# United States Cancer Incidence and Death Rates Explored: The Effects of Median State Age, Population Size, Political Alignment, and Recent Trends

Patrick Hinton, Jeff Chen, Megan Shin

Dalhousie University – STAT 3340: Final Project
Fall 2020

## ABSTRACT

The purpose of this study is to examine the prevalence of diagnoses and deaths from invasive cancer in the United States in relation to factors such as county population, median state age, recent trend coefficients, and statewide political alignment. The statistical methods used to analyze this data are visualization of data distributions, simple linear regression, multiple linear regression, stepwise selection, and outlier analysis. Here, we find that incidence and death rates are positively correlated with age, statewide political alignment, and five-year trend, whilst being negatively correlated with population size. We also find that Kentucky has the worst cancer rates while Utah has the best. Lastly, we find efficient predictive models based on the above parameters for incidence and death rates. This study should further enable researchers to understand the determinants of invasive cancer prognosis and death across the United States.

# 1. INTRODUCTION

Cancer is the second highest leading cause of death in the United States and affects millions of new individuals globally each year. Determining potential factors that influence cancer prevalence and death rates helps direct cancer research and funding. Factors with large effects can be used, for example, to make targeted investments in hospital infrastructure and facilitate proper distribution of academic funding. Recently, the COVID-19 pandemic has created large novel problems related to the delivery and access of healthcare, further understanding the determinants of cancer may lead to more efficient distribution of resources and personnel when, globally, we need it the most.

Here, we look to explore several relationships between cancer incidence/death rates (in the United States) and a number of spatially related variables. First, using data visualization techniques, we will investigate which of the 50 states has the highest incidence and death rates of cancer, respectively, as well as how these rates compare across republican and democratic states. Next, we will demonstrate potential individual relationships between incidence/death rates and, each of, state median age, population by county, 5-year trend coefficients (incidence and death), and statewide political alignment. Furthermore, we will look at how all these predictors can affect incidence/death rates in a single model. Following, we will determine the best predictor(s) of incidence and death rates, and create a final predictive model based on these. Lastly, we will briefly explore how outliers may affect state rates with a case-study.

# 2. METHODS

## 2.1 Dataset Description

The combined dataset analysed in this report includes data on age adjusted incidence and death rates of invasive cancers for each county in the United States. Information was downloaded from data.world, which included data from statecancerprofiles.cancer.gov. Data was collected by CDC's National Program of Cancer Registries Cancer Surveillance System (NPCR-CSS) and SEER data. The incidence of cancer dataset contained variables; County name, FIPS identification number, age adjusted incidence rate with 95% CI, average annual counts and five-year trend data with 95% CI. The death rate of cancer dataset contained many of the same variables, but exchanged age adjusted incidence rate for age adjusted death rate, as well as replaced average annual count with average deaths per year. Added variables include median age by state and county population, sourced from World Population Review and US Government Census Bureau, respectively (1). The compiled dataset had dimensions of 15 variables on 3140 observations.

*Introduction of additional data point: We introduced “Patmeff county” into the dataset. The variable associated values for Patmeff are as follows:

i) Incidence Rate

| County | State | Political Part | FIPS | MedianAgeb | CountyPopul | Age-Adjuste | Lower95%Co | Upper95%Co | AverageAnnu | RecentTrend | Recent5-Yea | Lower95%Co | Upper95%Co |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PatmeffCour | Alabama | R | NA | 38.9 | 1500 | 80.9 | NA | NA | 24 | stable | 1.5 | NA | NA |

ii) Death Rate

| County | State | Political Part | FIPS | Median Age | County Popul | Met Objectiv | Age Adjustec | Lower 95 Co | Upper 95 Co | Average Dea | Recent Trenc | 5 Year Trend | Lower 95 Co | Upper 95 Co |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Patmeff Cou | Alabama | R | NA | 38.9 | 1500 | NA | 496.7 | NA | NA | 350 | rising | 52.7 | NA | NA |

## 2.2 Plots

To explore how cancer incidence and death rates varied from state-to-state, counties were first subset according to state, and then boxplots were created to visualize the overall distribution of these rates (incidence, death) using the ggplot2 data visualization package in R. Moreover, states were then further subset according to their state-wide political alignment (2020 US election), and boxplots were created to show how incidence and death rates vary according to democratic and republican states.

## 2.3 Simple Linear Regression

To look at the relationship of predictors on age adjusted incidence and death rates, simple linear regression models were created. Incidence rates were regressed against state median age, population by county, 5-year trend coefficients, and statewide political alignment. These regressions were made with the lm() command in R, specifying the function to omit NA values when encountered. Political alignment corresponded to a two-level categorical factor, D for Democratic and R for Republican. Results use Republican as the determining coefficient. Eight simple linear regressions were performed in total, four for incidence and deaths, respectively.

## 2.4 Multiple Linear Regression

Incidence of cancer diagnoses was looked at in relation to available variable data. The multiple linear regression used age adjusted incidence rate regressed on median age by state, population by county, political party alignment, and five-year trend coefficients in one test. Cancer related death rates were regressed with the same set of predictor variables. Since some values in the dataset were NA, the function specified to omit them from the regression. Political alignment predictors were used the same as in the simple linear regressions.

## 2.5 Stepwise Selection to Select Best Predictor Variables

To ensure only the most significant predictor variables are used in the model, a stepwise selection approach was employed. To do this, age adjusted incidence rate was first regressed on median age by state, as this variable was the most significant from the simple linear regression tests. The next variable to be added was political party alignment, as the second lowest p-value. The remaining two variables were then added one at a time to the model, and the variable that produced a lower p-value, population by county, was included. The last predictor variable, five-year trend was added, and all coefficient p-values were checked for significance. Four steps were taken in the selection process. The same procedure was repeated on the cancer related death rate regression model.

## 2.6 Outlier Analysis

To explore how outliers can heavily influence state response, we subsetted the data to include only the counties within Alabama and performed outlier analysis via both Grubbs's test and Rosener's test using the 'outliers' and 'ENVstats' packages in R, respectively. Grubbs's test functions to identify outliers by testing the null hypothesis that the highest value is not an outlier versus the alternative hypothesis that it is. Alternatively, Rosner's test functions similar to Grubbs's, yet it looks for multiple outliers given a pre-specified number of suspected outliers. Alabama was chosen as the state of interest as we suspected a/several outlier(s) to be present as a result of adding in Patmeff County to the dataset.

# 3. RESULTS

## 3.1 Plots and Visualizations

Boxplots of cancer incidence and death rates across all 50 states indicate that the state of Kentucky has both the highest median incidence rate (Figure 1.) and death rate (Figure 2.) amongst all states. Conversely, the state of Utah showed both the lowest mean incidence rate (Figure 1.) and death rate (Figure 2) amongst all states. Next, boxplots displaying incidence and death rates across republican and democratic states, individually, demonstrate that democratic states have a lower median incidence (Figure 3.) and death rate (Figure 4.).

## 3.2 Simple Linear Regression

Age adjusted incident rate was regressed on median age by state, population by county, political alignment, and five-year trend coefficients. Incidence rates were regressed on four predictor variables, producing significant results for each model (Table 1). Median age by state was significant at p = 2.2e-16, and an adjusted $R^2$ value of 0.0496. The coefficient estimate was 2.01 (±SE =0.1682), proposing a significant positive relationship of age on cancer incidence rates. Statewide political election results were also a significant predictor of cancer incidence rates, at p = 2.2e-16, and an adjusted $R^2$ of 0.0309. Since the data were categorical, this regression used factor level "R", Republican, as a predictor. The coefficient estimate was 6.5 (±SE = 0.6975), indicating a positive relationship. This interpretation shows that cancer incidence rates are higher in Republican states. County population was regressed on incidence rates producing a p-value of 0.002, and an adjusted $R^2$ value of 0.0046. The coefficient estimate was -3.58e-6 (±SE = 9.72e-7), proposing a negative relationship where, as county population increases cancer incidence decreases. Lastly, recent trend coefficients from the previous five years were regressed against cancer incidence rates, and produced a significant relationship, with p = 0.0058, and an adjusted $R^2$ of 0.0025. The coefficient estimate was 0.12 (±SE = 0.0435), indicating a slight positive relationship where, as trend coefficient increases, current incidence rate also increases.

Age adjusted death rate was regressed on median age by state, population by county, statewide political alignment, and five-year trend coefficients. Death rates were regressed on four predictor variables, producing significant results for each model (Table 2). Median age by

state was significant at p = 8.05e-15, and an adjusted $R^2$ value of 0.0209. The coefficient estimate was 1.2378 (±SE = 0.1585), a positive relationship of age on death rate. As age increases, cancer related death rates increase by a factor of 1.2378. Population by county was regressed on death rate and found to be significant (p = 3.49e-5), with a coefficient estimate of -3.65e-6 (±SE = 8.8e-7), and an adjusted $R^2$ of 0.0057. This relationship indicates that as population increases, cancer related death rates decrease slightly. Five-year trend coefficient data were regressed on death rate, producing a coefficient estimate of 2.7109 (±SE = 0.1094) (p = 2.2e-16, adjusted $R^2$ = 0.1855). Lastly, the relationship of statewide political party results was regressed on death rate and found to be significantly positively correlated with Republican as the predictor (p = 2.2e-16, adjusted $R^2$ = 0.0522). If the state voted Republican in the most recent election, the death rate correlation increased by a factor of 7.6956 (±SE = 0.6161).

## 3.3 Multiple Linear Regression

To look at the combined relationship of predictor variables, a multiple linear regression was performed on both cancer incidence and cancer related death rates. A full model was used, regressing median age by state, population by county, political results by state, and recent five-year trends on age adjusted cancer rates. The final model produced coefficients as seen in equation 1. Overall relationship and correlations were consistent from simple linear regression results.

```
Incidence rate = -1.01e+1 + 2.2*Median Age by State – 2.48e-6*County
Population + 7.91*Republican State + 8.69e-2*Recent Five-Year Trend
```

Equation 1. Full multiple linear regression of cancer incidence rates on all predictor variables.

The same procedure was done with cancer related death rates, using all predictor variables in the initial regression. The final model produced coefficients as seen in Equation 2. Overall relationship and correlations were consistent from simple linear regression results.

```
Death rate = 2.61 + 1.49*Median Age by State – 1.23e-6*County
Population + 7.43*Republican State + 2.56*Recent Five-Year Trend
```

Equation 2. Full multiple linear regression of cancer related death rates on all predictor variables.

## 3.4 Stepwise Selection

After using stepwise selection to optimize multiple regression models, the most efficient predictors were kept in the reduced models. Incidence rates were best predicted by all four predictor variables, with significant p-values throughout the stepwise process. The reduced cancer related death rate model was reduced, deeming county population an insignificant variable. The final reduced model is seen in Equation 3.

$$\text{Death rate} = -4.9398 + 1.4846*\text{Median Age by State} + 7.5369*\text{Republican State} + 2.5706*\text{Recent Five-Year Trend}$$

Equation 3. Reduced multiple linear regression of cancer related death rates on significant predictor variables.

### 3.5 Outlier Analysis

Outlier analysis, via Grubbs's and Rosner's tests, performed within the subsetted Alabama dataset show that Patmeff county was identified as an outlier of death rate in both tests conducted (Grubbs; $p < 2.2e\text{-}16$). When including multiple outliers within Rosner's test (eg. k=2), only Patmeff county is identified as a significant outlier. We therefore removed Patmeff county in subsequent analyses in which the data were skewed by this outlier.

## 4. CONCLUSION

The results of our analyses provide us with a clearer depiction of cancer distribution and death within the present United States. We found that; (i) Kentucky has the highest rates of cancer incidence and death whilst Utah is the lowest in these categories, (ii) incidence rates are significantly positively correlated with state median age, republican states, and five-year trend, and negatively correlated with county population, (iii) death rates are significantly positively correlated with state median age, republican states, and five-year trend, and negatively correlated with county population. This suggests that older people and people in republican states have a higher probability of having cancer and dying from it. Additionally, this suggests that counties with larger populations are have less cancer and cancer-related death, normalized for the relative populations, and that the five-year trend is significantly predictive of the current death/incidence rate.

Results of these analyses may help inform future researchers about the potential social and physical determinants of cancer prognosis and how these are distributed across the United States.

## Figures and Tables

Table 1. Simple linear regression results for incidence rates. Tests were performed separately, results are compiled.

| Coefficient | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -6.4826 | 6.3936 | -1.014 | 0.311 |
| Median Age by State | 2.0106 | 0.1682 | 11.953 | <2e-16 |
| (Intercept) | 7.02E+01 | 3.59E-01 | 195.82 | < 2e-16 |
| County Population | -3.58E-06 | 9.72E-07 | -3.68 | 0.000237 |
| (Intercept) | 70.25081 | 0.34743 | 202.201 | < 2e-16 |
| Five-Year Trend | 0.12002 | 0.04345 | 2.762 | 0.00578 |
| (Intercept) | 65.7193 | 0.5533 | 118.77 | <2e-16 |
| Republican State | 6.538 | 0.6975 | 9.373 | <2e-16 |

Table 2. Simple linear regression results for death rates. Tests were performed separately, results are compiled.

| Coefficients | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 6.3567 | 6.0247 | 1.055 | 0.291 |
| Median Age by State | 1.2378 | 0.1585 | 7.809 | 8.05E-15 |
| (Intercept) | 5.38E+01 | 3.22E-01 | 166.729 | <2e-16 |
| County Population | -3.65E-06 | 8.80E-07 | -4.146 | 3.49E-05 |
| (Intercept) | 56.247 | 0.3027 | 185.8 | <2e-16 |
| Five-Year Trend | 2.7109 | 0.1094 | 24.78 | <2e-16 |
| (Intercept) | 48.6153 | 0.483 | 100.65 | <2e-16 |
| Republican State | 7.6956 | 0.6161 | 12.49 | <2e-16 |

Figure 1. Boxplot of the age adjusted incidence rate across the 50 states

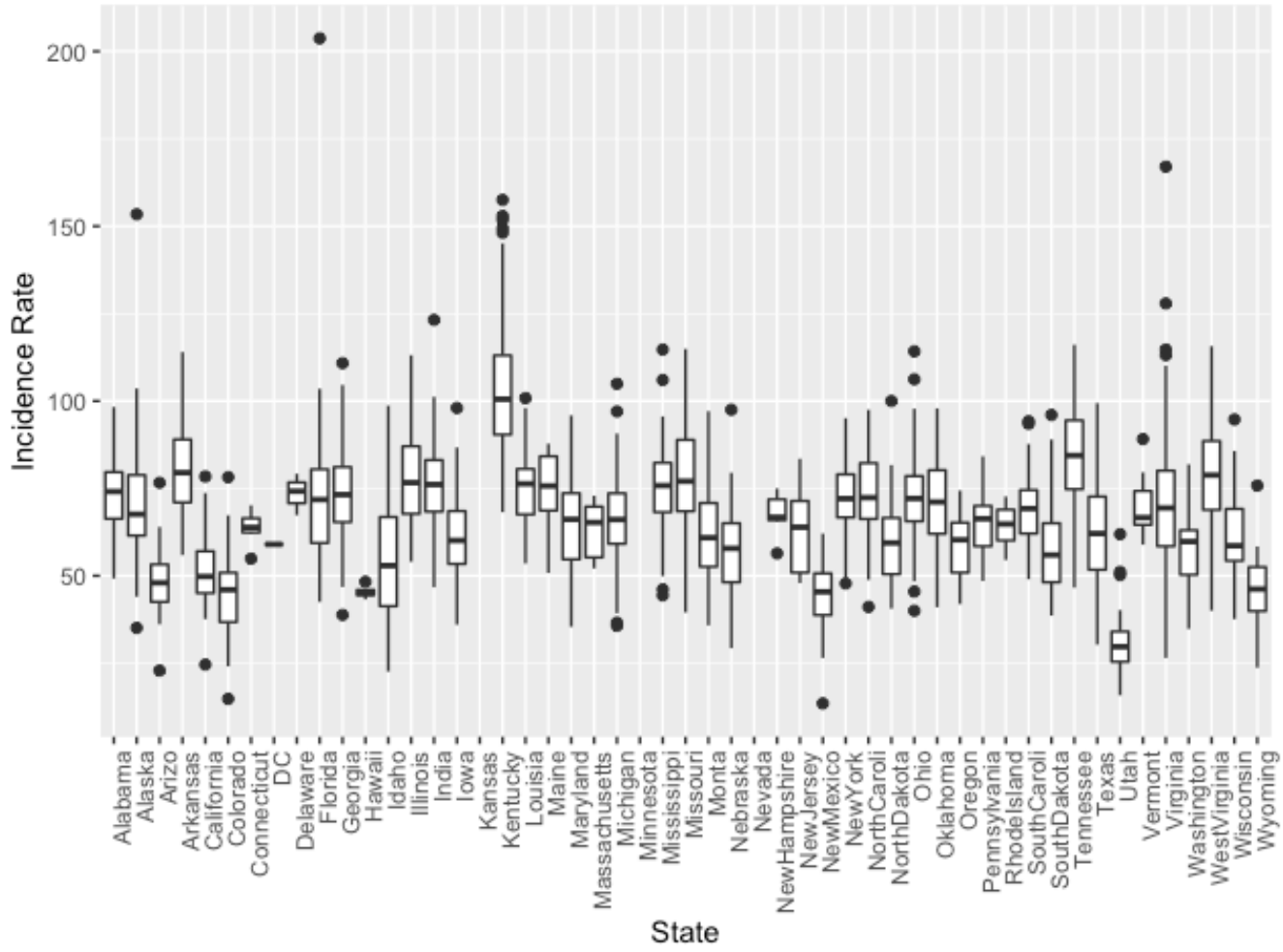

Figure 2. Boxplot of the age adjusted death rate across the 50 states

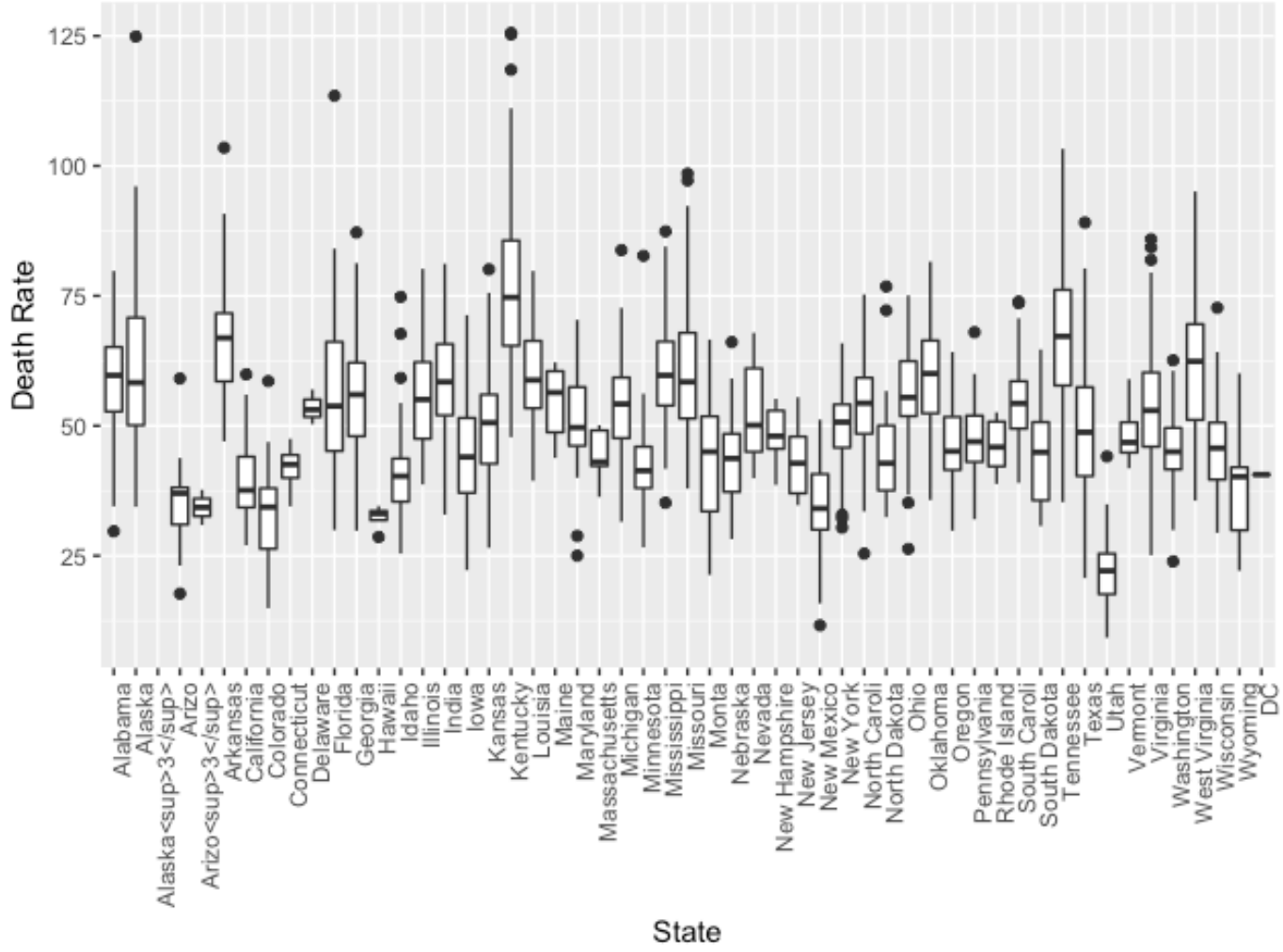

Figure 3. Boxplot of the age-adjusted incidence rate of cancer across democratic and republican states.

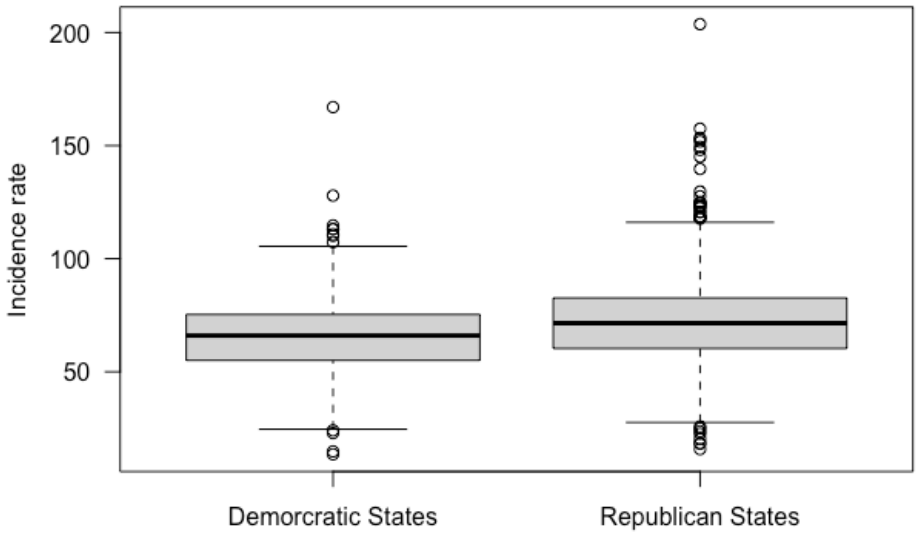


Figure 4. Boxplot of the age-adjusted death rate of cancer across democratic and republican states.