Examining the Role of Socioeconomic Status on Blood Pressure in African Americans

By

Brittany Marie Hollister

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Human Genetics
August 11, 2017
Nashville, Tennessee

Approved:
Melinda C. Aldrich, Ph.D., M.P.H.
Todd Edwards, Ph.D.
Dana Crawford, Ph.D.
Derek Griffith, Ph.D.
Amy Non, Ph.D., M.P.H.

To my parents, Bill and Karen, for their constant love and support.
This would not have been possible without you.

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the Human Genetics Training Grant (NIH/NIGMS Genetics Predoctoral Research Training Program 5T32GM080178-07) and the Human Genetics Training Program.

I am especially thankful to my mentors, Dr. Melinda Aldrich, Dr. Dana Crawford, and Dr. Amy Non. Their constant guidance throughout my graduate career, in scientific, professional, and personal realms, have made me the scientist I am today. They have each greatly contributed to my current success and my future success will be due to their efforts.

I am also grateful to everyone I have worked with during my time in graduate school including my committee members, fellow Human Genetics graduate students, and collaborators. I am indebted for my committee chair, Dr. Todd Edwards, who provided exceptional guidance and support from the moment we met. I am also beholden to my final committee member, Dr. Derek Griffith, for his help in understanding the social implications of my project and his positive attitude which helped me remain positive.

Finally, the support of my family and Michael Dial have been essential to the completion of this dissertation. Their emotional support and love made this dissertation happen.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION TO SOCIOECONOMIC STATUS AND BLOOD PRESSURE

Socioeconomic status and health

Definition of socioeconomic status

Socioeconomic status is a major determinant of variation in health outcomes worldwide [1]. It can be defined and measured using a variety of methods, but it is typically an aggregate measure which includes an assessment of economic status (normally in the form of income), social status (usually in the form of education), and work status (generally in the form of occupation) in the United States [2]. These measurements are broad and can be assessed at the individual, household, or neighborhood level. Additionally, socioeconomic status can be evaluated at an objective and subjective level. Objective dimensions, such as occupation, education, and income, include those measured without consideration of an individual's perspective.

The collection of these socioeconomic status data can vary. For example, occupation can be measured via occupational prestige, job income brackets, or types of employment (e.g. blue collar versus white collar), and these labels can be further be grouped into categories, leading to a large variety of types of occupational data.

Unlike occupation, the measurement of education tends to be more standardized in the United States: it is typically measured as the highest level an individual has achieved. Education can also be grouped into measurements such as the completion of high school or the completion of college. These categories can be country-specific and therefore educational groups may vary across countries. Income can be measured at the individual or household level and it can also be lumped into classifications, depending on the needs of the study or the details of the available information. Income is often related to wealth, although wealth encompasses more information than income alone. Wealth is defined as an individual or household's total assets including income, property, items, and debts [3]. While income can be used to assess socioeconomic status, it is not a direct proxy for wealth [4].

Subjective socioeconomic status is a measurement of an individual's perception of their place within society's socioeconomic structure [3]. Subjective socioeconomic status is measured by interviewing individuals and asking them to place themselves in a societal hierarchy, represented by the rungs of a ladder. One example of participant instructions for evaluating subjective socioeconomic status from Singh-Manoux, Marmot, and Adler is:

> Think of this ladder as representing where people stand in society. At the top of the ladder are the people who are best off—those who have the most money, most education and the best jobs. At the bottom are the people who are worst off—who have the least money, least education and the worst jobs or no job. The higher up you are on this ladder, the closer you are to people at the very top and the lower you are, the closer you are to the bottom. Where would you put yourself on the ladder? Please place a large 'X' on the rung where you think you stand.[3]

This measurement of socioeconomic status allows individuals to categorize themselves into a group. The ability of subjective socioeconomic status to capture psychosocial impacts of socioeconomic status, as well as the more precise measurement of social position reflected in subjective socioeconomic status, allows this measure to be a better predictor of health status and decline of health over time when compared with objective measures[5].

A final dimension of socioeconomic status that is utilized by investigators is neighborhood socioeconomic status. Like individual socioeconomic status, neighborhood socioeconomic status can be assessed using a diversity of indicators including household income, home values, availability of grocery stores, and others. Neighborhood socioeconomic status can influence an individual's health beyond the effects of the individual's socioeconomic status. For example, neighborhood socioeconomic status has been associated with an increased risk of coronary heart disease [6], increased mortality [7], makers for higher risk of cardiovascular disease [8], as well as many other health outcomes [9].

The large assortment of methods for assessing socioeconomic status makes choosing a method difficult and worth consideration. It is true that many measures of socioeconomic status are correlated, however the correlation is not always strong enough for measures to be used as proxies for each other [4]. Additionally, the relationship between the measures of socioeconomic status can vary by groups. For example, black and Mexican-American adults have significantly lower incomes when compared with white adults of the same educational level [4]. Another example is that income does not always equate to wealth. When examining US Census data, white adults are shown to have 400 times the wealth of black adults of the same income level [4]. Neighborhood socioeconomic status has similar constraints in that the measures of neighborhood socioeconomic status and their effects on health can vary by population and the measures are not always strongly correlated. Subjective socioeconomic status can reflect more aspects of the effects of socioeconomic status on health in a single measure; however, it requires survey data from participants and this is not always possible.

Socioeconomic status can be measured in various ways in epidemiological studies, ideally encompassing economic resources including income and wealth, as well as social prestige that can influence health at the individual, household, and neighborhood level [10]. Even though it is ideal to consider multiple measurements of socioeconomic status and incorporate all of them into a study design, this is not always possible. The best option for deciding on what socioeconomic measures to incorporate is to consider which measures are more likely to affect the outcome of interest and to take limitations of the measures into account when reflecting on the results of the study.

Potential pathways for how socioeconomic status can affect health

Over the years, the relationship between socioeconomic status and health has been studied in a few different approaches. Prior to the mid-1980s, socioeconomic status was considered a confounder to be controlled for in study designs, but socioeconomic status as a causal variable was not well studied [11]. During the 1980s, socioeconomic status was considered a dichotomous variable where individuals were categorized as either above or below the poverty line, implying a threshold view of the relationship between socioeconomic status and health. In this threshold model, health improved with increasing wealth until the poverty line where health became stable. This threshold model implied that there were no health differences between individuals living just above the poverty line and those who are wealthy [11]. In the mid-1980s this view began to change, as investigators influenced by Michael Marmot and the Whitehall study began to realize that the effects of socioeconomic status were on a continuum across all levels of socioeconomic status, rather than just above and below the poverty line [12]. The Whitehall study investigated morbidity and mortality among British civil servants and found a gradient among the occupational grades: more prestigious occupation grades had better morbidity and mortality when compared with less prestigious occupation grades. The gradient relationship between socioeconomic status and health exists across countries and across different health outcomes including infant mortality, mortality, and chronic diseases such as hypertension, cancer, and arthritis [11]. As socioeconomic status increases, these poor health outcomes decrease. In addition to health outcomes, health risk factors such as smoking, cholesterol, and sedentary lifestyles also show a socioeconomic status gradient, such that lower socioeconomic status individuals have higher smoking prevalence, higher cholesterol and more sedentary behaviors [11].

The association between socioeconomic status and health outcomes is clear but its interpretation is complex. Does lower socioeconomic status lead to poorer health outcomes or are poor health outcomes causing lower socioeconomic status? While some diseases can have an influence on socioeconomic status, more evidence exists for the hypothesis that socioeconomic status influences health [11]. Determining the mechanisms for how socioeconomic status affects health is an area of intense research. Socioeconomic status could affect health through several different pathways including reduced access to healthcare services, decreased knowledge of health behaviors, exposure to environmental stressors and hazards, limited financial resources, and less familial and social support [13; 14]. Figure 1 demonstrates an overview of potential pathways through which socioeconomic status may influence health. Within these broad overarching pathways there can be specific pathways. For example, environment encompasses physical environment such as exposures to pollutants and toxins, social environment such as support networks, and resources such as access to healthy food and healthcare [11]. Healthy People 2020, the United States' public health goals, includes a concise summary of the five domains of socioeconomic status that can influence health: economic stability, education, health care, neighborhood environment, and social context (Figure 2). Economic stability includes factors such as poverty, employment (or lack of), food security, and housing stability [15]. Education includes the access to education and higher education, literacy (including health literacy), and early childhood development [15]. Health care includes access to health care and primary care, as well and knowledge of health behaviors [15]. Neighborhood environment includes access to healthy foods and grocery stores, access to safe and affordable housing, low amounts of crime, violence, and exposure to toxins and pollutants [15]. Finally, social context includes civic participation, discrimination, equality, and incarceration [15]. Though the pathways between socioeconomic factors and health outcomes are difficult to distinguish and could be affecting different populations in varying degrees, it is important to consider socioeconomic status as a representation of these potential pathways.

Review of prior literature on relationship between socioeconomic status and health

Socioeconomic status has a strong association with health. This relationship has led to extensive studies on a variety of health outcomes and an assortment of dimensions of socioeconomic status. Although measurements of socioeconomic status can be correlated with each other, it is important to understand that there are different relationships between socioeconomic variables and that each measure is not necessarily a reflection of the effect of all measures of socioeconomic status on health. Due to these differences between measures, identifying the previously described relationships between socioeconomic variables and health outcomes is an significant step in studying these relationships.

Education

Education is an objective representation of social status. It can also be viewed as a reflection of earning potential, as individuals with higher education tend to have higher incomes. Education is also one of the more stable objective measures of socioeconomic status in adults, as it is less likely to change over time. For these reasons, as well as the relative ease of collection, educational attainment is often used in epidemiologic studies as a measure of socioeconomic status. Educational attainment affects health in a stepwise manner. For example, higher educational attainment is associated with longer life expectancy: men with a bachelor's degree or higher have a life expectancy 9.3 years longer than men without a high school diploma. Similarly, women with a bachelor's degree or higher have a life expectancy that is 8.6 years longer than women without a high school diploma [16]. In addition to mortality, the education gradient is negatively correlated with a number of biomarkers in the National Health and Nutrition Survey (NHANES). Fewer years of education is correlated with higher C-reactive protein (a general marker for inflammation), higher glycated hemoglobin (a general marker for type 2 diabetes risk), lower HDL cholesterol (a general marker for cardiovascular disease risk), higher waist-to-hip ratio (a general marker for obesity), higher systolic blood pressure (a general marker for hypertension), and higher resting pulse [13]. Education gradients also exist for other health outcomes and risk behaviors in the United States. Individuals with lower educational attainment are more obese compared with individuals with higher educational attainment (Figure 3) [16]. In addition to having higher obesity rates, when compared with individuals with a college degree or higher, individuals who did not complete high school are more likely to smoke (10% versus 32%), less likely to receive colorectal tests (68% versus 41%), have children who are more obese (9% versus 23%), and are less likely to breastfeed (75% versus 42%) [16].

Income and wealth

Income is often used as an objective measurement of the economic aspect of socioeconomic status. When included in study designs, income can be measured as continuous or individuals can be grouped into income categories such as above or below the federal poverty line. Additionally, in the United States, income can be reflected as a percentage of the federal poverty level. The federal poverty level is determined each year by the United States Census Bureau. This measurement is calculated by determining the minimum cost of a food diet (for individuals or families) and multiplied by three to account for other expenses [17]. An individual's or family's poverty level is determined by incorporating their earnings, which include: unemployment compensation, workers' compensation, Social Security, public assistance, veterans' payments, survivor benefits, pension, interest, rents, royalties, trusts, education assistance, alimony, child support, and any other sources of income [17]. The federal poverty level changes each year.

In the United States, many health outcomes follow an income gradient, with individuals in the lowest income categories having the worst health. Examples of this association between lower income and poorer health can be seen in children with asthma (Figure 4), depression prevalence (Figure 5), and middle age adults with two or more chronic diseases (Figure 6) [16]. This relationship is also observed in dental outcomes; when compared to individuals living at or above 400% of the poverty level, individuals living below the federal poverty level are less likely to take their children to the dentist (84% versus 70%) and more likely to lose their natural teeth by age 65 or older (41% versus 11%) [16]. While these are just some examples, this relationship has been observed for dozens of health outcomes and behaviors.

Homelessness is also an important aspect of socioeconomic status. The relationship between homelessness and health is bidirectional; poor health can lead to homelessness and homelessness can lead to poor health [18]. Individuals who are homeless may have mental health and/or other debilitating disorders which, in addition to lack of social and economic support, can lead to their homelessness [18]. Chronic, financial, and emotional stress suffered as a result of high poverty levels, in addition to the lack of proper healthcare, can lead individuals to develop mental illness [18]. Individuals who have experienced homelessness have an exceptionally high burden of poor health outcomes [19]; therefore, identifying these individuals is important and must be considered when studying health.

Occupation

Another useful measure of socioeconomic status is occupation, which can be utilized as an indicator of income or occupational prestige. Occupational prestige is a scale to describe how respected occupations are by society. Occupational prestige tends to positively correlate with income [20]. An important aspect captured by occupational prestige is the level of respect individuals perceive from others. Higher occupational prestige is correlated with better health outcomes and often the correlation between occupational prestige and health outcomes is stronger than the correlation between income and health outcomes [20]. One example of occupational prestige scores is determined from a National Opinion Research Center (NORC) survey where respondents were asked to rank occupations according to their prestige. In the survey, respondents are presented with a ladder, similar to the ladder used when determining subjective socioeconomic status. Respondents are then given cards with individual occupations listed on them and asked to place the cards on the ladder, with the highest rung representing the most respected occupations in society and the lowest rung representing the least respected occupations [20]. Higher occupational prestige is positively associated with better self-rated health[20; 21].

Health insurance

In the United States, health insurance coverage is a major determinant of access to healthcare [22]. Individuals who lack of health insurance are less likely to receive preventative medical care and needed care for chronic conditions, and are more likely to die prematurely from cancer or acute conditions such as heart attack or trauma [22]. Due to this association, it is important to consider health insurance coverage when assessing the relationship between socioeconomic status and health outcomes. A lack of health insurance is associated with poverty level for both children and adults in the United States (Figure 7) [16]. Furthermore, individuals who are less likely to have insurance due to poverty are more likely to delay needed medical care due to cost: 25% of individuals living below the poverty line delay care, versus only 6% of individuals living above the poverty line [16].

Medicaid, like occupation, can serve as a useful proxy for low-income level. In order to qualify for Medicaid assistance, households must have a maximum income at or below 133% of the federal poverty level[23]; the federal poverty level in 2016 for a household of four was $24,250. Thus Medicaid information can provide an upper limit of income for those receiving Medicaid, which is an important aspect of socioeconomic status.

Socioeconomic status, health and race

Racial health disparities

Racial disparities in health are differences in the burden of illness and mortality among racial groups [24]. The dramatic disparities experienced by black and other minorities are well documented in the 1985 Report of the Secretary's Task Force on Black and Minority Health and led to the development of expansive goals for improving minority health and the Office on Minority Health [24]. As race is a social construct with biological implications which is influenced by many factors, including an entanglement with socioeconomic status, it is vital to recognize the socioeconomic status differences that can exist between racial groups and to incorporate these measurements into studies of health outcomes with racial health disparities.

Differences in socioeconomic status for racial groups

Racial differences in wealth within the United States are extreme; for every $1.00 that whites have in wealth, Asians have $0.83, Hispanics have $0.07, and blacks have $0.06 [10]. For example, 23% of black individuals are living below the 100% poverty level versus 9% of white individuals [16]. Both black men and women have lower life expectancies when compared with whites of the same sex: 79.2 years for white men versus 72.0 years for black men, and 84.0 years for white women versus 78.1 years for black women [24]. Across racial groups, the percentage of adults in poor or fair health decreases for each increase in educational attainment. For example, 45% of black adults with less than a high school diploma report having fair/poor health, versus 29% of black adults with a high school diploma, 21% of black adults with some college, and 11% of black adults with a college degree or higher. This trend is similar in white adults, however the percentage of individuals in each education group with poor or faith health is lower when compared with black adults. Of whites with less than a high school education, 40% report poor or fair health. 19% of whites with a high school degree, 15% of whites with some college, and 5% of whites with a college degree or higher report poor or fair health [25]. This trend is observed in other underrepresented groups as well, including Hispanic and Asian populations [25]. In addition to life expectancy and self-reported health, racial disparities also exist for health outcomes such as preterm births (Figure 8) and hypertension (Figure 9) [24].

Beyond health outcomes, there are also large racial differences in receipt of healthcare. Black and Hispanic populations are more likely to be uninsured and less likely to receive needed dental care [24].Thus, racial differences exist in health outcomes and healthcare, even at similar levels of socioeconomic status.

Use of socioeconomic status in studies of racial health disparities

Due to racial differences in health outcomes as well as socioeconomic factors, measurements of socioeconomic status should be included when studying any health-related issue with observed racial differences. Typically, in genetic studies, race/ethnicity or genetic ancestry (a measurement of the population origin of genetic variants in an individual) are included in statistical models in order to avoid population stratification. The potential association of race/ethnicity with both socioeconomic status variables and health outcomes could lead to confounding by socioeconomic status in these genetic association studies. Consequently, it is possible that any associations between genetic ancestry and disease could actually reflect an association between socioeconomic status and disease. Race can also be a proxy for other environmental factors, such as racism, beyond socioeconomic status. By including socioeconomic status in studies of health outcomes, we can further elucidate the environmental factors affecting health and disentangle this complex relationship between social environment, genetics, and health.

Socioeconomic status in genetic studies

Gene-environment interactions

A gene-environment interaction is defined as “a different effect of an environmental exposure on disease risk in persons with different genotypes or a different effect of genotype on disease risk in persons with different environmental exposures” [26]. There are five potential models for how gene-environment interactions may affect biology (Figure 10). Model A describes an interaction where the genotype produces or increases expression of a disease risk factor that can also occur environmentally (Figure 10) [26]. In model B, the risk genotype can exacerbate the effect of an environmental exposure, but individuals with the risk genotype and without the environmental exposure are not affected

(Figure 10) [26]. Model C describes situations where the environmental exposure exacerbates the effect of the genotype, but individuals with the low risk genotype are not affected by the exposure (Figure 10) [26]. In model D, the genotype and the exposure are both needed to increase the risk of the poor outcome (Figure 10) [26]. Finally, in model E, the environmental exposure and the genotype can both contribute to risk; however the presence of both in an individual can either increase or decrease the risk (Figure 10) [26].

Socioeconomic status can serve as a proxy for many types of environmental exposures including increased stress due to lack of resources, lack of medical care, exposure to environmental toxins, lack of access to healthy foods and other pathways [13; 14]. There are four general mechanisms as to how socioeconomic status can moderate genetic effects: an individual's biological response to stress can be affected by genes, genes can affect how an individual adapts to the social environment, inherited characteristics can help make an individual more suited for certain environments, and inherited characteristics may only display in some environments [27]. It is important to note that the interpretation of interactions must be carefully considered as socioeconomic status is a broad category and can be representative of many factors such as stress due to low resources or financial strain.

Socioeconomic status in genetic studies of racial health disparities

Despite the overwhelming evidence that socioeconomic status affects health outcomes, measurements of socioeconomic status are not often included in genetic studies of disease and racial disparities. The lack of inclusion of socioeconomic status data may be due to the lack of available data in existing cohorts, as well as the additional time and resources it takes to collect socioeconomic status data for new studies. Even with the challenges, socioeconomic factors must still be included in studies of health outcomes with racial differences. In addition to the potential confounding by socioeconomic status that may occur due to the association of race/ethnicity with both socioeconomic status and health outcomes in the United States, factors represented by socioeconomic status have the potential to modify the effect of genetic variants on health outcomes as well as be the cause of health outcomes or health disparities. Therefore, the biology of disease is likely to be misunderstood without the inclusion of socioeconomic status data in association studies.

Blood pressure and hypertension

Blood pressure is defined as the force of blood pushing on arteries as the heart pumps blood. Systolic blood pressure is the force when the heart contracts when pumping blood. Diastolic blood pressure if the force of the blood when the heart is at rest [28]. Normal blood pressure for adults is defined as less than 120 mmHg for systolic blood pressure and less than 80 mmHg for diastolic blood pressure. Blood pressure varies throughout the day and can be affected by physical activity, sleep, stress and other factors.

Hypertension is a common disease defined by high blood pressure, affecting over one billion people throughout the world today [29]. Hypertension is defined as blood pressure higher than 120/80 mmHg. In the United States, hypertension is characterized by three stages: prehypertension, high blood pressure stage 1, and high blood pressure stage 2 [28]. Prehypertension has a range of 120-139 mmHg for systolic blood pressure or 80-89 mmHg for diastolic blood pressure. Stage 1 has a range of 140-159 mmHg for systolic blood pressure or 90-99 for diastolic blood pressure. Stage 2 is defined as systolic blood pressure of 160 mmHg or higher or diastolic blood pressure of 100 mmHg or higher [28]. A diagnosis of hypertension typically requires five measurements of clinically measured high blood pressure, as patients tend to have higher blood pressure in the clinic due to stress or illness [30].

The prevalence of hypertension in the United States increases with age, with individuals 60 years of age and older having the highest prevalence [31]. Non-Hispanic black men and women have a roughly 15% higher prevalence of hypertension than other racial groups in the United States [31]. While the overall prevalence of hypertension has not changed much in recent years, the percent of individuals with controlled hypertension has increased from 31.5% of adults in the United States in 2000 to 54% in 2014 [31]. Controlled hypertension indicates individuals with hypertension whose blood pressure measurements are below 140 mmHg for systolic blood pressure and 90 mmHg for diastolic blood pressure due to medication use. Despite the increase in controlled hypertension, underrepresented individuals with hypertension are less likely to have their hypertension under control when compared with whites; 55.7% of white individuals have their hypertension under control, compared with 47.5% of black individuals, 43.5% of Asian individuals, and 47.4% of Hispanic individuals [31].

Effect of hypertension on health

While hypertension in itself is a serious health problem, it can also lead to other life-threatening conditions including myocardial infarction, cardiac failure, and kidney disease [29]. As a result of the higher prevalence of hypertension among African Americans, they face a larger disease burden of comorbidities and conditions resulting from hypertension, such as stroke, heart failure and end-stage renal disease [32], as well as a three times higher death rate due to hypertension [29].

Genetics of blood pressure

Genome wide association (GWA) studies, to date, have led to the discovery of numerous genetic variants that may be contributing to blood pressure. The majority of these studies are in populations of European ancestry. Within this group, GWA studies have identified 83 loci associated with blood pressure, hypertension, or pulse pressure [33-35] [33; 36-38]. Additional studies have also been conducted in Asian populations. These GWA studies have identified a total of 23 loci associated with blood pressure or hypertension [39-43].

The first GWA study of hypertension in a black population (N=1,017) identified five loci associated with systolic blood pressure [29]; however, these findings were not replicated in a later study [44]. Another GWA study replicated three single nucleotide polymorphisms (SNPs) previously associated with blood pressure in Europeans in a black population [32]. An additional study in black adults using admixture mapping identified a locus that was significantly associated with systolic blood pressure and diastolic blood pressure in both an initial dataset and a replication dataset [45].

A more recent GWA study of BP in a black population performed a large meta-analysis including 29,378 individuals of African ancestry, as well as multi-ethnic replication cohorts, to find five additional loci that were significantly associated with either SBP or DBP across the cohorts studied [46]. This same dataset was used in a meta-analysis of the correlated traits SBP, DBP, and hypertension where four loci were significantly associate with blood pressure [47]. One of the more recent publications involved the use of three biobanks: Genetic Epidemiology Research on Adult Health and Aging (GERA), International Consortium for Blood Pressure (ICBP), and the UK Biobank (UKB). This large study identified (and replicated) 75 novel loci associated with blood pressure across these cohorts which consisted of multiple ancestral populations [48]. The most recent study to date conducted a large meta-analysis of 21 GWA studies, consisting of 31,968 individuals of African ancestry and a validation with 54,395 individuals from multi-ethnic studies[49]. This study found nine loci with eleven independent variants which associated with either systolic or diastolic blood pressure, hypertension, or combined traits. Among these associations, four variants were only common in African ancestry populations[49].

These GWAS-identified SNPs range in effect size from -1.0 mmHg to 3.28 mmHg. Despite these studies, the percent of variance explained by GWAS-identified SNPs to date is only around 25%[50]. Although these studies have utilized large populations and found many variants contributing to blood pressure in black individuals, none have controlled for or included socioeconomic status information within their analyses, but instead only control for variables such as age, sex, and principal components of genetic ancestry.

Due to the high health impact of hypertension, as well as the high heritability estimates of 30-70% [51], the genetics of blood pressure remain an important area of investigation. Although some of these SNPs have been confirmed to contribute to the estimated heritability of blood pressure, they still do not explain the total estimated heritability, resulting in a mystery of "missing heritability" [51]. The "missing" heritability estimates may be explained, in part, by other factors, such as socioeconomic status, that interact with genetic variation and contribute to the variance observed in blood pressure. Recent genetic studies of blood pressure and hypertension in black populations have focused only on demographic and medical factors that can affect blood pressure such as age, sex, body mass index (BMI), genetic ancestry and medications prescribed for hypertension [29; 32; 44; 46]. In general, these studies have identified a small number of SNPs that appear to contribute minimally to trait variance in blood pressure.

Socioeconomic status and blood pressure

Use of socioeconomic status in genetic studies of blood pressure

Low socioeconomic status is strongly associated with hypertension and related cardiovascular comorbidities and mortality [13; 52-57]. However, among the blood pressure and hypertension GWA studies conducted to date, only one has included any measurements of the social environment, which was in the form of education [58]. This neglect of socioeconomic status continues among genetic studies, despite the fact that numerous epidemiologic studies have found that social environment, specifically socioeconomic status, is associated with blood pressure and hypertension [13; 52-57; 59]. Education and income gradients are inversely correlated with markers of cardiovascular disease risk, including hypertension, such that people with lower education and lower income have a higher risk of hypertension [13].

Socioeconomic status may potentially affect blood pressure through a number of pathways including access to healthcare services, knowledge, awareness of hypertension as a disease, exposure to environmental hazards and stressors, limited financial resources, and less familial and social support [13; 14]. Many of these pathways can lead to an individual experiencing chronic stress. Genetic variants can potentially influence an individual’s biological response to chronic stress via stress response pathways. Therefore, these variants can affect the biological outcomes of exposure to different levels of socioeconomic status. Without the inclusion of socioeconomic status in genetic studies, we cannot elucidate the relationship between social environment and biological outcomes such as hypertension.

Blood pressure and gene-environment interactions

Within the past few years, studies have begun to examine interactions of environmental and demographic factors with genetic variants to determine if interactions account for a larger part of the heritability of blood pressure. Investigations have examined interactions between SNPs associated with blood pressure and age, alcohol consumption, smoking, and education in European ancestry populations [58; 60-62]. Together these studies identified a total of 31 novel loci significantly associated with blood pressure by testing for interactions between the environmental and demographic variables and the SNPs. These results indicate that investigation of gene-environment interactions holds promise in contributing to the knowledge of blood pressure etiology. One study investigated interactions between genetic factors and education, by examining 487,988 SNPs in the Framingham Heart Study [58]. Despite using only one limited measure of socioeconomic status, completion of high school or the completion of college, the study identified novel SNP x education interactions associated with blood pressure[58]. The effect sizes of these interactions ranged from -5.40 mmHg to 5.50 mmHg. Previous studies of blood pressure in the Framingham Heart Study had not detected the associations with the variants that were found when examining education-SNP interactions, suggesting that accounting for gene-environment interactions may reveal novel genetic associations with blood pressure. The Framingham study focused on white individuals, and only included a limited measure of the social environment. Investigating genetic interactions with more comprehensive measures of socioeconomic status, as applied to more diverse populations, has yet to be explored.

Potential models for gene-socioeconomic status interactions

In the case of gene-socioeconomic status interactions, there are several possible models to explain how an interaction may function in the case of hypertension [26] (Figure 11). In model I, a genotype may exacerbate the effect of the exposure, in this case, low socioeconomic status. Under this model, there may be a variant which further increases a person's risk of developing hypertension beyond the expected increase due to exposure to low socioeconomic status. When a person who has the risk variant is not exposed to low socioeconomic status, they would not have an increased risk of hypertension. Also under this first model, there could be an interaction in which the genotype suppresses the effect of the low socioeconomic status exposure. In model II, the exposure to low socioeconomic status and the risk genotype can both have a main effect on disease risk; however the risk may be higher if both occur together in one individual. Therefore, having both a risk genotype and exposure to low socioeconomic status could lead to an even greater risk of hypertension than either factor alone. Under this model, there is also the possibility that a genotype may interact with the exposure to reduce the risk of hypertension in individuals;

for example if an individual is exposed to low socioeconomic status but has normal blood pressure, the genotype may have a protective effect. In this model, both the genotype and the exposure have individual effects that can be combined to affect risk.

Summary

The scientific contribution of this project will be significant because it attempts to address the missing heritability of blood pressure in a population that has a high disease burden of hypertension. This research examines how the social environment, in the form of socioeconomic status, is contributing as a main effect and how it may interact with genetics to influence variation within blood pressure. Analyses of the interaction between genetics and social environment will lead to a better understanding of the etiology of hypertension. This information will be invaluable for motivation for social change or interventions to address socioeconomic disparities. Improved awareness of the biology of hypertension can lead to enhanced prevention, treatment, and decreased mortality for the high percentage of people within the United States that are affected. It is also likely that other common diseases are affected by socioeconomic status and gene-environment interactions. This research lays the groundwork to increase access to socioeconomic status information, and demonstrates the importance of incorporating this data into genetic studies of other common diseases.

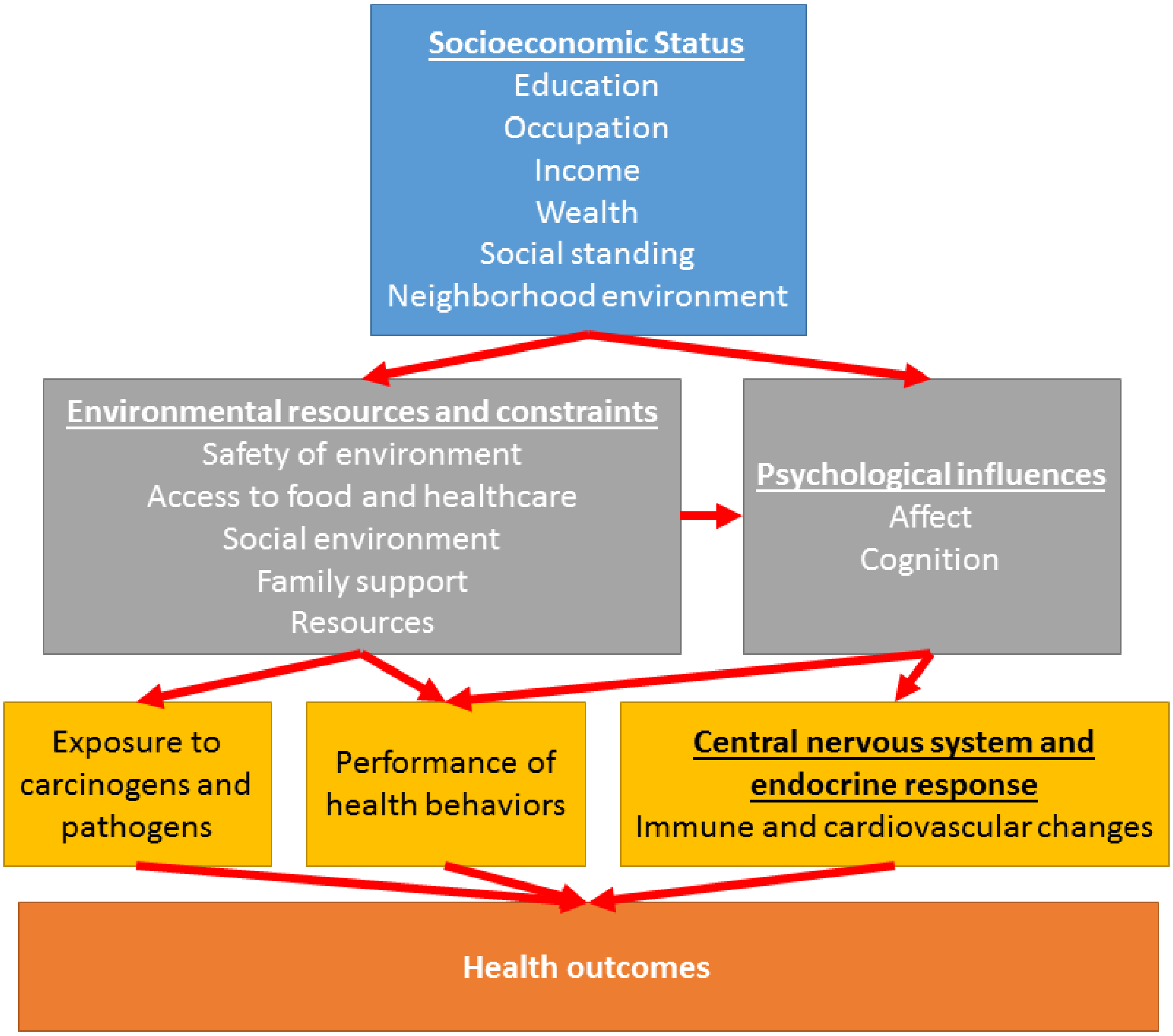
Socioeconomic Status
Education
Occupation
Income
Wealth
Social standing
Neighborhood environment
Environmental resources and constraints
Safety of environment
Access to food and healthcare
Social environment
Family support
Resources
Psychological influences
Affect
Cognition
Exposure to carcinogens and pathogens
Performance of health behaviors
Central nervous system and endocrine response
Immune and cardiovascular changes
Health outcomes

Figure 1

Model of pathways by which socioeconomic status affects health. Modified from [11].

Socioeconomic status can affect many realms in a person's life, but in general the effect on environment and psychology are two main pathways for influencing health outcomes.

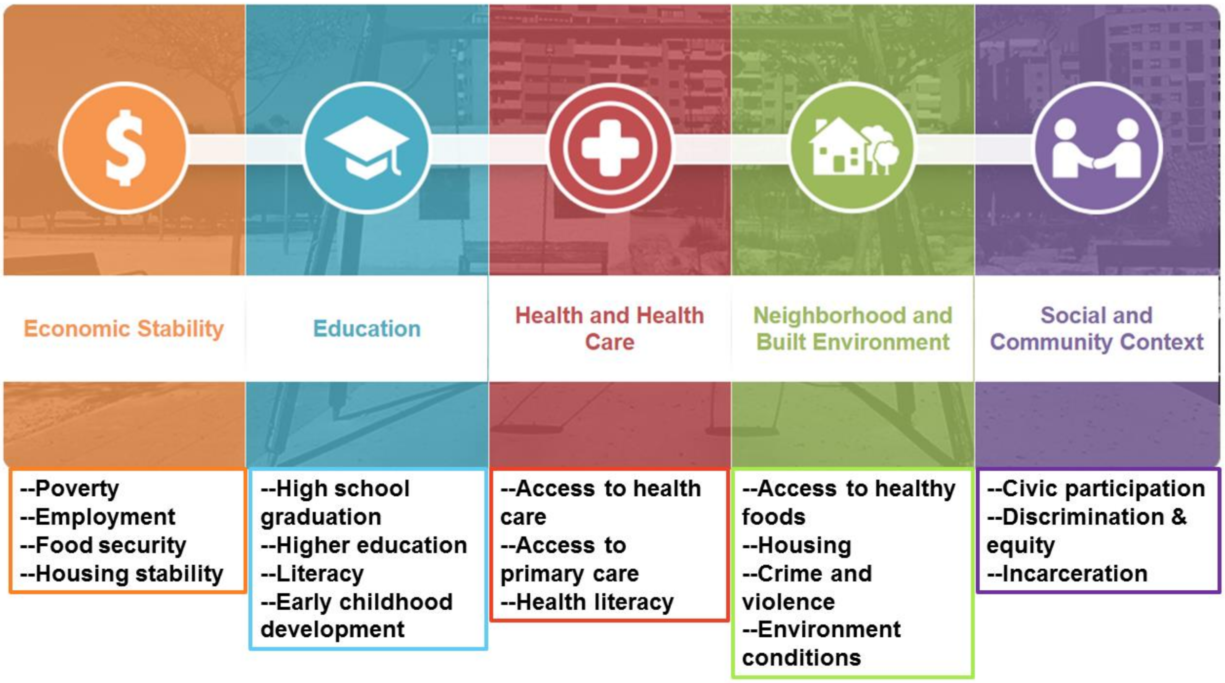

Figure 2

Socioeconomic status encompasses five realms which can affect health: economic stability, education, health care, neighborhood environment, and social context. Modified from[15].

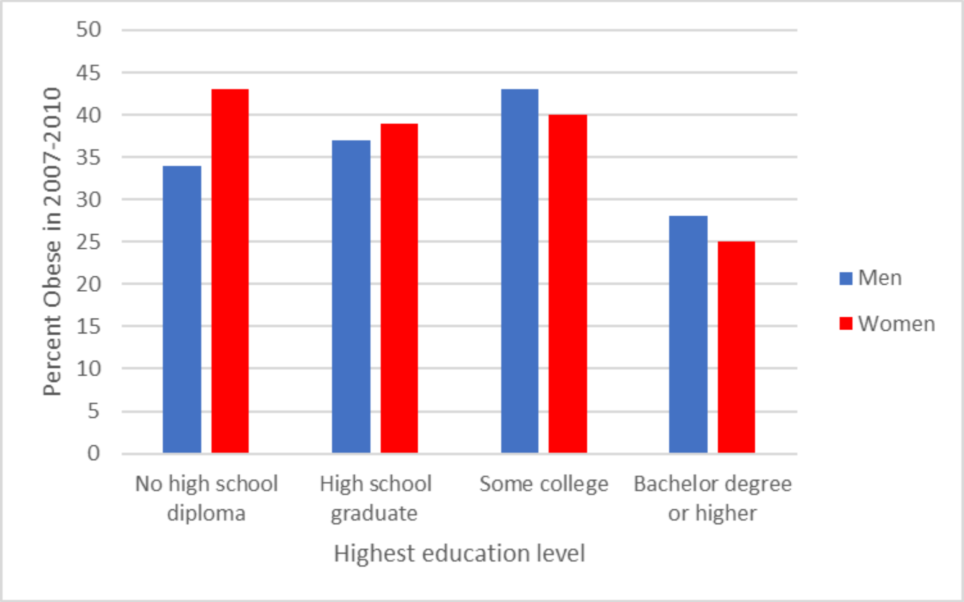

Figure 3

Obesity among adults 25 years and older by sex and education level: United States 2007-2010. Modified from [16].

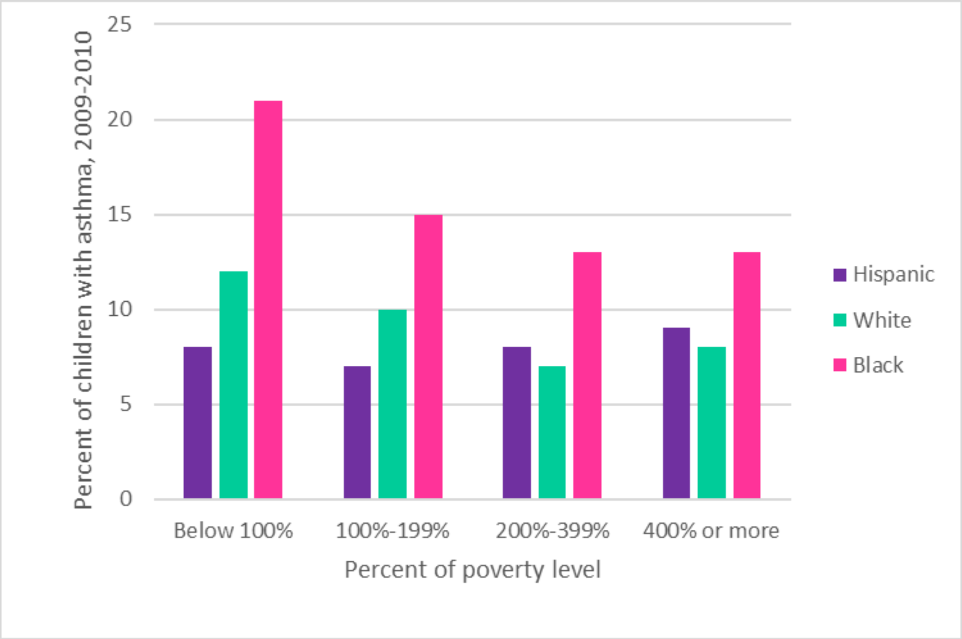

Figure 4

Current asthma among children under 18 years, by race/ethnicity and percent of poverty level, 2009-2010. Modified from[16].

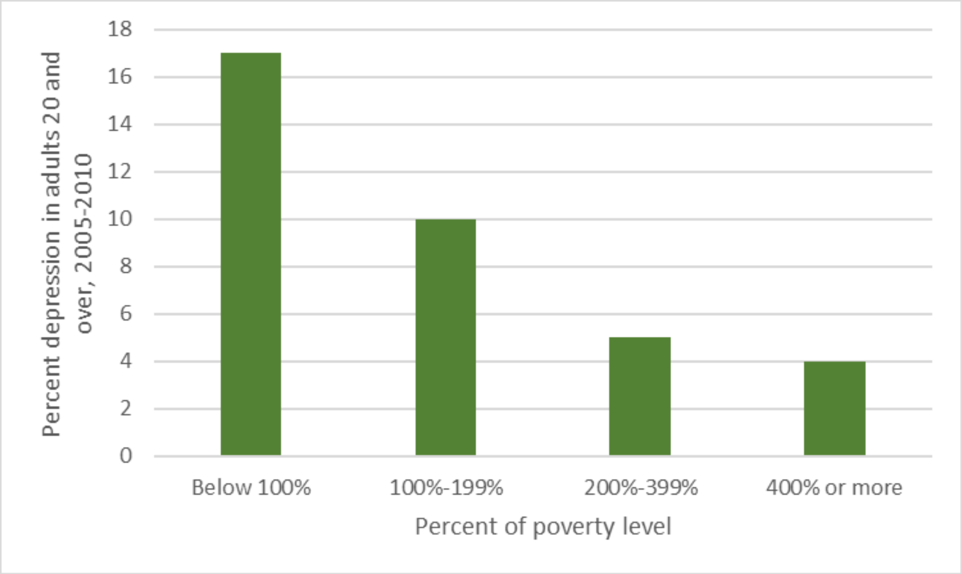

Figure 5

Depression among adults 20 years of age and over by percent of poverty level, 2005-2010. Modified from[16]*.*

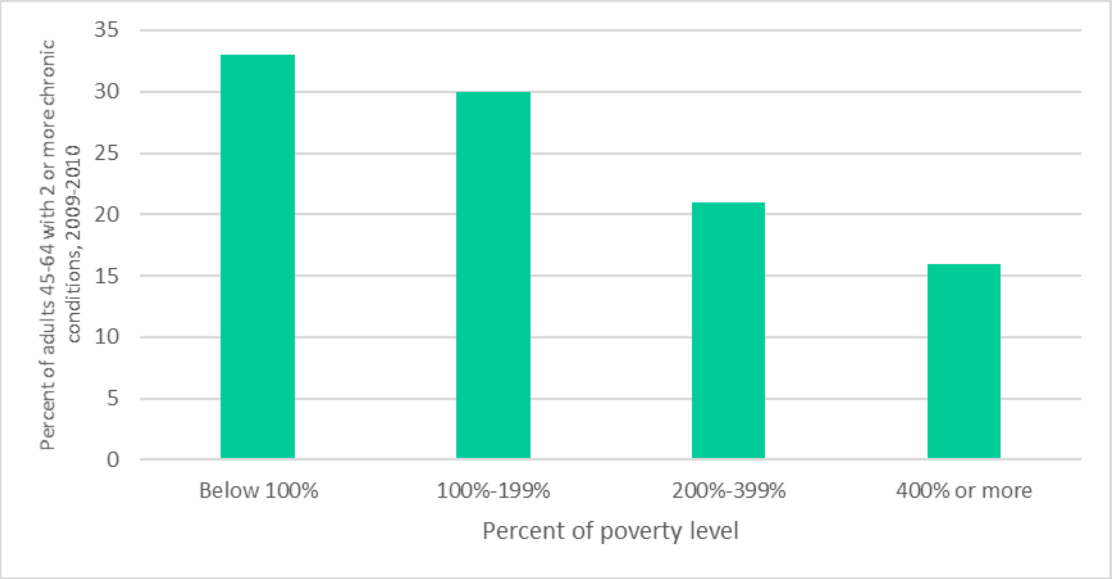

Figure 6

Adults between the ages of 45 and 65 years with two or more chronic health conditions by percent of poverty level, 2009-2010. Modified from [16].

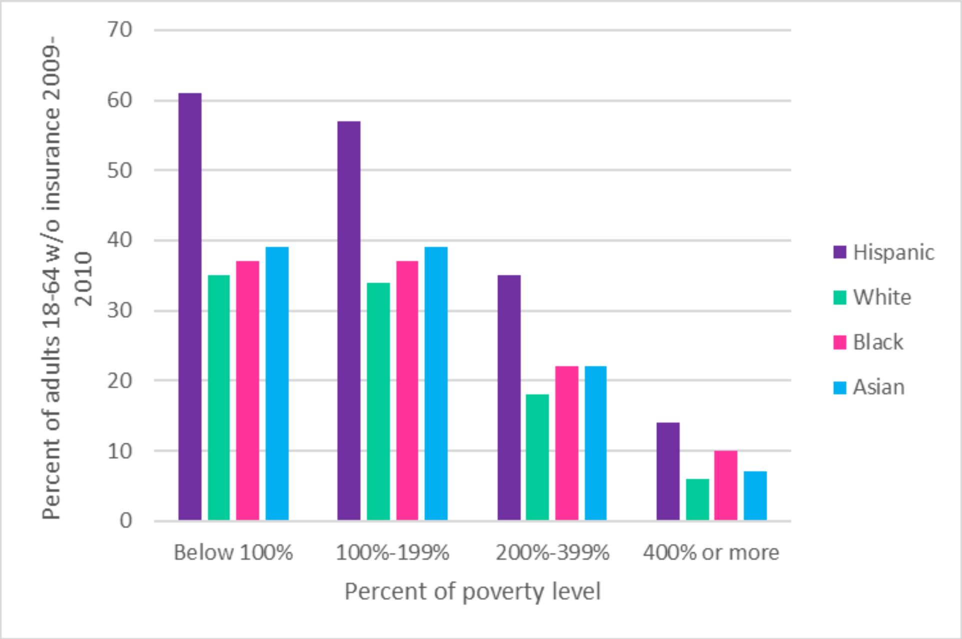

Figure 7

Adults between the ages of 18 and 64 years without health insurance coverage by percent of poverty level and race/ethnicity, 2000-2010. Modified from [16].

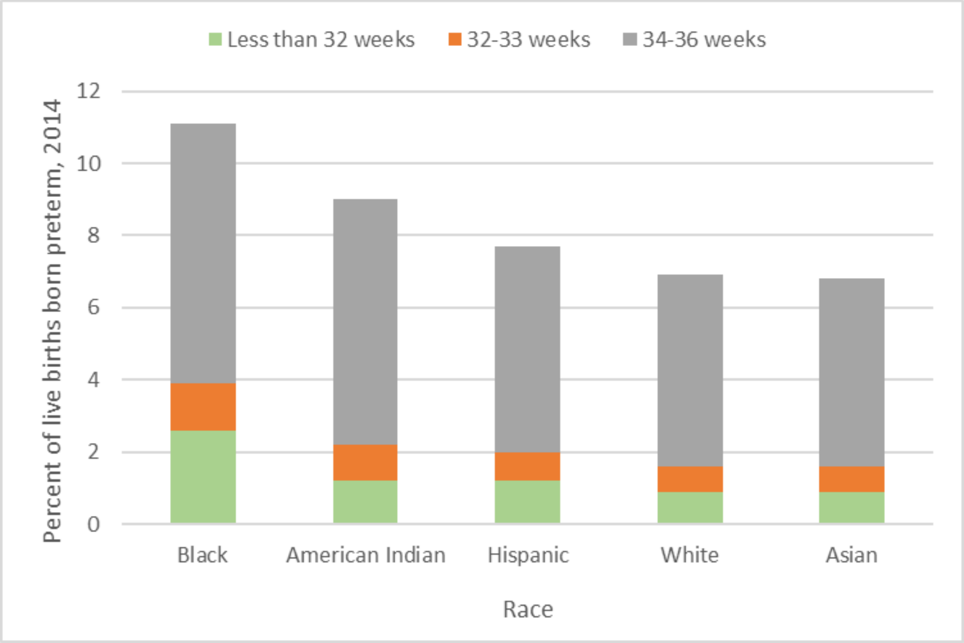

Figure 8

Preterm births by gestational age and race/ethnicity of mother, 2014. Modified from [24].

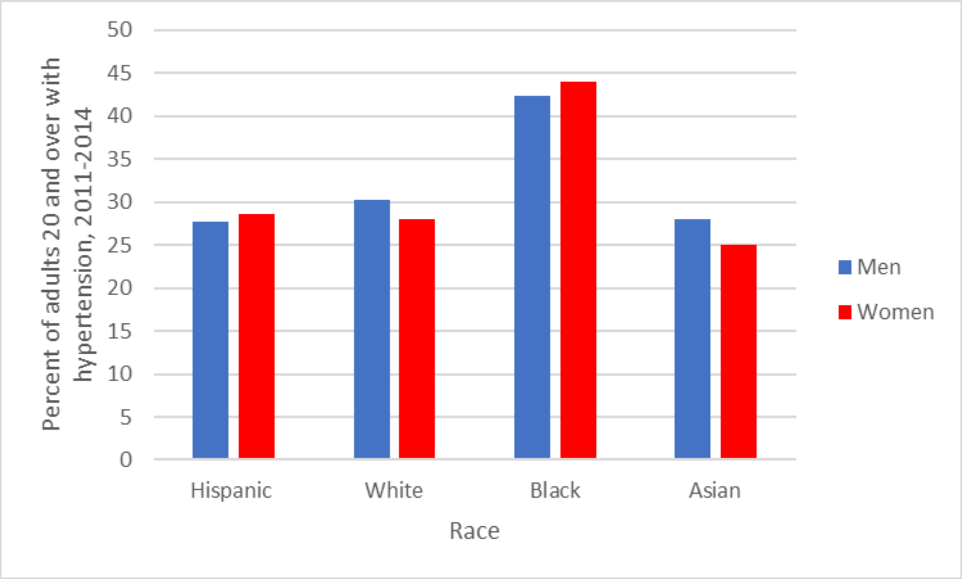

Figure 9

Hypertension among adults age 20 years and older, by sex and race/ethnicity, 2011-2014. Modified from [24].

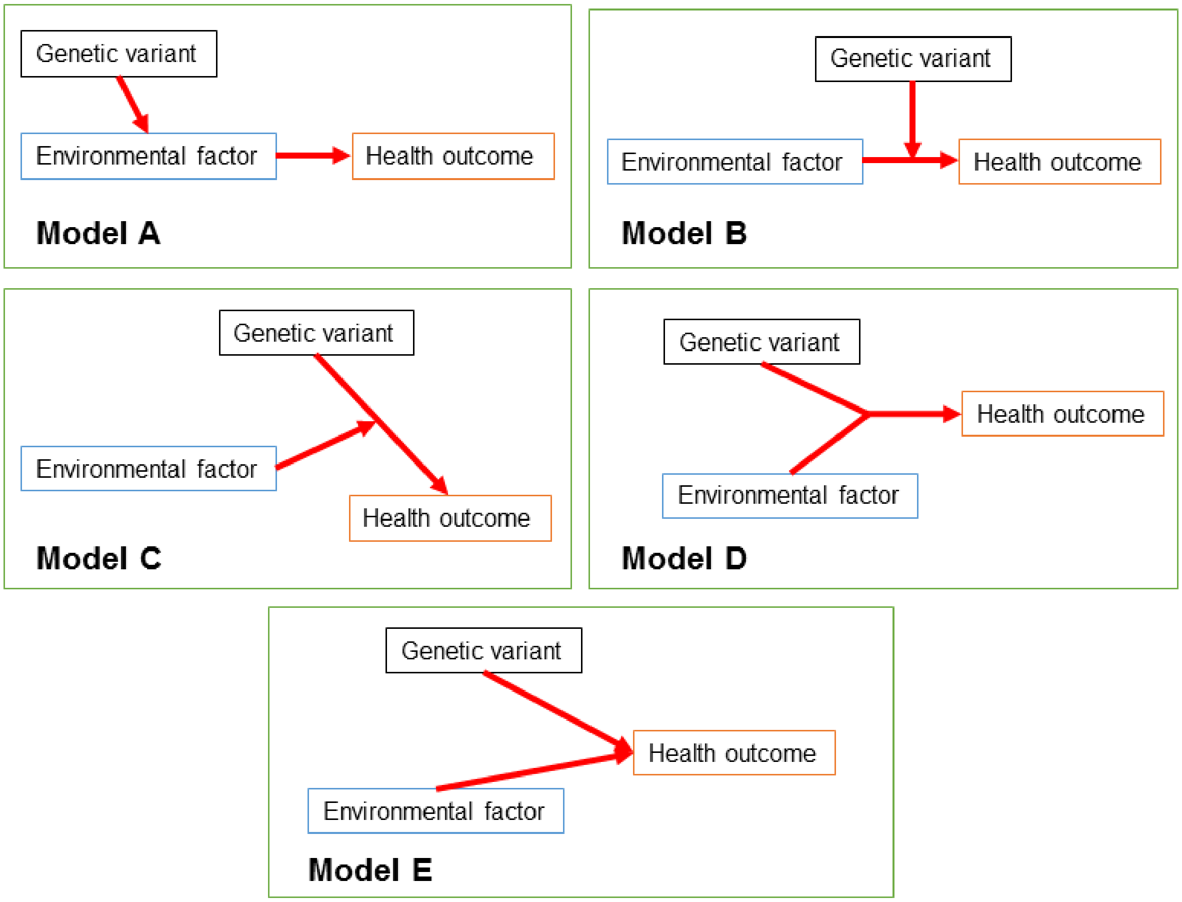

Figure 10

Potential gene-environment interaction models. Modified from [26].

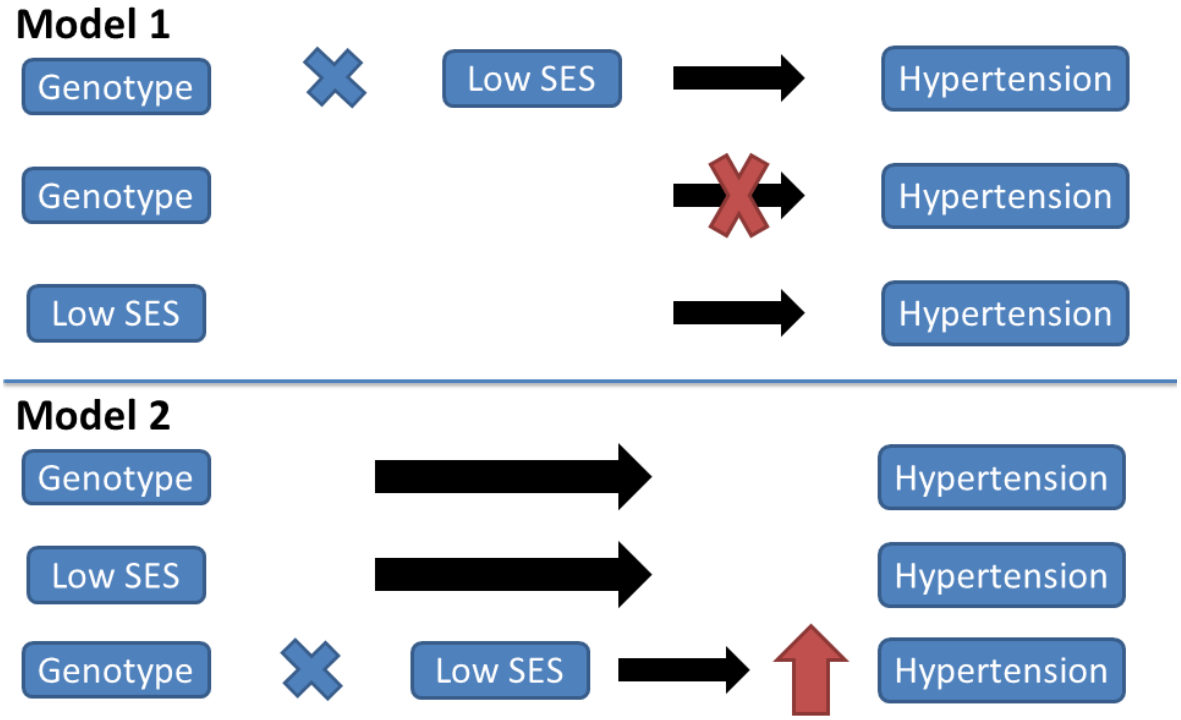

Figure 11

Potential models for the interaction between genotype and socioeconomic status which may affect hypertension.

CHAPTER TWO

DEVELOPMENT OF ALGORITHMS TO EXTRACT SOCIOECONOMIC STATUS VARIABLES FROM ELECTRONIC HEALTH RECORDS

Introduction

Socioeconomic status in research

As evidence demonstrates in the previous chapter, socioeconomic status is an important contributor to health outcomes and therefore must be included in studies of health. There are many different methods for examining socioeconomic status including at the individual, family, and neighborhood level. To date, socioeconomic status is typically captured by researchers through survey methods or utilizing government resources such as census data. Survey methods are useful in that investigators can be specific and comprehensive when collecting socioeconomic status information. However, collecting survey data from large populations can take a lot of time and be very expensive. These methods are also not very useful on existing large datasets. The use of census information to measure socioeconomic status is only useful if address or other location information for the participant is included in the study data. The movement to de-identified data in order to protect participants makes it impossible to utilize census level data to measure socioeconomic status.

Socioeconomic status data within electronic health records

The use of electronic health records (EHRs) for research purposes is becoming increasingly prevalent. The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 promoted the adoption of EHRs by clinical centers[63]. The increasing adoption of EHRs created a potential resource for large-scale epidemiological analyses. With the announcement of the Precision Medicine Initiative, now called All of Us, and its goal of recruiting one million participants with biological, environmental, and

EHR data, the research use of EHRs is anticipated to increase [64]. EHRs provide an attractive resource for biomedical researchers for many reasons, including their rich phenotypic and longitudinal data, as well as the lower cost of participant recruitment versus a traditional observational epidemiology study. Additionally, clinical biobanks that contain biological samples linked to EHRs are becoming an invaluable resource for conducting genetic epidemiology studies. Currently, the focus of EHR algorithms has been extracting clinical phenotype information for disease-focused study designs. When examining algorithm depositories such as PheKB, it is clear that the emphasis on EHR algorithm development has been disease phenotypes for case-control studies. Generally, these phenotype algorithms have been developed utilizing a combination of ICD-9 billing codes, CPT procedural codes, medication lists, laboratory and clinical values, and natural language processing. Despite the potential for EHRs in research settings, these clinical data repositories currently have noted deficits in the availability and completeness of important social and environmental data [65], including socioeconomic status, that are known to contribute independently to health status and could modify genetic effects [58].

In recognition of the importance of formally and systematically capturing social and behavioral measures in the EHR, the Institute of Medicine (IOM) recently recommended socioeconomic status measures, specifically educational attainment, financial resource strain, and neighborhood median household income be included in the EHR. The committee also recommended that a plan be developed by the NIH to expand the research use of EHRs to include social and behavioral data. Adoption of these recommendations will take time, and may not be universal across medical centers; therefore, there is a need to develop approaches and methods to access existing unstructured socioeconomic status data within the EHR for research purposes. Socioeconomic status data are almost entirely found within the free text clinical notes written by providers. We developed an approach for extracting available socioeconomic status information from the free text of a de-identified EHR. These algorithms will facilitate the immediate extraction of key socioeconomic status information from de-identified clinical biobanks for incorporation into future biomedical research.

BioVU

BioVU is the DNA biobank of the Vanderbilt University Medical Center (VUMC) linked to de-identified EHRs. DNA samples are extracted from discarded blood samples drawn for routine clinical care [66]. Sample collection began in 2007. When samples were first collected, patients were required to opt-out of BioVU. During their clinical visits, patients were presented with a consent form where they would need to indicate that they did not want to be in BioVU if they did not want their sample included in the biobank. As of 2015, BioVU has switched to an opt-in model where patients must indicate that they would like to be included in BioVU in order for their sample to be eligible. DNA samples are linked to the Synthetic Derivative (SD), the de-identified version of the VUMC EHR, by a unique study ID. Medical records within the SD are scrubbed of all Health Insurance Portability and Accountability Act (HIPAA) identifiers such as names, locations, zip codes, and social security numbers. Dates within each SD record are shifted to prevent re-identification of the records. Date shifting is consistent within a single patient's record. As previously described [67], data from BioVU are de-identified in accordance with provisions of Title 45, Code of Federal Regulations, part 46 (45 CFT 46); consequently, this study is considered non-human subjects research by the Vanderbilt University Institutional Review Board.

Methods

Population

The study population included all racial/ethnic minority patients ≥18 years old participating in BioVU as of 2011[68]. These patients were selected in order to explore the genetic variation within non-white populations, as the vast majority of large-scale genetic studies to date have focused on white populations[69]. The EHRs used for the development of the algorithms were updated in 2015 to include current information. Race/ethnicity is reported by the provider in BioVU and strongly correlated with genetic ancestry [70; 71]. The majority (81%) of patients in the dataset are black individuals. And the mean age is 50 years as of 2015 (Table 1). The mean number of clinic visits within the population in a patient's EHR record is 40.45 visits, and the mean number of days between patients' first and last visit within the EHR is 2,340 days (Table 1).

Development of algorithms

The goal is to develop algorithms to extract socioeconomic status information from structured and unstructured text in the de-identified EHRs. Seven algorithms were developed to extract education level, occupation, unemployment, retirement, insurance status, Medicaid status, and homelessness (Table 2). The initial development of the socioeconomic status algorithms began with a manual review of both structured and unstructured data within the de-identified EHR of 200 randomly selected minority patients to identify the following: 1) the categories of socioeconomic status information most frequently mentioned, 2) where in the EHR this information is noted, and 3) the semantic language used by clinical providers for socioeconomic information (Figure 12). The manual review revealed that the socioeconomic status data were found exclusively within the unstructured free text of the clinical notes, social history, and clinical communications. It was also noted that the most frequently mentioned semantic categories were employment, education, insurance status, and homelessness, and thus these categories were chosen for extraction. Semantic tags for each category were selected if they appeared more than once within the 200 development records.

Employment

Employment information was extracted using three different algorithms designed to capture data on occupation, unemployment, and retirement. The occupation algorithm extracts the occupation mentioned in a patient's record and translates it to an occupational prestige score (scale 0-100). This score represents how well-respected an occupation is within a society (i.e., subjective socioeconomic position). Occupational prestige scores were developed from a National Opinion Research Center (NORC) survey where respondents were asked to rank occupations according to their prestige [72]. The occupation tags utilized for the occupation algorithm were adopted from the most recent NORC report [72]. The algorithm's occupation tags were shortened to 678 occupations from the original NORC list of 860 occupations given that some of the occupations were highly specific with repetitive occupational prestige scores. As an example, “teacher, elementary school” and “teacher, secondary school” were collapsed to “teacher.”

The occupation algorithm was used to search the unstructured data of the original 200 patients for the initial occupation tags. This search identified a large number of false positives, where the algorithm tagged occupation-related words that were not indicative of the patient’s occupation, which we referred to as “false positives”. In this case, false positive refers to the inaccurate identification of socioeconomic status information by the algorithm. Several methods were used to filter these false positives. The first attempt removed any occupations that had more than 10 false positive entries. When this method was utilized, over half of the occupation tags were lost and only a minor set of occupation information was identified. This small dataset still had a high number of false positives. The second method was inclusion of prefix language filters. With this approach, the list of 678 occupations was used and 10 prefixes were added: is a, is an, works as, works in, works at, occupation, is the, as a, as an, former. These prefixes were selected based on previous occupation algorithm results where a random selection of 200 results were reviewed, accurate patient occupations were identified based on the context of the clinical note, the prefix language that was used by providers was noted, and then the prefix was added to the occupation list. Once this list was developed, the occupation algorithm was implemented by requiring results to be identified as a patient’s occupation only if one of the prefixes was found in front of the occupation word. This method reduced the number of false positives; however, 75% of the original occupation data was lost.

The third method included reviewing an additional selection of the occupation results from the first method (without the use of prefixes) for additional prefix language. After further review of a random set of 200 records, 15 additional prefixes were added: social history, retired, she is a, she is an, he is a, he is an, was a, was an, used to be, assistant, pt is, patient is, employment, employed, employ. The occupation algorithm was implements and one prefix was required to be present before an occupation tag in order for the algorithm to identify the occupation as the patient’s occupation. With this method, the number of results increased, while maintaining a low number of false positives. However, only a fraction of the results from the initial search were identified (Method 1).

In the fourth method, modifications were made to the occupation list that required the use of the prefixes. In this method, the assumption was that only occupations related to the medical field would require a prefix. It was assumed that these occupations were likely to have the highest rate of false positives, as medical occupations are frequently mentioned in a patient’s health record when related to the patient’s care. Therefore, the occupation algorithm was run with two separate groups: a list of occupation tags that did not require a prefix and a list of occupation tags that required a prefix. This method greatly increased the number of results, but it also slightly increased the number of false positives.

The final method that was used increased the number of occupations on the list which required the use of a prefix. After a review of the false positives from the fourth method, it was noted that there were additional occupations that had not been classified as medical occupations, which appeared in the patient’s record when related to the patient’s care. Medical occupations were added to the list of occupations which required a prefix. After running these results, the balance between number of results and number of false positives was optimal. There were still a small number of false positives, but much fewer than some of the earlier methods, while still maintaining a large number of results.

Unemployment data were extracted using semantic tags for unemployment (e.g., “unemployed,” “does not work,” “hasn’t worked since”). The unemployment algorithm was then tested on the unstructured data from the 200 records used for development, and a high number of false positives were returned. These false positives were often in reference to medications. Therefore the tags “if this does not work” and “if that does not work” were excluded to filter false positives. The addition of these tags essentially eliminated the false positives from the results. Unemployment was classified as ever/never (Table 2). Retirement was also extracted from the EHR using the tag “retired” and classified as ever/never (Table 2). The tag “retired” accurately extracted patients who were identified as retired within their health record, without the need for additional filtering.

Education

The education algorithm was designed to assign education level to a patient based on the highest education achieved and recorded in the EHR. The first method to classify education levels focused on searching for the term “education:” and then classifying a patient's education based on what came after that tag. However, this method lead to a large number of false positives, due to the other types of education found within the EHR such as diet and dialysis education. Additionally, this method missed a lot of the education information that was in the EHR because the majority of education information is found within the narrative of the provider notes, rather than a list format.

For the second method, a different approach was taken. Education levels were assigned to each relevant tag word or phrase found in the unstructured text of the EHR (Table 2). Sixty-two semantic tags were utilized and the highest level of education was determined for each patient. These tags were exclusive to an assigned education level. For example, the high school degree category of education level included tags such as "high school graduate" and "completed 12th grade," while the bachelor's degree category included terms such as "BS degree" and "completed college." The levels of education were based on U.S. census definitions with one modification such that all grade levels below high school graduate were collapsed into a "less than high school" category. Searches were conducted through the unstructured text of the 200 records used for development to determine if further filtering or modification was needed. Fifteen additional tags were used to filter false positive results related to types of medical education (e.g. "diet education," "dialysis education") and Vanderbilt Medical School students (e.g., "medical student," "pharmacy student," "student nurse").

Insurance status

Due to the nature of the de-identification process of the SD, specific insurance information is not included within the patients' records. It was therefore decided to identify patients who did not have insurance and those who are on Medicaid, as this information is likely to be found within the SD and also associated with health outcomes.

The extraction process for insurance status required two algorithms. The first algorithm was used to determine if there was any time point in the EHR when the patient did not have insurance based on the presence of five semantic tags (Table 2). These tags included "no insurance" and "does not have insurance." Some language was eliminated, mainly words that were used in a standard discharge letter at VUMC and therefore appeared frequently in the EHR. This discharge letter included a generic set of instructions for patients who may not have insurance or were on Medicaid. The exclusion of the language in the discharge letter allowed for a large reduction in the number of false positives.

A second insurance algorithm extracted Medicaid information using specific phrases or keywords such as “Medicaid” and “TennCare” (Tennessee’s version of Medicaid) and was classified as ever/never in order to determine if a patient was ever on Medicaid in their EHR (Table 2).

Homelessness

Homelessness information was extracted using the tags “homeless” and “shelter” among the 200 development EHRs. After this search, several false positives were returned relating to patients who worked or volunteered at homeless shelters. Therefore, exclusion tags were added such as “volunteer at homeless shelter,” “works at homeless shelter,” “works with homeless,” and “animal shelter.” Homelessness was classified as ever/never (Table 2).

Evaluation of algorithm performance

To evaluate the performance of these socioeconomic status algorithms, results were compared to findings from a manual review of 50 randomly selected patients. These 50 individuals were selected using random sampling without replacement. Two independent reviewers manually reviewed the clinical record of each patient and any discrepancies were resolved by discussion between the two reviewers. Comparison of results from the two independent reviewers was quantified using percent positive agreement, percent negative agreement, and kappa statistics for each of the seven categories and subcategories: education level, occupation, unemployment, retirement, uninsured, Medicaid, and homelessness. The manual review of 50 records was then compared to the algorithm results for each of the seven categories and subcategories. Sensitivity, specificity, and positive predictive value were estimated. The chi-square statistic was used to determine if the algorithms performed differently across racial/ethnic populations.

Results

Population characteristics

Among the total study population (N=9,977), at least one type of socioeconomic status information was extracted from 8,282 (83.0%) individuals. Additionally, education information for 3,780 individuals and occupation information for 7,296 individuals (Table 3) was also extracted. For the remaining categories, it was determined whether an individual was unemployed, retired, uninsured, on Medicaid, or homeless at any point in his or her record. Of the total population for which socioeconomic status data (n=8,282) was extracted, 1,978 individuals were unemployed, 1,742 individuals were retired, 1,839 individuals were uninsured, 1,865 were on Medicaid, and 318 were homeless at least one time in their EHR (Table 3). For each of the seven categories, the algorithms returned socioeconomic status information for a higher percentage of black patients than Hispanic or Asian patients ($p<0.00001$).

The five most frequently extracted occupations among those having occupation information (n=7,296) were manager, nurse, Army, manufacturer, and restaurant employee. Within the population with education information (n=3,780), the vast majority of individuals had a high school degree (n=2,066), followed by individuals without a high school degree (n=492), and individuals with a bachelor's degree (n=446).

Algorithm Performance

Prior to evaluating algorithm performance, the manual review results from the randomly selected records of 50 patients were compared between the two reviewers and any conflicts were resolved. The percent positive agreement between reviewers ranged from 98.0% to 100.0% and the percent negative agreement ranged from 94.7% to 100.0%. The Kappa statistic between reviewers ranged from 0.94 to 1.0.

Once all reviewer discrepancies were resolved, the manual review results were used as the gold standard and compared to the algorithm results. All the algorithms, with the exception of occupation, had high specificity levels >78%. The lower specificity for occupation (40%) is due to six of the ten individuals who did not have occupation information (as identified by manual review) but were identified as having occupation information by the algorithm. All the algorithms had high sensitivity levels (above 70%), with the exception of education level (66.7%) (Table 4). The lower sensitivity for education is driven by eight individuals who have an education level that was identified by manual review but not by the algorithm. The lower sensitivity for unemployment is due to the six individuals who were identified as unemployed by manual review but not by the algorithm. PPV values across the algorithms ranged from 23.1%-87.5%. The lower PPV for the retirement algorithm (63.6%) is due to the four individuals identified as retired by the algorithm but not retired by the manual review (Table 4). The low PPV for the uninsured algorithm (23.1%) is due to the ten individuals who were identified as uninsured by the algorithm, but not by manual review. The low PPV for homelessness (33.3%) was a result of the fact that the manual review only identified one patient with homelessness in his or her record, whereas the algorithm misidentified two others.

Missing data

Of the total population (n=9,977), the algorithm was not able to extract any socioeconomic status information for 1,695 individuals (17.0%). Of this group, there were 1,193 blacks, 309 Hispanics, and 193 Asians. Missing socioeconomic status data were more common among Hispanic and Asian individuals than among black individuals (p<0.001). The Hispanic and Asian populations represent 10.5% and 8.5% of the total dataset, respectively; however, these groups represent 18.2% and 11.4%, respectively, of the individuals with missing socioeconomic status data. Males represent 35.8% of the study population and 28.0% of those without extracted socioeconomic status data. The mean age for the total population is 49.9 years, and the mean age for the group without extracted socioeconomic status information is 46.7 years.

Discussion

Socioeconomic status is considered a fundamental cause of disease, because it affects so many proximate risk factors and disease outcomes [73]. It has been consistently associated with health outcomes such as mortality, cancer, and cardiovascular disease [74; 75]. Despite these consistent associations, socioeconomic status data are typically not included in genetic studies of health outcomes. For studies that utilize biobanks, the lack of socioeconomic status data is likely related to the difficulty in accessing these data within the EHR, where they are not usually recorded in structured fields. The algorithms described in this study are the first to extract these important data from EHRs for research purposes.

The socioeconomic status algorithms described here focus on the extraction of data related to four semantic categories: occupation, education, insurance status, and homelessness. The occupation algorithms extracted and classified data as occupational prestige, unemployment (ever/never), and retirement (ever/never). The occupational prestige algorithm had a strong sensitivity and PPV; however, it had a low specificity of 40% reflective of the difficulty in filtering the occupation information. Although steps were taken to remove false positives, it was difficult to completely eliminate all false positives without removing a large amount of accurate data. The unemployment and retirement algorithms had high sensitivity (70% and 100%) and specificity (93.3% and 90.7%). The unemployment algorithm had the highest PPV and the uninsured algorithm had the lowest PPV. Both unemployment and retirement were classified as ever/never because the EHR only captures a snapshot of time when the patient visits the clinic. It was not possible to accurately capture the length of time for unemployment or retirement as the patient’s visits to the clinic may not reflect the length of time he or she was unemployed or retired. The sensitivity of the unemployment algorithm was affected by the varying language used to describe unemployment, which was identified in manual review but not consistently recognized by the algorithm (“does not work outside the home”, “used to work in a restaurant”). The quality of the retirement algorithm was affected by false positives related to the identification of words related to retirement that were used in a context outside of the patient’s retirement from an occupation.

The education algorithm identified the highest level of education that a patient achieved over the course of their EHR. This algorithm had a high specificity and PPV, but a low sensitivity. The low sensitivity was due to the inability of the algorithm to detect variations in education level compared with the manual review. The variation in language used by clinical providers made it difficult to include every mention of education while still maintaining some level of precision. For example, some of the Vanderbilt Medical School students were excluded (“medical student,” “pharmacy student”) because of the frequent mention of these terms in the EHR related to patient care, rather than education level. The reviewers were able to infer education level based on occupation and context clues as well as identify the medical school students, while the algorithm was not able to do so. The algorithm that identified patients who were uninsured at some point in his or her record as well as the homelessness algorithm each had high sensitivity and specificity, but low PPV. Uninsured patients are the smallest proportion of patients within VUMC, making up only 4.7% of the patient population in 2015The low PPV of these algorithms may influenced by a low prevalence of uninsured patients and homeless individuals within the VUMC patient population. Within the randomly selected minority patient population used for evaluation, only four individuals were uninsured and one was homeless. These categories had the lowest prevalence within our evaluation dataset. The Medicaid algorithm was one of the highest performing algorithms, with a high sensitivity, specificity, and PPV.

The major challenges in utilizing EHR data in a research setting include missing data and the inconsistencies in the recording of socioeconomic status data by clinical providers. While the majority of individuals within the study population had some socioeconomic status information, a notable percentage of individuals did not have any socioeconomic status information within their records (17.0%). The missing socioeconomic status data could be a result of the lack of recording of information by the provider, either due to socioeconomic status factors not being discussed in conversation with the patient, a low number of visits in the patient’s EHR, or the willingness of the patient to provide socioeconomic status information. Additionally, when variables are missing within a patient’s record, it cannot be distinguished whether that patient truly is negative for the socioeconomic status information or just missing data. For example, if a patient does not have an occupation listed, it cannot be assumed that they are unemployed because it may have not been discussed with the provider or recorded by the provider. As a result, true negatives and false negatives cannot be identified. The higher level of missing data observed for Hispanic and Asian individuals in this dataset could be a reflection of the fact that the algorithms are optimized for the largest racial/ethnic population within the dataset (i.e., black patients).

The inconsistencies in the recording of the socioeconomic status data are typical for social and environmental exposure data contained within free clinical text [65]. In the development of these algorithms, it was noted that providers, in general, do not follow patterns when recording socioeconomic status data within their notes in the EHR. The lack of consistent language and the numerous variations used to describe the socioeconomic status information made extracting this information challenging. Furthermore, algorithms could also be limited by the accuracy of the selected filters and tags, rather than the information available within the EHR. While the aim of the algorithms was to include all possible semantic tags, there is a possibility that some information was missed by the algorithms or that information was captured inaccurately due to the limitations of the filtering process.

In addition to these general limitations, the algorithms developed here have specific limitations regarding portability. Even within the same dataset, a difference in tag retrieval for the socioeconomic status categories queried across the three major racial/ethnic groups has been noted. Additional studies are required to improve the algorithms' performances and retrieval of semantic tags in multiple populations as well as within different study sites. Indeed, some of the tags developed here (such as "TennCare" in reference to Medicaid) are specific to Tennessee and will require modification to ensure portability regardless of the state in which the algorithms are deployed. Furthermore, these algorithms were created in a de-identified EHR, which required the development of a free text algorithm for insurance status, as the structured insurance information is considered identifying information. An identified EHR may have this insurance information within the structured text. However, the other categories of socioeconomic status information are likely to only be found within the free text of an identified EHR.

Despite the many challenges faced with the extraction of socioeconomic status data from the EHR, these algorithms were able to successfully extract a large amount of data not previously accessible for research purposes. The sensitivities, specificities, and PPVs for the algorithms were high considering the limitations of the socioeconomic status data within the current EHR. Overall, these algorithms represent a first important step in incorporating socioeconomic status data from EHRs into precision medicine research, as envisioned by the Institute of Medicine and others.

Resources

Semantic tag and filter lists for each algorithm can be found on the Vanderbilt University Medical Center TREAT Lung Cancer Research Program website (https://medschool.vanderbilt.edu/treat-lung-cancer-program/) and the Institute for Computational Biology website (http://www.icompbio.net/?page_id=1654).

Characteristic	**N=9,977**
Sex	
Male	3,568 (36%)
Female	6,409 (64%)
Race/ethnicity	
Black	8,078 (81%)
Hispanic	1,049 (10.5%)
Asian	850 (8.5%)
Age (mean, years ± SD)	49.8 ± 18.1
Number of clinic visits (mean ± SD)	40.5 ± 55.0
Number of days between visits (mean ± SD)	2,340 ± 1,793.1

Table 1

Table 1. Vanderbilt BioVU racial/ethnic minority population characteristics as of 2015.

Semantic category	**Format of algorithm output**
Occupational prestige	0-100
Unemployment	Ever/never
Retirement	Ever/never
Education	-Never attended -Less than high school -High school graduate/GED -Associate’s degree -Bachelor’s degree -Master’s degree -Professional degree -Doctoral degree
Uninsured	Ever/never
Medicaid	Ever/ never
Homelessness	Ever/never

Table 2

Table 2. Variables extracted by socioeconomic status algorithms applied to de-identified electronic health records.

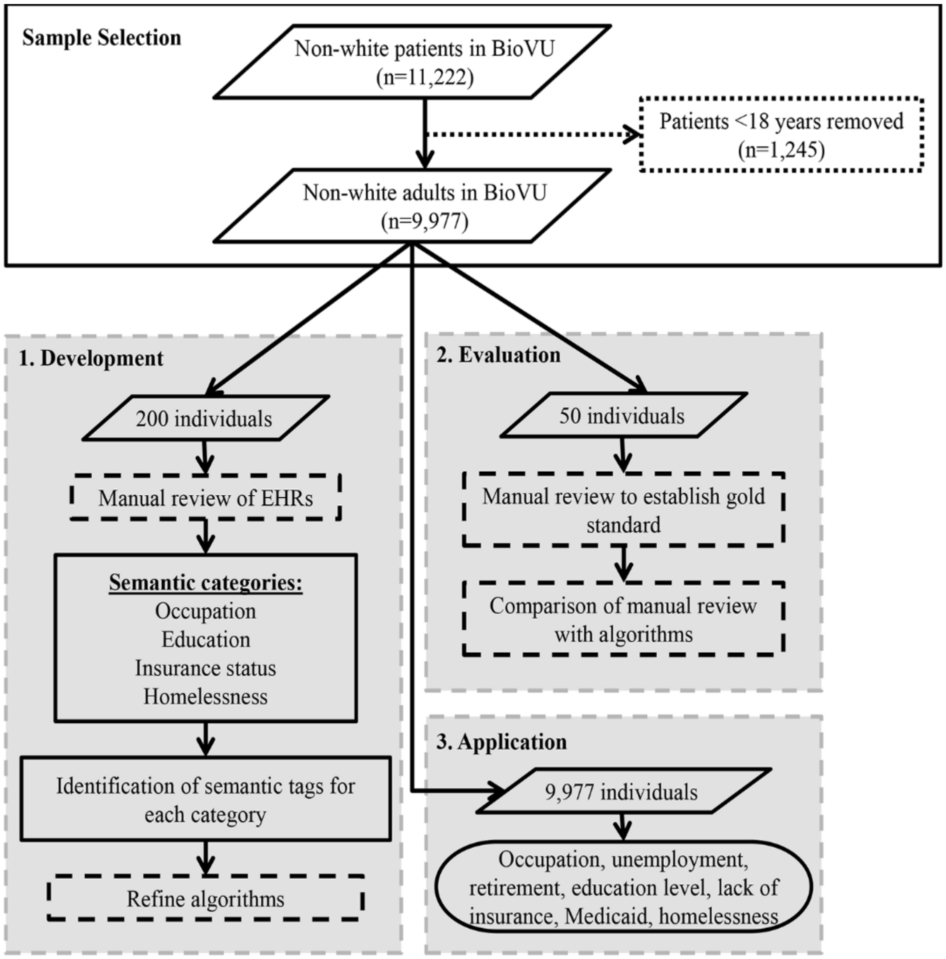
Sample Selection
Non-white patients in BioVU (n=11,222)
Patients <18 years removed (n=1,245)
Non-white adults in BioVU (n=9,977)
1. Development
200 individuals
Manual review of EHRs
Semantic categories:
Occupation
Education
Insurance status
Homelessness
Identification of semantic tags for each category
Refine algorithms
2. Evaluation
50 individuals
Manual review to establish gold standard
Comparison of manual review with algorithms
3. Application
9,977 individuals
Occupation, unemployment, retirement, education level, lack of insurance, Medicaid, homelessness

Figure 12

Figure 12. Overview of the development process for the socioeconomic status algorithms.

The creation of the socioeconomic status algorithms took place over three steps: development, evaluation, and application. Development involved the identification of categories and tags, followed by refinement. Evaluation involved the comparison of manual review results to algorithm results in order to determine a sensitivity and specificity for each algorithm. Application involved applying all of the algorithms to the full dataset of individuals.

Characteristics	**Race**			
	Black (n=8,078)	Hispanic (n=1,049)	Asian (n=850)	Total (n=9,977)
% with occupation	76.0	57.1	65.4	73.1
% unemployed	21.4	13.0	13.4	19.8
% retired	19.8	4.9	11.2	17.5
% with education	39.1	28.7	37.9	37.9
% uninsured	19.5	15.6	11.5	18.4
% on Medicaid	20.5	13.9	7.9	18.7
% homeless	3.7	1.3	1.0	3.2

Table 3

Table 3. Percent of records within the study population with algorithm-identified socioeconomic status characteristics.

These values represent the percent of individuals within each group that had algorithm identified socioeconomic status variables. For example, the individuals who had a term for Medicaid within their record are listed as part of the percentage in this table.

Semantic Category	Records with SES information (%)	Sensitivity (%)	Specificity (%)	PPV (%)
Education level	48.0	66.7	84.5	80.0
Occupation	80.0	87.5	40.0	85.4
Unemployment	40.0	70.0	93.3	87.5
Retirement	14.0	100.0	90.7	63.6
Uninsured	8.00	75.0	78.3	23.1
Medicaid	18.0	100.0	95.1	81.8
Homelessness	2.00	100.0	95.9	33.3

Table 4

Table 4. Comparison of manual review with algorithm results for each socioeconomic status algorithm in a subset of randomly selected individuals (n=50).

This table shows the percent of the 50 records that contained each type of socioeconomic status information, as well as the sensitivity, specificity, and positive predictive values calculated by analyzing the comparison of manual review results (gold standard) to algorithm results for the 50 randomly selected records.

CHAPTER THREE

GENE X EDUCATION INTERACTION: BLOOD PRESSURE IN BLACK ADULTS

Introduction

Blood pressure in black individuals

As reviewed in Chapter One, black Americans have a higher burden of hypertension than other racial/ethnic groups. Despite the higher burden of hypertension in black populations, there is limited knowledge about the genetic variants contributing to the estimated heritability. Several large-scale genetic studies have been done, but there is still much to be known about the genetic component of hypertension and blood pressure in black populations [29; 44-49]. These studies have focused on utilizing genome-wide common variants and examining hypertension, systolic blood pressure, diastolic blood pressure, or pulse pressure. They have also included meta-analyses, with the goal of a large population in order to examine smaller effect sizes. To date, the large-scale genetic studies of blood pressure and related outcomes only account for a maximum of 25% of the estimated heritability of blood pressure, which is up to 70%. While some of the SNPs in these large scale studies have been confirmed to contribute to the estimated heritability of blood pressure, they still do not explain the total estimated heritability, resulting in a mystery of "missing heritability" [51]. The "missing" heritability estimates may be explained, in part, by other factors, such as socioeconomic status, that interact with genetic variation and contribute to the variance observed in blood pressure.

Blood pressure and education

Socioeconomic status is usually defined as some combination of education, income, and occupation [13]. Low socioeconomic status is strongly associated with hypertension and related cardiovascular comorbidities and mortality [13; 52]. However, among the GWA studies conducted to date, only one study has included any measurements of the social environment, which was in the form of education [58]. This neglect of socioeconomic status continues among genetic studies, despite the fact that numerous epidemiological studies have found that social environment, specifically socioeconomic status, has a strong influence on blood pressure and hypertension [13; 52; 59]. Education and income gradients have both shown to be significantly inversely correlated with markers of cardiovascular disease risk, including hypertension, such that people with lower education and lower income have a higher risk of hypertension [13]. Education is one method of measuring socioeconomic status that is stable in adults, due to most individuals achieving their highest level of education early in life, and it is a reflection of long term earning potential as well as social status [76].

Electronic health record data

BioVU

BioVU is a DNA biobank of the Vanderbilt University Medical Center (VUMC) linked to de-identified EHRs. DNA samples are extracted from discarded blood samples drawn for routine clinical care [66]. DNA samples are linked to the Synthetic Derivative (SD), the de-identified version of the VUMC EHR, by a unique study ID. Medical records within the SD are scrubbed of all HIPAA identifiers such as names, locations, zip codes, and social security numbers. Dates within each SD record are shifted to prevent re-identification of the records. Date shifting is consistent within a single patient's record. As previously described [67], data from BioVU are de-identified in accordance with provisions of Title 45, Code of Federal Regulations, part 46 (45 CFT 46); consequently, this study is considered non-human subjects research by the Vanderbilt University Institutional Review Board.

Electronic health record blood pressure data

For this study, the choice was made to focus on measurements of blood pressure rather than classifying patients into cases or controls based on hypertension status for three reasons. The first is that utilizing continuous measurements as an outcome is more statistically powerful than a dichotomous outcome. The second is that measuring blood pressure is potentially closer to outcome that is more directly impacted by genetic variation[77]. The third is that defining hypertension status based on data within the electronic health record is challenging due to the potential for inaccurate or missing information. For example, when defining hypertension cases and controls in the electronic health record, ICD-9 codes, medication use, blood pressure measurements, and clinical notes would be used. However, ICD-9 codes are not always an accurate indication of a person's status. Individuals who have the ICD-9 code may not always be hypertensive. Additionally, defining cases and controls based on medication use may not always be accurate either due to the use of medications for multiple conditions, as well as hypertensive patients in the Synthetic Derivative who may not have their medications in their record. Finally, it is very challenging to extract usable and accurate data from the free text clinical notes, even if a patient's hypertension status is within the notes. Due to these challenges, classifying hypertension cases and controls would be prone to inaccuracies. For these reasons, blood pressure measurements are employed, rather than classifying patients into cases or controls for hypertension.

Electronic health record and socioeconomic status data

Socioeconomic status is considered a fundamental cause of disease, because it affects so many proximate risk factors and disease outcomes [73]. Despite these consistent associations, socioeconomic status data are typically not included in genetic studies of health outcomes. For studies that utilize biobanks, the lack of socioeconomic status data is likely related to the difficulty in accessing these data within the EHR, where they are not usually recorded in structured fields. The algorithms described in our study are the first to extract these important data from EHRs for research purposes [78].

The socioeconomic status algorithms described previously focus on the extraction of data related to four semantic categories: occupation, education, insurance status, and homelessness. The occupation algorithms extracted and classified data as occupational prestige, unemployment (ever/never), and retirement (ever/never). The education algorithm identified the highest level of education that a patient achieved over the course of their EHR. Uninsured patients, patients on Medicaid, and patients who experienced homelessness were described as ever or never in the algorithm.

The major challenges in utilizing EHR data in a research setting include missing data and the inconsistencies in the recording of socioeconomic status data by clinical providers. The missing socioeconomic status data could be a result of the lack of recording of information by the provider, either due to socioeconomic status factors not being discussed in conversation with the patient, a low number of visits in the patient's EHR, or the willingness of the patient to provide socioeconomic status information. The inconsistencies in the recording of the socioeconomic status data are typical for social and environmental exposure data contained within free clinical text [65]. Additional information regarding the limitations of our socioeconomic status algorithms were previously discussed in Chapter Two[78].

While there are many measurements of socioeconomic status, and several were extracted from Vanderbilt's Synthetic Derivative, the choice was made to focus on the measurement of education for the analyses. Based on the algorithm, it had a comparatively reasonable sensitivity (67%) and specificity (85%), while also representing one of the more stable measurements of socioeconomic status and a good reflection of earning potential [76]. The other measurements of socioeconomic status that we extracted from the EHR included occupation, retirement, unemployment, homelessness, and Medicaid use. While these measurements are helpful, they are more likely to be transient and change over time. For these reasons, the decision was made to focus on the relationship between education as a measurement of socioeconomic status and genetic variants contributing to blood pressure in black individuals.

Summary

Due to previously reported interactions between education levels and genetic variants associated with blood pressure in a white population[58], it is expected that interactions between education and genetic variants associated with blood pressure may also occur in a black population. As education is a component of socioeconomic status, which is known to be associated with health outcomes, we expect to be that the social environment may be interacting with genetic variants to affect blood pressure in black individuals. The inclusion of EHR-derived education in a large-scale genetic analysis of blood pressure in a black population is novel and could lead to a better understanding of the biology of blood pressure. Additionally, this study will lay the groundwork for additional investigations into gene-socioeconomic status interactions using EHR populations.

Methods

Population

The study population is a subset (n=2,577) of black adults >18 years old participating in BioVU as of 2011 [68]. The original population included all non-white patients in BioVU with DNA samples as of 2011 (N=15,863). Black adults who had passed the quality control procedures for the outcome and covariates were selected for investigation. Race/ethnicity is administratively reported in BioVU and strongly correlated with genetic ancestry [70; 71]. Individuals included within the analysis subset had available sex, age, smoking status, percent African ancestry, BMI, education, and pre-medication blood pressure values. Sex and age were extracted from provider-recorded values in the record. Smoking status was extracted using ICD-9 tobacco use codes [79]. Percent African ancestry was calculated using Metabochip genotype data which passed quality control and ADMIXTURE in an unsupervised analysis [80]. BMI was calculated by taking the median weight across all of the values in an individual's record and their height.

Blood pressure measurements used for each individual were the median of the values from all blood pressure measurement found within an individual's record prior to a recording of blood pressure-altering medications in the medication list. The medications included in the list of anti-hypertensives are ACEI/ARB, beta blockers, non-dihydropyridine CCBs, hydralazine, Minoxidil, central alpha antagonists, direct renin antagonists, aldosterone antagonists, alpha antagonists, and diuretics including thiazides, K-sparing, and loop diuretics. Any blood pressure measurement found after any of these medications were mentioned in the medication list were excluded from the blood pressure calculation.

Education was extracted using the algorithm described in Chapter 2. Education was examined in numerous ways including the eight-tier algorithm extracted variable, the dichotomous completion of high school or completion of college variable, and a three-tier variable which grouped individuals into those that had not completed high school, those who completed high school, and those with some college or above. The three-tier variable was chosen because it maintained some level of precision, while enabling a larger number of individuals within each group.

Genotyping

Genotyping was performed on the Metabochip, a custom Illumina genotyping chip which targets SNPs associated with metabolic traits and cardiovascular disease [81; 82]. The array includes 2,207 SNPs from the NHGRI GWAS catalog as of August 1, 2009. For each of the GWAS identified SNPs, SNPs with an $r^2 > 0.90$ in the CEU HapMap II population and up to four additional SNPs with an $r^2 > 0.50$ in the YRI HapMap II population were included on the array. The array also includes fine-mapping for SNPs of interest to the consortia which contributed to the development of the chip, X and Y chromosome SNPs, mitochondrial SNPs, and “wildcard” SNPs for a total of approximately 200,000 SNPs. After the removal of SNPs with a minor allele frequency of less than 5.0%, SNPs with a Hardy-Weinberg Equilibrium exact test p-value of less than 1×10^{-7}, and SNPs with a genotyping call rate of less than 95%, a total of 115,834 variants remained (Figure 13). All genotyping analyses were carried out in plink1.9 [83] or R [84]. African ancestry estimates were calculated using Metabochip data which passed quality control procedures in an unsupervised ADMIXTURE analysis, with only the individuals included in analysis, who were all identified as black in BioVU [80].

Regression models

Linear regression models were used to investigate the relationship between the genetic variants and both pre-medication systolic blood pressure and pre-medication diastolic blood pressure. The first model did not include any education information:

$$\text{Premedication systolic or diastolic blood pressure} = \beta_0 + \beta_{cov}{*}X_{cov} + \beta_1{*}SNP + e$$

The covariates in the model included age, age squared, sex, BMI, smoking status, and percent African ancestry. The second model included education as a categorical variable, with no education coded as 0, less than high school as 1, high school as 2, GED as 3, some college or associates degree as 4, bachelor degree as 5, master degree as 6, medical or law degree as 7, and PhD as 8. These classifications were recoded into three categories: 0 as less than high school, 1 as high school degree, and 2 as some college and higher, in order to better represent the earning potential as higher levels of education are associated with higher socioeconomic status.

In order to examine the interaction between genetic variants and education and how it may affect blood pressure, several models were conducted. The first model included education as a covariate and the SNP x education interaction term:

$$\text{Premedication SBP or DBP} = \beta_0 + \beta_{cov}*X_{cov} + \beta_1*\text{SNP} + \beta_2*\text{Education} + \beta_3*\text{SNP}*\text{Education} + e$$

The decision was made to focus on a set of SNPs which had a p-value of less than 1.4×10^{-5} from the main effects model in order to reduce issues with multiple testing. This cutoff was chosen based on a Bonferroni correction for the number of SNPs that would remain if SNPs with an r^2 value of greater than 0.1 were removed from our dataset. For this set of SNPs, the model which included the main effect of education as well as the interaction term was utilized. The p-value level for significance was based on the number of SNPs tested for premedication systolic blood pressure and premedication diastolic blood pressure.

Results

Population characteristics

The population was selected from a previously genotyped BioVU population [68]. This original population included all non-white individuals in BioVU as of 2011. For the study population, black individuals were the point of focus (n=11,301). During the quality control process, 967 individuals were removed for either ambiguous sex, missing genotypes (>5.0%), or relatedness (twins, full siblings, parent/offspring) (Figure 13). After individuals were removed during genotype-based filtering, individuals were removed based on covariate data. Once individuals with missing data on education, premedication systolic blood pressure, premedication diastolic blood pressure, body mass index, and individuals under 18 were removed, a population of 2,577 individuals remained (Figure 14). This population was mostly female (71% and mostly never-smokers (87%), with a median age of 38 years, a median African ancestry percentage of 81.7%, a median BMI of 26.8 kg/m^2, a median premedication systolic blood pressure of 122 mmHg, and a median premedication diastolic blood pressure of 74 mmHg (Table 5).

Individuals were represented in every level of education, with the majority of the individuals included in analyses having a high school degree (Figure 15). Within analyses, individuals were grouped into one of three categories of education: less than a high school degree (n=328), a high school degree or GED (n=1,518), or some college and above (n=731). Minor differences between educational groups in terms of premedication systolic blood pressure, premedication diastolic blood pressure, and body mass index were observed (Figure 16, Figure 17, Figure 18). These differences were not statistically significant, with the exception of an association between education and premedication diastolic blood pressure, where premedication diastolic blood pressure increases slightly with increasing education level. A trend within the education groups by age was observed: age steadily increased with higher levels of education (Figure 19).

When blood pressure measurements were examined across age groups, an increasing trend of systolic blood pressure with age was observed (Figure 20). When diastolic blood pressure across age groups was examined, an increase in diastolic blood pressure until around age 60 was noted, then diastolic blood pressure decreased (Figure 21). Prior to examining genetic data, the correlation of covariates with blood pressure measurements was analyzed (Table 6). Age, sex, and body mass index were significantly correlated with both premedication systolic and diastolic blood pressure in the dataset. Age and premedication diastolic blood pressure significantly covaried with education (Table 7).

Individuals excluded from analyses

The characteristics of the individuals excluded from the analyses because of missing education values were assessed. When comparing the populations included in the analyses with the population of individuals without education values, it was observed that the differences in sex, age, smoking status, and premedication blood pressure measurements were statistically different (Table 8). However, the range of differences was minimal, with the exception of age. Those included in analyses had a median age of 38 years, while those without education data had a median age of 57 years. This is likely due to a bias in provider recording; individuals who are younger and in school may be more likely to be asked by a provider about their education because they do not have an occupation. When the blood pressure measurements across age groups for individuals without education were examined, a similar pattern to what was observed in our individuals included in analyses was noticed. Systolic blood pressure increased with increasing age, while diastolic blood pressure increased until about 60 years of age, and then it began to decrease (Figure 22 and Figure 23). The variation in blood pressure of those excluded from analyses to those include in the analyses was similar.

Predictors of systolic and diastolic blood pressure

The initial models examined both premedication SBP and DBP without the inclusion of any education measurements. The following linear model was used:

$$\text{Premedication BP} = \beta_0 + \beta_{cov}*X_{cov} + \beta_1*\text{SNP} + e$$

Where X_{cov} refers to age, sex, BMI, smoking status, and percent African ancestry. The initial model did not include an age squared term. The second model included age squared in the list of covariates, which was included based on prior studies that found age squared to be significantly associated with BP[46; 58]. The Manhattan plots of these results are shown in Figure 24 and Figure 25. For systolic blood pressure, one SNP passed a Bonferroni correction: rs4593967. For diastolic blood pressure, one SNP passed a Bonferroni correction: rs950928.

Impact of education on predictors of systolic and diastolic blood pressure

The second set of models included a measure of education as a predictor in the model. Individuals were placed into three categories based on highest education level achieved in their health record: less than high school, high school, and some college or higher. This model was similar to the early model, with the exception of the addition of the education term.

$$\text{Premedication BP} = \beta_0 + \beta_{cov}*X_{cov} + \beta_1*\text{SNP} + \beta_2*\text{Education} + e$$

As with the earlier model, age, sex, BMI, smoking status and percent African ancestry were included as covariates, with age squared added as an additional covariate for a second set of models (Figure 26 and Figure 27).

In the results for systolic blood pressure, one SNP at chromosome 10 passed a Bonferroni correction (4.32×10^{-7}), with two other SNPs passing a suggestive correction line, based on a Bonferroni correction if SNPs with an r^2 value of higher than 0.6 were removed (7.24×10^{-6}). For diastolic blood pressure, a peak at chromosome 16 was seen, with two SNP barely passing a Bonferroni correction and another passing the suggestive correction. However, these SNPs have the same effect size and are in perfect linkage disequilibrium, so they are representing the same locus. The SNPs are shown in Table 9. The addition of education to the model did not change the most significantly associated SNPs.

Gene-environment interaction models

The initial examination of the interactions between SNPs and education included all SNPs across the dataset, utilizing the model:

$$\text{Premedication BP} = \beta_0 + \beta_{cov}*X_{cov} + \beta_1*\text{SNP} + \beta_2*\text{Education} + \beta_3*\text{SNP}*\text{Education} + e$$

In this model, age squared was not included as covariate. The Manhattan plot of the interaction term p-values can be seen in Figure 28 and Figure 29. There were no interactions which passed a Bonferroni correction. In a second model, I examined interactions with SNPs that passed a threshold of suggestive significance in the main effects model. This lowered the multiple testing burden, which is incredibly high when examining gene-environment interactions across a massive dataset. When selecting the suggestive SNPs, a p-value threshold was chosen based on the Bonferroni correction for the number of SNPs from the dataset with an r^2 value of less than 0.1 (n=36,762). This is a lenient cutoff point to allow a larger number of SNPs to test for interactions. SNPs with a p-value of less than 1.36 x 10^{-6} were included in the interaction analyses. The model was similar to the interaction model stated earlier, with the addition of age squared as a covariate. Table 10 and Table 11 show the SNPs selected for interaction testing and the p-value results for the interaction term. No statistically significant interactions between our education variable and the selected SNPs were observed.

Discussion

The aim of this section is to determine if education interacted with genetic variants to affect blood pressure in a black population, as was expected due to previous gene x education interactions associated with blood pressure that were observed in a white population. Associations between premedication systolic blood pressure or premedication diastolic blood pressure and genetic variants from the Metabochip were examined, while including known predictors of blood pressure (age, BMI, sex, percent African ancestry, and smoking status) in the model. These models were compared with models which included a main effect of education, and a main effect of education plus an interaction between genetic variants and education in order to determine if education affected the associations between genetic variants and blood pressure in a black population. While some significant novel associations were observed between genetic variants and blood pressure, these associations were not greatly affected by the addition of education information. Additionally, no significant gene x education interactions were observed.

Models without interaction

Premedication systolic blood pressure

The SNP rs4593967, which was found to be significantly associated with systolic blood pressure, has not previously been associated with blood pressure or hypertension. It is found within intron 3 of *ARHGAP22*. *ARHGAP22* has been associated with diabetic retinopathy, conduct disorder, daytime sleep, and self-employment [85-88], but these associations have not been replicated. The minor allele frequency in the 1000 genomes African populations is 18%. rs4593967 is in linkage disequilibrium with other intronic variants of ARHGAP22 and therefore may be tagging one of those variants. ARHGAP22 is a regulator of a RhoGTPase. The effect estimate of one minor allele of this variant is a decrease of 2.53 mmHg with a standard error of 0.48.

The SNP rs10921895, which passed a suggestive significance threshold in our systolic blood pressure association test, is found in an intergenic region on chromosome 1. The minor allele frequency in 1000 Genomes African populations is 35%. This SNP has not been associated with any other phenotypes, but it is found in a region with an H3K27Ac mark in K562 cells, which are derived from bone marrow. This indicates that this region may be involved in some type of gene regulation in bone marrow cells. The variant is found to be in linkage disequilibrium (within the 1000 Genomes African Americans of the Southwest) with other intergenic variants. The effect estimate of one copy of the minor allele is a decrease of 1.55 mmHg with a standard error of 0.36.

The final SNP from our systolic blood pressure analysis, which was suggestively significant, was rs3804485, which is found on chromosome 6. It has a 34% minor allele frequency in 1000 Genomes African populations and is found within an intron of LY86. While rs3804485 has not been previously associated with any phenotype, LY86 is a lymphocyte antigen that has been associated with coagulation, waist-to-hip ratio, depression, gastritis, response to radiotherapy in cancer, urate levels, diabetic kidney disease, and anxiety [89-98]. This SNP does not appear to be in high linkage disequilibrium with anything in the African Americans in the Southwest 1000 Genomes population. The effect estimate of one copy of the minor allele is an increase of 1.51 mmHg with a standard error of 0.33.

Premedication diastolic blood pressure

Our examination of diastolic blood pressure revealed a peak on chromosome 16, with rs950928 passing a Bonferroni correction, and rs8056711 passing suggestive significance. Both of these SNPs fall within introns of IQCK, a gene that is involved as an EF hand protein binding site. While, neither of these SNPs have been previously associated with any phenotypes, IQCK has previously been associated with blood pressure, body mass index, bone density, heart rate, chronic obstructive pulmonary disease, bipolar disorder, and a body mass index-education interaction [99-103]. Both SNPs have a minor allele frequency of 40% within the 1000 Genomes African populations and a decrease of 1.10 mmHg with the presence of one copy of the minor allele, with a standard error of 0.22.

Power

While the population size of our dataset is somewhat small (n=2,577), there was enough power to detect some significant variants. As shown in Figure 30, the study was powered to detect more common variants with moderate effect sizes. For an effect size (β) of 1.0, the study was at 80% for minor alleles with a frequency above 20%. For less common alleles, with a minor allele frequency between 10% and 15%, the study was powered to detect affect sizes of 1.5 or greater. In order to detect alleles with a minor allele frequency of 5%, an effect size of 2.0 or greater was needed. For systolic blood pressure, rs4593967 was discovered, which had a minor allele frequency of 13.91% in the dataset and an effect size of -2.53; we had 99% power to detect this association. The study was also at 99% power to detect rs10921895 (minor allele frequency of 37.14%, effect size of 1.55) and rs3804485 (minor allele frequency of 41.28% and effect size of 1.51). For diastolic blood pressure, the study was at 96.7% power for both rs950928 and rs8056711; both had a minor allele frequency of 36.35% and an effect size of -1.10.

Interactions

Due to previous gene x education interactions associated with blood pressure in a white population, it was anticipated that gene x education interactions may exist within a black population. As known blood pressure associated SNPs do not explain the full picture of heritability, it was expected that gene x environment interactions may be contributing to the estimated heritability of blood pressure. In light of the strong associations between socioeconomic status variables and blood pressure, we hypothesized that gene x education interactions may be associated with blood pressure, as education level is a measurement of socioeconomic status.

However, no significant interactions between the SNPs tested and the education variable were discerned. This result may be explained by a number of reasons. The first is that the null hypothesis is supported and interactions between education variables and the SNPs investigated do not exist. If instead type II error was present, it is possible that the

SNPs that do interact with education to affect blood pressure were not tested or that the education variable was not accurate enough, or a strong enough proxy of SES, to be able to capture interactions. The use of the algorithm-extracted education variable may have limited the ability to detect associations due to its limited accuracy[78]. As stated in Chapter One, the measurements of socioeconomic status are imperfect and may not be proxies for each other. In this study, increasing education is meant to represent increasing status in society, which can represent less stress and better health outcomes. However, the association between increasing education and increasing social status and wealth may not be consistent across racial groups. Higher education in a black population (which is our population of interest) may not equate to the same social mobility as higher education in a white population[104]. Therefore, education may not be the most appropriate measurement for capturing the relationship between social environment and health in our population.

In addition to the measurement challenges, there were also challenges of statistical power. The detection of gene-environment interactions often requires a large sample size and it is possible that we did not have power to detect these interactions. Gene x education interactions were examined without the main effect of education. The method used may not be robust enough to detect an interaction without a main effect.

Limitations

The analyses had several limitations. One main limitation of the study was the small sample size. While the BioVU population seems very large as whole, the population does have a problem with missing data and limitations of what is available within the EHR. Once the population is limited to individuals of a certain race, black individuals in this study, plus individuals with complete phenotype information, the population becomes very small. There may also be unknown biases which exist due to the selection of individuals who have complete phenotype data.

In addition to the limited sample size, the population in this study was also slightly different than previous populations used to examine blood pressure in a black population. While the proportion of the two sexes, median body mass index, median systolic blood pressure, and median diastolic blood pressure were similar to previous studies[105], the population in this study did have a much lower median age, over 15 years younger. This made the study population unique and may have reduced variability in blood pressure measurements, as blood pressure increases with age.

Another limitation was the lack of a replication dataset. Without the ability to replicate the findings, there is not enough support to be able to say if the findings are real or not. This is especially true since other studies with much larger populations did not observe an association between these SNPs and blood pressure [48]. Additional limitations to the study include a relatively small sample size, especially considering the typical effects sizes that are observed for SNPs associated with blood pressure and the fact that interaction studies require large samples, and limitations on genotype data. The genotype data is from a curated genotyping chip, rather than a true genome wide chip, so it has limited SNPs. This chip was also designed to include rare variation collected from the African ancestry 1000 genomes populations and therefore, many of the variants on the Metabochip were rare in African ancestry populations [81]. Due to the limited population size, we had to remove many of the SNPs on the chip during quality control, as the study was not powered to detect rare variation.

In addition to the limitations regarding the genotype data, there were also some limitations regarding the phenotype data. All of the variables were extracted from electronic health records. While these records have extensive amounts of data, the data recorded by healthcare provider is not always accurate and the ability to extract the data can be limited. For example, the use of ICD-9 billing codes for phenotyping within the electronic health record is vital. However, these codes do not always accurately describe the patient’s medical condition. The use of medication information within the patient’s medication list is also important for phenotyping. The medication lists are based on patients telling providers which medications they are taking, therefore errors in patient statements or a lack of updating of the medication information could lead to misleading information within the record. In the case of education, where the positive predictive value of our algorithm was 80% [78], there may have been inaccurate education information for the individuals within the dataset. Therefore, it is possible that there was inaccurate education data, as well as inaccurate blood pressure and covariate data in the analyses, limiting the ability to detect true associations.

Determining the blood pressure measurements from the EHR to use within a study is challenging. Measurements are subject to inaccuracies in recording by providers, as well as missing education information. Determining which measurements to use in the study is also a challenge, as measurements can vary widely across the EHR. The median blood pressure measurements were chosen for our study in an attempt to reduce the influence of this variation. Beyond the inaccuracies and decisions to be made regarding the information within the EHR, blood pressure is difficult to measure within the clinic. Measurements of blood pressure can vary due to the calibration of instruments, the time of day it is measured, and due to illness[106]. Patients also tend to have higher blood pressure within a clinical setting due to stress[106].

Strengths

Despite the limitations within the study, there were also a number of strengths. Primarily, this is the first study to incorporate electronic healthcare record-derived education information into a large scale genetic investigation. This is also the first analysis to incorporate education information into a large scale genetic study of blood pressure in a black population. This study paves the road for the incorporation of education, as well as other socioeconomic status information into genetic studies which utilize biobank populations. Additionally, despite the lack of interaction effects, we hope that this research encourages other investigators to continue to study health outcomes with racial health disparities and incorporate socioeconomic status information.

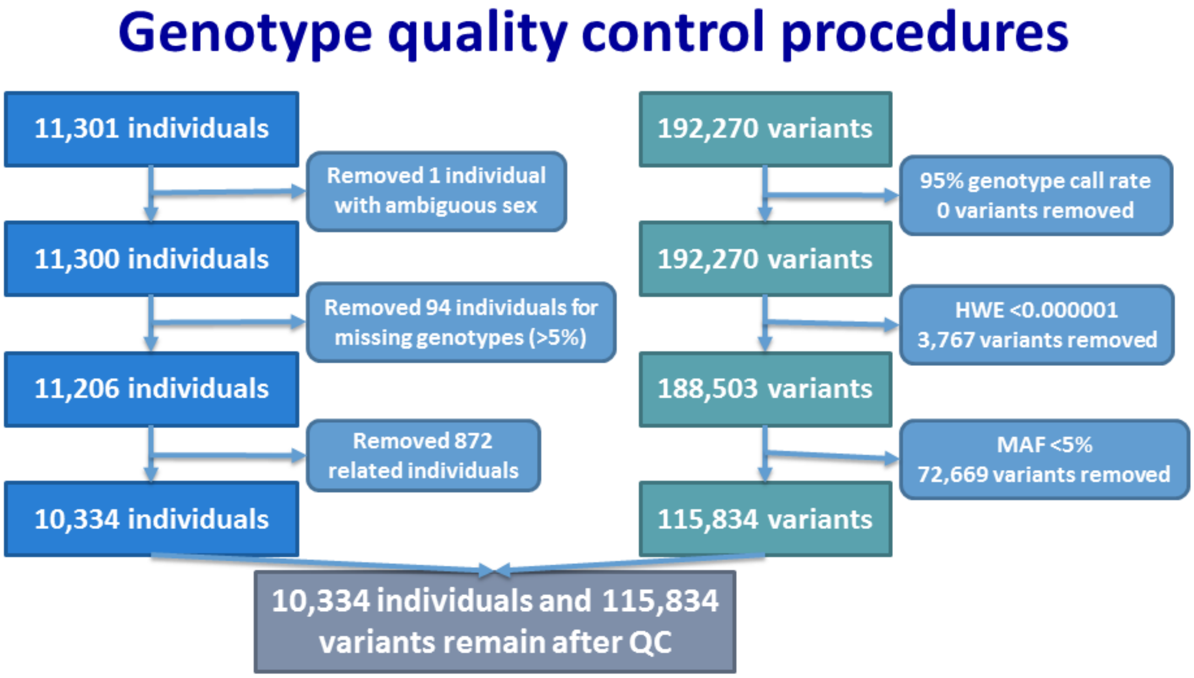
Genotype quality control procedures
11,301 individuals
Removed 1 individual with ambiguous sex
11,300 individuals
Removed 94 individuals for missing genotypes (>5%)
11,206 individuals
Removed 872 related individuals
10,334 individuals
192,270 variants
95% genotype call rate 0 variants removed
192,270 variants
HWE <0.000001 3,767 variants removed
188,503 variants
MAF <5% 72,669 variants removed
115,834 variants
10,334 individuals and 115,834 variants remain after QC

Figure 13

Genotype quality control procedures for full dataset

Individuals with ambiguous sex and greater than 5% missing genotypes were removed. One individual from each of twin pairs, parent-offspring pairs, and sibling pairs was also removed. Variants that had less than a 95% genotyping call rate, were significantly outside of Hardy-Weinberg equilibrium, or had less than a 5% minor allele frequency within our population were also removed.

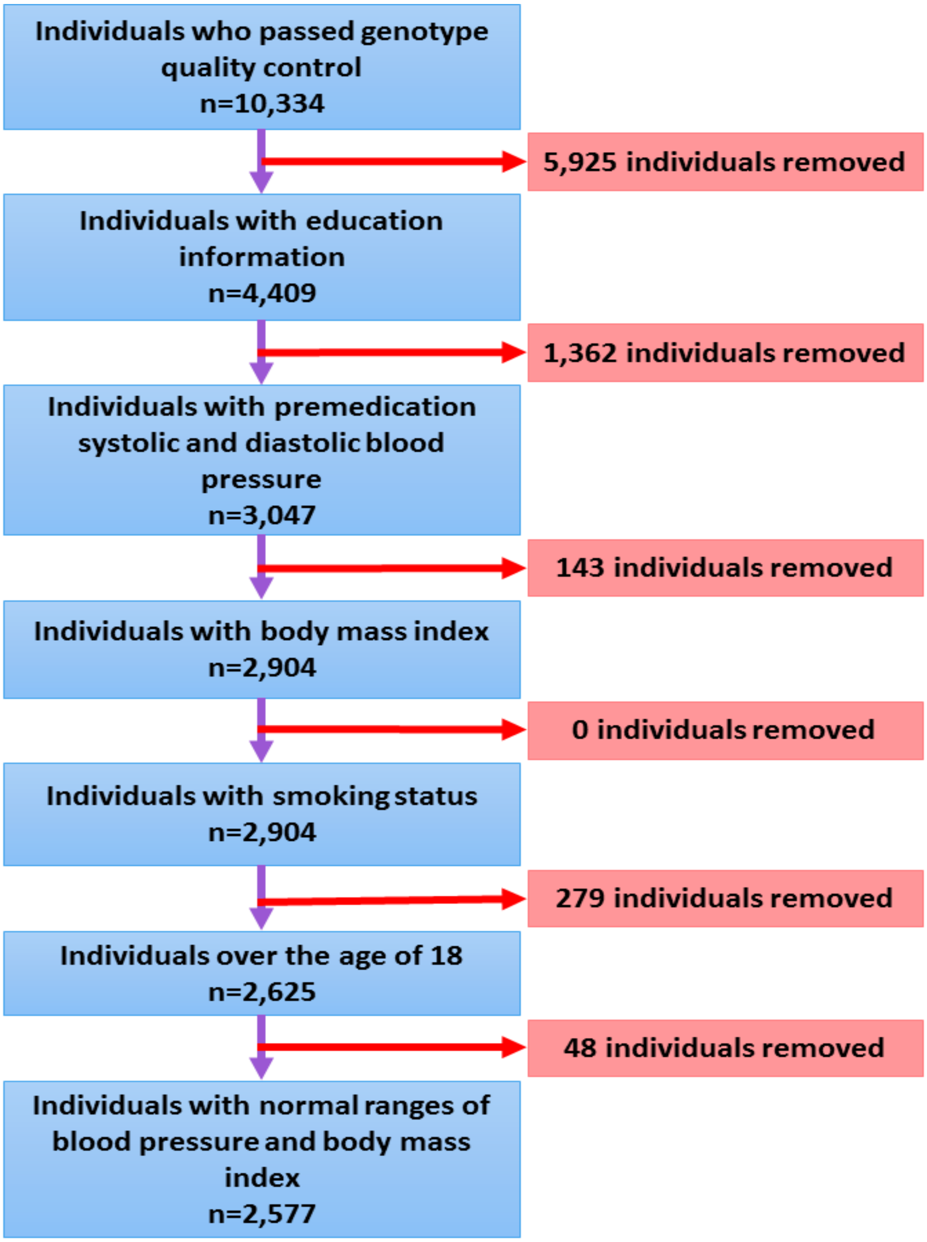
Individuals who passed genotype quality control
n=10,334
5,925 individuals removed
Individuals with education information
n=4,409
1,362 individuals removed
Individuals with premedication systolic and diastolic blood pressure
n=3,047
143 individuals removed
Individuals with body mass index
n=2,904
0 individuals removed
Individuals with smoking status
n=2,904
279 individuals removed
Individuals over the age of 18
n=2,625
48 individuals removed
Individuals with normal ranges of blood pressure and body mass index
n=2,577

Figure 14

Phenotype data quality control

Individuals with missing phenotype data were removed from the dataset. Children were removed. Individuals with blood pressure or body mass index values greater than three times the standard deviation of the mean were also removed.

Characteristic	Number of individuals n=2,577
Sex	
Male	753 (29%)
Female	1,824 (71%)
Smoking Status	
Non-smokers	2,242 (87%)
Smokers	335 (13%)
Age (median, years)	38
Percent African ancestry (median)	81.7%
Body mass index (median, kg/m^2)	26.8
Premedication systolic blood pressure (median, mmHg)	122
Premedication diastolic blood pressure (median, mmHg)	74

Table 5

Table 5. Characteristics of the population used in the study.

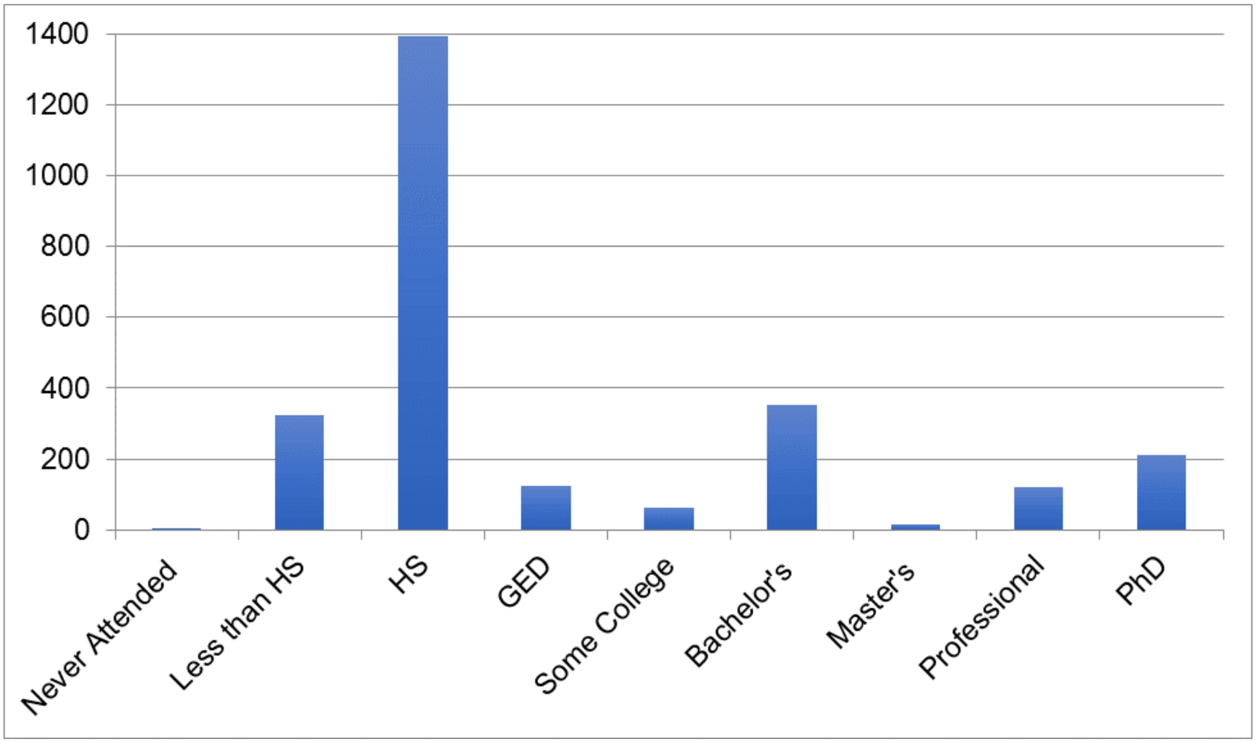

Figure 15

Graph of education level of individuals included in analyses.

The majority of individuals had a high school degree as their highest level of education. Levels of education are shown on the x-axis, number of individuals are shown on the y-axis. For analyses, individuals were grouped into less than a high school degree (n=328), high school degree and GED (n=1,518), and some college and above (n=731).

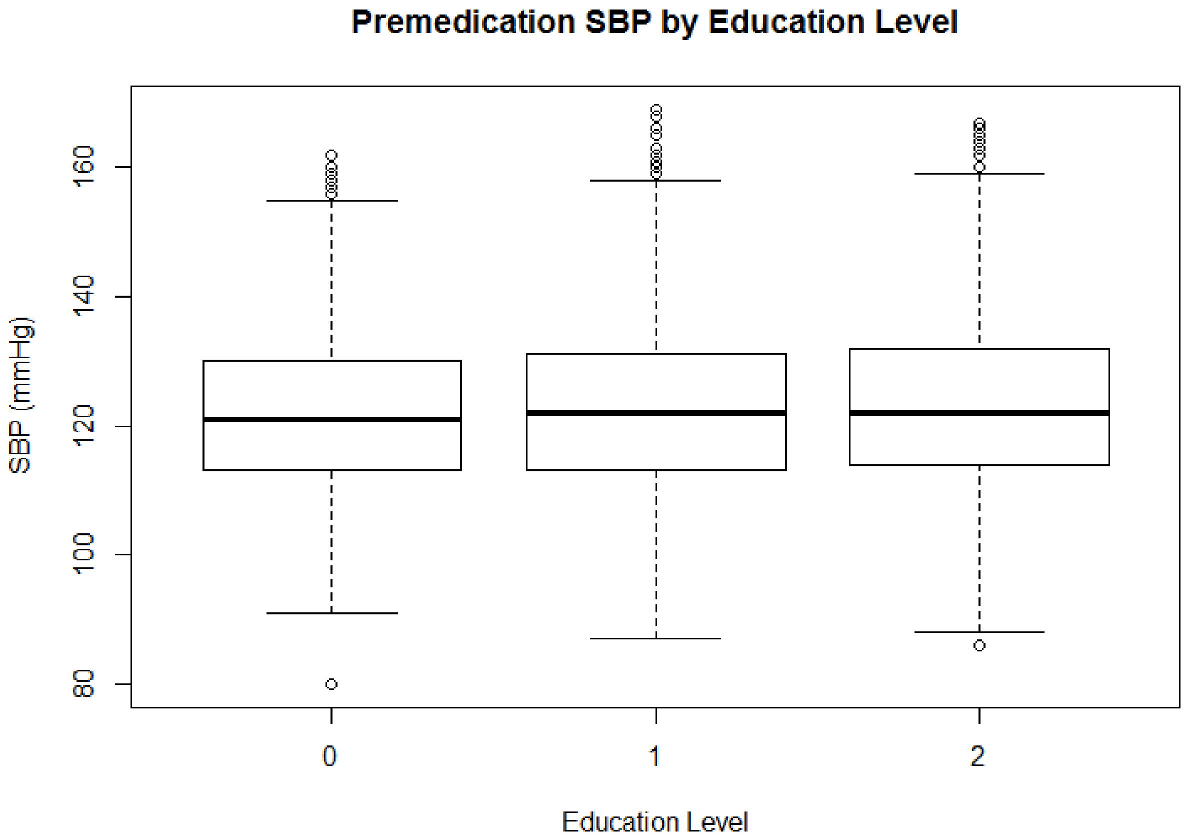

Figure 16

Premedication systolic blood pressure by education level.

The x-axis shows education level: Level 0 indicates less than high school, level 1 indicates high school degree and GED, level 2 indicates some college and above. The y-axis shows the median premedication systolic blood pressure.

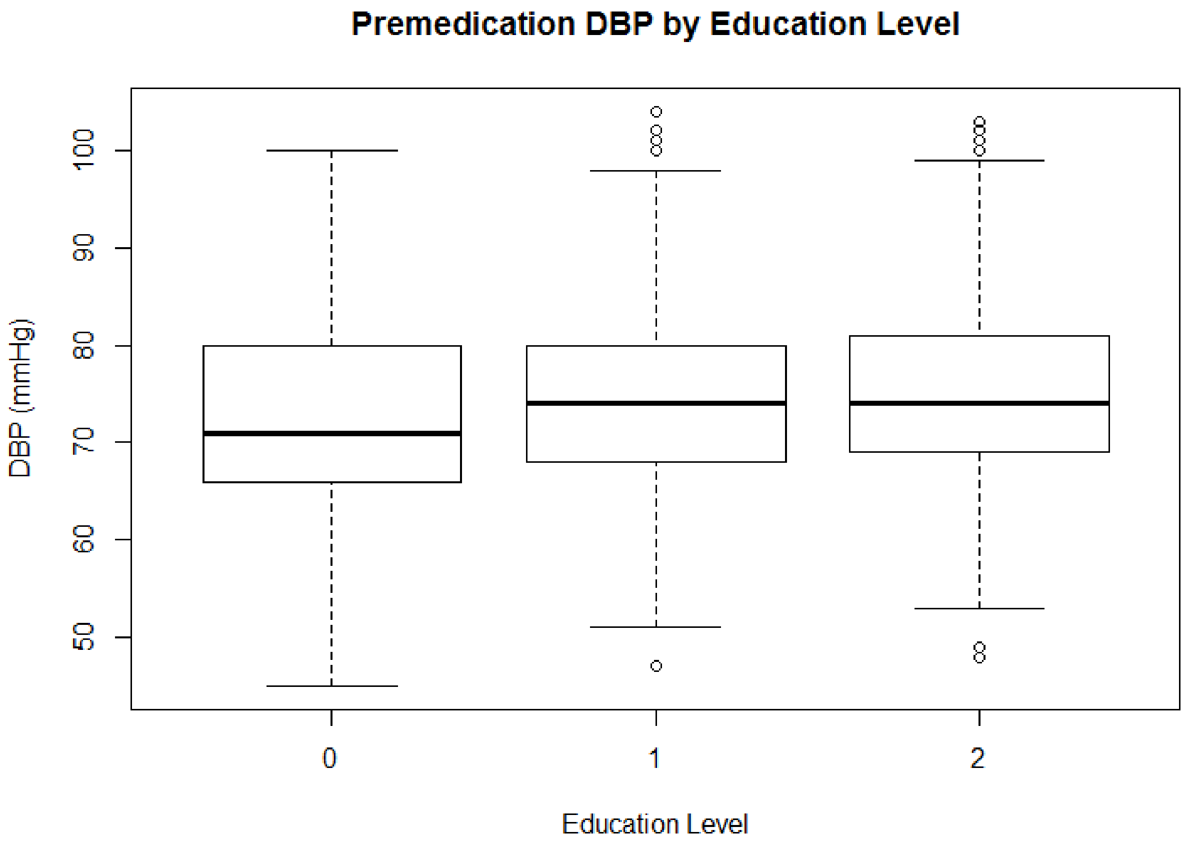

Figure 17

Premedication diastolic blood pressure by education level.

The x-axis shows education level: Level 0 indicates less than high school, level 1 indicates high school degree and GED, level 2 indicates some college and above. The y-axis shows median premedication diastolic blood pressure.

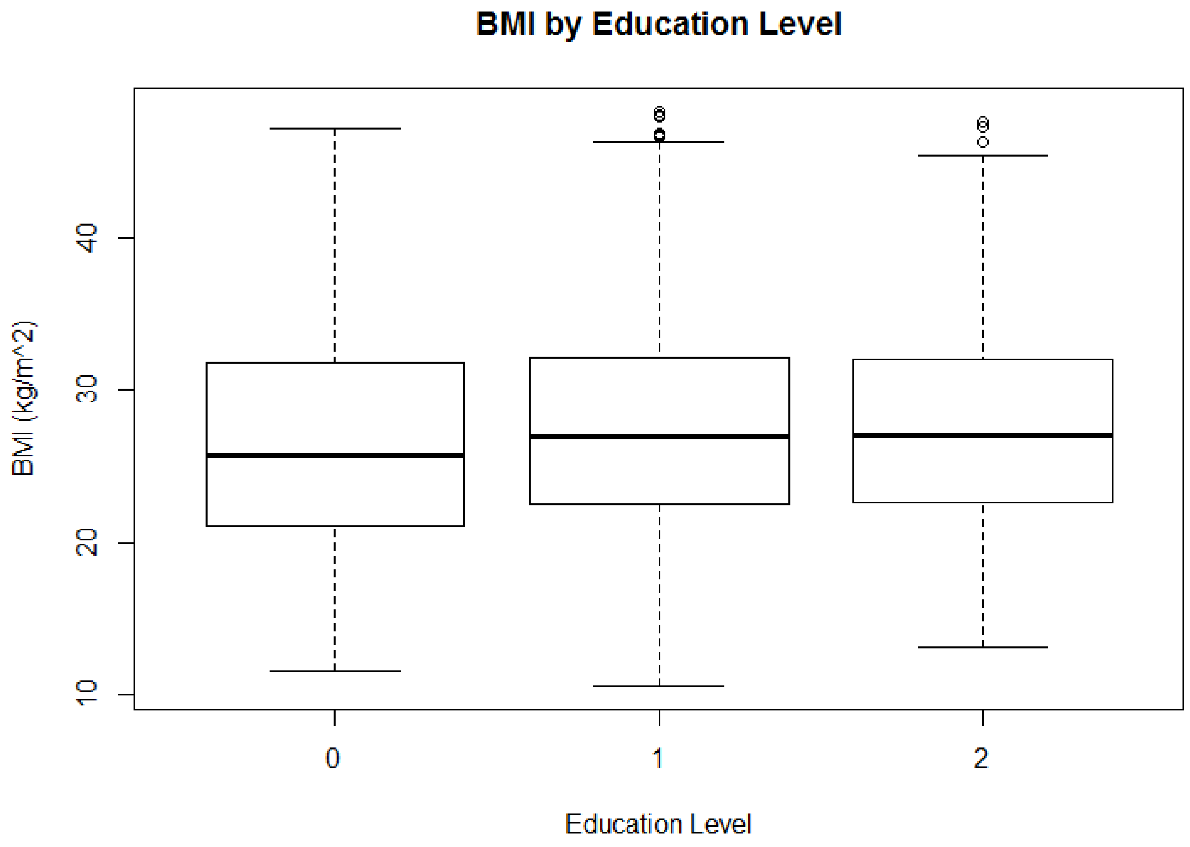

Figure 18

Body mass index by education level.

The x-axis shows education level: Level 0 indicates less than high school, level 1 indicates high school degree and GED, level 2 indicates some college and above. The y-axis shows body mass index, calculated using the median weight values extracted from the EHR.

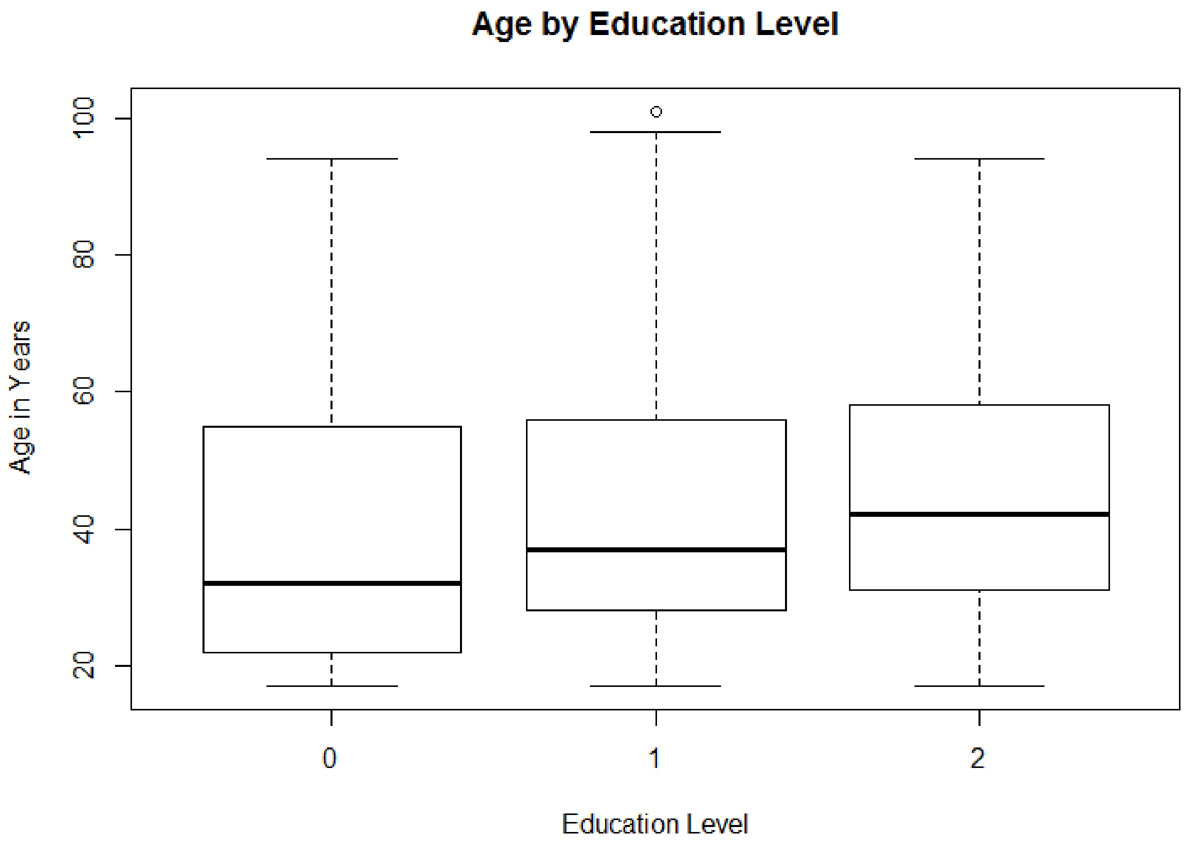

Figure 19

Age by education level.

The x-axis shows education level: Level 0 indicates less than high school, level 1 indicates high school degree and GED, level 2 indicates some college and above. Age increases with increasing education level. The y-axis shows the age of the participant as of 2015.

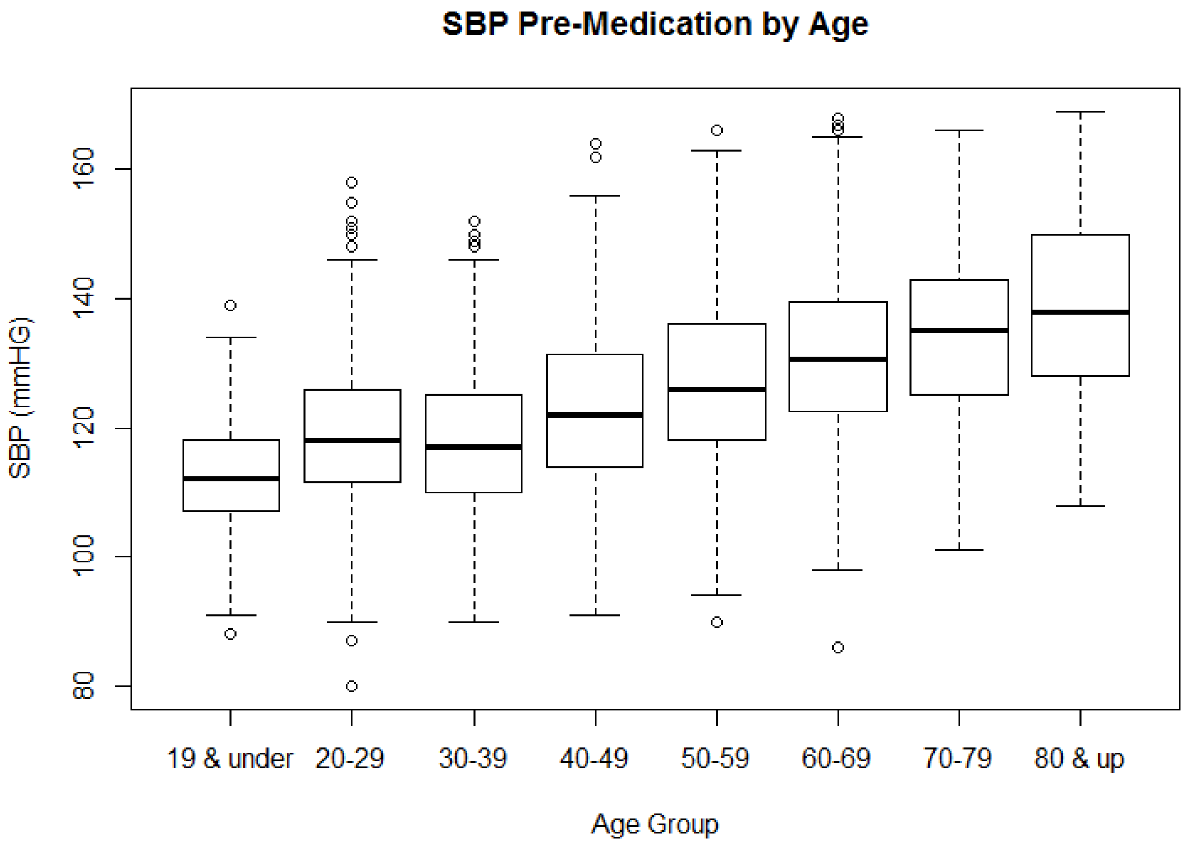

Figure 20

Premedication systolic blood pressure increases across age groups within our dataset. The x-axis shows age groups lumped by decade.

The y-axis shows the median premedication systolic blood pressure.

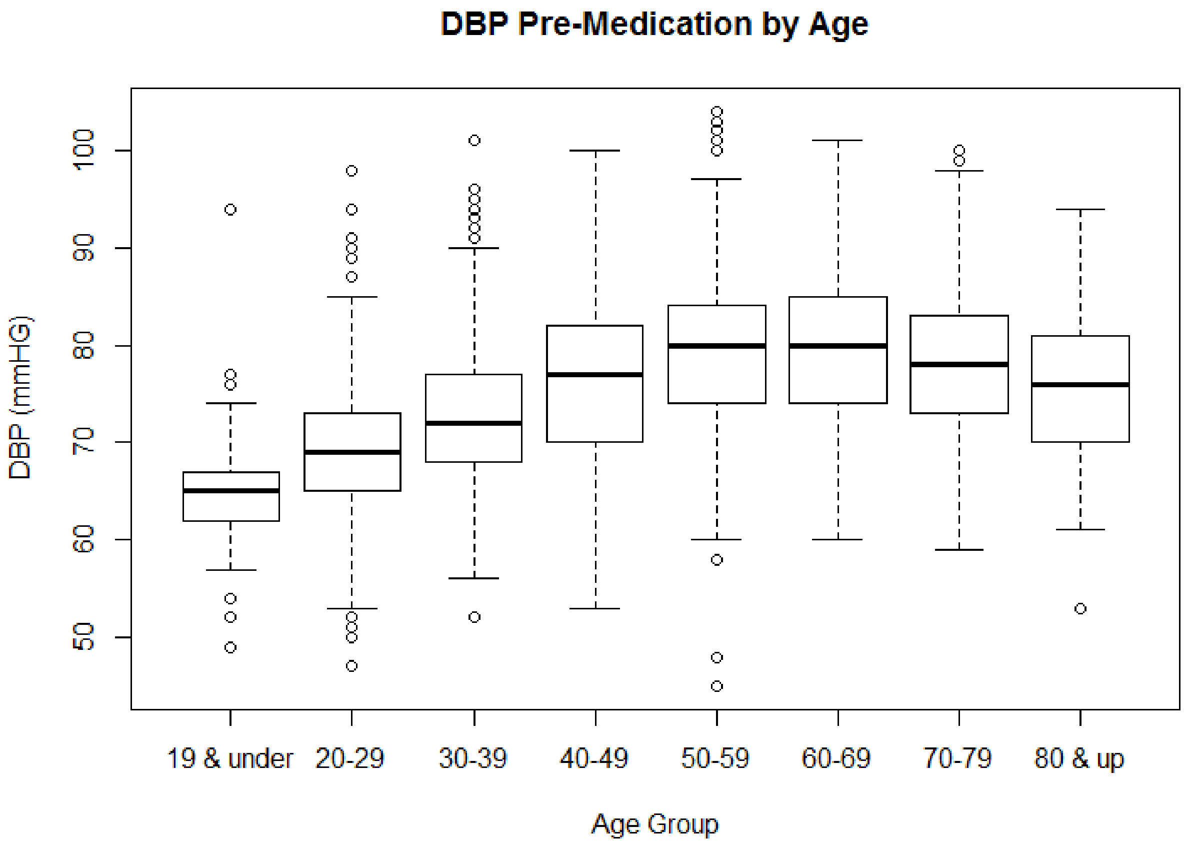

Figure 21

Premedication diastolic blood pressure increases until around age 60, then decreases in older individuals within our dataset.

The x-axis shows age groups lumped by decade. The y-axis shows the median premedication diastolic blood pressure.

Variable	Effect estimate (β)	Standard error	p-value
Premed SBP			
Education	0.02	0.11	0.84
Age	0.31	0.01	**<0.0001**
Sex (male is reference)	-4.52	0.52	**<0.0001**
BMI	0.38	0.04	**<0.0001**
Smoking status (nonsmoker is reference)	0.56	0.72	0.43
African ancestry	-0.46	1.89	0.809
Premed DBP			
Education	0.05	0.07	0.53
Age	0.19	0.01	**<0.0001**
Sex (male is reference)	-1.48	0.35	**0.0001**
BMI	0.25	0.02	**<0.0001**
Smoking status (nonsmoker is reference)	0.45	0.48	0.35
African ancestry	-1.24	1.24	0.32

Table 6

Table 6. Correlation of education and covariate variables with premedication systolic and premedication diastolic blood pressure.

Both systolic and diastolic blood pressure are correlated with age, sex, and body mass index.

Variable	**Degrees of freedom**	**Sum of squares**	**Mean squares**	**F value**	**p-value**
Age	1	8874	8874	28.45	**$1.05x10^{-7}$**
Sex	1	0.6	0.57	2.74	0.098
Smoking status	1	0.0	0.003	0.02	0.879
BMI	1	154	154	3.31	0.069
Premed SBP	1	545	545	2.94	0.087
Premed DBP	1	1221	1221.1	15.3	**$9.4x10^{-5}$**

Table 7

Table 7. Analysis of covariance (ANCOVA) between three level education variable and blood pressure, age, sex, smoking status, and body mass index.

Education is the independent variable and the other variables are examined individually, to see if they covary with age, without the other variables in the model. Education significantly co-varies with age and premedication diastolic blood pressure.

Included in analyses (n=2,577)	Individuals missing education (n=5,925)
Sex*** Male: 753 (29%) Female: 1,824 (71%)	Sex*** Male: 2,173 (37%) Female: 3,752 (63%)
Age*** Median, years: 38	Age*** Median, years: 57
African Ancestry Median: 81.7%	African Ancestry Median: 81.7%
Smoking Status*** Ever smokers: 335 (13%) Never smokers: 2,242 (87%)	Smoking Status (n=5,090; excluded missing data)*** Ever smokers: 795 (16%) Never smokers: 4,295 (84%)
BMI*** Median, kg/m^2: 26.8	BMI (n=4,522; dropped missings) *** Median, kg/m^2: 27.8
Premedication SBP*** Median, mmHg: 122	Premedication SBP (n=3,445; excluded missing data) *** Median, mmHg: 125
Premedication DBP *** Median, mmHg: 74	Premedication DBP (n=3,445; excluded missing data) *** Median, mmHg: 77

Table 8

Table 8. Comparison of population characteristics between individuals included in analyses and individuals excluded from analyses due to missing education information.

While there are statistically significant differences between sex, age, smoking status, body mass index, systolic blood pressure and diastolic blood pressure, the most striking difference is the median age between groups.

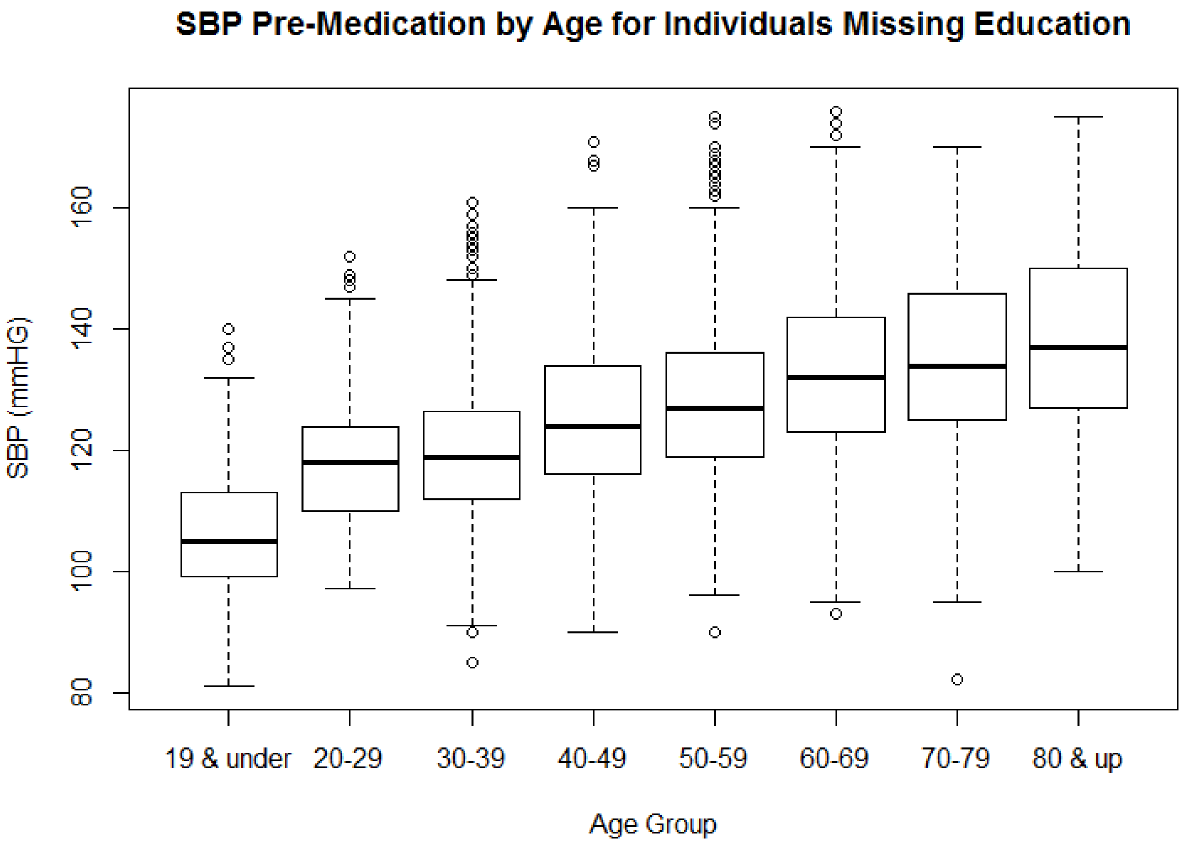

Figure 22

Premedication systolic blood pressure increases across age groups within individuals who are missing education information.

The x-axis shows age groups lumped by decade. The y-axis shows the median premedication systolic blood pressure.

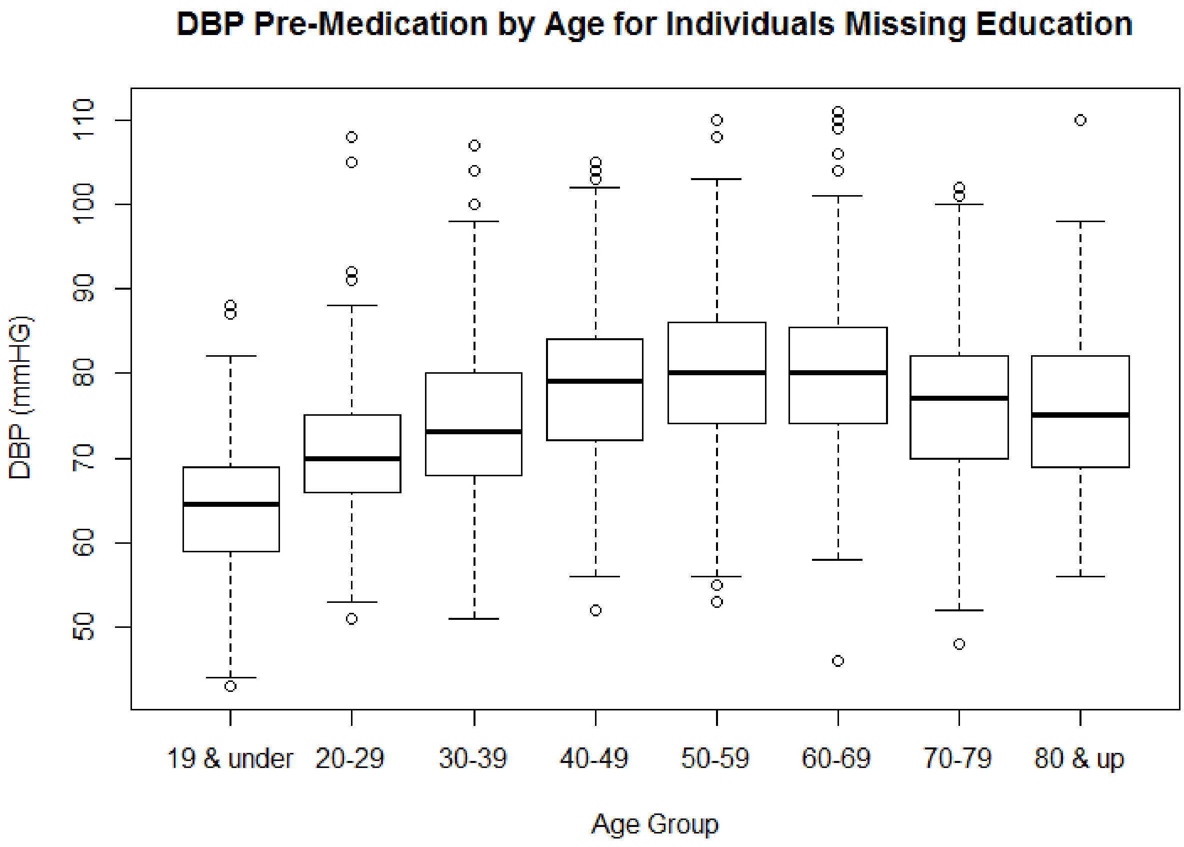

Figure 23

Premedication diastolic blood pressure increases across age groups within individuals who are missing education information.

The x-axis shows age groups lumped by decade. The y-axis shows the median premedication diastolic blood pressure.

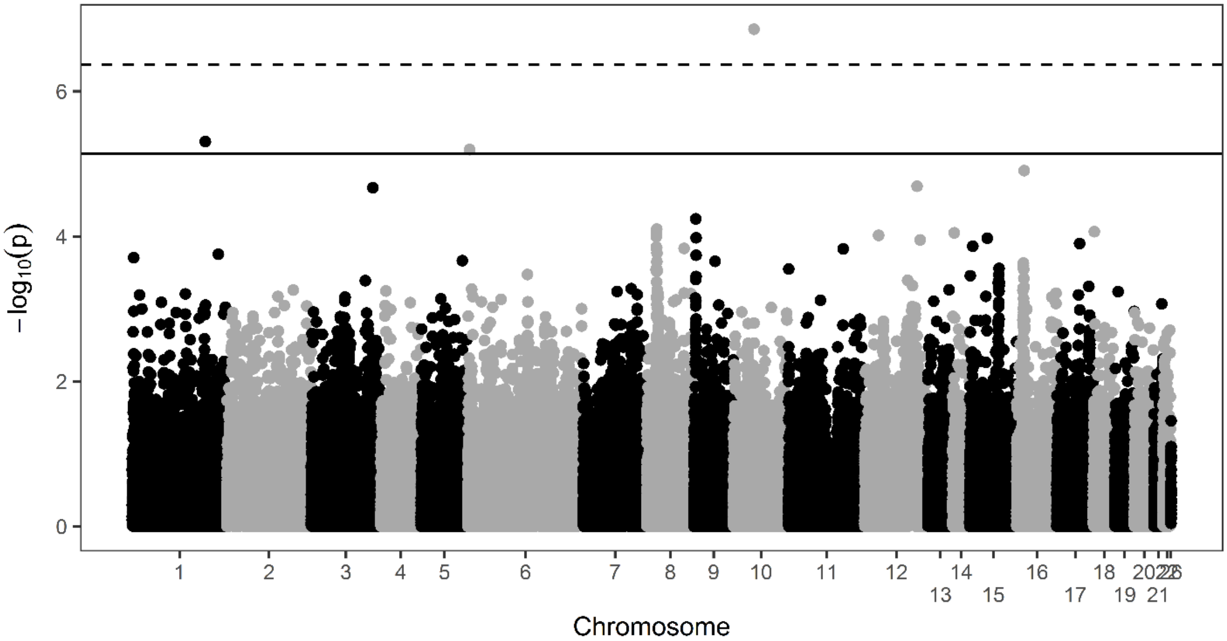

Figure 24

Manhattan plot of premedication systolic blood pressure.

Covariates include age, age squared, sex, body mass index, smoking status, and African ancestry. The x-axis shows SNP position grouped by chromosome number. The y-axis shows the -log$_{10}$ of the p-value for the SNP, which indicates that smaller p-values are higher on the axis. The dashed line is a Bonferroni correction, a p-value of 4.32 x 10^{-7}. The solid line is a suggestive line, which was calculated by removing SNPs with an r^{2} of higher than 0.6, was 7.24 x 10^{-6}.

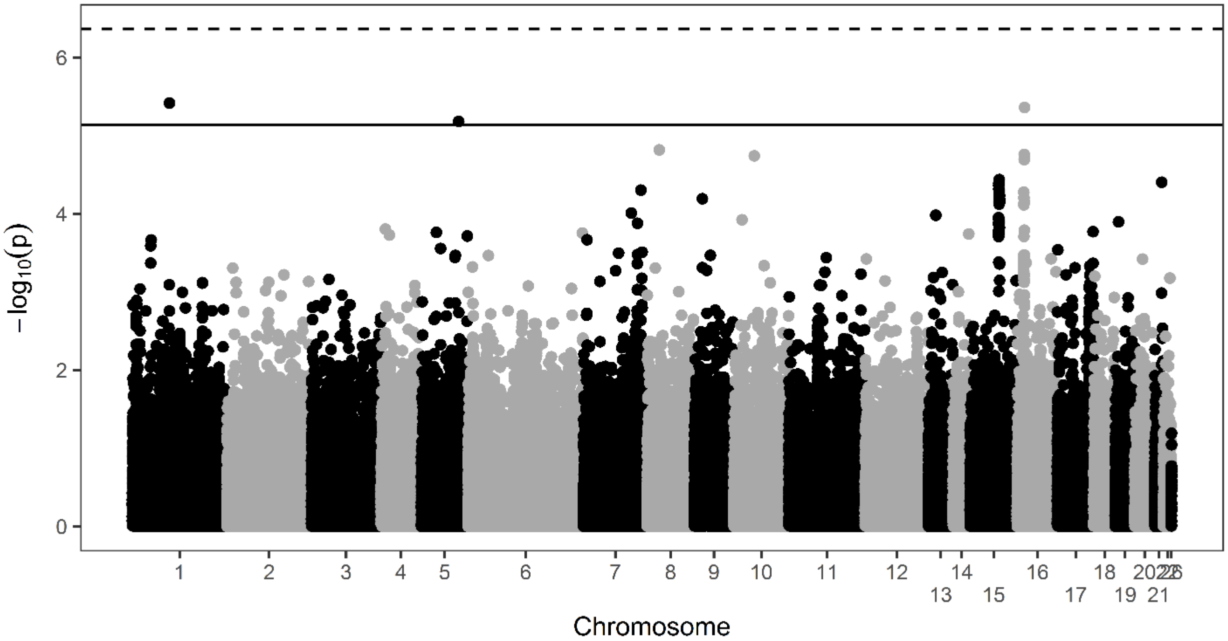

Figure 25

Manhattan plot of premedication diastolic blood pressure.

Covariates include age, age squared, sex, body mass index, smoking status, and African ancestry. The x-axis shows SNP position grouped by chromosome number. The y-axis shows the -log$_{10}$ of the p-value for the SNP, which indicates that smaller p-values are higher on the axis. The dashed line is a Bonferroni correction, a p-value of 4.32 x 10^{-7}. The solid line is a suggestive line, which was calculated by removing SNPs with an r^{2} of higher than 0.6, was 7.24 x 10^{-6}.

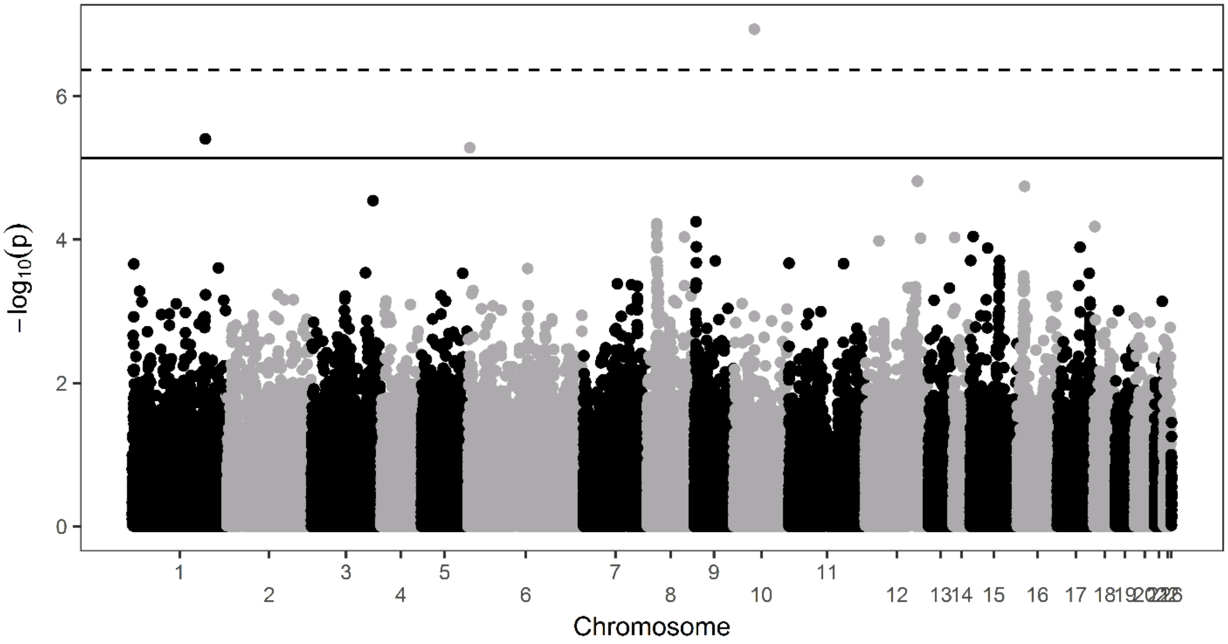

Figure 26

Manhattan plot of premedication systolic blood pressure.

Covariates include age, age squared, sex, body mass index, smoking status, African ancestry. Education is included in this model. The x-axis shows SNP position grouped by chromosome number. The y-axis shows the -log$_{10}$ *of the p-value for the SNP, which indicates that smaller p-values are higher on the axis. The dashed line is a Bonferroni correction, a p-value of 4.32 x 10*$^{-7}$*. The solid line is a suggestive line, which was calculated by removing SNPs with an r*^{2} *of higher than 0.6, was 7.24 x 10*$^{-6}$*.*

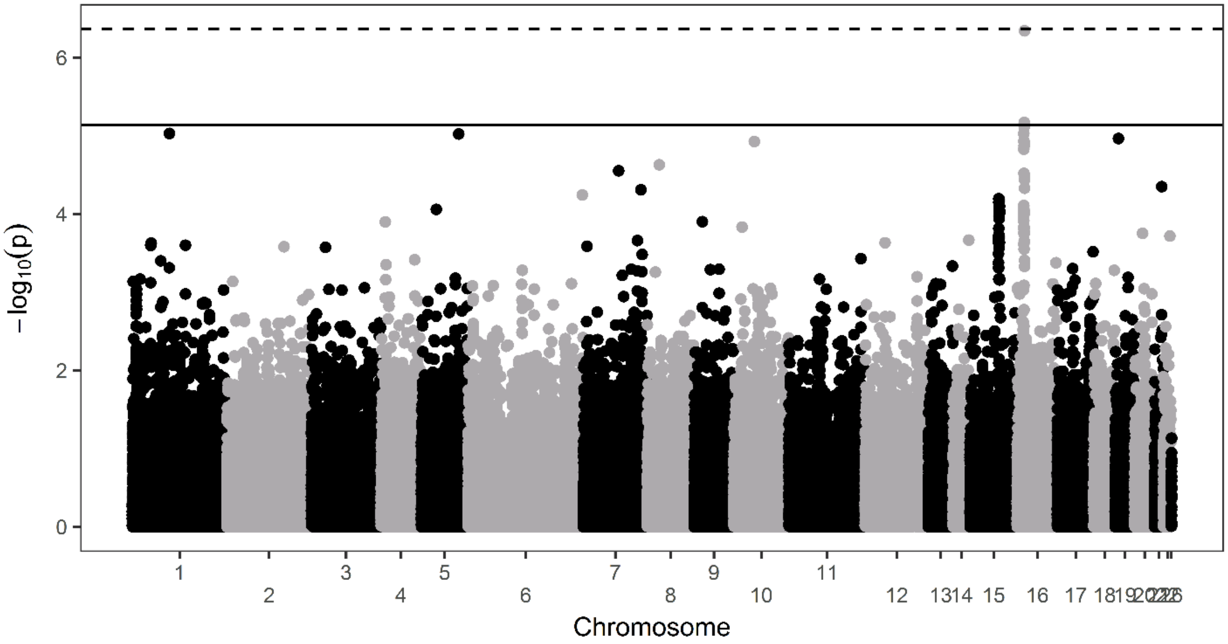

Figure 27

Manhattan plot of premedication diastolic blood pressure.

Covariates include age, age squared, sex, body mass index, smoking status, African ancestry. Education is included in this model. The x-axis shows SNP position grouped by chromosome number. The y-axis shows the -log$_{10}$ of the p-value for the SNP, which indicates that smaller p-values are higher on the axis. The dashed line is a Bonferroni correction, a p-value of 4.32 x 10^{-7}. The solid line is a suggestive line, which was calculated by removing SNPs with an r^{2} of higher than 0.6, was 7.24 x 10^{-6}.

SNP	**Assoc. with**	**Location**	**Associated gene**	**Minor allele frequency**	**Effect estimate**	**Std error**	**p-value**
rs4593967	SBP	Intron	ARHGAP22	13.91%	-2.53	0.48	**1.16×10^{-7}**
rs10921895	SBP	Intergenic		37.14%	-1.55	0.36	3.92×10^{-6}
rs3804485	SBP	Intron	LY86	41.28%	1.51	0.33	5.20×10^{-6}
rs950928	DBP	Intron	IQCK	36.35%	-1.10	0.22	**4.53×10^{-7}**
rs8056711	DBP	Intron	IQCK	36.35%	-1.10	0.22	4.53×10^{-7}

Table 9

Table 9. Summary of characteristics of SNPs associated with premedication systolic and diastolic blood pressure when education is included in the model.

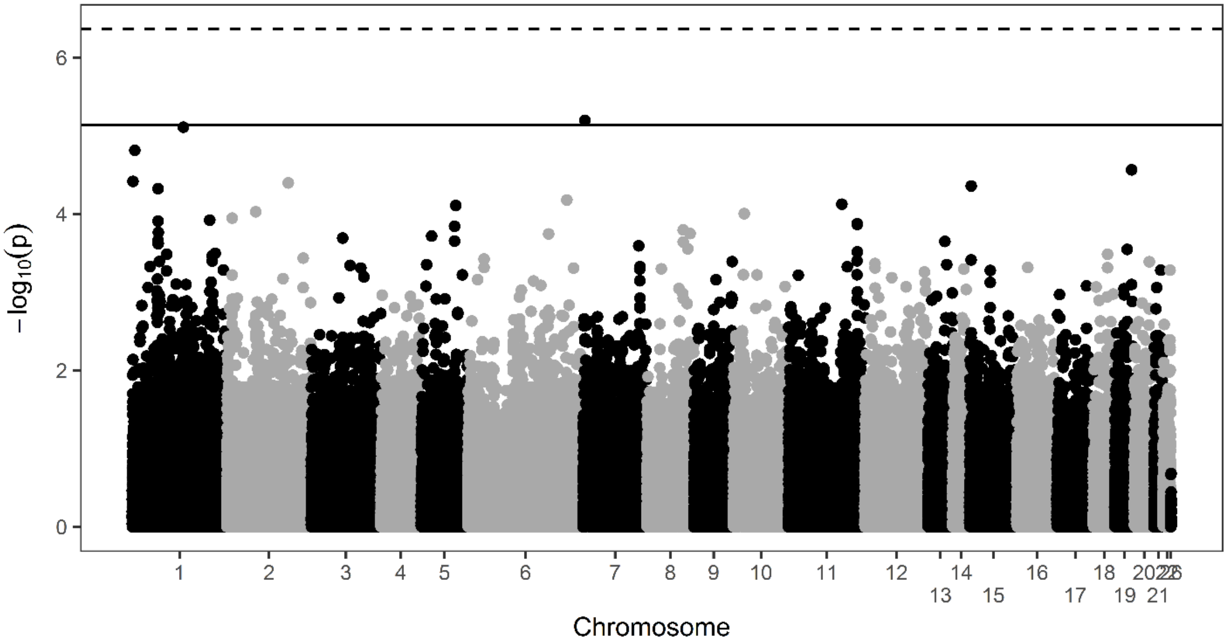

Figure 28

Manhattan plot of interaction term p-values for premedication systolic blood pressure analysis.

Covariates include age, sex, body mass index, smoking status, African ancestry. Education and SNP x education interactions were also included in this model. The x-axis shows SNP position grouped by chromosome number. The y-axis shows the -log_{10} of the p-value for the SNP, which indicates that smaller p-values are higher on the axis. The dashed line is a Bonferroni correction, a p-value of 4.32 x 10^{-7}. The solid line is a suggestive line, which was calculated by removing SNPs with an r^2 of higher than 0.6, was 7.24 x 10^{-6}.

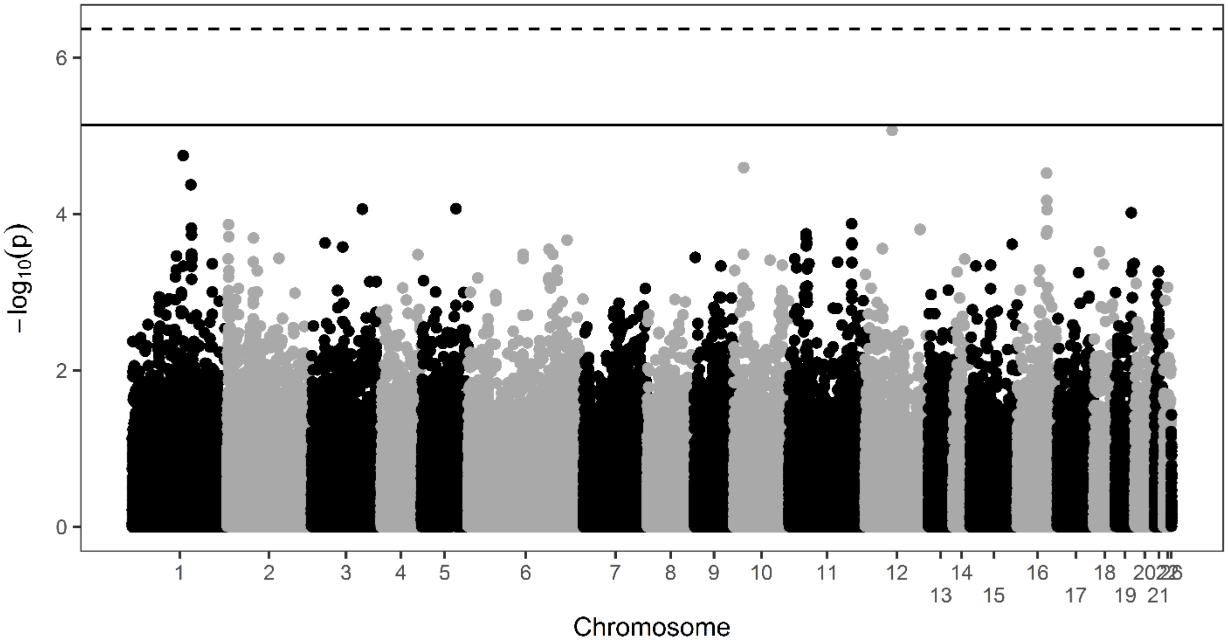

Figure 29

Manhattan plot of interaction term p-values for premedication diastolic blood pressure analysis.

Covariates include age, sex, body mass index, smoking status, African ancestry. Education and SNP x education interactions were also included in this model. The x-axis shows SNP position grouped by chromosome number. The y-axis shows the -log$_{10}$ of the p-value for the SNP, which indicates that smaller p-values are higher on the axis. The dashed line is a Bonferroni correction, a p-value of 4.32 x 10^{-7}. The solid line is a suggestive line, which was calculated by removing SNPs with an r^{2} of higher than 0.6, was 7.24 x 10^{-6}.

SNP in education interaction	p-value of the high school interaction term	p-value of the college interaction term
rs4593967_A	0.886	0.858
rs10921895_G	0.896	0.746
rs3804485_C	0.260	0.863
rs11066700_A	0.178	0.200

Table 10

Table 10. SNPs examined for education interactions impacting systolic blood pressure.

Covariates included in the model were age, age squared, sex, body mass index, smoking status, and African ancestry. The main effect of education and the SNP x education interaction term were also included in the model.

SNP in education interaction	p-value with education
chr16.19732139_G	High school: 0.175 College and higher: 0.298
chr16.19734035_C	High school: 0.175 College and higher: 0.297
chr16.19700099_A	High school: 0.927 College and higher: 0.503
chr16.19702910_T	High school: 0.940 College and higher: 0.478
chr16.19690303_A	High school: 0.941 College and higher: 0.461
rs6687976_A	High school: 0.052 College and higher: 0.996
chr16.19642355_G	High school: 0.094 College and higher: 0.110
rs3095994_A	High school: 0.738 College and higher: 0.726
rs1273518_G	High school: 0.218 College and higher: 0.648
chr16.19660835_A	High school: 0.076 College and higher: 0.091
rs4593967_A	High school: 0.512 College and higher: 0.768
chr16.19689461_C	High school: 0.181 College and higher: 0.316
chr16.19676895_A	High school: 0.062 College and higher: 0.085
chr16.19641087_A	High school: 0.064 College and higher: 0.076

Table 11

Table 11. SNPs examined for education interactions impacting diastolic blood pressure.

Covariates included in the model were age, age squared, sex, body mass index, smoking status, and African ancestry. The main effect of education and the SNP x education interaction term were also included in the model.

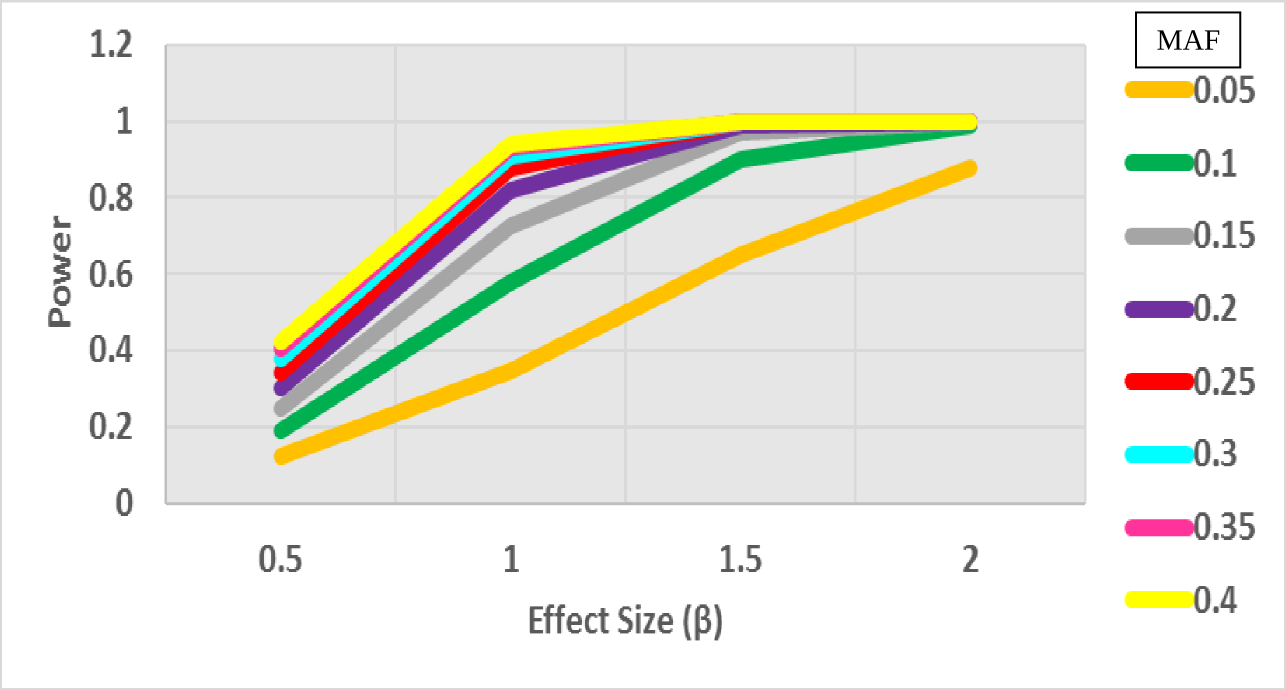

Figure 30

Figure 30. Power estimation for the detection of genetic associations based on a population of 2,577 individuals.

These power calculations were determined by assuming an additive genetic model and a continuous trait. The study is not well-powered to detect rarer variation or small effect sizes. The study is powered to detect effect sizes above 1 and minor allele frequencies (MAF) above 0.15.

CHAPTER FOUR

CONCLUSION

Summary of Chapter Two

Results

In Chapter Two, the aim was to develop a set of algorithms that could be used to extract existing socioeconomic data from electronic health record databases such as BioVU's Synthetic Derivative. A set of seven algorithms was created to extract occupation, unemployment, retirement, education level, Medicaid, a lack of health insurance, and homelessness. The algorithms for education, occupation, unemployment, and Medicaid all had a positive predictive value of over 80%. The algorithm for retirement had a positive predictive value of 64%. The algorithm for homelessness had a positive predictive value of 33% and the algorithm for uninsured status had a positive predictive value of 23%. While these values are somewhat low, it is important to consider that these three categories were also the least prevalent within our evaluation dataset.

Limitations

There were many challenges with the development of this algorithm. The most difficult to address was that of missing data. Unfortunately, it is not constant that medical providers include socioeconomic data within a patient's record. This lack of recording made it difficult to consistently extract every type of socioeconomic data from every patient's record. While it was common to be able to extract one or two variables from a record, it was rare to be able to extract information from one patient that contained all seven variables.

In addition to the data that was not recorded by providers, there were also challenges with the data that was recorded. As the Institute of Medicine recommendations are not currently applied at Vanderbilt University Medical Center, the information that was recorded by providers was not standardized, in the clinical narrative, and generally very inconsistent across records. These inconsistencies made extracting the socioeconomic data in a systematic way extremely difficult. The algorithms developed were limited in their methods and therefore not perfect in their ability to extract all the available socioeconomic data.

Strengths

Despite the challenges with the development of these algorithms, these algorithms were able to successfully extract a large amount of fairly accurate socioeconomic data. Prior to this investigation, algorithms to extract socioeconomic data from electronic health records had never been developed. The development of these algorithms is an important first step to incorporating socioeconomic data into electronic health record based studies and achieving the goals of precision medicine research.

Summary of Chapter Three

Results

In order to examine the effect of the inclusion of socioeconomic data (in the form of education) on a large scale genetic analysis of blood pressure in a black population, three different regression models were analyzed. In the first set of models, both premedication systolic blood pressure and diastolic blood pressure were examined as outcomes, with age, age squared, sex, body mass index, smoking status, and percent African ancestry as covariates. When education was added to the models, small changes in the significance of our most significant associations were discovered. This could indicate that whatever environmental impact the education variable is representing could be affecting to a small degree the associations between genetic variants and blood pressure measurements.

The investigation included SNP-education interactions by including an interaction term in the models with education. No statistically significant associations were observed and therefore the null hypothesis was supported. However, the limitations of the study may have contributed to type II error and therefore it is important to keep investigating potential gene-social environment interactions.

Limitations

Despite some interesting findings, this study had a number of limitations. Primarily, the main limitation was a lack of a replication dataset. Without this dataset, the null hypothesis cannot truly be rejected. Additionally, the investigation was limited by a relatively small sample size compared to some of the more recent large scale meta-analyses and the only access was to SNPs from a selective genotyping chip. While these genotype data were a great tool for the investigation, they are still limited in their focus. A large number of variants from our dataset had to be removed because many of the variants were rare in African ancestry populations; there were also limitations regarding the accuracy of phenotyping. Unfortunately, extracting phenotype data from an electronic health record can be limiting, as the study relies on the accuracy of the available data and the ability to extract the data accurately. While the algorithms utilized performed well in testing, it is always possible that remaining imperfections affected the results.

The age of our population may have also had an impact on the results of the analyses. The study population had a median age of 38 years, which is young when compared with other published study populations[105]. The young age of the participants led to a limited number of individuals who had developed high blood pressure and decreased variation in blood pressure measurements. This decreased variation and the young age of participants may have led to a lack of associations because individuals who may be at risk of developing high blood pressure have not developed it yet.

Beyond these limitations, there are also limitations in terms of the variables examined in analysis. Gene x education interactions were examined without a main effect of education on the outcome, blood pressure. The analysis method used is not robust enough to detect an interaction without a main effect. Additional analysis methods are likely to be more appropriate and provide more information regarding the effect of education on genetic associations with blood pressure.

The use of education as a measurement for socioeconomic status may also be a limitation. As alluded to earlier, higher education is assumed to be associated with increased social mobility and increased income. These associations may not be consistent across racial groups. For example, black individuals with higher levels of education may not have access the same levels of mobility as white individuals with the same level of education[104]. In terms of their effects on biology, socioeconomic status and education can be used as indicators of chronic stress. Utilizing these measurements as proxies, without measuring the actual level of stress individuals are experiencing, can lead to a lack of observed associations.

Strengths

Despite these limitations, the study did have multiple strengths. It was the first to examine the impact of socioeconomic data on a large scale genetic study of blood pressure in a black population. It was also the first to utilize socioeconomic data extracted from a de-identified electronic health record in a genetic study. While the study did not find any significant interactions, it did contribute to the field of health disparities by showing that it is possible to include social environment data in a large scale genetic study of an existing dataset. This ability is novel and an important step in utilizing genetic information to address health disparities. Without the ability to include any social environment data, it is difficult for geneticists to contribute meaningful findings to the field of health disparities. These novelties in this work lay the groundwork for the inclusion of additional socioeconomic data in future studies of health outcomes, especially those with disparities.

Future Directions

In order to continue to improve these investigations, it would be ideal to have an independent replication dataset to validate the findings. The validation of most significantly associated SNPs would support the confidence in these findings and the addition of even minimal socioeconomic data to large scale genetic studies has the power to elucidate new genetic variants. It may also be informative to explore other approaches to the genetic influences on blood pressure, such as utilizing genetic risk scores. The subtle effect sizes of typical blood pressure-related SNPs could limit the ability to investigate interactions. Investigating the use of methods which group genetic variants may create more power in examining blood pressure. These methods, such as genetic risk scores or pathway analysis, as well as examining variation which affects expression could be more productive since environmental factors may be acting on pathways and expression, rather than individual variants. Therefore, observing an interaction may be more likely. These methods can also help reduce multiple testing burden, which is an issue when examining genome-wide variants and interactions.

Investigating the effect of different types of socioeconomic variables (such as the other variables we extracted) would also be worthwhile. Exploring interactions with chronic stress variables such as biomarkers and survey data, or other environmental variables that are associated with socioeconomic status such as exposure to toxins, would be an ideal situation. Socioeconomic status in general is a proxy for these exposures, so measuring the direct variables would be more informative. However, socioeconomic status markers can be more easily collected for large scale studies than these variables, or extracted from EHRs as was demonstrated here. Therefore, continuing to explore the use of socioeconomic status in genetic studies is worthwhile.

With the continuation of exploring the use of socioeconomic status in genetic studies, it is vital to think carefully about which measurements to use. As discussed earlier, socioeconomic variables do not represent the same aspects of social environment across racial groups in the United States. It would be ideal to collect as much social environment information as possible, determine the limitations and covariance across variables, then determine which variables would best measure the variable of interest in the study. For example, measuring education may be intended to be a proxy for income or health behaviors. Measuring these variables directly would be more informative.

In terms of additional improvements to the study, including more sophisticated language processing techniques would greatly improve extraction outcomes of socioeconomic information from electronic health records. Encouraging the adoption of a standard set of social environment-related questions into the electronic health record would be even better as recommended by the Institute of Medicine[107]. The addition of this information would help clinicians better treat patients and help researchers conduct better research.

Within the context of the future, it is important to consider the Precision Medicine Initiative. This program has the potential to be very helpful with addressing genetic questions. The goal of the Precision Medicine Initiative cohort is one million participants, with genetic data, as well as EHR, and other environmental data. The collection of a large cohort with such extensive data will provide more power to address these questions of genetic variants associated with blood pressure in black individuals, as one of the goals of the cohort is to collect diverse individuals. Beyond investigating black populations and blood pressure, the Precision Medicine Initiative will also allow investigators to examine other diverse populations and other phenotypes with this incredibly rich dataset.

It is imperative to continue to investigate health disparities and move toward health equity. There are many different types of health disparities within the United States and conducting strong research and gathering evidence on the causes of these disparities is a central step in making societal and policy changes to reduce them. Without fully understanding the biology, the medical community cannot strive for the necessary societal changes that must occur.

BIBLIOGRAPHY

1. CSDH. (2008). Closing the gap in a generation: health equity through action on the social determinants of health. Final Report of the Commission on Social Determinants of Health. In. (Geneva.)

2. Adler, N.E., Boyce, T., Chesney, M.A., Cohen, S., Folkman, S., Kahn, R.L., and Syme, S.L. (1994). Socioeconomic status and health. The challenge of the gradient. Am Psychol 49, 15-24.

3. Singh-Manoux, A., Adler, N.E., and Marmot, M.G. (2003). Subjective social status: its determinants and its association with measures of ill-health in the Whitehall II study. Social Science & Medicine 56, 1321-1333.

4. Braveman, P.A., Cubbin, C., Egerter, S., Chideya, S., Marchi, K.S., Metzler, M., and Posner, S. (2005). Socioeconomic status in health research: one size does not fit all. JAMA 294, 2879-2888.

5. Singh-Manoux, A., Marmot, M.G., and Adler, N.E. (2005). Does subjective social status predict health and change in health status better than objective status? Psychosom Med 67, 855-861.

6. Diez Roux, A.V., Merkin, S.S., Arnett, D., Chambless, L., Massing, M., Nieto, F.J., Sorlie, P., Szklo, M., Tyroler, H.A., and Watson, R.L. (2001). Neighborhood of residence and incidence of coronary heart disease. The New England Journal of Medicine 345, 99-106.

7. Waitzman, N.J., and Smith, K.R. (1998). Phantom of the area: poverty-area residence and mortality in the United States. American Journal of Public Health 88, 973-976.

8. Smith, G.D., Hart, C., Watt, G., Hole, D., and Hawthorne, V. (1998). Individual social class, area-based deprivation, cardiovascular disease risk factors, and mortality: the Renfrew and Paisley Study. J Epidemiol Community Health 52, 399-405.

9. Pickett, K.E., and Pearl, M. (2001). Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. J Epidemiol Community Health 55, 111-122.

10. Williams, D.R., Priest, N., and Anderson, N.B. (2016). Understanding associations among race, socioeconomic status, and health: Patterns and prospects. Health Psychol 35, 407-411.

11. Adler, N.E., and Ostrove, J.M. (1999). Socioeconomic status and health: what we know and what we don't. Ann N Y Acad Sci 896, 3-15.

12. Marmot, M.G., Rose, G., Shipley, M., and Hamilton, P.J. (1978). Employment grade and coronary heart disease in British civil servants. J Epidemiol Community Health 32, 244-249.

13. Seeman, T., Merkin, S.S., Crimmins, E., Koretz, B., Charette, S., and Karlamangla, A. (2008). Education, income and ethnic differences in cumulative biological risk profiles in a national sample of US adults: NHANES III (1988-1994). Social science & medicine 66, 72-87.

14. Braveman, P., Egerter, S., and Williams, D.R. (2011). The social determinants of health: coming of age. Annual review of public health 32, 381-398.

15. (2010). Healthy People 2020: An Opportunity to Address the Societal Determinants of Health in the United States.

16. (2012). Health, United States, 2011: With special feature on socioeconomic status and health. In. (Hyattsville, Maryland.)

17. (2016). How the Census Bureau Measures Poverty. In, U.S.C. Bureau, ed.

18. Ventriglio, A., Mari, M., Bellomo, A., and Bhugra, D. (2015). Homelessness and mental health: A challenge. Int J Soc Psychiatry 61, 621-622.

19. Currie, L.B., Patterson, M.L., Moniruzzaman, A., McCandless, L.C., and Somers, J.M. (2014). Examining the relationship between health-related need and the receipt of care by participants experiencing homelessness and mental illness. BMC Health Serv Res 14, 404.

20. Hauser, R., and Warren, J. (1997). Socioeconomic indexes for occupations: A review, update, and critique. Sociological Methodology 27, 177-298.

21. Fujishiro, K., Xu, J., and Gong, F. (2010). What does "occupation" represent as an indicator of socioeconomic status?: exploring occupational prestige and health. Social Science & Medicine 71, 2100-2107.

22. (2009). America's Uninsured Crisis: Consequences for Health and Health Care. In, I.o. Medicine, ed.

23. (2016). Medicaid Eligibility In, U.S.D.o.H.a.H. Services, ed.

24. (2016). Health, United States, 2015: With special feature on racial and ethnic health disparities In. (Hyattsville, Maryland.)

25. Braveman, P., and Gottlieb, L. (2014). The social determinants of health: it's time to consider the causes of the causes. Public health reports 129 Suppl 2, 19-31.

26. Ottman, R. (1996). Gene-environment interaction: definitions and study designs. Prev Med 25, 764-770.
27. Reiss, D., Leve, L.D., and Neiderhiser, J.M. (2013). How genes and the social environment moderate each other. American Journal of Public Health 103 Suppl 1, S111-121.

28. Forouzanfar, M.H., Liu, P., Roth, G.A., Ng, M., Biryukov, S., Marczak, L., Alexander, L., Estep, K., Hassen Abate, K., Akinyemiju, T.F., et al. (2017). Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015. JAMA 317, 165-182.

29. Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., Zhou, J., Lashley, K., Chen, Y., Christman, M., et al. (2009). A genome-wide association study of hypertension and blood pressure in African Americans. PLoS Genet 5, e1000564.

30. Hackam, D.G., Quinn, R.R., Ravani, P., Rabi, D.M., Dasgupta, K., Daskalopoulou, S.S., Khan, N.A., Herman, R.J., Bacon, S.L., Cloutier, L., et al. (2013). The 2013 Canadian Hypertension Education Program recommendations for blood pressure measurement, diagnosis, assessment of risk, prevention, and treatment of hypertension. Can J Cardiol 29, 528-542.

31. (2014). National Health and Nutrition Examination Survey. In, C.f.D.C.a. Prevention, ed. (Hyattsville, MD.)

32. Fox, E.R., Young, J.H., Li, Y., Dreisbach, A.W., Keating, B.J., Musani, S.K., Liu, K., Morrison, A.C., Ganesh, S., Kutlar, A., et al. (2011). Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study. Hum Mol Genet 20, 2273-2284.

33. Levy, D., Ehret, G.B., Rice, K., Verwoert, G.C., Launer, L.J., Dehghan, A., Glazer, N.L., Morrison, A.C., Johnson, A.D., Aspelund, T., et al. (2009). Genome-wide association study of blood pressure and hypertension. Nat Genet 41, 677-687.

34. Padmanabhan, S., Melander, O., Johnson, T., Di Blasio, A.M., Lee, W.K., Gentilini, D., Hastie, C.E., Menni, C., Monti, M.C., Delles, C., et al. (2010). Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. PLoS Genet 6, e1001177.

35. Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S., et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 41, 666-676.

36. Wain, L.V., Verwoert, G.C., O'Reilly, P.F., Shi, G., Johnson, T., Johnson, A.D., Bochud, M., Rice, K.M., Henneman, P., Smith, A.V., et al. (2011). Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet 43, 1005-1011.

37. International Consortium for Blood Pressure Genome-Wide Association, S., Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478, 103-109.

38. Org, E., Eyheramendy, S., Juhanson, P., Gieger, C., Lichtner, P., Klopp, N., Veldre, G., Doring, A., Viigimaa, M., Sober, S., et al. (2009). Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. Hum Mol Genet 18, 2288-2296.

39. Yang, H.C., Liang, Y.J., Wu, Y.L., Chung, C.M., Chiang, K.M., Ho, H.Y., Ting, C.T., Lin, T.H., Sheu, S.H., Tsai, W.C., et al. (2009). Genome-wide association study of young-onset hypertension in the Han Chinese population of Taiwan. PloS One 4, e5459.

40. Liu, X., Hu, C., Bao, M., Li, J., Liu, X., Tan, X., Zhou, Y., Chen, Y., Wu, S., Chen, S., et al. (2016). Genome Wide Association Study Identifies L3MBTL4 as a Novel Susceptibility Gene for Hypertension. Sci Rep 6, 30811.

41. Lu, X., Wang, L., Lin, X., Huang, J., Charles Gu, C., He, M., Shen, H., He, J., Zhu, J., Li, H., et al. (2015). Genome-wide association study in Chinese identifies novel loci for blood pressure and hypertension. Hum Mol Genet 24, 865-874.

42. Kelly, T.N., Takeuchi, F., Tabara, Y., Edwards, T.L., Kim, Y.J., Chen, P., Li, H., Wu, Y., Yang, C.F., Zhang, Y., et al. (2013). Genome-wide association study meta-analysis reveals transethnic replication of mean arterial and pulse pressure loci. Hypertension 62, 853-859.

43. Guo, Y., Tomlinson, B., Chu, T., Fang, Y.J., Gui, H., Tang, C.S., Yip, B.H., Cherny, S.S., Hur, Y.M., Sham, P.C., et al. (2012). A genome-wide linkage and association scan reveals novel loci for hypertension and blood pressure traits. PloS One 7, e31489.

44. Kidambi, S., Ghosh, S., Kotchen, J.M., Grim, C.E., Krishnaswami, S., Kaldunski, M.L., Cowley, A.W., Jr., Patel, S.B., and Kotchen, T.A. (2012). Non-replication study of a genome-wide association study for hypertension and blood pressure in African Americans. BMC Med Genet 13, 27.

45. Zhu, X., Young, J.H., Fox, E., Keating, B.J., Franceschini, N., Kang, S., Tayo, B., Adeyemo, A., Sun, Y.V., Li, Y., et al. (2011). Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. Hum Mol Genet 20, 2285-2295.

46. Franceschini, N., Fox, E., Zhang, Z., Edwards, T.L., Nalls, M.A., Sung, Y.J., Tayo, B.O., Sun, Y.V., Gottesman, O., Adeyemo, A., et al. (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. Am J Hum Genet 93, 545-554.

47. Zhu, X., Feng, T., Tayo, B.O., Liang, J., Young, J.H., Franceschini, N., Smith, J.A., Yanek, L.R., Sun, Y.V., Edwards, T.L., et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet 96, 21-36.

48. Hoffmann, T.J., Ehret, G.B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P.Y., Iribarren, C., Chakravarti, A., and Risch, N. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nat Genet 49, 54-64.

49. Liang, J., Le, T.H., Edwards, D.R.V., Tayo, B.O., Gaulton, K.J., Smith, J.A., Lu, Y., Jensen, R.A., Chen, G., Yanek, L.R., et al. (2017). Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations. PLoS Genet 13, e1006728.

50. Aslibekyan, S., Wiener, H.W., Wu, G., Zhi, D., Shrestha, S., de Los Campos, G., and Vazquez, A.I. (2014). Estimating proportions of explained variance: a comparison of whole genome subsets. BMC Proc 8, S102.

51. Doris, P.A. (2011). The genetics of blood pressure and hypertension: the role of rare variation. Cardiovasc Ther 29, 37-45.

52. Cha, S.H., Park, H.S., and Cho, H.J. (2012). Socioeconomic disparities in prevalence, treatment, and control of hypertension in middle-aged Koreans. J Epidemiol 22, 425-432.

53. Talwar, A., Sahni, S., Talwar, A., Kohn, N., and Klinger, J.R. (2016). Socioeconomic status affects pulmonary hypertension disease severity at time of first evaluation. Pulm Circ 6, 191-195.

54. Leng, B., Jin, Y., Li, G., Chen, L., and Jin, N. (2015). Socioeconomic status and hypertension: a meta-analysis. J Hypertens 33, 221-229.

55. Wang, Z., Yue, X., Wang, H., Bao, C., Xu, W., Chen, L., and Qi, X. (2014). Relation of socioeconomic status to hypertension occurrence. Int J Cardiol 173, 544-545.

56. Wu, W.H., Yang, L., Peng, F.H., Yao, J., Zou, L.L., Liu, D., Jiang, X., Li, J., Gao, L., Qu, J.M., et al. (2013). Lower socioeconomic status is associated with worse outcomes in pulmonary arterial hypertension. Am J Respir Crit Care Med 187, 303-310.

57. Vathesatogkit, P., Woodward, M., Tanomsup, S., Hengprasith, B., Aekplakorn, W., Yamwong, S., and Sritara, P. (2012). Long-term effects of socioeconomic status on incident hypertension and progression of blood pressure. J Hypertens 30, 1347-1353.

58. Basson, J., Sung, Y.J., Schwander, K., Kume, R., Simino, J., de las Fuentes, L., and Rao, D. (2014). Gene-education interactions identify novel blood pressure loci in the Framingham Heart Study. American Journal of Hypertension 27, 431-444.

59. Non, A.L., Gravlee, C.C., and Mulligan, C.J. (2012). Education, genetic ancestry, and blood pressure in African Americans and Whites. American Journal of Public Health 102, 1559-1565.

60. Sung, Y.J., de Las Fuentes, L., Schwander, K.L., Simino, J., and Rao, D.C. (2015). Gene-smoking interactions identify several novel blood pressure loci in the Framingham Heart Study. American Journal of Hypertension 28, 343-354.

61. Simino, J., Sung, Y.J., Kume, R., Schwander, K., and Rao, D.C. (2013). Gene-alcohol interactions identify several novel blood pressure loci including a promising locus near SLC16A9. Front Genet 4, 277.

62. Simino, J., Shi, G., Bis, J.C., Chasman, D.I., Ehret, G.B., Gu, X., Guo, X., Hwang, S.J., Sijbrands, E., Smith, A.V., et al. (2014). Gene-age interactions in blood pressure regulation: a large-scale investigation with the CHARGE, Global BPgen, and ICBP Consortia. Am J Hum Genet 95, 24-38.

63. Adler-Milstein, J., DesRoches, C.M., Kralovec, P., Foster, G., Worzala, C., Charles, D., Searcy, T., and Jha, A.K. (2015). Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. Health Aff (Millwood) 34, 2174-2180.

64. Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. The New England Journal of Medicine 372, 793-795.

65. Kohane, I.S. (2011). Using electronic health records to drive discovery in disease genomics. Nature Reviews Genetics 12, 417-428.

66. Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys, D.R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clinical Pharmacology and Therapeutics 84, 362-369.

67. Pulley, J., Clayton, E., Bernard, G.R., Roden, D.M., and Masys, D.R. (2010). Principles of human subjects protections applied in an opt-out, de-identified biobank. Clinical and translational science 3, 42-48.

68. Crawford, D.C., Goodloe, R., Farber-Eger, E., Boston, J., Pendergrass, S.A., Haines, J.L., Ritchie, M.D., and Bush, W.S. (2015). Leveraging Epidemiologic and Clinical Collections for Genomic Studies of Complex Traits. Human Heredity 79, 137-146.

69. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature 538, 161-164.

70. Hall, J.B., Dumitrescu, L., Dilks, H.H., Crawford, D.C., and Bush, W.S. (2014). Accuracy of administratively-assigned ancestry for diverse populations in an electronic medical record-linked biobank. PloS One 9, e99161.

71. Dumitrescu, L., Ritchie, M.D., Brown-Gentry, K., Pulley, J.M., Basford, M., Denny, J.C., Oksenberg, J.R., Roden, D.M., Haines, J.L., and Crawford, D.C. (2010). Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. Genetics in medicine : official journal of the American College of Medical Genetics 12, 648-650.

72. Smith, T.K., J. (2014). Measuring Occupational Presitge on the 2012 General Social Survey. In. (Chicago, GSS Methodological Reports.

73. Link, B.G., and Phelan, J. (1995). Social conditions as fundamental causes of disease. J Health Soc Behav Spec No, 80-94.

74. Bethea, T.N., Palmer, J.R., Rosenberg, L., and Cozier, Y.C. (2016). Neighborhood Socioeconomic Status in Relation to All-Cause, Cancer, and Cardiovascular Mortality in the Black Women's Health Study. Ethnicity & Disease 26, 157-164.

75. Rawshani, A., Svensson, A.M., Zethelius, B., Eliasson, B., Rosengren, A., and Gudbjornsdottir, S. (2016). Association Between Socioeconomic Status and Mortality, Cardiovascular Disease, and Cancer in Patients With Type 2 Diabetes. JAMA Internal Medicine.

76. Shavers, V.L. (2007). Measurement of socioeconomic status in health disparities research. J Natl Med Assoc 99, 1013-1023.

77. Carlson, C.S., Eberle, M.A., Kruglyak, L., and Nickerson, D.A. (2004). Mapping complex disease loci in whole-genome association studies. Nature 429, 446-452.

78. Hollister, B.M., Restrepo, N.A., Farber-Eger, E., Crawford, D.C., Aldrich, M.C., and Non, A. (2016). Development and Performance of Text-Mining Algorithms to Extract Socioeconomic Status from De-Identified Electronic Health Records. Pac Symp Biocomput 22, 230-241.

79. Wiley, L.K., Shah, A., Xu, H., and Bush, W.S. (2013). ICD-9 tobacco use codes are effective identifiers of smoking status. J Am Med Inform Assoc 20, 652-658.

80. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19, 1655-1664.

81. Buyske, S., Wu, Y., Carty, C.L., Cheng, I., Assimes, T.L., Dumitrescu, L., Hindorff, L.A., Mitchell, S., Ambite, J.L., Boerwinkle, E., et al. (2012). Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. PloS One 7, e35651.

82. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8, e1002793.

83. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.

84. Team, R.D.C. (2008). R: A language and environment for statistical computing. In. (Vienna, Austria R Foundation for Statistical Computing.)

85. Spada, J., Scholz, M., Kirsten, H., Hensch, T., Horn, K., Jawinski, P., Ulke, C., Burkhardt, R., Wirkner, K., Loeffler, M., et al. (2016). Genome-wide association analysis of actigraphic sleep phenotypes in the LIFE Adult Study. J Sleep Res 25, 690-701.

86. van der Loos, M.J., Rietveld, C.A., Eklund, N., Koellinger, P.D., Rivadeneira, F., Abecasis, G.R., Ankra-Badu, G.A., Baumeister, S.E., Benjamin, D.J., Biffar, R., et al. (2013). The molecular genetic architecture of self-employment. PloS One 8, e60542.

87. Huang, Y.C., Lin, J.M., Lin, H.J., Chen, C.C., Chen, S.Y., Tsai, C.H., and Tsai, F.J. (2011). Genome-wide association study of diabetic retinopathy in a Taiwanese population. Ophthalmology 118, 642-648.

88. Dick, D.M., Aliev, F., Krueger, R.F., Edwards, A., Agrawal, A., Lynskey, M., Lin, P., Schuckit, M., Hesselbrock, V., Nurnberger, J., Jr., et al. (2011). Genome-wide association study of conduct disorder symptomatology. Mol Psychiatry 16, 800-808.

89. Shyn, S.I., Shi, J., Kraft, J.B., Potash, J.B., Knowles, J.A., Weissman, M.M., Garriock, H.A., Yokoyama, J.S., McGrath, P.J., Peters, E.J., et al. (2011). Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. Mol Psychiatry 16, 202-215.

90. Oh, S., and Oh, S. (2014). Epidemiological and genome-wide association study of gastritis or gastric ulcer in korean populations. Genomics Inform 12, 127-133.

91. Barnett, G.C., Thompson, D., Fachal, L., Kerns, S., Talbot, C., Elliott, R.M., Dorling, L., Coles, C.E., Dearnaley, D.P., Rosenstein, B.S., et al. (2014). A genome wide association study (GWAS) providing evidence of an association between common genetic variants and late radiotherapy toxicity. Radiother Oncol 111, 178-185.

92. Huffman, J.E., Albrecht, E., Teumer, A., Mangino, M., Kapur, K., Johnson, T., Kutalik, Z., Pirastu, N., Pistis, G., Lopez, L.M., et al. (2015). Modulation of genetic associations with serum urate levels by body-mass-index in humans. PloS One 10, e0119752.

93. Berndt, S.I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genet 45, 501-512.

94. Liu, C.T., Monda, K.L., Taylor, K.C., Lange, L., Demerath, E.W., Palmas, W., Wojczynski, M.K., Ellis, J.C., Vitolins, M.Z., Liu, S., et al. (2013). Genome-wide association of body fat distribution in African ancestry populations suggests new loci. PLoS Genet 9, e1003681.

95. Heid, I.M., Jackson, A.U., Randall, J.C., Winkler, T.W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M.C., Speliotes, E.K., Magi, R., et al. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet 42, 949-960.

96. Williams, F.M., Carter, A.M., Hysi, P.G., Surdulescu, G., Hodgkiss, D., Soranzo, N., Traylor, M., Bevan, S., Dichgans, M., Rothwell, P.M., et al. (2013). Ischemic stroke is associated with the ABO locus: the EuroCLOT study. Ann Neurol 73, 16-31.

97. Iyengar, S.K., Sedor, J.R., Freedman, B.I., Kao, W.H., Kretzler, M., Keller, B.J., Abboud, H.E., Adler, S.G., Best, L.G., Bowden, D.W., et al. (2015). Genome-Wide Association and Trans-ethnic Meta-Analysis for Advanced Diabetic Kidney Disease: Family Investigation of Nephropathy and Diabetes (FIND). PLoS Genet 11, e1005352.

98. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature 518, 187-196.

99. Liu, J.Z., Medland, S.E., Wright, M.J., Henders, A.K., Heath, A.C., Madden, P.A., Duncan, A., Montgomery, G.W., Martin, N.G., and McRae, A.F. (2010). Genome-wide association study of height and body mass index in Australian twin families. Twin Res Hum Genet 13, 179-193.

100. Wan, E.S., Cho, M.H., Boutaoui, N., Klanderman, B.J., Sylvia, J.S., Ziniti, J.P., Won, S., Lange, C., Pillai, S.G., Anderson, W.H., et al. (2011). Genome-wide association analysis of body mass in chronic obstructive pulmonary disease. Am J Respir Cell Mol Biol 45, 304-310.

101. Winham, S.J., Cuellar-Barboza, A.B., Oliveros, A., McElroy, S.L., Crow, S., Colby, C., Choi, D.S., Chauhan, M., Frye, M., and Biernacka, J.M. (2014). Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. Mol Psychiatry 19, 1010-1016.

102. Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M., et al. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41, 527-534.

103. Boardman, J.D., Domingue, B.W., Blalock, C.L., Haberstick, B.C., Harris, K.M., and McQueen, M.B. (2014). Is the gene-environment interaction paradigm relevant to genome-wide studies? The case of education and body mass index. Demography 51, 119-139.

104. Williams, D.R., Mohammed, S.A., Leavell, J., and Collins, C. (2010). Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. Ann N Y Acad Sci 1186, 69-101.

105. Franceschini, N., Carty, C.L., Lu, Y., Tao, R., Sung, Y.J., Manichaikul, A., Haessler, J., Fornage, M., Schwander, K., Zubair, N., et al. (2016). Variant Discovery and Fine Mapping of Genetic Loci Associated with Blood Pressure Traits in Hispanics and African Americans. PloS One 11, e0164132.

106. Jones, D.W., Appel, L.J., Sheps, S.G., Roccella, E.J., and Lenfant, C. (2003). Measuring blood pressure accurately: new and persistent challenges. JAMA 289, 1027-1030.

107. (2014). Capturing social and behavioral domains and measures in electronic health records: Phase 2. In. (Washington, D.C. , The National Academies Press.)