Analysis of treatment disparities for Nasopharynx cancer patients among the SEER database populations

Simu Huang, Masanao Yajima

2021/11/23

Abstract

In this project, we study the factors that affect whether patients receives standard treatment defined in the NCCN guidelines for Nasopharynx cancer population in the SEER Database. We found that factors cancer stage and age at diagnosis impacted whether the patient receives standardized treatment that meets the NCCN guidelines.

Introduction

Nasopharynx cancer is a type of head and neck cancer and it is the 23rd most common cancer worldwide. It starts in the Nasopharynx, which is located behind the nose and above the back of throat.

<_____> Please add more detailed information.

For this cancer, We use machine learning methods to study which factors affect whether patients receive standard treatment, especially the patients' own background, such as income and education. Besides, we try to find the impact of standard treatment and other factors on patient survival situation.

Data description and cleaning

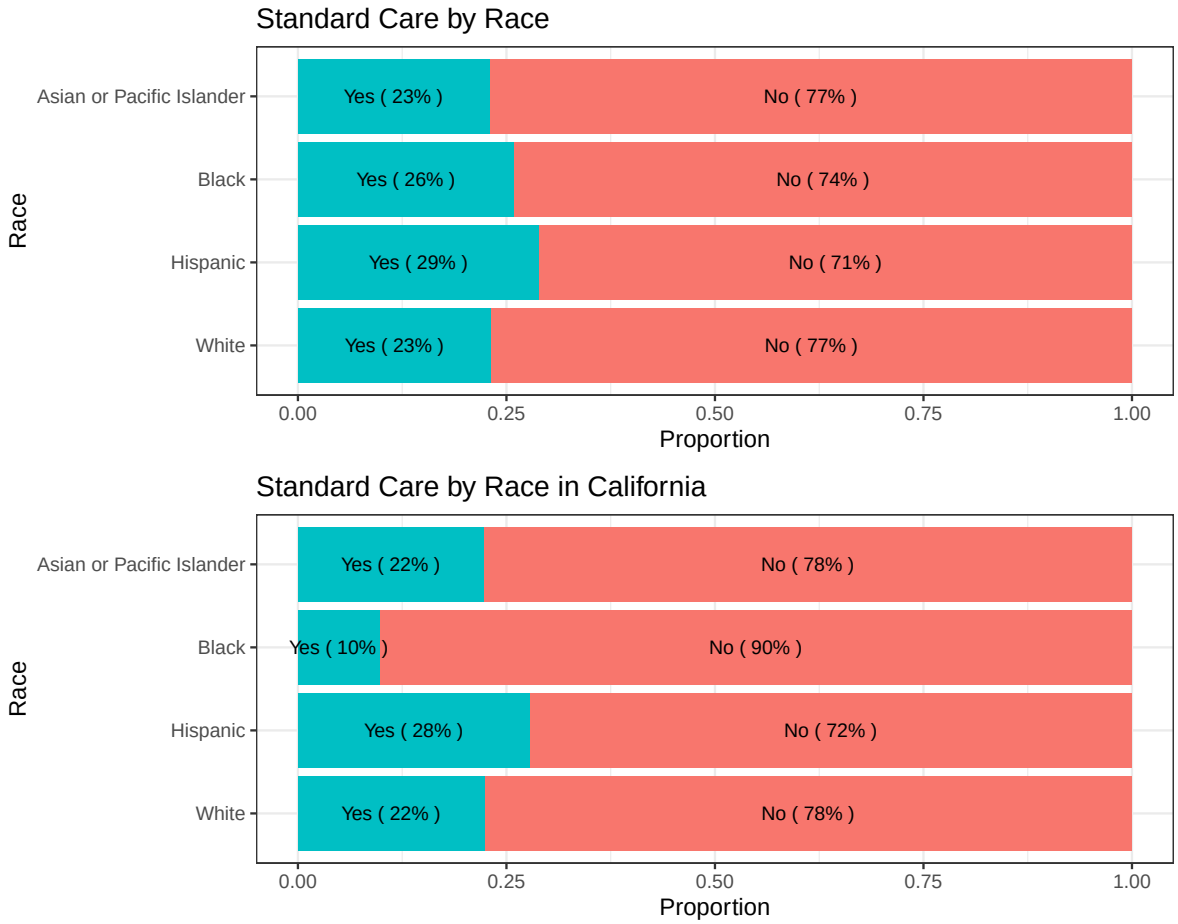
Our analysis is based on the SEER database sorted out by Anand Devaiah, Pratima Agarwal and Jacob Bloom. We query the data of patients with Nasopharynx cancer. Among four states, we deleted the data for the Alaska region and Connecticut region. Because the variance of the data in these areas are too small. For example, in Alaska region, all the people suffering from Nasopharynx in this area are American Indian/Alaska Native who bought insurance. And none of them received standard treatment.

Based on the NCCN guideline, the patients with Nasopharynx cancer in stage I are suggested to take radiation and patients in other stages are suggested to take the radiation and chemotherapy. We compare the situation in the SEER database with the NCCN guideline and then transfer the result into a binary variable `receive_standard`. If standard treatment has been given, it is 1; if not, it is 0.

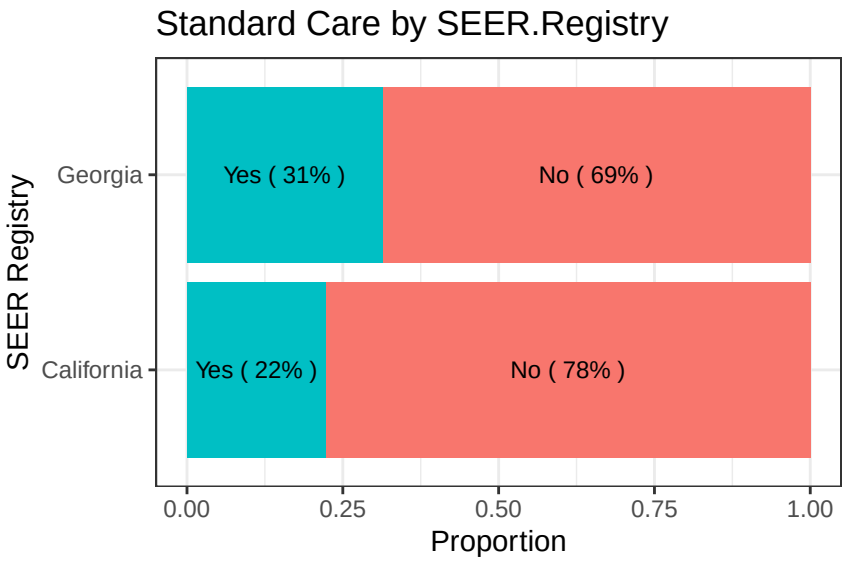
The value `Any medicaid` and `Insured/No specificsin` the `Insurance` column are clustered into the `Insured` and other values are clustered into the `Uninsured`.

The variables we use in the analysis are `Sex`, `std_age`(the difference between age of diagnosis and age 60), `Race`, `Insurance`, `SEER.Registry`, `log(median household income)`(log value of median household income.), `AJCC.7.Stage`, `std_edu`(the proportion of people who in the community with less than a high school education minus the proportion of people who in the state with less than a high school education.), `std_unemployed`(the proportion of people who in the community are unemployed minus the proportion of people who in the state are unemployed.), `std_language_isolation`(the proportion of people who in the community are language isolated minus the proportion of people who in the state are language isolated.), and `receive_standard`.

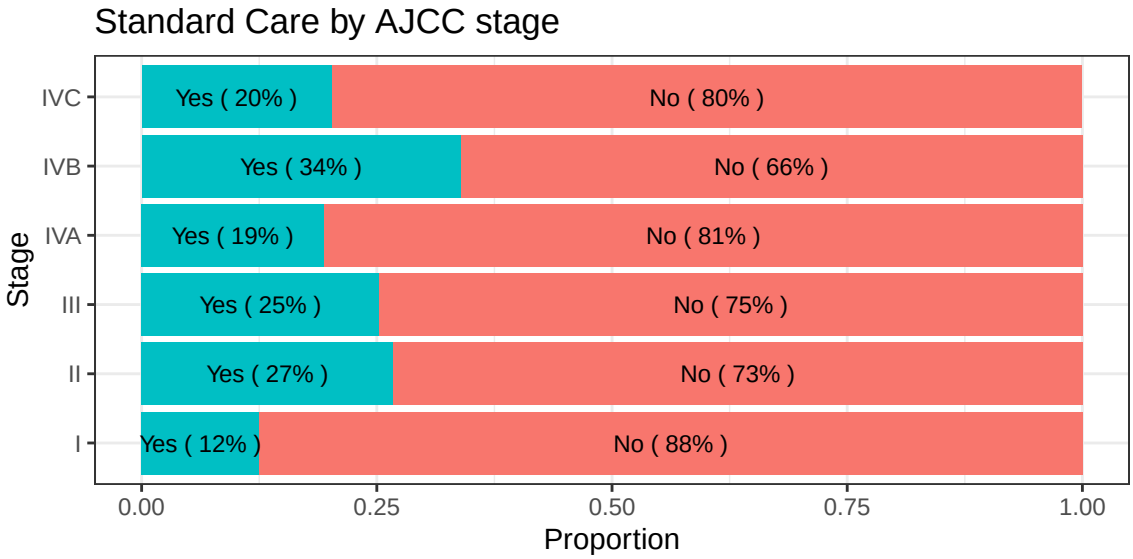
EDA



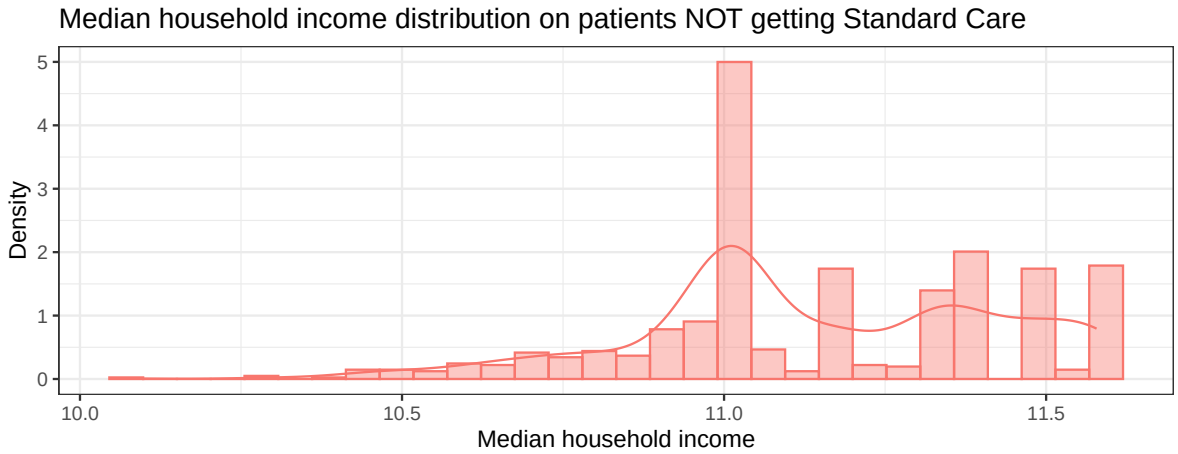
According to all the data, race has no major influence on receiving standard treatment. But in California, the proportion of blacks receiving standard treatment is significantly lower than that of others. This may because blacks in California are treated differently from other races. It is also possible that in the other region (George), white people have been treated differently, or relatively few blacks have been diagnosed with the Nasopharynx cancer.

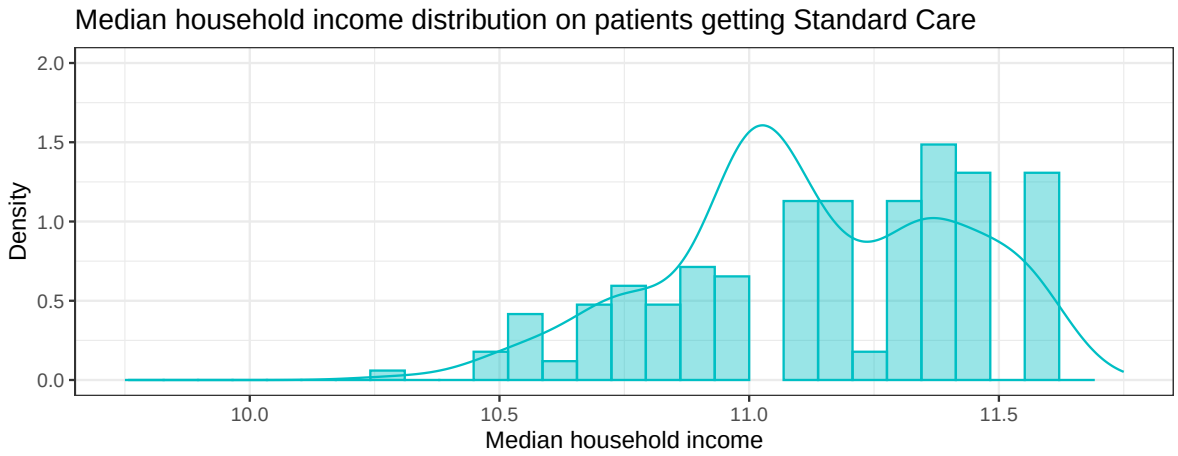


There are also differences between regions. Georgia has the highest rate of receiving standard treatment.

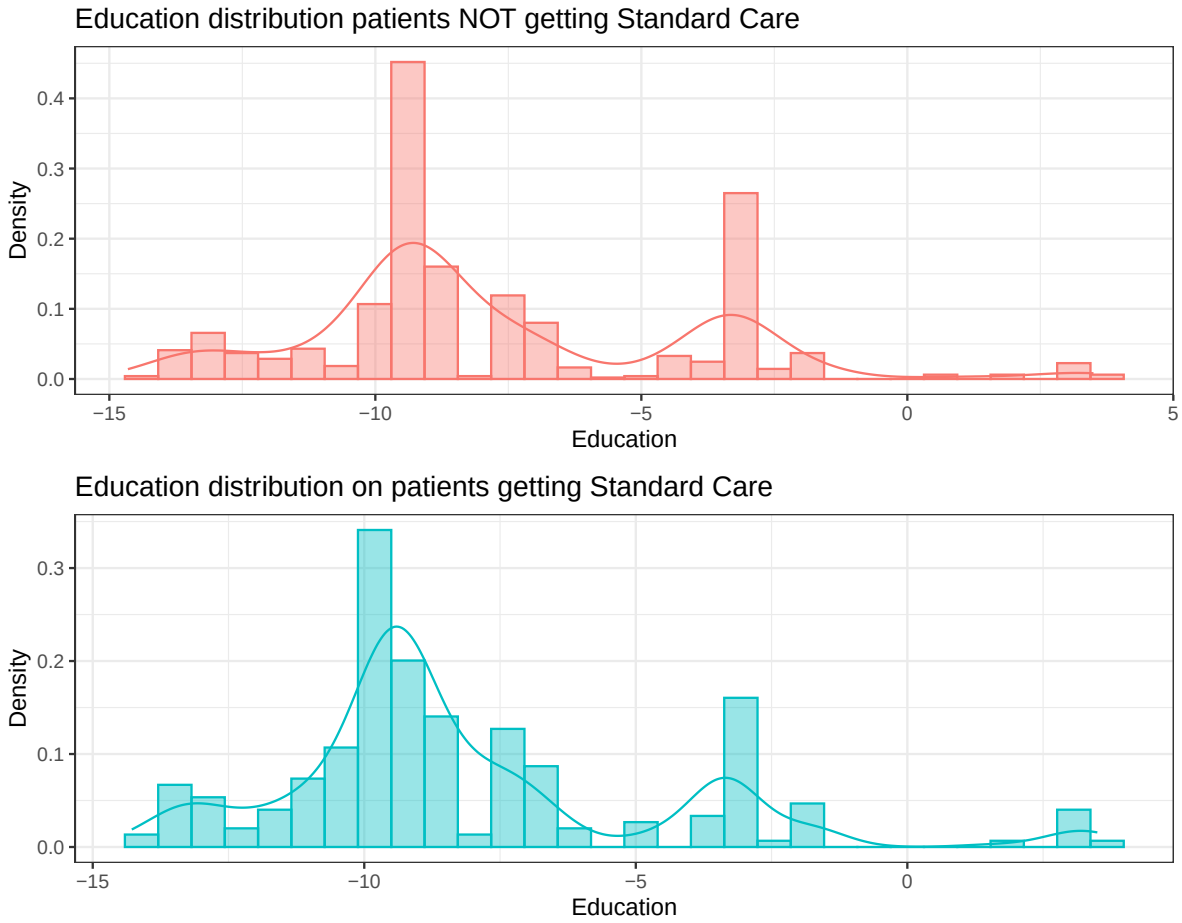


From the plot above, AJCC stage is also an important factor. Patients in stage II, III, IVB are more likely to be given standardized treatment.

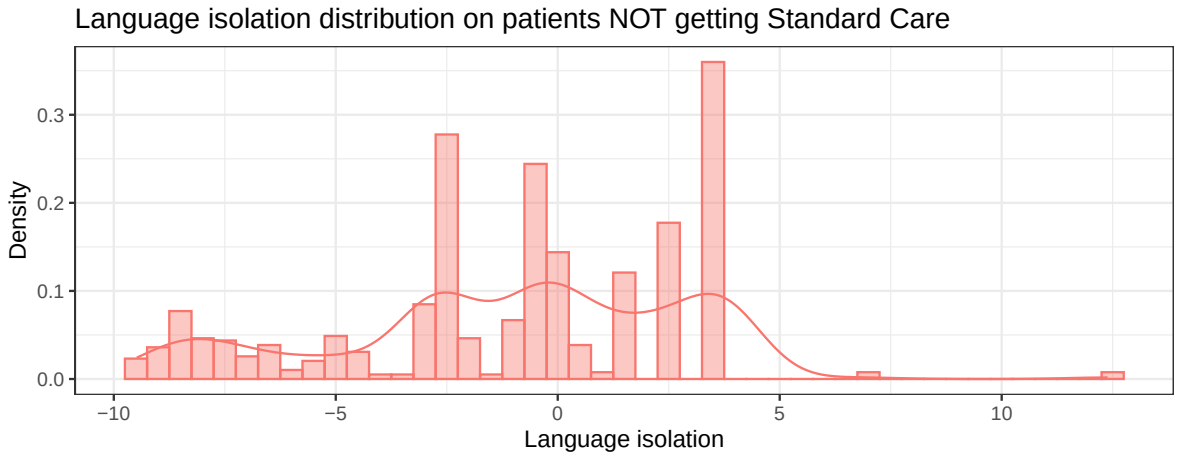


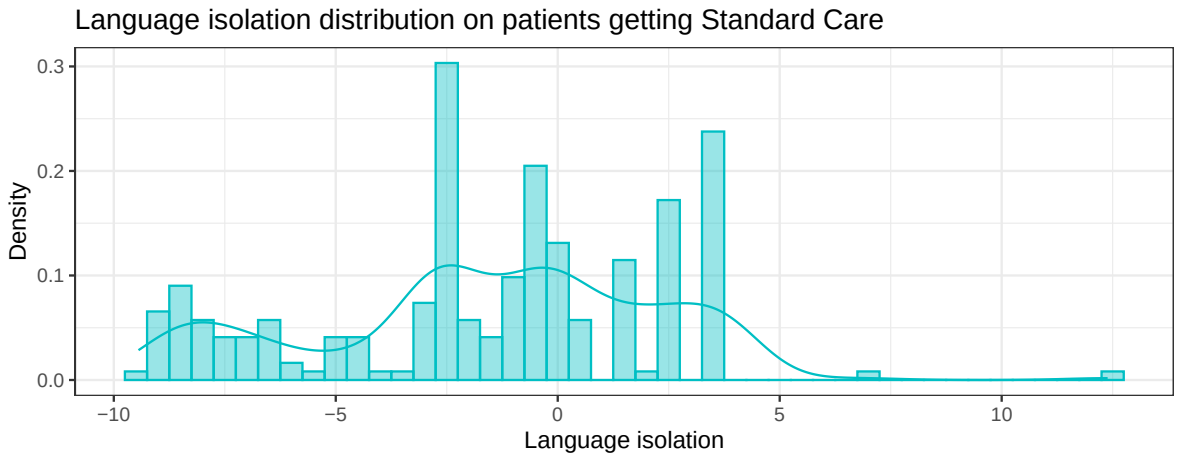


Generally speaking, among people who have received standard treatment, the median household income of the community where the patients live is higher.



The level of education and poverty have little effect on whether they are given standard treatment. We can see that the difference between the upper and lower plots is not very big.



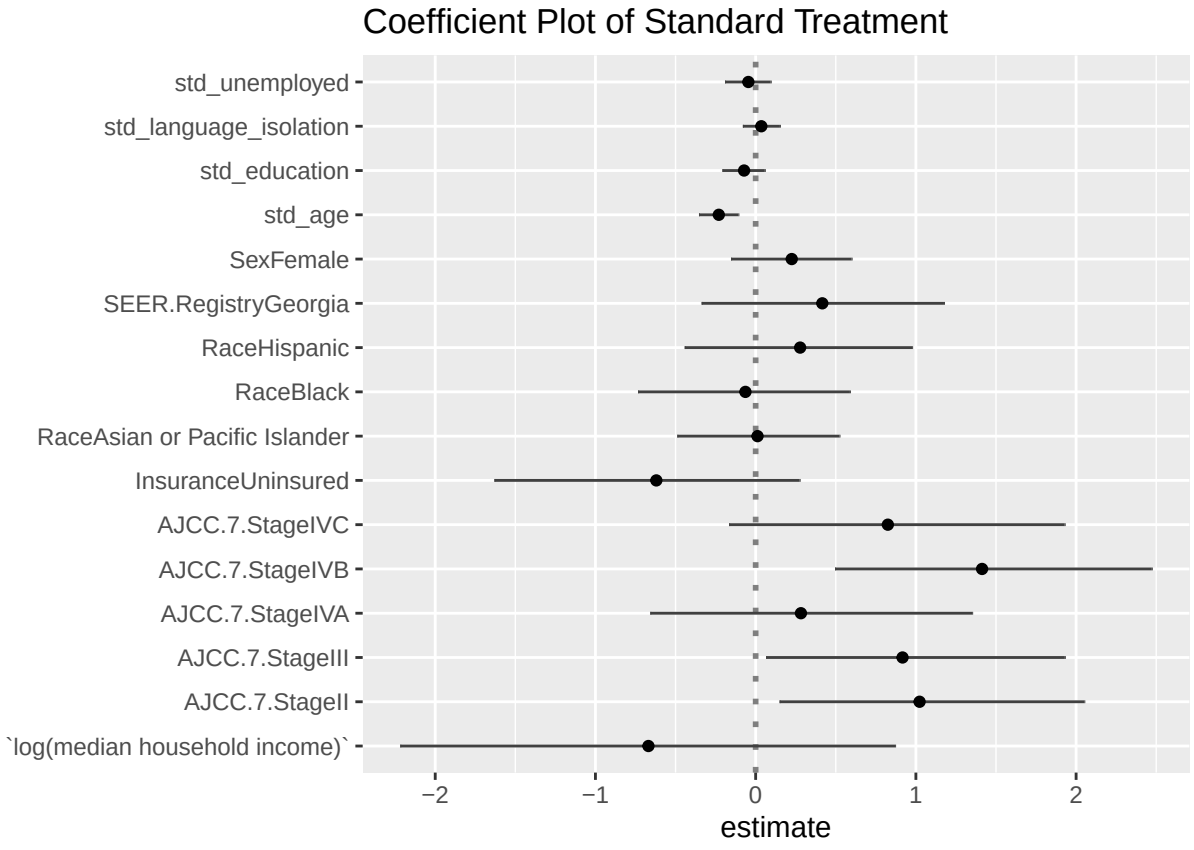


Language isolation has no significant impact on whether standardized treatment is given.

Modeling

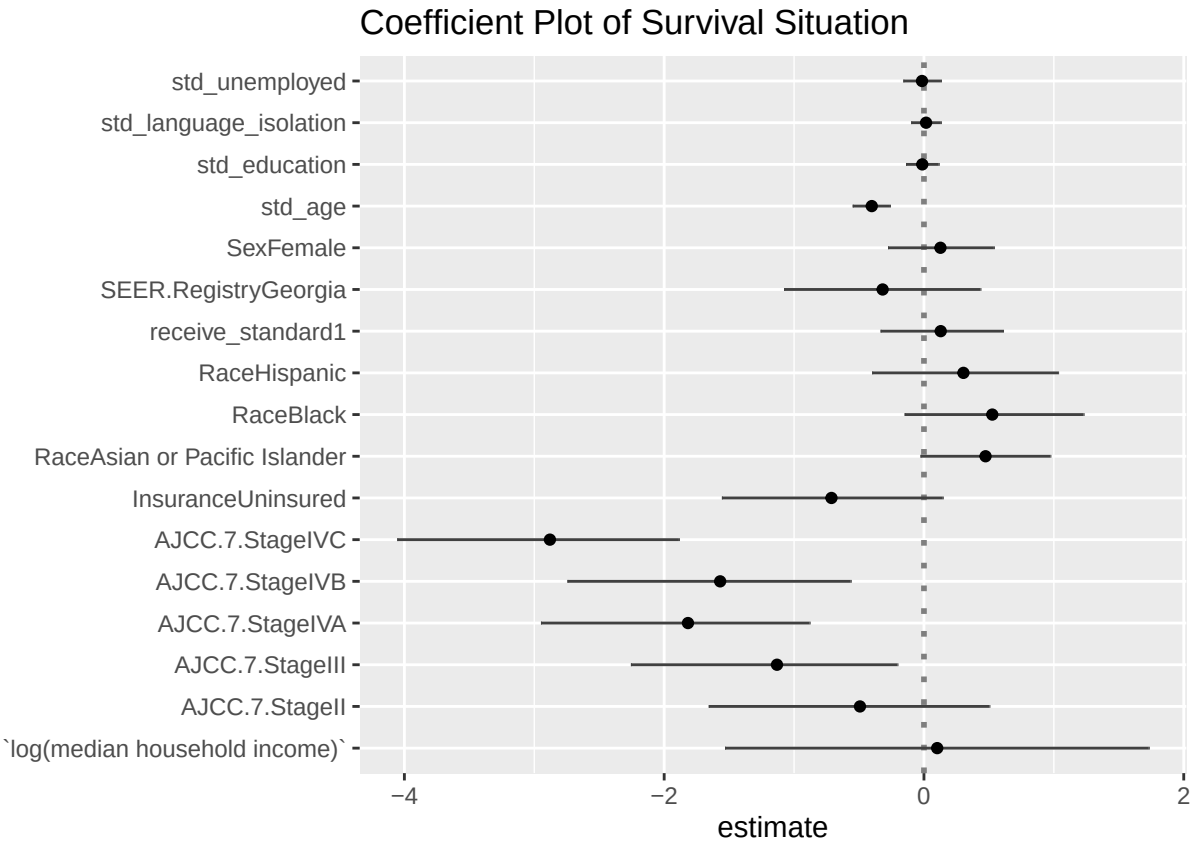
Standard therapy

We use the logistic model here to find that, whether the patient has received standard treatment is related to these independent factors. The coefficient plot below displays the point estimates and their confidence intervals. The x-axis represents the coefficient estimate. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease. The y-axis represents different independent variables. The line represents the 95% confidence interval, which means that we are 95% confident that the interval captured the true mean value.



Survival Situation

We use the logistic model here to find that, the survival situation is related to these independent factors, especially the factor `receive_standard`. The coefficient plot below displays the result of the model.



Discussion

Standard Therapy

Based on the result of modeling, we can find that the stage of cancer and age at diagnosis impact whether the patient receives standard treatment that meets the NCCN guidelines. When patients' cancer are in stage II, III, or IVB, they are more probably to receive standard treatment. The older the patient, the less likely it is to be given standardized treatment.

Meanwhile, compared to patients in other stages, patients in the Stage IVA and IVC are less likely to receive standard therapy. We can find that people in these two stages are suggested to not only receive radiation but also receive chemotherapy based on the NCCN guideline but most of patients in that stage only receive chemotherapy in our database. A small number of patients received radiation as recommended.

Survival Situation

According to the results of survival analysis, we can find that cancer stages, age at diagnosis and whether they have insurance have impact on the patient's survival status. With the exception of stage IVA, the more advanced the cancer, the worse the survival situation. The survival situation of stage IVA patients is worse than that of IVB stage patients and better than that of IVC stage patients. Besides, The older the patient is at the time of diagnosis,the more likely the patient is to die of the Nasopharynx cancer. Survival of uninsured patients is worse than that of patients with insurance.

Appendix

Modeling results

result of standard therapy modeling

```
summary(model_st_x)
```

```
##
## Call:
## glm(formula = receive_standard ~ Sex + AJCC.7.Stage + std_age +
##     SEER.Registry + `log(median household income)` + Insurance +
##     Race + std_education + std_unemployed + std_language_isolation,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3481  -0.7502  -0.6076  -0.3785   2.3980
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     4.52125    8.48351   0.533 0.594071
## SexFemale                       0.22517    0.19161   1.175 0.239926
## AJCC.7.StageII                  1.02301    0.47750   2.142 0.032159 *
## AJCC.7.StageIII                 0.91670    0.46785   1.959 0.050066 .
## AJCC.7.StageIVA                 0.28260    0.50435   0.560 0.575256
## AJCC.7.StageIVB                 1.41309    0.49686   2.844 0.004455 **
## AJCC.7.StageIVC                 0.82496    0.52629   1.568 0.116997
## std_age                        -0.23001    0.06265  -3.671 0.000241 ***
## SEER.RegistryGeorgia            0.41593    0.38490   1.081 0.279877
## `log(median household income)` -0.66927    0.78591  -0.852 0.394443
## InsuranceUninsured             -0.61998    0.48034  -1.291 0.196805
## RaceHispanic                    0.27765    0.36063   0.770 0.441361
## RaceBlack                      -0.06388    0.33627  -0.190 0.849345
## RaceAsian or Pacific Islander   0.01162    0.25766   0.045 0.964020
## std_education                  -0.07228    0.06732  -1.074 0.282927
## std_unemployed                 -0.04561    0.07237  -0.630 0.528528
## std_language_isolation          0.03532    0.05886   0.600 0.548448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 774.89  on 714  degrees of freedom
## Residual deviance: 730.45  on 698  degrees of freedom
## AIC: 764.45
##
## Number of Fisher Scoring iterations: 4
```

result of survival situation modeling

```
summary(model_sur_x)
```

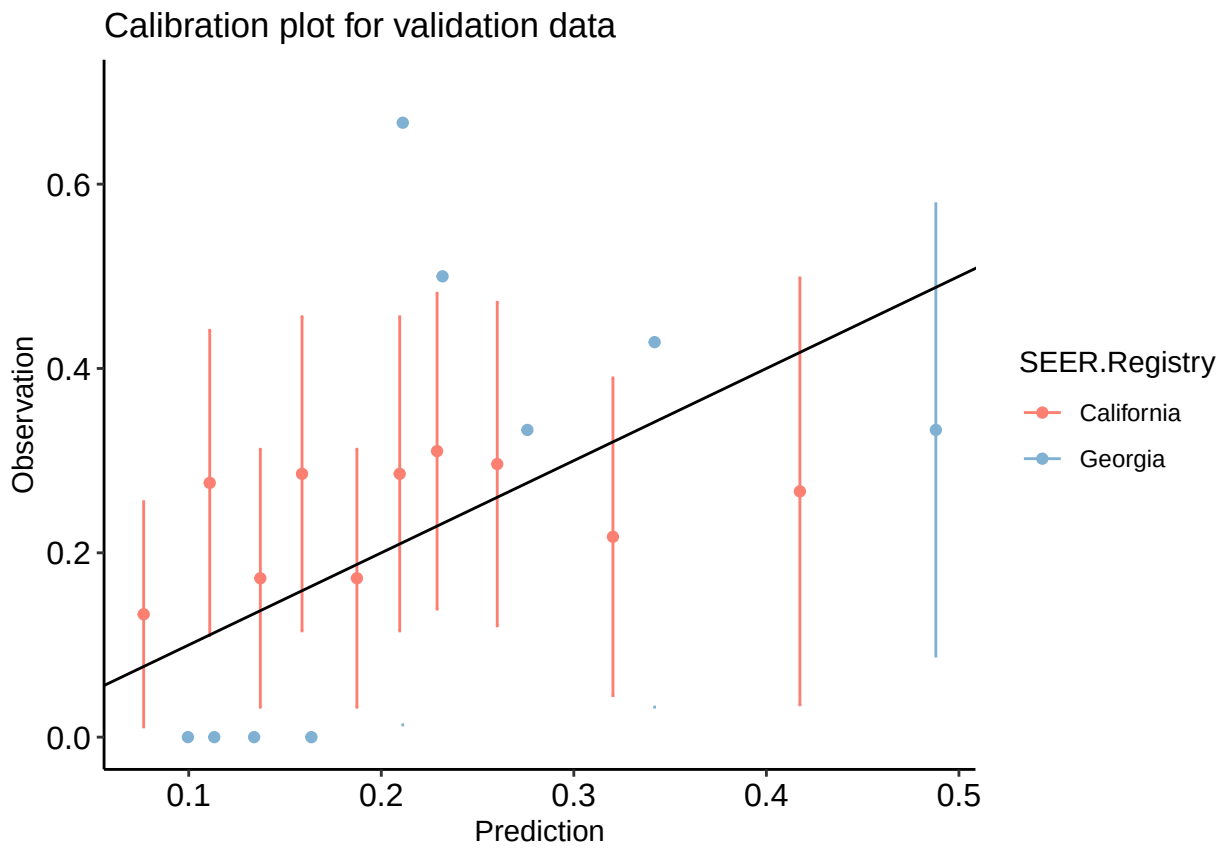
```
##
## Call:
```

```
## glm(formula = survival ~ receive_standard + Race + Sex + AJCC.7.Stage +
##     std_age + SEER.Registry + `log(median household income)` +
##     Insurance + std_education + std_unemployed + std_language_isolation,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4405   0.2053   0.4908   0.7118   1.7171
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     0.82579    8.98976   0.092 0.926810
## receive_standard1               0.12903    0.23980   0.538 0.590534
## RaceHispanic                    0.30379    0.36379   0.835 0.403690
## RaceBlack                       0.52594    0.35034   1.501 0.133297
## RaceAsian or Pacific Islander   0.47416    0.25468   1.862 0.062636 .
## SexFemale                       0.12697    0.20760   0.612 0.540787
## AJCC.7.StageII                 -0.49258    0.54088  -0.911 0.362444
## AJCC.7.StageIII                -1.13144    0.51372  -2.202 0.027633 *
## AJCC.7.StageIVA                -1.81675    0.51807  -3.507 0.000454 ***
## AJCC.7.StageIVB                -1.56893    0.54680  -2.869 0.004114 **
## AJCC.7.StageIVC                -2.87949    0.54432  -5.290 1.22e-07 ***
## std_age                        -0.40188    0.07308  -5.499 3.82e-08 ***
## SEER.RegistryGeorgia           -0.31819    0.38471  -0.827 0.408185
## `log(median household income)`  0.10193    0.83029   0.123 0.902290
## InsuranceUninsured             -0.71210    0.43103  -1.652 0.098519 .
## std_education                  -0.01298    0.06406  -0.203 0.839481
## std_unemployed                 -0.01522    0.07354  -0.207 0.836006
## std_language_isolation          0.01623    0.05811   0.279 0.780036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 800.27  on 714  degrees of freedom
## Residual deviance: 676.92  on 697  degrees of freedom
## AIC: 712.92
##
## Number of Fisher Scoring iterations: 5
```

Model validation

We can find that, in the calibration plot of standard therapy, these red points are evenly distributed around the straight line, which means that this model run very well for the data whose `SEER.Registry` is California. The c-stat score here is 0.676.

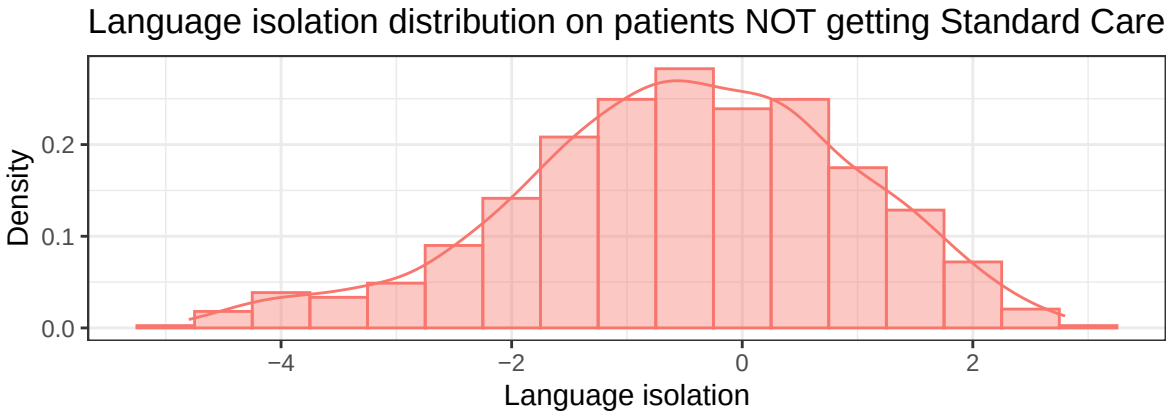
```
## $calibration_plot
```

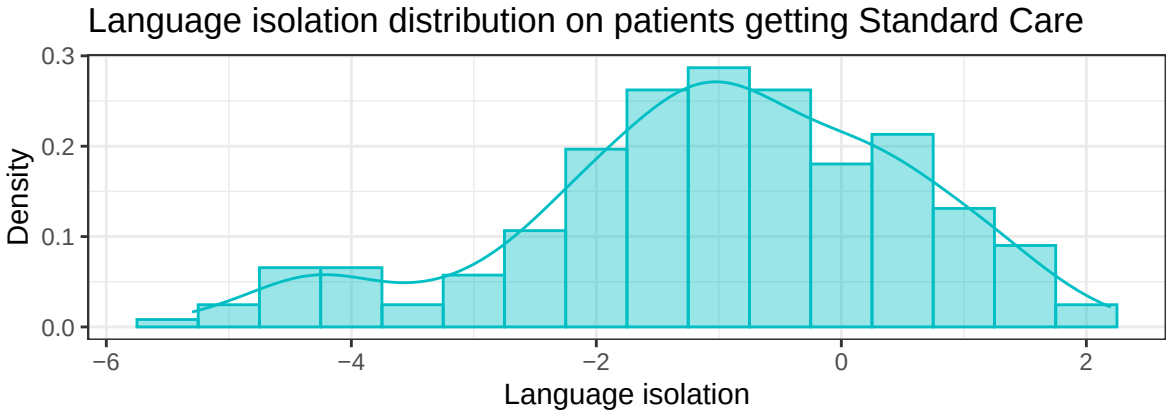


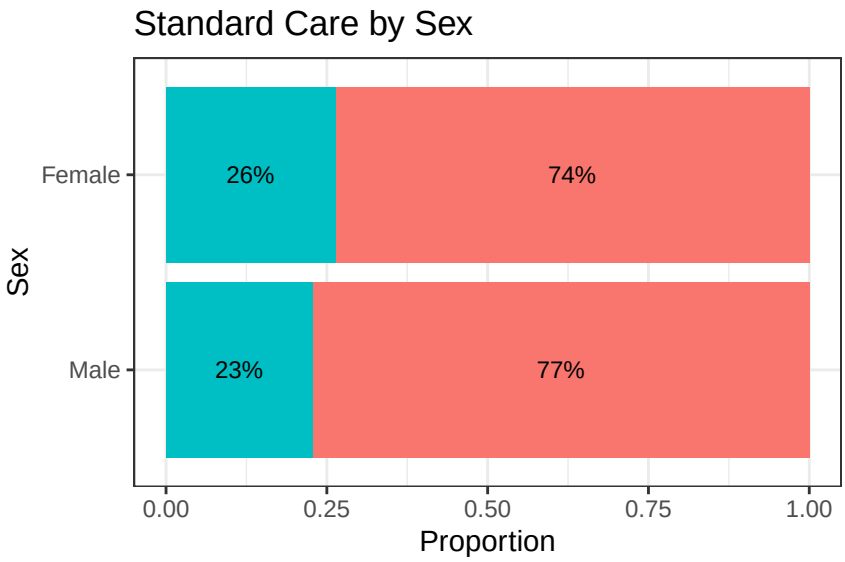

```
## The c-stat score here is 0.6610979
```

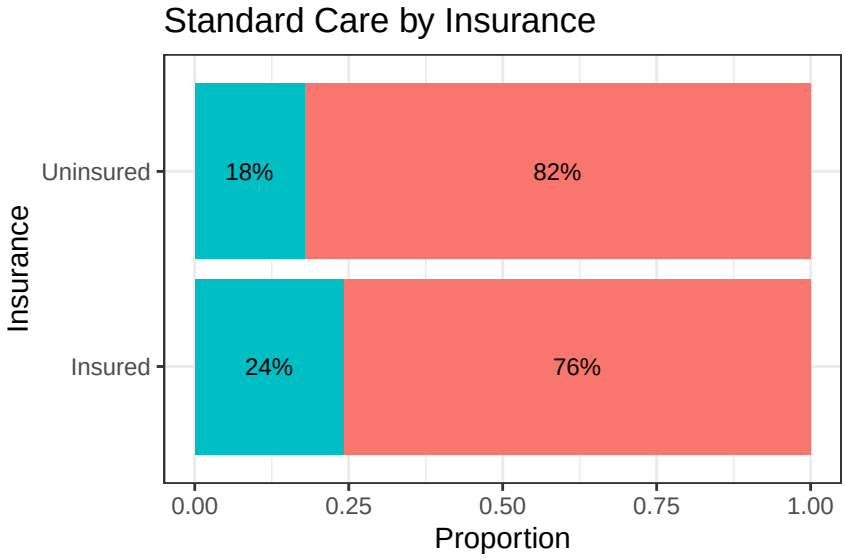
EDA

Following are some other EDA plots related to the factors we selected.


Language isolation distribution on patients NOT getting Standard Care
Density
0.0
0.1
0.2
−4
−2
0
2
Language isolation


Language isolation distribution on patients getting Standard Care
Density
0.0
0.1
0.2
0.3
−6
−4
−2
0
2
Language isolation


Standard Care by Sex
Female
Male
Sex
26%
74%
23%
77%
0.00
0.25
0.50
0.75
1.00
Proportion


Standard Care by Insurance
Uninsured
Insured
Insurance
18%
82%
24%
76%
0.00
0.25
0.50
0.75
1.00
Proportion