Socio-economic Inequality Due to Race Fails to Explain Variation of Cancer Mortality in US counties[*]

A multivariate regression study to examine the marginal effect of varying racial majorities on mortality due to cancer

Anshuman Agarwal

29 April 2022

Abstract

This paper aims to investiagte effect of the white privilige of a US county - measured by the concentration of white residents in the county; and their rate of mortality due to cancer. The paper analyses data from the National Cancer Institute and the US Census to create a multivariate model with age-adjusted mortality due to all types of cancer per 100,000 residents, as the response and the percentage of white residents of the county as the main predictor, while controlling for age-adjusted incidence rate, poverty, income, medical insurance, population, and social vulnerability. The findings suggest that while people of colour are unfairly burdened with health and socio-economic inequality in America, more white residents doesn't exactly associate with a lower cancer mortality rate.

Contents

[*] Code and data are available at: https://github.com/anshumanagarwal27/Cancer-Mortality-and-Race

1 Introduction

The United States of America is a country riddled with many socio-economic disparities that arise due to systemic racial biases, and one such area where such inequality is observed is healthcare. Accessibility to quality healthcare and medical insurance varies significantly for non-white people, which is made worse by the increasing wealth gap between the communities. Thus, when it comes to treating life threatening illnesses such as cancer, affordability of medical care becomes a key factor which determines if a person's battle against cancer is going to be a successful one. Furthermore, white communities have observed faster growth rates when compared to non-white communities, and on average also see greater development of resources such as specialty hospitals making essential care more available in some parts of the country than others. Due to varying standards of care, it is not surprising to see that mortality rates of cancer incorporate significant geographical variation, which is not entirely explained by the variation of the corresponding incidence rates.

Research has been done in the past to investigate the implication of socio-economic inequalities on healthcare, and cancer mortality is a popular topic in this field. For the US, papers from (Moss et al. 2020) and (Boscoe et al. 2014) study the effect of poverty on the cancer mortality rate of US counties using cross-tabulations and found that poverty was significantly associated with a greater incidence of specific cancers, which in turn corresponded to higher mortality rates. The findings of (Brown et al. 2021) and (Palencia et al. 2020) confirm the positive association found between poverty and cancer mortality in samples outside the US, while research in (Teng et al. 2017) point toward the significant share of social vulnerability in increasing the likelihood of mortality due to cancer. Past literature talks about how poverty and income is very significant to understanding mortality from cancer, however to my knowledge, research in the past do not mention race as a possible predictor.

Therefore, the aim of this study is to understand how the racial concentration of a US county explains the geographical variation in mortality rates due to cancer, after accounting for key determinants such as incidence rates, poverty levels, subscription of medical insurance, median income and the overall social vulnerability of the county.

1.1 Hypothesis

The aim of this analysis is to investigate if the health inequality caused by the systemic racial inequality in the United States can be used to explain the geographical variation in the cancer mortality rates for counties in the US. More, specifically, the analysis looks to estimate the marginal effect of varying concentration of white residents in a county on the age-adjusted mortality due to cancer per 100,000 residents, while controlling for known confounders such as incidence rate of cancer, poverty levels, health insurance coverage, median household income, the relative social vulnerability, and the population of the county.

Weighing in research done in the past on decomposing the geographical variation of cancer mortality rates, as well as research done in the field of health and racial inequality, the study makes the following hypotheses. Firstly, higher death rates in a county would be associated with higher poverty levels of the county, higher concentration of uninsured residents in the county, lower median household income of the county, higher social vulnerability of the county and lower population of the county. Secondly, poverty levels, concentration of uninsured residents and relative social vulnerability would decrease with an increasing concentration of white residents in the county, while median income would increase at the same time. Following this notion, it can be hypothesized that there would be negative relation between increasing white privilege and cancer mortality rates, meaning that counties with a higher concentration of white residents should on average report lower mortality rates due to cancer.

2 Data

2.1 Statistical Software

This report analyses data collected by the US Census Decennial Census 2020, American Community Survey 2015-2019, and data collected by the Surveillance, Epidemiology and End Results (SEER) program which is published by the National Cancer Institute (NCI) to estimate the marginal effect of variation in the concentration of white residents living in a US county on the age-adjusted mortality rate due to cancer. The statistical analysis done in this study which includes creation and cleaning of raw data, data visualization and linear regression analysis, was done using `R Statistical Programming Language` (R Core Team 2021). Apart from the base features present in R, the study also made use of numerous publicly available statistical packages. The dataset was created and cleaned using the package `tidyverse` (Wickham et al. 2019). The amalgamated dataset was then analysed using `dplyr` (Wickham et al. 2022), and robust linear regressions were done using `robustbase` (Maechler et al. 2022). Data was visualized using `ggplot2` (Wickham 2016), and tables of summaries were created using `modelsummary` (Arel-Bundock 2022) and `kableExtra` (Zhu 2021). This report was written in *R Markdown* and was formatted using `knitr` (Xie 2022b), `bookdown` (Xie 2022a), and `extrafont` (Chang 2022). Reproducibility was ensured using `here` (Müller 2020).

2.2 Description of the Dataset

The dataset analysed to construct the findings of this report was constructed by combining necessary columns from publicly accessible meta-datasets published by branches of the United States government. The cleaned dataset contains information on 2885 US counties (in the US mainland), where each instance in this cross-sectional dataset is a county and has information about this county in the form of 11 variables. Table 1 shows a slice of 10 observations from the full dataset as a preview.

Table 1: A slice of the Cancer Mortality Study Dataset

FIPS	pop2020	pct_w	pct_pov	med_inc	pct_unins	incidence	death	svi
1023	12665	55.85472	23.3	35892	10.3	447.9	147.8	0.8748
1025	23087	52.10292	23.2	37404	9.7	471.3	175.8	0.8471
1027	14236	79.90306	17.9	40845	10.3	478.9	183.3	0.5640
1029	15056	91.78401	17.3	44741	12.3	440.2	146.2	0.5417
1031	53465	69.35378	15.1	55637	9.8	417.5	163.3	0.5408
1033	57227	76.24198	15.5	48065	7.2	442.7	178.4	0.4274
1035	11597	50.97870	17.6	37837	12.7	451.7	180.9	0.8086
1037	10387	65.69751	13.5	38990	8.0	359.0	163.2	0.5261
1039	37570	82.18525	18.6	42189	11.1	456.3	189.7	0.7723
1041	13194	71.15355	15.2	43163	8.7	534.8	203.7	0.6873

The first column records the unique geo-identifier code given to each US county known as the FIPS code. In some datasets which report county-level information, the FIPS code is replaced by the full geo-id of the county, however the FIPS can be easily extracted from the geo-id of the county, as it is the last 5 digits of the geo-id. This code is important in the context of the creation of the dataset used for the analysis as necessary columns from separate meta-datasets were merged based on this code. More information about the creation of the dataset can be found in GitHub repo linked in this report.

The second column contains information about the total population of the county in 2020. The third column contains information about the primary predictor which is the percentage of white (non-Hispanic) residents in the county. The column representing the total population of the county was taken from one of the datasets, `P1`, made from the data collected by the United States Decennial Census conducted in 2020. This dataset can easily be downloaded from the US Census website. The meta-dataset contained information about the total population (column code `P1_001N`) and the population of all the different races. The variable `pct_w` of the complied dataset was constructed by using the information in `P1_001N` and the total white (non-Hispanic) population of the county (column code `P1_003N`).

The fourth column contains information about the percentage of residents in the county who live below the federal baseline poverty level. The fifth column records the median household income of the county. The sixth column records the percentage of residents (non-institutionalized) without health insurance. These variables represent columns taken from various datasets constructed from the data collected by the American Community Survey for the years 2015-2019 (Table IDs are - `S1701`, `S1903` and `S2701` respectively). These datasets are also available for public use in the US Census website and can be found by inputting the Table ID in the search bar. The variable `pct_pov` is column `S1701_C03_001E` of the dataset `S1701`, which contains information about the poverty status in the past 12 months. The dataset also has information based on various categories such as age, employment, sex, and race. The variable `med_inc` is column `S1903_C03_001E` of the `S19031` dataset. This dataset reports information about the median income of the households in US dollars, in the past 12 months. Like the poverty dataset, it too contains this information for various categories. Lastly, the variable `pct_unins` is column `S2701_C05_001E` of the `S2701` dataset. This dataset contains information about health insurance coverage characteristics for different categories. For this analysis, the percent uninsured for the total civilian non-institutionalized population was used.

The seventh column contains the age-adjusted incidence rate of all types of cancer per 100,000 residents in the county. The eight column hosts the primary response variable in the analysis which is the age-adjusted mortality rate due to all types of cancer per 100,000 residents in the county. These variables were downloaded individually from the National Cancer Institute (NCI) websites as county level data, for the entire United States (excluding Puerto Rico), for all ages, sexes, and races. The data on age-adjusted incidence rates were collected by the SEER program, the National Program for Cancer Registries Cancer Surveillance System (NPCR-CSS) and the CDC and were made available to the public by the NCI. The data on the age-adjusted death rates were extracted by the NCI from data collected and reported by the National Vital Statistics System.

The ninth column of the dataset contains the Social Vulnerability Index rank of the county as calculated by the Agency for Toxic Substances and Disease Registry (ATSDR) wing of the CDC. The Social Vulnerability Index is an aggregate ranking that the CDC gives to every US Census tract, using US Census data on 15 social factors such as poverty, access to a vehicle, etc, and groups this data in four related themes – Socioeconomic, Household Composition and Disability, Minority Status and Language, and Housing Type and Transportation ("CDC/ATSDR Svi Fact Sheet" 2021). Using the information from these 15 factors, the CDC provides a rank to each census tract for each category, before assigning each census tract with an overall rank for their relative social vulnerability. These ranking are based on percentiles and range from 0-1, with an increasing rank representing an increase in vulnerability ("CDC/ATSDR Svi Fact Sheet" 2021). These rankings, for all categories are available for the public to download as one dataset from the CDC website. For this analysis we use the overall composite index which has the column code `RPL_THEMES`.

2.3 Sources of Limitations from the data

Majority of the data used to construct models in this study was collected by the American Community Survey 2020, which publishes data as 5-year estimates for 2015-2019. According to the US Census Bureau, errors arise in the data collected by the ACS in due to sampling error and non-sampling error. Sampling error arises due to the incorporation of probability sampling in the survey methodology, and this was accounted for by including confidence intervals and margin of error estimates with the data. Non-sampling error arises due to various reasons such as data-entry error, or systematic sources such as non-response bias, because unlike the Decennial Census which surveys the entire population of the United States, the ACS surveys a selected sample, thus the data representative of the respondents of the survey might be different for the non-respondents, and for the 2020 ACS survey, this effect was heightened due to the lower response rates that arose as a result of the COVID-19 pandemic (Bureau 2022). The Census Bureau aims to correct this non-response bias by adjusting the weights in the survey methodology such that the age and race statistics match those of the population estimates acquired from the decennial census. (Bureau 2022). Additionally, coverage bias also becomes an issue as the entire population is not being surveyed and hence some households could be omitted.

Data on the incidence and death rates also carry some limitations with them. Due to lack of availability, data on incidence rates are compiled from different sources. For both death and incidence rates data, availability issues also pose a challenge to the time period coherence in the data, as time periods may differ for some counties. Lastly, for both incidence and death rates, data was suppressed for counties recording annual counts less than 16 incidents and deaths respectively, to ensure confidentiality and stability of the estimates of the rates. Thus, data on incidence and death rates are not published for such counties and the study needs to account for the issue of missing observations. Thus, the dataset used for the analysis only contains information on 2885 counties.

3 Methodology

To test the hypotheses stated above this study will employ the use of a robust multivariate regression model to fit the data and make inferences about the marginal effects of the predictors in the model. The multivariate regression model uses the OLS framework to estimate an equation which is linear in its parameters:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \mu$$

Here, Y is the age-adjusted mortality rate due to cancer per 100,000, and X_k are the predictors outlined in the section before. The β_k represents the marginal effect of each predictor, which is the effect of a unit increase in the predictor on the response, while holding the influence of all the remaining predictors constant. μ is the residual which quantifies the effects of the variables that belong in the model but were not added. Most of the times the residual contains variables which are tough to quantify or unfeasible to collect data on, or are just unobservable.

As mentioned in Section 2.3, the data has missing observation due to data suppression by the NCI, so prior to any analysis, those observations were removed, bringing the total sample size of the data to 2885. In the first step of the analysis, scatterplots are used to assess the relationship between mortality rates and the predictors, and to also see the relationship between percentage of white residents and the remaining predictors in the model. In the plots, a linear trendline is used to visually assess these relationships. Scatterplots also come in handy to see obvious violations of the Gauss-Markov assumptions which are crucial for making inferences from the estimates obtained from the OLS process. This should also provide some information regarding model selection and if it is feasible to remove a variable from the model without the loss of statistical significance. Linear regression framework is also notorious for not being robust to the presence of influential observations, thus scatterplots also help identifying if there are such points that need to be addressed.

A robust simple linear regression (SLR) model is also considered to see if percentage of white residents alone has a statistically significant association with cancer mortality rates. Furthermore, a SLR model can help us validate the use of a multivariate linear regression (MLR) model if model assumptions fail to hold, and subsequently point out if there is omitted variable bias in using just percentage of white residents as a sole predictor. Finally, the analysis uses a robust MLR model as shown above, to create the model whose findings this analysis is most interested in. Robust linear regressions are chosen primarily because the assumption of homoscedasticity is unrealistic and restrictive, and under if this assumption is violated, the standard errors of this estimates would be biased, and results of subsequent hypothesis tests would be unreliable. Robust linear regression relaxes this assumption, and in turn provides accurate estimates of the standard errors, while preserving the size and direction of the estimated marginal effects to a great extent. Under the framework of the robust linear regression, results of hypothesis tests can be trusted, and inference can be made without the risk of bias.

4 Results

4.1 Data Visualization and Scatterplots

To see how cancer mortality rates and the concentration of white residents in a county relate, a scatterplot was constructed with mortality on the y-axis and percentage white residents on the x-axis. This scatterplot can be found as Figure 1 in Appendix A. The first result that can be taken from this graph is about the spread of the data points. Data seems to be more spread-out for smaller percentages of white residents and they get concentrated as we move right on the x-axis. This is indicative of the fact that most of the counties in the sample have a majority of white residents living in them (greater than 50%). This can also be observed in the subsequent figures which plot percentage of white residents on the x-axis. Furthermore, no concerning patterns are noticed, however it won't to incorrect to assume that the assumption of homoskedasticity would hold. 2 counties can be identified as having abnormally high mortality rates (greater than 350 per 100,000), and the trendline seems to be slightly upward sloping, however no definite signs of an association can be noticed right off the bat.

Moving onto the relationship between incidence rates and mortality rates (see Figure 2), the scatterplot which plots the mortality rates on the y-axis and the incidence rates on the x-axis (Plot 1, Figure 2) clearly shows that incidence rates and mortality rates have a strong positive correlation. Most counties fall between the range of 200 incidents to 600 incidents per 100,000, and the two outliers seen in the previous figures are also outliers in terms of incidence, thus their high death rates can be attributed to high incidence rates. (Plot 2, Figure 2) shows a scatterplot for the relationship between the percentage of white residents and incidence rates, with the latter being plotted on the x-axis. As seen in Figure 1 the data points for this plot are spread out in the same way. Incidence rates also show signs of a slight correlation with percentage of white residents, however no definite evidence is present.

For mortality rates and percentage of residents without health insurance, (see Figure 3). In (Plot 1, Figure 3), the overall scatter of data in the plot indicate that there is no relationship between mortality rates and lack of health insurance. The furthest outlier county which has a death rate greater than 400 deaths per 100,000, and an incidence rate greater than 1000 cases per 100,000, has only about 10% of residents without health insurance. The trendline has a slight upward slope, but that does not provide a definitive result about the correlation between these two variables. Outliers in terms of counties with abnormally high percentage of residents without medical insurance, can be seen (with almost 40% of residents without insurance) however these counties have a death rate close to the average. (Plot 2, Figure 3) shows the relationship between percent without insurance and percent white residents in a county. Here, a clear negative correlation is noticed, by the trendline and by how the data is spread out. Most of the outliers noticed in Plot 1 are found to be counties with less than 30% white residents, however the furthest outlier is also extremely concentrated with white residents. As we move right on the x-axis, the data points become congested, indicating that whiter counties on average have higher coverage rates.

Furthermore, when poverty is examined in a similar way, (see Figure 4), a positive correlation is noticed by the upward sloping trendline as well as the spread of the data ((Plot 1, Figure 4). The trendline seems to be heavily influenced by the presence of two outlier counties with almost 50% residents living below poverty level, thus the strength of the correlation might be stronger than what is seen here. This result validates findings of past research. The two outlier counties with abnormal mortality rates seem to have average percentage of residents living below poverty, with most data points being congested towards the bottom left corner of the graph, indicating that majority of counties in the sample have smaller levels of residents living in poverty. This result ties in with the fact that majority of counties in the sample are majorly white, as seen in (Plot 2, Figure 4) the relationship between percentage living below poverty and the percentage of white residents is a negative correlation. The spread of the data is a mirror image of the adjacent graph, where points become congested to the bottom right indicating that as counties become whiter, on average the percentage of residents living below poverty falls. The downward sloping trendline confirms all the other hints seen in the plot. Homoscedasticity is very unlikely to hold in any of the four relationships we have assessed so far due to systemic clustering of data points. It is not wrong to assume that this issue will continue to show up in the remaining figures too.

When median income is looked at, (see Figure 5), the notoriously skewed nature of this variable can be clearly seen in both the figures. This is why a non-linear curve is noticed in (Plot 1 and Plot 2, Figure 4). Furthermore, the 2 outliers counties with very high death rates have lower than average median household income, however getting a sense of the central tendency from a scatterplot of a variable with high right skew can be misleading. The trendlines for both the plots are misleading as clearly the relationship between these variables is not linear. But if past literature and the evidence in front of us are consulted simultaneously, it won't be a spurious assumption that in a multivariate model, median income will be an important addition and removal of it could seriously affect the overall fit.

The plots for the relationship between our variables of interest and the social vulnerability index provide some interesting findings (see Figure 6). (Plot 1, Figure 6) shows that there is a strong positive correlation between worsening social vulnerability and high mortality rates. The spread of the data and the upward sloping trendline exhibit a textbook linear relationship, so much so, the two outlier counties we have been talking about have very high social vulnerability (between 0.75 and 1). (Plot 2, Figure 6), shows that on average counties with higher white residents are less socially vulnerable. Most of the counties which don't have a white majority are found alarmingly close to the upper bound of the SVI (rank of 1.00), and as we move onto a county which is majorly white, their SVI rankings fall. Thus, the association between higher mortality and higher vulnerability does exist, alongside the association between higher white residents and lower social vulnerability.

Lastly, the relationship between population and the variables of interest don't provide any useful findings (see Figure 7). The population of the counties seem to be extremely right skewed, and have no observable relationship with death rates (Plot 1, Figure 7). The same can be said about the relationship between population and percentage of white residents (Plot 2, Figure 7). Population seems to be exogenously spread out for all levels of white concentrations, due to the extreme left skew. Controlling for population however seems to be warranted by past literature.

4.2 Regression Analysis

Keeping the results found in the previous section in mind, 5 regression models were fit. The model summaries for these five models can be found in Table 2 in the next section. The first model examines the simple linear model with death rates as response and percent white residents as the predictor. The second third and fourth models use a subset of chosen predictors; excluding median income in model 2, population in model 3, and both median income and population in model 4. Model 5 uses the entire set of predictors examined in the section above.

The first model looks at a simple linear regression model for percent white residents explaining death rates. This regression corresponds to the green trendline seen in Figure 1. Immediately, the first thing to notice in the model summary is the R-sq value of 0.001, which indicates that only 0.1% of the variation in death rates was explained by the variation in percentage of white residents in a county. Clearly the model suffers from omitted variable bias as the residual explains 99.9% of the variation in the response. Furthermore, the estimate for the slope is not statistically significant to any feasible levels of statistical significance.

The second model looks at a multivariate linear regression model with death rates as the response and a subset of all the predictors in the analysis excluding median income of the households. The inclusion of 5 additional predictors have increased the overall R-sq of the model to 0.535, meaning that now the model explains about 53.5% variation in the response and the residual explains 46.5%. Furthermore, the estimate of the slope coefficient on percentage of white residents in a county is not statistically significant to all feasible levels of significance. All the estimates for the marginal effects of other predictors in the model are statistically significant. The estimated effect of population is approximately 0, and it is statistically significant to a 95% level of significance. Off the bat, the key result to note from this summary is that the direction of the estimate on the percent white residents is positive, which is opposite from what was hypothesized.

The third model looks at a multivariate linear regression model with death rates as the response and a subset of all the predictors in the analysis excluding population of the county. The inclusion of median income and the exclusion of population has increased the overall R-sq to 0.540, which is marginally better than model 2. Like model 2, all the estimates in this model are statistically significant, and the estimate on population is 0 while being statistically significant to the 95% level. The estimate on the percent of white residents is highly significant, and while the direction of the estimate is preserved, the estimate is smaller than that of model 2, while also having a smaller standard error.

The fourth model is similar to models 2 and 3, however in this case both median income and population are excluded from the model. The exclusion costed the total explanatory power of the model, with a R-sq of 0.532, which is lower than both models 2 and 3. All the estimated marginal effects are highly statistically significant, with the estimate of the effect of percent uninsured becoming significant to a 99.9% level compared to 99% level for model 2 and a 95% level for model 3. There are no dramatic changes to the sizes and directions of all the estimates in the summary, however the estimate on percent white residents has increased, while its standard error has decreased. The estimates in this model have the lowest standard errors relative to any model in Table 2.

The fifth model looks at a multivariate linear regression model with death rates as the response and a subset of all the predictors in the analysis. This model has the highest R-sq out of any models in Table 2, with an R-sq of 0.541. This is only marginally higher than that of model 3 and can be considered comparable with those of models 2 and 4. All the estimates in the model are highly significant, however inclusion of both median income and population seems to have decreased the significance of the estimate on percentage uninsured to 95% from 99.9% in model 4. The estimate on percent white residents is the lowest in comparison to models 2,3 and 4, while the standard error of the estimate is ever so slightly larger than models 3 and 4. Overall, this model, like the models before it, suffers from omitted variable bias. However, the results of the regression analysis have provided enough evidence that the hypothesis made in Section 1.1 can be tested.

The analysis found evidence to support the hypothesis made about higher mortality rates being positively associated with higher poverty levels (0.632), as when other factors are controlled for, on average a 10-percentage point increase in the percentage of residents living under poverty is associated with a rise of 6.32 deaths per 100,000. This result is highly significant so the null hypothesis can be comfortably rejected. Higher mortality rates are also positively associated with poorer health insurance coverage (0.222) and higher social vulnerability (24.559), and their significance makes it so that their nulls can be rejected. Furthermore, the null stating that percentage of white residents don't affect mortality rates can be rejected as the estimate is highly significant, however the hypothesis about the direction of the estimate was rendered incorrect.

5 Discussion

The discussion done in this section is based on the results found from Figures 1 to 7 in Appendix A, and the table of model summaries for the models mentioned above. This table is reported below.

Table 2: Model Summaries of Regression Analysis

	Model 1	Model 2	Model 3	Model 4	Model 5
Percentage of White Residents	0.043	0.352***	0.287***	0.409***	0.262***
	(0.035)	(0.038)	(0.034)	(0.027)	(0.035)
Age-adjusted Incidence rate per 100,000		0.235***	0.236***	0.234***	0.237***
		(0.009)	(0.009)	(0.009)	(0.009)
Percentage without Health Insurance		0.265**	0.257*	0.329***	0.222*
		(0.102)	(0.101)	(0.098)	(0.102)
Percentage living below Poverty Level		1.057***	0.634***	1.135***	0.632***
		(0.116)	(0.150)	(0.112)	(0.149)
Social Vulnerability Index of the county		29.621***	24.055***	29.775***	24.559***
		(2.601)	(2.794)	(2.615)	(2.770)
Population of the county		0.000*			0.000*
		(0.000)			(0.000)
Median Income of Households			0.000***		0.000***
			(0.000)		(0.000)
R2	0.001	0.535	0.540	0.532	0.541

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.1 Implication of socio-economic vulnerability

The topic for this analysis was chosen as it is well known that socio-economic inequality is prevalent between different races in the United States, which then translate to inequality in healthcare quality and access.

Given that getting treatment for cancer is expensive and in severe cases require specialised care which is only offered at few hospitals located in populous metropolitan cities. Thus, theoretically, differences in income and access to health care should then affect the likelihood of a cancer patients' chances of mortality. Even if incidence of cancer is considered majorly exogenous or correlated with biological unobservable such as genetic mutation, the dynamics of the county the resident is in should also affect a person's chances to get cancer in the first place. For example, a person is more likely to get lung cancer if they live in an industrial town, and more specifically a person of colour due to their propensity to work in low-income job fields such as factory labour. Additionally, specialty hospitals and overall development of hospitals should also be related to the racial make-up of the county, as if more white patients can afford specialty care than patients of colour, then economically speaking it would not be surprising to see the supply of specialised care to be centered around high demand areas. Thus, mortality due to cancer does not only depend on incidence of said cancer, but also the patient's socio-economic vulnerability, and the demographics of the patient's county of residence.

So, when the relationship between mortality rates and concentration of white residents of a county were directly tested (Model 1 in Table 2, Figure 1), no significant relationship was found. Unsurprisingly, the R-sq of model 1 was 0.001, indicating that it is important to account for omitted variables in the model to gain explanatory power. This was also supported by the results seen in Plot 1 of Figures 2, 3, 4 and 6, suggesting that incidence rate, percentage of residents living below poverty level, percentage of residents without medical insurance and the relative social vulnerability of the county are correlated with the mortality rate due to cancer. The population of the county and the median household income of the county didn't seem to affect death rates, probably due to the skewed nature of the data, but when they were included in the model (Models 2,3 and 5) their estimated statistically significant marginal effects were almost 0, however the overall explanatory power of the model increased from their inclusion. This probably happened due to median income and population being correlated with poverty, insurance coverage and the relative social vulnerability of the county, which in turn do affect the death rates. Furthermore, the socio-economic parameters in the model (poverty, insurance coverage and SVI) remain significant throughout models 2 to 5, and their inclusion in the model is not deemed unnecessary. This section of the discussion only validates what has been already found by previous literature, and how these factors might cause racial inequality in cancer mortality rates is discussed in the section below.

5.2 Implication of White Privilige

As explained in the section above, socio-economic parameters are necessary to explain the geographical variation in cancer mortality rates, when the geographical variation of cancer incidence is controlled for, however what usually goes undiscussed is if race plays an important role. Speaking from the intuition explained above, people may believe that white privilege applies to cancer mortality rates, as it is evidently correlated with higher income, better standards of living and relative ease to access specialty medical care, the findings of this study say otherwise.

The findings of Plot 2 in Figures 2, 3, 4 and 6, convey that the same socio-economic parameters that correlate with death rates, share the opposite association with how white the county is. For example, increasing poverty levels in a county correlated positively with higher death rates, while also being negatively correlated with the percentage of white residents in a county. Thus, ideally if the percentage of white residents were to increase for a county, then holding other factors constant, the county should see lower levels of poverty and subsequently also benefit from a lower death rate. This idea can be extended towards percentage uninsured, and relative social vulnerability as the data supports the association. However, looking at Model 5 (as it has the highest explanatory power relative to all the models), a 10-percentage points increase in the concentration of white residents in a county is on average associated with the mortality rate due to cancer being 2.62 deaths per 100,000 higher, while other factors affecting mortality rates are controlled for. This result is surprising, based on previous findings of the paper, and can be attributed to an incomplete model, as even with a R-sq of 0.541, almost 45.9% of the variation remains unexplained and this model too suffers gravely due to uncontrolled heterogeneity and omitted variable bias. Additionally, there are some limitations surrounding the data and the methodology used in this analysis, and if these limitations are accounted for, future research may reveal results suiting the intuition that racial and health inequality increases the likelihood of mortality from cancer.

5.3 Limitations and Scope for Future Analysis

In conclusion, this study found that an increasing concentration of white residents in a US county is associated with a higher mortality rate due to cancer, on average after controlling for the incidence rate of cancer and important socio-economic parameters such as the percentage of residents living below poverty, the percentage of residents without health insurance, the relative social vulnerability of the county, the median income of households in the county and the population of the county. This finding is counter intuitive to what was hypothesised, that white privilege would associate with a lower death rate, however due to the presence of limitations, this result might not represent the true effect of white privilege on the cancer mortality rates.

Firstly, the dataset that was being used was a cross-sectional observational dataset. Thus, realistically speaking the problem of uncontrolled heterogeneity will never go away, thus the association found by the study, accounting for its statistical significance, cannot be interpreted as a causal relationship. That being said, the model could benefit from the inclusion of important variables for which data was not publicly available at the time this analysis was being conducted. For example, oncologist per capita, the presence of specialty hospitals in the county, hospital capacity can be added to the model to increase its explanatory power. Another limitation that this study faced was due to the data suppression done by NCI for the incidence and mortality data (for details see section 2.3). As counties for whom data is suppressed are systemically excluded from the analysis, the findings of this study cannot to be extended to these counties due to the coverage bias in the data. Coverage bias also exists for data collected by the ACS 2020, and the same limitation extends to the findings for the variables retrieved from data collected by the survey. In the future, the Census Bureau might correct for this bias, so future researchers should investigate this if they plan on getting their data from this source. Lastly, even though the study employed the use of a robust regression framework, some model violations are still unaccounted for. Transforming skewed variables such as population and median income might help increase the explanatory power of the model.

Additionally, researchers in the future could opt for a tangential model to the one used in this analysis. This analysis used data for all types of invasive cancers, however past research has found a link between the incidence of site-specific cancers and race. For example, a higher incidence of testicular cancer was found for Black American men when compared to other races. People of colour are also subjected to generational poverty and poor standards of living thus making them susceptible to cancers such as lung cancer (bad working conditions and smoking trends). Therefore, if incidence and mortality data for such site-specific cancers are used, the positive effect of white privilege on decreasing cancer mortality rates might show. Accessing data through the SEER program directly could also benefit future analysis, however approval is needed before data is shared, and the laws surrounding confidentiality of medical records might make reproducibility of the study an issue.

Thus, this analysis failed to prove that 'whiter' counties see lower mortality rates due to white privilege, however this study successfully proved the influence that socio-economic parameters of a patients surroundings have on their likelihood of mortality from cancer.

Appendix

A Figures

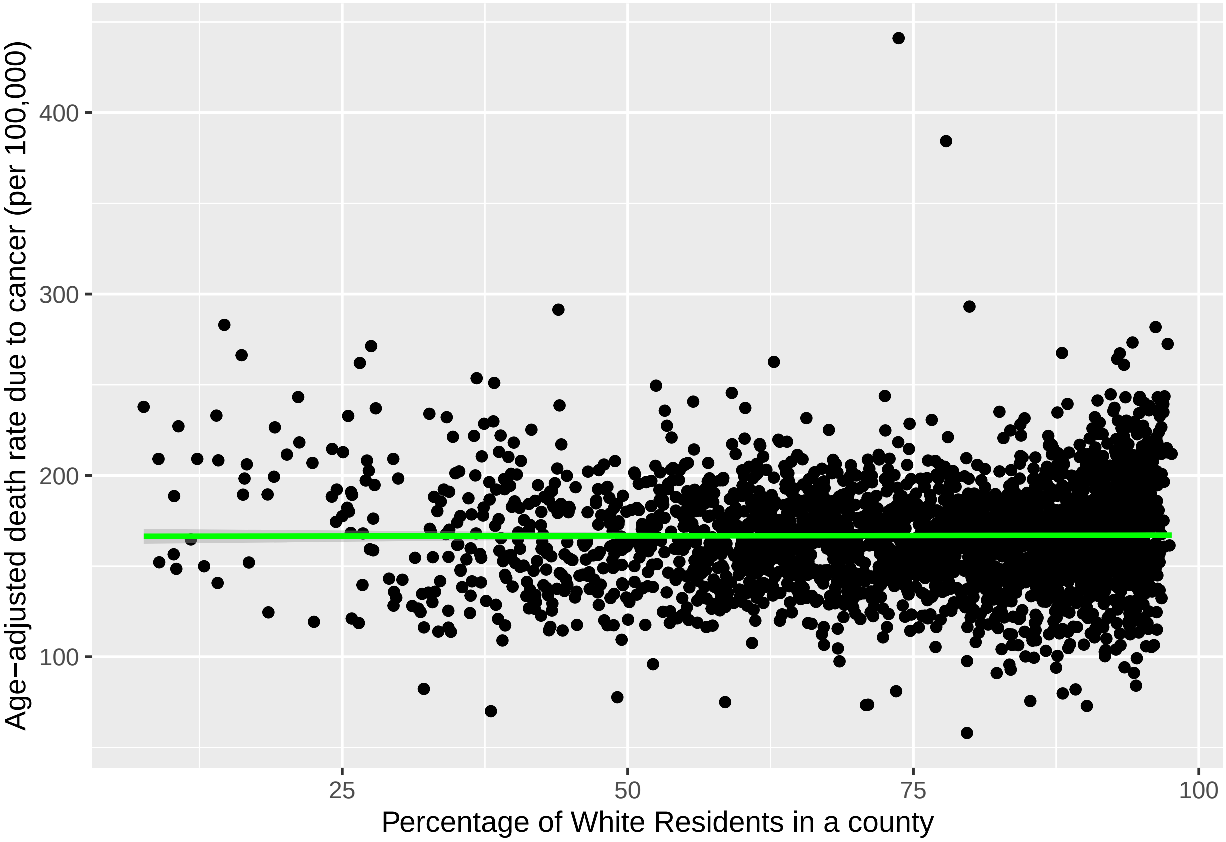

Figure 1: A Scatterplot showing the Relationship between the death rate due to cancer and the concentration of white residents in the county

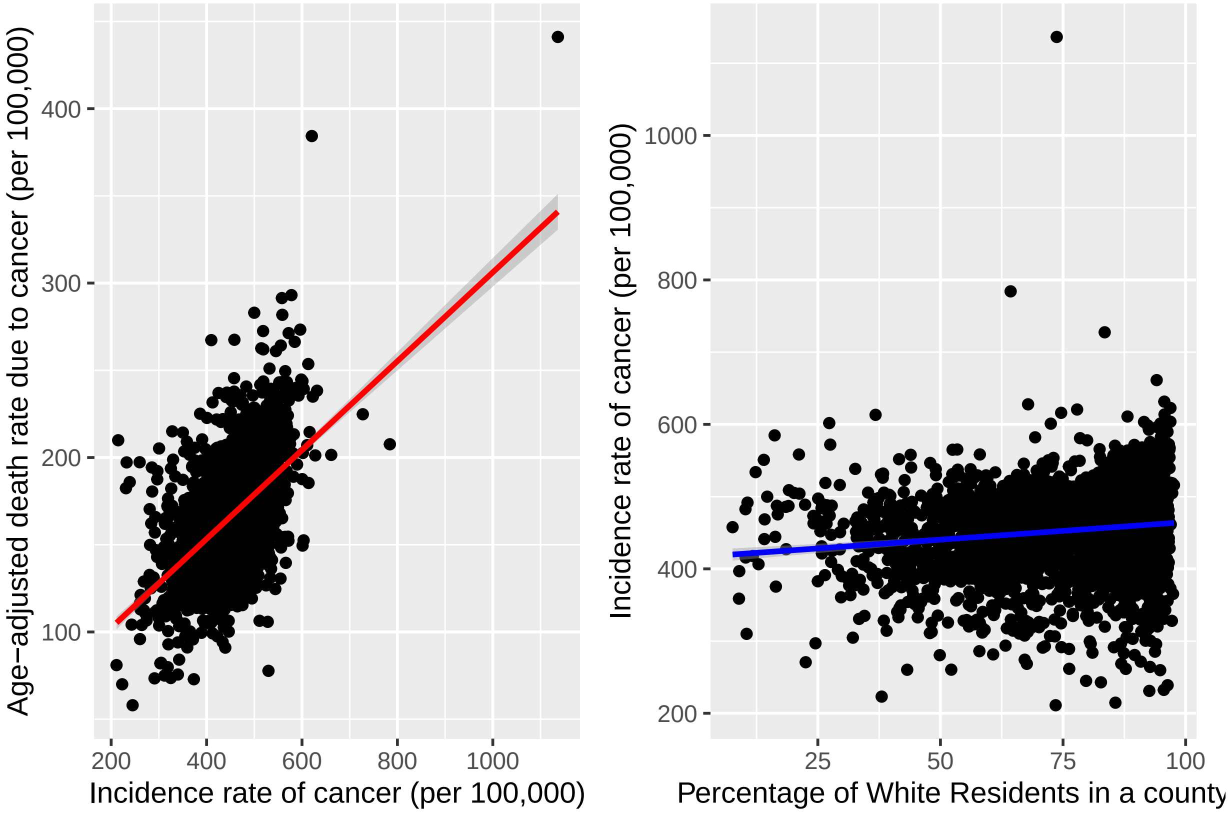

Figure 2: A Scatterplot showing the Relationship between the death rate due to cancer and the incidence rate of cancer

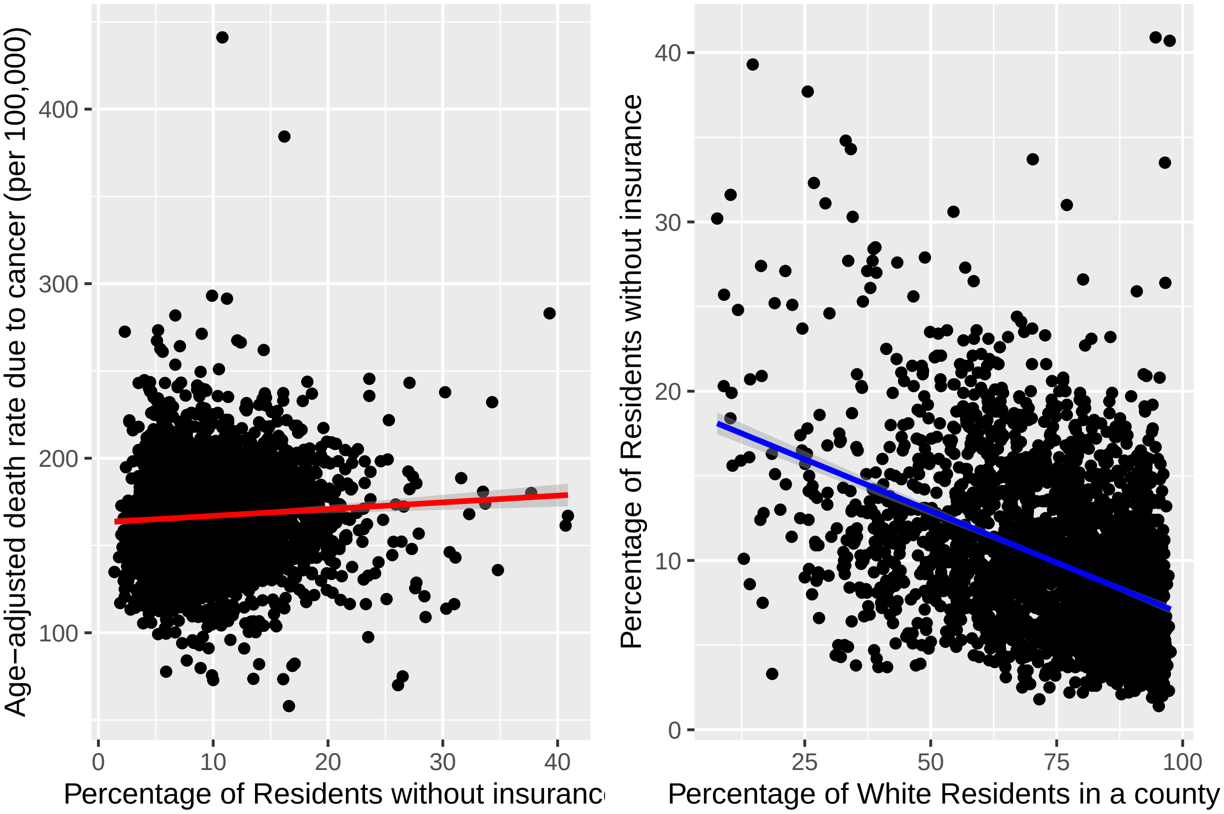

Figure 3: A Scatterplot showing the Relationship between the death rate due to cancer and the percentage of uninsured residents in the county

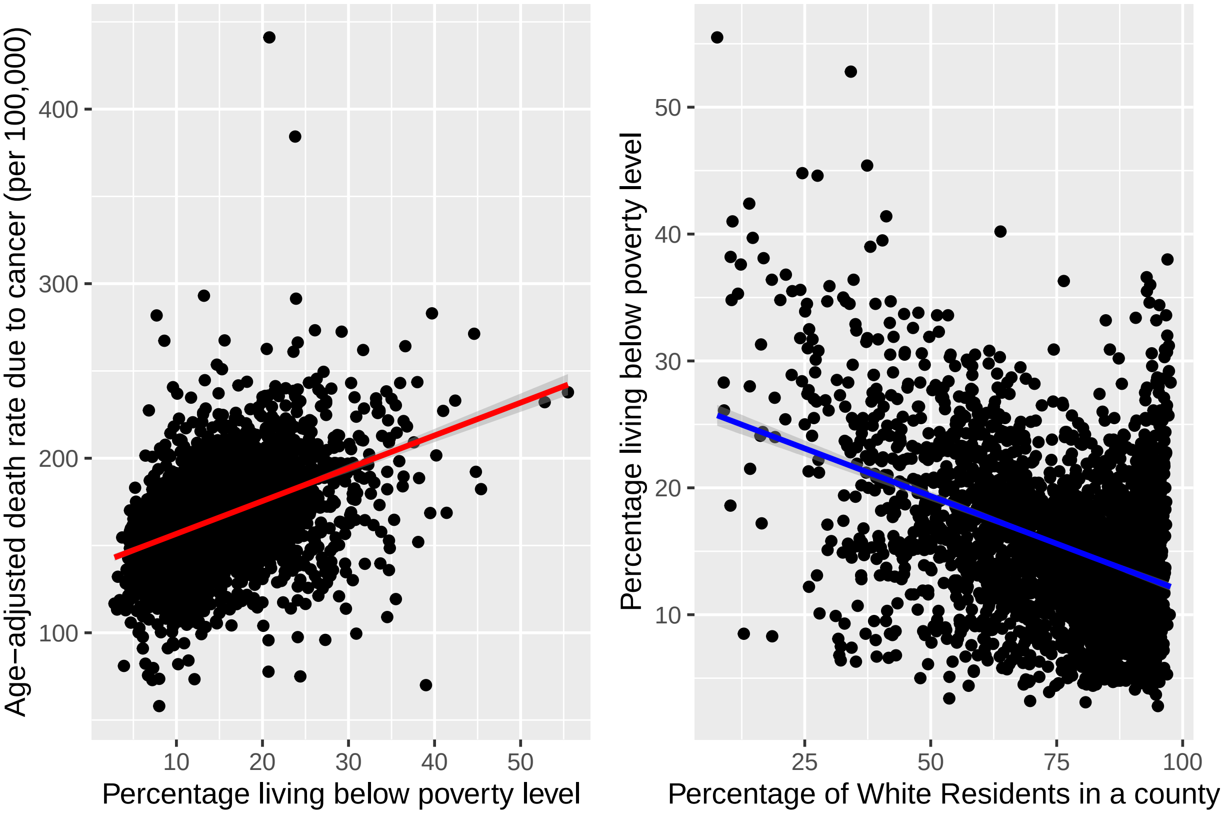

Figure 4: A Scatterplot showing the Relationship between the death rate due to cancer and the percentage of residents living below poverty level in the county

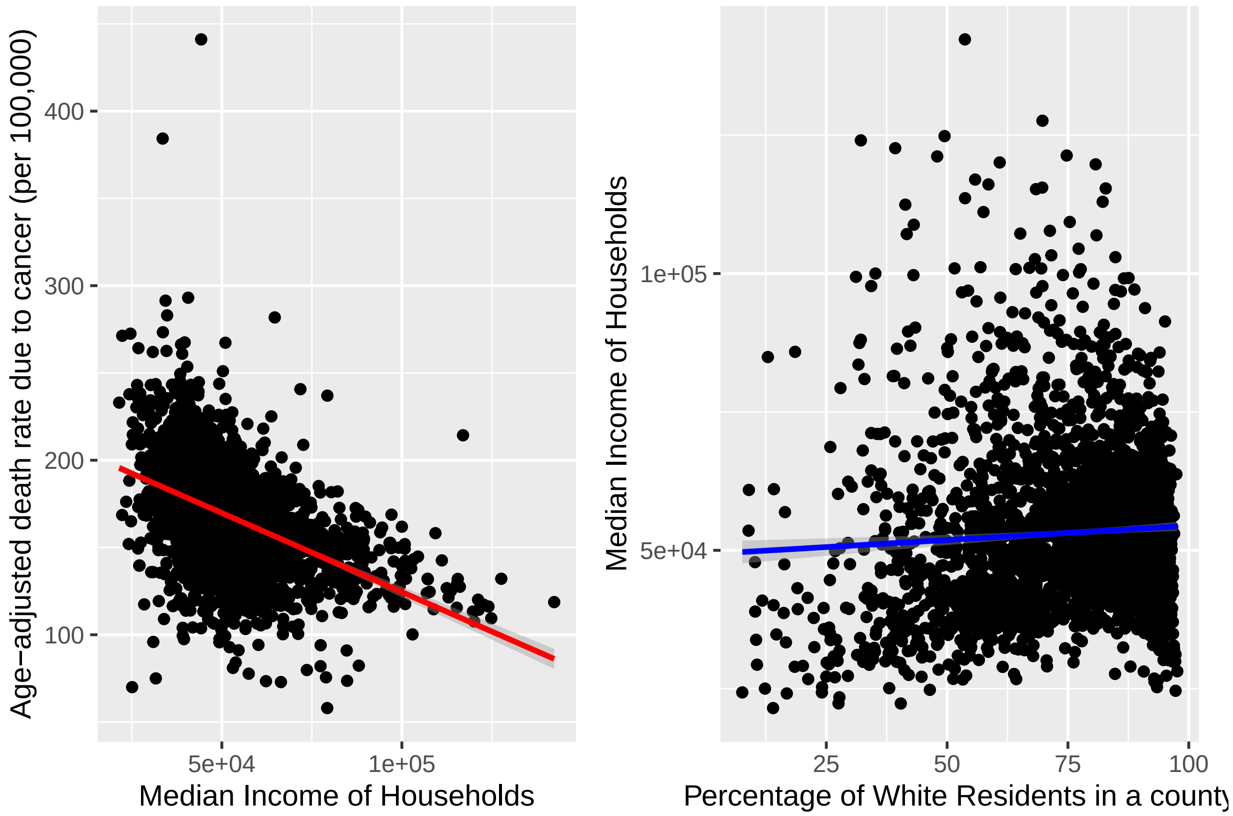

Figure 5: A Scatterplot showing the Relationship between the death rate due to cancer and the median income of households in the county

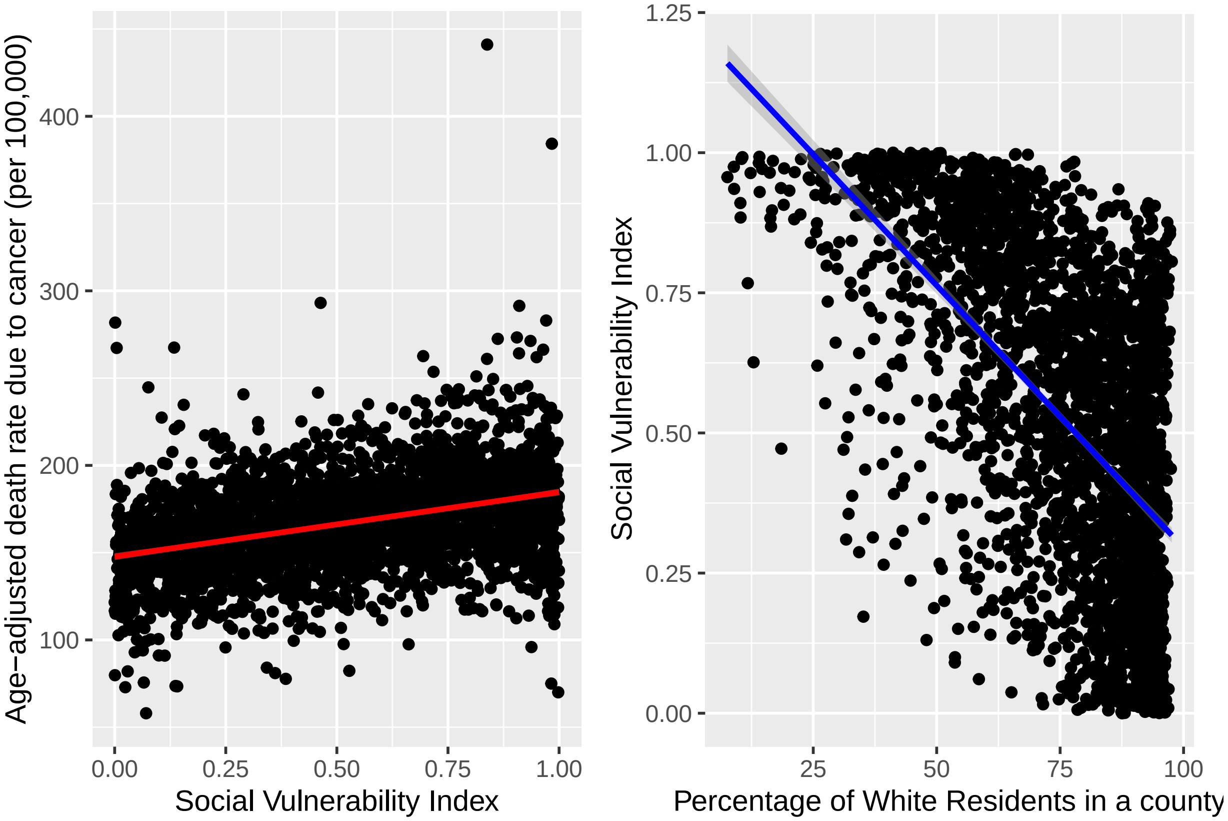

Figure 6: A Scatterplot showing the Relationship between the death rate due to cancer and the Social Vulnerability Index rank of the county

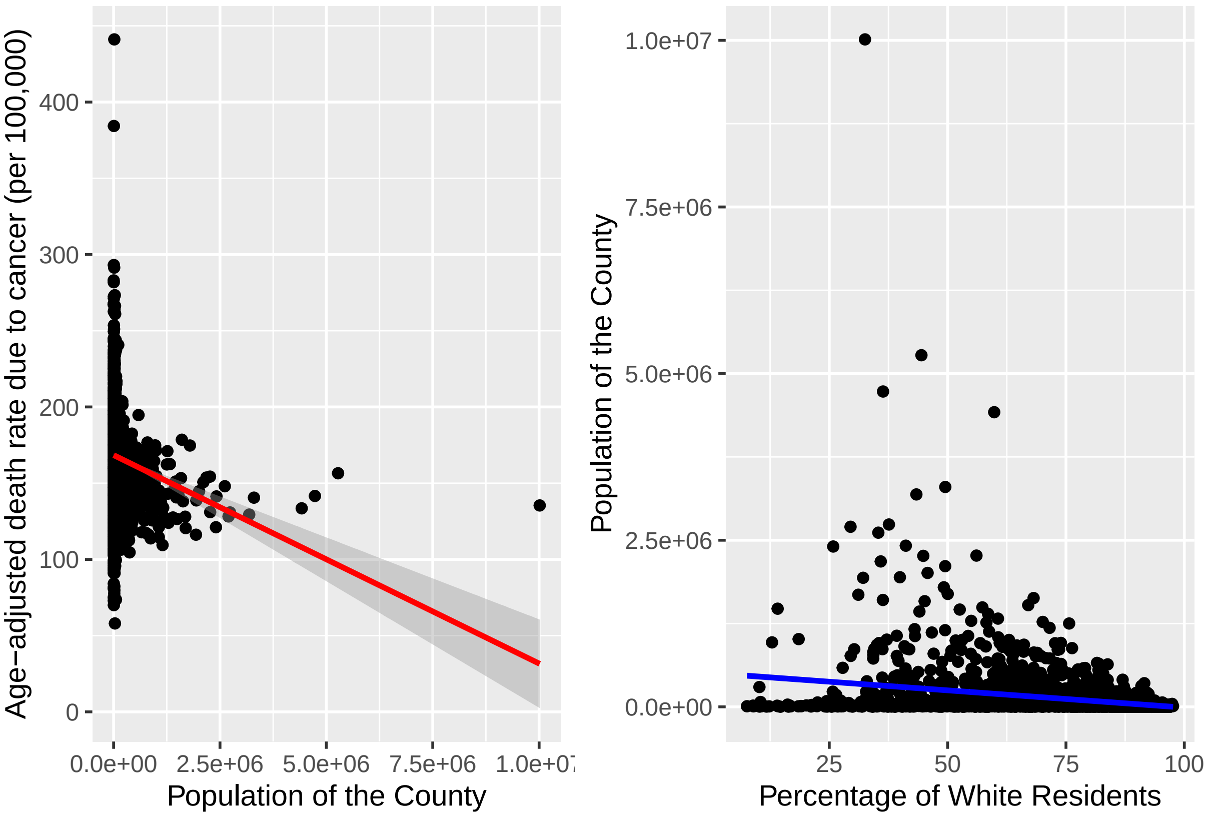

Figure 7: A Scatterplot showing the Relationship between the death rate due to cancer and the population of the county

B Datasheet

B.1 Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- This dataset was created to investigate the effect of the concentration of white residents in a US county on the age-adjusted mortality rate due to cancer per 100,000, while controlling for correlated variables.

2. Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

- This dataset was created by the author by merging publicly available data from US Government sources

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

- The components of this data set were created by the Govt of USA

B.2 Composition

1. What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

- A county in the United States of America

2. How many instances are there in total (of each type, if appropriate)?

- 2885 instances

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

- The dataset is a sample of 2885 counties. Some information were missing for a few counties so they were excluded. The sample is representative

4. What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

- Each instance consists of 9 variables, with 2 identifier variables

5. Is there a label or target associated with each instance? If so, please provide a description.

- Yes, there is an county identifier variable for each instance

6. Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

- No, relationships are not made explicit

7. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

- None, other than possible measurement error

8. Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

- No

9. Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

- No

10. Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

- No

11. Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

- No, only aggregate data for counties are used

B.3 Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

- Data was acquired by the Census survey

2. What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

- Nationwide census survey was used. Data was downloaded from publicly accessible websites

3. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

- ACS 5 year estimates from 2015-2019

4. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

- Webistes

5. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

- Data was collected by census survey

B.4 Processing/Cleaning/Labeling

1. Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- Missing values were deleted. Datasets were merged to create the data

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

- Raw data is available in the GitHub repo: https://github.com/anshumanagarwal27/Cancer-Mortality-and-Race"

3. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

- R was used

B.5 Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

- No

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

- https://github.com/anshumanagarwal27/Cancer-Mortality-and-Race"

3. What (other) tasks could the dataset be used for?

- To investigate the relationship between socio-economic parameters and cancer incidence and mortality

4. Are there tasks for which the dataset should not be used? If so, please provide a description.

- To obtain a causal inference, as data is observational

B.6 Distribution

1. Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

- The dataset is available for public use via GitHub. Individual raw datasets are publicly accessible

2. How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

- GitHub: https://github.com/anshumanagarwal27/Cancer-Mortality-and-Race"

3. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

- No restrictions have been imposed

4. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

- No

B.7 Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

- Anshuman Agarwal

2. How can the owner/curator/manager of the dataset be contacted (for example, email address)?

- Owner can be contacted using GitHub: https://github.com/anshumanagarwal27/Cancer-Mortality-and-Race"

3. Is there an erratum? If so, please provide a link or other access point.

- No erratum available

4. Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

- Currently no plans to update dataset, if in the future it were to be updated the changes would be reflected on GitHub: https://github.com/anshumanagarwal27/Cancer-Mortality-and-Race"

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

- Dataset does not relate to people

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

- Additional variables can be added to the dataset by merging them with the FIPS code, if they are collected on a county level.

C Additional Information

There is a GitHub repo corresponding to this paper: https://github.com/anshumanagarwal27/Cancer-Mortality-and-Race" The author of this paper can be contacted at: anshuman.agarwal@mail.utoronto.ca

References

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Boscoe, Francis P, Christopher J Johnson, Recinda L Sherman, David G Stinchcomb, Ge Lin, and Kevin A Henry. 2014. "The Relationship Between Area Poverty Rate and Site-Specific Cancer Incidence in the United States." *Cancer* 120 (14): 2191–98.

Brown, Denise, David I Conway, Alex D McMahon, Ruth Dundas, and Alastair H Leyland. 2021. "Cancer Mortality 1981–2016 and Contribution of Specific Cancers to Current Socioeconomic Inequalities in All Cancer Mortality: A Population-Based Study." *Cancer Epidemiology* 74: 102010.

Bureau, US Census. 2022. "An Overview of Addressing Nonresponse Bias in the American Community Survey During the COVID-19 Pandemic Using Administrative Data." *The United States Census Bureau*. https://www.census.gov/newsroom/blogs/random-samplings/2021/11/nonresponse-acs-covid-administrative-data.html.

"CDC/ATSDR Svi Fact Sheet." 2021. *Centers for Disease Control and Prevention*. Centers for Disease Control; Prevention. https://www.atsdr.cdc.gov/placeandhealth/svi/fact_sheet/fact_sheet.html.

Chang, Winston. 2022. *Extrafont: Tools for Using Fonts*. https://CRAN.R-project.org/package=extrafont.

Maechler, Martin, Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Eduardo L. T. Conceicao, and Maria Anna di Palma. 2022. *Robustbase: Basic Robust Statistics*. http://robustbase.r-forge.r-project.org/.

Moss, Jennifer L, Casey N Pinto, Shobha Srinivasan, Kathleen A Cronin, and Robert T Croyle. 2020. "Persistent Poverty and Cancer Mortality Rates: An Analysis of County-Level Poverty Designations." *Cancer Epidemiology and Prevention Biomarkers* 29 (10): 1949–54.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Palencia, Laia, Josep Ferrando, Marc Mari-Dell'Olmo, Merce Gotsens, Joana Morrison, Dagmar Dzurova, Michala Lustigova, et al. 2020. "Socio-Economic Inequalities on Cancer Mortality in Nine European Areas: The Effect of the Last Economic Recession." *Cancer Epidemiology* 69: 101827.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Teng, Andrea M, June Atkinson, George Disney, Nick Wilson, and Tony Blakely. 2017. "Changing Socioeconomic Inequalities in Cancer Incidence and Mortality: Cohort Study with 54 Million Person-Years Follow-up 1981–2011." *International Journal of Cancer* 140 (6): 1306–16.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2022a. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://github.com/rstudio/bookdown.

———. 2022b. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.