2022

The Evolution of Racial and Ethnic Disparities in Health Outcomes

Megan T. Hoang
University of Central Florida

THE EVOLUTION OF RACIAL AND ETHNIC DISPARITIES IN HEALTH OUTCOMES

by

MEGAN HOANG

A thesis submitted in partial fulfillment of the requirements
for the Honors in the Major Program in Economics
in the College of Business Administration
and in the Burnett Honors College
at the University of Central Florida
Orlando, Florida

Spring Term, 2022
Thesis Chair: Melanie Guldi, Ph.D.

Abstract

Health disparities between different racial/ethnic groups in the United States are substantial. When reviewed across an extensive body of literature, these disparities have been demonstrated to persist even when socioeconomic status, geographic region, health conditions, treatment methods, and patient access-related variables are controlled for. This ultimately leads to higher mortality rates among minority patients, making disparities in health a highly prevalent issue. However, the literature suggests that while racial and ethnic disparities in health have been widely examined, research documenting the evolution of these changes over time is lacking. This motivates the research questions: (1) *How has the impact of racial biases on disparities in health outcomes evolved over the past decade?*; (2) *To what extent do race and ethnicity impact variation in health outcomes?*; and (3) *To what extent are race and ethnicity correlated with the socioeconomic gradient in health?;* Last, (4) *How present were these disparities when looking at outcomes related to the COVID-19 Pandemic?* This thesis aims to address these questions through a two-part empirical analysis using publicly available data from the National Health Interview Survey (NHIS) and the COVID-19 Case Surveillance Public Use Dataset from the Centers for Disease Control and Prevention (CDC).

Dedication

For my friends and family, who support me in all I do. You inspire me to stay true to myself, always. I would not be where I am today without you.

Acknowledgements

I would like to express my deepest thanks to my thesis chair, Dr. Melanie Guldi, for her guidance throughout my undergraduate studies. She has been an outstanding resource, and her continuous dedication, encouragement, and support have greatly influenced my academic success.

I would also like to thank my committee, Dr. David Scrogin, for providing insightful feedback throughout my research process.

Contents

List of Tables

List of Figures

List of Acronyms

Acronym	*Definition*
ACE	Angiotensin-Converting Enzyme
CDC	Centers for Disease Control
COVID-19	Coronavirus Disease
ICU	Intensive Care Unit
IOM	Institute of Medicine
IPUMS	Integrated Public Use Microdata Series
NCHS	National Center for Health Statistics
NHIS	National Health Interview Survey
OB/GYN	Obstetrician-Gynecologist
OLS	Ordinary Least Squares
PID	Physician Induced Demand
PROGRESS	Place of Residence, Race/ethnicity, Occupation, Gender, Religion/culture, Education, Socioeconomic status, Social capital/networks
RLM	Robust Linear Model
SARS	Severe Acute Respiratory Syndrome
SES	Socio-economic Status
SQL	Structured Query Language
WHO	World Health Organization

Chapter 1: Introduction

Over the course of history in the United States, racial and ethnic minorities have faced deep-rooted discrimination. Residual effects of racism and prejudice have created persisting effects on the health of minorities. Experiences of racism are associated with adverse effects on mental and physical health, ultimately resulting in poorer health outcomes for minority patients (Paradies et al., 2015). Racial and ethnic disparities in health have been well-documented across a large body of literature, as patients from racial/ethnic minority backgrounds statistically experience worse health outcomes than their non-minority counterparts. Many factors contribute to the widening health disparity gap – at the patient level, variables include patient attitudes, preferences, treatment compliance, and use of healthcare services. At the practitioner level, bias and discrimination – whether overt or implicit – affect the quality of patient care. Finally, at the health systems level, convoluted clinical bureaucracies, administrative processes, and insurance market inefficiencies further contribute to disparities in health outcomes, as patients of racial/ethnic minority backgrounds are less likely to have the resources to navigate these healthcare systems effectively.

When reviewed across an extensive body of literature, these disparities have been demonstrated to persist even when socioeconomic status, geographic region, health conditions, treatment methods, and patient access-related variables are controlled for. These discrepancies ultimately lead to higher mortality rates among minority patients, making disparities in health a highly prevalent issue. The goal of this study is to contribute to this body of literature,

expanding upon current knowledge. As noted in Shavers et al. (2012), while racial and ethnic disparities in health have been widely examined, research documenting the evolution of these changes over time is lacking. This motivates the research questions: (1) *How has the impact of racial biases on disparities in health outcomes evolved over the past decade?*; (2) *To what extent do race and ethnicity impact variation in health outcomes?*; and (3) *To what extent are race and ethnicity correlated with the socioeconomic gradient in health?;* Last, (4) *How present were these disparities when looking at outcomes related to the COVID-19 Pandemic?* This thesis aims to address these questions through a two-part empirical analysis using publicly available data. The first analytical component will be an analysis of data from the National Health Interview Survey (NHIS), pulled from the Integrated Public Use Microdata Series (IPUMS) database through a data extract. The second component will use the COVID-19 Case Surveillance Public Use Dataset from the Centers for Disease Control and Prevention (CDC). To address each of the research questions outlined above, this paper intends to apply a combination of generalized linear and multilevel (or hierarchical) regression models to evaluate the different mechanisms affecting patient health, as well as the levels at which they are associated with health outcomes.[1]

[1] The analysis code discussed in this thesis can be found at: https://github.com/meganthoang/healthdisparities

Chapter 2: Literature Review

Health disparities due to economic inequality and differences in socio-economic status (SES) in the United States are substantial. Much of the literature on this topic illustrates the relationship between various factors and their relation to disparities in patient health. The source of these disparities is rooted in a wide variety of factors and differences between patients and healthcare providers. In its 2003 report, *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care,* the Institute of Medicine (IOM)[2] provides an in-depth review and analysis of a large body of previous publications and gives a comprehensive overview of the various factors that affect health outcomes. Disparities in health care, as stated by the IOM, are defined to be "racial or ethnic differences in the quality of care not due to access" (IOM, 2003, p. 3). Generally speaking, the IOM asserts that racial and ethnic minorities are prone to receiving lower quality health care than non-minorities. Evidence of these disparities is present across a variety of health conditions, regions, and treatment methods. Additionally, "the majority of studies […] find that racial and ethnic disparities remain even after adjustment for socioeconomic differences and other healthcare access-related factors" (IOM, 2003, p. 5). These disparities ultimately lead to higher mortality rates/lower survival rates among minorities, even at equivalent levels of access to care. These differences in health partially stem from variables at the patient level, from patient attitudes, preferences, refusal of treatment, and use of healthcare

[2] This paper heavily relies on conclusions drawn from the IOM's 2003 report throughout the Literature Review chapter, as it provides the most relevant and comprehensive overview of the topic.

services. However, these disparities can also be caused at the practitioner level due to provider beliefs and stereotyping, prejudice, statistical discrimination, and clinical uncertainty (IOM, 2003). A third source of disparities is at the system level, as patients frequently experience difficulties navigating through health systems due to clinical bureaucracies. Practitioners are expected to be reliable agents of patient health; however, they may often be unable to fully aid patients in maneuvering through bureaucracies and administrative processes. Healthcare systems become further convoluted through complex payment systems and insurance market inefficiencies.

Patient-Level Factors

The IOM (2003) suggests several patient-level sources of disparities in health outcomes. Patients of minority backgrounds are more likely to approach care-seeking with negative attitudes, more likely to refuse healthcare services, less likely to adhere to treatment regimens, and simultaneously more likely to delay seeking care. These differences in approach to care are partially due to patient preferences. Patients' preferences regarding their care are directly related to their trust in practitioner authority and advice. However, patients of racial and ethnic minority backgrounds are more likely to mistrust health professionals due to a history of racial discrimination and inferior care. Additionally, negative experiences with physicians and other healthcare professionals can directly influence patient preferences, making them less likely to trust recommendations for more invasive procedures, however necessary. Racial concordance (i.e., when a patient receives care from a same-race provider) may play a role in patient trust, as patients are more likely to feel that their values and expectations for care match with those of their providers. Studies have demonstrated a correlation between racial concordance and

"greater participatory decision-making, greater patient-centered care, lower levels of physician verbal dominance, and greater patient satisfaction," which, consequently, may increase patient compliance with treatment recommendations (IOM, 2003, p. 134). Furthermore, minority patients' utilization of healthcare services statistically differs from that of non-minority patients. White patients are actually more likely to overuse clinical services, which may be due to higher levels of education and access to information, which, in turn, may make them more informed consumers (IOM, 2003).

However, there are also several external factors that can impact health disparities. Biological differences are another patient-level factor that can affect health outcomes and may justify differences in treatment methods. Genetic differences between racial groups can impact the efficacy of therapeutic and pharmacologic treatments. Variability in polymorphic traits such as drug-metabolizing enzymes can affect treatment responses, and in such cases, equal treatment between racial groups can result in differing health outcomes. This may contribute to health differences. However, differences in treatments prescribed across racial groups is still evidenced in regimens that are effective across minority and non-minority populations (IOM, 2003). Additionally, language barriers can contribute to difficulties establishing patient trust. As with patient-physician concordance, language concordance is essential for effective communication between physicians and patients. Lack of effective communication can lead to patient misunderstanding of care, which, in turn, can lead to poor compliance and reduced patient satisfaction. It is evidenced that language mismatches significantly influence patient use of services and clinical outcomes (IOM, 2003).

Considering more recent studies, Speybroeck et al. (2013) explores the most common modeling techniques used in studies examining factors that affect patient health outcomes. In

this study, the authors define eight specific factors that lead to disparities in health – following the PROGRESS acronym, "Place of Residence, Race/ethnicity, Occupation, Gender, Religion/culture, Education, Socioeconomic status, Social capital/networks," (p. 5751). These disparities are generally referred to as the socioeconomic gradient in health. Further research has been conducted regarding these specific factors, providing additional evidence demonstrating the relationship between these factors and health. In an empirical application of the Grossman model, Galama et al. (2018) discusses how having higher levels of SES leads to having a healthier lifestyle. Using the Grossman model's definition of health as a durable capital stock that depreciates over time, it is concluded that having higher levels of education leads to being more efficient consumers and producers of health. Using a Method of Simulated Moments approach to modeling the relationship between income and health also leads to the conclusion that having higher levels of wealth and income leads to increased health investment (Galama et al., 2018). An exploration into the relationship between regional health care utilization and mortality rates found that location accounted for nearly 50% of the difference in health care utilization. However, applications of this study are restricted due to modeling limitations and assumptions, as well as difficulties predicting regional demand for health care (Godøy et al., 2020). In another study, Grönqvist et al. (2012) examines the relationship between income inequality on health outcomes, citing two theories explaining the link between the two. The "strong" income inequality hypothesis states that the inequality between income levels directly affects overall health, regardless of the actual individual levels of income. This hypothesis can be seen through spheres of political influence, where more wealthy individuals have more influence over policies that affect health care. The "strong" theory also asserts that income inequality weakens interpersonal trust societally, which impacts general health through psychological stress.

The second theory, the "weak" income inequality hypothesis, refers to relative income levels. This theory asserts that differences in income levels across a society result in a divide between more advantaged populations and those who are less advantaged. Being disadvantaged relative to the rest of the population adversely impacts individual health (Grönqvist et al., 2012). Buckles et al. (2016) explores the impact of college education on health through a study of draft-avoidance induced college enrollment during the Vietnam War. This study found that increased levels of education decreased mortality rates. However, the study asserted that the effects of increased education may in part be indirectly caused, as increased college education contributes to higher earnings and is correlated with higher rates of health insurance (Buckles et al., 2016). Generally, these studies show a strong correlation between PROGRESS factors and health.

In a 2018 study, Moscelli et al. investigates the idea that patient choice is the primary driver behind health inequalities and is what is affected by patient-level variables, as opposed to prejudice and discrimination. This study examined patient responses to waiting time inequality between two expensive treatments for a severe disease. Patients with varying levels of SES differ in the way they exercise choice – wealthier and more educated individuals are more likely to be willing to travel further for lower wait times and higher quality of care, whereas patients from lower SES backgrounds are more likely to tolerate longer wait times when in the same environment. This is due to a combination of factors, as patients from higher SES backgrounds have fewer financial constraints and limitations. However, patient choice only accounted for 12% of wait time inequalities, which still leaves a statistically significant SES gradient not caused by patient choice.

Practitioner-Level Factors

Another potential source of health disparities, as identified by the IOM (2003), are variables at the practitioner level. In the clinical setting, physicians and other healthcare providers are required to make medical decisions under many constraints. Limited supply of health resources and practitioner availability can lead to mismatches with patient demand for health services. Excess demand and patient competition for healthcare services can lead to extensive queues and wait times. Additionally, cost-containment incentives can lead to more frugal practices, potentially exacerbating supply-demand mismatches. Physicians and other healthcare providers also often face limited time constraints and must evaluate a large amount of information, both from patient disclosure and diagnostic testing. Operating under these constraints, they need to quickly make decisions regarding a patient's care with information that may be inaccurate or incomplete. This creates a level of clinical uncertainty and ambiguity in the decision-making process, which allows subjectivity and provider bias to influence clinical decisions. As stated by the IOM (2003), "Under conditions of time pressure, problem complexity, and high cognitive demand, physicians' attitudes may therefore shape their interpretation of this information and their expectations for treatment" (p. 161). In such cases, provider prejudice or bias can negatively impact patient care.

Socially, humans are inclined towards classifying others into categories, often on the basis of race, gender, or age. These categories allow us to form stereotypes – a heuristic method of forming judgements about others based on their categorization. This method of classification leads to the concept of "group membership," in which specific groups are deemed to generally exhibit certain characteristics. Even in cases where practitioners believe they are unbiased and negative attitudes are not overtly expressed, they can still affect patient care. This set of clinical

heuristics is one mechanism through which physician beliefs affect patient care - through provider beliefs and stereotypes. Social stereotypes tend to be systemically biased and can unconsciously affect physician judgements during clinical interactions, even among practitioners who believe they are not prejudiced. In these cases, regardless of intent, clinical heuristics can negatively impact patient care as providers may use stereotypes to make assumptions about patient compliance or a patient's ability to afford treatment solely based on their race or ethnicity.

Classification can lead to prejudice and bias in cases where negative attitudes regarding an individual's group membership are the sole reason for differential treatment. Prejudice, in this context, refers to taste-based differences in treatment for patients on the basis of race or ethnicity. There is substantial evidence supporting bias or prejudiced attitudes among healthcare providers, whether conscious or unconscious. Schulman et al. (1999) used Black and White actors to present a set of symptoms and characteristics in scripted interviews with healthcare providers. With all other variables controlled for, they found that physicians were less likely to recommend the same treatment for Black patients as for White patients. The study concluded that the race of the patient influenced the treatment method a physician was likely to recommend, and that physician diagnostic and treatment decisions may be influenced by racial biases.

As with patient-level factors, several external factors can impact the clinical decision-making process. Biological differences present across different racial and ethnic groups can justify differences in treatments and diagnostic methods recommended by a physician (IOM, 2003). Burroughs et al. (2002) documents a large body of pharmacological studies indicating differences in responses to pharmaceutical treatments across different racial and ethnic groups.

Genetic polymorphisms can influence the way certain medications are metabolized in the body, which affects their therapeutic effects. For instance, ACE inhibitors like lisinopril and enalapril are more effective for treating hypertension and heart failure in Whites than in African Americans. In contrast, diuretic medications like hydrochlorothiazide have greater antihypertensive effects in African Americans than other racial groups. Physicians should generally be aware of variation in drug responses across different racial and ethnic groups and take those differences into account when making prescription and dosage recommendations (Burroughs et al., 2002). Additionally, language barriers between physicians and patients can complicate the clinical encounter, making it more difficult for physicians to establish patient trust or understand patient needs. This adds a layer of uncertainty to clinical decisions made by physicians and healthcare providers, which can lead to misdiagnoses, improper treatment, or further miscommunications between practitioners and patients. Language discordance can also affect healthcare practitioners' ability to adequately understand patient needs or properly communicate details of patient care (including treatment regimens and potential side effects). This can contribute to differences in patient treatment across different racial and ethnic groups in the clinical setting, which, in turn, leads to patient mistrust of healthcare systems.

Regarding more recent research, many studies exploring discrimination in health care also examine the effects of PROGRESS factors, as defined in Speybroeck et al. (2013). Johar et al. (2013) examined patient wait times for non-emergency treatments. This study was conducted using data from public hospitals in Australia, where a universal health system is in place. Even without financial incentives for the hospital (as all patients were non-paying), the study found that patients with higher SES experienced shorter wait times before receiving treatment than lower SES patients. This effect was evidenced across the entire wait time distribution, and the

study found that high SES patients were consistently prioritized over low SES patients, even among patients requiring urgent care. Balsa et al. (2003) analyzes the clinical encounter and proposes three potential mechanisms through which disparities in health can be created during clinical encounters and how race/ethnicity plays a role. The first is through physician prejudice and a physician's preference towards non-minority patients. The second is through clinical uncertainty and interpretation of patient symptoms. The third is through physician-held stereotypes regarding minority health-related behavior.

Discrimination in healthcare settings can create barriers to care for minority groups, leading members of these groups to forgo seeking care, despite necessity. In Rivenbark (2020), analysis of a French nationally conducted survey found a positive correlation between experiences of discriminatory care in health settings and a patient's likelihood to forgo necessary care. These trends were observed across groups who were socially disadvantaged due to gender, race or ethnicity, religion, or immigration status (Rivenbark, 2020). Discrimination has also been demonstrated to lead to adverse effects in health. Kim (2013) conducted a study of Asian Americans and concluded that experiences of racial discrimination contribute to higher stress and increased depressive symptoms. Additionally, discriminatory practices are evidenced to result in significant disparities in the physical and psychological health of African American patients, especially when contrasted with non-Hispanic Whites (Kim, 2013).

However, while discrimination and racism are prevalent issues, there remains a large gap in the research in this area. In a 2012 article, Shavers et al. conducted a review of recent literature studying the effects of racism and discrimination in healthcare settings. This study found that although implicit bias and discrimination in healthcare settings are frequently examined, few studies have researched the overall prevalence of the issue or examined the

changes in these trends over time. Furthermore, many studies rely on survey data on patient perceptions of discrimination. Shavers et al. (2012) notes that aspects of the clinical encounter that drive these perceptions should be more systematically examined. Overall, "there is a continuing need for innovative methodology, better instrumentation, and strategies for identifying racial/ethnic and other types of discrimination in healthcare settings, particularly because of the somewhat subjective manner in which health care is delivered" (Shavers et al., 2012, p. 963).

System-Level Factors

The final category of factors that impact the socioeconomic gradient in health exists at the health systems level. These factors exist due to the way health systems are structured. According to the IOM (2003), "aspects of health systems—such as the ways in which systems are organized and financed, and the "ease" of accessing services—may exert different effects on patient care, particularly for racial and ethnic minorities" (p. 140). Regional variation in health service availability also contributes to disparities in healthcare access as well as differences in care received. For instance, minority patients are more likely to live in areas with physician shortages, limiting health service accessibility. These geographical factors impact racial and ethnic minorities differently, further widening the health disparity gap. Furthermore, navigation of complex health systems can present challenges for patients with low English proficiency, limiting access to care. English proficiency in the United States is especially limited among certain racial and ethnic groups, which can complicate patient-practitioner interactions during clinical encounters. Linguistic discordance has been evidenced to negatively impact health

outcomes, as ineffective communication and misunderstandings between physicians and patients lead to clinical uncertainty, misdiagnoses, and poor patient compliance, among other issues.

While health systems attempt to address language barriers, assisted communication services are limited. Patients attempting to navigate health systems with low English proficiency often encounter situations in which they are either required to provide their own interpreter (via family members or friends), or in which care is denied or delayed due to lack of available interpreter services. Additionally, a lack of standardization across interpreter services may contribute to patient miscommunications (i.e., a family member may not have the appropriate medical vocabulary to communicate patient symptoms or translate physician orders). Failure to provide effective assisted communication services can lead to severe consequences, and in some cases, death. The IOM (2003) asserts that due to barriers to access, minority patients are less likely than White patients to have a regular healthcare provider, and lack of consistent care can affect medical follow-up and reduce the likelihood of referral to specialty care. Lack of consistent care also leads to incomplete health information, which limits physician ability to provide comprehensive assessments of patient health.

Dramatic changes in health systems over the years, as well as the changes in delivery of care, disproportionately affect racial and ethnic minority groups. The evolution of healthcare policies and regulations throughout the years has further convoluted the clinical bureaucracy, making health systems a complex "maze" that is difficult for patients to navigate. Physicians and other healthcare practitioners are essential agents in assisting patients as they navigate these systems (IOM, 2003). However, financial incentives and time constraints can limit a practitioners' ability to advocate for patient health. Practitioner advocacy can be adversely affected by the practitioner-level variables previously discussed (i.e., clinical uncertainty,

stereotyping, language barriers, etc.), but these effects can be further influenced by external pressures. Gruber and Owings (1994) explore how financial incentives lead to Physician-Induced Demand (PID), a situation in which physicians prescribe treatments where they are not needed, thus inducing unnecessary demand. The study examines an exogenous change in the 1970s, where declining fertility rates created external financial pressure for OB/GYNs. A strong correlation was found between the decrease in fertility and increase in c-section deliveries during this time period, insinuating that physicians used their agency relationship with patients to prescribe more expensive treatments for financial gain. Instances of PID as shown in this study have the potential to disproportionately affect patients with racial/ethnic minority backgrounds, as they are less likely to have access to knowledge and resources to question the necessity of treatments prescribed. Due to this, effects of PID may be more severe for underrepresented groups.

The complexity of health payment and health insurance systems also contributes to disparities in patient care. Patients may encounter difficulties with fragmentation of healthcare systems, as different levels of insurance coverage affect the range of services available to patients. In low-coverage policies, patients may experience greater constraints in provider choice as well as service coverage. These differences mean that health systems are segmented into different sectors based on patient wealth and coverage levels. Minority patient groups are disproportionately more likely to hold less expensive/more restrictive plans and thus receive lower levels of care, furthering racial disparities in health (IOM, 2003).

In recent years, several studies have illustrated that while health policies have evolved over the years, the health systems landscape is still extremely complex, and many of the same issues described by the IOM in 2003 still remain. Angerer et al. (2019) explores the relationship

between patient SES and access to care. The study conducted an experiment in which patients of varying levels of education requested appointments with physicians across Austria. Patients with a university degree experienced substantially shorter response times and waiting times when requesting appoints with physicians. Results concluded that discrimination in healthcare access exists based on patient SES, and the study asserts that this statistical disparity is created by financial incentives for physicians (Angerer et al., 2019). In addition to socioeconomic differences, language barriers and linguistic discordance between health systems and patients also creates further disparities in care. Dillender (2017) examines how patient level of English proficiency affects access to insurance coverage and health services among immigrant populations. The study finds that higher levels of English proficiency led to a higher likelihood of access to employer-sponsored health insurance. The remaining patients who do not have employer-sponsored health insurance may have access to Medicaid coverage, however, immigrants with poor English proficiency are more likely to be entirely uninsured. The study further examines responses to Medicaid expansions and finds that among patients who satisfy Medicaid income requirements, immigrants with the lowest levels of English proficiency receive the lowest levels of coverage, suggesting that coverage effects are not entirely caused by income differences (Dillender, 2017).

Additionally, in recent years, financial incentives have increasingly motivated changes in patient care. In the American fee-for-service healthcare reimbursement model, payment and compensation are progressively becoming prioritized over patient health. Health services are being offered at an increasing rate, regardless of their benefit to patient health, causing the healthcare industry to become exceedingly profit driven. Consequently, PID is a prevalent issue, as physicians are inducing demand for healthcare services, incentivized by higher compensation

levels (Vedantam, 2020). Doyle et al. (2010) conducted an experimental comparison of two sets of physicians from different medical programs: one group was affiliated with a highly ranked medical school and the other was affiliated with a lower ranked institution. The study found that physicians from the highly ranked group incurred significantly lower costs due to reduced diagnostic testing and concluded that physicians have substantial impact on medical costs. The potential for significant physician impact on healthcare costs, combined with external incentives for medical providers to induce demand for healthcare services, explains the rising cost of healthcare in past years. Higher costs of health care disproportionately affect racial and ethnic minority groups, potentially widening the socioeconomic gradient of health.

Insurance markets create an additional layer of complexity to healthcare systems, as information asymmetry within these markets can potentially lead to adverse selection or moral hazard, both of which contribute to market inefficiencies. As stated in Gruber (2017), insurance markets frequently face risk of adverse selection – where patients with higher health risks are more likely to enroll in health plans with higher coverage. However, shifts in insurance provisions to address this problem and an overabundance of choice can lead to many issues. While a variety of insurance coverage plans creates more consumer choice, having too many options can lead to "choice inconsistencies" – where many consumers select insurance coverage plans that do not reflect their preferences (Gruber, 2017). Another study attempts to explore the cause behind these suboptimal consumer choices. Ericson et al. (2017) asserts that the majority of people purchasing insurance plans have limited knowledge of insurance markets or available coverage options, have limited knowledge of potential future medical expenses, and can often become overwhelmed when faced with an overabundance of options. They find that this consumer confusion may lead to selection of suboptimal coverage plans, which, in turn, creates

allocative inefficiencies and increases market volatility. The overabundance of insurance choices noted in both studies exacerbates the fragmentation of health systems previously discussed, further contributing to inequalities in health.

Historically, disparities in health have been evidenced across varying levels of SES and across different racial/ethnic groups, contributing to the socioeconomic gradient in health. Statistically, patients from racial and ethnic backgrounds receive a lower quality of health care than non-minority patients. Health disparities are evidenced across health conditions, geographic location, and various treatment methods, and have been demonstrated to persist even after adjustment for differences in socioeconomic status and access-related factors. These discrepancies have severe long-term consequences, leading to higher mortality rates among minorities. This topic has been widely researched across a wide variety of variables – for the purposes of this chapter, they have been organized into Patient-Level, Practitioner-Level, and System-Level Factors. Examining these factors through a review of literature from a more recent period has indicated that despite cultural changes in society and policy changes across healthcare systems, disparities in health remain. However, while there is a large body of literature examining racial and ethnic disparities in health, there remain significant gaps in the existing knowledge. As aforementioned, Shavers et al. (2012) notes that research exploring overall prevalence and examining changes in these trends over time is lacking. This reveals a few areas requiring additional investigation. Further research should not only examine the state of racial and ethnic disparities in recent years but should also consider the evolution of these disparities over time.

Chapter 2.1: Differential Effects of COVID-19 by Race/Ethnicity

In early 2020, the World Health Organization (WHO) declared a global public health emergency regarding the rapid spread of the SARS-CoV-2 virus. In the following months, the spread of the virus escalated, and the number of cases worldwide skyrocketed, leading to the emergence of a global pandemic. By the end of 2020, the number of cases of coronavirus (COVID-19) reported in the United States exceeded 20 million, resulting in over 346,000 deaths (AJMC, 2021). This led to drastic changes in healthcare systems in the United States, presenting novel challenges for healthcare facilities and practitioners across the nation. Several studies have since shown the COVID-19 crisis to have disproportionate effects on racial/ethnic minority groups. In an examination of COVID-19 infections, hospitalizations, and deaths by race/ethnicity, Mackey et al. (2021) stated that the COVID-19 pandemic had differential effects between varying racial and ethnic groups, with minority groups being more heavily impacted by the pandemic. In early 2020, Azar et al. (2020) performed a retrospective cohort analysis of COVID-19 patients and found that African American patients were significantly more likely to be hospitalized due to severe symptom onset than their White counterparts, even after adjustment for age, sex, comorbidities, and income level. In a later study, Miller et al. (2021) examined changes in mortality rates caused by the COVID-19 pandemic, and found that patient race/ethnicity, occupation, insurance coverage, and income level all affected the level of increase in mortality. The presence of these disparities motivates an additional component to the topic investigated in this paper, to give insight into the effect of the pandemic on trends between patient race/ethnicity and health outcomes.

Chapter 3: Data

The empirical analysis for this study was conducted in two sections. The first component aims to address research questions (1) through (3) – *How has the impact of racial biases on disparities in health outcomes evolved over the past decade?*; *To what extent do race and ethnicity impact variation in health outcomes?*; and *To what extent are race and ethnicity correlated with the socioeconomic gradient in health?* – by conducting a set of regressions examining the relationship between patient characteristic variables and reported health, using a dataset from the National Health Interview Survey (NHIS). The second component aims to address research question (4) – *How present were these disparities when looking at outcomes related to the COVID-19 Pandemic?* To directly examine the impact of racial and ethnic disparities on health throughout the pandemic, this component of the empirical analysis was conducted using data from the COVID-19 Case Surveillance Public Use Dataset from the Centers for Disease Control and Prevention (CDC).

Variable Selection for the NHIS Dataset

In the first component of the analysis for this study, data from the NHIS was pulled from the Integrated Public Use Microdata Series (IPUMS) database through a data extract, selecting the variables detailed in *Table 3.1* below over the time period from 2010 to 2020. These variables were selected based on the PROGRESS acronym, which is denoted in Speybroeck, (2013) to represent "Place of Residence, Race/ethnicity, Occupation, Gender, Religion/culture,

Education, Socioeconomic status, Social capital/networks" (p. 5751). The corresponding variables selected were *region, race, hispeth, age, sex, edu,* and *incfam07on.* An additional variable, *intervlang*, was selected based on literature supporting a correlation between patient-physician language concordance and health outcomes (IOM, 2003).

Table 3.1 – Variables Selected from the NHIS

Variable Name	*Description*
health	**Health status** – rates an individual's general health (as self-reported by the person in question or evaluated by a family member). The scale ranges from 1 = "Excellent" to 5 = "Poor."
year	**Survey Year** – YEAR is a four-digit variable reporting the calendar year (e.g., 2003) the survey was conducted and the data were collected. YEAR indicates the survey year reported on the household record.
region	**Region of Residence** – reports the region of the U.S. where the housing unit containing survey participants was located.
age	**Patient Age** – reports the individual's age, in years since their last birthday.
sex	**Patient Sex** – indicates whether the person was male or female.
race	**Main Racial Background** (Pre-1997 Revised OMB Standards), self-reported or interviewer reported
hispeth	**Hispanic ethnicity** – identifies and classifies persons of Hispanic/Spanish/Latino origin or ancestry.
intervlang	**Language of interview** – reports the language in which the interview was conducted.
edu	**Educational attainment** – reports the highest level of schooling an individual had completed, in terms of completed grades for persons with less than a high school degree, and in terms of degrees attained for high school graduates and those with higher education.
pooryn	**Above or below poverty threshold** – indicates whether family income was above or below poverty level.
incfam07on	**Total combined family income** (2007+) – provides total grouped family income using an income bracket methodology introduced in 2007.

Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*

The variables highlighted in blue in *Table 3.1* above, (*'health'* and *'mortstat'*), are indicators of individual health and mortality. The other variables, (*age, sex, race, hispeth, intervlang, pooryn,* and *incfam07on*), represent individual characteristics. In all models evaluated in this study, health

outcomes are the dependent variables, while individual characteristics are the independent variables. The dependent variable examined in this study is *'health'* – a categorical ranking variable which operates on a five-point Likert scale (1 – Excellent, 2 – Very Good, 3 – Good, 4 – Fair, 5 – Poor). The primary independent variable examined in this study is racial/ethnic background, *'race'*.

In 2019, the structure and content of the NHIS survey were significantly redesigned to reduce survey length and implement more modern survey methodologies. This set of changes impacted the data for this study in several ways: (1) the *'race'* variable was adjusted to include categories for "Asian" and "American Indian/Alaskan Native". (2) the *'intervlang'* and *'pooryn'* variables are no longer available. (3) the way in which the *'health'* variable was measured also changed slightly. Due to these adjustments, the National Center for Health Statistics (NCHS) advises against pooling data from before and after the survey change, so for the purposes of this study, separate analyses are conducted for the 2010-2018 and 2019-2020 panels.

The data for this component was preprocessed using the *'sqlite3'* module in Python – the data was read and loaded into a Structured Query Language (SQL) database, to simplify data access through database querying. This also provided an added benefit of reduced data size, allowing for faster processing and more optimized performance. The database was then queried to select only the variables relevant for analysis, outlined in *Table 3.1* above. Additionally, referencing the NHIS codebook, observations in the dataset associated with "Unknown" or "Missing" values were removed, for simplicity. Summary statistics for the subset of data used for this project are listed in *Tables 3.2 and 3.3* below. The demographic composition of this dataset was predominantly White and English-speaking. Data across regions, sex, education, and income levels appears to be more evenly distributed.

Table 3.2 – Summary Statistics, NHIS Panel A (2010-2018)

	year	region	age	sex	race	hispeth	lang	edu	pooryn	famincome	health
count	772996	772996	772996	772996	772996	772996	772996	772996	772996	772996	772996
mean	2013.766	2.739	37.210	0.516	137.466	0.209	0.053	247.479	0.165	17.721	2.126
std	2.454	1.029	22.753	0.499	81.565	0.407	0.225	133.637	0.371	8.043	1.056
min	2010	1	0	0	100	0	0	100	0	10	1
25%	2012	2	17	0	100	0	0	100	0	10	1
50%	2014	3	36	1	100	0	0	200	0	20	2
75%	2016	4	55	1	100	0	0	300	0	20	3
max	2018	4	85	1	400	1	1	600	1	30	5

Table 3.3 – Summary Statistics, NHIS Panel B (2019-2020)

	year	region	age	sex	race	hispeth	lang	edu	pooryn	famincome	health
count	72349	72349	72349	72349	72349	72349	— —	72349	— —	72349	72349
mean	2019.477	2.648	46.809	0.530	133.414	0.096	— —	317.998	— —	19.366	2.220
std	0.499	1.018	44.029	0.499	78.496	0.294	— —	124.450	— —	8.288	1.076
min	2019	1	0	0	100	0	— —	100	— —	10	1
25%	2019	2	27	0	100	0	— —	200	— —	10	1
50%	2019	3	48	1	100	0	— —	300	— —	20	2
75%	2020	3	65	1	100	0	— —	400	— —	30	3
max	2020	4	85	1	400	1	— —	600	— —	30	5

Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Notes: Observations with missing values were omitted, for simplicity. Prior to omitting "unknown" or "missing" values, there were 947024 observations in the dataset. After omission, there were 772996 observations in Panel A (2010-2018), and 72349 observations in Panel B (2019-2020).

Data Processing for the NHIS Dataset

Due to the categorical nature of the data, the variables outlined in *Table 3.1* were converted to binary indicator (dummy) variables with true/false values to represent multiple groups within each regression. In the creation of the dummy variables, one category was omitted in each group to avoid the "dummy variable trap" – a situation in which the independent or exogenous variables in a model are perfectly multicollinear. The variables created in each category are outlined in *Table 3.4 – Dummy Variables, NHIS*, below. The *'Dataset Values'* column delineates an abridged summary of the values in the dataset, as well as their corresponding definitions, as stated in the NHIS codebook file. These categories were then grouped into distinct individual variables with 1 or 0 values indicating the categorical presence as True or False. Summary statistics were then output for the dummy variables and compared to the initial variable set to verify accuracy and ensure the variable assignment process did not alter the data.

Table 3.4 – Dummy Variables, NHIS

Variable	*Dataset Values*		*Dummy Variables*
region	01	Northeast	northeast
	02	North Central/Midwest	midwest
	03	South	south
	04	West	west (omitted)
	08	NO DATA IN ROUND	
	09	Unknown	
sex	1	Male	female
	2	Female	male (omitted)
	7	Unknown-refused	
	8	Unknown-not ascertained	
	9	Unknown-don't know	
race	100	White	white
	200	Black/African-American	black
	300-390	Alaskan Native/American Indian	native
	400-490	Asian or Pacific Islander	asian
	500-590	Other Race	other/mixed (omitted)
	600-690	Multiple Race, No Primary	
	900	Unknown	
	970	Unknown-refused	
	980	Unknown-not ascertained	

Variable	*Dataset Values*		*Dummy Variables*
hispeth	10	Not Hispanic/Spanish origin	hisp
	20-70	Hispanic/Latino/Spanish	non-hispanic
	90	Unknown	(omitted)
	91	Unknown if Hispanic/Spanish origin	
	92	Two origins	
	93	Origin unknown	
	99	NIU	
intervlang	1	English	english
	2	Spanish	spanish
	3	English and Spanish	other (omitted)
	4	Other	
	8	Unknown-not ascertained	
	9	Inapplicable	
edu	000	NIU	highschool
	100-116	Grade 12 or less	somecollege
	200-202	High school diploma or GED	collegedegree
	300-303	Some college, no 4yr degree	(omitted)
	400	Bachelor's degree (BA,AB,BS,BBA)	
	500	Master's, Professional, or Doctoral	
	501	Master's degree (MA,MS,Med,MBA)	
	502	Professional (MD,DDS,DVM,JD)	
	503	Doctoral degree (PhD, EdD)	
	504	Other degree	
	996	No degree, years of education unknown	
	997	Unknown--refused	
	998	Unknown--not ascertained	
	999	Unknown--don't know	
pooryn	1	At or above poverty threshold	poor
	2	Below poverty threshold	not poor (omitted)
	9	Unk (1997+: incl. Undefined)	
faminc	10	$0 - $49,999	lowinc (< 50,000)
	11	$0 - $34,999	midinc (between
	12	$35,000 - $49,999	50,000 & 100,000)
	20	$50,000 and over	highinc (> 100,000)
	21	$50,000 - $99,999	(omitted)
	22	$50,000 - $74,999	
	23	$75,000 - $99,999	
	24	$100,000 and over	
	96	Undefined	
	99	Unknown	
health	1	Excellent	Response/endogenous
	2	Very Good	variable, so no dummy
	3	Good	variables created
	4	Fair	
	5	Poor	

Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Notes: Variables denoted as "omitted" (in the *Dummy Variables* column) designate which dummy variables are omitted from the regression equation(s). Dummy values were designated as 1 to indicate categorical presence of the variable, and 0 to indicate a lack of categorical presence.

Variable Selection for the CDC Dataset

The second component of the analysis aims to examine differential effects on health from the coronavirus pandemic. In early 2020, the emergence of COVID-19 drastically impacted health systems in the United States. To examine the impact of racial and ethnic disparities on health throughout the pandemic, this component of the empirical analysis will be conducted using data from the COVID-19 Case Surveillance Public Use Dataset from the Centers for Disease Control and Prevention (CDC). This aggregate dataset includes the variables defined in *Table 3.4* below.

Table 3.5 – Variables from the COVID-19 Case Surveillance Public Use Data

Variable Name	*Definition*
hosp_yn	Hospitalization status
icu_yn	ICU admission status
death_yn	Death status
cdc_case_earliest_dt	The earlier of the Clinical Date (date related to the illness or specimen collection) or the Date Received by CDC
cdc_report_dt	Date case was first reported to the CDC
pos_spec_dt	Date of first positive specimen collection
onset_dt	Symptom onset date, if symptomatic
current_status	Case status
sex	Patient Sex
age_group	Age Group: 0 - 9 Years; 10 - 19 Years; 20 - 39 Years; 40 - 49 Years; 50 - 59 Years; 60 - 69 Years; 70 - 79 Years; 80 + Years
race_ethnicity_combined	Race and ethnicity (combined)
medcond_yn	Presence of underlying comorbidity or disease

Source: CDC COVID-19 Case Surveillance Public Use Data *(Lee, 2021)*

The dependent variables, highlighted above in blue, are hospitalization status, and ICU admission status, and death status or mortality (denoted as *hosp_yn, icu_yn,* and *death_yn,* respectively). The independent variables are sex, age group, race/ethnicity, and presence of underlying comorbidities (denoted as *sex, age_group, race_ethnicity_combined,* and *medcond_yn, respectively*).

Table 3.6 – Summary Statistics, CDC

	status	female	male	child	youth	adult	senior	white	black
count	833640	833640	833640	833640	833640	833640	833640	833640	833640
mean	1	0.531	0.468	0.043	0.113	0.281	0.198	0.624	0.104
std	0	0.499	0.499	0.204	0.316	0.449	0.398	0.484	0.306
min	1	0	0	0	0	0	0	0	0
25%	1	0	0	0	0	0	0	0	0
50%	1	1	0	0	0	0	0	1	0
75%	1	1	1	0	0	1	0	1	0
max	1	1	1	1	1	1	1	1	1

(continued)	hisp	native	asian	hosp	icu	death	medcond
count	833640	833640	833640	833640	833640	833640	833640
mean	0.192	0.005	0.035	0.174	0.057	0.058	0.442
std	0.394	0.067	0.185	0.379	0.232	0.235	0.496
min	0	0	0	0	0	0	0
25%	0	0	0	0	0	0	0
50%	0	0	0	0	0	0	0
75%	0	0	0	0	0	0	1
max	1	1	1	1	1	1	1

Source: CDC COVID-19 Case Surveillance Public Use Data *(Lee, 2021)*
Notes: Values of 1 in the table above indicate categorical presence of each variable, while 0 values indicate a lack of categorical presence. Observations with missing values were omitted, for simplicity. After omitting "unknown" or "missing" values, there were 833640 remaining observations in the dataset.

Data Processing for the CDC Dataset

Similar to the processing of the NHIS dataset, the categorical variables in the COVID-19 Case Surveillance Public Use Dataset (outlined in outlined in *Table 3.5)* were converted to binary indicator (dummy) variables with true/false values indicating categorical presence. The variables created in each category are outlined in *Table 3.7 – Dummy Variables, CDC*, below. The *'Dataset Values'* column delineates an abridged summary of the values in the dataset, as well as their corresponding definitions. These categories were then grouped into distinct individual variables with 1 or 0 values indicating the categorical presence as True or False, omitting one variable in each group to avoid the "dummy variable trap". Summary statistics were then output for the dummy variables and compared to the initial variable set to verify accuracy.

Table 3.7 – Dummy Variables, CDC

Variable	*Dataset Values*	*Dummy Variables*
status	Laboratory-confirmed case	confirmed
	Probable case	probable (omitted)
sex	Male	female
	Female	male (omitted)
age	0 - 9 Years	Child (0-9)
	10 - 19 Years	Youth (10-19)
	20 – 39 Years	Adult (20-59)
	40 – 49 Years	(omitted)
	50 – 59 Years	Senior (60+)
	60 – 69 Years	
	70 – 79 Years	
	80 +	
race	White, Non-Hispanic; Hispanic/Latino	white (omitted)
	Black, Non-Hispanic	black
	Hispanic/Latino	hisp
	American Indian/Alaska Native, Non-Hispanic	native
	Asian, Non-Hispanic	asian
	Native Hawaiian/Other Pacific Islander, Non-Hispanic	
hosp	Yes (hospital admittance)	hosp
	No (no hospital admittance)	nohosp (omitted)
icu	Yes (ICU admittance)	icu
	No (no ICU admittance)	noicu (omitted)
death	Yes (patient reported dead)	death
	No (patient reported alive)	

Source: CDC COVID-19 Case Surveillance Public Use Data *(Lee, 2021)*
Notes: Variables denoted as "omitted" (in the *Dummy Variables* column) designate which dummy variables are omitted from the regression equation(s).

Chapter 4: Methods

Model Specification

This study applies several Huber Robust Linear Models (RLMs) using the RLM function from the *'statsmodels'* module in Python to account for heteroskedasticity and potential outliers in the data. As healthcare systems are inherently complex and affected by a wide range of variables, applying several different models would best allow for the identification of different mechanisms that affect health outcomes. Different model specifications are applied to answer each of the research questions previously stated.

To address the first research question, *(1) "How has the impact of racial biases on disparities in health outcomes evolved over the past decade?"*, this study visually examines a series of bar charts to examine changes in the mean health rating for various population groups over the 2010-2020 time period, as given by the NHIS dataset.

To address the second research question, *(2) "To what extent do race and ethnicity impact variation in health outcomes?"*, this study examines the relationship between the independent variable of individual attributes (i.e., '*race*') and the dependent variable, health outcome (i.e., *'health'*). RLM regressions in the form of (eq. 1) below were fitted between the two to weigh the impact of patient race/ethnicity on health outcomes. Due to the restructuring of the NHIS survey in 2019, separate analyses were conducted for Panel A (2010-2018) and Panel B (2019-2020).

(eq. 1) $$health = \alpha + \beta_1\, white + \beta_2\, black + \beta_3\, hisp + \beta_4\, asian + \beta_5\, native + \varepsilon$$

To address the third question, *(3) "To what extent are race and ethnicity correlated with the socioeconomic gradient in health?"*, this study evaluates the impact of various individual characteristics on health outcomes. The covariates for this model were patient characteristics selected based on the PROGRESS acronym (as defined in Speybroeck, 2013). Independent variables included in this model were *region, race, age, sex, hispeth, intervlang, edu,* and *incfam07on*. This segment also implements a set of RLM regressions in the form of (eq. 2) below, which are then compared to the initial regression in a descriptive analysis to evaluate the strength of correlation between the two.

(eq. 2) $$health = \alpha + \beta_1\, lowinc + \beta_2\, midinc + \beta_3\, highschool + \beta_4\, somecollege + \varepsilon$$

An additional component was then conducted to answer the fourth question, *(4) "How present were these disparities when looking at outcomes related to the COVID-19 Pandemic?,"* incorporating the CDC's COVID-19 Case Surveillance Public Use Dataset. Similar to the methodology used to address research question *(1)*, a series of bar charts were created to examine the number of reported cases, hospitalizations, ICU admittances, and deaths in each population group.[3]

[3] The analysis code discussed in this chapter can be found at: https://github.com/meganthoang/healthdisparities

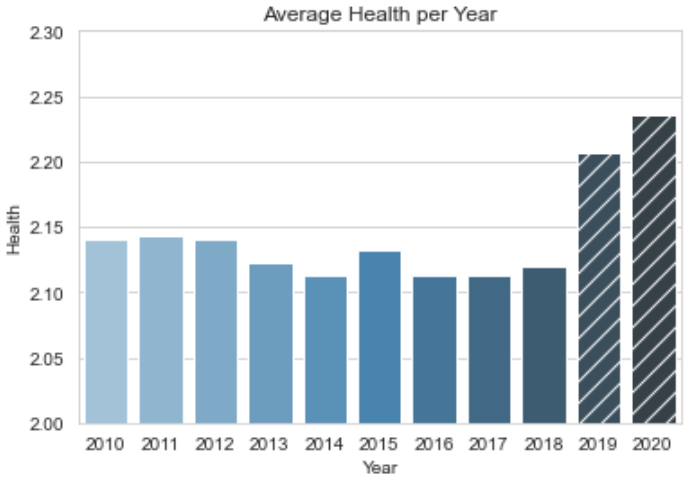
Chapter 5: Results

As discussed in *Chapter 4: Methods*, a series of visualizations were developed and a set of Robust Linear Models (RLMs) were estimated using different variables selected from the National Health Interview Survey (NHIS). Each of the models examines the relationship between the endogenous response variable, *'health,'* and different combinations of exogenous individual characteristic variables. Due to the NHIS survey redesign noted in *Chapter 3: Data*, separate analyses are conducted for Panel A (2010-2018) and Panel B (2019-2020), as it would otherwise be difficult to distinguish between actual trends in health and differences due to changes in data collection.

To address Research Question *(1),* "*How has the impact of racial biases on disparities in health outcomes evolved over the past decade?*," the following bar charts illustrate Average Health Rating (the mean of the *'health'* variable) for each year in the 2010-2020 range, broken down into categories by individual variables. Due to the NHIS survey change in 2019, the data for 2010-2018 cannot be directly compared to the data from 2019-2020. The results for the years affected by the survey change are indicated in *Figures 5.1 – 5.7* below with the '//' hatching pattern. Additionally, the *'health'* response variable is measured on a 5-point Likert scale where a value of 1 corresponds to excellent health and a value of 5 corresponds to poor health, so 'larger' bars indicate worse health, while 'shorter' bars indicate better health. Bar charts were created for the following categories: region, gender, race, interview language, education, and income.

Figure 5.1 below displays average health per year to provide a general overview of health trends across all groups. From 2010-2018, the average health rating is observed to generally remain within the same range, but decreases slightly, indicating a slight improvement in reported health. However, in 2019, due to the survey change (noted in *Chapter 3: Data*), the average reported health rating is higher. An increase in overall health rating is observed from 2019-2020, indicating a decline in health which may be attributed to an increase in illness due to the emergence of the COVID-19 pandemic.

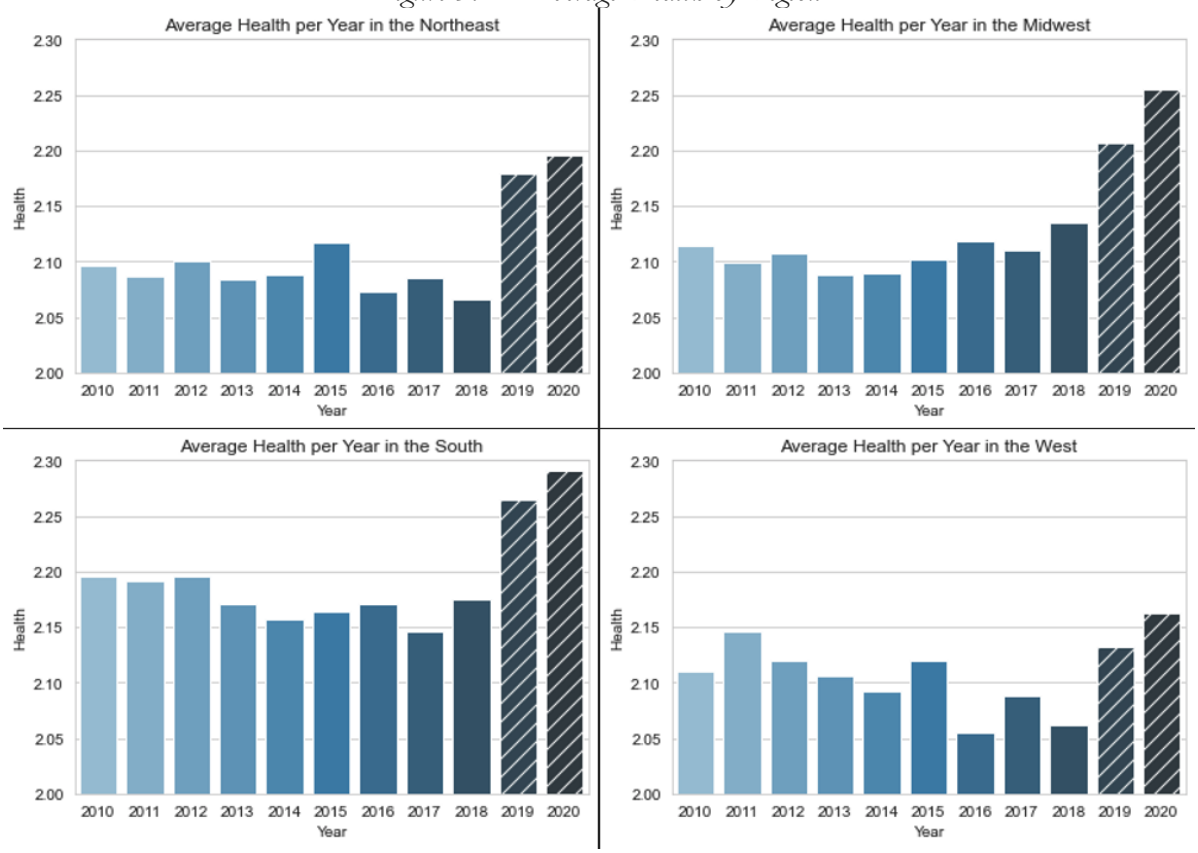
Figure 5.1 – Average Health per Year



Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Note: Due to the NHIS survey administration, results for 2019-2020 are not directly comparable to pre-2019 data.

Figure 5.2 below displays average health by region (Northeast, Midwest, South, West). From 2010-2018, it is observed that the average health rating is generally lowest in the Northeast and West, indicating better health in those regions. The highest health ratings are in the South, indicating worse health in that region. Health worsens across all regional groups in the 2019-2020 period, which may potentially be an effect of the COVID-19 pandemic, which collectively worsened health across the nation. Note: the 'high' and 'low' bars have counter-intuitive meanings, as a lower *'health'* value indicates better health, while a higher *'health'* bar indicates worse health.

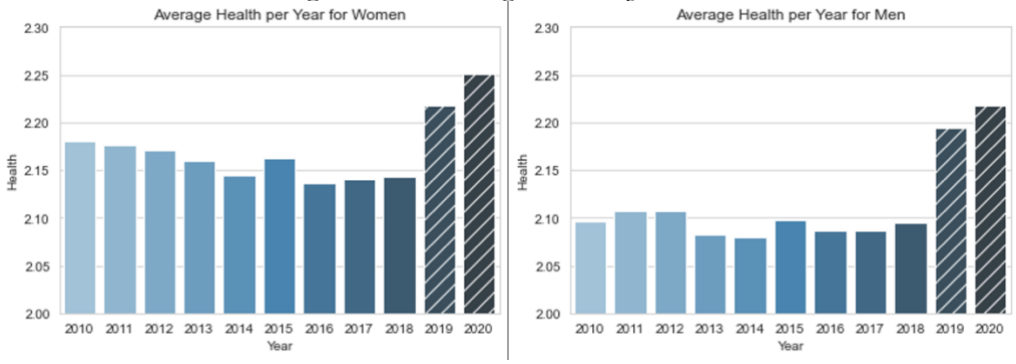
Figure 5.2 – Average Health by Region



Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Note: Due to the NHIS survey administration, results for 2019-2020 are not directly comparable to pre-2019 data.

Figure 5.3 below displays average health for each gender group. For the entire 2010-2020 period, the average health rating for women is consistently higher than for men, indicating that women, on average, reported worse health than men.

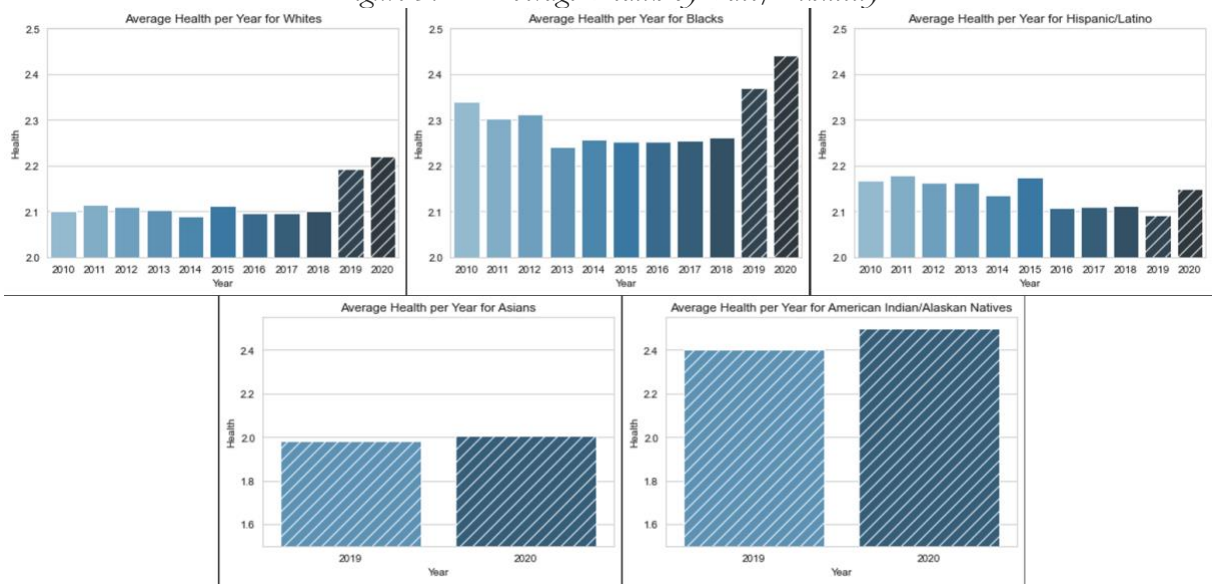
Figure 5.3 – Average Health by Gender



Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Note: Due to the NHIS survey administration, results for 2019-2020 are not directly comparable to pre-2019 data.

Figure 5.4 below displays average health for each racial/ethnic group (White, Black, Hispanic/Latino, Asian, and American Indian/Alaskan Native). From 2010-2018, it is observed that average health reported by White respondents is notably better than the other groups, while average health reported by Blacks was worst overall. Data on the Asian and American Indian/Alaskan Native racial groups was only available in 2019-2020, after the NHIS survey redesign, which makes it difficult to compare health trends for the two with the other racial groups. However, in 2019-2020, health notably worsened across all racial groups.

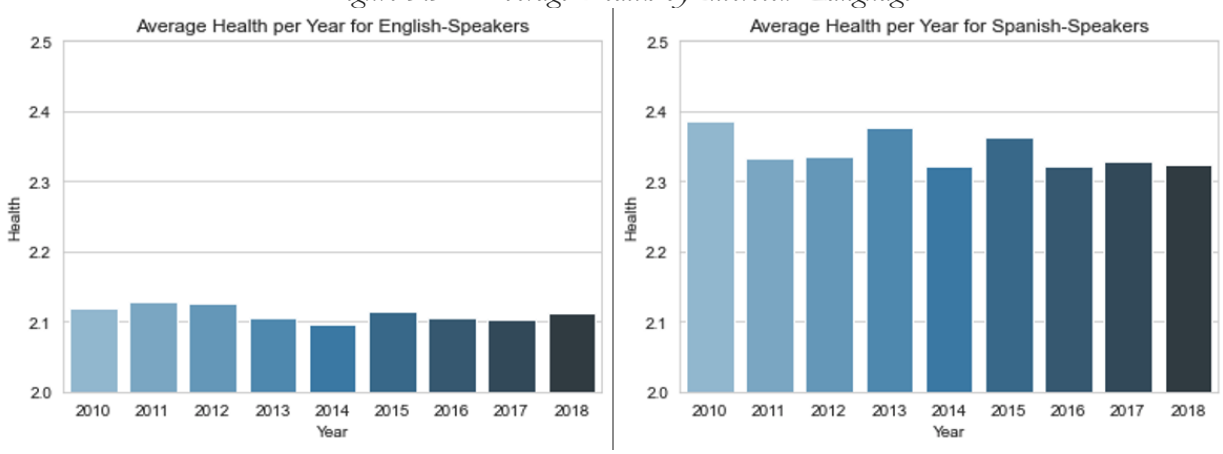
Figure 5.4 – Average Health by Race/Ethnicity



Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Note: Due to the NHIS survey administration, results for 2019-2020 are not directly comparable to pre-2019 data.

Figure 5.5 below displays average health by interview language (Spanish and English). The Interview Language data was only available through the 2010-2018 range, but the notable trend between these two charts is that respondents who spoke English reported significantly better health than respondents who spoke Spanish (the second-most common language).

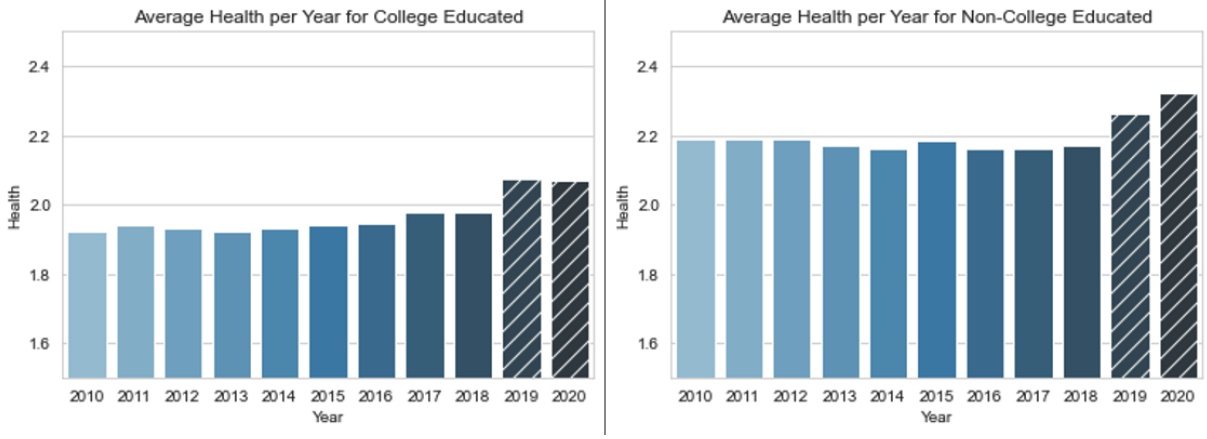
Figure 5.5 – Average Health by Interview Language



Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Note: Due to the NHIS survey administration, results for 2019-2020 are not directly comparable to pre-2019 data.

Figure 5.6 below displays average health by educational attainment (college educated and non-college educated). Respondents who received a college degree reported lower health ratings (corresponding to better health) when compared to their non-college educated counterparts.

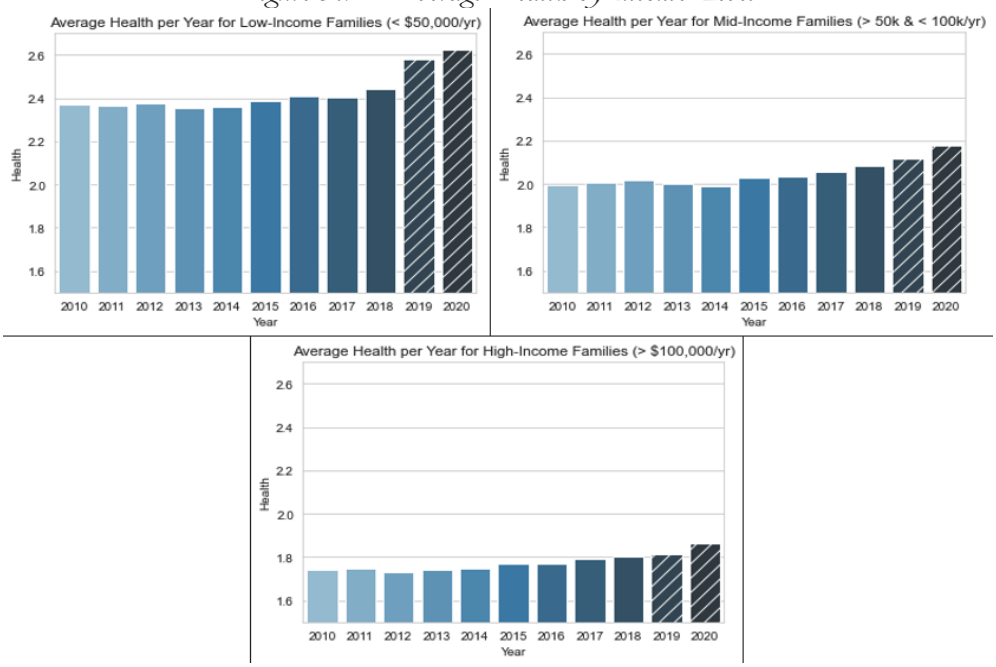
Figure 5.6 – Average Health by Education



Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Note: Due to the NHIS survey administration, results for 2019-2020 are not directly comparable to pre-2019 data.

Figure 5.7 below displays average health by income level (low income, mid-income, and high income). Respondents in the lowest income group reported having the worst health, while respondents in the highest income group reported the best health of the three (which was significantly better).

Figure 5.7 – Average Health by Income Level



Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Note: Due to the NHIS survey administration, results for 2019-2020 are not directly comparable to pre-2019 data.

In addition to the figures above, a series of regressions were fitted to examine specific relationships between individual variables and health. The following models aim to specifically address the research questions outlined in *Chapter 1: Introduction*. To address Research Question *(2), "To what extent do race and ethnicity impact variation in health outcomes?," 'health'* is modelled as a function of the race variables *'white,' 'black,' 'native,' 'asian,'* and *'hisp,'* using the "Other" race group as the omitted variable.

Table 5.1– RLMs with Race Covariates

Model 1.1 – Health vs. Race
Panel A (2010-2018)

Covariate	*Parameter Coefficient*	*Std. Error*
white	2.038	0.002
black	2.243	0.004
native	— —	— —
asian	— —	— —
hisp	0.157	0.003

Model 1.2 – Health vs. Race
Panel B (2019-2020)

Covariate	*Parameter Coefficient*	*Std. Error*
white	2.162	0.005
black	2.372	0.012
native	2.416	0.044
asian	1.955	0.017
hisp	-0.082	0.014

Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Notes: Parameter coefficients are in respect to the 'base' group, which was omitted (i.e., Other/Mixed race)

As shown in *Table 5.1* above, the parameter estimates for the race variables are positive, for the most part. This indicates an inverse effect on patient *'health,'* the response variable. Racial health data for the 2010-2018 range is limited, however, as the NHIS did not collect data on the *'native'* or *'asian'* racial groups prior to its 2019 survey redesign. From these two tables, it is observed that the estimated parameter for *'white'* is consistently lower than that of *'black'*, indicating that Black survey respondents report lower health than White survey respondents. An interesting observation is that in both *Models 1.1 and 1.2*, the coefficient for the *'hisp*' variable is lower than that of all other groups, indicating better health.

To address Research Question *(3) "To what extent are race and ethnicity correlated with the socioeconomic gradient in health?,"* a set of RLMs was fitted using education and income variables to determine socioeconomic status (SES). The *'health'* endogenous variable was modelled as a function of the exogenous variables *'college,' 'lowinc,'* and *'midinc.'*

Table 5.2 – RLMs with Education and Income Covariates

Model 2.1 – Health vs. Edu/Income
Panel A (2010-2018)

Covariate	*Parameter Coefficient*	*Std. Error*
highschool	1.013	0.003
somecollege	1.194	0.004
lowinc	1.378	0.003
midinc	1.154	0.003

Model 2.2 – Health vs. Edu/Income
Panel B (2019-2020)

Covariate	*Parameter Coefficient*	*Std. Error*
highschool	0.969	0.011
somecollege	1.215	0.013
lowinc	1.683	0.012
midinc	1.390	0.011

Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Notes: Parameter coefficients are in respect to the 'base' group, which was omitted (i.e., the *'collegedegree'* group for the education variables, and the *'highinc'* group for the income variables)

As observed in *Table 5.2* above, the parameter estimates for the education and income variables are positive, indicating an inverse effect on individual health. In both models, the *lowinc* variable has a larger coefficient than the *midinc* variable, indicating that survey respondents who reported lower income experienced worse health. Interestingly, it appears that the reverse is true

for education – survey respondents with no college education appear to report better health than those who received some college education (but no bachelor's degree).

A general set of models, *Model 3.1 and Model 3.2*, was then fitted to the data to provide an overview of how each variable correlates with the response variable, *'health.'* Parameter coefficients are displayed in *Table 5.3,* below. The *'health'* response variable is measured on a 5-point Likert scale, so positive coefficients indicate an inverse or negative correlation with health, while negative coefficients indicate a direct or positive correlation with health.

$$\begin{aligned} health = \ & \alpha + \beta_1(year) + \beta_2(age) + \beta_3(northeast) + \beta_4(midwest) + \ \beta_5(south) \\ & + \beta_6(female) + \beta_7(white) + \beta_8(black) + \ \beta_9(native) + \beta_{10}(asian) \\ & + \beta_{11}(hisp) + \beta_{12}(spanish) + \beta_{13}(nocollege) + \beta_{14}(somecollege) \\ & + \beta_{15}(lowinc) + \beta_{16}(midinc) + \ \varepsilon \end{aligned}$$

Table 5.3 – RLMs with All Covariates

Model 3.1 – Health vs. All Covariates
Panel A (2010-2018)

Covariate	*Parameter Coefficient*	*Std. Error*
year	0.000	0.000
age	0.019	0.000
northeast	-0.020	0.003
midwest	0.007	0.003
south	0.014	0.003
female	0.020	0.002
white	-0.101	0.004
black	0.048	0.005
native	— —	— —
asian	— —	— —
hisp	0.054	0.003
spanish	0.138	0.005
highschool	0.376	0.003
somecollege	0.245	0.003
lowinc	0.449	0.003
midinc	0.178	0.003

Model 3.2 – Health vs. All Covariates
Panel B (2019-2020)

Covariate	*Parameter Coefficient*	*Std. Error*
year	0.001	0.000
age	0.013	0.000
northeast	0.019	0.011
midwest	0.037	0.011
south	0.043	0.010
female	-0.049	0.007
white	-0.153	0.035
black	-0.046	0.036
native	-0.003	0.051
asian	-0.163	0.038
hisp	-0.035	0.013
spanish	0.000	0.000
highschool	0.163	0.009
somecollege	0.219	0.010
lowinc	0.566	0.009
midinc	0.211	0.009

Source: National Health Interview Survey 2010-2020 *(Blewett et al., 2019)*
Notes: Parameter coefficients are in respect to the 'base' group, which was omitted (i.e., the *'west'* group for regional variables, *'male'* for gender variables, *'other/mixed'* for the race variables, the *'collegedegree'* group for the education variables, and the *'highinc'* group for the income variables)

Interpretations of the different parameter coefficients are grouped into the following categories: region, gender, race, interview language, education, and income. As observed in *Table 5.3* above, regional correlation in health is relatively consistent, with the exception of *'northeast'* in 2010-2018 – the coefficient is negative, which indicates that the health in the northeast region was better compared to other regions. Health variation is positively correlated in other regions, indicating worse health. In considering health variation by gender, the *'female'* variable has a positive coefficient in 2010-2018, which indicates that health for women is worse. In 2019-2020, however, the *'female'* variable has a negative coefficient, indicating that health for women was positively correlated during these years.

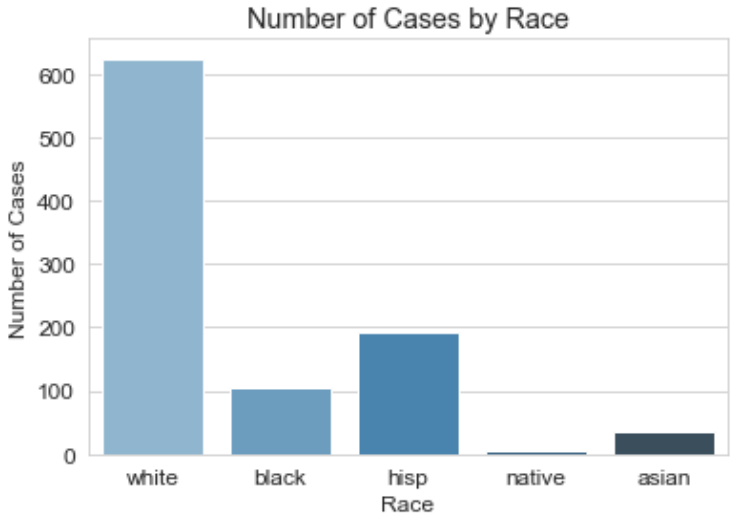
Racial/Ethnic correlation with health is signified by the *'white,' 'black,' 'native,' 'asian,'* and *'hisp'* variables above. Prior to the 2019 survey adjustment, the NHIS did not collect data on the *'native'* or *'asian'* racial groups, so data on these racial/ethnic groups does not exist for the 2010-2018 time period. In both *Model 3.1* and *Model 3.2*, the *'white'* variable has a strong negative correlation with *'health*', indicating that health was significantly better for Whites when compared to other groups. Health for the Black and Hispanic demographic groups is positively correlated with health in 2010-2018, indicating worse health for both groups. In 2019-2020, all racial/ethnic groups are negatively correlated with health. Additionally, the coefficient for the interview language variable, *'spanish,'* is positive for 2010-2018, indicating that survey respondents who did not speak English reported worse health than respondents who spoke English as a primary language.

Considering socioeconomic characteristics, the coefficient of the education variables, *'highschool'* and *'somecollege,'* are negative in both models when compared against the base group

who attained a bachelor's degree or above. This indicates that educational attainment is correlated with improved health. The coefficients of the household income variables, *'lowinc'* and *'midinc,'* are positive in both models, with the coefficients for *'lowinc'* being significantly higher than those for *'midinc.'* This indicates that lower income households report significantly worse health, while mid-income households report moderately worse health.

In addressing Research Question *(4) How present were these disparities when looking at outcomes related to the COVID-19 Pandemic?*, a series of bar charts were created illustrating the distribution of confirmed COVID-19 cases for various case outcomes, broken down by race. Bar charts were created for the following categories: number of cases, hospitalizations, ICU admissions, and deaths.

Figure 5.8 – Number of COVID-19 Cases by Race

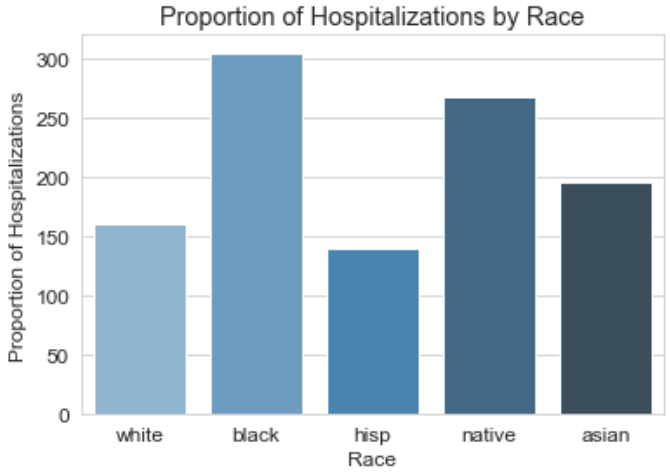


Source: CDC COVID-19 Case Surveillance Public Use Data *(Lee, 2021)*
Notes: *Figure 5.8* illustrates the number of confirmed COVID-19 cases by race per 1000 people (omitting probable and unknown cases).

Figure 5.8 above displays the distribution of confirmed cases of COVID-19 by race per 1000 people (i.e., in a sample of 1000 positive cases of COVID-19, the count of cases in each race group). From this graph, it appears that the vast majority of positive-confirmed COVID

cases in the United States are accounted for by White Americans, while the minority groups trail far behind. It is likely that this trend is largely due to White Americans accounting for a significantly larger percentage of the population – according to the U.S. Census Bureau, as of July 2021, White Americans accounted for 76.3% of the population of the United States (U.S. Census Bureau, 2021).

Figure 5.9 – Proportion of COVID-Related Hospitalizations by Race

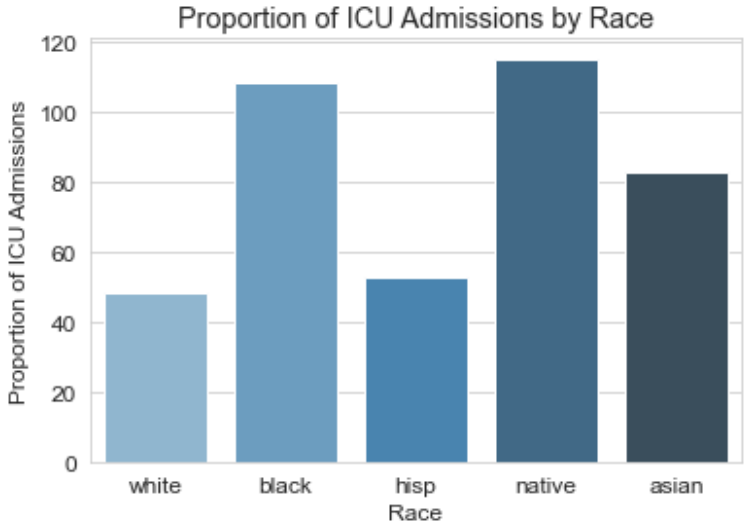


(Source: CDC COVID-19 Case Surveillance Public Use Data *(Lee, 2021)*
Notes: *Figure 5.9* illustrates the proportion of hospitalizations by race per 1000 confirmed cases of COVID-19 in each race group (omitting probable and unknown cases).

Figure 5.9 above displays the proportion of hospitalizations by race per 1000 confirmed cases of COVID-19 in each race group. From this chart, it is apparent that there is a significant difference in the effects of COVID-19 in Whites when compared with minority groups. While White Americans account for a larger number of positive-confirmed cases of COVID-19, the percentage of those confirmed cases that resulted in hospitalization is strikingly different when compared to other groups (as seen in *Figure 5.9*, 160/1000 confirmed cases in White Americans resulted in hospitalization, while 305/1000 confirmed cases in Black Americans resulted in hospitalizations). This suggests that minority individuals who tested positive for COVID-19 were overall more likely to be hospitalized when compared against non-minority (White)

individuals. The one notable exception to this is the Hispanic/Latino group, who reported 139 hospitalizations per 1000 confirmed cases of COVID-19. From this figure alone, it appears that minority patients are more affected by the coronavirus, requiring disproportionate amounts of hospitalizations related to COVID-19 than White patients.

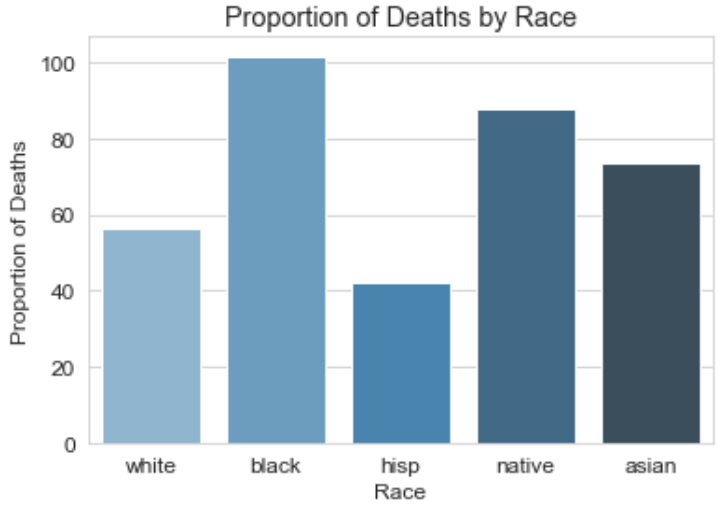
Figure 5.10 – Proportion of COVID-Related ICU Admissions by Race



Source: CDC COVID-19 Case Surveillance Public Use Data *(Lee, 2021)*
Notes: *Figure 5.10* illustrates the proportion of ICU admissions by race per 1000 confirmed cases of COVID-19 in each race group (omitting probable and unknown cases).

Figure 5.10 above displays the proportion of ICU admissions by race per 1000 confirmed cases of COVID-19 in each race group. Similar to *Figure 5.9*, this chart illustrates a notable trend in cases of COVID-19 in Whites when compared with minority groups. The proportion of confirmed cases that resulted in ICU admission for the White group is strikingly different when compared to the other minority groups (48/1000 Whites who tested positive for COVID-19 were admitted to the ICU, compared to 108 Blacks, 52 Hispanic/Latinos, 115 Native Americans, and 82 Asians). This suggests that minority patients are more severely affected by COVID-19, requiring disproportionate amounts ICU admissions due to complications related to the coronavirus, when compared against White patients.

Figure 5.11 – Proportion of COVID-Related Deaths by Race



Source: CDC COVID-19 Case Surveillance Public Use Data *(Lee, 2021)*
Notes: *Figure 5.11* illustrates the proportion of deaths by race per 1000 confirmed cases of COVID-19 in each race group (omitting probable and unknown cases).

Figure 5.11 above displays the proportion of deaths by race per 1000 confirmed cases of COVID-19 in each race group. Similar to *Figures 5.9 and 5.10*, this chart reveals a strikingly disproportionate effect in minority race groups when compared to Whites. Of 1000 confirmed cases of COVID-19 in each race group, the Black and Native American groups are most likely to die due to illness-related causes, while the White and Hispanic/Latino groups report the lowest death rates.

Chapter 6: Discussion and Conclusion

Historically, disparities in health have been well documented across the literature as patients from racial/ethnic minority backgrounds have statistically been demonstrated to experience worse health outcomes than their non-minority counterparts. These disparities ultimately lead to higher mortality rates among racial/ethnic minority groups, motivating further research in this area. Many factors may contribute to these differences – with variables at the patient-level, practitioner-level, and system-level. However, as much of the literature surrounding this topic examines data from several decades ago, this study aimed to examine these trends in health in a more recent period to determine whether these disparities continue to persist, or whether they have lessened. Additionally, in light of the COVID-19 pandemic, this study aims to examine differential effects of the pandemic across various racial and ethnic groups.

This study analyzed a series of bar charts and robust regression models comparing the evaluating the change in *'health'* rating across various groups over the 2010-2020 time period. It appears that the mean health rating generally declined over the 2010-2018 time period, indicating improved health across the population. However, when examining changes in average health within specific categories, some interesting trends emerged. Women, on average, reported poorer health than men. White respondents generally reported better health than other racial/ethnic groups, while Black respondents reported poorer health than all other racial/ethnic

groups. This finding supports the trends previously established in the literature, as racial minority patients have historically been documented to experience poorer health outcomes than non-minority patients (Balsa, 2001). In a similar vein, survey respondents who spoke English as their primary language reported better health overall than respondents who only spoke Spanish or another language. This supports previous findings suggesting that patient-physician language concordance has a positive effect on health, while language discordance may negatively impact health outcomes (IOM, 2003) and (Dillender, 2017).

Additionally, when examining health reported at different levels of education, respondents who obtained a Bachelor's degree or above reported better health than the other two groups (those who did not attend college, and those who attended some college). This finding matches with the results from Buckles et al. (2016), who found that completion of college is correlated with an improvement in health, leading to a decline in mortality. However, when examining the other education groups, an interesting observation emerged – those who attended some college but did not complete their degree actually reported *worse* health than the group who did not attend college at all. This finding actually contradicts what was found in Buckles et al. (2016), however, the contributing cause of this difference is not known.

When assessing health results reported across different income groups, low-income households (earning less than $50,000 per year) reported significantly worse health than other groups, while mid-income households (earning between $50,000 and $100,000 annually) reported moderate health. The high-income group (earning over $100,000 annually) reported the best health rating of the three groups. These findings substantiate the findings of Galama et al. (2018), which states that higher income and SES contribute to improved health. These findings

also support Grönqvist et al. (2012), which indicated that income inequality contributes to differences in health outcomes.

Furthermore, in examining trends in health during the COVID-19 pandemic, an examination of the CDC Case Surveillance Public Use Dataset found that there appear to be differential effects across various racial/ethnic groups due to the pandemic. While Whites accounted for the vast majority of positive-confirmed COVID-19 cases, they were reported to have disproportionately lower hospitalization, ICU admission, and death rates when compared with the corresponding rates for other minority groups. Conversely, Blacks and Native Americans were reported to have significantly higher hospitalization, ICU admittance, and death rates than all other groups.

Overall, the findings from this study substantiated findings from prior literature and demonstrated that disparities in health across these various groups (gender, race/ethnicity, language, education, and income) still exist in the 2010-2018 time period. The specific cause of these differences in health is not identified, however. Additionally, in examining the data from the NHIS 2019-2020 panel, it is observed that the COVID-19 pandemic had a negative impact on health across all groups and categories. However, in examining more specific trends in positive cases reported by the CDC, it appears that the COVID-19 pandemic had a disproportionately greater impact on the health of minorities.

Appendix A: Analysis Code

The analysis code discussed in this thesis can also be found at: https://github.com/meganthoang/healthdisparities

```
# NHIS 2010-2020 Analysis
#### Megan Hoang | HUT Script | 2-13-2022
# > Data extract from IPUMS NHIS. Codebook found at:
https://live.nhis.datadownload.ipums.org/web/extracts/nhis/1750331/nhis_00003.cbk

# import all necessary modules
import pandas as pd
import numpy as np
import sqlite3 # for SQL queries
import csv
#import requests # for API call
import matplotlib
from matplotlib import pyplot as plt # import matplotlib.pyplot as plt
from matplotlib import cm #Colormap
import seaborn as sns # visualization
#import glob
import os # directory
#from sodapy import Socrata # to read in the CDC Dataset
from itertools import combinations
import statsmodels.api as sm
import numpy as np
import statsmodels.formula.api as smf
from statsmodels.api import add_constant
import matplotlib.pyplot as plt
import sklearn
from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import cross_validate
from sklearn.linear_model import LinearRegression
import itertools

### First, let's read in our data.
#
# Steps in this section:
# * Set our Directory
# * Use Pandas to read in the CSV
# * View our data to make sure everything looks good
# * View some basic summary statistics

# set our directory
print(os.getcwd())
path = "/Users/meganhoang/Desktop/"
os.chdir(path)
```

```
print(os.getcwd())

# read in the CSV
data = pd.read_csv("nhis_00003.csv", low_memory=False)

# let's view our data to make sure everything looks good so far
print(data.head())

# let's look at some basic summary statistics
df = pd.DataFrame(data)
print(df.describe())

# ***
### Preprocessing
#
# > I chose to use the sqlite3 module in python in order to use SQL queries to
simplify the preprocessing process. (This way I can select the specific data I need
each time.) Additionally, I chose to omit observations in the dataset that were
designated as "unknown" values, for simplicity.
#
# Steps in this section:
# * Create a SQL database
# * Create the NHIS table to insert the data
# * Read the CSV into the database
# * Query the database for selected variables & compare sample statistics to above to
make sure everything still looks OK.
# * Create dummy variables and re-query

# create a SQL database to store the data so we can query it
con = sqlite3.connect('nhis.db')
cur = con.cursor()

# create our SQL table and insert the data
cur.execute("""create table NHIS
            (year       INTEGER, serial      INTEGER, strata      INTEGER,
            psu         INTEGER, nhishid     INTEGER, hhweight    INTEGER,
            region      INTEGER, pernum      INTEGER, nhispid     INTEGER,
            hhx         INTEGER, fmx         INTEGER, px          INTEGER,
            perweight   INTEGER, sampweight  INTEGER, longweight  INTEGER,
            partweight  INTEGER, fweight     INTEGER, astatflg    INTEGER,
            cstatflg    INTEGER, age         INTEGER, sex         INTEGER,
            race        INTEGER, hispeth     INTEGER, lang        INTEGER,
            edu         INTEGER, pooryn      INTEGER, famincome   INTEGER,
            health      INTEGER, mortstat    INTEGER, mortwt      INTEGER)""")

# read the csv into the database
file = open('nhis_00003.csv')
data = csv.reader(file)
```

```
cur.executemany('insert into NHIS values(?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?,
?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)', data)

# check to make sure a test query works
# cur.execute("select * from NHIS WHERE year = 2010")
# for row in cur.fetchall():
#  print(row)

# The commit method saves the changes.
con.commit()

# let's store the variables I want to query in a string:
select = """
        select
                year,
                region,
                age,
                case
                        when sex = 1 then 0
                        when sex = 2 then 1
                end as sex,
                case
                        when race = 100 or race = 200 then race
                        when race between 300 and 350 then 300
                        when race between 400 and 434 then 400
                end as race,
                case
                        when hispeth = 10 then 0
                        when hispeth between 20 and 70 then 1
                end as hispeth,
                case
                        when lang = 1 or lang = 3 then 0
                        when lang = 2 then 1
                end as lang,
                case
                        when edu between 100 and 116 then 100
                        when edu between 200 and 202 then 200
                        when edu between 300 and 303 then 300
                        when edu = 400 then edu
                        when edu between 500 and 501 then 500
                        when edu between 502 and 503 then 600
                end as edu,
                case
                        when pooryn = 1 then 0
                        when pooryn = 2 then 1
                end as pooryn,
                case
                        when famincome between 10 and 12 then 10
```

```
                        when famincome between 20 and 23 then 20
                        when famincome = 24 then 30
                end as famincome,
                health,
                mortstat
        """

# edit the SQL query to clean the data and omit "unknown" values per IPUMS codebook
remove = """ and region < 08
        and age < 999
        and sex < 3
        and race < 500
        and hispeth < 70
        and edu < 600
        and famincome < 96
        and health < 6"""

# remove = ""

df_query = pd.read_sql_query(select + "from NHIS where year between 2010 and 2020" +
remove, con)
df_query.describe()

# compare to the results from the summary statistics for df above (the non-queried
dataframe)

df_query = pd.read_sql_query(select + "from NHIS where year between 2010 and 2018" +
remove, con)
df_query.describe()


df_query = pd.read_sql_query(select + "from NHIS where year between 2019 and 2020" +
remove, con)
df_query.describe()


# __Dummy variables: (0 = F, 1 = T)__
#
# * region: northeast, midwest, south (omitted: west)
# * sex: female (omitted: male)
# * race: white, black, native, asian (omitted: other/mixed)
#     * Hisp: hisp = true (omitted: non-hispanic)
# * lang: english, spanish (omitted: other)
# * edu: college (omitted: no college)
# * famincome: lowinc (< 50,000), midinc (between 50,000 & 100,000), highinc (>
100,000) (omitted: highinc)
#
# __Variables left as is:__
```

```
# * year
# * age
# * famincome
# * health
# * mortstat
#
# > 19 variables in total: year, age, northeast, midwest, south, female, white, black,
native, asian, hisp, spanish, college, lowinc, midinc, highinc
#

# let's store the variables I want to query in a string:
# dummy variables: sex: region (West omitted) 1 = female
select_dummy = """
        select
                year,

                case
                    when region = 01 then 1
                    else 0
                end as northeast,
                case
                    when region = 02 then 1
                    else 0
                end as midwest,
                case
                    when region = 03 then 1
                    else 0
                end as south,
                case
                    when region = 04 then 1
                    else 0
                end as west,

                age,

                case
                    when sex = 2 then 1
                    else 0
                end as female,


                case
                    when race = 100 then 1
                    else 0
                end as white,
                case
                    when race = 200 then 1
                    else 0
```

```
end as black,
case
    when race = 300 then 1
    else 0
end as native,
case
    when race = 400 then 1
    else 0
end as asian,
case
        when hispeth = 10 then 0
        when hispeth between 20 and 70 then 1
end as hisp,

case
        when lang = 1 or lang = 3 then 1
        else 0
end as english,
case
    when lang = 2 then 1
    else 0
end as spanish,

case
    when edu <= 202 then 1
    else 0
end as nocollege,
case
    when edu between 300 and 399 then 1
    else 0
end as somecollege,
case
    when edu < 400 then 0
    when edu >= 400 then 1
end as collegedegree,


case
    when famincome between 10 and 12 then 1
    else 0
end as lowinc,
case
    when famincome between 20 and 23 then 1
    else 0
end as midinc,
case
    when famincome = 24 then 1
    else 0
```

```
                end as highinc,

                health,
                mortstat
        """

df_dummy = pd.read_sql_query(select_dummy + "from NHIS where year between 2010 and
2020" + remove, con)
df_dummy.describe()

# The commit method saves the changes.
con.commit()

# Close the connection when finished.
con.close()

#
# ***
### Visualizations
#
# > *independent variable: health*
#
# Visualizations in this section:
# * Average Health per Year by Race:
#     * Whites
#     * Blacks
#     * American Indian/Alaskan Native
#     * Asian
#     * Hispanic/Latino
# * Average Health per Year by Gender:
#     * Women
#     * Men
#
#

white = df_dummy.loc[df_dummy['white'] == 1, ['year', 'health']]
white.describe()

plt.figure(figsize = (10, 90))
white.groupby('year').mean().plot()
plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Whites')

#### Health by Race

white = df_dummy.loc[df_dummy['white'] == 1, ['year', 'health']]
white.describe()
```

```
year_health = white.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")
for i, bar in enumerate(ax.patches):
    if i > 8:
        hatch = next(hatches)
        bar.set_hatch(hatch)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.5)

# matplotlib.rc('xtick', labelsize=10)
# matplotlib.rc('ytick', labelsize=10)
# plt.rcParams.update({'font.size': 14})

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Whites')

black = df_dummy.loc[df_dummy['black'] == 1, ['year', 'health']]
black.describe()

year_health = black.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.5)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Blacks')

native = df_dummy.loc[df_dummy['native'] == 1, ['year', 'health']]
native.describe()

year_health = native.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")
```

```
for i, bar in enumerate(ax.patches):
    if i >= 0:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(1.5, 2.55)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for American Indian/Alaskan Natives')

asian = df_dummy.loc[df_dummy['asian'] == 1, ['year', 'health']]
asian.describe()

year_health = asian.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i >= 0:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(1.5, 2.55)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Asians')

hisp = df_dummy.loc[df_dummy['hisp'] == 1, ['year', 'health']]
hisp.describe()

year_health = hisp.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.5)
```

```
plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Hispanic/Latino')

#### Health by Gender

female = df_dummy.loc[df_dummy['female'] == 1, ['year', 'health']]
female.describe()

# plt.figure(figsize = (10, 90))
# female.groupby('year').mean().plot()
# plt.xlabel('Year')
# plt.ylabel('Health')
# plt.title('Average Health per Year for Women')

year_health = female.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.3)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Women')

male = df_dummy.loc[df_dummy['female'] == 0, ['year', 'health']]
male.describe()

year_health = male.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.3)
```

```
plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Men')

#### Health by Education

college = df_dummy.loc[df_dummy['college'] == 1, ['year', 'health']]
college.describe()

year_health = college.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(1.5, 2.5)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for College Educated')

college = df_dummy.loc[df_dummy['college'] == 0, ['year', 'health']]
college.describe()

year_health = college.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(1.5, 2.5)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Non-College Educated')

#### Health by Income
```

```
low = df_dummy.loc[df_dummy['lowinc'] == 1, ['year', 'health']]
low.describe()

year_health = low.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(1.5, 2.7)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Low-Income Families (< $50,000/yr)')

mid = df_dummy.loc[df_dummy['midinc'] == 1, ['year', 'health']]
mid.describe()

year_health = mid.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(1.5, 2.7)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Mid-Income Families (> 50k & < 100k/yr)')

high = df_dummy.loc[df_dummy['highinc'] == 1, ['year', 'health']]
high.describe()

year_health = high.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")
```

```
for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(1.5, 2.7)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for High-Income Families (> $100,000/yr)')

#### Health by Interview Language

english = df_dummy.loc[df_dummy['english'] == 1, ['year', 'health']]
english.describe()

year_health = english.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.5)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for English-Speakers')

spanish = df_dummy.loc[df_dummy['spanish'] == 1, ['year', 'health']]
spanish.describe()

year_health = spanish.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)
```

```
sns.set_style("whitegrid")
ax.set_ylim(2, 2.5)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year for Spanish-Speakers')

#### Health by Region

northeast = df_dummy.loc[df_dummy['northeast'] == 1, ['year', 'health']]
northeast.describe()

year_health = northeast.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.3)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year in the Northeast')

midwest = df_dummy.loc[df_dummy['midwest'] == 1, ['year', 'health']]
midwest.describe()

year_health = midwest.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.3)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year in the Midwest')
```

```
south = df_dummy.loc[df_dummy['south'] == 1, ['year', 'health']]
south.describe()

year_health = south.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.3)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year in the South')

west = df_dummy.loc[df_dummy['west'] == 1, ['year', 'health']]
west.describe()

year_health = west.groupby('year').mean().reset_index()

hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.3)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year in the West')

#### General Health per Year

all = df_dummy
all.describe()

year_health = all.groupby('year').mean().reset_index()
```

```
hatches = itertools.cycle(['//', '//'])
ax = sns.barplot(x=year_health['year'], y=year_health['health'], palette = "Blues_d")

for i, bar in enumerate(ax.patches):
    if i > 8:
        bar.set_hatch(hatch)
        hatch = next(hatches)

sns.set_style("whitegrid")
ax.set_ylim(2, 2.3)

plt.xlabel('Year')
plt.ylabel('Health')
plt.title('Average Health per Year')

# ***
### Models
#
# > For my models, I chose to run two models: a Basic OLS and a Robust Linear Model
(RLM) to determine some baseline coefficients.
#
# *independent variable: health,
# dependent variables: region, age, sex, racea, edu, pooryn, & famincome*
#
# Models in this section:
# * Basic OLS (from last week) -- commented out
# * Basic OLS with Dummy Variables
# * Robust Linear Model (from last week) -- commented out
# * Robust Linear Model with Dummy Variables

# now let's try fitting our RLM using the dummy variables & print summary
df_dummy = df_dummy.dropna()
x = df_dummy[['year', 'age', 'northeast', 'midwest', 'south', 'female', 'white',
'black', 'native', 'asian', 'hisp', 'spanish', 'college', 'lowinc', 'midinc']]
y = df_dummy['health']
rlm_model = sm.RLM(y, x, M=sm.robust.norms.HuberT())
rlm_results = rlm_model.fit()
print("Parameters:")
print(rlm_results.params)
print("\n")
print(rlm_results.summary())

#### Interpretation of 'health' response variable:
# > health: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor
#

#### Models 1.1 and 1.2 - Health v. Race
```

```
# only race variables 2010-2018
df_dummy = df_dummy.dropna()
df4 = df_dummy.loc[df_dummy['year'] <= 2018]
df4.describe()

x = df4[['white', 'black', 'native', 'asian', 'hisp']]
y = df4['health']
rlm_model4 = sm.RLM(y, x, M=sm.robust.norms.HuberT())
rlm_results4 = rlm_model4.fit()
print("Parameters:")
print(rlm_results4.params)
print("\n")
print(rlm_results4.summary())

# race variables 2019-2020
df_dummy = df_dummy.dropna()
df5 = df_dummy.loc[df_dummy['year'] > 2018]
df5.describe()

x = df5[['white', 'black', 'native', 'asian', 'hisp']]
y = df5['health']
rlm_model5 = sm.RLM(y, x, M=sm.robust.norms.HuberT())
rlm_results5 = rlm_model5.fit()
print("Parameters:")
print(rlm_results5.params)
print("\n")
print(rlm_results5.summary())

#### Models 2.1 and 2.2 - Socioeconomic Variables (education + income) v. Health

# socioeconomic

df_dummy = df_dummy.dropna()
df6 = df_dummy.loc[df_dummy['year'] <= 2018]
df6.describe()

x = df6[['nocollege', 'somecollege' 'lowinc', 'midinc']]
y = df6['health']
rlm_model6 = sm.RLM(y, x, M=sm.robust.norms.HuberT())
rlm_results6 = rlm_model6.fit()
print("Parameters:")
print(rlm_results6.params)
print("\n")
print(rlm_results6.summary())

# socioeconomic

df_dummy = df_dummy.dropna()
```

```
df7 = df_dummy.loc[df_dummy['year'] > 2018]
df7.describe()

x = df7[['college', 'lowinc', 'midinc']]
y = df7['health']
rlm_model7 = sm.RLM(y, x, M=sm.robust.norms.HuberT())
rlm_results7 = rlm_model7.fit()
print("Parameters:")
print(rlm_results7.params)
print("\n")
print(rlm_results7.summary())

#### Models 3.1 & 3.2 -- Robust Linear Model all covariates

# Panel A (2010-2018)
df_dummy = df_dummy.dropna()
df2 = df_dummy.loc[df_dummy['year'] <= 2018]
df2.describe()

x = df2[['year', 'age', 'northeast', 'midwest', 'south', 'female', 'white', 'black',
'native', 'asian', 'hisp', 'spanish', 'college', 'lowinc', 'midinc']]
y = df2['health']
rlm_model2 = sm.RLM(y, x, M=sm.robust.norms.HuberT())
rlm_results2 = rlm_model2.fit()
print("Parameters:")
print(rlm_results2.params)
print("\n")
print(rlm_results2.summary())

# Panel B (2019-2020)
df_dummy = df_dummy.dropna()
df3 = df_dummy.loc[df_dummy['year'] > 2018]
df3.describe()

x = df3[['year', 'age', 'northeast', 'midwest', 'south', 'female', 'white', 'black',
'native', 'asian', 'hisp', 'spanish', 'college', 'lowinc', 'midinc']]
y = df3['health']
rlm_model3 = sm.RLM(y, x, M=sm.robust.norms.HuberT())
rlm_results3 = rlm_model3.fit()
print("Parameters:")
print(rlm_results3.params)
print("\n")
print(rlm_results3.summary())
```

```
# CDC COVID-19 Case Surveillance Public Use Data 2020 Analysis
#### Megan Hoang | HUT Script | 4-7-2022
# > Data from https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-
Public-Use-Data/vbim-akqf

# import all necessary modules
import pandas as pd
import numpy as np
import sqlite3 # for SQL queries
import csv
import matplotlib
from matplotlib import pyplot as plt # import matplotlib.pyplot as plt
from matplotlib import cm #Colormap
import seaborn as sns # visualization
import os # directory
import statsmodels.api as sm
from sklearn.linear_model import LogisticRegression

# set our directory to the SSD
print(os.getcwd())
path = "/Volumes/Extreme SSD/Megan Windows Backup 1.6.2022/Honors Undergraduate
Thesis/Analysis/Data/CDC/Case Surveillance Public Use Data/"
os.chdir(path)
print(os.getcwd())

# read in the CSV to see if we can access it properly -- the dataset is too large, so
processing using pandas "chunks"
# data = pd.read_csv("COVID-19_Case_Surveillance_Public_Use_Data.csv",
low_memory=False)
# print(data.head())

# for chunk in pd.read_csv("COVID-19_Case_Surveillance_Public_Use_Data.csv",
chunksize=10):
#     print(chunk)
# now that we can access the data, let's set up the database:

# set our directory
print(os.getcwd())
path = "/Users/meganhoang/Desktop/"
os.chdir(path)
print(os.getcwd())


con = sqlite3.connect('cdc.db')
```

```
cur = con.cursor()
print(os.getcwd())
path = "/Volumes/Extreme SSD/Megan Windows Backup 1.6.2022/Honors Undergraduate
Thesis/Analysis/Data/CDC/Case Surveillance Public Use Data/"
os.chdir(path)
print(os.getcwd())

cur.execute("""create table CDC
            (cdc_case_earliest_dt   DATETIME,
            cdc_report_dt           DATETIME,
            pos_spec_dt             DATETIME,
            onset_dt                DATETIME,
            status                  TEXT,
            sex                     TEXT,
            age                     TEXT,
            race                    TEXT,
            hosp                    TEXT,
            icu                     TEXT,
            death                   TEXT,
            medcond                 TEXT)""")

# read the csv into the database
file = open('COVID-19_Case_Surveillance_Public_Use_Data.csv')
data = csv.reader(file)
cur.executemany('insert into CDC values(?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)', data)
print("success!!")
# let's store the variables I want to query in a string:
# dummy variables: sex: region (West omitted) 1 = female
# Age naming bracket is as follows: Child (0-9), Youth (10-19), Adult (20-59), Senior
(60+)


select = """
        select
            cdc_report_dt,
            case
                when status = 'Laboratory-confirmed case' then 1
                else 0
            end as confirmed_case,

            case
                when sex = 'Female' then 1
                when sex = 'Male' then 0
                else 99999
            end as female,
            case
                when sex = 'Female' then 0
                when sex = 'Male' then 1
```

```
            else 99999
        end as male,

        case
            when age = '0 - 9 Years' then 1
            else 0
        end as child,
        case
            when age = '10 - 19 Years' then 1
            else 0
        end as youth,
        case
            when age = '20 - 39 Years' then 1
            when age = '40 - 49 Years' then 1
            when age = '50 - 59 Years' then 1
            else 0
        end as adult,
        case
            when age = '60 - 69 Years' then 1
            when age = '70 - 79 Years' then 1
            when age = '80 + Years' then 1
            else 0
        end as senior,

        case
            when race = 'White, Non-Hispanic' then 1
            else 0
        end as white,
        case
            when race = 'Black, Non-Hispanic' then 1
            else 0
        end as black,
        case
            when race = 'Hispanic/Latino' then 1
            else 0
        end as hisp,
        case
            when race = 'American Indian/Alaska Native, Non-Hispanic' then 1
            else 0
        end as native,
        case
            when race = 'Asian, Non-Hispanic' then 1
            when race = 'Native Hawaiian/Other Pacific Islander, Non-Hispanic'
then 1
            else 0
        end as asian,

        case
```

```
            when hosp = 'Yes' then 1
            when hosp = 'No' then 0
        end as hosp,

        case
            when icu = 'Yes' then 1
            else 0
        end as icu,

        case
            when death = 'Yes' then 1
            when death = 'No' then 0
        end as death,

        case
            when medcond = 'Yes' then 1
            when medcond = 'No' then 0
        end as medcond
    """

# edit the SQL query to clean the data and omit "unknown" values per CDC codebook
# remove = """ and sex != 'Unknown' and sex != 'Other' and sex != 'Missing' and sex !=
'NA' """

remove = """ and sex != 'Unknown' and sex != 'Other' and sex != 'Missing' and sex !=
'NA'
        and age != 'Missing' and age != 'NA'
        and race != 'Unknown' and race != 'Missing' and race != 'NA'
        and hosp != 'Unknown' and hosp != 'Missing'
        and icu != 'Unknown' and icu != 'Missing'
        and death != 'Missing' and death != 'Unknown'
        and medcond != 'Unknown' and medcond != 'Missing'"""

# remove = ""
# cur.execute("select * from CDC where race_ethnicity_combined like 'Asian, Non-
Hispanic'")
# for row in cur.fetchall():
#  print(row)

# Use the commit method to save changes.
con.commit()
df_query = pd.read_sql_query(select + "from CDC where status = 'Laboratory-confirmed
case'" + remove, con)
df_query.describe()
df_query.dropna()
df_query.head()
# df_query.describe(include='all')
```

```
#### Visualizations
# * Number of Cases by Race
# * Hospitalizations by Race
# * ICU Admittance by Race
# * Deaths by Race
# Number of Cases by Race

val_counts = []

for col in ['white', 'black', 'hisp', 'native', 'asian']:
    count = df_query[col].value_counts()
    val_counts.append(count[1] / (count[1] + count[0]) * 1000)

import matplotlib.pyplot as plt; plt.rc("font", size=12)
y_pos = np.arange(len(['white', 'black', 'hisp', 'native', 'asian']))

p = reversed(sns.color_palette('Blues_d', n_colors=5))
sns.barplot(y_pos, val_counts, palette = p)

plt.xticks(y_pos, ['white', 'black', 'hisp', 'native', 'asian'])
plt.ylabel('Number of Cases')
plt.xlabel('Race')
plt.title('Number of Cases by Race')

plt.show()
# Proportion of Hospitalizations by Race

val_counts = []

for col in ['white', 'black', 'hisp', 'native', 'asian']:
    counts_df = df_query.groupby(col)['hosp'].value_counts()
    try:
        print(counts_df[1][1] / (counts_df[1][1] + counts_df[1][0]) * 1000)
        val_counts.append(counts_df[1][1] / (counts_df[1][1] + counts_df[1][0]) *
1000)
    except:
        val_counts.append(0)

import matplotlib.pyplot as plt; plt.rc("font", size=12)
y_pos = np.arange(len(['white', 'black', 'hisp', 'native', 'asian']))

p = reversed(sns.color_palette('Blues_d', n_colors=5))
sns.barplot(y_pos, val_counts, palette = p)
# plt.bar(y_pos, val_counts, align='center', alpha=0.5)
plt.xticks(y_pos, ['white', 'black', 'hisp', 'native', 'asian'])
plt.ylabel('Proportion of Hospitalizations')
plt.xlabel('Race')
plt.title('Proportion of Hospitalizations by Race')
```

```
plt.show()
# Proportion of ICU Cases by Race

val_counts = []

for col in ['white', 'black', 'hisp', 'native', 'asian']:
    counts_df = df_query.groupby(col)['icu'].value_counts()
    try:
        print(counts_df[1][1] / (counts_df[1][1] + counts_df[1][0]) * 1000)
        val_counts.append(counts_df[1][1] / (counts_df[1][1] + counts_df[1][0]) *
1000)
    except:
        val_counts.append(0)

import matplotlib.pyplot as plt; plt.rc("font", size=12)
y_pos = np.arange(len(['white', 'black', 'hisp', 'native', 'asian']))

p = reversed(sns.color_palette('Blues_d', n_colors=5))
sns.barplot(y_pos, val_counts, palette = p)
# plt.bar(y_pos, val_counts, align='center', alpha=0.5)
plt.xticks(y_pos, ['white', 'black', 'hisp', 'native', 'asian'])
plt.ylabel('Proportion of ICU Admissions')
plt.xlabel('Race')
plt.title('Proportion of ICU Admissions by Race')

plt.show()
# Proportion of Deaths by Race

val_counts = []

for col in ['white', 'black', 'hisp', 'native', 'asian']:
    counts_df = df_query.groupby(col)['death'].value_counts()
    try:
        print(counts_df[1][1] / (counts_df[1][1] + counts_df[1][0]) * 1000)
        val_counts.append(counts_df[1][1] / (counts_df[1][1] + counts_df[1][0]) *
1000)
    except:
        val_counts.append(0)

import matplotlib.pyplot as plt; plt.rc("font", size=12)
y_pos = np.arange(len(['white', 'black', 'hisp', 'native', 'asian']))

p = reversed(sns.color_palette('Blues_d', n_colors=5))
sns.barplot(y_pos, val_counts, palette = p)

plt.xticks(y_pos, ['white', 'black', 'hisp', 'native', 'asian'])
plt.ylabel('Proportion of Deaths')
```

```
plt.xlabel('Race')
plt.title('Proportion of Deaths by Race')

plt.show()

#### Model Specification
# * Logistic Model
#   * independent variables: 'white', 'black', 'hisp', 'native', 'asian'
#   * dependent variable: 'death'
# bar chart for visualization
GroupedData = df_query.groupby(by='jobsatis').size()
GroupedData.plot.bar(x='lab', y='val', rot=0)
plt.xlabel('jobsatis')
plt.ylabel('observations')
plt.title('Distribution of the Response')
X = df_query[['female', 'black', 'hisp', 'native', 'asian', 'child', 'youth',
'senior', 'hosp', 'icu', 'medcond']]
# omitted group: male, white, adult
y = df_query['death']

import statsmodels.api as sm
logit_model=sm.Logit(y,X)
result=logit_model.fit()
print(result.summary2())

# logit_model=sm.Logit(y,X)
# result=logit_model.fit()
# print(result.summary2())
X = df_query[['black', 'hisp', 'native', 'asian']]
# omitted group: male, white, adult
y = df_query['icu']

import statsmodels.api as sm
logit_model=sm.Logit(y,X)
result=logit_model.fit()
print(result.summary2())
# Close the connection when finished.
con.close()
```

References

AJMC Staff. (2021, January 1). A timeline of covid-19 developments in 2020. AJMC. Retrieved November 3, 2021, from https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020.

Angerer, S., Waibel, C., Stummer, H. (2019). Discrimination in health care: A field experiment on the impact of patients' socioeconomic status on access to care. American Journal of Health Economics, 5(4), 407–427. https://doi.org/10.1162/ajhe_a_00124

Azar, K. M., Shen, Z., Romanelli, R. J., Lockhart, S. H., Smits, K., Robinson, S., Brown, S., Pressman, A. R. (2020). Disparities in outcomes among COVID-19 patients in a large health care system in California. Health Affairs, 39(7), 1253–1262. https://doi.org/10.1377/hlthaff.2020.00598

Balsa, A. I., McGuire, T. G. (2001). Statistical discrimination in health care. Journal of Health Economics, 20(6), 881–907. https://doi.org/10.1016/s016 7-6296(01)00101-1

Balsa, A. I., McGuire, T. G. (2003). Prejudice, clinical uncertainty and stereotyping as sources of health disparities. Journal of Health Economics, 22(1), 89–116. https://doi.org/10.1016/s0167-6296(02)00098-x

Balsa, A. I., Cao, Z., McGuire, T. G. (2007). Does managed health care reduce health care disparities between minorities and Whites? Journal of Health Economics, 26(1), 101–121. https://doi.org/10.1016/j.jhe aleco.2006.06.001

Blewett, L. A., Drew, J. A., King, M. L., Williams, K. C. W. (2019). IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D070.V6.4

Buckles, K., Hagemann, A., Malamud, O., Morrill, M., Wozniak, A. (2016). The effect of college education on mortality. Journal of Health Economics, 50, 99–114. https://doi.org/10.1016/j.jhealeco.2016.08.002

Burroughs, V. J., Maxey, R. W., Levy, R. A. (2002). Racial and ethnic differences in response to medicines: towards individualized pharmaceutical treatment. Journal of the National Medical Association, 94(10 Suppl), 1–26.

Dillender, M. (2017). English skills and the health insurance coverage of immigrants. American Journal of Health Economics, 3(3), 312–345. https://doi.org/10.1162/ajhe_a_00077

Doyle, J. J., Ewer, S. M., Wagner, T. H. (2010). Returns to physician human capital: Evidence from patients randomized to physician teams. Journal of Health Economics, 29(6), 866–882. https://doi.org/10.1016/j.jhealeco.2010.08.004

Ericson, K. M., Sydnor, J. (2017). The Questionable Value of Having a Choice of Levels of Health Insurance Coverage. Journal of Economic Perspectives, 31(4), 51–72. https://doi.org/10.1257/jep.31.4.51

Galama, T. J., van Kippersluis, H. (2018). A Theory of Socio-economic Disparities in Health over the Life Cycle. The Economic Journal, 129(617), 338–374. https://doi.org/10.1111/ecoj.12577

Godøy, A., Huitfeldt, I. (2020). Regional variation in health care utilization and mortality. Journal of Health Economics, 71, 102254. https://doi.org/10.1016/j.jhealeco.2019.102254

Grönqvist, H., Johansson, P., Niknami, S. (2012). Income inequality and health: Lessons from a refugee residential assignment program. Journal of Health Economics, 31(4), 617–629. https://doi.org/10.1016/j.jhealeco.2012.05.003

Gruber, J. (2017). Delivering Public Health Insurance Through Private Plan Choice in the United States. Journal of Economic Perspectives, 31(4), 3–22. https://doi.org/10.1257/jep.31.4.3

Gruber, J., Owings, M. (1994). Physician financial incentives and Cesarean Section Delivery. https://doi.org/10.3386/w4933

Institute of Medicine (2003). Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: The National Academies Press. https://doi.org/10.17226/12875.

Johar, M., Jones, G., Keane, M. P., Savage, E., Stavrunova, O. (2013). Discrimination in a universal health system: Explaining socioeconomic waiting time gaps. Journal of Health Economics, 32(1), 181–194. https://doi.org/10.1016/j.jhealeco.2012.09.004

Kim, I. (2013). The relationship between critical ethnic awareness and racial discrimination: Multiple indirect effects of coping strategies among Asian Americans. Journal of the Society for Social Work and Research, 4(3), 261–277. https://doi.org/10.5243/jsswr.2013.17

Lee, B. (2021). COVID-19 Case Surveillance Public Use Data. [Data file]. Retrieved from https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

Mackey, K., Ayers, C. K., Kondo, K. K., Saha, S., Advani, S. M., Young, S., Spencer, H., Rusek, M., Anderson, J., Veazie, S., Smith, M., Kansagara, D. (2021). Racial and ethnic disparities in COVID-19–related infections, hospitalizations, and deaths. Annals of Internal Medicine, 174(3), 362–373. https://doi.org/10.7326/m20-6306

Miller, S., Wherry, L. R., Mazumder, B. (2021). Estimated mortality increases during the COVID-19 pandemic by socioeconomic status, race, and ethnicity. Health Affairs, 40(8), 1252–1260. https://doi.org/10.1377/hlthaff.2021.00414

Moscelli, G., Siciliani, L., Gutacker, N., Cookson, R. (2018). Socioeconomic inequality of access to healthcare: Does choice explain the gradient? Journal of Health Economics, 57, 290–314. https://doi.org/10.1016/j.jhealeco.2017.06.005

Paradies, Y., Ben, J., Denson, N., Elias, A., Priest, N., Pieterse, A., Gupta, A., Kelaher, M., Gee, G. (2015). Racism as a Determinant of Health: A Systematic Review and Meta-Analysis. PloS one, 10(9), e0138511. https://doi.org/10.1371/journal.pone.0138511

Rivenbark, J.G., Ichou, M. (2020). Discrimination in healthcare as a barrier to care: experiences of socially disadvantaged populations in France from a nationally representative survey. BMC Public Health 20, 31. https://doi.org/10.1186/s12889-019-8124-z

Schulman, K. A., Berlin, J. A., Harless, W., Kerner, J. F., Sistrunk, S., Gersh, B. J., Dubé, R., Taleghani, C. K., Burke, J. E., Williams, S., Eisenberg, J. M., Ayers, W., Escarce, J. J. (1999). The effect of race and sex on physicians' recommendations for cardiac catheterization. New England Journal of Medicine, 340(8), 618–626. https://doi.org/10.1056/nejm199902253400806

Shavers, V. L., Fagan, P., Jones, D., Klein, W. M., Boyington, J., Moten, C., Rorie, E. (2012). The state of research on racial/ethnic discrimination in the receipt of Health Care. American Journal of Public Health, 102(5), 953–966. https://doi.org/10.2105/ajph.2012.300773

Speybroeck, Niko. (2013) "Simulation models for socioeconomic inequalities in health: a systematic review." International journal of environmental research and public health vol. 10,11 5750-80. 4 Nov. 2013, doi:10.3390/ijerph10115750

Team, M. P. C. U. X. U. I. (n.d.). 2019 NHIS redesign. IPUMS NHIS. Retrieved March 15, 2022, from https://nhis.ipums.org/nhis/userNotes_2019_NHIS_Redesign.shtml

U.S. Census Bureau Quickfacts: United States. (2021). Retrieved April 8, 2022, from https://www.census.gov/quickfacts/fact/table/US/PST045221

Vedantam, S. (Host). (2020, September 7). The Fee-for-Service Monster [Audio podcast episode]. In *Hidden Brain*. NPR. https://podcasts.apple.com/gb/podcast/the-fee-for-service-monster/id1028908750?i=1000490388408