our predictive models by race. An understanding of how our longitudinal SDOH exposures, clusters, and models varied within race subgroups allowed us to generate hypotheses related to the unique contribution of SDOH to CVH health disparities.

Associations of SDOH exposures with important CVD outcomes

Through the creation of our novel SDOH exposures, we were also able to examine the longitudinal association of a representative set of key SDOH domains and issues with subclinical CVD measures and CVD events. As mentioned, previous work traditionally examined a one or just a few SDOH measures cross-sectionally or used longitudinal statistical methodologies with limitations. We gained an improved understanding of the longitudinal relationship between SDOH and important CVD health outcomes. We also explored whether the association of SDOH with CVD events was maintained after adjusting for subclinical CVD or CVH status. This helped us to determine whether there may be independent pathways through which SDOH influence subclinical or clinical CVD. These findings will be helpful in informing future work addressing health disparities through SDOH.

CHAPTER TWO: APPLICATION OF A NOVEL SEQUENTIAL PATTERN MINING METHOD TO STUDY LONGITUDINAL SOCIAL DETERMINANTS OF HEALTH: THE CORONARY ARTERY RISK DEVELOPMENT IN YOUNG ADULTS (CARDIA) STUDY

2.1 Abstract

Objective: To identify patterns of social determinant of health (SDOH) exposure from young adulthood to middle age, define exposure clusters among the cohort, and evaluate whether the patterns and clusters improve prediction of mid-life cardiovascular health (CVH).

Materials and Methods: We analyzed SDOH data from participants recruited in the longitudinal prospective CARDIA study. Using subgraph augmented non-negative matrix factorization (SANMF), we identified frequent, time-dependent patterns among 48 SDOH and psychosocial variables across four age windows from young adulthood to middle age. We then generated and characterized clusters of the patterns and examined whether they were predictive of mid-life CVH.

Results: Among the 3,522 participants included in the study, we identified 502 frequent patterns of SDOH variables and generated five unique clusters. The clusters incorporated patterns from the five SDOH domains and were interpretable. In predictive modeling using multiple machine learning algorithms, the models incorporating the SDOH patterns and clusters as unique predictors met or slightly exceeded predictive performance of the base models.

Discussion: Using the SANMF method, we were able to generate time-dependent SDOH patterns and clusters that were associated with and predictive of mid-life CVH. The SANMF model improves upon existing methods used to study SDOH by generating interpretable clusters and predictors of multiple, varied SDOH variables over time.

Conclusion: The patterns and clusters reduce the complexity of SDOH exposures and, following replication in other settings, may be used to develop targeted and time-specific social interventions and policies to address adverse SDOH and potentially improve low CVH in individuals and populations.

2.2 Introduction

Over the past decade, there has been increased discussion around the influence of social determinants of health (SDOH) on overall health and clinical outcomes. Traditionally, researchers have focused on individual-level factors, including genetics and access to healthcare, as major contributors to health. While these factors are important, they may not be the main drivers of health.[51–53] In recent years, the 2010 Affordable Care Act, increasing health disparities, and the COVID-19 pandemic have drawn attention to the influence of nonclinical, population-level factors on well-being, and highlighted how social inequities lead to health disparities.[54] These factors, defined by the World Health Organization, are the historical and “structural determinants and conditions in which people are born, grow, live, work, and age” shaped by the distribution of money, power, and resources— known as SDOH.[23] The five key domains of SDOH include economic stability, neighborhood and built environment, education, social and community context, and health and health care.[24] SDOH are the upstream factors that drive institutional inequities, living conditions, risk behaviors, and ultimately disease, morbidity, and mortality.[42,55]

While increasing recognition of these factors is critical to design more targeted health interventions and improve public health, methodological approaches to study SDOH have been limited. Prior studies have generally focused on the cross-sectional association between one or only a few social determinants, typically from the economic stability or education domains, and one health outcome. However, there are complex relationships between diverse SDOH that are not considered when only examining one

SDOH variable within one domain.[51,56] There is also increasing recognition that SDOH change over time and influence health outcomes differently over the life course.[56] It is important to consider which SDOH are most important, and at what time, when assessing their impact on health outcomes. To accomplish this goal, new methods are needed for defining and studying how patterns of SDOH across domains and over the life course impact critical health outcomes.

Cardiovascular disease is a burdensome public health problem and the leading cause of death in the United States.[57] Reducing morbidity and mortality from cardiovascular disease and improving cardiovascular health (CVH) has been a key focus for the past decade.[9] CVH is a broader and more positive construct than cardiovascular disease, and is a multi-component score made up of seven key health factors and behaviors.[9] While there has been a focus on improving CVH, disparities remain and may be due to differences in SDOH.[22] Limited evidence demonstrates an association between a selected set of SDOH and the individual CVH indicators, cross-sectionally and longitudinally.[29,30,46] Despite this existing work, data are sparse regarding the associations and predictive power of the full spectrum of SDOH with overall CVH.

2.3 Objectives

In this study, we aimed to identify time-dependent patterns of SDOH exposure from young adulthood to middle age in the longitudinal Coronary Artery Risk Development in Young Adults (CARDIA) Study using a novel sequential pattern mining method.[49,58] We created and characterized clusters of SDOH exposure patterns with sparse non-negative matrix factorization (NMF), and determined whether the SDOH patterns and clusters improved prediction of mid-life CVH using several analytic approaches.

2.4 Methods

2.4.1 Dataset, Predictors, and Outcome

The Coronary Artery Risk Development in Young Adults (CARDIA) study is a prospective community-based cohort study with detailed information on cardiovascular risk factors and disease in a geographically and racially diverse sample of young adults.[58] The cohort is comprised of 5,115 black and white men and women, ages 18-30 at baseline, recruited from four metropolitan areas: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. Participants, originally recruited in 1985-1986, were enrolled with similar numbers of people within self-identified race (Black or white), sex (female or male), education (high school or less vs. more), and age (18-24 and 25-30) groups at each of the four centers. CARDIA participants have undergone in-person examinations at baseline (Year 0: Y0) and at Y2, Y5, Y7, Y10, Y15, Y20, Y25, and Y30. Retention rates among surviving participants at each in-person examination were 91%, 86%, 81%, 79%, 74%, 72%, 72%, and 71%, respectively, and >90% of participant have been contacted within the last 5 years.

Primary predictor variables included 48 time-dependent SDOH and psychosocial variables collected during the examination cycles. Data regarding both individual- and neighborhood-level variables from all five SDOH domains were collected across seven CARDIA exam years spanning ages 18-45 years. Individual-level SDOH were available for all exams of interest. Neighborhood-level SDOH were available for Year 0 (Y0), Y7, Y10, Y15, and Y20. Table 1 displays the individual-level and neighborhood-level SDOH variables by domain and two psychosocial variables that were also included in the analysis.

The primary outcome for this study was CVH measured at age 45 years or older, as defined by the American Heart Association, measured at age 45 years or older.[9] High mid-life CVH is associated with many favorable cross-sectional and longitudinal outcomes including markedly lower incidence of total mortality, CVD-related mortality and non-fatal events, improved quality of life, and compression of morbidity.[9,10,12,14] The CVH construct incorporates seven metrics of current smoking, physical activity, diet, body mass index, total cholesterol, blood pressure, and fasting glucose. In this analysis, each component metric was categorized as poor, intermediate, or ideal status, as done previously[59,60], assigning either 0, 1, or 2 points, respectively. The status of each metric was then summed to create the overall CVH score and then categorized into low (0-7 points), moderate (8-11 points), and high (12-14 points) CVH. We further dichotomized the outcome into low vs. moderate/high CVH for our classification models.

For this analysis, we only included those participants who reached age 45 or beyond during CARDIA and had a CVH measure at Y20 or later. For our outcome, we also looked at the first CVH measurement at or after age 45. Diet, a key component of the composite CVH metric, was only assessed at Y0, Y7, and Y20. Because diet tracks over time[9], we used Y7 diet for those participants missing a Y20 assessment and/or carried forward the Y20 assessment for those who reached age 45 at the Y25 or Y30 exam.

2.4.2 Overall Workflow

We followed the SANMF process as described in Luo et al.[49] The overall workflow is pictured in Figure 3. There were four primary steps to our analysis: 1) graph construction; 2) frequent subgraph mining; 3) sparse non-negative matrix factorization (NMF); and 4) predictive modeling. The SANMF method specifically encompasses steps two and three and requires the SDOH variables to be represented as graphs, necessitating pre-processing to discretize the time and measurement axes. The primary outputs of SANMF are a set of time-dependent patterns or subgraphs (used interchangeably for this chapter) and clusters representing groups of the subgraphs.

2.4.3 Graph Construction

In order to discretize the time axis, we created four age windows. Instead of using the exam-based CARDIA design for our exposures, we assigned values based on age rather than by exam year to more appropriately reflect the timing of exposures in individuals, as has been done in numerous prior studies.[61,62] This approach allowed us to identify potential critical periods when SDOH may be more predictive of CVH, and to normalize our predictor windows. We used four age windows: 18-24, 25-34, 35-44, and 45 years or older. We focused on SDOH only up to the CVH measurement in the 45 or older age window, in order to avoid issues of reverse causation (protopathic bias). If there were multiple assessments of a variable within an age window, we averaged the values for the continuous variables and took the earliest available value for the categorical variables. For the categorical variables, if the first assessment was missing, we took the next available value.

To discretize the measurement axes, we took different approaches based on the variable type. For the continuous values, we either created categories based on what has been generally accepted in the literature, or created tertiles representing lower, mid-range, and higher exposure categories. We took this approach because many of the variables did not have externally validated categorizations representing lower or higher exposure. For the categorical variables, we created categories based on what other researchers have used or what we determined would aid in interpretability.

We generated graphs for each time-dependent SDOH variable by connecting the adjacent discretized measurement values on the time axis. For the categorical variables and continuous variables coded using external categorizations, we created edges to represent the changes between adjacent nodes. For the categorical variables discretized using the tertile method, we created edges labeled as “up,” “down,” and “same” to encode the directionality of the graphs and changes between age windows.

2.4.4 Identification of Time-Dependent SDOH Patterns with Frequent Subgraph Mining

We created node-edge lists for the 3,522 participants who met our inclusion criteria in CARDIA. We used the Molecular Substructure Miner (MoSS) to identify common subgraphs or patterns in the SDOH variables among the participants.[63] In order to identify as many patterns as possible, we ran MoSS without a complement group or pre-defined frequency threshold.

We filtered out shorter subgraphs contained within larger subgraphs, as described by Luo et al., in order to remove the redundant information.[49] We also only included subgraphs occurring in ≥5% of the participants, which was the chosen threshold based on internal discussion identifying a meaningful prevalence for intervention. Because we had subgraphs of lengths one, two, three, or four age windows, the frequency threshold calculation varied based on which age windows were represented in the subgraph, which is different from other previous applications of SANMF. The numerator of the frequency calculation was always the number of participants with the pattern. The denominator, however, was based on the number of participants with data available in the age window. Because almost half of the participants were enrolled at age 25 and later, there were a substantial number of participants without information in the 18-24 age window. There were1,393 participants with SDOH data available in the 18-24 age window, 3,504 in the 25-34 age window, 3,397 in the 35-44 age window, and 3,522 in the 45 and beyond age window. We used the smallest denominator for the frequency threshold calculation in order to include the most patterns possible, including the most patterns from each age window in the creation of the clusters described below. For example, if the subgraph spanned all age windows, 1,393 was used as the denominator. Following exclusion of the subgraphs, the participant by subgraph matrix was used as the input for NMF.

2.4.5 Creation of Clusters with Sparse Non-negative Matrix Factorization

NMF is a commonly used unsupervised machine learning method to cluster patients or variables to create meaningful groups from a set of high-dimensional data.[64–66] NMF serves as an ideal grouping method for analyzing count data that is non-negative by providing simple and additive clusters, as opposed to k-means or principal components analysis, which do not have the benefit of interpretability.[49]

As presented in Figure 4, SANMF approximates the participant-by-subgraph count matrix, X, the patterns matrix, *(*dimensions of $P\ x\ S$, number of participants by number of subgraphs) into lower ranked matrices W and H.[49] W, the features matrix, is of dimension $P\ x\ S_g$ (S_g is the number of subgraph groups) and H, the coefficient matrix, is of dimension S_g x S. W identifies the participants with exposure to the patterns in each cluster. H identifies the clusters, or the groups of common patterns in the SDOH variables.

We randomly partitioned our cleaned dataset using a 70% training and 30% test set ratio, stratified by the dichotomous CVH outcome. We examined descriptive statistics for the training and test sets to ensure the datasets were similar across the demographic and SDOH and psychosocial variables. Because there is no way to evaluate the clusters independently, we determined the number of clusters by evaluating their utility in predictive models. We assumed useful features will improve predictive performance of our outcome, therefore we chose the number of clusters based on area-under-the-curve (AUC) values from logistic regression modeling using five-fold cross-validation on the training set.

To conduct NMF, we used the projected gradient NMF solver[67] implemented in Scikit-learn.[68] We implemented NMF on the training set enforcing sparsity using the SNMF/R algorithm with Nonnegative Double Singular Value Decomposition (nndsvd) seeding, as has been implemented and described previously.[49,69–71] The SNMF/R algorithm enforces sparsity on H matrix, minimizing the number of non-zero entries or limiting the number of subgraphs with membership coefficients for each cluster. We used 5-fold cross-validation to evaluate the predictive performance of the clusters of size 2 through 10, using logistic regression models, controlling for maximum education level, baseline CVH score, baseline age, sex, CARDIA center, and race. The models included the specified covariates, in addition to the columns in the W matrices for each NMF iteration representing the likelihood of each participant being exposed to the SDOH patterns within each cluster. We chose the minimum number of clusters from the model with the highest AUC value and continued with the analysis using the W and H matrices from the NMF iteration for the chosen number of clusters.

2.4.6 Characterization of SDOH Clusters

We characterized the generated SDOH clusters by identifying the top ten subgraphs by identifying the ten largest membership coefficients within each cluster, in order to maintain interpretability. We did not force each participant into one SDOH cluster based on the highest coefficient in W because we wanted to account for a participant's exposure to multiple clusters and the subgraphs within each cluster. The subgraphs could be one node in length or up to four nodes in length, representing the four age windows under study. The subgraphs were categorized, *a priori*, by domain to characterize the clusters by their unique subgraph makeup and domain.

2.4.7 Prediction of Cardiovascular Health

To determine whether SDOH improve longitudinal prediction of CVH statistically, we used supervised machine learning methods to predict CVH from the identified frequent subgraphs and generated SDOH clusters on the training and held-out test sets. We utilized four predictor groups for modeling: 1) Base model: baseline age, sex, CARDIA center, and race; 2) Base + CVH model: baseline CVH and base model predictors; 3) Base + SDOH Clusters model: base model with all SDOH clusters as independent features and maximum education achieved; 4) Base + CVH + SDOH Clusters model: all SDOH clusters as independent features with Base + CVH model predictors and maximum education achieved; and 5) Base + CVH + Subgraphs model: all frequent subgraphs modeled as independent features with Base + CVH model predictors and maximum education achieved. We included the subgraphs as independent features in model five to further explore the added value of creating more interpretable clusters of SDOH compared with modeling the subgraphs separately. To understand the predictive utility of the SDOH clusters for CVH under diverse statistical models and assumptions, we used logistic, Lasso, and Ridge regression, in addition to machine learning classifiers including random forests and multi-layer perceptron neural networks. All models were tuned on the training set using five-fold cross validation with 10 repeats. We applied the chosen NMF model to the held-out test set prior to application of the final tuned models to create the clusters in the test set for utilization in the Base + CVH + SDOH clusters models. We compared AUC values for all nested models applied to the test set to assess their predictive performance above the base model using Delong's test for two correlated ROC curves.[72] All predictive modeling was conducted using R version 3.6.1.[73] and the caret package.[74]

To assess how each predictor contributed to classification, we examined permutation feature importance[75] using classification error as the loss function on the test set using the iml package.[76] We also examined misclassification rates for the Base + CVH logistic regression model and conducted an error analysis with SHAP plots[77] from the iml package to determine the features driving misclassification.

2.5 Results

There were 3,522 participants included in the study cohort, who were split into a training and test set ensuring an equal distribution of participants in the dichotomous CVH outcome categories as shown in Table 2. Characteristics of the training and test set were similar. Following frequent subgraph mining of the 48 time-dependent SDOH variables, we identified 502 frequent subgraphs of various lengths: 109 patterns spanning one age window, 237 patterns spanning two, 134 spanning three, and 22 spanning four age windows. We ultimately chose five clusters for sparse NMF after comparing AUC values from five-fold cross-validation on the training data with a range of groups sizes: two through ten and 20 through 100 (at increments of ten) groups.

2.5.1 SDOH Clusters

Table 3 presents the top ten subgraphs identified within each SDOH cluster. The clusters can be characterized as follows: Cluster 1) economically stable with less psychologically demanding job, mid-range social support, zero health care access barriers, and no change in residence; Cluster 2) economically stable with employment, mid-range social support, and zero health care access barriers; Cluster 3) some difficulty economically with lower status job, one change in residence during late 20s and early 30s, and low social support; Cluster 4) no difficulty meeting demands, but with vulnerable neighborhood environment, and Cluster 5) economically wealthy with high status job, higher social support, and change in residence during late 20s and early 30s.

2.5.2 Prediction of CVH

The results from predictive modeling are shown in Table 4 with the AUC values for the training and held-out-test sets. Typically for prediction, an AUC of <0.70 is considered inadequate discrimination, 0.70-0.79 is satisfactory, and 0.80-0.89 is excellent.[78] The Base + CVH model offered improved predictive performance over the Base model (p<0.001) and satisfactory discrimination. The Base + SDOH Clusters model improved the AUC by ~0.07 from the Base model (p<0.001) and offered satisfactory discrimination over inadequate discrimination in the Base model. For the Base + CVH + SDOH Clusters models, the logistic regression had the highest AUC value and exceeded performance of both the Base and Base + CVH models, although statistical significance was not achieved for comparison with the Base + CVH model (p<0.001 and p=0.068, respectively). For the Base + CVH + subgraphs models, Lasso regression offered the best performance (satisfactory discrimination) and exceeded performance of the Base + CVH model somewhat, but not statistically significantly (p=0.390).

2.5.3 Important Clusters and Subgraphs

In the Base + CVH + SDOH clusters logistic regression model, 0.10-unit greater likelihood of being exposed (i.e., 0.10 SD greater likelihood of being exposed) to the subgraphs in Clusters 1, 2, and 5 were associated with lower odds of low mid-life CVH (adjusted Odds Ratio, 95% confidence interval: Cluster 1: 0.71, 0.60 – 0.83; Cluster 2: 0.83, 0.71 – 0.97; Cluster 5: 0.75, 0.65 – 0.85). In Figure 5, the variable importance plots from the SDOH + CVH + Clusters logistic regression model highlight the predictive power of the SDOH clusters. Clusters 5, 2, and 3 offered improved predictive performance over some of the traditional predictors such as sex and age, and all of the clusters offered improved predictive performance over self-identified race. In the Base + CVH + Subgraphs Lasso regression model, seven subgraphs were identified in the top ten most important variables (Figure 6). SDOH from all domains and one psychosocial variable improved predictive performance.

The Base + CVH + SDOH clusters logistic regression model offered the best predictive performance, correctly classifying 772 participants (75.5%), but it misclassified 251 participants when applied to the test set; there were 184 participants with low mid-life CVH who were predicted as having moderate or high mid-life CVH status and 67 participants with moderate or high mid-life CVH who were predicted as having low mid-life CVH. Among the misclassified participants, 61.8% were Black, which is higher than the prevalence of self-identified Black participants in the test set (46.6%). When examining SHAP plots for the misclassified participants, baseline CVH score was the primary driver of prediction.

2.6 Discussion

By applying the novel SANMF method to complex longitudinal SDOH data, we were able to identify frequent patterns of SDOH variables and generate five clusters of the time-dependent patterns. The subgraphs and clusters incorporated into predictive models for CVH met and slightly exceeded performance above the base model without the SDOH predictors, offered satisfactory discrimination, and better represented the social exposures of the participants. We characterized and identified the important clusters associated with CVH and patterns and clusters which improved classification of the dichotomous CVH outcome. We were able to distinguish subgraphs and clusters offering improved predictive performance of the CVH outcome, including those from all five SDOH domains and psychosocial factors. In logistic regression modeling, three out of the five clusters were associated with mid-life CVH.

This study represents the first time the SANMF method has been applied to either the field of SDOH as an exposure or CVH as an outcome. SANMF has been previously used in other longitudinal scenarios. Luo et al.[49] used SANMF to group patients admitted to the intensive care unit based on their temporal trends in multiple physiologic variables, and Sanchez et al.[50] created phenotypes of multiple organ dysfunction in children. These use cases focused on stratifying patients into groups based on similar types of exposures. In our current use case, we further widened the applicability of the method by developing clusters using 48 categorical (nominal and ordinal) and continuous individual- and neighborhood-level SDOH variables, the first extension of this method to social and epidemiological science fields. We also did not assign participants exclusively to one SDOH exposure cluster, but rather kept the focus on the SDOH exposure clusters—emphasizing that participants could be exposed to multiple clusters—and included the clusters as unique predictors in modeling.

SANMF offers a novel approach for characterizing a diverse set of longitudinal SDOH exposures from all five domains. Most traditional analyses focus on singular SDOH, like poverty or education, which limits our understanding of the complex social experiences of individuals and populations. If multiple SDOH are considered, they are often represented as a composite index of socioeconomic status (SES) measures at one time point in regression-based models.[37–39] These indices simplify analysis and reduce multicollinearity between the SES exposures, but are difficult to interpret because they are unit-less values and the association between the outcome and distinct SES variables can no longer be interpreted. In regard to timing of the exposures, SES variables are often assessed at the same time as a health outcome and not longitudinally.[33] SANMF allows for the analysis of longitudinal SDOH, from all five domains and psychosocial factors, and illuminates the intersectionality SDOH factors. The application of SANMF to SDOH data represents a step towards furthering our understanding of the multiple, complex, and intersectional nature of SDOH in association with CVH across the life course and points the way to its use in the context of other health outcomes in the future.

Baseline CVH offered the greatest improvement in predictive performance for mid-life CVH over the base model of demographic variables, with additional significant, although modest, improvements in prediction of mid-life CVH after the addition of the

SDOH clusters. We judge that there may be a few reasons why the SDOH predictors did not offer greater improvement in the overall AUC value. The improvement in the Base model after addition of CVH is not surprising; in biology and epidemiology, we often observe that the best predictor of a future outcome is the baseline value, rather than more proximal values.[79,80] It is important to note, however, that baseline CVH was assessed when the study participants were between the ages of 18 and 30 and is determined by several factors, including childhood SDOH exposures, prior to observation in CARDIA.[81,82] This implies that future work should also focus on these childhood exposures and their relationship with CVH throughout the life course. The association between SDOH and CVH may also be primarily mediated, although not fully, by the health factors and behaviors that make up the composite CVH score.[43–46] We used a composite measure of CVH as our outcome instead of modeling the individual health behaviors and factors as outcomes in separate models. SDOH may offer improved prediction of certain CVH components over others, which should be examined in future work.

Measurement of the predictors and use of the predictive models may be other reasons why we did not see large improvements in the AUC values after addition of the SDOH clusters. The SDOH captured in CARDIA are also primarily self-reported by the participants. Because of this, the precision of the measurements may be lower than expected, which would introduce noise into the models. In addition, there were significant associations between Clusters 1, 2, and 5 and mid-life CVH, but relatively large and precise measures of association for novel variables are needed to generate meaningful changes in AUC values with addition of those new variables.[83] In clinical risk prediction models, novel predictors rarely make large improvements in classification for those at low or high risk from the base model. New predictors, however, are often most useful for people in the intermediate risk category from baseline models when using a sequential screening approach.[84,85] For example, we may use the Base + CVH model to generate a pretest probability to identify participants at low, intermediate, and high-risk for low mid-life CVH, and subsequently use the clusters to reclassify the risk of the participants at intermediate risk of low mid-life CVH. Separate from the sequential testing approach, we may need to create stratified models for the prediction of mid-life CVH within race groups based on the error analysis and higher rates of misclassification among the Black participants.

There are limitations to this work and the study cohort used. CARDIA is a unique dataset with a population of Black and white men and women and detailed information on individual- and neighborhood-level SDOH factors over time. Data capturing exposures to the SDOH factors collected in CARDIA may not always be available in other settings or within the chosen age windows, limiting the generalizability of the clusters. Because the effects of SDOH are generally felt through stress, knowledge, and time, wider age windows were best suited for this analysis, but may not be applicable to other cross-sectional or short-term cohort studies. This work also first focused on identifying frequent patterns of individual SDOH variables and then grouped the patterns into clusters. We did not examine fully the complex relationships between the patterns, which should be assessed in future work. We also selected five clusters by comparing AUC values to models with a different number of clusters and proceeded with all subsequent analyses using five clusters. Future models, examining the association of SDOH with other health outcomes, may benefit from a examining a different number of clusters.

2.7 Conclusion

Using SANMF, we identified time-dependent patterns of 48 SDOH variables and generated five SDOH exposure clusters. Predictive models for CVH incorporating SDOH patterns and clusters as predictors met or exceeded the predictive performance of our Base model, while also improving representative power. The SDOH patterns and clusters that improved predictive performance may assist clinicians and public health professionals in developing targeted and timely interventions for groups and populations at risk for having low CVH in mid-life. Further work is needed to validate the SDOH clusters in cohorts assessing CVH and other health outcomes. An understanding of which social determinants cluster together may help in making inferences about SDOH exposures observed among new study cohorts and their relationship with other social exposures. The clusters may also be used in the prediction of other health outcomes, helping to target those at risk of disease and to minimize health disparities.

CHAPTER THREE: NOVEL TIME-DEPENDENT CLUSTERS OF SOCIAL DETERMINANTS THROUGH YOUNG ADULTHOOD AND ASSOCIATIONS WITH MID-LIFE CARDIOVASCULAR HEALTH: THE CARDIA STUDY

3.1 Abstract

Importance: Social determinants of health (SDOH) may be important contributors to substantial disparities in mid-life cardiovascular health (CVH).

Objective: To identify frequent, time-dependent individual- and neighborhood-level SDOH exposure patterns among young adults up to age 45, and to examine associations of SDOH exposure clusters with mid-life CVH.

Design: Prospective cohort study.

Setting: The CARDIA study recruited young adult participants from Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota; and Oakland, California in 1985-86.

Participants: 3,522 Black and white men and women ages 18-30 at enrollment who had CVH measured at age 45 or older.

Exposures: Individual- and neighborhood-level SDOH exposure clusters through young adulthood, generated via sequential pattern mining and non-negative matrix factorization.

Main Outcomes and Measures: The primary outcome was poor (vs. moderate/high) CVH, a composite metric made up of seven health factors and behaviors, at age 45 or older. Logistic regression analysis was used to adjust for baseline age, CVH score, race, sex, and study site.

Results: Among 1,632 Black and 1,890 white participants, mean age at baseline was 25.3 years and 55.8% were female. From 48 SDOH variables measured repeatedly over time, we observed 502 frequently occurring, time-dependent SDOH patterns. Using machine learning methods, five data-driven clusters of those patterns were identified and characterized. In the multivariable model, Cluster 1 (economically stable/less psychologically demanding job/mid-range social support/no health care access barriers/no change in residence), Cluster 2 (economically stable/mid-range social support/no health care access barriers), and Cluster 5 (economically wealthy with high status job/higher social support/change in residence during late 20s and early 30s) were each associated with lower odds of poor mid-life CVH: Cluster 1 (adjusted odds ratio 0.76, 95% confidence interval 0.67 – 0.87), Cluster 2 (0.78, 0.64 – 0.94), Cluster 5 (0.74, 0.65 – 0.84).

Conclusions and Relevance: We identified five clusters of time-dependent SDOH patterns, defining novel patterns of social determinants that co-vary through young adulthood and are associated with mid-life CVH. Patterns related to food security, health care access, and ability to pay for medical care were common across the five

clusters, indicating these SDOH may be especially important for targeted intervention to preserve CVH.

3.2 Introduction

In 2010, the American Heart Association defined a novel construct of "cardiovascular health" (CVH),[9] which is a composite measure determined by seven health factors and behaviors—smoking, body mass index, physical activity, diet, total cholesterol, blood pressure, and fasting plasma glucose—that can be measured for individuals and populations. CVH generally declines throughout the life course[57,86], but maintaining high CVH into mid-life has been associated with numerous highly favorable health outcomes.[10,12–14,16–18,20] Despite the recent focus on improving CVH in all Americans, major racial disparities remain and are not well understood.[22,87] For example, non-Hispanic Black adults have the lowest prevalence (12%) of meeting ≥5 of 7 criteria for ideal CVH, compared with 26% of Asian adults, 20% of Non-Hispanic white adults, and 13% of Hispanic adults.[57]

Social determinants of health (SDOH) likely contribute to persistent CVH disparities across self-reported race groups. SDOH are defined as the "structural determinants and conditions in which people are born, grow, live, work, and age" that affect health, functioning, and quality of life.[23] SDOH have commonly been stratified into five key domains: economic stability, neighborhood and built environment, education, social and community context, and health and health care.[24] Recent National Heart, Lung, and Blood Institute and American Heart Association statements highlight the importance of understanding and addressing SDOH to improve CVH.[28,46] There are known cross-sectional associations between individual economic stability or education variables and individual CVH health factors.[29,32,46] There is also some evidence that racial differences in modifiable CVH behaviors may be primarily mediated by socioeconomic factors.[8] There may also be a link between neighborhood environment and overall CVH.[30,34] However, it is unclear how the complex interplay of individual- and neighborhood-level SDOH exposures across all five domains changes across early life and how these patterns of exposure may influence mid-life CVH. Furthermore, the types of SDOH that impact CVH may vary by race and ethnicity, and more work is needed to understand how SDOH vary by race and are associated with CVH.[88]

3.3 Objectives

In the longitudinal, biracial Coronary Artery Risk Development in Young Adults (CARDIA) Study,[58] we elucidated and characterized time-dependent clusters of SDOH exposure using novel, unsupervised machine learning methods to address the complex covariate structure among SDOH factors over time. We further assessed associations of SDOH clusters through young adulthood with mid-life CVH overall and by self-identified race groups.

3.4 Methods

3.4.1 Study Design and Population

CARDIA[58] recruited 5,115 Black and white men and women aged 18 to 30 years at baseline from four metropolitan areas: Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota; and Oakland, California. Participants were enrolled in 1985-1986, with a balance of participants across education, age, and self-identified race and sex groups at each center. The study collected detailed information on health behaviors and cardiovascular risk factors, in addition to SDOH, at nine exams over thirty years (baseline/year 0 (Y0), Y2, Y5, Y7, Y10, Y15, Y20, Y25, and Y30) with high levels of retention at each in-person exam (e.g., >71% participation among surviving participants at Y30). Contact is maintained with participants via telephone, mail, or email every 6 months, with annual interim medical history ascertainment. Since 2016, >90% of surviving cohort members have been directly contacted, and follow up for vital status is virtually complete. Each participant provided written informed consent at each study visit, and the institutional review boards at each site approved the study annually.

3.4.2 Primary Exposures

SDOH Variables. The primary exposures included 48 individual- and neighborhood-level SDOH variables from all five domains and psychosocial factors (Table 1) collected throughout CARDIA's nine in-person exam cycles (eTables 1 and 2). In order to reflect exposures in the context of age, we used the rich repeated measures of SDOH and CVH across time to assign values based on age at measurement; this provides a more appropriate means for representing exposures and outcomes that are related to age or have critical periods of exposure.[61,62] Four age intervals were used from young adulthood to middle age: 18-24, 25-34, 35-44, and up to the determination of the CVH outcome of interest (age 45 years or older). SDOH variable values were categorized as listed in Table 1. Categorical variables were coded at each age interval based on clinical thresholds or consensus among the authors for improved interpretability. Continuous variables were split into tertiles representing lower, mid-range, and higher exposure levels. Exposures were examined as time-dependent patterns (across the age intervals) and grouped into exposure clusters, as described below.

*Generation of SDOH Clusters***.** We used a novel method of machine learning, subgraph-augmented nonnegative matrix factorization[49] (SANMF) (Figure 7), to generate data-driven clusters of SDOH from young adulthood to middle age. The SANMF process occurred in two parts: first, sequential pattern mining was used to identify frequently occurring (≥5% of the cohort exposed) time-dependent SDOH patterns. Next, non-negative matrix factorization (NMF) was used to group SDOH exposure patterns into clusters, as has been done previously.[49,64,69,70] We chose the number of clusters by comparing area-under-the-curve values for models including two through ten clusters, baseline CVH, and a set of covariates (described in detail below). As described in Figure 7, the resulting NMF output was two matrices. W, the features matrix, contained the likelihood values for each participant of being in each cluster. Participants were not forced into one cluster and could have likelihood values for multiple clusters. We standardized cluster likelihood values using z-scores and these values were used to represent a participant's likelihood of being exposed to the SDOH patterns in each cluster. Thus, the focus of analysis was SDOH cluster exposures, not individual participant assignment to a given cluster. This allowed us to control for a participant's exposure to multiple clusters in regression modeling. H, the coefficient matrix, contained the membership coefficients for each SDOH pattern in each cluster. Characterization of the SDOH clusters is described in greater detail in the Supplement.

3.4.3 Primary Outcome

The primary outcome of interest was CVH status, as defined by the American Heart Association,[9] at age 45 years or older. CVH is defined by seven component metrics including current smoking, body mass index, physical activity, diet, total cholesterol, blood pressure, and fasting plasma glucose. We assigned each metric 0, 1, or 2 points for poor, intermediate, and ideal levels of each metric (eTable 3), as in prior studies.[59,60] We then summed these points to create the overall CVH score (range, 0-14 points), and defined levels of poor (0-7 points), moderate (8-11 points), and high (12-14 points) CVH. For this analysis, the primary outcome was defined as poor CVH (vs. moderate/high CVH). If there were multiple CVH measurements after age 45, we used the first assessment because CVH status at age 45 is an important surrogate marker for future health and longevity, regardless of later changes.[89] Because full dietary history data needed to characterize CVH were only available at Y0, Y7, and Y20, we restricted our sample to those with mid-life CVH measured at Y20 or later. If the participant reached age 45 or beyond at Y25 or Y30, the diet data were carried forward from Y20 to the later exams. Additionally, for those without a diet assessment at Y20, we carried forward diet from Y7 because diet score is known to track over time.[9]

3.4.4 Statistical Analysis

All analysis was conducted using R version 3.6.1.[73] To standardize the participants' likelihood values of being in each cluster, we created z-scores. We performed bivariate logistic regression analysis to examine the unadjusted association of the clusters with the dichotomous CVH outcome measured at age 45 years or after. We performed multivariable logistic regression adjusting for covariates with a two-tailed P value of <0.05 to determine statistical significance.

We evaluated the association of SDOH clusters with CVH in mid-life. Each cluster was incorporated as a continuous variable represented as a z-score likelihood value for each participant being exposed to the SDOH patterns in each cluster. We assessed models with different sets of predictor variables: 1) unadjusted models with each cluster evaluated separately; 2) Base models: each cluster separately, plus maximum education level, baseline (Y0) age, sex, CARDIA center, and race; 3) Base + CVH models: base models, in addition to baseline (Y0) CVH score, as a secondary analysis to assess whether baseline CVH attenuated the individual associations of the SDOH clusters with mid-life CVH; 4) Full models: all clusters, maximum education level, baseline CVH score, baseline age, sex, CARDIA center, and race, to determine how the individual cluster associations were attenuated after adjustment for the other four clusters; and 5) Full models (removing race as a covariate) stratified by race to assess the association between the clusters, created among all participants, and CVH within race groups. We also examined the coefficient matrix to better understand the likelihood of exposure to the patterns in each cluster in race and sex groups, defined by having a standardized likelihood value greater than one standard deviation (SD) above the mean.

In sensitivity analyses, we included average Center for Epidemiological Studies-Depression (CES-D) values across all age intervals as a covariate because psychosocial variables such as depression score did not achieve inclusion in any of the clusters. We also examined the Pearson correlation coefficients among each of the standardized likelihood values for each cluster. Finally, we included cluster pairs, triplets, and quadruplets in the Base + CVH model to better understand how correlation between the clusters may have affected our outcomes.

3.5 Results

3.5.1 Study Sample

There were 3,522 participants in the study sample (eFigure 1), including 1,632 (46.3%) Black and 1,890 (53.7%) white participants. There were modest differences between the participants excluded and the final cohort included for analysis (eTable 4). Excluded participants were more likely to be Black, have lower baseline CVH, and less favorable SDOH. Participant demographic information, education, and CVH measures are presented in Table 5 and eFigure 2 for all participants and stratified by self-reported race groups. Among all participants, the mean age at baseline was 25.3 years and 55.8% were female. For the entire cohort, there were 30.4% participants in the high CVH category at baseline, falling to 13.7% at age 45 and beyond. There were fewer Black than white participants in the high CVH category at baseline (18.1% vs. 40.9%), and this disparity persisted at mid-life (4.4% vs. 21.7%). As demonstrated in eFigure 3, there were also differences in the CVH component metrics by race at baseline and outcome assessment. There were also differences in time-dependent SDOH variables between the Black and white participants (eTables 5 through 8), further underscoring the need for stratified analyses.

3.5.2 Characterization of Clusters

For each of the five SDOH clusters identified in the total cohort, the general description and the top 10 time-dependent SDOH patterns within each cluster are depicted in Figure 8. Psychosocial measures did not load highly into any cluster because there were only two variables in the psychosocial domain and they were not prevalent among the full cohort. There were three SDOH patterns that appeared in multiple clusters: high food security (Clusters 1, 2, and 5), zero health care access barriers (Clusters 1, 2, 4, and 5), and no difficulties paying for medical care (Clusters 1, 2, and 5) from age 35 years until CVH assessment.

3.5.3 Associations of SDOH Clusters with Mid-Life CVH

When examining each cluster separately in unadjusted and adjusted models (eFigure 4), Cluster 1 and Cluster 5 were significantly associated with lower odds of poor CVH in mid-life in the Base + CVH models (adjusted Odds Ratio [aOR], 95% confidence interval [CI]): Cluster 1 (0.83, 0.75 – 0.92) and Cluster 5 (0.76, 0.68 – 0.85)). Cluster 3 was significantly associated with higher odds of poor CVH in mid-life (1.37, 1.26 – 1.49). The aOR represent the odds of poor CVH in mid-life per one standard deviation (SD) difference in the likelihood of being exposed to that cluster. In other words, there were 17% and 24% lower odds of having poor CVH in mid-life associated with a one-SD higher likelihood of being exposed to Cluster 1 and 5 SDOH patterns, respectively. Conversely, the odds of having poor CVH in mid-life were 1.37 times higher for a one-SD higher likelihood of being exposed to Cluster 3 SDOH patterns.

When examining the aOR from the full model with all five clusters, the association between Cluster 3 and mid-life CVH was attenuated after adjustment for the other clusters, and Cluster 2 became significantly associated with lower odds of poor mid-life CVH in the full model (Table 6).

We also evaluated the five SDOH clusters created among all participants in models that included only white or Black participants (Table 7). Among the Black participants, Clusters 1, 2, 4, and 5 were significantly associated with lower odds of poor mid-life CVH. Among the white participants, Clusters 1 and 5 were significantly associated with lower odds of poor mid-life CVH. The prevalence of exposure to the patterns in each cluster were different by race and sex groups (eTable 9). For example, Black participants were somewhat more likely to be exposed to Cluster 3 patterns than white participants, and white participants were more likely to be exposed to Cluster 1 and 5 patterns.

In our sensitivity analyses, results remained consistent. Adjustment for CES-D score did not change overall associations substantially in overall or race-specific analyses. Correlation coefficients between the clusters were generally modest, with most ≤0.31, and two somewhat higher (Clusters 1 & 2: -0.52, Clusters 3 & 5: -0.38). We also did not see qualitative evidence of first-order or higher-order interactions between the clusters (eFigures 5, 6, and 7) when analyzing cluster pairs, triplets, and quadruplets in the full model.

3.6 Discussion

Using novel methods to account for complex SDOH exposure patterns, we observed five unique time-dependent clusters through young adulthood among CARDIA participants. The clusters encompassed variables from different SDOH domains across the life course. There were significant associations between several clusters and mid-life CVH. The strength of the associations varied modestly when evaluated in race-specific models, and we may be underpowered to assess significant differences in the effect of SDOH on CVH across race groups.

The observed clusters represent SDOH patterns that were frequent (≥5% of participants) and tended to occur together. We observed that economic stability, higher social support, and increased health care access often occur in tandem, in addition to the converse—lower economic stability with lower social support and decreased health care access. There were SDOH patterns that consistently appeared in the five clusters because of their high prevalence (patterns occurred in 60-64% of participants), including high food security, zero health care access barriers, and no difficulties paying for medical care from age 35 years to the participant’s CVH outcome. We also observed a cluster (Cluster 4) primarily comprised of variables indicating a vulnerable neighborhood environment. This finding highlights how vulnerable neighborhood environments may need additional resources across a variety of SDOH domains including education, income, employment, and overall investment to increase housing unit values. It also draws attention to a potential lack of social mobility among some participants; there were participants who lived in consistently vulnerable neighborhoods throughout young adulthood.

All of the clusters were significantly associated with mid-life CVH—except for Cluster 4 (the neighborhood-focused cluster)—in either unadjusted or full models. Clusters 1, 2, and 5 are made up of SDOH patterns that confer benefit to the participants (higher economic stability, mid-range to high social support, and no health care access barriers). These clusters were associated with lower odds of poor mid-life CVH. Cluster 3, consisting of SDOH patterns that may cause vulnerability for participants (lower occupation status, some economic difficulties, lower social support, and health care access barriers), was associated with higher odds of poor mid-life CVH. The directionality of the cluster associations with CVH is consistent with previous research linking lower socioeconomic status, lower social support, and health care access barriers with higher CVD risk factors and disease.[46,90–92]

Our results suggest that clusters of SDOH may act differently within race groups, but this finding needs to be explored further. The neighborhood cluster (Cluster 4), when evaluated in race-specific models, was significantly associated with poor mid-life CVH among Black but not white participants, suggesting that neighborhood-level factors may be more of a contributor to poor CVH among Black than white participants. There were also differences in strengths of association with mid-life CVH by race. Cluster 1 was associated with lower odds of poor mid-life CVH among both race groups, but the beneficial association was stronger among Black participants. Effect estimates do overlap among Black and white participants, indicating the clusters created among all participants are appropriate for characterizing SDOH exposures among both Black and white populations.

Our study is novel and has numerous strengths because of the distinct nature of the CARDIA dataset and our methodological approach. CARDIA contains longitudinal SDOH data across all five SDOH domains encompassing a critical period for loss of CVH. SDOH are also assessed both on an individual and neighborhood level for the same set of participants. Additionally, the longitudinal nature of the study and biracial study cohort allow for the examination of the effects of SDOH over time and within Black and white participants. Our methodological approach allowed us to leverage the complexity of CARDIA's data structure to identify frequent, time-dependent SDOH exposures, with 48 variables across all five SDOH domains, and cluster them among the full cohort to examine their associations with CVH overall and by race. Existing methods that examine many SDOH variables together often do so by creating composite scores, which are limited in their ability to examine the association of distinct SDOH variables and their longitudinal patterns with the outcome of interest.[37–39] We did not restrict each participant to one cluster, emphasizing the fact that different individuals can be exposed to multiple types of SDOH clusters. In our analysis, the clusters contained variables that were most frequent and occurred concurrently, which is beneficial for identifying SDOH factors that may be targeted for improvement. Additionally, by first generating the clusters and later assessing their association with CVH, the SDOH clusters can be used to assess a variety of health outcomes.

Our study does have limitations. The data from CARDIA are limited to two race groups—Black and white participants—from four geographic centers, which may limit the generalizability of our findings. Although CVH at age 45-50 years is an important surrogate marker for longevity and healthy longevity during the remainder of the life course,[89] we used only the first available measure of CVH at age 45 and beyond, which may limit our findings. Excluded participants had greater prevalence of adverse SDOH, which may indicate we are underestimating true associations with our findings. We were also limited to the SDOH that were collected in CARDIA. Half of the participants did not have SDOH information in the 18-24 years age window because they were recruited at baseline between the ages of 25 and 30. Our approach, and algorithm applied, focused on examining longitudinal patterns of each SDOH variable. We did not look at the interplay between the SDOH variable patterns and how these changed over time, which should be considered in future work.

3.7 Conclusions

In this study, we used novel methods to elucidate, characterize, and evaluate five novel SDOH clusters not previously described, made up of distinct, data-driven time-dependent SDOH patterns. These clusters, representing changes in SDOH exposure from young adulthood to middle age, were associated significantly with mid-life CVH, which is itself an important indicator of remaining longevity and health. Our approach is novel in its ability to identify longitudinal SDOH patterns across all five domains and define exposure clusters that have an association with CVH. Further work is needed to validate these clusters in other settings and explore potential causal pathways linking SDOH and CVH, especially within racial groups.

Ultimately, these data may inform efforts to target improvements in SDOH at critical periods in order to increase CVH across the life course and reduce the burden of disparities in major health outcomes. We observed frequent patterns related to food security, health care access, and ability to pay for medical care in our clusters. These findings would support natural experiments and policy interventions examining effects of social policies related to these patterns like increasing minimum wage, expanding eligibility criteria for the Supplemental Nutrition Assistant Program, and working towards universal healthcare coverage and greater access to quality longitudinal health care.

3.8 Supplement

Methods Appendix

Characterization of SDOH Clusters. In order to characterize the clusters, we examined the top ten time-dependent SDOH patterns with the highest membership coefficients within each cluster. In the NMF results, SDOH patterns were not restricted to membership in only one cluster, and patterns may span only one or up to four of the age intervals across young adulthood.

After performing frequent subgraph mining on the 48 SDOH variables and excluding the rare and redundant patterns, we observed 502 time-dependent SDOH patterns occurring in ≥5% of participants. There were 109 variables representing exposures from one age interval, 237 from two age intervals, 134 from three age intervals, and 22 from all four age intervals. We performed sparse NMF on the total cohort using the counts of the 502 time-dependent SDOH variables as the input, which resulted in the five main exposure clusters, using methods as described previously. [1–3]

eTable 1: Timing of assessment of individual-Level SDOH variables from CARDIA, by domain and exam

SDOH Domains and Issues	Y0	Y2	Y5	Y7	Y10	Y15	Y20	Y25	Y30
Education									
Education	X	X	X	X	X	X	X	X	X
Economic Stability									
Income			X	X	X	X	X	X	X
Home Ownership			X	X	X	X	X	X	X
Hard to Meet Demands	X	X							
Hard to Pay For Basics	X	X		X	X	X	X	X	X
Trouble Making Ends Meet	X	X							
Hard to Pay For Medical Care					X	X	X	X	X
Assets						X	X	X	X
Debt						X	X	X	X
Food Security						X	X	X	X
Employment Status	X	X	X	X	X	X	X	X	X
Occupation Status- TSEI	X	X	X	X	X	X	X		
Karasek Job Strain Questionnaire		X			X				
Social And Community Context									
Household Size	X	X	X	X	X	X	X	X	X
Children	X	X	X	X	X	X	X	X	X
Marital Status	X	X	X	X	X	X	X	X	X
Social Support Questionnaire	X	X							
Discrimination				X		X		X	X
Social Network						X	X	X	
Subjective Social Standing						X	X		
Social Support and Conflict Questionnaire						X	X		
Neighborhood and Built Environment (Self-Reported by Participant)									
Change in Residence	X	X	X						
Neighborhood Cohesion						X	X		X
Neighborhood Environment							X		X
Health and Health Care									
Health Care Barriers				X	X	X	X	X	X
Health Insurance Coverage				X	X	X	X	X	X
Psychosocial									
Material and Psychological Wellbeing (CES-D)					X	X	X	X	X
Chronic Burden						X	X	X	

Abbreviations: CES-D = Center for Epidemiologic Studies Depression Scale; SDOH = Social Determinants of Health; TSEI = Total-based Socioeconomic Index;

eTable 2: Neighborhood-level SDOH variables assessed in CARDIA[a]

Percent Population White Race
Percent Population Education < High School
Percent Population <150% Federal Poverty Level
Median Income
Percent Population Professional/Management Occupation
Percent Population Unemployed
Median Rent
Percent Owner-Occupied Housing Units
Percent Vacant Housing Units
Aggregate Value Housing Units
Racial Segregation (Gi* statistic)
SES Deprivation Score
Fast Food and Convenience Stores
Supermarkets
Physical Activity Facilities

[a] All Neighborhood-Level information was available during exams Y0, Y7, Y10, Y15, and Y20

Abbreviations: SDOH = Social Determinants of Health;

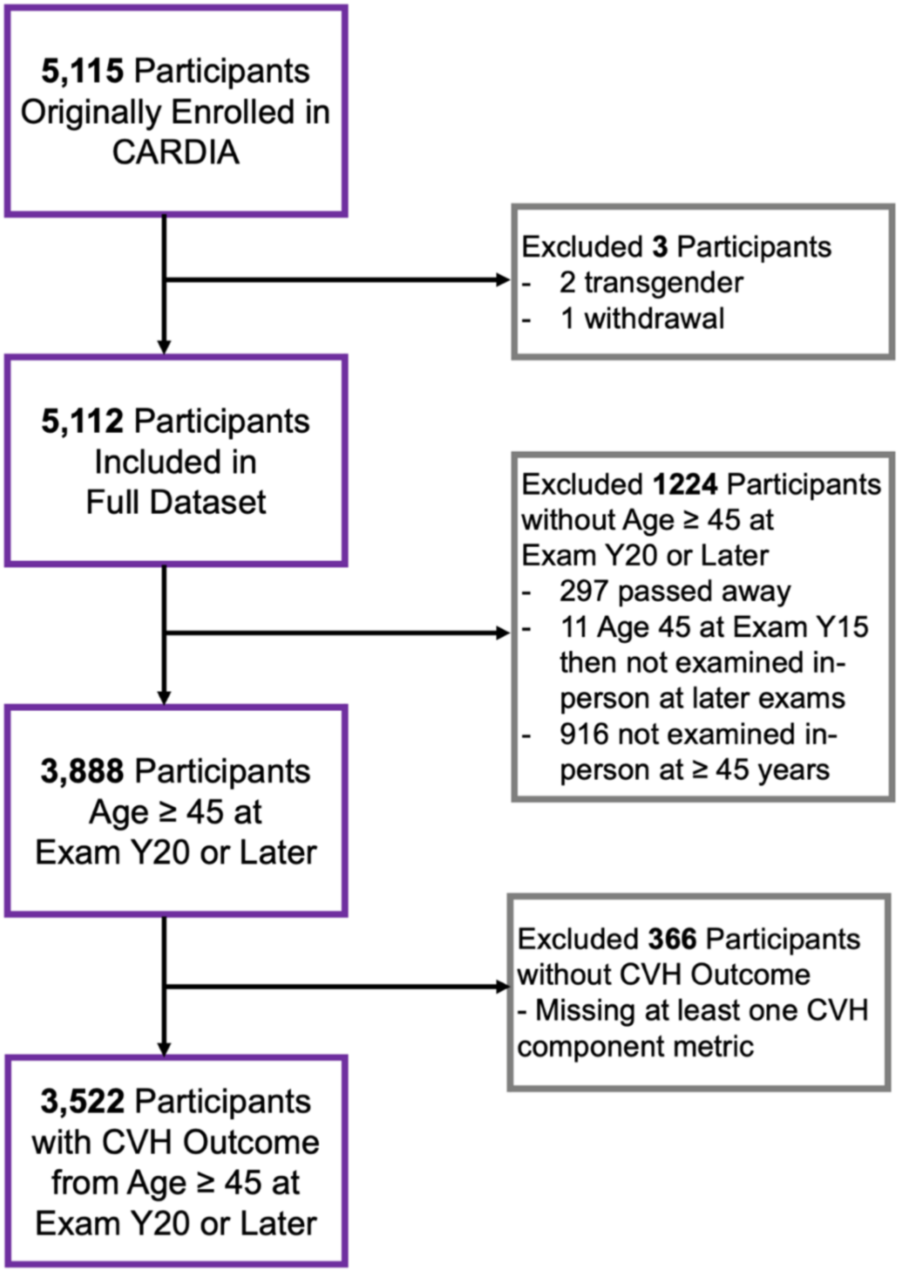
eTable 3: Definitions of poor, intermediate, and ideal for cardiovascular health seven component metrics

	Poor (0 points)	**Intermediate (1 point)**	**Ideal (2 points)**
Current smoking	Yes	Former ≥12 months	Never or quit >12 months
Body mass index	≥30 kg/m^2	25–29.9 kg/m^2	<25 kg/m^2
Physical activity	None	1–149 min/wk moderate or 1–74 min/wk vigorous or 1–149 min/wk moderate + 2x vigorous	≥150 min/wk moderate or ≥75 min/wk vigorous or ≥150 min/wk moderate + 2× vigorous
Healthy diet pattern, No. of components (AHA diet score)[a]	<2	2–3	4–5
Total cholesterol, mg/dL	≥240	200–239 or treated to goal	<200
Blood pressure	SBP ≥140 mm Hg or DBP ≥90 mm Hg	SBP 120–139 mm Hg or DBP 80–89 mm Hg or treated to goal	<120 mm Hg/<80 mm Hg
Fasting plasma glucose, mg/dL	≥126	100–125 or treated to goal	<100

Abbreviations: AHA = American Heart Association; DBP = Diastolic Blood Pressure; SBP = Systolic Blood Pressure;

[a] Diet based on 5 components: consume ≥4.5 cups/day of fruits and vegetables, ≥2 servings/week of fish, and ≥3 servings/day of whole grains and no more than 36 ounces/week of sugar-sweetened beverages and 1500 mg/day of sodium

Adapted from Lloyd-Jones et al.[1]


5,115 Participants Originally Enrolled in CARDIA
Excluded **3** Participants
- 2 transgender
- 1 withdrawal
5,112 Participants Included in Full Dataset
Excluded **1224** Participants without Age ≥ 45 at Exam Y20 or Later
- 297 passed away
- 11 Age 45 at Exam Y15 then not examined in-person at later exams
- 916 not examined in-person at ≥ 45 years
3,888 Participants Age ≥ 45 at Exam Y20 or Later
Excluded **366** Participants without CVH Outcome
- Missing at least one CVH component metric
3,522 Participants with CVH Outcome from Age ≥ 45 at Exam Y20 or Later

eFigure 1: STROBE Diagram outlining cohort for inclusion. Abbreviations: CVH = Cardiovascular Health;

eTable 4: Characteristics of included and excluded participants: demographics, cardiovascular health measures, and select SDOH variables overall and by cohort

	Included - Participants with CVH Outcome (CVH)	**Excluded - Participants without CVH Outcome (NoCVH)**	**Excluded - Participants without Age ≥45 at Exam Y20 or Later (No45)**
	N=3,522	N=366	N=1,224
Race			
Black	46.3%	65.0%	62.7%
Sex			
Female	55.8%	61.5%	48.4%
Mean Age at Baseline (years, SD)	25.3, 3.5	24.7, 3.7	23.7, 3.7
Mean Age at Outcome Measurement (years, SD)	48.1, 2.3	49.9, 3.5	--
Cardiovascular Health at Baseline, Ages 18-30[a]			
Score- mean, SD, out of 14 points	10.4, 1.8	9.9, 1.9	9.8, 1.9
Low (0-7 points)	6.8%	10.5%	10.2%
Moderate (8-11 points)	62.9%	65.2%	70.5%
High (12-14 points)	30.4%	24.2%	19.4%
Cardiovascular Health Components at Baseline, Ages 18-30[b]			
Poor Physical Activity	19.2%	22.1%	21.6%
Poor Body Mass Index	11.1%	14.6%	12.6%
Poor Smoking	26.5%	36.9%	39.9%
Poor Diet	30.0%	35.0%	40.4%
Poor Blood Pressure	1.9%	2.2%	3.4%
Poor Total Cholesterol	4.1%	5.5%	4.5%
Poor Fasting Plasma Glucose	0.5%	0.0%	1.2%
Cardiovascular Health at Outcome Measurement, Ages ≥45			
Score- mean, SD, out of 14 points	8.9, 2.3	--	--
Low (0-7 points)	28.8%	--	--
Moderate (8-11 points)	57.5%	--	--
High (12-14 points)	13.7%	--	--
Cardiovascular Health Components at Outcome Measurement, Ages ≥45			
Poor Physical Activity	20.4%	23.6%	--
Poor Body Mass Index	40.9%	48.1%	--

Poor Smoking	18.1%	27.4%	--
Poor Diet	22.3%	30.7%	--
Poor Blood Pressure	11.6%	16.0%	--
Poor Total Cholesterol	7.7%	12.2%	--
Poor Fasting Plasma Glucose	5.8%	10.2%	--
Highest Degree Earned			
Elementary/Jr. High/ Some High School	2.0%	4.4%	7.4%
High School Graduate	11.8%	23.2%	28.8%
Some College	29.7%	34.4%	38.2%
College Graduate (4-Year)	23.0%	17.5%	13.8%
Graduate School	33.5%	20.5%	11.9%
Economic Stability- Income[d]			
<5k to $15,999	18.8%	27.7%	29.8%
$16k to $34,999	36.1%	34.9%	37.5%
$35k to $49,999	19.4%	17.2%	15.4%
$50k to $74,999	15.3%	13.4%	11.3%
≥$75k	10.4%	6.7%	6.0%
Social and Community Context- Mean Household Size (people, SD)[e]	3.0, 1.7	3.3, 1.7	3.2, 1.8
Neighborhood and Built Environment- Mean Percent <150% Federal Poverty Level (percent, SD)[f]	0.3, 0.2	0.3, 0.2	0.3, 0.2
Health and Health Care- Health Care Barriers[g]			
0 Barriers	70.1%	74.4%	65.1%
Psychosocial- CES-D Questionnaire[h]			
Yes- Depressed (CES-D ≥16)	22.7%	31.4%	29.0%

Data are % unless otherwise noted.

Abbreviations: CES-D = Center for Epidemiologic Studies Depression Scale; CVH = Cardiovascular Health; SD = Standard Deviation;

[a] Participants missing CVH at Baseline: All- 101, NoCVH- 15, No45- 46;

[b] Participants missing CVH Component Values at Baseline: Physical Activity: All- 1, NoCVH- 0, No45- 1; Body Mass Index: All- 14, NoCVH- 2, No45- 8; Smoking: All- 22, NoCVH- 3, No45- 11; Diet: All- 4, NoCVH- 0, No45- 0; Blood Pressure: All- 0, NoCVH- 0, No45- 1; Total Cholesterol: All- 28, NoCVH- 2, No45- 19; Fasting Plasma Glucose: All- 51, NoCVH- 8, No45- 22;

[c] Participants missing CVH Component Values at Outcome: Physical Activity: NoCVH- 57; Body Mass Index: NoCVH- 17; Smoking: NoCVH- 52; Diet: NoCVH- 226; Blood Pressure: NoCVH- 9; Total Cholesterol: NoCVH- 38; Fasting Plasma Glucose: NoCVH- 44;

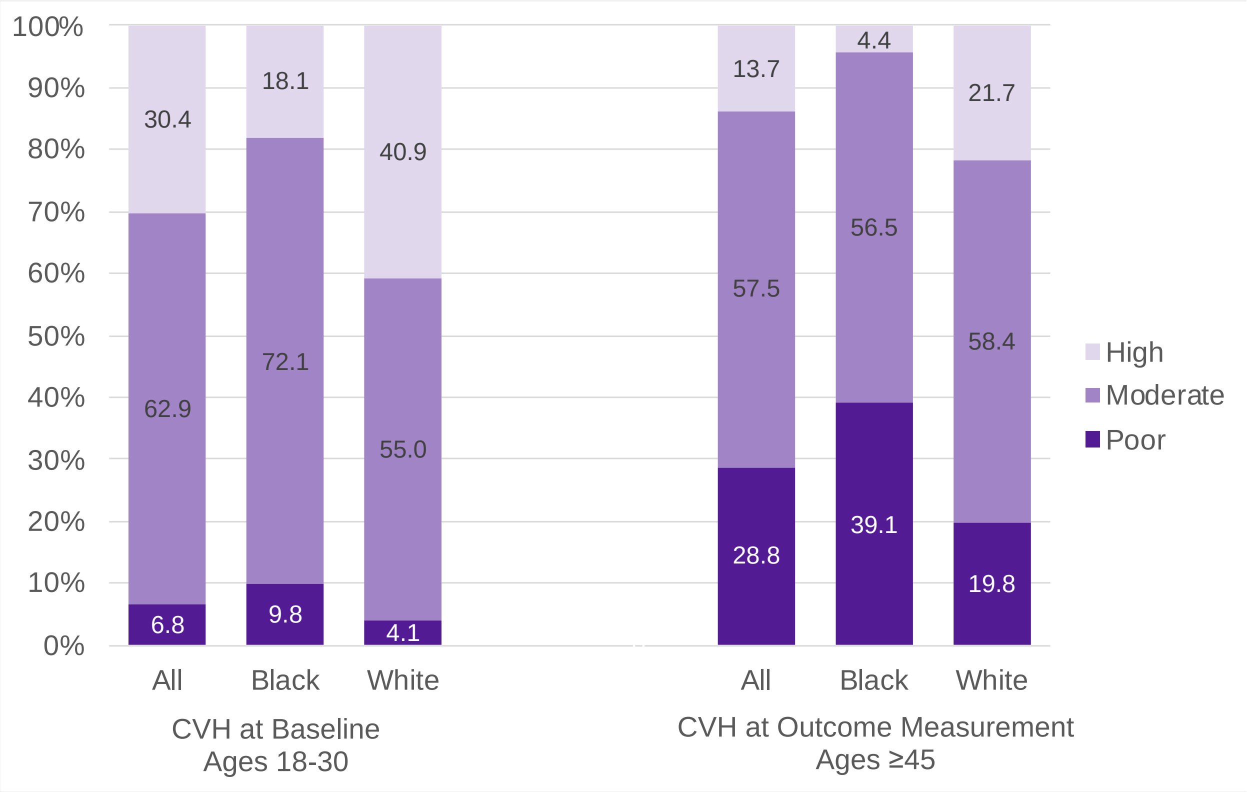
[d] Participants missing income at Y5: CVH- 252, NoCVH- 128, No45- 463;

[e] Participants missing household size at Y0: CVH-1, NoCVH- 1, No45- 2;

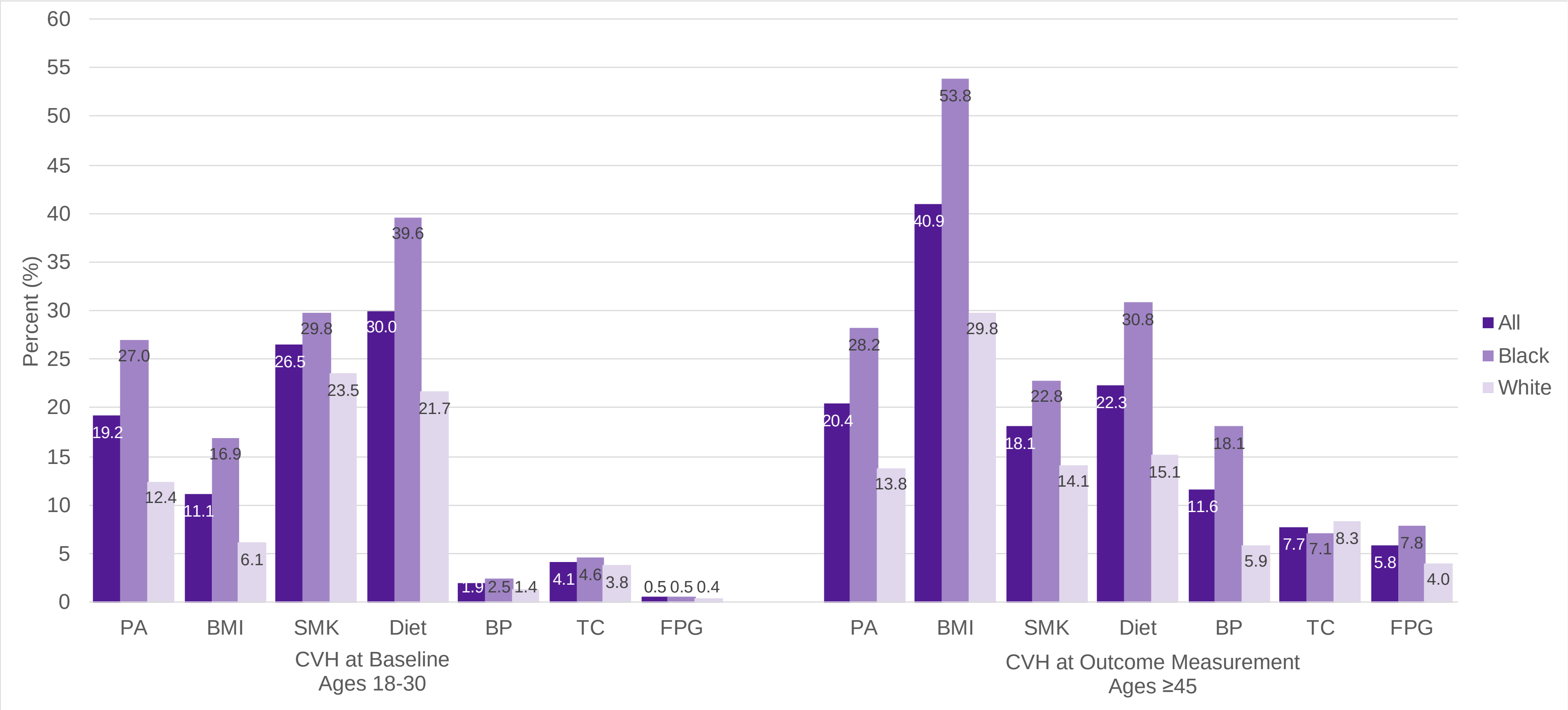
[f] Participants missing percent population in census tract at <150% of federal poverty level at Y0: CVH-25, NoCVH- 3, No45-6;

[g] Participants missing health care barriers at Y7: CVH- 271, NoCVH- 206, No45- 587;

[h] Participants missing CES-D questionnaire at Y5: CVH- 239, NoCVH- 121, No45- 463;



eFigure 2: Prevalence of poor, moderate, and high cardiovascular health (CVH) by race at baseline and outcome measurement.



eFigure 3: Prevalence of poor status for the cardiovascular health (CVH) component metrics by race at baseline and the outcome measurement. Abbreviations: BMI = Body Mass Index; BP = Blood Pressure; FPG = Fasting Plasma Glucose; PA = Physical Activity; SMK = Smoking; TC = Total Cholesterol;

eTable 5: Economic Stability SDOH variables at baseline, overall and by race[a]

	All	Black Participants	White Participants
	N=3,522	N=1,632 (46.3%)	N=1,890 (53.7%)
Income (Y5)			
<5k to $15,999	18.8%	27.6%	11.8%
$16k to $34,999	36.1%	40.0%	33.0%
$35k to $49,999	19.4%	16.6%	21.6%
$50k to $74,999	15.3%	12.0%	17.9%
≥$75k	10.4%	3.8%	15.8%
Home Ownership (Y5)			
Owned or Being Bought	47.5%	38.5%	54.8%
Renter for Money	48.1%	56.0%	41.7%
Occupied without Payment	4.0%	5.1%	3.1%
Other	0.3%	0.3%	0.3%
Hard to Meet Demands (Y0)			
Very Hard	2.7%	2.8%	2.7%
Hard	10.3%	7.7%	12.5%
Somewhat Hard	35.3%	31.1%	38.8%
Not Very Hard	51.7%	58.4%	45.9%
Hard to Pay for Basics (Y0)			
Very Hard	3.7%	5.2%	2.4%
Hard	6.6%	6.9%	6.4%
Somewhat Hard	22.8%	26.3%	19.8%
Not Very Hard	66.9%	61.6%	71.5%
Trouble Making Ends Meet (Y0)			
Frequent Trouble	7.4%	8.3%	6.6%
Occasional Trouble	41.4%	47.1%	36.5%
Hardly Ever Trouble	29.5%	28.2%	30.5%
Never Trouble	21.7%	16.4%	26.4%
Hard to Pay for Medical Care (Y10)			
Very Hard	6.4%	8.7%	4.5%
Hard	4.5%	5.9%	3.4%
Somewhat Hard	11.9%	12.0%	11.8%
Not Very Hard	77.2%	73.5%	80.3%
Assets (Y15)			
<$500-$4,999	14.8%	24.8%	7.0%
$5k - $19,999	10.8%	14.5%	7.9%
$20k - $49,999	9.3%	10.4%	8.4%
$50k - $99,999	14.0%	16.4%	12.1%

$100k - $199,999	15.3%	15.4%	15.2%
$200k - $499,999	21.2%	13.9%	27.0%
≥$500k	14.6%	4.5%	22.4%
Debt (Y15)			
<$500	25.2%	15.7%	32.7%
$500 - $4,999	29.0%	33.3%	25.8%
$5k - $9,999	16.4%	19.3%	14.1%
$10k - $19,999	14.3%	17.0%	12.2%
$20k - $49,999	10.4%	11.1%	9.9%
≥$50k	4.7%	4.0%	5.3%
Food Security (Y15)			
Enough Food and Kinds	86.5%	80.4%	91.4%
Not Always Enough or Kinds	13.5%	19.6%	8.6%
Employment Status (Y0)			
Employed	73.4%	62.6%	82.7%
Unemployed	26.6%	37.4%	17.3%
Mean Occupation Status (Y0), (TSEI, SD)	37.2, 18.5	30.8, 14.9	42.5, 19.4
Mean Job Decision Latitude (Y2), (mean score, SD)	35.6, 6.5	34.2, 6.3	36.7, 6.5
Mean Psychological Job Demands (Y2), (mean score, SD)	32.0, 6.3	31.2, 6.3	32.5, 6.2

[a]We selected exams as close to Y0 as possible for the baseline measure.

Data are % unless otherwise noted.

Abbreviations: SD = Standard Deviation; SDOH = Social Determinants of Health; TSEI = Total-based Socioeconomic Index;

eTable 6: Social and community context SDOH variables at baseline, overall and by race[a]

	All	Black Participants	White Participants
	N=3,522	N=1,632 (46.3%)	N=1,890 (53.7%)
Mean Household Size (Y0), (people, SD)	3.0, 1.7	3.6, 1.8	2.4, 1.4
Children or Step-Children (Y5)			
Yes	50.2%	64.4%	38.5%
Children or Step-Children Living in House (Y5)			
Yes	87.9%	84.8%	92.0%
Hard	10.3%	7.7%	12.5%
Hard to Pay for Basics (Y0)			
Very Hard	3.7%	5.2%	2.4%
Marital Status (Y0)			
Marriage-Like Relationship	10.6%	9.3%	11.7%
Married	24.0%	21.3%	26.3%
Never Married	58.4%	60.6%	56.5%
Separated or Divorced	6.7%	8.4%	5.4%
Widowed	0.3%	0.4%	0.2%
Mean Instrumental Support (Y0), (mean score, SD)	5.8, 2.7	5.7, 2.9	5.8, 2.6
Mean Emotional Support (Y0), (mean score, SD)	1.7, 1.0	1.5, 1.0	1.8, 0.9
Mean Network Adequacy (Y0), (mean score, SD)	10.5, 2.1	10.6, 2.3	10.5, 2.0
Any Discrimination (Y7)			
0 Domains	25.1%	14.8%	33.5%
1-2 Domains	34.8%	29.7%	39.0%
≥3 Domains	40.1%	55.5%	27.5%
Racial Discrimination (Y7)			
0 Domains	48.8%	21.9%	70.8%
1-2 Domains	28.2%	30.9%	25.9%
≥3 Domains	23.1%	47.2%	3.3%
Social Network (Y15)			
0-3 Ties	16.0%	19.7%	13.0%
4-7 Ties	22.8%	23.9%	21.9%
8-10 Ties	25.4%	22.5%	27.7%
11-14 Ties	22.7%	22.7%	22.8%
≥15 Ties	13.1%	11.3%	14.6%

Mean Supportive Interactions (Y15), (mean score, SD)	14.1, 2.3	13.8, 2.5	14.3, 2.0
Mean Negative Interactions (Y15), (mean score, SD)	8.2, 2.5	8.6, 2.7	7.9, 2.3

[a]We selected exams as close to Y0 as possible for the baseline measure.

Data are % unless otherwise noted.

Abbreviations: SD = Standard Deviation; SDOH = Social Determinants of Health;

eTable 7: Neighborhood and built environment SDOH variables at baseline, overall and by race[a]

	All	Black Participants	White Participants
	N=3,522	N=1,632 (46.3%)	N=1,890 (53.7%)
Self-Reported by Participant			
Mean Change in Residence (Y0), (moves, SD)	1.0, 1.5	0.8, 1.3	1.1, 1.6
Mean Neighborhood Cohesion (Y15), (cohesion score, SD)	3.6, 0.7	3.4, 0.7	3.8, 0.7
Mean Neighborhood Environment Resources (Y20), (resources, SD)	5.3, 1.7	5.5, 1.5	5.2, 1.9
Census Tract Level[b]			
Mean Percent Population White Race (percent, SD)	0.6, 0.3	0.6, 0.3	0.6, 0.3
Mean Percent Population Education < High School (percent, SD)	0.3, 0.2	0.3, 0.2	0.3, 0.2
Mean Percent Population <150% Federal Poverty Level (percent, SD)	0.3, 0.2	0.3, 0.2	0.3, 0.2
Mean Median Income (dollars, SD)	42,195, 18,292	42,477, 18,594	41,950, 18,028
Mean Percent Population Professional Occupation (percent, SD)	0.2, 0.1	0.3, 0.1	0.2, 0.1
Mean Percent Population Unemployed (percent, SD)	0.1, 0.1	0.1, 0.1	0.1, 0.1
Mean Median Rent (dollars, SD)	694, 223	694, 229	695, 217
Mean Percent Owner-Occupied Housing Units (percent, SD)	0.5, 0.2	0.5, 0.2	0.5, 0.3

Mean Percent Vacant Housing Units (percent, SD)	0.1, 0.0	0.1, 0.0	0.1, 0.0
Mean Aggregate Value Housing Units (dollars- millions, SD)	63.7, 74.6	64.2, 75.8	63.4, 73.4
Mean Racial Segregation (Gi* statistic) (z-score, SD)	1.9, 3.5	4.7, 3.0	-0.4, 1.8
Mean SES Deprivation Score (first factor score, SD)	0.2, 1.1	0.2, 1.1	0.2, 1.1
Mean Percent Fast Food/Convenience Stores w/in 3km (percent, SD)	0.0, 0.0	0.0, 0.0	0.0, 0.0
Mean Percent Supermarkets with 5km (percent, SD)	0.0, 0.0	0.0, 0.0	0.0, 0.0
Mean Physical Activity Facilities with 3km (count, SD)	47.2, 30.9	45.4, 26.8	48.7, 33.9

Abbreviations: SD = Standard Deviation; SDOH = Social Determinants of Health; KM = Kilometers;

[a] We selected exams as close to Y0 as possible for the baseline measure.

[b] All census tract level variables were reported at Y0.

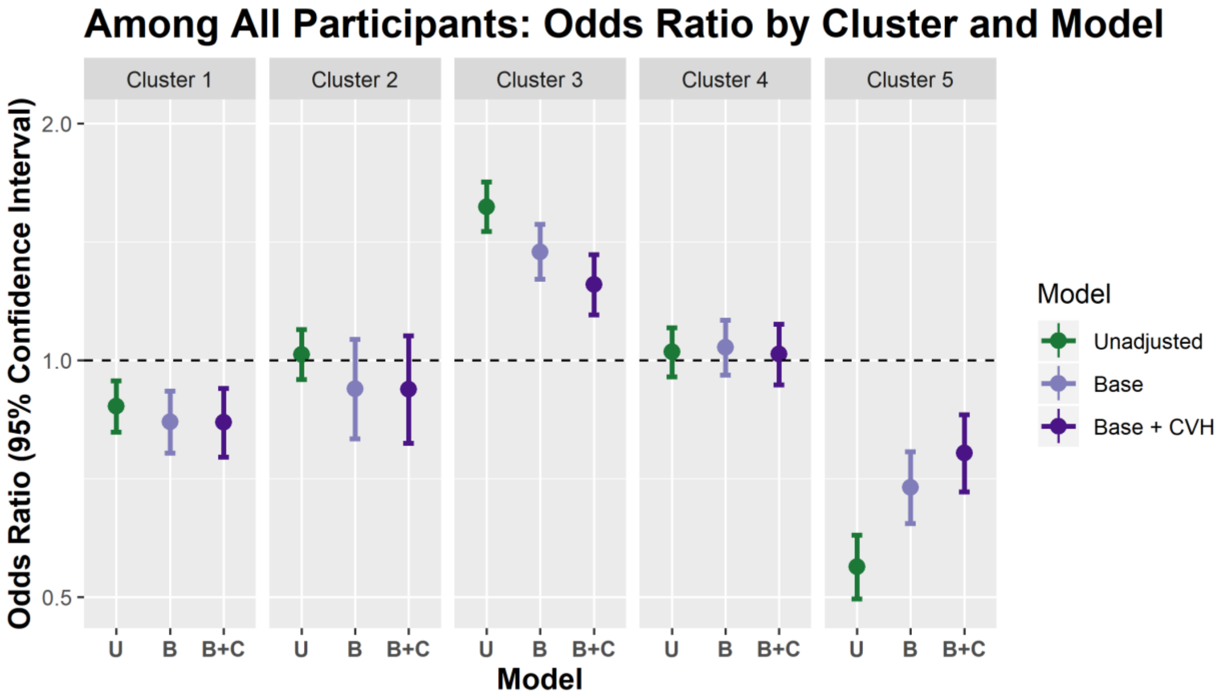
eTable 8: Health and health care and psychosocial SDOH variables at baseline, overall and by race[a]

	All	**Black Participants**	**White Participants**
	N=3,522	N=1,632 (46.3%)	N=1,890 (53.7%)
Health and Health Care			
Health Care Barriers (Y7)			
0 Barriers	70.1%	70.9%	69.4%
1 Barrier	19.7%	18.3%	20.9%
2 Barriers	7.9%	8.5%	7.4%
3 Barriers	2.3%	2.3%	2.4%
Mean Time Without Health Insurance Coverage (Y7), (months, SD)	15.8, 8.2	17.3, 7.7	14.4, 8.5
Psychosocial			
CES-D Questionnaire (Y5)			
Yes- Depressed (CES-D ≥16)	22.7%	29.0%	17.5%
Mean Chronic Burden Scale (Y15), (num domains stress, SD)	1.3, 1.3	1.2, 1.3	1.4, 1.3

[a]We selected exams as close to Y0 as possible for the baseline measure.

Data are % unless otherwise noted.

Abbreviations: CES-D = Center for Epidemiologic Studies Depression Scale; SD = Standard Deviation; SDOH = Social Determinants of Health;

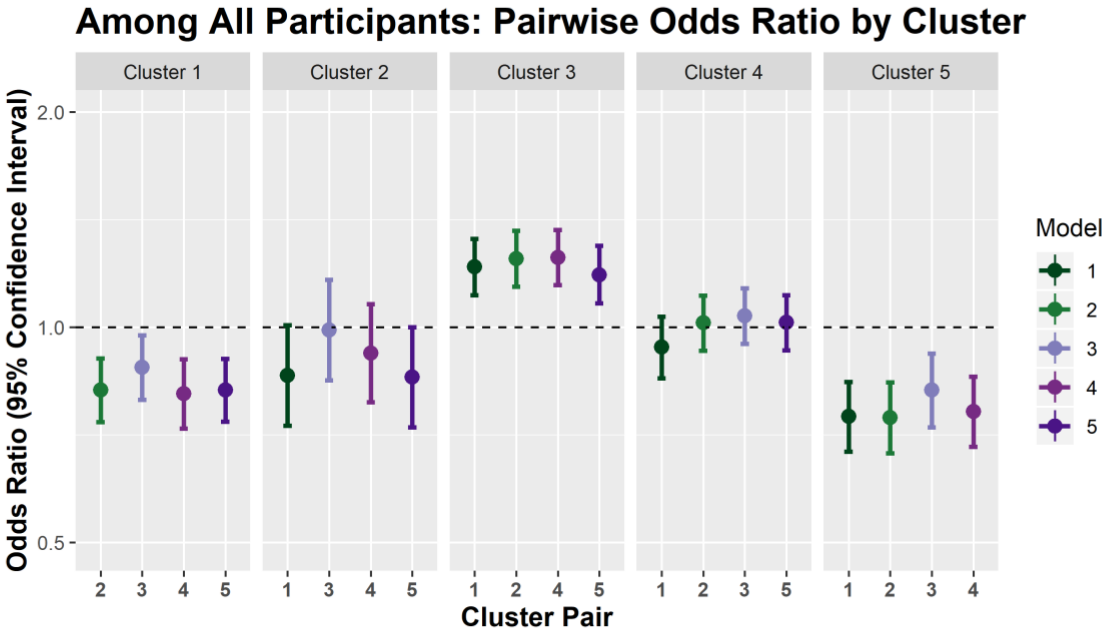


eFigure 4: Multivariable-adjusted logistic regression analysis of poor cardiovascular health incorporating each cluster created among all participants separately in unadjusted (U), Base (B), and Base + Cardiovascular Health (B+CVH) models.

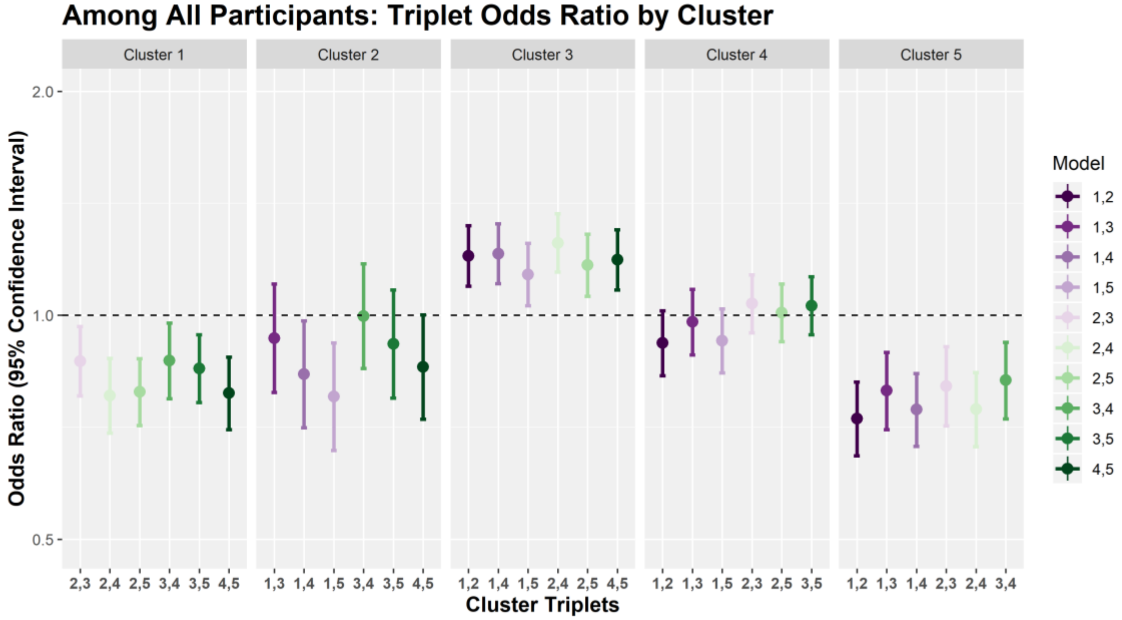
eTable 9: Prevalence of exposure to the patterns in each cluster by race and sex groups (standardized likelihood values ≥ 1SD in the coefficient matrix)

	Black Female	**Black Male**	**White Female**	**White Male**
	N = 965 (27.4%)	N = 667 (18.9%)	N = 1,002 (28.4%)	N = 888 (25.2%)
Cluster 1	171 (17.7%)	119 (17.8%)	216 (21.6%)	161 (18.1%)
Cluster 2	292 (30.3%)	192 (28.8%)	250 (25.0%)	222 (25.0%)
Cluster 3	231 (23.9%)	157 (23.5%)	106 (10.6%)	94 (10.6%)
Cluster 4	133 (13.8%)	83 (12.4%)	139 (13.9%)	137 (15.4%)
Cluster 5	38 (3.9%)	39 (5.8%)	241 (24.1%)	252 (28.4%)

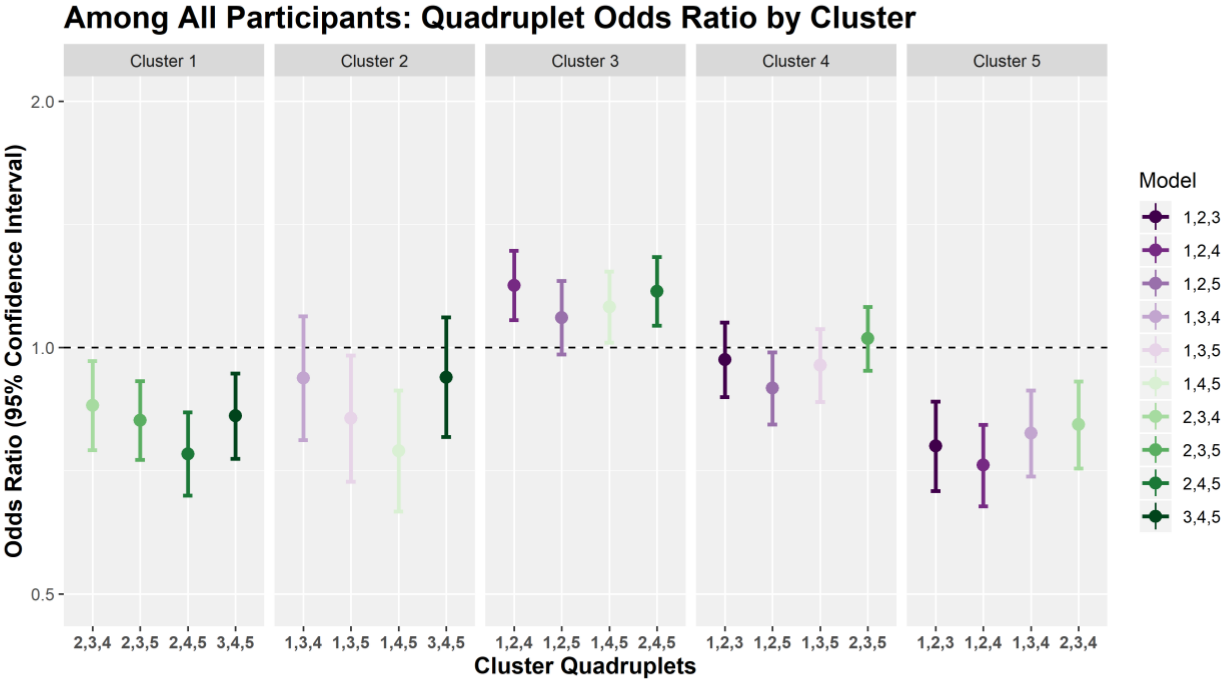
Abbreviations: SD = Standard Deviation;



eFigure 5: Multivariable-adjusted logistic regression analysis of poor cardiovascular health (adjusting for age, sex, center, education, baseline cardiovascular health score, and race) incorporating each cluster created among all participants separately with one other cluster. The odds ratios (OR) shown represent the changes in the primary cluster OR (represented by the cluster identified in the top grey bar) when including one other cluster in the model (identified on the X axis).



eFigure 6: Multivariable-adjusted logistic regression analysis of poor cardiovascular health (adjusting for age, sex, center, education, baseline cardiovascular health score, and race) incorporating each cluster created among all participants separately with two other clusters. The odds ratios (OR) shown represent the changes in the primary cluster OR (represented by the cluster identified in the top grey bar) when including two other clusters in the model (identified on the X axis).



eFigure 7: Multivariable-adjusted logistic regression analysis of poor cardiovascular health (adjusting for age, sex, center, education, baseline cardiovascular health score, and race) incorporating each cluster created among all participants separately with three other clusters. The odds ratios (OR) shown represent the changes in the primary cluster OR (represented by the cluster identified in the top grey bar) when including three other clusters in the model (identified on the X axis).

Supplement References

1. Luo Y, Xin Y, Joshi R, Celi LA, Szolovits P. Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements. In: *AAAI*. ; 2016. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewFile/11843/11562

2. Chao G, Mao C, Wang F, Zhao Y, Luo Y. Supervised Nonnegative Matrix Factorization to Predict ICU Mortality Risk. *CoRR*. 2018;abs/1809.10680. Accessed January 11, 2021. https://arxiv.org/abs/1809.10680

3. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*. 2010;11:367. doi:10.1186/1471-2105-11-367

4. Tapia Granados JA, Christine PJ, Ionides EL, et al. Cardiovascular Risk Factors, Depression, and Alcohol Consumption During Joblessness and During Recessions Among Young Adults in CARDIA. *Am J Epidemiol*. 2018;187(11):2339-2345. doi:10.1093/aje/kwy127

5. Janicki-Deverts D, Cohen S, Matthews KA, Gross MD, Jacobs DR. Socioeconomic Status, Antioxidant Micronutrients, and Correlates of Oxidative Damage: The Coronary Artery Risk Development in Young Adults (CARDIA) Study: *Psychosomatic Medicine*. 2009;71(5):541-548. doi:10.1097/PSY.0b013e31819e7526

6. Stevens G, Cho JH. Socioeconomic indexes and the new 1980 census occupational classification scheme. *Social Science Research*. 1985;14(2):142-168. doi:10.1016/0049-089X(85)90008-0

7. Greenlund KJ, Kiefe CI, Giles WH, Liu K. Associations of job strain and occupation with subclinical atherosclerosis: The CARDIA Study. *Ann Epidemiol*. 2010;20(5):323-331. doi:10.1016/j.annepidem.2010.02.007

8. Allen J, Markovitz J, Jacobs DR, Knox SS. Social support and health behavior in hostile black and white men and women in CARDIA. Coronary Artery Risk Development in Young Adults. *Psychosom Med*. 2001;63(4):609-618. doi:10.1097/00006842-200107000-00014

9. Krieger N, Sidney S. Racial discrimination and blood pressure: the CARDIA Study of young black and white adults. *Am J Public Health*. 1996;86(10):1370-1378. doi:10.2105/ajph.86.10.1370

10. Seeman TE, Gruenewald TL, Cohen S, Williams DR, Matthews KA. Social relationships and their biological correlates: Coronary Artery Risk Development in Young Adults (CARDIA) study. *Psychoneuroendocrinology*. 2014;43:126-138. doi:10.1016/j.psyneuen.2014.02.008

11. Dhurandhar EJ, Pavela G, Kaiser KA, et al. Body Mass Index and Subjective Social Status: The Coronary Artery Risk Development in Young Adults Study. *Obesity (Silver Spring)*. 2018;26(2):426-431. doi:10.1002/oby.22047

12. Kershaw KN, Hankinson AL, Liu K, et al. Social relationships and longitudinal changes in body mass index and waist circumference: the coronary artery risk development in young adults study. *Am J Epidemiol*. 2014;179(5):567-575. doi:10.1093/aje/kwt311

13. Kim D, Diez Roux AV, Kiefe CI, Kawachi I, Liu K. Do neighborhood socioeconomic deprivation and low social cohesion predict coronary calcification?: the CARDIA study. *Am J Epidemiol*. 2010;172(3):288-298. doi:10.1093/aje/kwq098

14. Whitaker KM, Jacobs DR, Kershaw KN, et al. Racial Disparities in Cardiovascular Health Behaviors: The Coronary Artery Risk Development in Young Adults Study. *Am J Prev Med*. 2018;55(1):63-71. doi:10.1016/j.amepre.2018.03.017

15. Richardson AS, Meyer KA, Howard AG, et al. Neighborhood socioeconomic status and food environment: a 20-year longitudinal latent class analysis among CARDIA participants. *Health Place*. 2014;30:145-153. doi:10.1016/j.healthplace.2014.08.011

16. Kershaw KN, Robinson WR, Gordon-Larsen P, et al. Association of Changes in Neighborhood-Level Racial Residential Segregation With Changes in Blood Pressure Among Black Adults: The CARDIA Study. *JAMA Intern Med*. 2017;177(7):996-1002. doi:10.1001/jamainternmed.2017.1226

17. Boone-Heinonen J, Diez Roux AV, Kiefe CI, Lewis CE, Guilkey DK, Gordon-Larsen P. Neighborhood socioeconomic status predictors of physical activity through young to middle adulthood: the CARDIA study. *Soc Sci Med*. 2011;72(5):641-649. doi:10.1016/j.socscimed.2010.12.013

18. Boone-Heinonen J, Gordon-Larsen P, Kiefe CI, Shikany JM, Lewis CE, Popkin BM. Fast food restaurants and food stores: longitudinal associations with diet in young to middle-aged adults: the CARDIA study. *Arch Intern Med*. 2011;171(13):1162-1170. doi:10.1001/archinternmed.2011.283

19. Kiefe CI, Williams OD, Greenlund KJ, Ulene V, Gardin JM, Raczynski JM. Health care access and seven-year change in cigarette smoking. The CARDIA Study. *Am J Prev Med*. 1998;15(2):146-154. doi:10.1016/s0749-3797(98)00044-0

20. Carroll AJ, Carnethon MR, Liu K, et al. Interaction between smoking and depressive symptoms with subclinical heart disease in the Coronary Artery Risk Development in Young Adults (CARDIA) study. *Health Psychol*. 2017;36(2):101-111. doi:10.1037/hea0000425

21. Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*. 1977;1(3):385-401. doi:10.1177/014662167700100306

22. Bromberger JT, Matthews KA. A longitudinal study of the effects of pessimism, trait anxiety, and life stress on depressive symptoms in middle-aged women. *Psychol Aging*. 1996;11(2):207-213. doi:10.1037//0882-7974.11.2.207

23. Everson-Rose SA, Roetker NS, Lutsey PL, et al. Chronic stress, depressive symptoms, anger, hostility, and risk of stroke and transient ischemic attack in the multi-ethnic study of atherosclerosis. *Stroke*. 2014;45(8):2318-2323. doi:10.1161/STROKEAHA.114.004815

CHAPTER FOUR: TIME-DEPENDENT CLUSTERS OF SOCIAL DETERMINANTS OF HEALTH IN YOUNG ADULTHOOD AND MID-LIFE CARDIOVASCULAR DISEASE EVENTS: THE CORONARY ARTERY RISK DEVELOPMENT IN YOUNG ADULTS (CARDIA) STUDY

4.1 Abstract

Background: Social determinants of health (SDOH) may be factors contributing to CVD disparities, and warrant additional attention in research. SDOH are traditionally studied in isolation, limiting our understanding of the complex associations between SDOH factors and their associations with disease, and need to be studied using new methods accounting for the relationships between social exposures.

Methods: Our primary objective was to create time-dependent SDOH clusters using a novel machine learning method, Subgraph-augmented Non-negative Matrix Factorization (SANMF), and evaluate whether the clusters were associated with mid-life CVD events before and after adjustment for mid-life subclinical CVD and cardiovascular health (CVH). We used data from the bi-racial CARDIA study, a prospective cohort study with detailed longitudinal information on individual- and neighborhood-level SDOH and CVD risk factors and disease. We included 4,853 Black men and women in our cohort with CVD events or follow-up time through age 45 or later. Cox proportional hazards modeling was used to fit unadjusted models and models adjusted for baseline age, CVH score, race, sex, and subsequently education, coronary artery calcification (CAC), left-ventricular mass index (LVMI), and mid-life CVH status.

Results: In the CVD Events cohort, there were 2,460 Black and 2,393 white participants with a mean age at baseline of 24.9 years, of whom 55.6% were female. After sequential pattern mining of 48 SDOH variables with repeated measures over time, we identified 353 frequently occurring time-dependent SDOH patterns, and created five unique and data-driven SDOH clusters of the patterns. In the multivariable model adjusting for baseline CVH score, baseline age, sex, race, and maximum education, Cluster 3—representing SDOH exposures of lower assets but food secure, rented home, lower job decision latitude job, mid-range neighborhood cohesion, no children, higher discrimination, lower social support, and higher chronic burden—was significantly associated with 13% higher hazards for CVD events (adjusted hazard ratio, 95% confidence interval: 1.13, 1.01 – 1.27). After adjustment for mid-life subclinical CVD and CVH status in separate models, the clusters were no longer significantly associated with mid-life CVD events.

Conclusions: We identified five novel SDOH clusters that were associated with mid-life CVD events that appear to be associated with CVD largely through known intermediary pathways of risk factors and subclinical disease development. This may lend further support for a primordial prevention strategy, specifically addressing upstream SDOH to prevent the development of subclinical and clinical CVD and development of health disparities.

4.2 Introduction

Although there was a sharp decline in cardiovascular disease (CVD) mortality rates since the 1970s, recent mortality rates are stagnant and the overall prevalence of CVD is rising, with major racial, ethnic, economic, and geographic disparities.[2–4] Non-Hispanic Black Americans are especially vulnerable to CVD with higher prevalence and levels of traditional risk factors and higher mortality rates compared with non-Hispanic whites.[57,93,94] The social determinants of health (SDOH) have received increasing attention in the past decade as modifiable, non-clinical factors that may confer benefit or harm among those at risk for CVD. SDOH are commonly defined as the "structural determinants and conditions in which people are born, grow, live, work, and age" and are made up of five key domains: economic stability, education, neighborhood and built environment, social and community context, and health and health care.[23,24] Limited data suggest that SDOH are associated with CVD incidence, treatment, and outcomes and may help to explain some of the persistent CVD disparities.[46]

It is well established that individual SDOH variables from the economic stability and education domains are associated with CVD risk factors,[30,31] subclinical CVD,[95–98] and CVD events cross-sectionally and longitudinally.[29,46,99,100] Less attention has been paid to studying SDOH from all five domains, the complex relationships between SDOH, and their associations with CVD. SDOH are traditionally examined in isolation, which only provides a narrow picture of an individual's social experiences and environment, and may prevent a true understanding of the co-occurrence and interactions between many SDOH and their associations with health outcomes. Additionally, the physiological, behavioral, and biological pathways linking SDOH and CVD are complex and not fully understood. For example, racial differences in the modifiable and behavioral CVD risk factors may be primarily explained by differences in socioeconomic status, but behavioral factors do not fully account for the ultimate disparities in CVD outcomes.[8,46] It is unclear how multiple SDOH from all five domains are associated with CVD over time and whether there are unique pathways linking SDOH and CVD outside of the traditional CVD risk factors and subclinical disease.

4.3 Objectives

Using the longitudinal, biracial Coronary Artery Risk Development in Young Adults (CARDIA) study cohort, we created clusters of time-dependent SDOH exposures through young adulthood by applying a novel machine learning method, Subgraph-Augmented Non-Negative Matrix Factorization (SANMF), and evaluated whether these clusters were associated with mid-life CVD events. Our secondary objective was to assess whether associations between young adult SDOH and mid-life CVD events persisted after adjustment for cardiovascular health (CVH) status or presence of subclinical CVD in middle age.

4.4 Methods

4.4.1 Study Design and Population

We used data from CARDIA,[58] a longitudinal cohort study with detailed information on cardiovascular risk factors and disease in a sample of 5,115 Black and white men and women aged 18 to 30 years at baseline from four metropolitan areas: Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota; and Oakland, California. Participants were originally enrolled between 1985 and 1986, and balanced across education, age, and self-identified race and sex groups at each center. Information on SDOH and CVD risk factors and disease was collected at nine exams over 30 years (baseline/year 0 (Y0), Y2, Y5, Y7, Y10, Y15, Y20, Y25, and Y30). Over 71% of surviving participants were present at the Y30 exam. Between exams, participants are contacted by telephone, mail, or e-mail every 6 months and medical history is ascertained every year. We included all participants with events or follow-up information at age 45 and later. In secondary analyses, we included participants with subclinical CVD and CVH measures as close to age 45 as possible and prior to a CVD event.

4.4.2 Exposures

To generate the time-dependent SDOH clusters, we used a novel machine learning method, subgraph-augmented non-negative matrix factorization (SANMF),[49] and followed the same process as performed in Chapters 2 and 3 using the full cohort with CVD events. We included the same 48 individual- and neighborhood-level SDOH from all five domains and psychosocial factors, as described in the Table 1. Instead of four age-intervals, we used three age-intervals: 18-24, 25-34, and 35-44 to ensure all SDOH were measured prior to all mid-life subclinical CVD, CVH, and CVD event outcome measures, and consistent with the primary and secondary analyses.

Our ultimate output from SANMF was two matrices: W, the features matrix, and H, the coefficient matrix. W contained the likelihood of each participant being exposed to the SDOH patterns from young adulthood through middle age in each cluster. These values were standardized into z-scores for interpretability during modeling. As in the previous chapters, our focus was on defining different clusters of SDOH exposure versus forcing each participant into a certain cluster. H contained the membership coefficients for each SDOH pattern in each cluster and was used to characterize the clusters based on the patterns with the top ten highest membership coefficients in each cluster.

4.4.3 Outcomes

The primary outcome for this study was CVD events at age 45 or later. CVD events were defined as the first occurrence of any of: 1) nonfatal myocardial infarction or stroke, (2) hospitalization for acute coronary syndrome not resulting in infarction, heart failure, or transient ischemic attack, (3) hospitalization for heart failure, (4) revascularization for or demonstration of obstruction of carotid arteries or peripheral arterial disease on angiographic or ultrasonographic findings, or (5) underlying cause of death of CVD, as defined previously.[101] Study participants were contacted annually about hospitalizations or procedures and vital status was assessed every 6-months.[101] For any CVD events, medical records were adjudicated by two physicians and any conflicts were addressed in a full committee review.[101] Over 90% of the surviving CARDIA participants have been contacted directly in the last 5 years, and vital status ascertainment is virtually complete through these methods and periodic queries of the National Death Index.

In order to understand the role of potential intermediary pathways between SDOH and CVD events, we also adjusted for mid-life CVH and subclinical CVD in secondary analyses. Subclinical CVD was measured in two ways: 1) presence or absence of coronary artery calcification (CAC) and 2) continuous left ventricular mass index (LVMI). CAC, LVMI, and CVH were measured as close to age 45 as possible and prior to any CVD events. CAC is highly correlated with the degree of coronary atherosclerosis and strongly associated with CVD risk factors and the rate of future cardiovascular events.[102–104] In CARDIA, CAC was measured by computed tomography (CT) of the chest at year 20 and 25.[95,105] Scan data from two sequential electrocardiogram-gated scans were transmitted to the CARDIA CT Reading Center and examined by a trained technician using image-processing software. The technician identified presence of calcified plaque. An expert examined and adjudicated any discordant scan pairs. LVMI is associated with traditional CV risk factors, and is a known risk factor for CVD events.[106–109] LV structure and function was measured with 2-dimensional echocardiography at years 25 and 30.[110] For our outcome, LVMI was measured in the units of $g/m^{2.7}$ to index for height. Mid-life CVH was measured as a continuous score, as defined by the American Heart Association.[9] CVH is made up of seven clinical and behavioral factors including current smoking, body mass index, physical activity, diet, total cholesterol, blood pressure, and fasting plasma glucose. Each factor was assigned 0, 1, or 2 points based on cutpoints for poor, intermediate, and ideal levels of each factor yielding a composite CVH score (range 0-14 points), as has been done in other studies.[59,60]

4.4.4 Statistical Analysis

We used R version 3.6.1 for all analyses.[73] We performed unadjusted and adjusted Cox proportional hazards modeling, ensuring the proportional hazards assumption was met prior to analyses. We censored participants who died from causes other than a CVD event on their date of death and those who did not have an event during the follow-up period on their last contact date. In multivariable models, a two-tailed P value of <0.05 was used to determine statistical significance.

For the primary analyses, we created two sets of models to assess the associations between the SDOH clusters and CVD events. The SDOH clusters were modeled as continuous z-scores representing each participant's standardized likelihood of exposure to the SDOH patterns in each cluster. In set one, we compared models with the SDOH clusters 1) unadjusted, 2) adjusted for age at baseline, sex, and self-reported race, and 3) adjusted for age at baseline, sex, self-reported race, and maximum years of education achieved. In set two, we built upon the models in set one, but adjusted for baseline CVH in addition to the other covariates. We also conducted race-stratified analyses to understand whether associations identified in the primary analyses were consistent among Black and white participants.

In secondary analyses, we modeled the association between the SDOH clusters and CVD events, after adjustment for mid-life subclinical CVD and CVH. In order to understand whether SDOH clusters retained independent associations with CVD after adjustment for intermediary markers of disease pathogenesis, we created three models including the SDOH clusters, maximum years of education achieved, and covariates (baseline CVH, baseline age, sex, race), with further adjustment for CAC, LVMI, and CVH in separate models.

4.5 Results

4.5.1 Study Sample

There were 4,853 total participants at risk for a CVD event at age 45 and older and analyzed in the primary analysis (eFigure 1). For the secondary analyses, 3,100 participants were included when adjusting for mid-life CAC, 3,349 participants when adjusting for mid-life LVMI, and 3,448 participants when adjusting for mid-life CVH. Table 8 reflects the demographic, CVD risk factor, and SDOH characteristics of each cohort. A summary of the CVD events, subclinical CVD, and CVH measures by cohort are presented in Table 9. Those in the primary CVD events sample were similar to all CARDIA participants across all measures. Participants excluded from the CAC, LVMI, and CVH cohorts were more likely to be people who were Black, smokers, and had less favorable SDOH.

4.5.2 Characterization of SDOH Clusters

There were 353 total time-dependent SDOH patterns found after frequent pattern mining in the CVD events cohort. Among the 353 patterns, 112 spanned one age interval, 175 spanned two age intervals, and 66 spanned three age intervals. Following non-negative matrix factorization, the five SDOH clusters identified are represented and described in Figure 9. All domains other than education, which was included in modeling as a separate covariate, were represented in at least one cluster. All clusters except Cluster 5 had SDOH patterns spanning multiple age intervals. Cluster 2 was characterized by patterns primarily among the 18-24 age interval and the remaining four clusters had patterns primarily from the 25-34 and 35-44 age intervals. There were three patterns from the 35-44 age interval present in multiple clusters: 1) patterns related to having enough food and kinds of food (Clusters 1, 3, and 4); 2) patterns related to no difficulties paying for medical care (Clusters 1, 4, and 5); and zero healthcare access barriers (Clusters 1, 4, and 5).

Different race and sex groups were more likely to be exposed to the SDOH patterns in certain clusters, as shown in eTable 1. For example, more white males and females were exposed to Cluster 1, whereas more Black males and females were exposed to Cluster 3.

4.5.3 Primary Analyses of SDOH-CVD Event Associations

For the CVD Events cohort, there were 252 total events during a median follow-up time of 33.8 years, for an unadjusted event rate of 1.56 events per 1,000 person-years. On average, the age at the participants' first CVD event was 52.4 years in this young adult cohort at inception. Unadjusted event rates were slightly different across race and sex subgroups and SDOH clusters, as shown in Figure 10. In the initial unadjusted analyses presented in Table 10, Clusters 1 and 3 were significantly associated with incident CVD events (p=0.004 and p=0.006, respectively). Cluster 1 was associated with lower risk, indicating that a one standard deviation higher likelihood of being exposed to the patterns in Cluster 1 was associated with 19% lower unadjusted hazards for CVD events. For Cluster 3, a one standard deviation higher likelihood of being exposed to the patterns in the cluster was associated with 17% higher unadjusted hazards for CVD events. After adjusting for age at baseline, sex, and race in Model 2, the results were consistent. In Model 2, self-reported Black race compared to white race was significantly associated with 38% higher hazard of CVD events even with the SDOH clusters in the model (p=0.022). In Model 3, adjustment for education along with the existing covariates from Model 2, Cluster 3 and race were no longer significantly associated with CVD events, but one-year greater education attainment was significantly associated with 10% lower adjusted hazards for CVD events.

In Table 11, we present the results after further adjustment for the baseline CVH score along with age, sex, and race in Model 4 and also education in Model 5. Across both models, Cluster 3 was significantly associated with higher hazards for CVD events (p=0.038 in Model 4 and p=0.039 in Model 5). A one standard deviation greater likelihood of being exposed to the patterns in Cluster 3 was associated with 13% higher hazards for CVD events in both models, independent of baseline CVH score and the additional covariates. In Models 4 and 5, race was no longer significantly associated with CVD events.

In the race-stratified analyses presented in Table 12, Cluster 3 was significantly associated with higher hazards for CVD events among Black participants (fully-adjusted hazard ratio (aHR): 1.17, 95% Confidence Interval (CI): 1.01 – 1.35). Among white participants, Cluster 1 was significantly associated with lower hazards for CVD events in the partially adjusted model (aHR, 95% CI: 0.76, 0.62 – 0.93), but this relationship was no longer significant after adjustment for baseline CVH score and education.

4.5.4 Secondary Analyses Adjusting for Mid-life CVH and Subclinical CVD

Table 13 presents the secondary analyses of the SDOH-CVD event associations after adjustment for mid-life CAC, LVMI, and CVH in separate models. Presence or absence of CAC at mid-life and a one $g/m^{2.7}$ higher LVMI were both significantly associated with greater hazards for CVD events (aHR, 95% CI for CAC: 2.28, 1.62 – 3.19; and for LVMI: 1.03, 1.02 – 1.04). Additionally, a one point higher mid-life CVH score was significantly associated with lower hazards for CVD events (aHR, 95%: 0.82, 0.75, 0.89). There were, however, no significant associations between the SDOH clusters and CVD events after adjusting for mid-life CVH and subclinical CVD.

4.6 Discussion

In this study, we identified five clusters of SDOH exposures made up of different time-dependent SDOH patterns from multiple domains through young adulthood. There were significant associations between Clusters 1 and 3 and incident CVD events; exposure to the patterns in Cluster 1 was associated with lower CVD risk, and Cluster 3 with higher risk for CVD events. In race-stratified analyses, we observed that associations between Cluster 3 and mid-life CVD events were primarily seen among the Black participants, whereas the white participants drove the Cluster 1 associations. We may, however, be underpowered to detect the different associations between the SDOH clusters and CVD events within race groups.

As part of our secondary objective, we sought to understand potential pathways linking the SDOH clusters and mid-life CVD events. Prior to adjustment for education, and after adjustment for the SDOH clusters, self-identified race was still significantly associated with incident CVD events. Further exploration is warranted as to whether there may be other modifiable pathways linking race and mid-life CVD events outside of the SDOH from young adulthood through middle age. Of note, after adjustment for baseline CVH, Cluster 3 was still significantly associated with mid-life CVD events. Baseline CVH was assessed when participants were between the ages of 18 and 30 and was influenced by childhood SDOH, indicating that early life exposures may not fully determine future incident CVD events in mid-life.[81,82] This finding suggests that attention should still be paid to SDOH exposures from young adulthood through middle age, as they may modify the risk of mid-life CVD events.

When including mid-life CVH and subclinical CVD, the SDOH clusters were no longer significantly associated with incident CVD events in mid-life. Thus, the relationship between young adult SDOH exposures and mid-life CVD events may work primarily through pathways mediated by mid-life subclinical CVD and/or mid-life CVH risk factors, which are more proximal indicators.

Our findings further support a broad emphasis on primordial prevention, defined as the maintenance of high CVH and the prevention of the development of CVD risk factors, in combating CVD disease.[111,112] If SDOH primarily work through mid-life risk factors and subclinical disease, preventing the development of risk factors and subsequent subclinical and clinical disease by addressing SDOH could be an effective disease prevention strategy. As shown previously, treating risk factors back to optimal levels cannot restore risk of CVD to ideal levels and there is a point of no return after which damage to the cardiovascular system may not be fully reversible.[113] Because SDOH are known to impact risk factors through stress, knowledge, and time, it is critical to reverse these stressors before they begin.[43,44,99] By moving further upstream to create interventions and policies targeted towards SDOH that confer risk, like structural racism, we may achieve lower CVD event rates.

Our study has distinct strengths, separate from what has been mentioned in Chapters 2 and 3. This is the first study, to our knowledge, which has examined longitudinal SDOH exposures, incorporating factors from all five domains, in relation to incident CVD events in mid-life. Because of the rich measures collected in CARDIA, we were able to assess how the SDOH-CVD event associations changed or were maintained by adjustment for subclinical CVD and CVH status. We were also able to adjust for baseline CVH status, which differentiates the effect of early-life SDOH and other factors from young adult and later factors. While we may not be powered to fully understand SDOH-CVD event associations within and between Black and white groups differentially, this study does indicate that SDOH are associated with the development of CVD events in both groups.

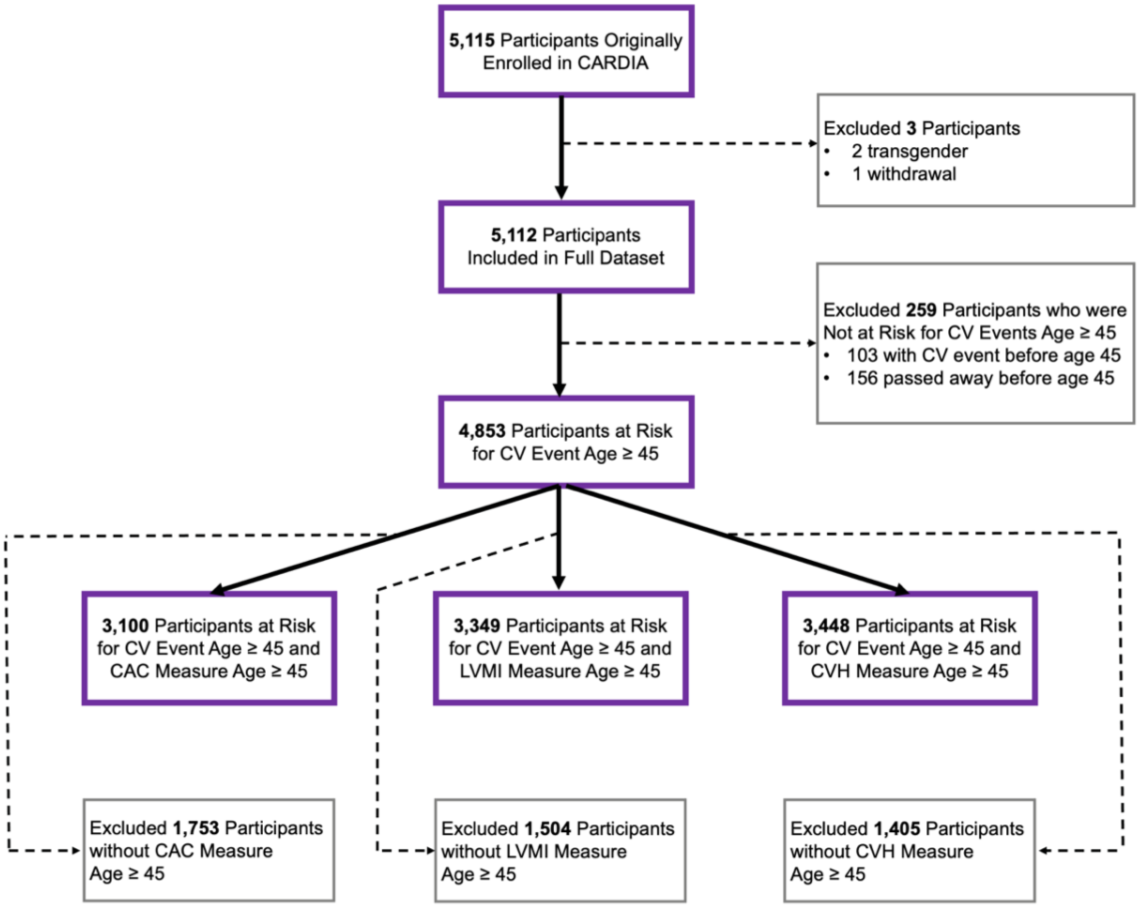
There are also limitations to this study. We focused on mid-life CVD events because of the length of follow-up in the younger CARDIA cohort, and the importance of mid-life CVH as a marker for health and longevity for the remainder of the life course.[89] Future studies should examine later-life CVD events and their relationship with SDOH. As in Chapters 2 and 3, we may be underpowered in the race-stratified analyses. In secondary analyses, those excluded from the CAC, LVMI, and mid-life CVH cohorts generally had a higher prevalence of SDOH that confer harm, meaning we may be underestimating the effect of the SDOH on CVD events. We were also limited by when the subclinical CVD and CVH measures were collected. Because we were concerned about reverse causality, we excluded all participants without subclinical CVD and CVH measured prior to a CVD event after age 45 or later, restricting our sample size for analysis.

4.7 Conclusions

Using the novel SANMF method, we generated and characterized five SDOH clusters, each made up of time-dependent SDOH patterns from multiple SDOH domains through young adulthood. The clusters were significantly associated with mid-life CVD events, even after adjustment for baseline covariates and CVH status. Race-stratified analyses indicate the clusters may be useful for both Black and white populations. After consideration of mid-life CVH and subclinical CVD status, the SDOH clusters were no longer associated with CVD events, indicating the pathways linking SDOH and incident CVD are primarily through detectable intermediate subclinical disease.

This work further underscores the need to focus on primordial prevention of CVD risk factors and focus on upstream SDOH to minimize CVD disparities. As in Chapters 2 and 3, food security, paying for medical care, and health care access were three patterns present in multiple SDOH clusters. Natural experiences of interventions and policies focused on these three SDOH exposures may be the best place to start to address SDOH that may substantially influence CVD risk factors and disease.

4.8 Supplement


5,115 Participants Originally Enrolled in CARDIA
Excluded **3** Participants
• 2 transgender
• 1 withdrawal
5,112 Participants Included in Full Dataset
Excluded **259** Participants who were Not at Risk for CV Events Age ≥ 45
• 103 with CV event before age 45
• 156 passed away before age 45
4,853 Participants at Risk for CV Event Age ≥ 45
3,100 Participants at Risk for CV Event Age ≥ 45 and CAC Measure Age ≥ 45
3,349 Participants at Risk for CV Event Age ≥ 45 and LVMI Measure Age ≥ 45
3,448 Participants at Risk for CV Event Age ≥ 45 and CVH Measure Age ≥ 45
Excluded **1,753** Participants without CAC Measure Age ≥ 45
Excluded **1,504** Participants without LVMI Measure Age ≥ 45
Excluded **1,405** Participants without CVH Measure Age ≥ 45

eFigure 1: **STROBE diagram for included and excluded participants**

eTable 1: Prevalence of SDOH clusters overall and by race and sex groups (cluster likelihood values ≥ 1SD)*

	Overall	**Black Female**	**Black Male**	**White Female**	**White Male**
	N = 4,853	N = 1,415 (29.2%)	N = 1,045 (21.5%)	N = 1,283 (26.4%)	N = 1,110 (22.9%)
Cluster 1	953 (19.6%)	113 (8.0%)	81 (7.8%)	384 (29.9%)	375 (33.8%)
Cluster 2	1,223 (25.2%)	397 (28.1%)	265 (25.4%)	297 (23.1%)	264 (23.8%)
Cluster 3	851 (17.5%)	294 (20.8%)	229 (21.9%)	181 (14.1%)	147 (13.2%)
Cluster 4	1,070 (22.0%)	432 (30.5%)	204 (19.5%)	262 (20.4%)	172 (15.5%)
Cluster 5	863 (17.8%)	240 (17.0%)	165 (15.8%)	235 (18.3%)	223 (20.1%)

Abbreviations: SD = Standard Deviation;

*Percentages do not add up to 100%. There can be multiple clusters where likelihood of exposure to the patterns in each cluster are ≥ 1SD

CHAPTER FIVE: CONCLUSION AND FUTURE DIRECTIONS

This work was motivated by a desire to understand the complex nature of SDOH and their associations with important health outcomes, in order to begin to determine effective interventions to lower the burden and disparities of cardiovascular disease. While there is existing literature describing the associations between a limited set of SDOH, most often from the economic stability and education domains, and CVD, data and research are sparse linking complex SDOH exposures from multiple domains with CVH and CVD over time. Additionally, the methods used to study SDOH and CVD are limited, and do not allow for a clear understanding of how multiple SDOH are associated with CVH and CVD over time.

5.1 Principal Findings

The dissertation contains three chapters that support our original hypothesis: 1) we can use SANMF, a novel machine learning method, to identify time-dependent SDOH patterns from young adulthood to middle age and create novel SDOH clusters of the patterns and 2) the clusters were predictive of mid-life CVH and associated with mid-life CVH status and CVD events. The data-driven sets of SDOH clusters that we observed were interpretable and unique, representing frequent patterns of SDOH from the economic stability, neighborhood and built environment, social and community context, and health and health care domains, along with psychosocial factors. The clusters also represented patterns of exposure from different age intervals from young adulthood to middle age. Food security, access to medical care, and the ability to pay for medical care were three patterns that consistently appeared in the clusters, indicating interventions and policies targeting these SDOH exposures may be important for improving population-level CVH.

Without the SANMF method, we would not have been able to understand the complexity of the 48 different individual- and neighborhood-level social exposure variables over time. Certain clusters were associated with mid-life CVH and did offer modest improvement in the prediction of mid-life CVH along with baseline CVH. The clusters were associated with mid-life CVD events, but this association became no longer significant after adjustment for mid-life subclinical CVD and mid-life CVH status, suggesting that SDOH may ultimately work through traditional risk factors and known disease pathways. In Chapters 3 and 4, we explored race-stratified analyses; while we may be underpowered to assess significant differences of the effect of SDOH on mid-life CVH or CVD events across race groups, our initial findings indicate the clusters are associated with health outcomes within both groups.

5.2 Methodological Approach

There are similarities and differences between the three manuscripts in this dissertation. We used the same SANMF method to generate the novel time-dependent SDOH clusters within different study cohorts across the three chapters. By applying SANMF, we identified frequent patterns of SDOH from young adulthood through middle age and clustered the patterns into meaningful groups in order to better represent the complex associations between longitudinal social exposures from multiple domains. We used the same mid-life CVH outcome in Chapters 2 and 3, but split the study cohort into a training and test set for predictive modeling in Chapter 2. We then used CVD events as our outcome, and introduced mid-life subclinical CVD measures for adjustment, in our Chapter 4 models. The broad goals of Chapter 2 and Chapters 3 and 4 were different; we focused on prediction in Chapter 2 and then associations in Chapters 3 and 4. Chapter 2 allowed us to think more about how we might target high-risk populations to improve or maintain their CVH status based on their social exposures, whereas Chapters 3 and 4 assisted in helping to improve our understanding of the complex nature of SDOH, the magnitude of the associations between SDOH and CVH/CVD events, and the pathways linking SDOH with CVH and CVD.

5.2.1 Strengths and Limitations

The strengths of this dissertation come from the rich data, the overall study design, and the methods applied. Whereas many previous studies have used single, cross-sectional SDOH measures to examine associations with health outcomes, we used SDOH data from young adulthood through middle age, representing exposures from all five SDOH domains. This allowed for a broader view of the longitudinal influences of social exposures on health outcomes during mid-life. As mentioned repeatedly in this work, examining SDOH in isolation does not provide the complete picture of how social exposures work to shape health outcomes. Our novel methodological approach, leveraging the complexity of the CARDIA data, allows a more comprehensive look at how 48 different SDOH, on both an individual- and neighborhood-level, may influence mid-life CVH. Additionally, we focused on a critical period of CVH during mid-life, which determines an individual's remaining lifetime risk of cardiovascular disease. This work also allowed for the generation of hypotheses to be explored further in future studies. In Chapter 4, we explored whether there were potentially independent pathways linking SDOH and CVD outside of subclinical CVD and CVH during mid-life. Although we did not perform formal statistical mediation analysis, which has its own inferential limitations, our data indicate that the association of younger adult SDOH exposures with mid-life CVD events is attenuated by intermediate factors such as CVH status and presence of subclinical CVD, the usual pathways from health to CVD incidence. And leveraging CARDIA's biracial cohort, we conducted an initial set of race-stratified analyses to understand whether the relationships between the novel SDOH clusters, CVH, and CVD events were consistent among Black and white subgroups. We believe this work lays the foundation for future studies described below.

Despite the strengths listed above, this dissertation has potential limitations. The CARDIA dataset is unique, which may limit generalizability. Participants were recruited from four specific geographic areas with a focus on balancing the composition of the cohort by sex, race, educational attainment, and age subgroups. Data were only collected among Black and white participants, and it would be useful to extrapolate our findings more broadly in other race groups. However, the 48 individual- and neighborhood-level SDOH collected at nine exams over 30 years may not be captured in other settings. While many studies are utilizing neighborhood-level measures by geocoding participant addresses to study SDOH, the self-reported SDOH information on the individual-level is much less widely available outside of traditional observational studies. Moreover, there are very few cohort studies with detailed information on SDOH *and* well-adjudicated CVD outcomes.

Despite the strengths of the CARDIA dataset, there were other limitations related to the timing of the measures and our study design. We were limited by the windows in which SDOH were collected. Each SDOH measure was not collected across all nine exams and age windows; half of the participants did not have information in the 18-24 age window because they were recruited initially between the ages of 25 and 30. We only used the first measurement of CVH and CVD events as our outcomes in each chapter, which limits our understanding of the effect of SDOH on repeated outcomes measures. The application of SANMF to study SDOH and CVD was novel, but still has its own boundaries on improving our understanding. For example, we focused on examining longitudinal patterns for each SDOH variable but did not examine the interplay of the SDOH patterns and subsequent changes in the patterns over time. The clusters of the patterns improved our understanding of the complexity inherent in SDOH exposures but may introduce some difficulty in implementing simple solutions to address SDOH that confer harm.

5.3 Implications

The primary implications of this work are three-fold: 1) machine learning can be beneficial to understand the complexity and intersectionality of SDOH exposures, 2) incorporating the longitudinal nature of SDOH is important to understand their effect on health outcomes, and 3) a focus on primordial prevention will continue to be a critical strategy to minimize the burden of CVD. By applying SANMF to SDOH factors, we generated novel clusters representing the frequent SDOH patterns that occurred concurrently. This emphasizes the intersectional nature of these factors; for example, a participant may not just be low-income, but may be exposed to low social support and low neighborhood cohesion that may have implications for health outcomes.

5.3.1 Intersectionality of SDOH exposures

Kimberlé Crenshaw, an African-American feminist and legal scholar, first coined the term “intersectionality” in 1989 to draw attention to the interactions of race and sex/gender in an individual’s lived experience.[118] This dissertation demonstrates one new method to study the intersectionality and complexity of SDOH exposures. Using our data-driven approach, the clusters generated were made up of unique SDOH patterns from various age intervals. Our findings suggest the need to focus on the complexity of the SDOH exposures, and also when they occur and how social exposures change over time to determine how to intervene.

5.3.2 SDOH as causal factors for CVH and CVD

Dr. Thomas McKeown, a British physician, first used the term “determinants of health” in 1972 during his thesis attributing increases in life expectancy during the 19th century to modern medical advancements, including antibiotics and intensive care units, and improved living conditions from enhancements in nutrition, sanitation, and clean water.[114,115] There is a large body of evidence demonstrating the association between individual social factors and a variety of health outcomes.[23,46,115,116] The causal pathways, however, between SDOH and health are still being explored more broadly and by health condition. Multiple conceptual frameworks have been developed linking SDOH with morbidity and mortality, highlighting the relationship between macrosocial conditions and political contexts, individual socioeconomic position (shaped by the distribution of money, power, and resources), and the development of risk factors and behaviors and disease.[29,42,55] If we assume CVH status is causal for CVD and mortality (since it consists of 7 risk factors that are each individually causal), this dissertation demonstrates that there is an association between SDOH and CVH, which provides evidence to support existing conceptual models linking upstream social factors with downstream disease through risk factors and subclinical disease. This dissertation provides support for other Bradford Hill criteria for causation[117] including: the temporality criterion based on the distinct timing of the exposures and their associations with CVH and CVD events, and also the plausibility criterion as demonstrated by support for the causal pathway linking SDOH and CVD events through risk factors and subclinical disease. This first body of work sets the stage for future studies to help us understand causality by intervening on SDOH and subsequently studying the short and long-term effects on CVH and disease outcomes. As mentioned in the Introduction, there is existing evidence supporting the idea that intervening on SDOH can improve CVH, but more work is needed in this area.[35,36]

5.3.3 Support for a primordial prevention strategy

As highlighted in Chapter 4, this dissertation also supports a primordial prevention strategy to address upstream SDOH before they lead to loss of CVH (onset of CVD risk factors), subclinical CVD, and future CVD events. Once risk factors are present, it is very difficult to minimize the risk of future disease, even with effective treatment.[113,119] Public health professionals, clinicians, and researchers will need to collaborate to develop targeted interventions and policies to address SDOH. Described further below, natural experiments may be the first steps toward this ultimate goal of creating impactful SDOH interventions and policies.

5.4 Future Research

While this dissertation adds to the literature, there are remaining knowledge gaps and future research left to explore. The CARDIA study incorporates very detailed measures of over 48 individual- and neighborhood-level SDOH. Some previously validated instruments were used to capture the SDOH variables, but other instruments are specific to CARDIA and were designed by the CARDIA investigators. There is a need to standardize measurement of SDOH variables more broadly, primarily in relatively unexplored social and community context, neighborhood and built environment, and health and health care domains. This will allow researchers to continue to monitor and understand the effects of SDOH based on standardized and widely disseminated definitions. Traditional epidemiologic studies currently are the best source for SDOH asked directly of participants, but there has been a push in recent years to systematically collect SDOH in electronic health records.[120,121] These efforts will be critical as we continue to study SDOH and intervene to address social risks.

We must also continue to focus on the intersectionality of SDOH and health outcomes more broadly. Recently, there has been a call to incorporate intersectionality research and methods in public health research.[122,123] SANMF is just one of many methods that can be applied to understand the complexity of social exposures. It will be important to describe quantitatively how social factors interact and affect health outcomes, but there is no consensus on the best methods to use for this type of research. As highlighted throughout this dissertation, we must continue to fund research focused on upstream SDOH. We did see in our work that they have significant, and potentially causal associations with health outcomes, but there are many questions left unexplored. For example, we saw a potential time dependency of the effects of SDOH. While we focused on SDOH from young adulthood to middle age, it is unclear how early-life SDOH impact mid-life health outcomes. As shown by Allen et al., there are distinct and identifiable CVH trajectories beginning as early as age 8.[86] Havranek et al. highlighted the need to not only focus on early-life SDOH, but to also study the intergenerational transmission of social advantage and to examine the effects on CVD.[46]

Having identified and described five unique SDOH clusters, obvious questions are whether we can now go back with a reductionist approach to find simple, easily measured variables that carry a similar information content and whether timing of the SDOH exposures was an important aspect of our main findings. We plan on exploring whether the clusters created are associated more strongly with some individual CVH components than others to determine whether there are one or a few components driving the cluster associations with the composite CVH measure. In this initial set of chapters, we used all 48 variables and all frequent patterns of the variables to create the clusters. We would like to refine the clusters and determine whether there is a minimum number of variables that offer the same level of information. In future work, other approaches—such as term frequency-inverse document frequency[124] (TF-IDF)—can be used to create clusters of less frequent, but unique SDOH patterns to target a smaller group of individuals at highest risk for low CVH. A reduction in the number of variables used to define the clusters may help with extrapolation to other settings. It is also possible that screening for food security, access to medical care, and the ability to pay for medical care may help to allow for streamlined screening for social risks in clinical settings. By screening for these factors, and understanding which SDOH occur concurrently in the clusters, these variables may provide insight into the other SDOH domains. By focusing on these three SDOH exposures, we are not moving back to the traditional approach of studying SDOH isolation, but are instead identifying which SDOH factors may be most important in determining CVH and CVD outcomes in the context of all other social exposures.

We may also wish to quantitatively assess whether timing of the SDOH exposure matters. In practice, an understanding of the timing of exposure to given SDOH and their associations with CVH and CVD is helpful to create timely and targeted interventions to address SDOH; however, it is unclear whether incorporation of the

SDOH exposure timing offers improved predictive performance. Empirically, timing did matter because we showed distinct timing of exposures in the clusters that were associated with CVH and CVD events. The next goal will be to validate the clusters in other settings, so that they may be used for streamlined screening of social risks and to examine their association with other health outcomes.

Ultimately, we want this work to inform the design, implementation, and evaluation of timely and multi-component social interventions and policies to address SDOH in order to improve CVH and reduce the burden and disparities of incident CVD. There are very clear distinctions between the interventions needed to address SDOH, social risk factors, and social needs. As highlighted by Green and Zook, interventions targeting SDOH "can be categorized as an upstream, communitywide intervention to address the root causes and conditions (for example, economic instability) that contribute to poor health."[125] Natural experiments, such as policy implementations, may assist researchers and public health professionals in identifying the interventions which address SDOH and minimize the burden of CVD.[35,36,126,127] For example, Rehkopf et al. have demonstrated an association between Earned Income Tax Credit funds and a decrease in diastolic blood pressure in the short-term.[36] As part of the Biden Administration's American Rescue Plan Act, signed into law in March 2021, the Child Tax Credit was expanded, impacting all families except the families in the highest income categories.[128] The tax credits were increased by $1,000 for children under age 18 and by $1,600 for children under age 6. All children 17 years old and younger now qualify, as opposed to 16 years old and younger previously. This type of broad policy intervention may have direct impacts on public health and cardiovascular health. Columbia University estimates that this plan could cut child poverty by more than half, but the cardiovascular health implications for parents and children in the short- and long-term are unclear.[129] Policy models, such as the validated microsimulation IMPACT Model developed at the University of Liverpool,[130,131] can help in simulating the long-term impact of certain interventions and policies on CVD outcomes and assess whether the potential benefits or harms differ by demographic groups.[132] This work is much needed to generate support for new social policies, and the expansion of existing policies, to decrease the public health burden of CVD and minimize health disparities.

5.5 Summary

This dissertation demonstrates that a novel machine learning method can be used to understand the complexity of SDOH exposures over time. The novel clusters representing patterns of SDOH from young adulthood to middle age were predictive of mid-life CVH and associated with mid-life CVH status and CVD events. More work must be done to validate these clusters in other settings, with the ultimate goal of informing programs looking to develop targeted, timely, and multi-component interventions to address SDOH and improve CVH.

TABLES AND FIGURES

Table 1: Variable descriptions for each of the 48 individual- and neighborhood-level social determinant of health (SDOH) and psychosocial variables assessed in CARDIA participants

SDOH Domain	Variable	Description
Education	Maximum Education Achieved	Maximum value of highest grade completed across all exams. Range: 1 – 20+
Economic Stability	Income	Total combined gross family income for the past 12 months from all sources. Categories: <$5k to $15,999; $16k to $34,999; $35k to $49,999; $50k to $74,999; $75k and greater;
	Home Ownership	Type of ownership for participant's current home. Categories: Owned or Being Bought; Rented for Money; Occupied without Payment; Other;
	Hard to Meet Demands	Self-reported difficulty meeting demands from job, family, friends, or school. Categories: Very Hard; Hard; Somewhat Hard; Not Very Hard;
	Hard to Pay for Basics	Self-reported difficulty paying for basics like food, medical care (separate question after Exams Y0, Y2, Y7), and heating. Categories: Very Hard; Hard; Somewhat Hard; Not Very Hard;
	Trouble Making Ends Meet	Self-reported trouble making ends meet (MEM). Categories: Frequent Trouble MEM; Occasional Trouble MEM; Hardly Ever Trouble MEM; Never Trouble MEM;
	Hard to Pay for Medical Care	Self-reported difficulty paying for medical care for participant and family.

Economic Stability		Categories: Very Hard; Hard; Somewhat Hard; Not Very Hard;
	Assets	All assets including family's checking and savings accounts, any stocks and bonds, and real estate (including principal home). Categories: <$500-$4,999; $5k-$19,999; 20k-$49,999; 50k-$99,999; 100k-$199,999; 200k-$499,999; 500k+;
	Debt	Total family debt from household for things like credit card charges, medical or legal bills, and loans from banks or relatives. Categories: <$500; $500-$4,999; 5k-$9,999; 10k-$19,999; 20k-$49,999; 50k+;
	Food Security	Self-reported food security related to amount of food and kinds of food. Categories: Enough Food and Kinds; Not Always Enough or Kinds;
	Employment Status	Self-report of being unemployed, laid off, or looking for work.[1] Categories: No; Yes;
	Occupation Status- TSEI	The Stevens and Cho total-based Socioeconomic Index (TSEI) was used to measure occupational class as a "predicted prestige ranking" linked to each occupation code (higher scores indicate higher predicted prestige rankings).[2,3] Categories: Lower, Mid-Range, Higher;
	Karasek Job Strain Questionnaire- Job Decision Latitude	Job "decision latitude" is based on a sum of the composite skill discretion and decision authority scores. The skill discretion score is a weighted sum of the responses to questions about learning new things on the job, being creative on the job, the job requiring a high level of skill, doing a variety of things on the job,

		and opportunities for developing one's own special abilities on the job. The decision authority score is a weighted sum of the responses to questions about the job allowing one to make a lot of his/her own decisions, having a lot of say about what happens on the job, and having freedom to decide how to do one's job.[4] Categories: Lower, Mid-Range, Higher;
	Karasek Job Strain Questionnaire- Psychological Job Demands	"Psychological job demands" is based on the weighted sum of questions related to working fast, working hard, excessive amounts of work, enough time to get the job done, and freedom from conflicting demands of others.[4] Categories: Lower, Mid-Range, Higher;
Social and Community Context	Household Size	Number of people currently living in participant's the household, including the participant. Categories: 1 Person; 2 People; 3-5 People; ≥6 People;
	Children- Yes/No	Yes or no to having children or step-children. Categories: No; Yes;
	Children- Living in House	Children or step-children living in the participants home; Categories: No; Yes;
	Marital Status	Current marital status. Categories: Marriage-Like Relationship; Married; Never Married; Separated or Divorced; Widowed;
	Social Support Questionnaire- Instrumental Support	Self-reported instrumental support focused on if participants have support of friends, spouse/mate, and/or family member when needing help with household tasks, a ride, help when too sick to take care of themselves, and a loan of money.[5] Categories: Lower, Mid-Range, Higher;
	Social Support Questionnaire- Emotional Support	Self-reported emotional support focused on if participants have support of friends, spouse/mate,

Social and Community Context		and/or family member when worried about personal problems.[5] Categories: Lower, Mid-Range, Higher;
	Social Support Questionnaire- Network Adequacy	Self-reported network adequacy focused on how frequently participants feel lonely, find themselves wishing someone would comfort them with a hug or other physical sign of affection, feel other people really care for them, and wish they had more close friends.[5] Categories: Lower, Mid-Range, Higher;
	Discrimination- Any Discrimination	Number of domains where participants experienced any discrimination (based on gender, race, social class, sexual preference, religion, weight, or age). Domains include at school, getting a job, at work, getting housing, getting medical care, and on the street or in a public setting.[6] Categories: 0 Domains; 1-2 Domains; ≥3 Domains;
	Discrimination- Racial Discrimination	Number of domains where participants experienced racial discrimination. Domains include at school, getting a job, at work, getting housing, getting medical care, and on the street or in a public setting.[6] Categories: 0 Domains; 1-2 Domains; ≥3 Domains;
	Social Network	Self-reported social network size (number of close friends and relatives).[7] Categories: 0-3 Ties, 4-7 Ties; 8-10 Ties; 11-14 Ties; ≥15 Ties.
	Subjective Social Standing	Perceived rank on social hierarchy in the United States. Participants were shown a picture of a ladder with 10 rungs representing where people stand in the United States; the top represented those who were the best-off and have the most money, education, and the most

		respected jobs. Participants were asked to place themselves on the ladder.[8] Categories: Lower, Mid-Range, Higher;
	Social Support and Conflict Questionnaire- Supportive Interactions	Frequency of exposure to supportive social interactions. Questions were related to how much family or friends care about you, how much they understand the way you feel, how much you can rely on them if you need to talk about worries, and how much you can open up to them if you have a serious problem.[9] Categories: Lower, Mid-Range, Higher;
	Social Support and Conflict Questionnaire- Negative Interactions	Frequency of exposure to negative social interactions. Questions were related to how much family and friends criticize you, let you down when you are counting on them, and get on your nerves.[9] Categories: Lower, Mid-Range, Higher;
Neighborhood and Built Environment- Self-Reported by Participant	Change in Residence	Number of times the participant changed residence in the past two years or since last exam. Categories: 0 Times; 1 Time; 2 or more times;
	Neighborhood Cohesion	Perceived neighborhood cohesion based on whether people are willing to help neighborhoods, whether it's a close-knit neighborhood, whether people can be trusted, whether people don't get along, and whether people do not share the same values.[10] Categories: Lower, Mid-Range, Higher;
	Neighborhood Environment	Number of resources available in participants neighborhood (within 10-15 minute walk from home) including exercise facilities, parks, grocery stores, fast food restaurants (reverse coded), sit-down restaurants, subway/bus/trolley stops, sidewalks, walking/bike paths.[11]

		Categories: 0-3 Resources; 4-6 Resources; 7-8 Resources;
Neighborhood and Built Environment- Census Tract Level	Percent Population White Race	Percent of population in participant’s census tract who identify as white race.[12] Categories: Lower, Mid-Range, Higher;
	Percent Population Education <High School	Percent of population in participant’s census tract with less than high school education among those aged 25 and older.[12] Categories: Lower, Mid-Range, Higher;
	Percent Population <150% Federal Poverty Level	Percent of population in the participant’s census tract with income below 150% of the federal poverty level.[12] Categories: Lower, Mid-Range, Higher;
	Median Income	Median income in the participant’s census tract.[12] Categories: Lower, Mid-Range, Higher;
	Percent Population Professional/Management Occupation	Percent of population in participant’s census tract in professional or management occupations among those ages 16 years or older.[12] Categories: Lower, Mid-Range, Higher;
	Percent Population Unemployed	Percent unemployed in participant’s census tract among those ages 16 years and older.[12] Categories: Lower, Mid-Range, Higher;
	Median Rent	Median gross rent (renter-occupied housing) in participant’s census tract.[12] Categories: Lower, Mid-Range, Higher;
	Percent Owner-Occupied Housing Units	Percent owner-occupied housing units out of all occupied housing units in participant’s census tract.[12] Categories: Lower, Mid-Range, Higher;
Neighborhood and Built Environment- Census Tract Level	Percent Vacant Housing Units	Percent vacant housing units out of all housing units in participant’s census tract.[12] Categories: Lower, Mid-Range, Higher;

Neighborhood and Built Environment- Census Tract Level	Aggregate Value Housing Units	Aggregate value of owner-occupied housing units in participant’s census tract.[12] Categories: Lower, Mid-Range, Higher;
	Racial Segregation (Gi* Statistic)	Own-group racial segregation measured as the local Getis-Ord Gi* statistic. The Gi* statistic is a z-score indicating how many standard deviations the participant’s census tract and its neighboring tracts is from the racial composition of the larger metropolitan area or county.[13] Categories: <0 SD- Lower own-group representation in census tract compared to larger area; 0-1.96 SD- Racial integration or similar own-group representation in census tract compared to larger area: >1.96 SD- Greater own-group representation in census tract compared to larger area;
	SES Deprivation Score	SES deprivation score calculated as the first factor score from a principal component analysis of four tract-level indicators from participant’s census tract: median household income, proportion of the population at or below the 150% federal poverty level, proportion of the population aged 25 or greater with less than a high school education, and proportion of the population aged 25 or greater with a college degree or higher.[14] Categories: Lower, Mid-Range, Higher;
	Fast Food and Convenience Stores	Percent of all stores in participant’s census tract that are fast food or convenience stores within 3km.[15] Categories: Lower, Mid-Range, Higher;
	Supermarkets	Percent of all food facilities that are supermarkets within 5km of participant’s census tract.[15] Categories: Lower, Mid-Range, Higher;

	Physical Activity Facilities	Count of physical activity facilities within 3km of participant's census tract.[14] Categories: Lower, Mid-Range, Higher;
Health and Health Care	Health Care Barriers	Numbers of barriers reported for health care: insurance barrier (lack of coverage), regular care barrier (no usual source of care), or expense barrier (did not seek care because it was too expensive or health insurance did not cover it).[16] Categories: 0 barriers; 1 barrier; 2 barriers; 3 barriers;
	Health Insurance Coverage	Number of months without health care coverage in the past two years.[16] Categories: 1-6 months; 7 months – 1 year; >1 – 2 years;
Psychosocial	Material and Psychological Wellbeing- CES-D Questionnaire	Score from Center for Epidemiologic Studies Depression (CES-D) scale. The score ranges from 0 to 60 and higher scores indicate more depressive symptoms. The score was dichotomized to indicate depression (≥16 score) or not.[17,18] Categories: No; Yes;
	Chronic Burden Scale	Number of domains where stress reported for longer than 6 months. Domains include 1) personal serious ongoing health problem, 2) serious ongoing health problem of parents, child, or others close to you, 3) ongoing difficulties with job or ability to work, 4) ongoing financial strain, and 5) ongoing difficulties in a relationship with someone close to you.[19,20] Categories: 0 domains; 1 – 2 domains; 3 – 5 domains;

Table 2: Summary of demographic, education, cardiovascular health, and select SDOH characteristics for train and test sets among CARDIA study participants

	Train		Test		Overall Cohort	
	N=2,467		N=1,055		N=3,522	
Race						
Black	1,140	46.2%	492	46.6%	1,632	46.3%
White	1,327	53.8%	563	53.4%	1,890	53.7%
Sex						
Female	1,386	56.2%	581	55.1%	1,967	55.9%
Male	1,081	43.8%	474	44.9%	1,555	44.1%
Highest Degree Earned						
Elementary/Jr. High/Some High School	57	2.3%	14	1.3%	71	2.0%
High School Graduate	285	11.6%	131	12.4%	416	11.8%
Some College	715	29.0%	332	31.5%	1,047	29.7%
College Graduate (4-Year)	565	22.9%	244	23.1%	809	23.0%
Graduate School	845	34.3%	334	31.7%	1,179	33.5%
Age at Baseline						
mean SD	25.3	3.5	25.1	3.5	25.3	3.5
Age at Outcome						
mean SD	48.1	2.3	48.1	2.3	48.1	2.3
Cardiovascular Health at Baseline, Ages 18-30[a]						
Score, mean SD (Range 0-14 points)	10.4	1.9	10.4	1.9	10.4	1.8
Prevalence of Low CVH	162	6.8%	69	6.7%	231	6.8%
Prevalence of Moderate/High CVH	2,232	93.2%	958	93.3%	3,190	93.3%
Cardiovascular Health at Outcome, Ages ≥45						

Score, mean SD (Range 0-14 points)	8.9	2.3	8.8	2.3	8.9	2.3
Prevalence of Low CVH	710	28.8%	303	28.7%	1,013	28.8%
Prevalence of Moderate/High CVH	1,757	71.2%	752	71.3%	2,509	71.2%
Income at Y5[b]						
<$5k to $15,999	431	18.9%	185	18.8%	616	18.8%
$16k to $34,999	818	35.8%	364	36.9%	1,182	36.2%
$35k to $49,999	450	19.7%	183	18.6%	633	19.4%
$50k to $74,999	350	15.3%	149	15.1%	499	15.3%
$75k and greater	235	10.3%	105	10.7%	340	10.4%
Household Size at Baseline, mean SD[c]	2.9	1.7	3.0	1.7	3.00	1.7
Neighborhood Population <150% FPL at Baseline, mean SD %[d]	29.0%	17.0%	29.0%	17.0%	29.0%	17.0%
Healthcare Access Barriers at Y7[e]						
0 Barriers	1589	70.1%	689	70.1%	2,278	70.1%
1 Barrier	441	19.4%	200	20.4%	641	19.7%
2 Barriers	184	8.1%	72	7.3%	256	7.9%
3 Barriers	54	2.4%	22	2.2%	76	2.3%

Data are N % unless otherwise noted.

Abbreviations: CVH = Cardiovascular Health; SDOH = Social Determinants of Health; SD = Standard Deviation; FPL = Federal Poverty Level;

[a] Participants missing CVH at Baseline: Training- 73, Testing- 28

[b] Participants missing Income at Y5: Training- 183, Testing- 69

[c] Participants missing Household Size at Baseline: Training- 1, Testing- 0

[d] Participants missing proportion <150% Federal Poverty Level at Baseline: Training- 18, Testing- 7

[e] Participants missing Healthcare Access Barriers at Y7: Training- 199, Testing- 72

Table 3: Characterization of the five SDOH clusters with their top 10 time-dependent SDOH patterns from NMF

Cluster 1

Domain	MC	Pattern
ES	2.49	Not Very Hard to Meet Demands
	1.59	Lower Psychological Job Demands
	1.88	Food Security: EFK – EFK
	1.79	Paying for Medical Care: NVH – NVH
NBE	1.86	No Change in Residence
SCC	2.00	Mid-Range Network Adequacy
	2.01	Mid-Range Subjective Social Standing
	2.08	Mid-Range Subjective Social Standing
	1.81	Higher Supportive Interactions
HHC	2.09	Health Care Access Barriers: Zero – Zero

Cluster 2

Domain	MC	Pattern
ES	1.81	Not Very Hard to Meet Demands
	1.75	Food Security: EFK – EFK
	1.39	Paying for Medical Care: NVH – NVH
	1.32	Paying for Basics: NVH – NVH – NVH – NVH
	1.56	Employed: Yes – Yes – Yes – Yes
SCC	1.91	Mid-Range Network Adequacy
	1.39	Mid-Range Emotional Support
	1.46	Mid-Range Instrumental Support
	1.30	Mid-Range Subjective Social Standing
HHC	1.59	Health Care Access Barriers: Zero – Zero

Cluster 3

Domain	MC	Pattern
ES	1.13	Somewhat Hard to Meet Demands
	1.26	Lower Job Decision Latitude
	1.51	Occasional Trouble Making Ends Meet
	1.27	Occupation Status: Lower – Lower
NBE	1.12	One Change in Residence
SCC	1.59	Lower Emotional Support
	1.26	Lower Instrumental Support
	1.36	Lower Network Adequacy
	1.46	Higher Negative Interactions
	1.26	Higher Negative Interactions

Cluster 4

Domain	MC	Pattern
ES	1.19	Not Very Hard to Meet Demands
NBE	2.02	% HS Graduates: Higher – Higher – Higher
	1.73	% Below 150% Poverty: Higher – Higher – Higher
	1.57	Median Income: Lower – Lower – Lower
	1.98	% Professional Occupations: Lower – Lower – Lower
	1.65	SES Deprivation: Higher – Higher – Lower
	1.17	% Unemployed: Higher – Higher – Higher
	1.22	Housing Unit Value: Lower – Lower – Lower
	1.41	Median Rent: Lower – Lower – Lower
SCC	1.09	Higher Supportive Interactions

Cluster 5

Domain		Pattern
ES	2.00	Higher Job Decision Latitude
	1.57	Occupation Status: Higher – Higher
	2.13	Food Security: EFK – EFK
	2.11	Paying for Medical Care: NVH – NVH
	1.55	Home Ownership: Owned -- Owned – Owned
NBE	1.59	One Change in Residence

Key

Abbreviations: SDOH = Social Determinants of Health; NMF = Non-negative Matrix Factorization; MC = Membership Coefficient; EFK = Enough Food & Kinds; NVH = Not Very Hard; HS = High School; SES = Socioeconomic Status;

Age Windows: 18-24, 25-34, 35-44, 45 up to outcome

Domains:

ES: Economic Stability

	2.22	Higher Subjective Social Standing
SCC	1.56	Higher Supportive Interactions
	1.82	Higher Subjective Social Standing
HHC	1.55	Health Care Access Barriers: Zero – Zero

NBE: Neighborhood and Build Environment
SCC: Social and Community Context
HHC: Health and Health Care

Table 4: Predictive performance (discrimination) of selected models for mid-life cardiovascular health by predictor group

Predictor Group	Model	Set	AUC
Base	Logistic Regression	Train	0.648
		Test	0.635
Base + CVH	Logistic Regression	Train	0.768
		Test	0.764
Base + SDOH Clusters	Logistic Regression	Train	0.713
		Test	0.703
Base + CVH + SDOH Clusters	Logistic Regression[a]	Train	0.784
		Test	0.776
	Neural Network	Train	0.801
		Test	0.770
	Random Forest	Train	1.000
		Test	0.759
Base + CVH + Subgraphs	Neural Network	Train	0.875
		Test	0.734
	Random Forest	Train	1.000
		Test	0.717
	Ridge Regression	Train	0.862
		Test	0.738
	Lasso Regression[b]	Train	0.805
		Test	0.770

Abbreviations: CVH = Cardiovascular Health; AUC = Area Under the Curve;
[a]Highest AUC value for the Base + CVH + Clusters models
[b]Highest AUC value for the Base + CVH + Subgraphs models

Table 5: Summary of demographic and education measures overall and by race

	All	Black Participants	White Participants
	N=3,522	N=1,632 (46.3%)	N=1,890 (53.7%)
Sex			
Female	55.8%	59.1%	53.0%
Highest Degree Earned			
Elementary/Jr. High/Some High School	2.0%	2.9%	1.2%
High School Graduate	11.8%	16.2%	8.0%
Some College	29.7%	40.1%	20.7%
College Graduate (4-Year)	23.0%	20.2%	25.4%
Graduate School	33.5%	20.6%	44.6%
Mean Age at Baseline (years, SD)	25.3, 3.5	24.7, 3.7	25.7, 3.3
Mean Age at Outcome Measurement (years, SD)	48.1, 2.3	48.1, 2.4	48.0, 2.1

Data are % unless otherwise noted.

Abbreviations: SD = Standard Deviation;

Table 6: Multivariable-adjusted logistic regression analysis of associations of SDOH clusters through young adulthood with poor CVH status in middle age among all participants

SDOH Cluster	MV-Adjusted[a] Odds Ratio for Poor CVH (95%CI)	P-Value	Time-Dependent Patterns in SDOH Cluster
SDOH Cluster 1	0.76 (0.67 - 0.87)	<0.001	Cluster representing individual factors of being able to meet demands, lower psychological job demands, enough food to eat, being able to pay for medical care, zero health care access barriers, no change in residence, mid-range network adequacy, mid-range subjective social standing, and higher supportive interactions.
SDOH Cluster 2	0.78 (0.64 - 0.94)	0.008	Cluster representing individual factors of being able to meet demands, enough food to eat, being able to pay for medical care, being able to pay for basics, being employed, zero health care access barriers, mid-range network adequacy, mid-range emotional support, mid-range instrumental support, and higher supportive interactions.
SDOH Cluster 3	1.05 (0.95 - 1.17)	0.377	Cluster representing individual factors of having some trouble meeting demands, lower job decision latitude, occasional trouble making ends meet, lower occupation status, >1 to 2 years without health insurance coverage, lower emotional support, lower instrumental support, lower network adequacy, higher negative interactions, lower and subjective social standing.
SDOH Cluster 4	0.91 (0.81 - 1.01)	0.080	Cluster representing individual factors of being able to meet demands and zero health care access barriers, and neighborhood-level factors of higher percentage of high school graduates, higher percentage below 150% of poverty, lower median income, lower percentage with professional occupations, higher and later lower socioeconomic deprivation, higher percentage unemployed, lower housing unit value, and lower median rent.

SDOH Cluster 5	0.74 (0.65 - 0.84)	<0.001	Cluster representing higher job decision latitude, higher occupation status, enough food to eat, being able to pay for medical care, owning a home, one change in residence, zero health care access barriers, higher subjective social standing, and higher supportive interactions.

[a]Adjusted for age, sex, race, center, education, and baseline CVH score

Abbreviations: CI = Confidence Interval; CVH = Cardiovascular Health; MV = Multivariable; SDOH = Social Determinants of Health;

Table 7: Multivariable-adjusted logistic regression analysis of associations of SDOH clusters (created using all participants) through young adulthood with poor CVH status in middle age among Black and white participants

	Black Participants		White Participants	
SDOH Cluster	**MV-Adjusted[a] Odds Ratio for Poor CVH (95%CI)**	**P**	**MV-Adjusted[a] Odds Ratio for Poor CVH (95%CI)**	**P**
SDOH Cluster 1	0.71 (0.59 - 0.85)	<0.001	0.81 (0.66 - 0.99)	0.041
SDOH Cluster 2	0.72 (0.56 - 0.92)	0.008	0.83 (0.61 - 1.13)	0.238
SDOH Cluster 3	1.04 (0.91 - 1.20)	0.553	1.07 (0.89 - 1.29)	0.453
SDOH Cluster 4	0.86 (0.75 - 0.99)	0.043	0.95 (0.80 - 1.13)	0.591
SDOH Cluster 5	0.82 (0.67 - 1.00)	0.049	0.71 (0.59 - 0.86)	<0.001

[a] Adjusted for age, sex, center, education, and baseline CVH score

Abbreviations: CI = Confidence Interval; CVH = Cardiovascular Health; MV = Multivariable; SDOH = Social Determinants of Health;

Cluster Summaries:

Cluster 1) economically stable with less psychologically demanding job, mid-range social support, zero health care access barriers, and no change in residence;

Cluster 2) economically stable with employment, mid-range social support, and zero health care access barriers;

Cluster 3) some difficulty economically with lower status job, low social support, and years without health insurance;

Cluster 4) no difficulty meeting demands, with vulnerable neighborhood environment, and

Cluster 5) economically wealthy with high status job, higher social support, and change in residence during late 20s and early 30s.

Table 8: Summary of demographic, education, CVD risk factor, and social determinants of health measures overall and by cohort

	All Participants	CVD Events Cohort	CAC Cohort	LVMI Cohort	CVH Cohort
	N=5,112	N=4,853	N=3,100	N=3,349	N=3,448
Race					
Black	2,637 (51.6%)	2,460 (50.7%)	1,384 (44.6%)	1,578 (47.1%)	1,589 (46.1%)
White	2,475 (48.4%)	2,393 (49.3%)	1,716 (55.4%)	1,771 (52.9%)	1,859 (53.9%)
Sex					
Female	2,785 (54.5%)	2,698 (55.6%)	1,763 (56.9%)	1,930 (57.6%)	1,940 (56.3%)
Male	2,327 (45.5%)	2,155 (44.4%)	1,337 (43.1%)	1,419 (42.4%)	1,508 (43.7%)
Baseline Age in Years (mean, SD)	24.8 (3.7)	24.9 (3.7)	25.9 (3.1)	25.1 (3.5)	25.2 (3.5)
Highest Degree Earned					
Elementary/Jr. High/ Some High School	177 (3.5%)	154 (3.2%)	62 (2.0%)	57 (1.7%)	66 (1.9%)
High School Graduate	853 (16.7%)	771 (15.9%)	357 (11.5%)	395 (11.8%)	399 (11.6%)
Some College	1,640 (32.1%)	1,553 (32.0%)	927 (29.9%)	978 (29.2%)	1,025 (29.7%)
College Graduate (4-Year)	1,042 (20.4%)	1,007 (20.8%)	703 (22.7%)	775 (23.1%)	796 (23.1%)
Graduate School	1,400 (27.4%)	1,368 (28.2%)	1,051 (33.9%)	1,144 (34.1%)	1,162 (33.7%)
Smoking at Baseline					
Current Smoker	1,544 (30.4%)	1,431 (29.7%)	835 (27.1%)	866 (26.0%)	904 (26.4%)
Former Smoker ≥ 12 months	676 (13.3%)	642 (13.3%)	458 (14.9%)	463 (13.9%)	468 (13.7%)
Never or Quit ≥ 12 months	2,856 (56.3%)	2,747 (57.0%)	1,788 (58.0%)	2,003 (60.1%)	2,055 (60.0%)
Diabetes Prevalence at Baseline	32 (0.6%)	20 (0.4%)	13 (0.4%)	9 (0.3%)	14 (0.4%)
Total Cholesterol- mg/dL (mean, SD)	176.8 (33.5)	176.7 (33.1)	178.3 (32.9)	176.7 (32.7)	177.2 (32.9)
HDL Cholesterol- mg/dL (mean, SD)	53.2 (13.2)	53.3 (13.2)	53.6 (13.0)	53.8 (12.7)	53.5 (12.9)
SBP- mmHg (mean, SD)	110.4 (11.0)	110.2 (10.8)	110.0 (10.7)	109.7 (10.6)	110.1 (10.7)
Hypertension Medication Use	115 (2.2%)	102 (2.1%)	67 (2.2%)	58 (1.7%)	66 (1.9%)

Prevalence at Baseline					
CVH Score at Baseline (mean, SD)*	10.2, 1.9	10.3, 1.9	10.4, 1.9	10.5, 1.8	10.4, 1.8
Income at Y5					
<$5k to $15,999	909 (21.3%)	842 (20.6%)	499 (17.5%)	574 (18.9%)	597 (18.6%)
$16k to $34,999	1,550 (36.3%)	1,486 (36.4%)	1,008 (35.3%)	1,074 (35.3%)	1,156 (36.1%)
$35k to $49,999	791 (18.5%)	766 (18.8%)	583 (20.4%)	591 (19.4%)	627 (19.6%)
$50k to $74,999	617 (14.5%)	595 (14.6%)	462 (16.2%)	470 (15.5%)	486 (15.2%)
$75k and greater	402 (9.4%)	396 (9.7%)	305 (10.7%)	332 (10.9%)	338 (10.5%)
Household Size at Baseline, mean SD	3.1, 1.7	3.0, 1.7	2.9 (1.7)	3.0 (1.7)	3.0 (1.7)
Neighborhood Population <150% FPL at Baseline, mean SD %	29.6%, 17.1%	29.6%, 17.1%	29.2%, 17.0%	29.3% (17.1%)	29.4% (17.1%)
Healthcare Access Barriers at Y7					
0 Barriers	2,812 (69.5%)	2,705 (69.6%)	1,963 (70.7%)	2,111 (71.1%)	2,235 (70.2%)
1 Barrier	801 (19.8%)	770 (19.8%)	536 (19.3%)	569 (19.2%)	623 (19.6%)
2 Barriers	341 (8.4%)	318 (8.2%)	210 (7.6%)	220 (7.4%)	250 (7.9%)
3 Barriers	94 (2.3%)	92 (2.4%)	67 (2.4%)	68 (2.3%)	74 (2.3%)

Data are N (%) unless otherwise noted.

Abbreviations: CVD = Cardiovascular Disease; CVH = Cardiovascular Health; CAC = Coronary Artery Calcification; LVMI = Left Ventricular Mass Index; SD = Standard Deviation; HDL = High-density Lipoprotein; SBP = Systolic Blood Pressure; Y = Exam Year;

*Range of CVH Score at Baseline is 0-14.

Table 9: Summary of CVD events, subclinical CVD, and CVH measures by cohort

	CVD Events Cohort	CAC Cohort	LVMI Cohort	CVH Cohort
	N=4,853	N=3,100	N=3,349	N=3,448
CVD Events Outcome				
CVD Event	252 (5.2%)	160 (5.2%)	108 (3.2%)	149 (4.6%)
Censored	4,601 (94.8%)	2,940 (94.8%)	3,241 (96.8%)	3,289 (95.4%)
Age at Event (mean, SD)	52.4, 4.3	53.8, 4.2	54.6, 4.1	53.6, 4.1
Age for Censored Participants (mean, SD)	58.4, 3.8	59.6, 3.2	58.9, 3.5	58.9, 3.6
CAC Outcome				
Presence of CAC	--	716 (23.1%)	--	--
Absence of CAC	--	2,384 (76.9%)	--	--
LVMI Outcome- g/m$^{2.7}$ (mean, SD)	--	--	40.1, 11.6	--
Mid-Life CVH Outcome (mean, SD)*	--	--	--	8.9, 2.3

Abbreviations: CVD = Cardiovascular Disease; CVH = Cardiovascular Health; CAC = Coronary Artery Calcification; LVMI = Left Ventricular Mass Index; SD= Standard Deviation;

*Range of CVH Score at Outcome is 0-14.

Table 10: Hazard ratios for CVD events, unadjusted and adjusted for age at baseline, sex, race, and education

	Model 1		Model 2		Model 3	
	Unadjusted HR (95% CI)	**P-Value**	**Adjusted HR (95% CI)**	**P-Value**	**Adjusted HR (95% CI)**	**P-Value**
SDOH Clusters						
Cluster 1	0.81 (0.71 - 0.94)	0.004	0.85 (0.73 - 0.99)	0.039	0.93 (0.80 - 1.09)	0.385
Cluster 2	0.86 (0.71 - 1.06)	0.153	1.06 (0.83 - 1.37)	0.633	1.10 (0.85 - 1.41)	0.470
Cluster 3	1.17 (1.05 - 1.31)	0.006	1.17 (1.03 - 1.31)	0.009	1.16 (1.03 - 1.30)	0.012
Cluster 4	0.99 (0.88 - 1.12)	0.913	1.01 (0.89 - 1.15)	0.857	1.01 (0.89 - 1.15)	0.849
Cluster 5	0.99 (0.82 - 1.19)	0.923	0.87 (0.71 - 1.07)	0.188	0.88 (0.73 - 1.08)	0.230
Age at Baseline Exam (per 1 year higher)	--	--	1.11 (1.04 - 1.20)	0.004	1.13 (1.05 - 1.22)	<0.001
Sex						
Female	--	--	0.54 (0.42 - 0.70)	<0.001	0.57 (0.44 - 0.73)	<0.001
Male	--	--	1.0 (ref)	--	1.0 (ref)	--
Race						
Black	--	--	1.38 (1.05 - 1.81)	0.022	1.23 (0.93 - 1.62)	0.149
White	--	--	1.0 (ref)	--	1.0 (ref)	--
Maximum Years of Education (per year)	--	--	--	--	0.90 (0.85 - 0.95)	<0.001

Abbreviations: HR = Hazard Ratio; CI = Confidence Interval;

Table 11: Hazard ratios for CVD events, adjusted for baseline CVH score, age at baseline, sex, race, and education

	Model 4		Model 5	
	Adjusted HR (95% CI)	**P-Value**	**Adjusted HR (95% CI)**	**P-Value**
SDOH Clusters				
Cluster 1	0.97 (0.83 - 1.14)	0.745	1.00 (0.85 - 1.17)	0.975
Cluster 2	1.05 (0.81 - 1.36)	0.693	1.06 (0.82 - 1.37)	0.648
Cluster 3	1.13 (1.01 - 1.27)	0.038	1.13 (1.01 - 1.27)	0.039
Cluster 4	0.99 (0.87 - 1.13)	0.897	0.99 (0.87 - 1.13)	0.899
Cluster 5	0.88 (0.72 - 1.09)	0.239	0.89 (0.72 - 1.09)	0.254
Baseline CVH Score	0.71 (0.66 - 0.77)	<0.001	0.72 (0.68 - 0.78)	<0.001
Age at Baseline Exam (per 1 year higher)	1.08 (1.00 - 1.16)	0.045	1.08 (1.01 - 1.17)	0.035
Sex				
Female	0.54 (0.42 - 0.70)	<0.001	0.54 (0.42 - 0.71)	<0.001
Male	1.0 (ref)	--	1.0 (ref)	--
Race				
Black	1.10 (0.83 - 1.45)	0.519	1.07 (0.81 - 1.43)	0.625
White	1.0 (ref)	--	1.0 (ref)	--
Maximum Years of Education (per year)	--	--	0.97 (0.92 - 1.03)	0.251

Abbreviations: HR = Hazard Ratio; CI = Confidence Interval;

Table 12: Race-stratified analyses partially and fully adjusted by covariates

	Black Participants			
	Partially Adjusted		Fully Adjusted	
	Adjusted HR (95% CI)	P-Value	Adjusted HR (95% CI)	P-Value
SDOH Clusters				
Cluster 1	0.99 (0.79 - 1.24)	0.922	1.11 (0.87 - 1.41)	0.419
Cluster 2	1.01 (0.73 - 1.42)	0.934	0.98 (0.69 - 1.39)	0.916
Cluster 3	1.20 (1.04 - 1.39)	0.014	1.17 (1.01 - 1.35)	0.042
Cluster 4	1.01 (0.86 - 1.19)	0.915	1.03 (0.87 - 1.22)	0.737
Cluster 5	0.92 (0.71 - 1.20)	0.555	0.92 (0.71 - 1.20)	0.559
Age at Baseline Exam (per 1 year higher)	1.12 (1.02 - 1.23)	0.014	1.07 (0.98 - 1.18)	0.131
Sex				
Female	0.65 (0.47 - 0.90)	0.009	0.58 (0.41 - 0.81)	0.002
Male	1.0 (ref)	--	1.0 (ref)	--
Baseline CVH Score	--	--	0.74 (0.67 - 0.81)	<0.001
Maximum Years of Education (per year)	--	--	0.99 (0.92 - 1.07)	0.837
	White Participants			
	Partially Adjusted		**Fully Adjusted**	
	Adjusted HR (95% CI)	**P-Value**	**Adjusted HR (95% CI)**	**P-Value**
SDOH Clusters				
Cluster 1	0.76 (0.62 - 0.93)	0.006	0.93 (0.75 - 1.14)	0.484
Cluster 2	1.09 (0.74 - 1.61)	0.656	1.14 (0.77 - 1.69)	0.505
Cluster 3	1.11 (0.92 - 1.35)	0.268	1.08 (0.88 - 1.31)	0.471
Cluster 4	1.00 (0.81 - 1.22)	0.962	0.91 (0.73 - 1.13)	0.385
Cluster 5	0.78 (0.56 - 1.08)	0.136	0.78 (0.55 - 1.11)	0.164
Age at Baseline Exam (per 1 year higher)	1.10 (0.97 - 1.25)	0.141	1.11 (0.97 - 1.27)	0.119
Sex				
Female	0.42 (0.28 - 0.62)	<0.001	0.50 (0.33 - 0.76)	0.001
Male	1.0 (ref)	--	1.0 (ref)	--
Baseline CVH Score	--	--	0.72 (0.65 - 0.79)	<0.001
Maximum Years of Education (per year)	--	--	0.94 (0.87 - 1.02)	0.132

Abbreviations: HR = Hazard Ratio; CI = Confidence Interval;

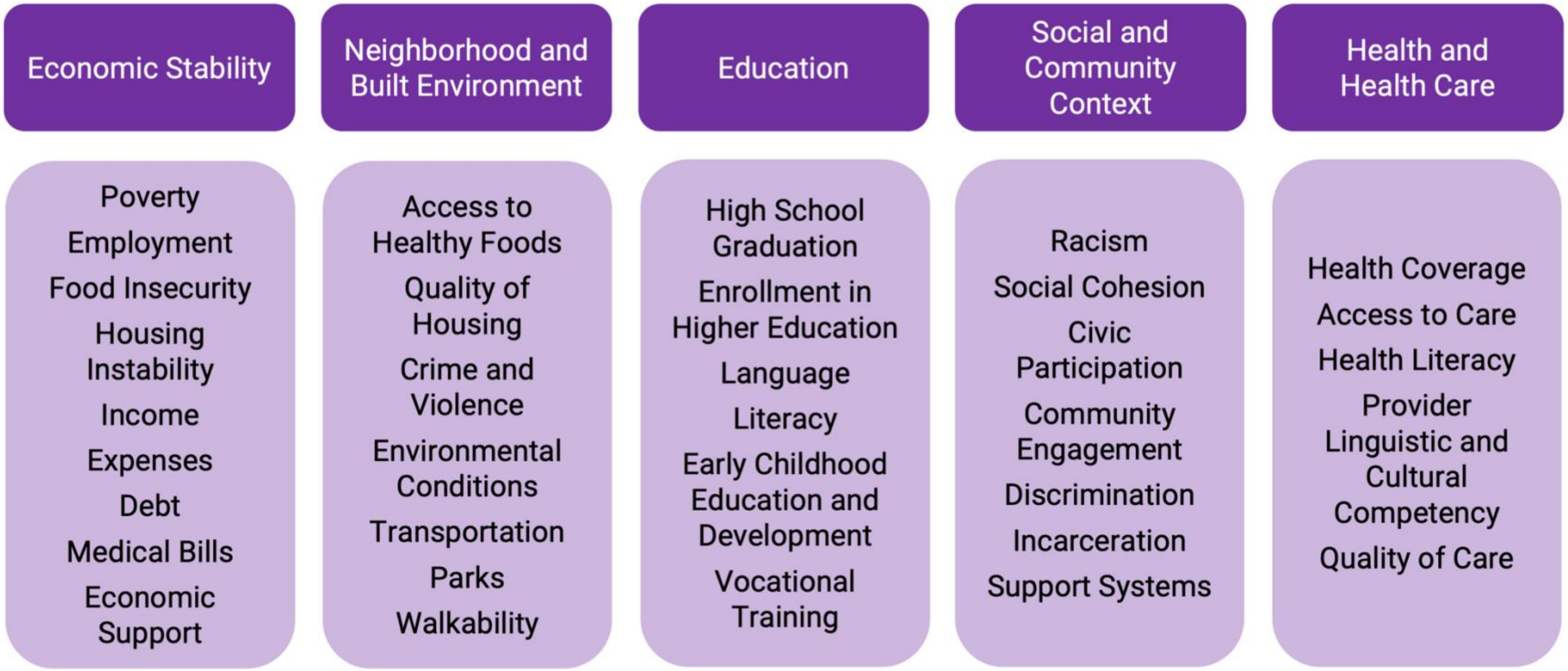
Table 13: Secondary analyses for SDOH-CVD event associations adjusting for mid-life CVH and subclinical CVD

	Model 6			Model 7	
	Adjusted HR (95% CI)	P-Value		Adjusted HR (95% CI)	P-Value
SDOH Clusters			**SDOH Clusters**		
Cluster 1	0.97 (0.79 - 1.20)	0.785	Cluster 1	0.95 (0.74 - 1.23)	0.725
Cluster 2	0.87 (0.62 - 1.21)	0.412	Cluster 2	1.07 (0.74 - 1.56)	0.708
Cluster 3	1.08 (0.92 - 1.26)	0.373	Cluster 3	1.14 (0.93 - 1.39)	0.195
Cluster 4	0.97 (0.83 - 1.15)	0.737	Cluster 4	1.10 (0.90 - 1.34)	0.349
Cluster 5	0.82 (0.62 - 1.08)	0.154	Cluster 5	0.83 (0.58 - 1.19)	0.320
Baseline CVH Score	0.75 (0.69 - 0.82)	<0.001	**Baseline CVH Score**	0.83 (0.74 - 0.92)	<0.001
Age at Baseline Exam (per 1 year higher)	0.98 (0.88 - 1.10)	0.776	**Age at Baseline Exam (per 1 year higher)**	0.92 (0.80 - 1.05)	0.198
Sex			**Sex**		
Female	0.62 (0.44 - 0.87)	0.006	Female	0.44 (0.29 - 0.66)	<0.001
Male	1.0 (ref)	--	Male	1.0 (ref)	--
Race			**Race**		
Black	1.26 (0.87 - 1.82)	0.216	Black	0.95 (0.61 - 1.49)	0.835
White	1.0 (ref)	--	White	1.0 (ref)	--
Maximum Years of Education (per year)	0.98 (0.91 - 1.05)	0.492	**Maximum Years of Education (per year)**	1.00 (0.92 - 1.09)	0.977
CAC at Mid-life			**LVMI at Mid-life ($g/m^{2.7}$)**	1.03 (1.02 - 1.04)	<0.001
Presence	2.28 (1.62 - 3.19)	<0.001			
Absence	1.0 (ref)	--			

	Model 8	
	Adjusted HR (95% CI)	P-Value
SDOH Clusters		
Cluster 1	1.06 (0.85 - 1.31)	0.614

Cluster 2	0.81 (0.57 - 1.13)	0.216
Cluster 3	1.09 (0.92 - 1.29)	0.306
Cluster 4	0.98 (0.83 - 1.16)	0.822
Cluster 5	0.85 (0.64 - 1.13)	0.270
CVH Score at Mid-Life	0.82 (0.75 - 0.89)	<0.001
Baseline CVH Score	0.80 (0.73 - 0.89)	<0.001
Age at Baseline Exam (per 1 year higher)	0.99 (0.89 - 1.11)	0.925
Sex		
Female	0.59 (0.42 - 0.82)	0.002
Male	1.0 (ref)	--
Race		
Black	1.13 (0.78 - 1.62)	0.521
White	1.0 (ref)	--
Maximum Years of Education (per year)	1.00 (0.94 - 1.07)	0.990

Abbreviations: HR = Hazard Ratio; CI = Confidence Interval; CAC = Coronary Artery Calcification; LVMI = Left Ventricular Mass Index;
CVH = Cardiovascular Health;



Economic Stability	Neighborhood and Built Environment	Education	Social and Community Context	Health and Health Care
Poverty Employment Food Insecurity Housing Instability Income Expenses Debt Medical Bills Economic Support	Access to Healthy Foods Quality of Housing Crime and Violence Environmental Conditions Transportation Parks Walkability	High School Graduation Enrollment in Higher Education Language Literacy Early Childhood Education and Development Vocational Training	Racism Social Cohesion Civic Participation Community Engagement Discrimination Incarceration Support Systems	Health Coverage Access to Care Health Literacy Provider Linguistic and Cultural Competency Quality of Care

Figure 1: Social determinants of health – key domains and issue

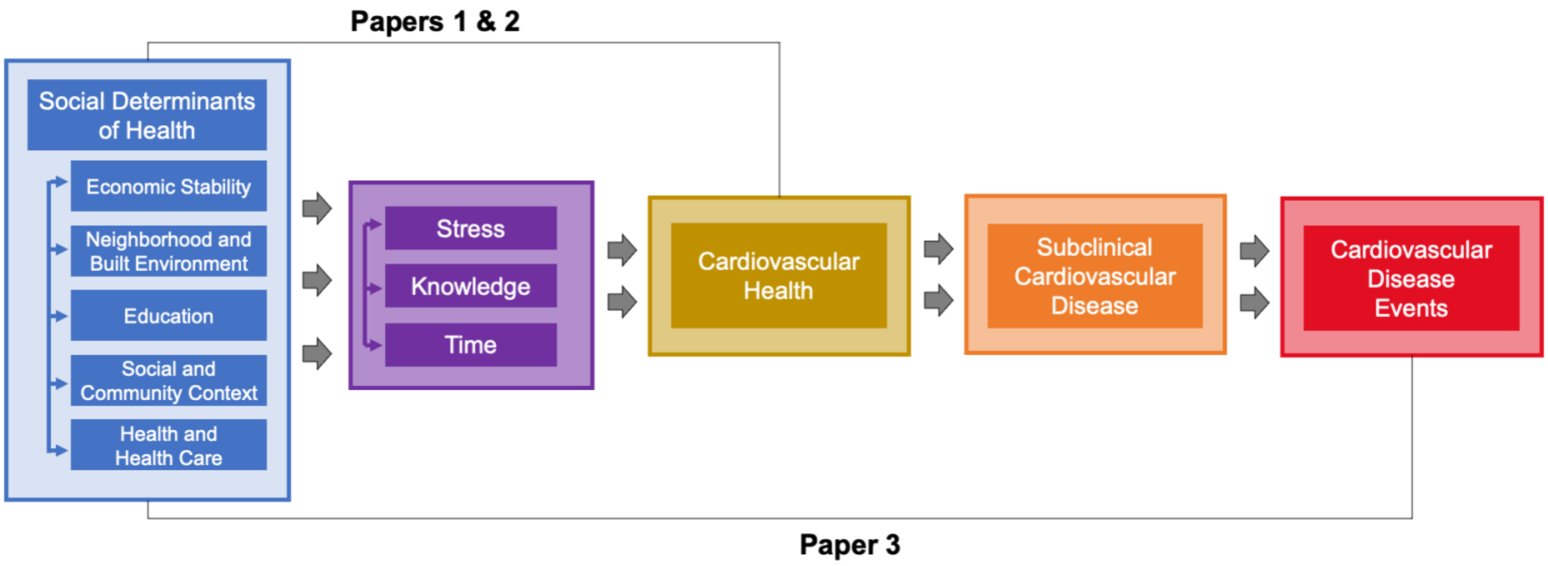

Papers 1 & 2
Social Determinants of Health
Economic Stability
Neighborhood and Built Environment
Education
Social and Community Context
Health and Health Care
Stress
Knowledge
Time
Cardiovascular Health
Subclinical Cardiovascular Disease
Cardiovascular Disease Events
Paper 3

Figure 2: SDOH and CVH conceptual framework
(Modified from WHO)[42]

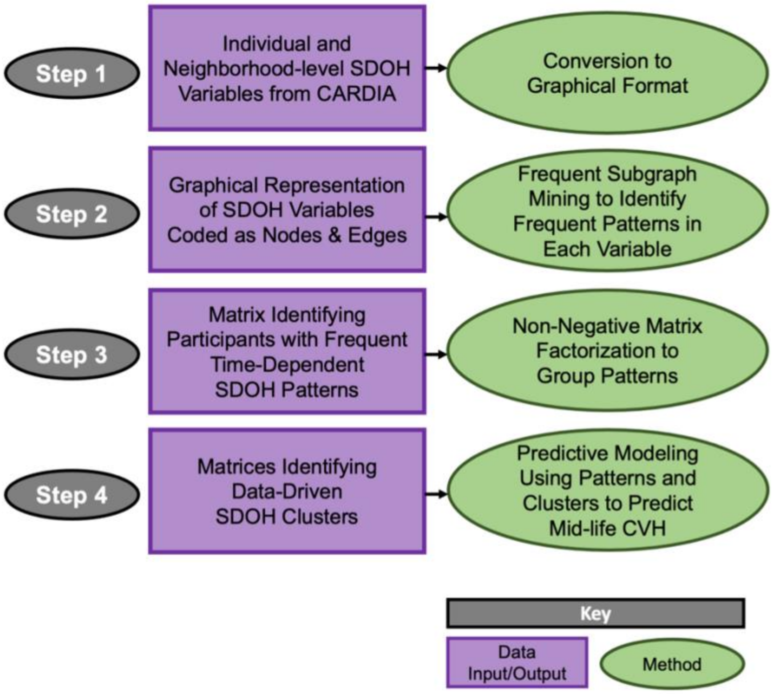

Step 1
Individual and Neighborhood-level SDOH Variables from CARDIA
Conversion to Graphical Format
Step 2
Graphical Representation of SDOH Variables Coded as Nodes & Edges
Frequent Subgraph Mining to Identify Frequent Patterns in Each Variable
Step 3
Matrix Identifying Participants with Frequent Time-Dependent SDOH Patterns
Non-Negative Matrix Factorization to Group Patterns
Step 4
Matrices Identifying Data-Driven SDOH Clusters
Predictive Modeling Using Patterns and Clusters to Predict Mid-life CVH
Key
Data Input/Output
Method

Figure 3: Subgraph augmented non-negative matrix factorization (SANMF) workflow

Abbreviations: CVH = Cardiovascular Health; SDOH = Social Determinants of Health;

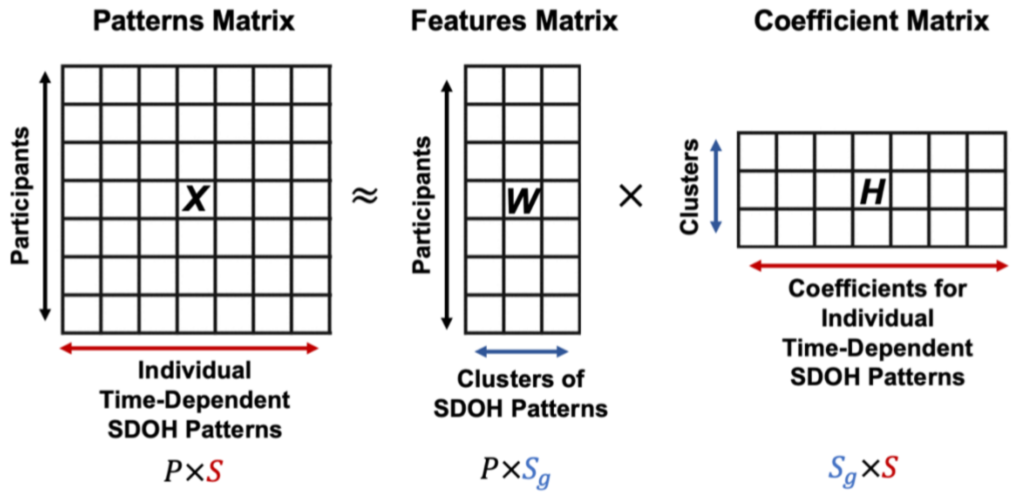


Figure 4: Subgraph augmented non-negative matrix factorization model
Abbreviations: SDOH = Social Determinants of Health;

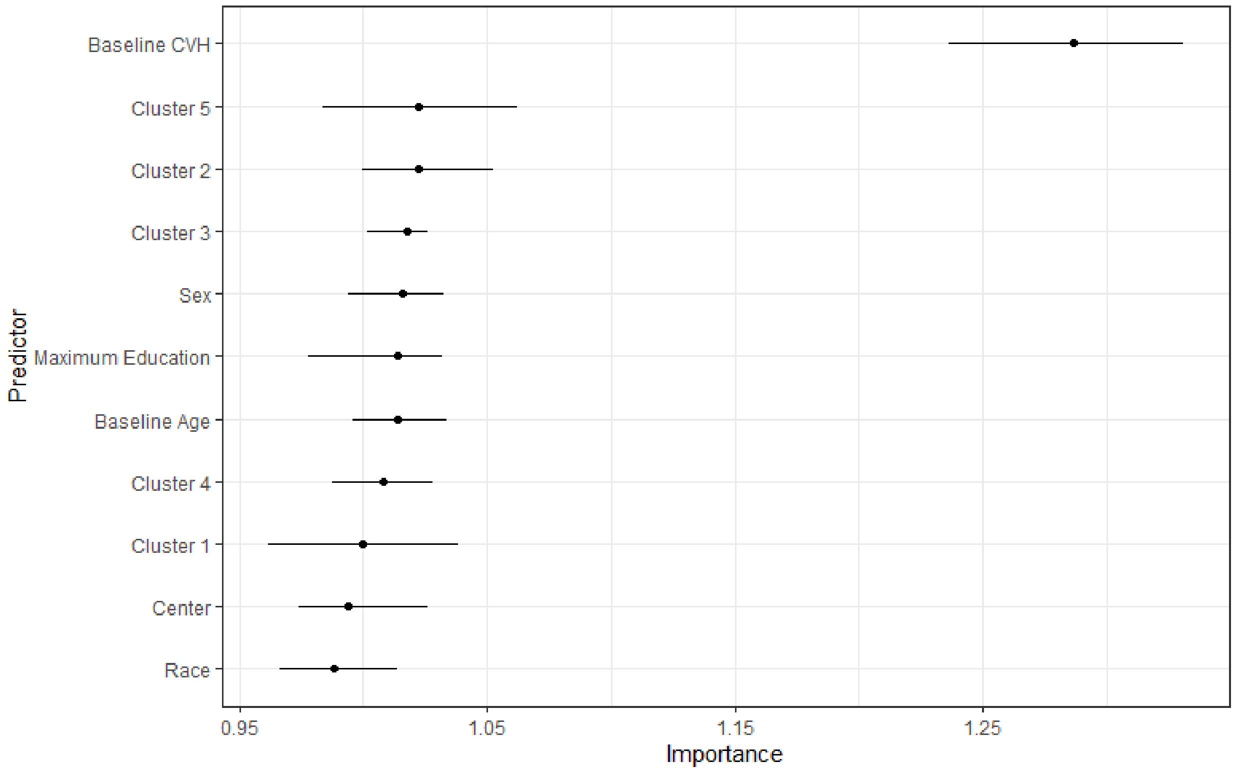


Figure 5: Variable importance plot from the Base + CVH + SDOH clusters logistic regression model

Importance is measured as the classification error in the model when the feature is permuted vs. the original, non-permuted, model. Abbreviations: CVH = Cardiovascular Health; SDOH = Social Determinants of Health;

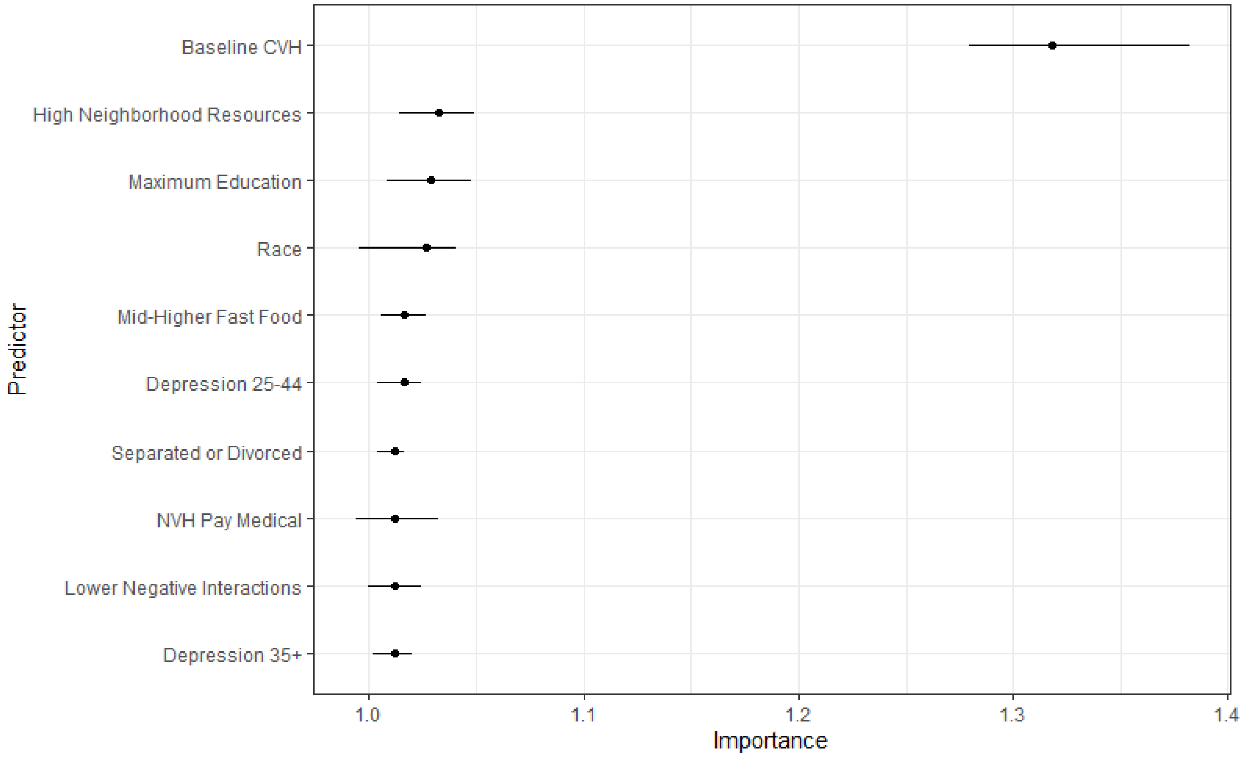


Figure 6: Variable importance plot from the Base + CVH + Subgraphs Lasso regression model

Importance is measured as the classification error in the model when the feature is permuted vs. the original, non-permuted, model.

Description of subgraphs in order of importance:

1) Baseline CVH- Baseline cardiovascular health;

2) High Neighborhood Resources- 7-8 resources available in a participant's neighborhood including exercise facilities, parks, grocery stores, bus stops, etc. in the 45 and beyond age window;

3) Maximum Education level across all age windows;

4) Race;

5) Mid-Higher Fast Food- Mid-range availability to fast food and convenience stores in a participant's neighborhood in the 25-34 age window and then Higher availability in the 34-44 age window;

6) Depression 25-44- Depression in the 25-34 and 35-44 age window;

7) Separated or Divorced- Separated or divorced marital status in the 25-34 age window;

8) NVH Pay Medical- Not very hard to pay for medical care in the 35-44 and 45 and beyond age windows;

9) Lower Negative Interactions- Lower negative interactions in the 35-44 age window.

10) Depression 35+- Depression in the 34-44 and 45 and beyond age windows;

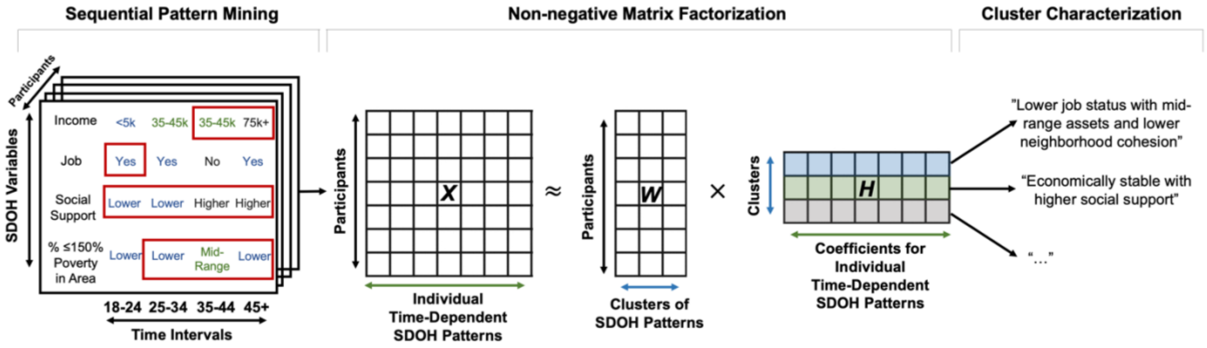
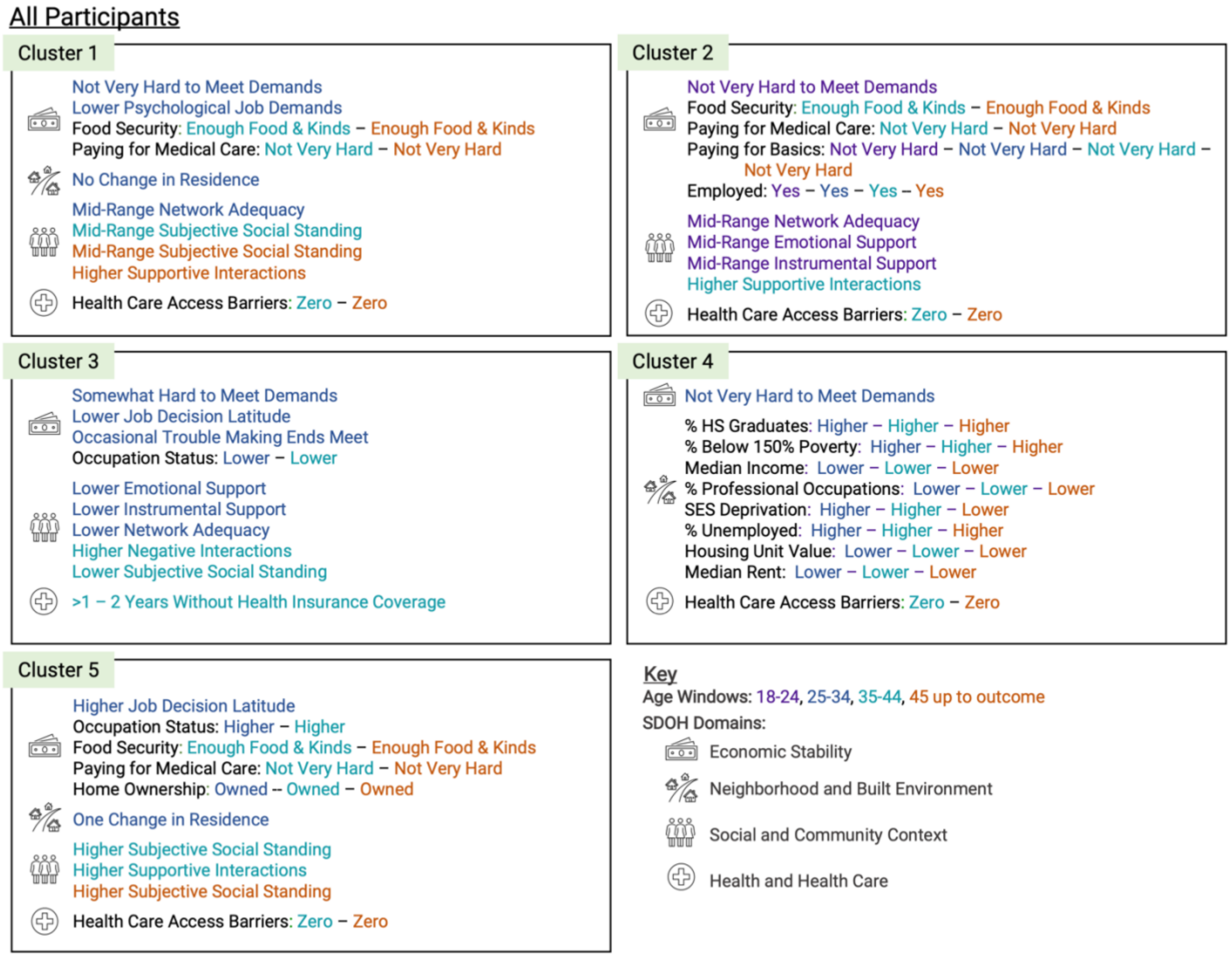

Sequential Pattern Mining
Non-negative Matrix Factorization
Cluster Characterization
Participants
SDOH Variables
Income
<5k
35-45k
35-45k
75k+
Job
Yes
Yes
No
Yes
Social Support
Lower
Lower
Higher
Higher
% ≤150% Poverty in Area
Lower
Lower
Mid-Range
Lower
18-24
25-34
35-44
45+
Time Intervals
Participants
X
Individual Time-Dependent SDOH Patterns
≈
Participants
W
Clusters of SDOH Patterns
×
Clusters
H
Coefficients for Individual Time-Dependent SDOH Patterns
"Lower job status with mid-range assets and lower neighborhood cohesion"
"Economically stable with higher social support"
"…"

Figure 7: Subgraph-augmented non-negative matrix factorization (SANMF) methodology to create and characterize social determinants of health (SDOH) clusters

(Adapted from Sanchez-Pinto et al.[50])



All Participants

Cluster 1

Not Very Hard to Meet Demands
Lower Psychological Job Demands
Food Security: Enough Food & Kinds – Enough Food & Kinds
Paying for Medical Care: Not Very Hard – Not Very Hard

No Change in Residence

Mid-Range Network Adequacy
Mid-Range Subjective Social Standing
Mid-Range Subjective Social Standing
Higher Supportive Interactions

Health Care Access Barriers: Zero – Zero

Cluster 2

Not Very Hard to Meet Demands
Food Security: Enough Food & Kinds – Enough Food & Kinds
Paying for Medical Care: Not Very Hard – Not Very Hard
Paying for Basics: Not Very Hard – Not Very Hard – Not Very Hard – Not Very Hard
Employed: Yes – Yes – Yes – Yes

Mid-Range Network Adequacy
Mid-Range Emotional Support
Mid-Range Instrumental Support
Higher Supportive Interactions

Health Care Access Barriers: Zero – Zero

Cluster 3

Somewhat Hard to Meet Demands
Lower Job Decision Latitude
Occasional Trouble Making Ends Meet
Occupation Status: Lower – Lower

Lower Emotional Support
Lower Instrumental Support
Lower Network Adequacy
Higher Negative Interactions
Lower Subjective Social Standing

>1 – 2 Years Without Health Insurance Coverage

Cluster 4

Not Very Hard to Meet Demands

% HS Graduates: Higher – Higher – Higher
% Below 150% Poverty: Higher – Higher – Higher
Median Income: Lower – Lower – Lower
% Professional Occupations: Lower – Lower – Lower
SES Deprivation: Higher – Higher – Lower
% Unemployed: Higher – Higher – Higher
Housing Unit Value: Lower – Lower – Lower
Median Rent: Lower – Lower – Lower

Health Care Access Barriers: Zero – Zero

Cluster 5

Higher Job Decision Latitude
Occupation Status: Higher – Higher
Food Security: Enough Food & Kinds – Enough Food & Kinds
Paying for Medical Care: Not Very Hard – Not Very Hard
Home Ownership: Owned – Owned – Owned

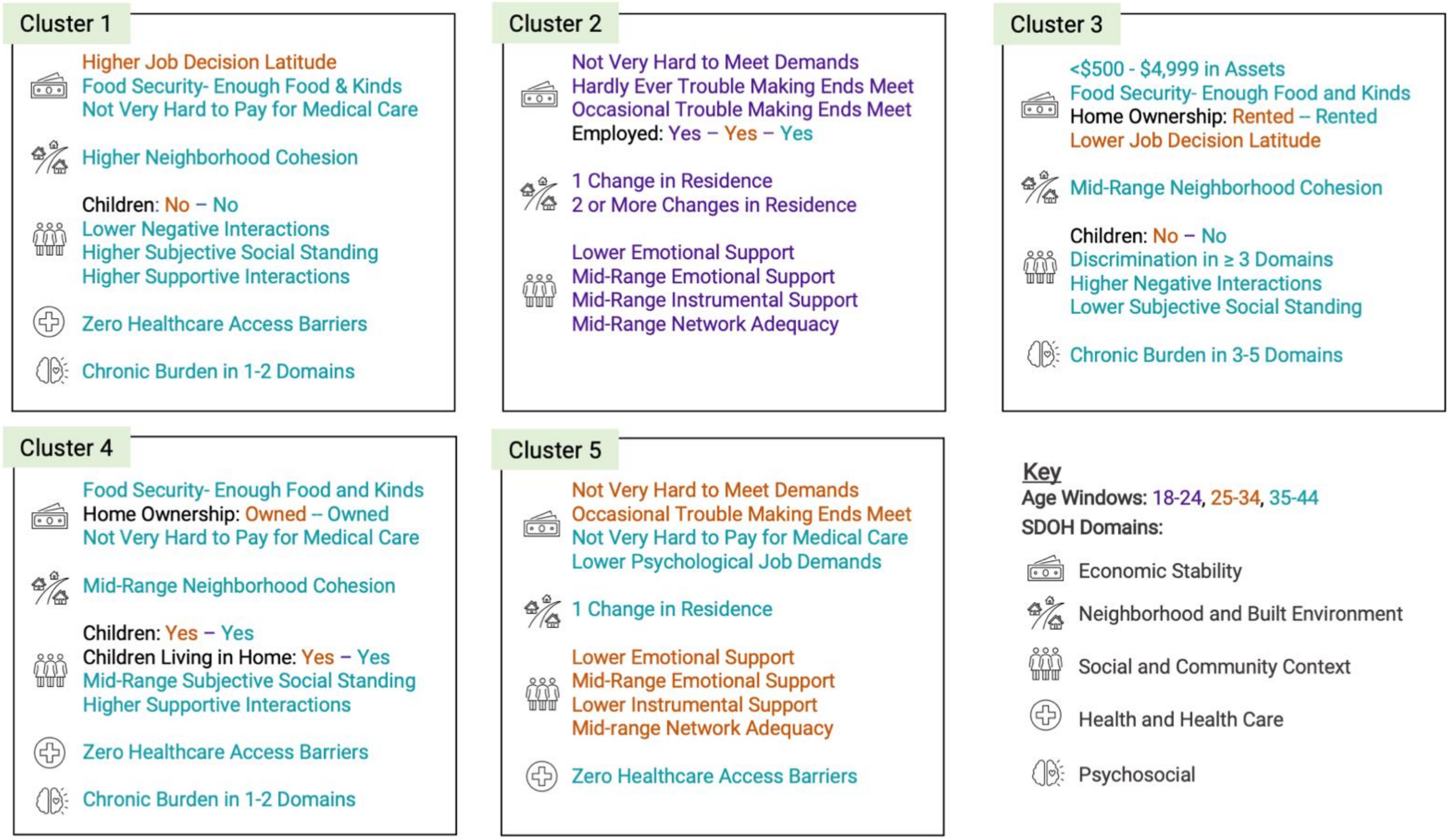
One Change in Residence

Higher Subjective Social Standing
Higher Supportive Interactions
Higher Subjective Social Standing

Health Care Access Barriers: Zero – Zero

Key
Age Windows: 18-24, 25-34, 35-44, 45 up to outcome
SDOH Domains:
Economic Stability
Neighborhood and Built Environment
Social and Community Context
Health and Health Care

Figure 8: Characterization of the five social determinants of health (SDOH) clusters created among all participants.
We present each cluster's top 10 time-dependent SDOH patterns. Abbreviations: HS = High School; SES = Socioeconomic Status; The five clusters can be described as follows: Cluster 1) economically stable with less psychologically demanding job, mid-range social support, zero health care access barriers, and no change in residence; Cluster 2) economically stable with employment, mid-range social support, and zero health care access barriers; Cluster 3) some difficulty economically with lower status job, low social support, and years without health insurance; Cluster 4) no difficulty meeting demands, with vulnerable neighborhood environment, and Cluster 5) economically wealthy with high status job, higher social support, and change in residence during late 20s and early 30s.



Cluster 1

Higher Job Decision Latitude
Food Security- Enough Food & Kinds
Not Very Hard to Pay for Medical Care

Higher Neighborhood Cohesion

Children: No – No
Lower Negative Interactions
Higher Subjective Social Standing
Higher Supportive Interactions

Zero Healthcare Access Barriers

Chronic Burden in 1-2 Domains

Cluster 2

Not Very Hard to Meet Demands
Hardly Ever Trouble Making Ends Meet
Occasional Trouble Making Ends Meet
Employed: Yes – Yes – Yes

1 Change in Residence
2 or More Changes in Residence

Lower Emotional Support
Mid-Range Emotional Support
Mid-Range Instrumental Support
Mid-Range Network Adequacy

Cluster 3

<$500 - $4,999 in Assets
Food Security- Enough Food and Kinds
Home Ownership: Rented – Rented
Lower Job Decision Latitude

Mid-Range Neighborhood Cohesion

Children: No – No
Discrimination in ≥ 3 Domains
Higher Negative Interactions
Lower Subjective Social Standing

Chronic Burden in 3-5 Domains

Cluster 4

Food Security- Enough Food and Kinds
Home Ownership: Owned – Owned
Not Very Hard to Pay for Medical Care

Mid-Range Neighborhood Cohesion

Children: Yes – Yes
Children Living in Home: Yes – Yes
Mid-Range Subjective Social Standing
Higher Supportive Interactions

Zero Healthcare Access Barriers

Chronic Burden in 1-2 Domains

Cluster 5

Not Very Hard to Meet Demands
Occasional Trouble Making Ends Meet
Not Very Hard to Pay for Medical Care
Lower Psychological Job Demands

1 Change in Residence

Lower Emotional Support
Mid-Range Emotional Support
Lower Instrumental Support
Mid-range Network Adequacy

Zero Healthcare Access Barriers

Key
Age Windows: 18-24, 25-34, 35-44
SDOH Domains:
Economic Stability
Neighborhood and Built Environment
Social and Community Context
Health and Health Care
Psychosocial

Figure 9: Characterization of SDOH clusters presenting patterns with the top ten highest membership coefficients

The five clusters can be described as: Cluster 1) higher job decision latitude, financially stable, higher neighborhood cohesion, no children, higher social support, no healthcare access barriers, and some chronic burden; Cluster 2) stable employment throughout early to mid-life, some difficulty making ends meet, one or more changes in residence, and lower to mid-range social support; Cluster 3) lower assets but food secure, rented home, lower job decision latitude job, mid-range neighborhood cohesion, no children, higher discrimination, lower social support, and higher chronic burden; Cluster 4) financially stable and owned home, mid-range neighborhood cohesion, children who are living in the home, mid-range to higher social support, zero healthcare access barriers, and some chronic burden; Cluster 5) none or some economic difficulties, lower job decision latitude, one change in residence, lower to mid-range social support, and zero healthcare access barriers;

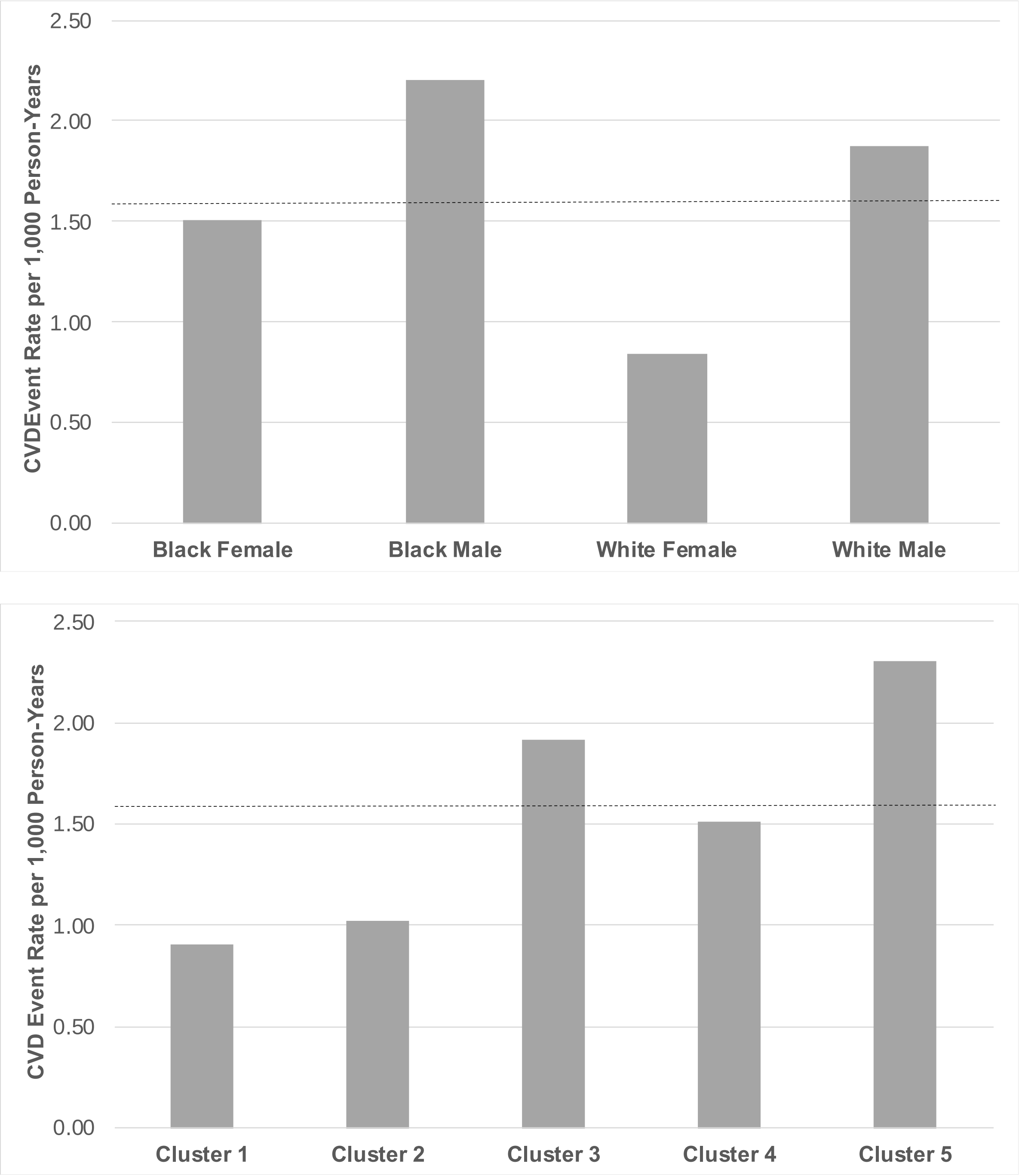


Figure 10: CVD event rates per 1,000 person-years by race and sex subgroups and SDOH cluster

Cluster event rates were based on participants with likelihood values for each SDOH cluster ≥1 standard deviation. Dashed line indicates overall unadjusted event rate (1.56 CVD events per 1,000 person-years).

REFERENCES

1. Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*. 2019;139(10):e56-e528. doi:10.1161/CIR.0000000000000659

2. Heidenreich PA, Trogdon JG, Khavjou OA, et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. *Circulation*. 2011;123(8):933-944. doi:10.1161/CIR.0b013e31820a55f5

3. Graham G. Disparities in cardiovascular disease risk in the United States. *Curr Cardiol Rev*. 2015;11(3):238-245. doi:10.2174/1573403x11666141122220003

4. Okwuosa IS, Lewsey SC, Adesiyun T, Blumenthal RS, Yancy CW. Worldwide disparities in cardiovascular disease: Challenges and solutions. *Int J Cardiol*. 2016;202:433-440. doi:10.1016/j.ijcard.2015.08.172

5. Daviglus ML, Talavera GA, Avilés-Santa ML, et al. Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. *JAMA*. 2012;308(17):1775-1784. doi:10.1001/jama.2012.14517

6. Kramer H, Han C, Post W, et al. Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis

(MESA). *Am J Hypertens*. 2004;17(10):963-970. doi:10.1016/j.amjhyper.2004.06.001

7. Mensah GA, Dunbar SB. A framework for addressing disparities in cardiovascular health. *J Cardiovasc Nurs*. 2006;21(6):451-456. doi:10.1097/00005082-200611000-00007

8. Whitaker KM, Jacobs DR, Kershaw KN, et al. Racial Disparities in Cardiovascular Health Behaviors: The Coronary Artery Risk Development in Young Adults Study. *Am J Prev Med*. 2018;55(1):63-71. doi:10.1016/j.amepre.2018.03.017

9. Lloyd-Jones DM, Hong Y, Labarthe D, et al. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond. *Circulation*. 2010;121(4):586-613. doi:10.1161/CIRCULATIONAHA.109.192703

10. Allen NB, Zhao L, Liu L, et al. Favorable Cardiovascular Health, Compression of Morbidity, and Healthcare Costs: Forty-Year Follow-Up of the CHA Study (Chicago Heart Association Detection Project in Industry). *Circulation*. 2017;135(18):1693-1701. doi:10.1161/CIRCULATIONAHA.116.026252

11. Dhamoon MS, Dong C, Elkind MSV, Sacco RL. Ideal cardiovascular health predicts functional status independently of vascular events: the Northern Manhattan Study. *J Am Heart Assoc*. 2015;4(2). doi:10.1161/JAHA.114.001322

12. Dong C, Rundek T, Wright CB, Anwar Z, Elkind MSV, Sacco RL. Ideal cardiovascular health predicts lower risks of myocardial infarction, stroke, and vascular death across whites, blacks, and hispanics: the northern Manhattan study. *Circulation*. 2012;125(24):2975-2984. doi:10.1161/CIRCULATIONAHA.111.081083

13. España-Romero V, Artero EG, Lee D-C, et al. A prospective study of ideal cardiovascular health and depressive symptoms. *Psychosomatics*. 2013;54(6):525-535. doi:10.1016/j.psym.2013.06.016

14. Ford ES, Greenlund KJ, Hong Y. Ideal cardiovascular health and mortality from all causes and diseases of the circulatory system among adults in the United States. *Circulation*. 2012;125(8):987-995. doi:10.1161/CIRCULATIONAHA.111.049122

15. Folsom AR, Yatsuya H, Nettleton JA, et al. Community prevalence of ideal cardiovascular health, by the American Heart Association definition, and relationship with cardiovascular disease incidence. *J Am Coll Cardiol*. 2011;57(16):1690-1696. doi:10.1016/j.jacc.2010.11.041

16. Rasmussen-Torvik LJ, Shay CM, Abramson JG, et al. Ideal cardiovascular health is inversely associated with incident cancer: the Atherosclerosis Risk In Communities study. *Circulation*. 2013;127(12):1270-1275. doi:10.1161/CIRCULATIONAHA.112.001183

17. Reis JP, Loria CM, Launer LJ, et al. Cardiovascular health through young adulthood and cognitive functioning in midlife. *Ann Neurol*. 2013;73(2):170-179. doi:10.1002/ana.23836

18. Yang Q, Cogswell ME, Flanders WD, et al. Trends in cardiovascular health metrics and associations with all-cause and CVD mortality among US adults. *JAMA*. 2012;307(12):1273-1283. doi:10.1001/jama.2012.339

19. Bambs C, Kip KE, Dinga A, Mulukutla SR, Aiyer AN, Reis SE. Low prevalence of “ideal cardiovascular health” in a community-based population: the heart strategies concentrating on risk evaluation (Heart SCORE) study. *Circulation*. 2011;123(8):850-857. doi:10.1161/CIRCULATIONAHA.110.980151

20. Lloyd-Jones DM, Dyer AR, Wang R, Daviglus ML, Greenland P. Risk factor burden in middle age and lifetime risks for cardiovascular and non-cardiovascular death (Chicago Heart Association Detection Project in Industry). *Am J Cardiol*. 2007;99(4):535-540. doi:10.1016/j.amjcard.2006.09.099

21. Stampfer MJ, Hu FB, Manson JE, Rimm EB, Willett WC. Primary prevention of coronary heart disease in women through diet and lifestyle. *N Engl J Med*. 2000;343(1):16-22. doi:10.1056/NEJM200007063430103

22. Brown AF, Liang L-J, Vassar SD, et al. Trends in Racial/Ethnic and Nativity Disparities in Cardiovascular Health Among Adults Without Prevalent

Cardiovascular Disease in the United States, 1988 to 2014. *Ann Intern Med*. 2018;168(8):541-549. doi:10.7326/M17-0996

23. WHO Commission on Social Determinants of Health, World Health Organization, eds. *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health: Commission on Social Determinants of Health Final Report*. World Health Organization, Commission on Social Determinants of Health; 2008.

24. Healthy People 2020. Social Determinants of Health. Accessed December 19, 2019. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

25. American Community Survey (ACS). Accessed December 19, 2019. https://www.census.gov/programs-surveys/acs

26. Cantor MN, Chandras R, Pulgarin C. FACETS: using open data to measure community social determinants of health. *J Am Med Inform Assoc*. 2018;25(4):419-422. doi:10.1093/jamia/ocx117

27. Havranek EP, Mujahid MS, Barr DA, et al. Social Determinants of Risk and Outcomes for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation*. 2015;132(9):873-898. doi:10.1161/CIR.0000000000000228

28. Strategic Vision | National Heart, Lung, and Blood Institute (NHLBI). Accessed December 19, 2019. https://www.nhlbi.nih.gov/about/strategic-vision

29. Harper S, Lynch J, Smith GD. Social determinants and the decline of cardiovascular diseases: understanding the links. *Annu Rev Public Health*. 2011;32:39-69. doi:10.1146/annurev-publhealth-031210-101234

30. Unger E, Diez-Roux AV, Lloyd-Jones DM, et al. Association of neighborhood characteristics with cardiovascular health in the multi-ethnic study of atherosclerosis. *Circ Cardiovasc Qual Outcomes*. 2014;7(4):524-531. doi:10.1161/CIRCOUTCOMES.113.000698

31. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *Am J Public Health*. 1992;82(6):816-820. doi:10.2105/ajph.82.6.816

32. Caleyachetty R, Echouffo-Tcheugui JB, Muennig P, Zhu W, Muntner P, Shimbo D. Association between cumulative social risk and ideal cardiovascular health in US adults: NHANES 1999–2006. *International Journal of Cardiology*. 2015;191:296-300. doi:10.1016/j.ijcard.2015.05.007

33. Pollitt RA, Rose KM, Kaufman JS. Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. *BMC Public Health*. 2005;5:7. doi:10.1186/1471-2458-5-7

34. Mujahid MS, Moore LV, Petito LC, Kershaw KN, Watson K, Diez Roux AV. Neighborhoods and racial/ethnic differences in ideal cardiovascular health (the Multi-Ethnic Study of Atherosclerosis). *Health Place*. 2017;44:61-69. doi:10.1016/j.healthplace.2017.01.005

35. Khan SS, Lloyd-Jones DM, Carnethon M, Pool LR. Medicaid Expansion and State-Level Differences in Premature Cardiovascular Mortality by Subtype, 2010-2017. *Hypertension*. 2020;76(5):e37-e38. doi:10.1161/HYPERTENSIONAHA.120.15968

36. Rehkopf DH, Strully KW, Dow WH. The short-term impacts of Earned Income Tax Credit disbursement on health. *Int J Epidemiol*. 2014;43(6):1884-1894. doi:10.1093/ije/dyu172

37. Krishnan V. Constructing an Area-based Socioeconomic Index: A Principal Components Analysis Approach. Published online 2010.

38. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plan*. 2006;21(6):459-468. doi:10.1093/heapol/czl029

39. Messer LC, Laraia BA, Kaufman JS, et al. The development of a standardized neighborhood deprivation index. *J Urban Health*. 2006;83(6):1041-1062. doi:10.1007/s11524-006-9094-x

40. Weyers S, Dragano N, Möbus S, et al. Low socio-economic position is associated with poor social networks and social support: results from the Heinz Nixdorf Recall Study. *Int J Equity Health*. 2008;7(1):13. doi:10.1186/1475-9276-7-13

41. Richardson AS, Meyer KA, Howard AG, et al. Neighborhood socioeconomic status and food environment: a 20-year longitudinal latent class analysis among CARDIA participants. *Health Place*. 2014;30:145-153. doi:10.1016/j.healthplace.2014.08.011

42. World Health Organization. *A Conceptual Framework for Action on the Social Determinants of Health: Debates, Policy & Practice, Case Studies.*; 2010. Accessed December 19, 2019. http://apps.who.int/iris/bitstream/10665/44489/1/9789241500852_eng.pdf

43. Williams DR, Sternthal M. Understanding racial-ethnic disparities in health: sociological contributions. *J Health Soc Behav*. 2010;51 Suppl:S15-27. doi:10.1177/0022146510383838

44. Gee GC, Hing A, Mohammed S, Tabor DC, Williams DR. Racism and the Life Course: Taking Time Seriously. *Am J Public Health*. 2019;109(S1):S43-S47. doi:10.2105/AJPH.2018.304766

45. Schulz AJ, Kannan S, Dvonch JT, et al. Social and physical environments and disparities in risk for cardiovascular disease: the healthy environments partnership

conceptual model. *Environ Health Perspect*. 2005;113(12):1817-1825. doi:10.1289/ehp.7913

46. Havranek EP, Mujahid MS, Barr DA, et al. Social Determinants of Risk and Outcomes for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation*. 2015;132(9):873-898. doi:10.1161/CIR.0000000000000228

47. Bailey ZD, Feldman JM, Bassett MT. How Structural Racism Works — Racist Policies as a Root Cause of U.S. Racial Health Inequities. Malina D, ed. *N Engl J Med*. Published online December 16, 2020:NEJMms2025396. doi:10.1056/NEJMms2025396

48. Wyatt SB, Williams DR, Calvin R, Henderson FC, Walker ER, Winters K. Racism and cardiovascular disease in African Americans. *Am J Med Sci*. 2003;325(6):315-331. doi:10.1097/00000441-200306000-00003

49. Luo Y, Xin Y, Joshi R, Celi LA, Szolovits P. Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements. In: *AAAI*. ; 2016. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewFile/11843/11562

50. Sanchez-Pinto LN, Stroup EK, Pendergrast T, Pinto N, Luo Y. Derivation and Validation of Novel Phenotypes of Multiple Organ Dysfunction Syndrome in

Critically Ill Children. *JAMA Netw Open*. 2020;3(8):e209271. doi:10.1001/jamanetworkopen.2020.9271

51. McGovern L. The Relative Contribution of Multiple Determinants to Health. *Health Affairs*. Published online August 21, 2014. doi:10.1377/hpb2014.17

52. Galea S, Tracy M, Hoggatt KJ, Dimaggio C, Karpati A. Estimated deaths attributable to social factors in the United States. *Am J Public Health*. 2011;101(8):1456-1465. doi:10.2105/AJPH.2010.300086

53. McGinnis JM, Foege WH. Actual causes of death in the United States. *JAMA*. 1993;270(18):2207-2212.

54. *The Patient Protection and Affordable Care Act of 2010.* Public Law 111-148, 111th Congress, 124 Stat 119, HR 3590; 2010.

55. The BARHII Framework. Bay Area Regional Health Inequities Initiative (BARHII). Accessed December 14, 2020. https://www.barhii.org/barhii-framework

56. Fuchs VR. Social Determinants of Health: Caveats and Nuances. *JAMA*. 2017;317(1):25-26. doi:10.1001/jama.2016.17335

57. 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-e596. doi:10.1161/CIR.0000000000000757

58. Friedman GD, Cutter GR, Donahue RP, et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J Clin Epidemiol*. 1988;41(11):1105-1116. doi:10.1016/0895-4356(88)90080-7

59. Bancks MP, Allen NB, Dubey P, et al. Cardiovascular health in young adulthood and structural brain MRI in midlife: The CARDIA study. *Neurology*. 2017;89(7):680-686. doi:10.1212/WNL.0000000000004222

60. Polonsky TS, Ning H, Daviglus ML, et al. Association of Cardiovascular Health With Subclinical Disease and Incident Events: The Multi-Ethnic Study of Atherosclerosis. *J Am Heart Assoc*. 2017;6(3). doi:10.1161/JAHA.116.004894

61. Huang C-C, Fornage M, Lloyd-Jones DM, Wei GS, Boerwinkle E, Liu K. Longitudinal association of PCSK9 sequence variations with low-density lipoprotein cholesterol levels: the Coronary Artery Risk Development in Young Adults Study. *Circ Cardiovasc Genet*. 2009;2(4):354-361. doi:10.1161/CIRCGENETICS.108.828467

62. Allen NB, Siddique J, Wilkins JT, et al. Blood pressure trajectories in early adulthood and subclinical atherosclerosis in middle age. *JAMA*. 2014;311(5):490-497. doi:10.1001/jama.2013.285122

63. Borgelt C, Berthold MR. Mining molecular fragments: finding relevant substructures of molecules. In: *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* ; 2002:51-58. doi:10.1109/ICDM.2002.1183885

64. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401(6755):788-791. doi:10.1038/44565

65. Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol*. 2008;4(7):e1000029. doi:10.1371/journal.pcbi.1000029

66. Gillis N. The Why and How of Nonnegative Matrix Factorization. *arXiv:14015226 [cs, math, stat]*. Published online March 7, 2014. Accessed December 19, 2019. http://arxiv.org/abs/1401.5226

67. Lin C-J. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*. 2007;19(10):2756-2779. doi:10.1162/neco.2007.19.10.2756

68. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12(85):2825-2830.

69. Chao G, Mao C, Wang F, Zhao Y, Luo Y. Supervised Nonnegative Matrix Factorization to Predict ICU Mortality Risk. *CoRR*. 2018;abs/1809.10680. Accessed January 11, 2021. https://arxiv.org/abs/1809.10680

70. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*. 2010;11:367. doi:10.1186/1471-2105-11-367

71. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*. 2007;23(12):1495-1502. doi:10.1093/bioinformatics/btm134

72. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

73. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2017. https://www.R-project.org/

74. Kuhn M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*. 2008;28(1):1-26. doi:10.18637/jss.v028.i05

75. Molnar C. Permutation Feature Importance. In: *Interpretable Machine Learning*. ; 2019. https://christophm.github.io/interpretable-ml-book/

76. Molnar C, Casalicchio G, Bischl B. iml: An R package for interpretable machine learning. *Journal of Open Source Software*. 2018;3(26):786.

77. Molnar C. Shapley Values. In: *Interpretable Machine Learning*. ; 2019. https://christophm.github.io/interpretable-ml-book/shapley.html

78. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*. 2010;5(9):1315-1316. doi:10.1097/JTO.0b013e3181ec173d

79. Lloyd-Jones DM. Cardiovascular Risk Prediction: Basic Concepts, Current Status, and Future Directions. *Circulation*. 2010;121(15):1768-1777. doi:10.1161/CIRCULATIONAHA.109.849166

80. Loria CM, Liu K, Lewis CE, et al. Early adult risk factor levels and subsequent coronary artery calcification: the CARDIA Study. *J Am Coll Cardiol*. 2007;49(20):2013-2020. doi:10.1016/j.jacc.2007.03.009

81. Steinberger J, Daniels SR, Hagberg N, et al. Cardiovascular Health Promotion in Children: Challenges and Opportunities for 2020 and Beyond: A Scientific Statement From the American Heart Association. *Circulation*. 2016;134(12). doi:10.1161/CIR.0000000000000441

82. Suglia SF, Campo RA, Brown AGM, et al. Social Determinants of Cardiovascular Health: Early Life Adversity as a Contributor to Disparities in Cardiovascular Diseases. *The Journal of Pediatrics*. 2020;219:267-273. doi:10.1016/j.jpeds.2019.12.063

83. Walter SD, Sinuff T. Studies reporting ROC curves of diagnostic and prediction data can be incorporated into meta-analyses using corresponding odds ratios. *J Clin Epidemiol*. 2007;60(5):530-534. doi:10.1016/j.jclinepi.2006.09.002

84. Wang MC, Lloyd-Jones DM. Cardiovascular Risk Assessment in Hypertensive Patients. *Am J Hypertens*. Published online January 27, 2021. doi:10.1093/ajh/hpab021

85. Lloyd-Jones DM, Braun LT, Ndumele CE, et al. Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology. *J Am Coll Cardiol*. 2019;73(24):3153-3167. doi:10.1016/j.jacc.2018.11.005

86. Allen NB, Krefman AE, Labarthe D, et al. Cardiovascular Health Trajectories From Childhood Through Middle Age and Their Association With Subclinical Atherosclerosis. *JAMA Cardiol*. 2020;5(5):557-566. doi:10.1001/jamacardio.2020.0140

87. Pool LR, Ning H, Lloyd-Jones DM, Allen NB. Trends in Racial/Ethnic Disparities in Cardiovascular Health Among US Adults From 1999-2012. *J Am Heart Assoc*. 2017;6(9). doi:10.1161/JAHA.117.006027

88. Connolly Sean D, Perak Amanda, Pool Lindsay, Ning Hongyan, Marino Bradley, Lloyd-Jones Donald M. Abstract 15163: Social Determinants of Cardiovascular Health in US Adolescents: National Health and Nutrition Examination Surveys (NHANES) 1999-2014. *Circulation*. 2018;138(Suppl_1):A15163-A15163. doi:10.1161/circ.138.suppl_1.15163

89. Wilkins JT, Ning H, Berry J, Zhao L, Dyer AR, Lloyd-Jones DM. Lifetime risk and years lived free of total cardiovascular disease. *JAMA*. 2012;308(17):1795-1801. doi:10.1001/jama.2012.14312

90. Kaplan GA, Keil JE. Socioeconomic factors and cardiovascular disease: a review of the literature. *Circulation*. 1993;88(4 Pt 1):1973-1998. doi:10.1161/01.cir.88.4.1973

91. Kawachi I, Colditz GA, Ascherio A, et al. A prospective study of social networks in relation to total mortality and cardiovascular disease in men in the USA. *J Epidemiol Community Health*. 1996;50(3):245-251. doi:10.1136/jech.50.3.245

92. Kiefe CI, Williams OD, Greenlund KJ, Ulene V, Gardin JM, Raczynski JM. Health care access and seven-year change in cigarette smoking. The CARDIA Study. *Am J Prev Med*. 1998;15(2):146-154. doi:10.1016/s0749-3797(98)00044-0

93. Van Dyke M, Greer S, Odom E, et al. Heart Disease Death Rates Among Blacks and Whites Aged ≥35 Years - United States, 1968-2015. *MMWR Surveill Summ*. 2018;67(5):1-11. doi:10.15585/mmwr.ss6705a1

94. Carnethon MR, Pu J, Howard G, et al. Cardiovascular Health in African Americans: A Scientific Statement From the American Heart Association. *Circulation*. 2017;136(21). doi:10.1161/CIR.0000000000000534

95. Matthews KA, Schwartz JE, Cohen S. Indices of socioeconomic position across the life course as predictors of coronary calcification in black and white men and women: coronary artery risk development in young adults study. *Soc Sci Med*. 2011;73(5):768-774. doi:10.1016/j.socscimed.2011.06.017

96. Kim D, Diez Roux AV, Kiefe CI, Kawachi I, Liu K. Do neighborhood socioeconomic deprivation and low social cohesion predict coronary calcification?: the CARDIA study. *Am J Epidemiol*. 2010;172(3):288-298. doi:10.1093/aje/kwq098

97. Medenwald D, Tiller D, Nuding S, et al. Educational status and differences in left ventricular mass and ejection fraction - The role of BMI and parameters related to the metabolic syndrome: A longitudinal analysis from the population-based CARLA cohort. *Nutr Metab Cardiovasc Dis*. 2016;26(9):815-823. doi:10.1016/j.numecd.2016.05.001

98. Carson AP, Rose KM, Catellier DJ, et al. Cumulative socioeconomic status across the life course and subclinical atherosclerosis. *Ann Epidemiol*. 2007;17(4):296-303. doi:10.1016/j.annepidem.2006.07.009

99. Schultz WM, Kelli HM, Lisko JC, et al. Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. *Circulation*. 2018;137(20):2166-2178. doi:10.1161/CIRCULATIONAHA.117.029652

100. Galobardes B, Smith GD, Lynch JW. Systematic review of the influence of childhood socioeconomic circumstances on risk for cardiovascular disease in adulthood. *Ann Epidemiol*. 2006;16(2):91-104. doi:10.1016/j.annepidem.2005.06.053

101. Cuttica MJ, Colangelo LA, Dransfield MT, et al. Lung Function in Young Adults and Risk of Cardiovascular Events Over 29 Years: The CARDIA Study. *J Am Heart Assoc*. 2018;7(24):e010672. doi:10.1161/JAHA.118.010672

102. Greenland P, LaBree L, Azen SP, Doherty TM, Detrano RC. Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals. *JAMA*. 2004;291(2):210-215. doi:10.1001/jama.291.2.210

103. Demer LL, Tintut Y. Vascular calcification: pathobiology of a multifaceted disease. *Circulation*. 2008;117(22):2938-2948. doi:10.1161/CIRCULATIONAHA.107.743161

104. Liu W, Zhang Y, Yu C-M, et al. Current understanding of coronary artery calcification. *J Geriatr Cardiol*. 2015;12(6):668-675. doi:10.11909/j.issn.1671-5411.2015.06.012

105. Carr JJ, Nelson JC, Wong ND, et al. Calcified coronary artery plaque measurement with cardiac CT in population-based studies: standardized protocol of Multi-Ethnic Study of Atherosclerosis (MESA) and Coronary Artery Risk Development in Young Adults (CARDIA) study. *Radiology*. 2005;234(1):35-43. doi:10.1148/radiol.2341040439

106. Heckbert SR, Post W, Pearson GDN, et al. Traditional cardiovascular risk factors in relation to left ventricular mass, volume, and systolic function by cardiac

magnetic resonance imaging: the Multiethnic Study of Atherosclerosis. *J Am Coll Cardiol*. 2006;48(11):2285-2292. doi:10.1016/j.jacc.2006.03.072

107. Guleri N, Rana S, Chauhan RS, Negi PC, Diwan Y, Diwan D. Study of Left Ventricular Mass and Its Determinants on Echocardiography. *J Clin Diagn Res*. 2017;11(9):OC13-OC16. doi:10.7860/JCDR/2017/28048.10576

108. Casale PN, Devereux RB, Milner M, et al. Value of echocardiographic measurement of left ventricular mass in predicting cardiovascular morbid events in hypertensive men. *Ann Intern Med*. 1986;105(2):173-178. doi:10.7326/0003-4819-105-2-173

109. Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. *N Engl J Med*. 1990;322(22):1561-1566. doi:10.1056/NEJM199005313222203

110. Yared GS, Moreira HT, Ambale-Venkatesh B, et al. Coronary Artery Calcium From Early Adulthood to Middle Age and Left Ventricular Structure and Function. *Circ Cardiovasc Imaging*. 2019;12(6):e009228. doi:10.1161/CIRCIMAGING.119.009228

111. Strasser T. Reflections on Cardiovascular Diseases. *Interdisciplinary Science Reviews*. 1978;3(3):225-230. doi:10.1179/030801878791925921

112. Weintraub WS, Daniels SR, Burke LE, et al. Value of Primordial and Primary Prevention for Cardiovascular Disease: A Policy Statement From the American Heart Association. *Circulation*. 2011;124(8):967-990. doi:10.1161/CIR.0b013e3182285a81

113. Liu K, Colangelo LA, Daviglus ML, et al. Can Antihypertensive Treatment Restore the Risk of Cardiovascular Disease to Ideal Levels?: The Coronary Artery Risk Development in Young Adults (CARDIA) Study and the Multi-Ethnic Study of Atherosclerosis (MESA). *JAHA*. 2015;4(9). doi:10.1161/JAHA.115.002275

114. Mackenbach JP. The contribution of medical care to mortality decline: McKeown revisited. *J Clin Epidemiol*. 1996;49(11):1207-1213. doi:10.1016/s0895-4356(96)00200-4

115. Braveman P, Gottlieb L. The social determinants of health: it’s time to consider the causes of the causes. *Public Health Rep*. 2014;129 Suppl 2:19-31. doi:10.1177/00333549141291S206

116. Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav*. 1995;Spec No:80-94.

117. Hill AB. THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? *Proc R Soc Med*. 1965;58:295-300.

118. Crenshaw K. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. In: Vol 1989. University of Chicago Legal Forum.

119. Liu K, Wilkins JT, Colangelo LA, Lloyd-Jones DM. Does Lowering Low-Density Lipoprotein Cholesterol With Statin Restore Low Risk in Middle-Aged Adults? Analysis of the Observational MESA Study. *J Am Heart Assoc*. Published online May 17, 2021:e019695. doi:10.1161/JAHA.120.019695

120. Cantor MN, Thorpe L. Integrating Data On Social Determinants Of Health Into Electronic Health Records. *Health Affairs*. 2018;37(4):585-590. doi:10.1377/hlthaff.2017.1252

121. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *Journal of the American Medical Informatics Association*. 2020;27(11):1764-1773. doi:10.1093/jamia/ocaa143

122. Bauer GR. Incorporating intersectionality theory into population health research methodology: Challenges and the potential to advance health equity. *Social Science & Medicine*. 2014;110:10-17. doi:10.1016/j.socscimed.2014.03.022

123. Turan JM, Elafros MA, Logie CH, et al. Challenges and opportunities in examining and addressing intersectional stigma and health. *BMC Med*. 2019;17(1):7. doi:10.1186/s12916-018-1246-9

124. Moussa M, Măndoiu II. Single cell RNA-seq data clustering using TF-IDF based methods. *BMC Genomics*. 2018;19(Suppl 6):569. doi:10.1186/s12864-018-4922-4

125. Green K, Zook M. When Talking About Social Determinants, Precision Matters. Health Affairs. Accessed May 27, 2021. https://www.healthaffairs.org/do/10.1377/hblog20191025.776011/full/

126. Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J Epidemiol Community Health*. 2012;66(12):1182-1186. doi:10.1136/jech-2011-200375

127. Crane M, Bohn-Goldbaum E, Grunseit A, Bauman A. Using natural experiments to improve public health evidence: a review of context and utility for obesity prevention. *Health Res Policy Sys*. 2020;18(1):48. doi:10.1186/s12961-020-00564-2

128. FACT SHEET: The American Rescue Plan Will Deliver Immediate Economic Relief to Families. U.S. Department of the Treasury. Published March 18, 2021. Accessed June 2, 2021. https://home.treasury.gov/news/featured-stories/fact-sheet-the-american-rescue-plan-will-deliver-immediate-economic-relief-to-families

129. Parolin Z, Collyer S, Curran M, Wimer C. The Potential Poverty Reduction Effect of the American Rescue Plan. Center on Poverty and Social Policy at Columbia University. Published March 11, 2021. Accessed June 2, 2021.

https://www.povertycenter.columbia.edu/news-internal/2021/presidential-policy/biden-economic-relief-proposal-poverty-impact

130. Capewell S, Morrison CE, McMurray JJ. Contribution of modern cardiovascular treatment and risk factor changes to the decline in coronary heart disease mortality in Scotland between 1975 and 1994. *Heart*. 1999;81(4):380-386. doi:10.1136/hrt.81.4.380

131. Huang Y, Kypridemos C, Liu J, et al. Cost-Effectiveness of the US Food and Drug Administration Added Sugar Labeling Policy for Improving Diet and Health. *Circulation*. 2019;139(23):2613-2624. doi:10.1161/CIRCULATIONAHA.118.036751

132. Unal B, Capewell S, Critchley JA. Coronary heart disease policy models: a systematic review. *BMC Public Health*. 2006;6:213. doi:10.1186/1471-2458-6-213