MA678 Midterm Project Report

Investigation of brain cancer incidence

Kayla Choi

MSSP Fall 2021

Abstract

Glioblastoma is an aggressive type of cancer that targets the central nervous system. According to literature, it afflicts two main groups of people: babies to toddlers and older people. Using data pulled from the CDC database on cancer statistics in the United States, I observed the relationship between cancer incidence and mortality rates depending on age and confirmed that there is a bimodal distribution, with a peak at baby age and another at older age. I obtained a mixed effects model, conditioned on age grouping, year, and region in the United States. There are mixed results from the modeling and visualization. For one, though incidence rate generally decreases over time, death rate does not. One would expect that, with the development of newer cancer therapy, that this would not be the case. There are also some interesting state trends that do not correspond with

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 from 1999-2018, 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.

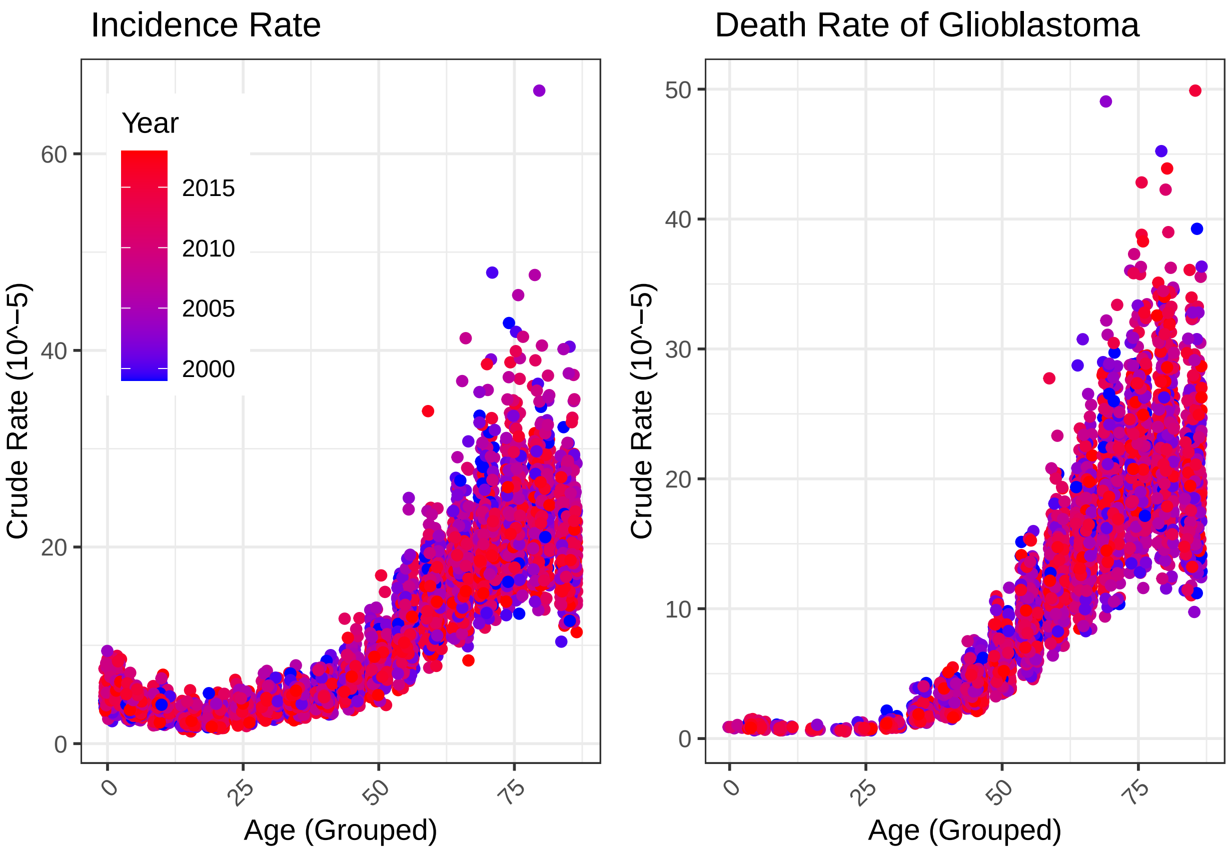
Method

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 10ˆ-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.

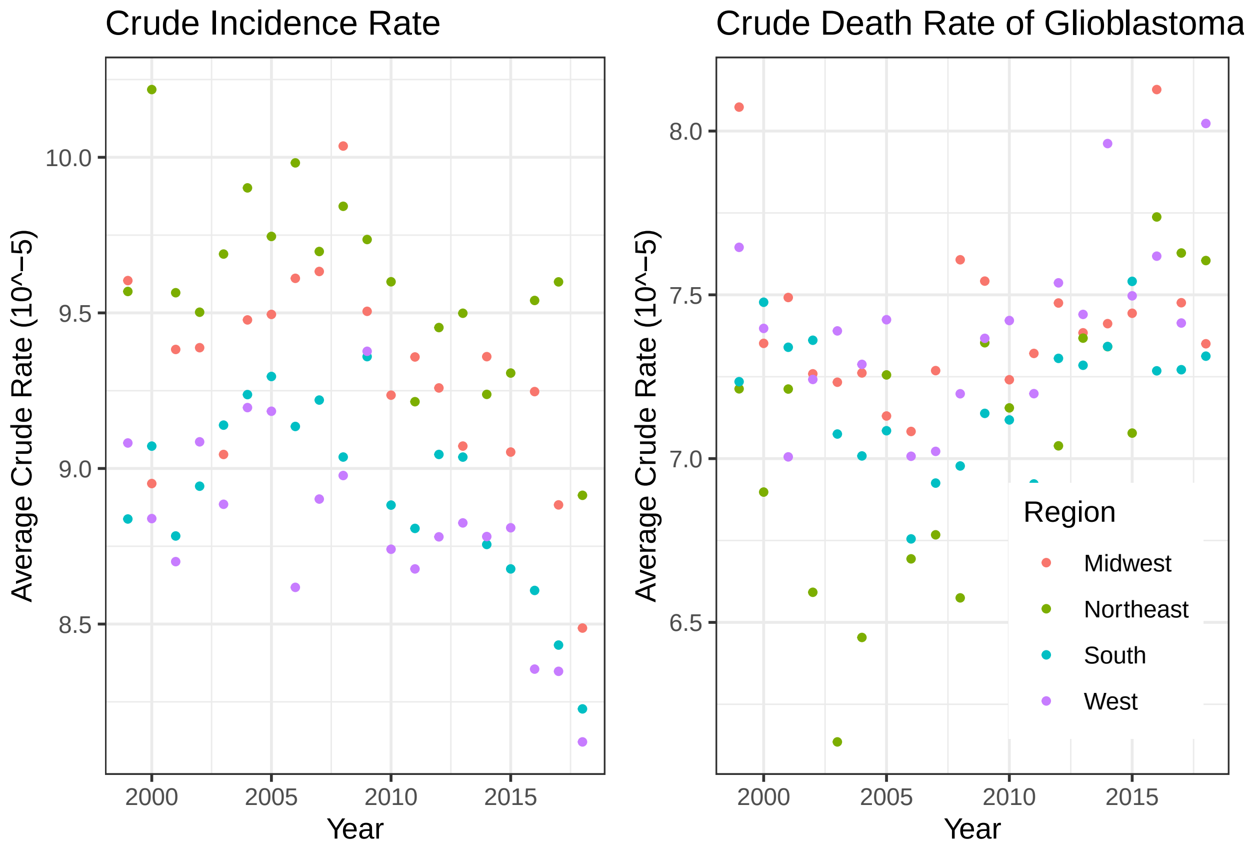
Data Visualization

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

Though I knew from EDA that there is a distribution that is not Gaussian or Poisson, I began with simple linear regressions with crude rates as the outcome and different combinations of year, states, and/or age groups as predictors. The resulting diagnostic plots showed trends in the residuals and non-normality. The posterior prediction did not capture the data trends. Therefore, I moved to mixed effects models for correlated data. I settled on using the predictors age group, state, and year, with state as the grouping.

Results

The final mixed effects models for glioblastoma incidence included the square root transformation of the predictor age group to account for unconventional data distribution mentioned above. This would scale the higher ages down, so that the linear age group predictor would not underestimate the peak at the younger age.
I found that the intercept for the states (random effect) ranged from values of approximately -2 to 10. Interestingly, the areas with the lowest populations compared to the rest of the United States generally had a higher intercept, meaning a higher baseline incidence rate of glioblastoma.

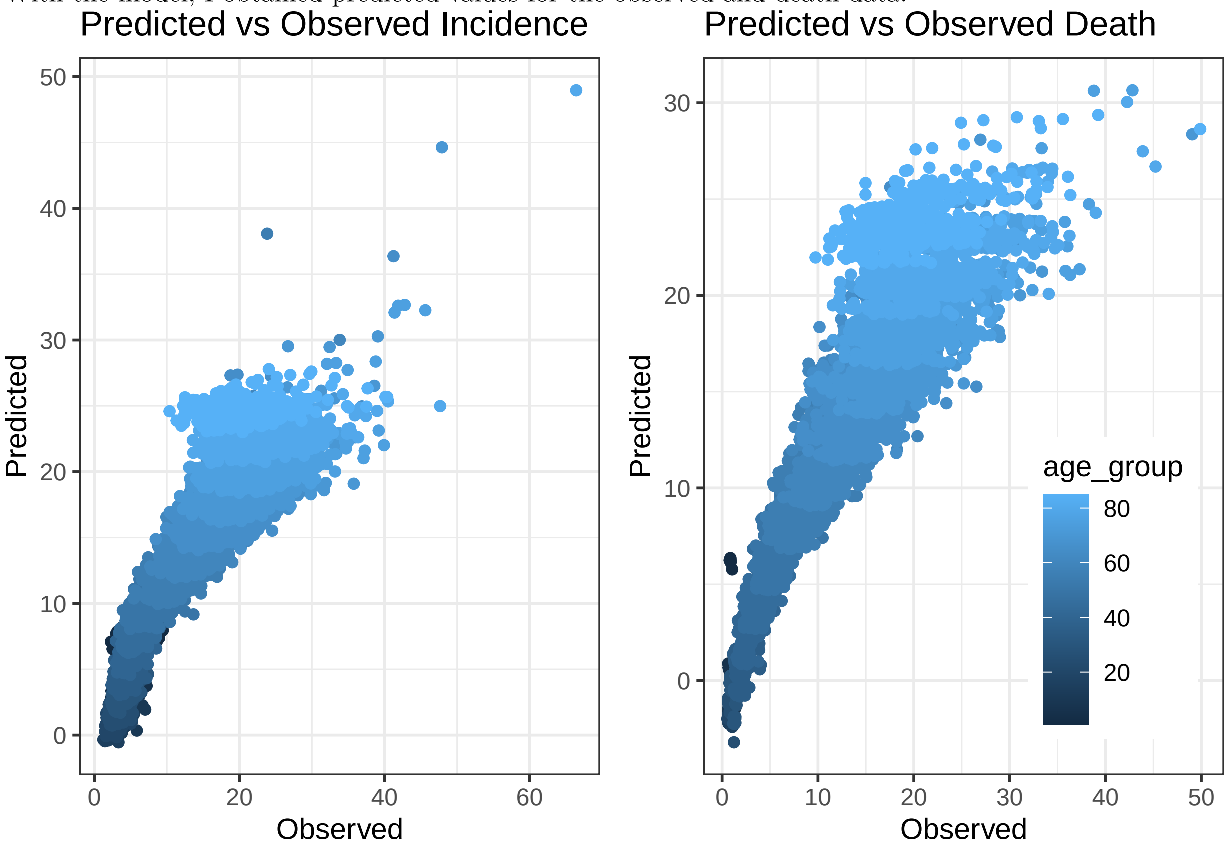
For the incidence rate:

```
##     (Intercept)       age_group sqrt(age_group)            Year
##    136.86289989      0.79607180     -6.00597729     -0.06040502
```

For the death rate:

```
##     (Intercept)       age_group sqrt(age_group)            Year 
##      58.4694432       0.9438269      -7.4851695      -0.0217849 
```

With the model, I obtained predicted values for the observed and death data.



The heteroskedasticity dictates that the model could further be improved.

Discussion

The results obtained were mixed and/or unexpected. I observed that the mortality rate does not decrease over the years, and it is unclear why. There were also unexpectedly high glioblastoma rates for some states, such as Rhode Island mentioned before. Literature search does not reveal any cancer-related phenomena in these states in the respective years. Without further information, it is impossible to hypothesize the reason for this. The model was able to capture the general trend, but there is heteroskedasticity with increasing age. Though I attempted to control for this with age group predictor transformation, this may simply come down to an intuitive reason that some states have a generally higher population of elderly, and we observed that increasing age corresponds to increasing risk of glioblastoma. Therefore, it makes sense that there is some spread. Next steps may be to dive further into investigating the age demographics of states over the years.

In regards to the data itself and the limitations posed, there was only a little over 13000 observations, once I cleaned the data of null values and only extracted glioblastoma, the cancer of interest. This may contribute to some uneven data, not missing completely at random, but there was no reasoning posted on the database website. From the map shown below in the appendix, it seems that the extremely lowly-populated states often missed data, which makes sense.

Appendix

The different models tried and their respective diagnostic plots are in the “glioblastoma_regression.R” file, for reference.

Here are the model details for the final models.

For the incidence rate:

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: ‘Crude Rate‘ ~ age_group + sqrt(age_group) + Year + (1 + Year |
##     States)
##    Data: brain_incident
## REML criterion at convergence: 41547.99
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  States   (Intercept) 2.6845
##           Year        0.0016   0.90
##  Residual             3.0616
## Number of obs: 8130, groups:  States, 46
## Fixed Effects:
##     (Intercept)        age_group  sqrt(age_group)             Year
##          136.86             0.80            -6.01            -0.06
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 2 lme4 warnings
```

```
##     (Intercept)       age_group sqrt(age_group)            Year
##    136.86289989      0.79607180     -6.00597729     -0.06040502
```

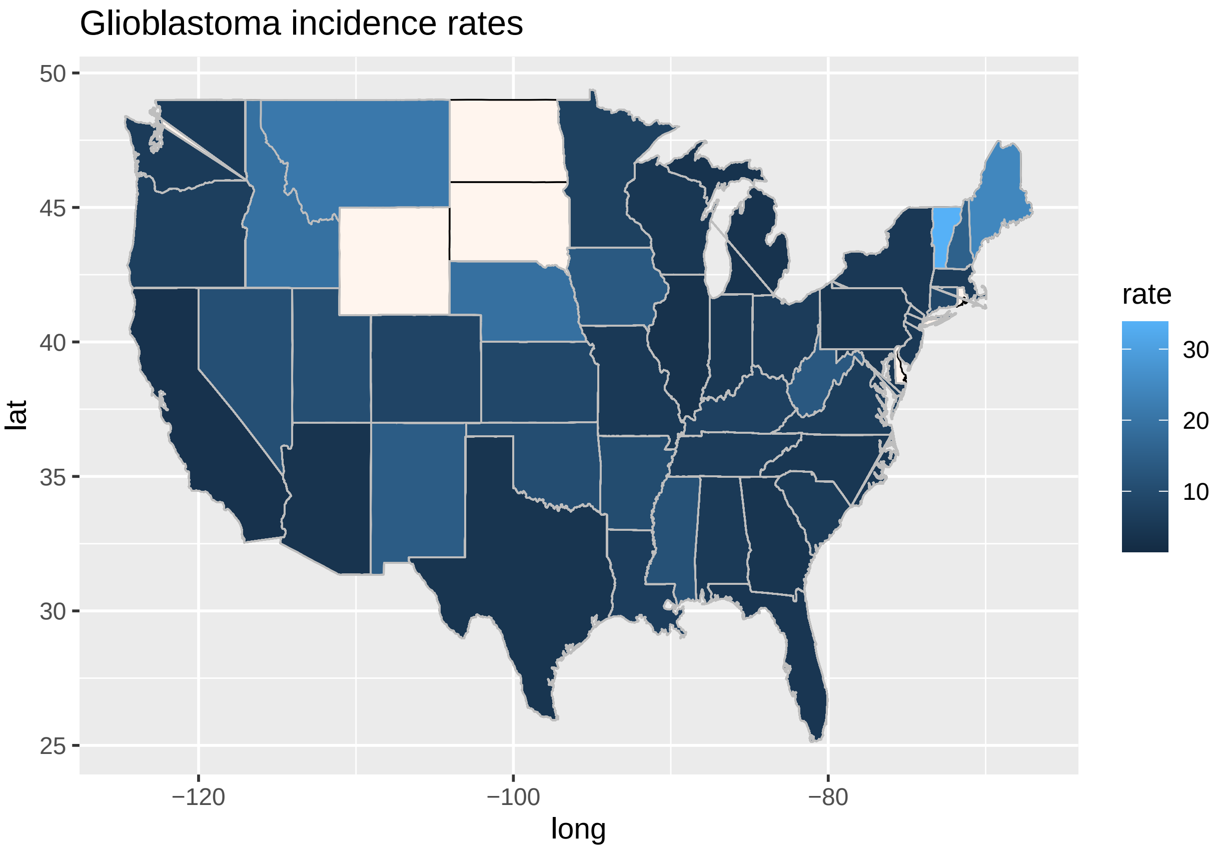
```
## $States
##                (Intercept)          Year
## Alabama        -1.16403978 -7.098289e-04
## Arizona        -1.62317450 -9.932293e-04
## Arkansas       -0.19942147 -1.202592e-04
## California     -1.84863360 -1.121188e-03
## Colorado       -1.16998503 -7.207610e-04
## Connecticut    -0.52845217 -3.309489e-04
## Delaware        7.84947081  4.793608e-03
## Florida        -1.54084821 -9.411088e-04
## Georgia        -1.75711227 -1.073339e-03
## Hawaii          0.63242050  3.862533e-04
## Idaho           2.47920544  1.513323e-03
## Illinois       -1.69844700 -1.026926e-03
## Indiana        -1.34574032 -8.238105e-04
## Iowa           -0.08520299 -5.431172e-05
## Kansas         -0.21457893 -1.407430e-04
## Kentucky       -0.91166919 -5.382928e-04
## Louisiana      -1.71997623 -1.047855e-03
## Maine           4.00082465  2.431609e-03
## Maryland       -1.51617319 -9.388415e-04
## Massachusetts  -1.19205303 -7.249819e-04
## Michigan       -1.31501831 -8.164846e-04
## Minnesota      -1.42904083 -8.619102e-04
## Mississippi    -0.71157339 -4.374214e-04
## Missouri       -1.49327903 -9.107124e-04
## Montana         2.67163180  1.631682e-03
## Nebraska        2.08974848  1.269911e-03
## Nevada         -0.30033117 -1.868198e-04
## New Hampshire   2.67389304  1.631063e-03
## New Jersey     -1.24660832 -7.595805e-04
## New Mexico     -0.08623483 -5.890837e-05
## New York       -1.51288207 -9.246399e-04
## North Carolina -1.60785071 -9.708018e-04
## Ohio           -1.32455609 -7.953447e-04
## Oklahoma       -0.76632623 -4.733885e-04
## Oregon         -0.42280602 -2.592786e-04
## Pennsylvania   -1.26274844 -7.477856e-04
## Rhode Island   10.40424893  6.351662e-03
## South Carolina -1.05860735 -6.493904e-04
## Tennessee      -1.54526189 -9.384257e-04
## Texas          -1.56192489 -9.604683e-04
## Utah            0.65558017  3.940317e-04
## Vermont         6.15679327  3.760763e-03
## Virginia       -1.91281141 -1.160888e-03
## Washington     -1.01395150 -6.209463e-04
## West Virginia   0.47651380  2.790850e-04
## Wisconsin      -1.00301050 -6.033694e-04
##
## with conditional variances for "States"
```

For the death rate:

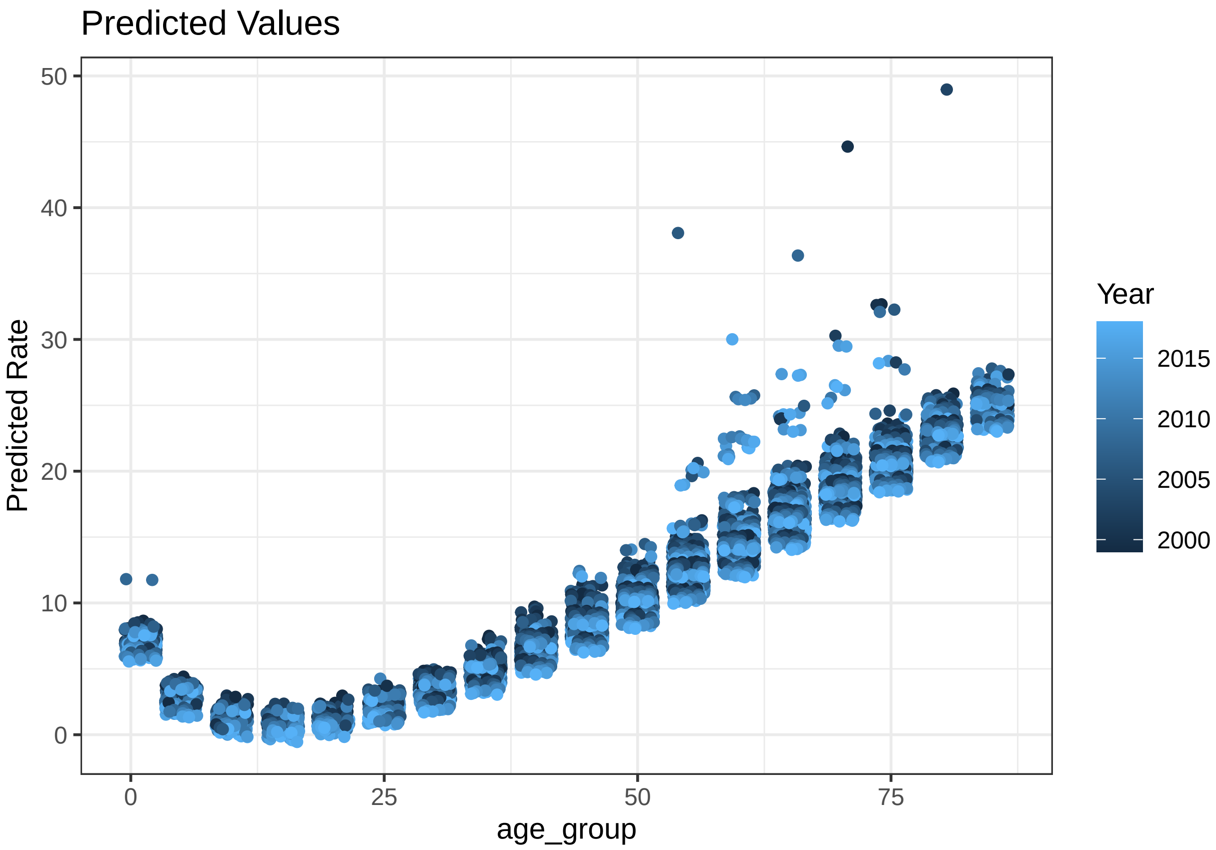
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: ‘Crude Rate‘ ~ age_group + sqrt(age_group) + Year + (1 + Year |
##     State)
##    Data: brain_death
## REML criterion at convergence: 28715.77
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  State    (Intercept) 3.15882
##           Year        0.00011  1.00
##  Residual             3.15765
## Number of obs: 5549, groups:  State, 41
## Fixed Effects:
##     (Intercept)        age_group  sqrt(age_group)             Year
##          58.469            0.944           -7.485           -0.022
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warnings
```

```
##     (Intercept)       age_group sqrt(age_group)            Year
##      58.4694432       0.9438269      -7.4851695      -0.0217849
```

```
## $State
##                (Intercept)          Year
## Alabama         -0.1557725 -5.255828e-06
## Arizona         -2.4724449 -8.342128e-05
## Arkansas         3.0286754  1.021887e-04
## California      -1.8536280 -6.254215e-05
## Colorado        -0.8304973 -2.802131e-05
## Connecticut     -0.4758167 -1.605425e-05
## Florida         -2.9307995 -9.888634e-05
## Georgia         -2.8550677 -9.633112e-05
## Idaho            9.4484915  3.187959e-04
## Illinois        -3.3560764 -1.132354e-04
## Indiana         -1.8767362 -6.332183e-05
## Iowa             1.6780084  5.661667e-05
## Kansas           2.9747032  1.003677e-04
## Kentucky        -0.4540644 -1.532031e-05
## Louisiana       -0.8510028 -2.871317e-05
## Maine            5.1429928  1.735266e-04
## Maryland        -2.8544027 -9.630869e-05
## Massachusetts   -2.3822991 -8.037972e-05
## Michigan        -1.6706949 -5.636991e-05
## Minnesota       -1.6187812 -5.461832e-05
## Mississippi      2.6650915  8.992125e-05
## Missouri        -2.1121632 -7.126523e-05
## Nebraska         6.3904496  2.156163e-04
## Nevada           1.0241125  3.455396e-05
## New Hampshire    6.9550702  2.346668e-04
## New Jersey      -3.6823106 -1.242426e-04
## New Mexico       4.1140402  1.388094e-04
## New York        -3.8233111 -1.290000e-04
## North Carolina  -2.4193639 -8.163030e-05
## Ohio            -2.4301191 -8.199319e-05
## Oklahoma         0.3482466  1.174998e-05
## Oregon           0.6450851  2.176543e-05
## Pennsylvania    -2.9845040 -1.006984e-04
## South Carolina  -0.2662193 -8.982346e-06
## Tennessee       -0.5186693 -1.750011e-05
## Texas           -2.2605932 -7.627331e-05
## Utah             3.9683026  1.338921e-04
## Virginia        -2.8541340 -9.629962e-05
## Washington      -0.1223781 -4.129085e-06
## West Virginia    2.9726475  1.002983e-04
## Wisconsin       -1.2440667 -4.197531e-05
##
## with conditional variances for "State"
```



This is an example of glioblastoma incidence rates (crude rate 10ˆ-5) in 2017 across the states. The excluded states (North Dakota, South Dakota, and Wyoming), are missing data. The pooled incidence rates look generally similar across the US, except a few states in the NorthWest and in the NorthEast. These were confirmed in the results section.



This plot using our final model. It depicts the prediction of brain incidence plotted against age group, which corresponds to the original EDA. The general shape is similar; however, we see more and more deviation with the higher age groups, and the younger age is also over-estimated.