Cancer Death Rate Comparisons in the United States

Anna Hubbard
Computer Science Department
Santa Clara University
Santa Clara, United States
ahubbard@scu.edu

Andrew Schulz
Computer Science Department
Santa Clara University
Santa Clara, United States
anschulz@scu.edu

Camden Rothleder
Computer Science Department
Santa Clara University
Santa Clara, United States
crothleder@scu.edu

***Abstract*—Our project focuses on analyzing data on the cancer death rate in the United States using machine learning. We focus on four main variables —age, gender, cancer type, and race. We express how our predictions can be useful in the medical field. Using linear regression and data visualization, we compare each of these features to total cancer death rates for all of the U.S. states. We go into detail about the design and implementation of each of our learning models. The results we find conclude in correlation, predictions, and observations surrounding our data. We aim for our interpretation of these outcomes to be beneficial in detecting cancer at an earlier stage to increase survival rates.**

***Index Terms*—Machine Learning, Supervised Learning, Prediction Model, Linear Regression, Data Visualization, Feature Variables, Target Variable**

I. Introduction

More than 1.8 million people in the United States are diagnosed with cancer each year and 0.6 million of those are terminal (Abdullah). However, the distribution of who is diagnosed is still under research. While cancer affects people of all ages, races, and genders, we still do not have a clear indication of any correlations. As cancer research continues, it is valuable to understand any prevalence or relationships in who is getting cancer as it may help answer questions of why cancer may occur in an individual and thus, aid in deriving better treatment plans. Research in this field is aimed not only to identify cancers but also to predict them in individuals who may be diagnosed in the future.

We want to compare cancer death rates against multiple variables including types of cancer, age, race, state, and gender. In the data set, we are given a lot of different categories, sorted by race, gender, age, and types of cancer, and all of the combinations between these variables. This data is also given state by state. It is accompanied by the total out of the population as well as the total based on each different variable. The data set shows the number of total deaths, not separated by these variables so we will be able to use this data and compare it to other variables.

Through this project, we aim to identify common trends found in comparing these different variables to cancer death rates in each state of America. We want to explore and analyze the data to determine if there is a link of causation between these variables and the total death rates. We would use this to determine if certain individuals are more susceptible to cancer than others. Ideally, we would also be able to use the outcome of this project to help predict if there is a correlation between any of these variables and the probability of surviving cancer. For instance, you may think that location has an effect on the type of cancer and its death rates based on environmental factors. We wish to test this theory and others to see if the hypotheses hold.

II. Background

A. Usefulness of Predictions

With an increase in the number of people visiting healthcare providers comes an improved data set. Since more data is available and being collected, we are able to have a more diverse set of data that can allow for more accurate predictions. In terms of our cancer prediction models, this yields a better understanding of possible variables that may affect one's chances of developing and surviving cancer. Detecting the patterns in the data is how medical professionals can deduce some of the common causes and help create better treatment plans.

Once we figure out the main causes of cancer and the variables that play a role in its development, we will be able to better predict and diagnose cancer in patients. With this information, we will hopefully be able to diagnosis individuals at an earlier stage in the disease and increase their chances at survival. "Early-stage cancers require less complex treatment regimens and reduced hospital utilization, resulting in reduced healthcare costs, whereas late-stage cancers require complex multimodal management, several rounds of extremely expensive drugs over significant periods of time, and the treatment of recurrences, equating to a staggering economic burden"

(Abdullah). This suggests that our results could increase survival rates by helping with early cancer diagnoses.

By looking at certain aspects of the data set we will be able to select certain aspects that stand out, having influence over cancer death rates. By analyzing this data we can compare it to outside sources and determine if our findings are significant. Our goal is to show the data in a way that is easy to digest and interpret rather than a standalone CSV file that is very hard to use for interpreting data. From these visualizations, learning models, and analysis we will be able to predict cancer death rates more efficiently when taking these other aspects into account.

Our data set was derived from the CORGIS Dataset Project. The visualizer uses data from social explorer which looks at medical records and computes statistics from this information to create a data set.

III. Design and Implementation

A. Overview

The data set we worked with was provided through The World Population Review. It covered various cancer death rates across all 50 U.S states. The rows of our data were the states and the columns varied between total population and cancer rates. The cancer death rates provided were divided by age, race, cancer type, and gender. Moreover, rates displayed also included combinations of these various categories —e.g. Types.Breast.Race.Asian. This distinction helped give us a range of variables to explore when analyzing the data in that we could see how two or more factors worked with one another. However, this layout of data also provided its unique challenges. Because the cancer death rates (the columns) of our data set varied between 73 different rates, it would not have been beneficial to find a correlation across the data set as a whole. Instead, we had to divide the set into different data frames for more accurate analysis.

We have decided to use supervised machine learning in the form of Linear Regression as one of our experimental methods. This means that we provide the machine with a data set in which it trains and tests the values in order to produce an algorithm to help make future predictions. Linear regression is an approach that uses the hypothesis function to make predictions for an output with a given input. We chose this approach since we wish to predict the value of a variable based on the value of another variable. This is a form of supervised learning since the training data includes our desired outputs. For this project, our desired output, or target variable, is the total rates of cancer for each state in the United States. We use this as our y variable in our models. A linear regression model is an equation that describes the straight-line relationship between dependent and independent variables. This linear model allows us to make future predictions by plugging in an input value. The accuracy of these models are tested by the loss (or error) function. This function measures the difference between the prediction and the actual output. Minimizing this function allows us to find the most ideal learning model that makes the most accurate predictions.

For the analysis of cancer death rates and the relations to race, we focused on data visualization and interpreting based on different graphs and data. Combining these methods or multiple graphs can be extremely powerful as it not only provides more data with different perspectives but it also provides more of an emphasis to the specified data sets. Connections can be drawn between multiple items creating a better argument for analysis.

The main features we are working with include age, type of cancer, and race. These selections were chosen since they are most related to the target variable we aim to predict —total cancer death rates. We use these feature variables to help predict the total death rate of cancer in each state in the USA.

B. Age - Data Visualization

To start our understanding our data, we needed to perform basic visualization. In this instance, we looked at age being the only factor to explore which age-group had the highest cancer death rate across all 50 U.S states. We believed that a scatter plot would be best used to represent this. As seen in figure 1, across all 50 states, the age group greater than 64 (red) has the highest total death rate across all states. On the x-axis is our independent variable being 50 tic-marks representing each of the 50 states. The y-axis is our dependent variable being the cancer death rate. Because there were technically 3 variables being compared, we had to utilize color to designate different age groups. This plot clearly shows that cancer death rates steadily increase with age.

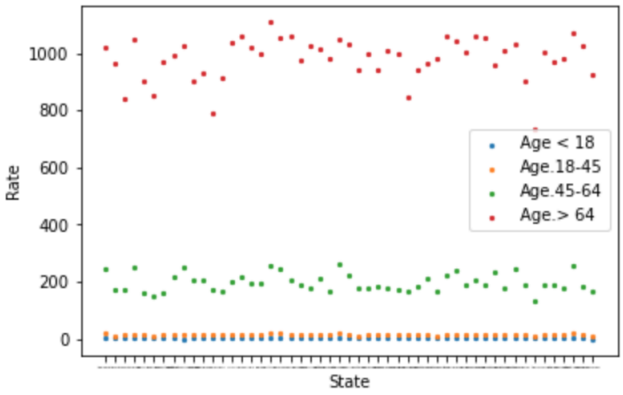

Fig. 1. Death Rates for Different Age Groups Across All 50 States

C. Gender - Data Visualization

We decided to take the age range with highest rates (ages 64) and look at distinctions between male and female death rates. As seen in figure 2, death rates for Men is significantly higher than for women across all total death rates within the age group. Once the total deaths passes about 975, it is clear that there is an increasing difference between male and female death rates. These findings suggest that men may be more at risk than females especially at high age groups.

Moreover, looking at figure 2, we noticed some subtle linearity between the two genders. This linearity simply suggests

that as the total death rates increase, so does the death rates unique to both genders.

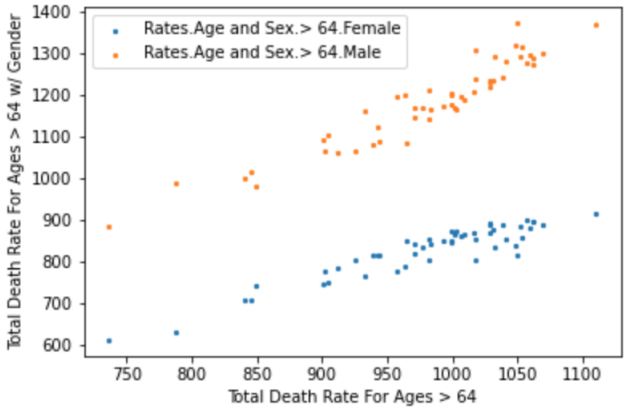

Fig. 2. Total Cancer Death Rates Ages Greater Than 64 vs Male/Female Death Rates Ages Greater Than 64

D. Age Groups vs Total Rate Linear Regression

To test age against total death rates, we decided to use linear regression. We took California out of the data set because we want to test our model using that data. To begin, we separated the data set to only contain the various age groups: <18, 18-45, 45-64, and >64. These features were grouped by state, excluding California. Next, we examined the correlation matrix and identified that there is a high correlation between total rates and the age groups of 18-45, 45-4, and >64. With this information, we graphed three scatter plots to detect which features appear to have a linear relationship to total rates. We found this to be the age group of >64. Using linear regression, we fit a learning model to the data and used that to predict new y values to find a line of best fit, shown in figure 3.

To test this learning model, we plugged in the data for California and the >64 age group. The data set shows California has a Total.Rates value of 150.9 and a Rates.Age.¿64 value of 902.4. Using our prediction model, we can estimate the total rates value using the known Rates.Age.>64 value of 902.4. The model predicts the total rate to be approximately -104.285 + 0.3(902.4) = 166.435. This is similar, but not precise, to the actual total rates value of 150.9, suggesting our algorithm is a mostly successful predictor of total death rates for states in the United States with an input of the cancer death rate for the age group >64.

As you can see, the error value of the linear regression model in figure 3 is around 300. This number resembles the sum of the squared errors between the predicted value and the actual value. Our objective to configure an optimal algorithm is to minimize this loss function. Since the number is relatively high, we can assume that this is not the most accurate predictor. Hence, the reason our actual and predicted values for California were not an exact match.

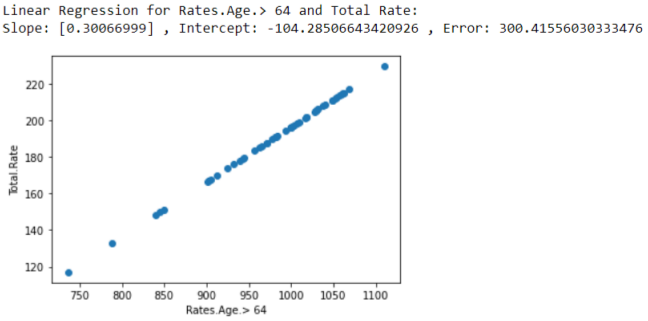

Fig. 3. Age Group Rates vs. Total Rate Prediction Model

E. Types of Cancer vs Total Rate Linear Regression

For our test comparing types of cancer to total rates, we took a similar approach as we did when comparing total rates with various age groups. We still used linear regression and removed California from the data set so we could use it to test our prediction at the end. After separating the data by the three types of cancer in our data set, we ended up with values for lung cancer, colorectal cancer, and breast cancer. Following this, we analyzed the correlation matrix for these features and found that all of them had high correlation values. Looking at the scatter plots we graphed for these features, we found that lung cancer appeared to be the most linear in relation to total rates. With this, we graphed a linear regression model of lung cancer versus total rates.

Using California to test our learning model, we should expect a total rate of around 150.9 with a given lung cancer total of 34.5. With our prediction model, we can estimate the total death rate to be roughly 76.36 + 2.15(34.5) = 150.535. This total is very close to the true value of 150.9. This implies that our algorithm is a successful predictor of total death rates for states in the United States when given the total rate for lung cancer.

The linear regression model in figure 4 shows the error value to be approximately 83. Since this value approximates the sum of the squared errors between the predicted value and the actual value, we can expect our predicated values to be slightly different from the actual values. This is the reason our prediction for California is reasonably close to the actual value.

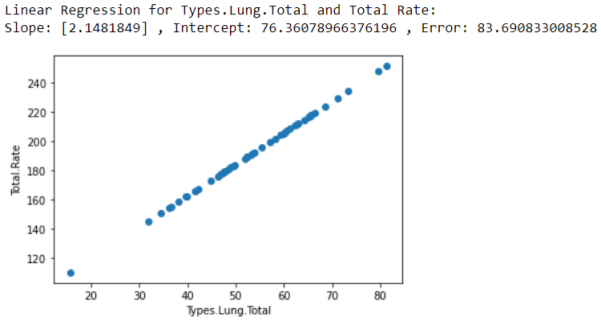

Fig. 4. Cancer Types vs. Total Rate Prediction Model

F. Cancer Death and Race Data Visualization

To take a look into how total cancer rates are compared to individual races, we created a graph storing the data on White, White non-Hispanic, Black, Asian, Indigenous, and Hispanic rates. Using this data we can visualize any separation or data points that do not seem correct. Theoretically, if every state was said to be similar, we might expect the cancer rates of every race to be somewhat similar across the board.

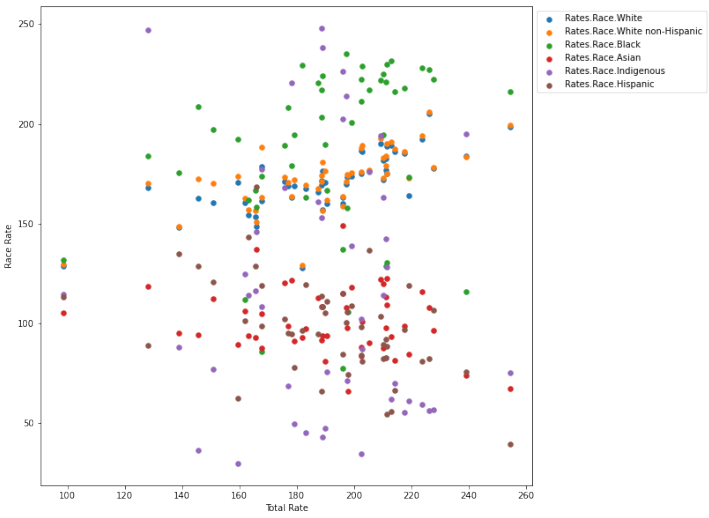

Fig. 5. Total Cancer Rates vs. Individual Race Rates

By looking at this visualization, many generalizations can be made. Interestingly while most races keep fairly steady with White/Black being high and Asian/Hispanic being lower, not dependent on the total cancer rate, Indigenous is very spread out. Before cleaning the data, there were also missing values in the data set resulting in zeros for the indigenous rate.

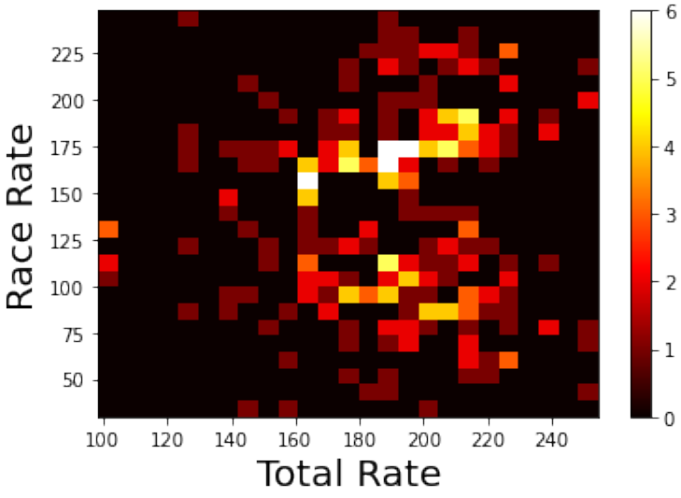

Fig. 6. Heat map

The divide is easier to see when put into a heat map. Here there are two distinct centers showing an emphasis on the divide between races. Comparing this to the previous graph we can see just how the races are divided. The Hispanic and Asian races make up the majority of the lower rate cluster while the upper cluster is mainly white and black races. This can likewise be seen in a more concrete form using the mean values of the rates but this form of visualization does take away the variance that the previous figures supplied.

Race	Cancer Death Rate (per 100,000)
White	170.999
White non-Hispanic	173.190
Black	186.499
Asian	100.972
Indigenous	118.056
Hispanic	98.156

Our interpretations for cancer death rate and race correlation are further backed up when looking at the averages per group. White and Black races have the highest cancer death rates, nearly doubling the rate of Hispanics. With this being steady across all states, there must be some factor that attributes to this other than coincidence. Because correlation does not always mean causation, looking at other articles can determine whether this data set can be used to show that cancer death rates differ between races or not.

IV. Results

A. Age

Our discovery surrounding the relationship between various age groups and total cancer rates shows a high correlation between total rates and the age group >64. This implies a high association between the two variables. Since this is the age group with the highest correlation, then it is also the variable with the most linear relationship to the total death rate. Thus, the reason we decided to use linear regression as the learning model.

B. Types of Cancer

Given the total for a particular type of cancer for a given state in the U.S., we are able to predict the total cancer death rate in that state. The positive correlation between types of cancer and total death rates led to our line of best fit to have a positive slope. Our use of linear regression and identification of total rates being the target variable has allowed us to establish a successful machine learning predictor.

C. Race

Through our findings along with added research, we were able to see the disparity between race and cancer death rates. This disparity comes from many factors, of which environment and genetics play a large role. These findings provide ideas such as diet, pollution exposure, and general lifestyles increase the risk of cancer death. Smoking, drinking and lack of activity all increase cancer death rates, in fact, these lifestyle activities were all prominent in the White and Black races

over others (Weiner, Winn), confirming our previous findings. Furthermore, while white and black cancer rates remain high, death rates among black people remain higher than the white counterpart regardless of gender. There is also a significant factor with cancer death and the general rate being genetics. As people of similar races come from generally a similar lineage which means they share a lot of similar genetic makeup and difference from other races. While this disparity may help with things like endurance or common physical features, it also predisposed certain races to disease and illness. Like a disease ”running in the family” cancer also seems to run in races. This is due to the genetic makeup of different races with some genes being more vulnerable to modification and abnormal growth.

V. Interpretation

The main objective of our cancer prediction models is to assist in detecting if a person is likely to develop cancer in their lifetime. The interpretation of our results is ideal to guide medical professionals in diagnosing cancer in their patients and making decisions for their treatment. Generalizing this information will allow us to apply our results to a wide variety of patients with a high precision rate.

“The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients” (Konstantina). This is why studying cancer in relation to other variables is beneficial to saving lives with early prognosis. Once we know what causes cancer and the aspects of our lives that enhance our chances of getting the disease, we will better be able to treat and prevent others from getting sick. Machine learning plays an important role in this. As soon as we are able to gather enough data to build a model, we then will be able to identify the key features and predict who is likely to be diagnosed. Predictions like this and the ones discovered from our learning models can help us make precise decisions on treatment methods.

References

[1] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, ”Machine learning applications in cancer prognosis and prediction”, Computational and Structural Biotechnology Journal, vol. 13, pp. 8-17, 2015.

[2] Abdullah Alfayez, Asma et al. “Predicting the risk of cancer in adults using supervised machine learning: a scoping review.” BMJ open, vol. 11,9 e047755. 14 Sep. 2021.

[3] Liu, Peter. “How Data Science Enables Early Cancer Diagnosis.” *Medium*, Towards Data Science, 12 Nov. 2019.

[4] Kafura, Denis. “Cancer CSV File.” CORGIS Datasets Project, 27 June 2019.

[5] George J. Weiner, Robert A. Winn, Disparate groups share cancer disparities, Trends in Cancer, 10.1016/j.trecan.2022.01.012, (2022).