Joseph Bui
Nathan Tsai
David Laub

Science Track: CDC Chronic Disease Data Analysis

Introduction

Diabetes is one of the most prevalent diseases that affects many people yearly. However, according to "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care," there is evidence demonstrating the unequal medical treatment towards racial minorities compared to whites in the U.S. healthcare system. Furthermore, Penner *et al.* report that "on average, White, Hispanic, and Asian physicians all display relatively large implicit racial preferences for Whites over Blacks. Physicians also hold implicit stereotypes, characterizing Whites as more compliant and 'better patients' than Blacks." This led us to want to verify these studies by finding evidence in the CDC Chronic Disease dataset. Further studies found that "African Americans, Hispanics, and Native Americans are 50-100% more likely to experience an illness or mortality from diabetes than white Americans" (Chow *et al.*, 2012). Due to this fact, we wanted to focus our research towards disparities between African Americans and other races. We also chose to limit our data analysis to normalized mortality rates as this metric is more robust to the influence of confounding factors and bypasses established evidence that, according to the Mayo Clinic, African Americans are more likely to contract diabetes, although there is not yet evidence that they are more at risk to die from it.

Edward A. Chow, MD, Henry Foster, MD, Victor Gonzalez, MD and LaShawn McIver, MD, MPH. The Disparate Impact of Diabetes on Racial/Ethnic Minority Populations.
Clinical Diabetes 2012 Jul; 30(3): 130-133. https://doi.org/10.2337/diaclin.30.3.130

Penner, L. A., Blair, I. V., Albrecht, T. L., & Dovidio, J. F. (2014). Reducing Racial Health Care Disparities: A Social Psychological Analysis. Policy Insights from the Behavioral and Brain Sciences, 1(1), 204–212. https://doi.org/10.1177/2372732214548430

Data Cleaning and Processing

```
clean_df = df.drop(["Response", "ResponseID", "StratificationCategory2", "Stratification
2", "StratificationCategory3", "Stratification3", "StratificationCategoryID2", "Stratific
ationID2", "StratificationID3", "StratificationCategoryID3"], axis=1)
cleaned['DataValue'] = pd.to_numeric(cleaned['DataValue'], errors = 'coerce')
cleaned['DataValueAlt'] = pd.to_numeric(cleaned['DataValueAlt'], errors = 'coerce')
cleaned = cleaned.dropna(subset = ['DataValue', 'DataValueAlt'], how='all')
diabetes = cleaned.loc[cleaned['Topic'] == 'Diabetes'].reset_index(drop=True)
```

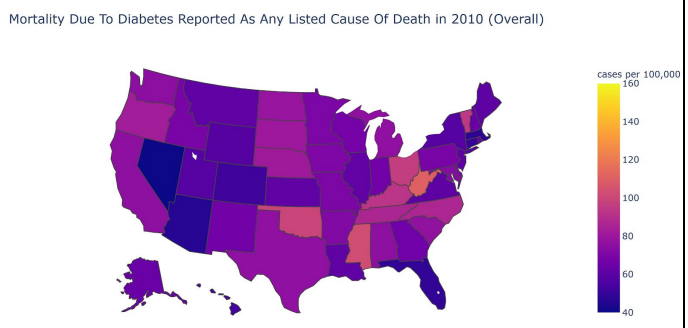
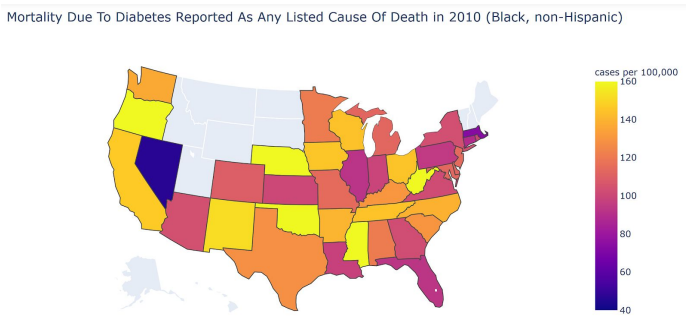
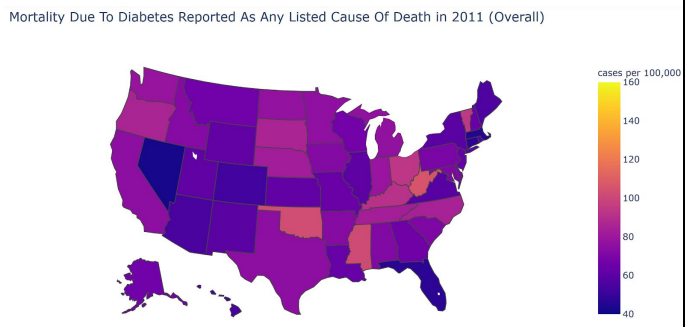
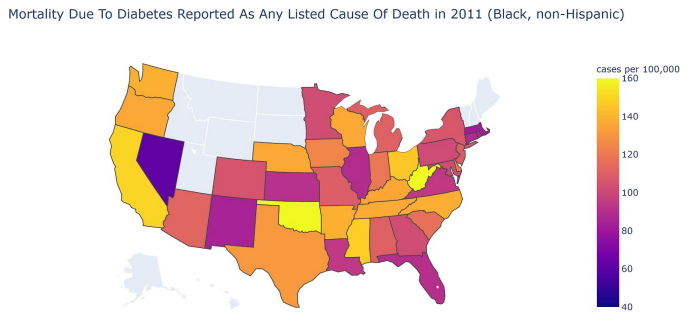
Assumptions

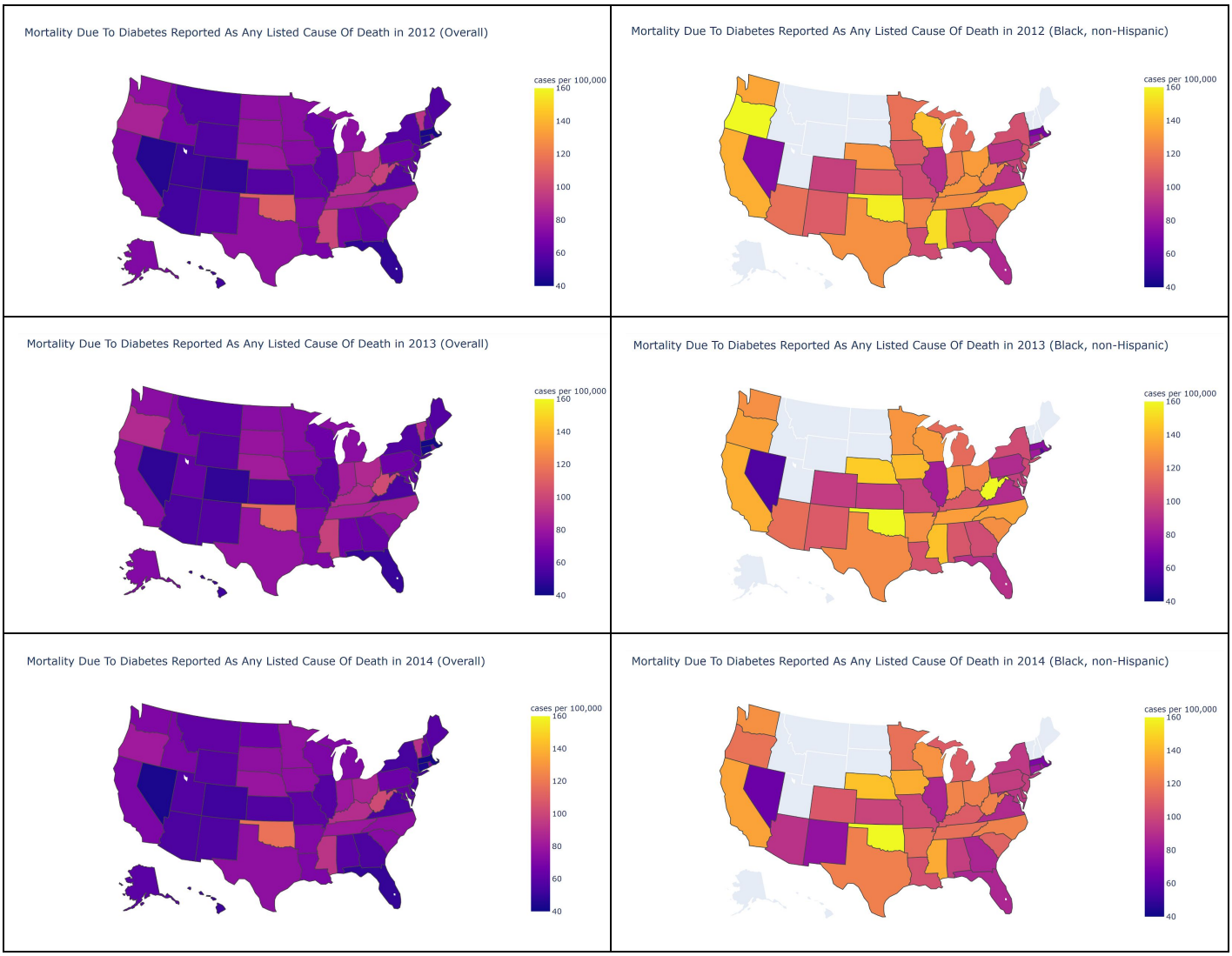
For our data cleaning process, we dropped columns that contained solely null or blank values which we thought of as meaningless. The columns that we dropped are: Response, ResponseID, StratificationCategory2, Stratification2, StratificationCategory3, Stratification3, StratificationCategoryID2, StratificationID2, StratificationCategoryID3, and StratificationID3. We dropped most Stratification columns because Stratification1 contained 0 null values and gave us enough information to help us identify a person's ethnicity or gender. Moreover, we dropped DataValueFootnoteSymbol because it contained weird symbols which were incomprehensible. We also dropped columns like TopicID, DataValueTypeID, and LocationID because we usually referred to these columns by their original name, e.g. Topic, DataValueType and Location, instead of their IDs. We deleted rows from the dataframe that contained null values for both DataValue and DataValueAlt because we knew that if both of these values did not have data, the row wouldn't be useful in data analysis. We dropped the non-mainland states like 'US,' 'DC,' 'PR,' and 'GU' because it cannot be graphed on the map. We were more concerned about the

states within the US. One challenge we encountered was when we were trying to delete rows that contained null values for DataValue and DataValueAlt, we wanted to delete rows that were null for both columns, not either or. We resolved this issue by using ".dropna(subset=['DataValue, 'DataValueAlt], how = 'all')" to delete NaN values that are in that subset.

Visualizations

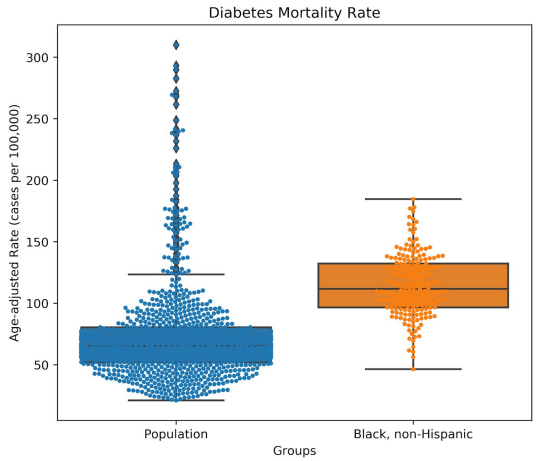
Overall vs. African-American Age-Adjusted Diabetes Mortality Rate (2010 - 2014)

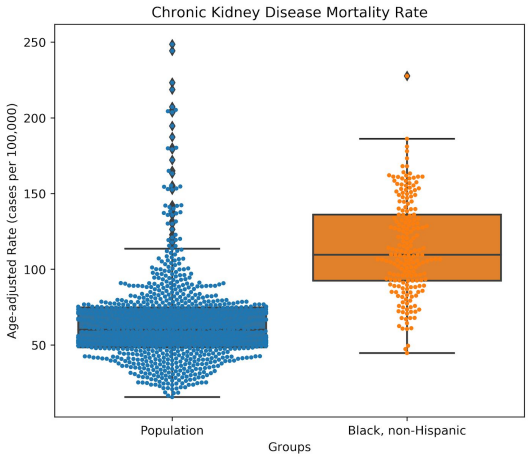
Overall Stratification	Black, not Hispanic Stratification
	
	

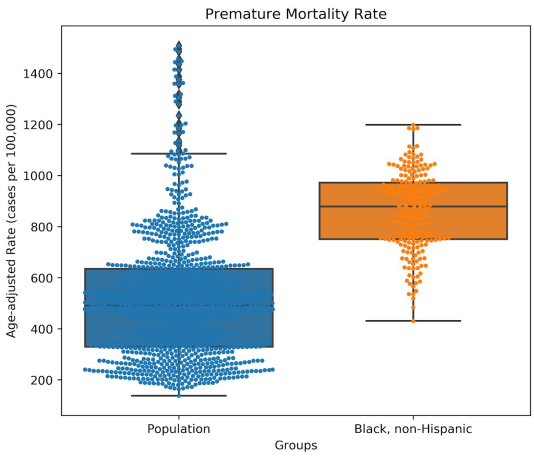


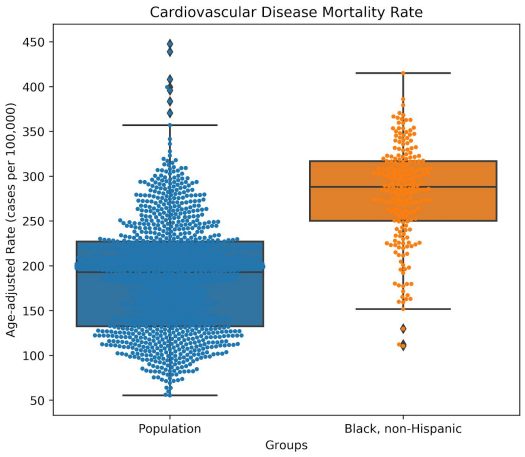
Looking at the Diabetes question, "Mortality due to diabetes reported as any listed cause of death", across 2010 to 2014, we saw that there was a significant increase this statistic in the African-American, or 'Black, non-Hispanic', stratification across the entire United States. We used the Plotly Express library to create choropleths of the age-adjusted diabetes mortality rate, comparing the overall and African-American stratifications. Looking at the choropleths, we can see there is a clear difference between the two stratifications. African-Americans seem to be dying from diabetes at a higher rate across the country.

Black, non-Hispanic vs. Population Mortality Rates across Diseases, Conditions ($p < 0.05$)









While the focus on our analysis was on African Americans within the topic of diabetes, we also noticed that the African-American population had significantly greater mortality rates across many different diseases. Overlaid box & swarm plots clearly show this significant difference. This supports our claim that there are systemic issues in the U.S. that may have a fatal effect on the African-American population.

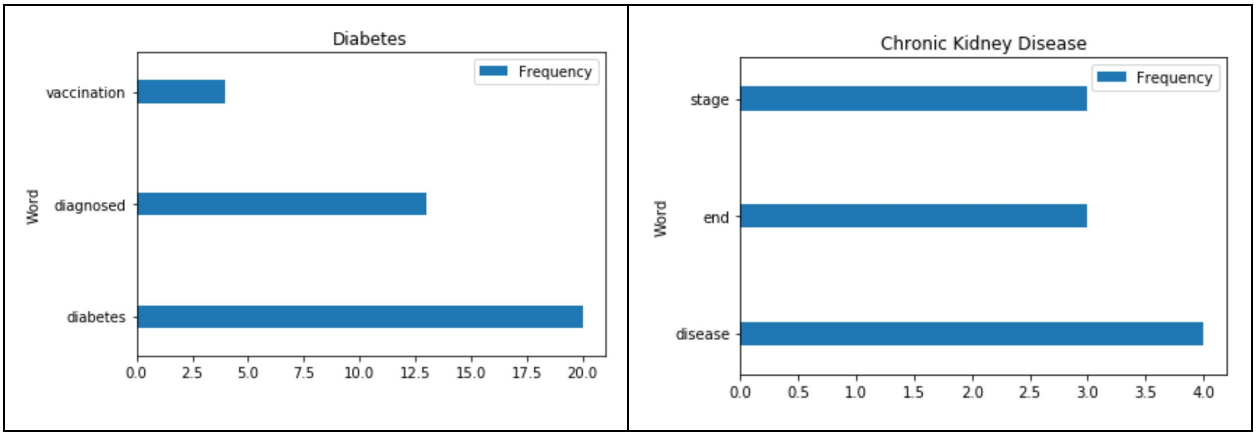
NLP Text Frequency Analysis

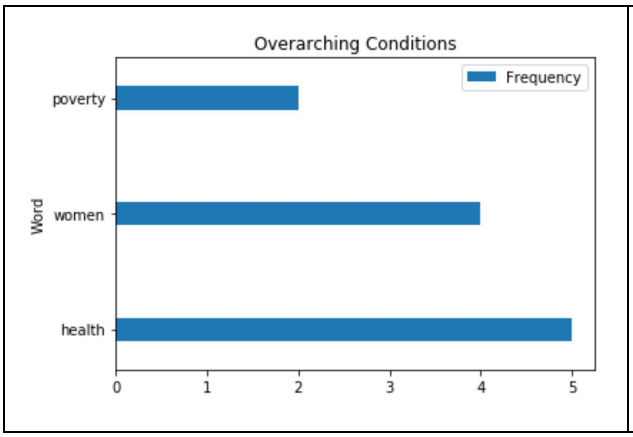
Topic	**Most Common Words**
Alcohol	'alcohol', 'binge', 'drinking', 'excise', 'tax', 'beverage', 'drink', 'heavy', 'women', 'use'
Nutrition, Physical Activity, and Weight Status	'school', 'students', 'high', 'physical', 'activity', 'care', 'obesity', 'daily', 'consumption', 'education'
Cardiovascular Disease	'heart', 'disease', 'mortality', 'high', 'coronary', 'stroke', 'blood', 'pressure', 'vaccination', 'noninstitutionalized'
Arthritis	'arthritis', 'doctor', 'diagnosed', 'limitation', 'diabetes', 'physical', 'inactivity', 'obese', 'fair', 'poor'
Diabetes	'diabetes', 'diagnosed', 'vaccination', 'noninstitutionalized', 'listed', 'examination', 'mortality', 'reported', 'cause', 'death'
Tobacco	'tobacco', 'smoking', 'cigarette', 'smoke', 'smokeless', 'use', 'states', 'youth', 'pneumococcal', 'vaccination'
Overarching Conditions	'health', 'women', 'poverty', 'self', 'rated', 'status', 'insurance', 'high', 'school', 'completion'
Chronic Obstructive Pulmonary Disease	'chronic', 'obstructive', 'pulmonary', 'disease', 'diagnosis', 'diagnosed', 'hospitalization', 'first', 'listed', 'vaccination'
Cancer	'cancer', 'incidence', 'mortality', 'invasive', 'use', 'women', 'female', 'papanicolaou', 'smear', 'cervix'
Chronic Kidney Disease	'disease', 'end', 'stage', 'renal', 'incidence', 'treated', 'mortality', 'chronic', 'kidney', 'attributed'
Asthma	'asthma', 'vaccination', 'noninstitutionalized', 'rate', 'influenza', 'pneumococcal', 'hospitalizations', 'emergency', 'department', 'visit'
Disability	‘disability’
Oral Health	'dental', 'visits', 'teeth', 'lost', 'health', 'preventive', 'children', 'adolescents', 'water', 'six'
Older Adults	'proportion', 'older', 'date', 'core', 'clinical', 'preventive', 'services', 'medicare', 'persons', 'hospitalization'
Mental Health	'mentally', 'unhealthy', 'days', 'least', 'women', 'postpartum',

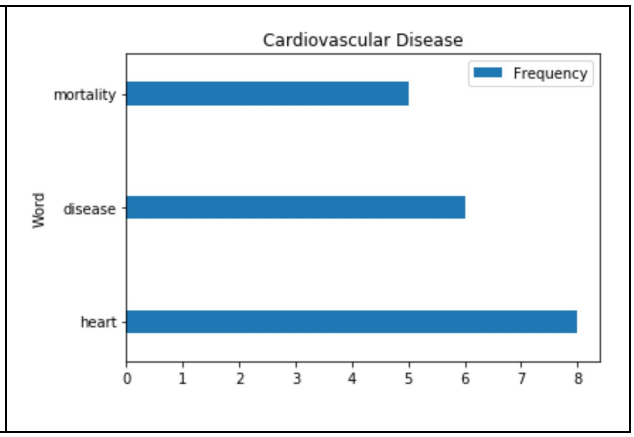
	'depressive', 'symptoms'
Immunization	'influenza', 'vaccination', 'noninstitutionalized'
Reproductive Health	'checkup', 'timeliness', 'routine', 'health', 'care', 'women', 'postpartum', 'folic', 'acid', 'supplementation'

Using Regex and NLTK, we analyzed the questions in each topic to find the top ten most common words in each topic. The main complication we ran into was performing analysis on limited text data. For topics like ‘Immunization’ and ‘Disability’, there was only one unique question, so the text data for those topics were just one sentence. This made it difficult to find frequent keywords among all 17 topics because the usable text data was so limited. We then used matplotlib to plot the top 3 most frequent words in the four topics we explored above, shown below.

Top 3 Most Frequent Words Horizontal Bar Chart







Hypothesis Testing

Research Question

We wanted to see if there was a difference in Age-adjusted Rate between African Americans and other races which could have contributed to mortality due to diabetes reported as any listed cause of death. Therefore, we asked the question: **Are African Americans more likely to have a higher rate of mortality due to diabetes than other races?** More formally, we statistically tested whether the mean age-adjusted rates of mortality (deaths per 100,000) between Black, non-Hispanic Americans and the remaining population were different.

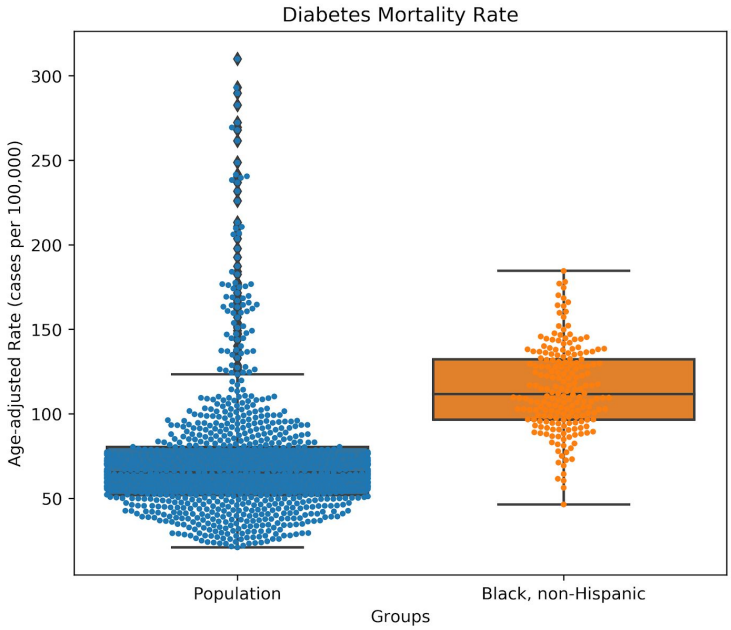
Null Hypothesis

African Americans have the same mean Age-adjusted Rate of diabetes as the population.

Alternative Hypothesis

African Americans have a higher mean Age-adjusted Rate of diabetes than the population.

Analysis



For our implementation, we made our observed mean to be the mean of the data values that were Age-adjusted Rates for people who had 'Black, non-Hispanic' as their 'Stratification1.' We made 1000 trials where we randomly sampled from data values of Age-adjusted Rates from the whole population with replacement. We would find the mean values of these rates and append it to our averages list. Then, we performed an unpaired t-test on the two groups, African-American and Overall stratifications. Our p-value was 0.0 which is less than the significance level of 0.05. Therefore, we rejected the null hypothesis. Our t-statistics were also consistently positive, so we conclude that African Americans have a significantly higher Age-adjusted Rate of mortality due to diabetes compared to the rest of the population.

Conclusion

There is established evidence of implicit biases in the American healthcare system that disfavor African Americans. Analysis of the CDC Chronic Disease dataset shows that African Americans consistently face significantly higher mortality rates from diabetes across both geography and time. Outside diabetes, African Americans are also at greater risk for death from chronic kidney disease, cardiovascular disease, and are more likely to face premature death than the population. From our analysis, we conclude that African Americans are significantly more likely to die from chronic illness and disease. This strengthens the body of evidence that points to systemic social justice issues in U.S. healthcare.

Research Proposal

During our data analysis, we quickly realized that the CDC dataset was too limited to allow us to rigorously determine possible causes for why there is a racial bias against African Americans in the US medical system. We suggest that more data collection and research into the medical care of African Americans is necessary to understand the mechanisms that drive the mortality trends we discovered. Specifically, data related to patient care in the clinic and how physicians treat different racial and ethnic groups would be very informative for understanding this issue and how to resolve it.

Further Study

If we were to further improve our research, we would want to use NLP to help predict some classification problems such as what the topic is based on the most common words. We

would also want to test if Hispanics or Native Americans were more likely to receive higher mortality rates than white Americans to confirm that there is racial bias against minorities. Then, we could find any discrepancies in their nutrition, physical activity, and weight status that could contribute to mortality rates for diabetes among minorities.

We also recognize limitations of our study, as there are variables in the dataset that are known risk factors for diabetes that can be controlled for in a future study. However, this analysis was beyond the scope of our work.