MA678 Midterm Project Report

Investigation of brain cancer incidence

Kayla Choi

MSSP Fall 2021

Abstract

Introduction

Cancer is a disease in which cells begin to rapidly proliferate abnormally. These cells can form clumps called tumors, which can progressively become deadly as it grows if located in a vital organ such as the lungs or the brain. Cancer of the brain or spinal cord, which are part of the central nervous system, is called glioblastoma. The median age of glioblastoma diagnosis is 61; however, the most common solid tumors that affect children are glioblastoma tumors.
To investigate this phenomena, I wanted to look at the incidence and mortality rates for different age groups. From the CDC WONDER cancer statistics database, I obtained demographic information and the relevant cancer counts and rates, grouped by year and US state. I hypothesized that mortality rates may be different between the younger and older age groups, and that these rates could show a time-dependent trend. There may also be a difference in rates across different states; higher populated states with a greater biomedical research impact may see a decrease in mortality rates with new cancer research, and the other states may lag behind as the technology gets distributed. I look to investigate all these relationships.

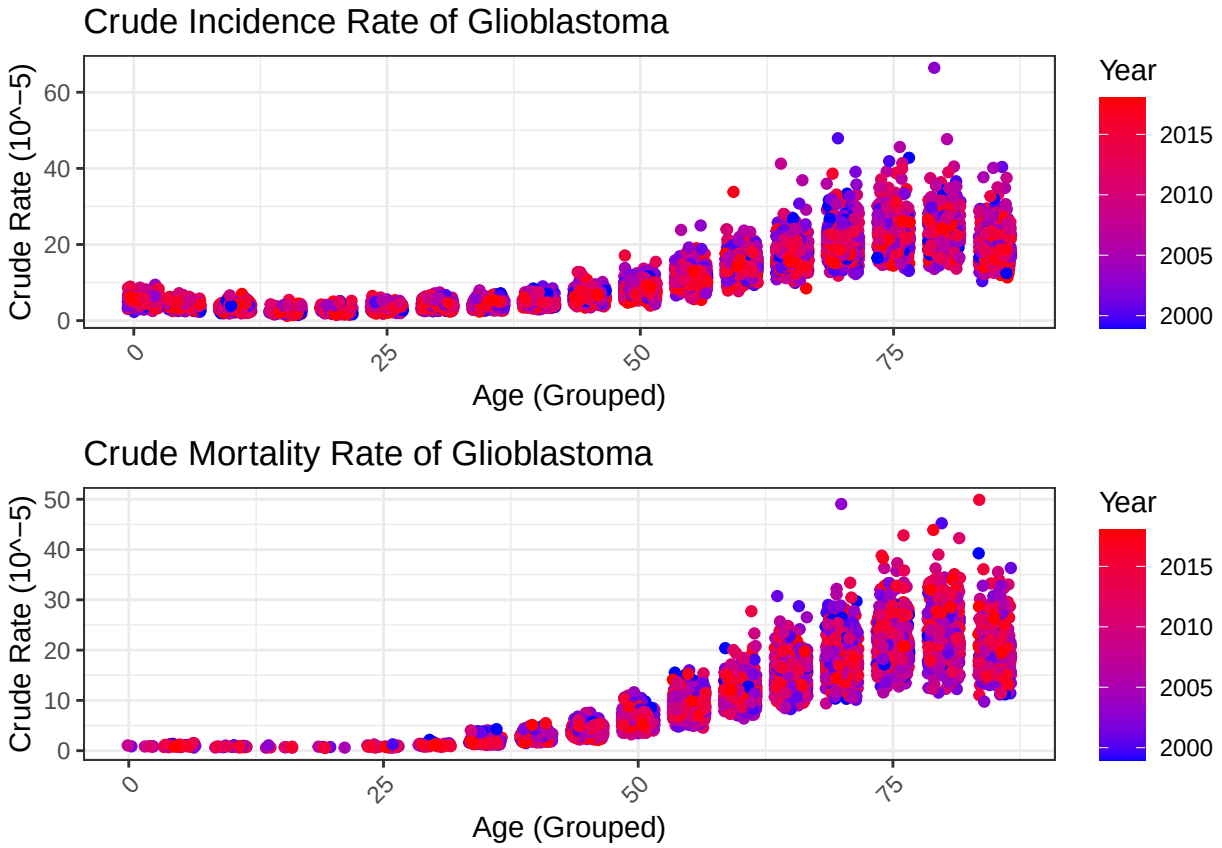
Method

Data Cleaning

The CDC database allows one to request incidence or mortality data grouped by different factors, such as race, age group, year, state, and type of cancer. I manipulated the data frame to add a code to age groups (e.g. turning "1-4 years" into "1"), obtained the crude rate (i.e. count/population, in the scale of 10^-5) and the respective standard error. The database had an option to request mortality rate data; however, it was not grouped by age and there were far fewer observations, so the incidence and death data was obtained and manipulated separately. Furthermore, some states did not have complete counts over the years, so I focused on US regions (Midwest, Northeast, South, and West), but you may see maps in the Appendix.

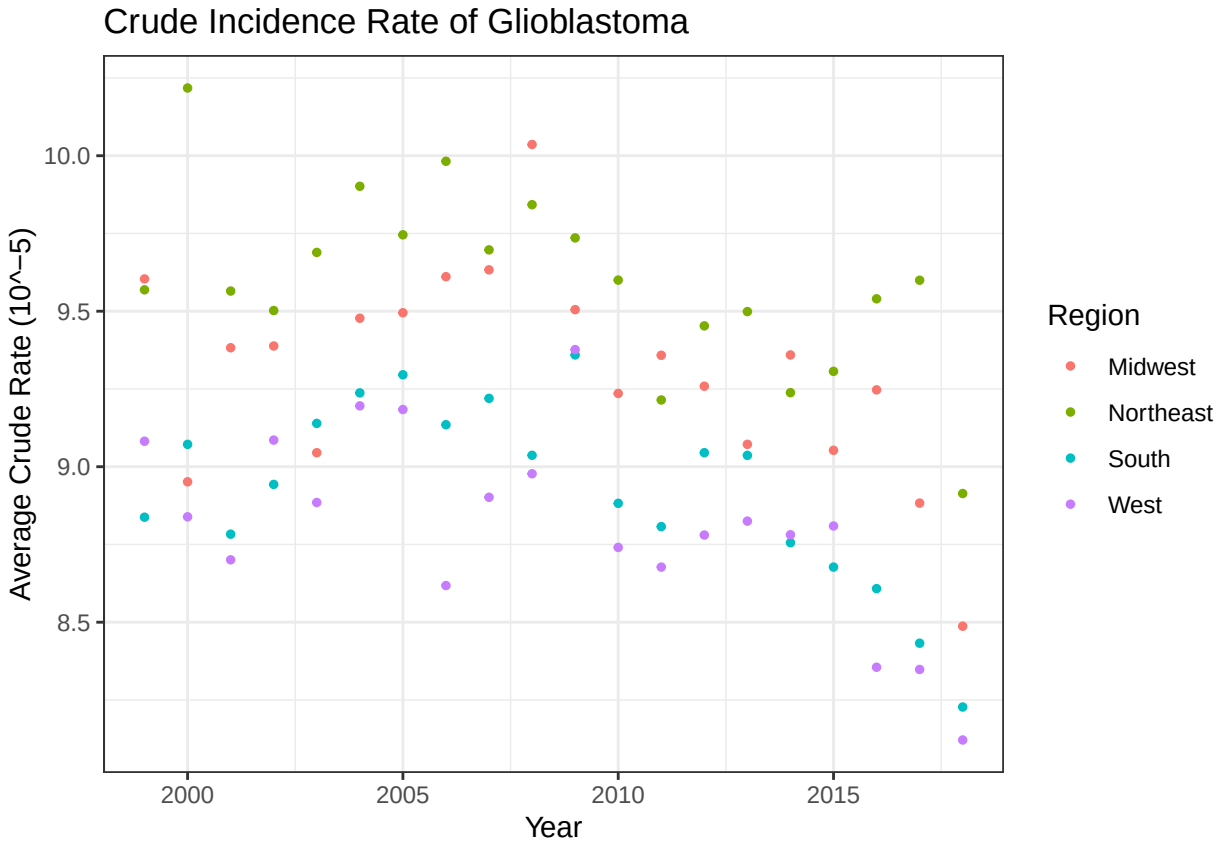
Data Visualization

I performed basic EDA to quickly surmise the visible trends in the data. I plotted the incidence and death rate by age group, shown below.



As expected from literature, in the incidence rate graph, there is a local peak at a very young age, then a tapering in the middle of life, then an increase as one gets older. There is an outlier, corresponding to an extremely high incidence count in Rhode Island in 2003 in 80-84 year-olds. In the mortality rate graph, we see a slightly different trend, of fewer young children dying from glioblastoma compared to a high rate in older people.

I also looked at the distribution of glioblastoma rates by regions of the US.



Upon first glance, it seems that the Northeast has the highest incidence rates in general, followed by the Midwest, then the South, then the West. We see that generally the rates decrease over the years for the West and South. However, the mortality rates do not decrease; in fact, they seem to slightly increase. This is a surprising find that I will explore.

Modeling

Results

Discussion