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 such as cancer stage, patient's location, household income, age at diagnosis, and whether they have insurance impact whether the patient receives standardized treatment that meets the NCCN guidelines. Patients in stage II, IVA, IVB and IVC, SEER registry in Georgia, have insurance are more likely to receive standard treatment.

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.
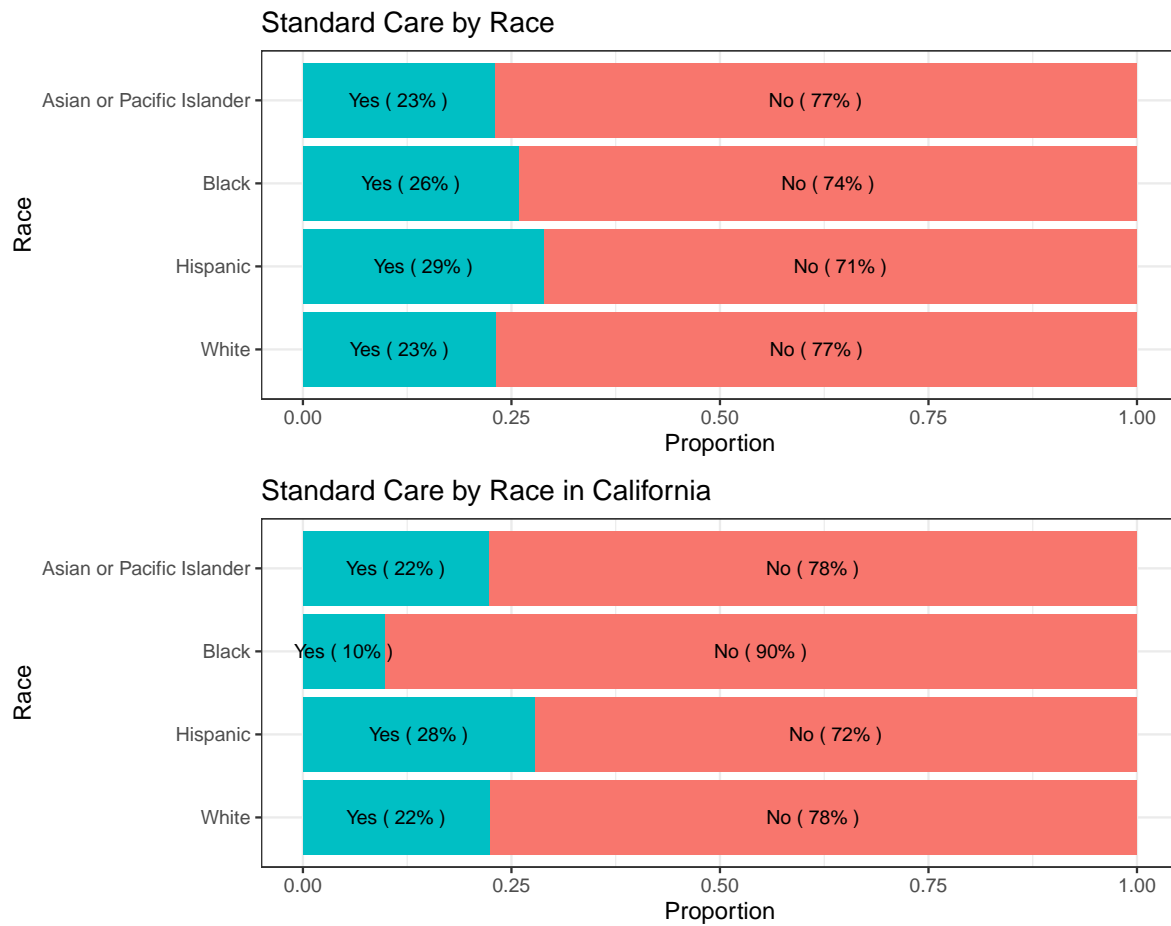
Data description and cleaning

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 specifics` in 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

Standard Care by Race

Race		
Asian or Pacific Islander	Yes (23%)	No (77%)
Black	Yes (26%)	No (74%)
Hispanic	Yes (29%)	No (71%)
White	Yes (23%)	No (77%)

Standard Care by Race in California

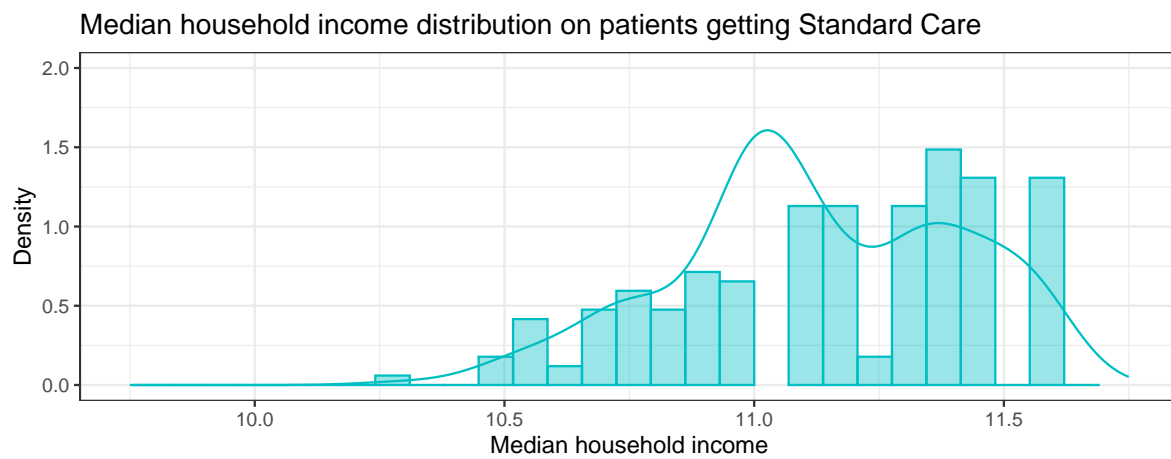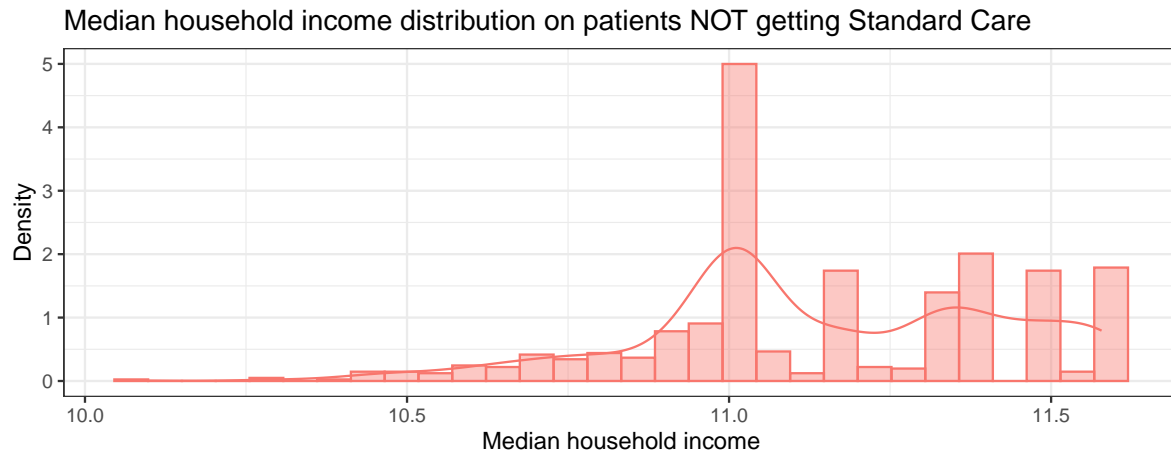| Race		
Asian or Pacific Islander	Yes (22%)	No (78%)
Black	Yes (10%)	No (90%)
Hispanic	Yes (28%)	No (72%)
White	Yes (22%)	No (78%)

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.
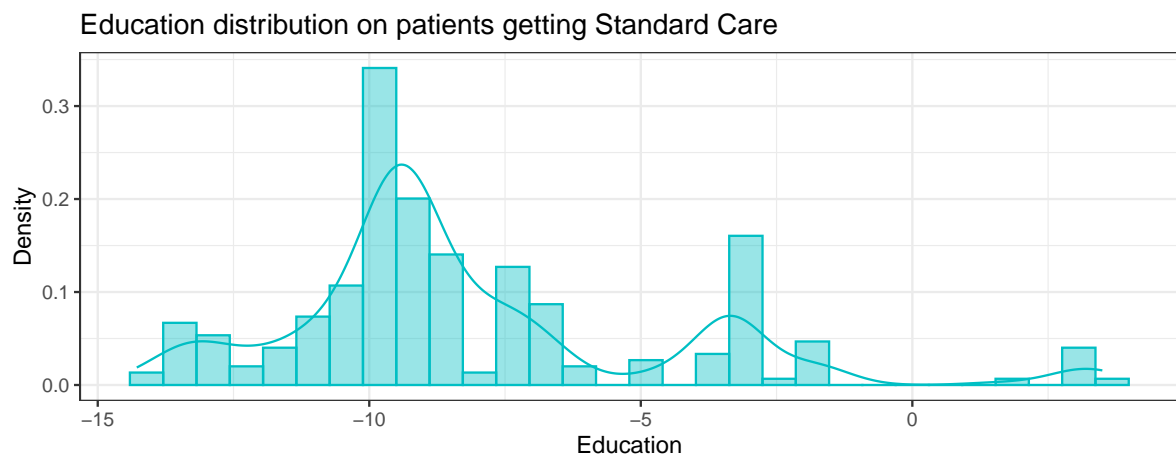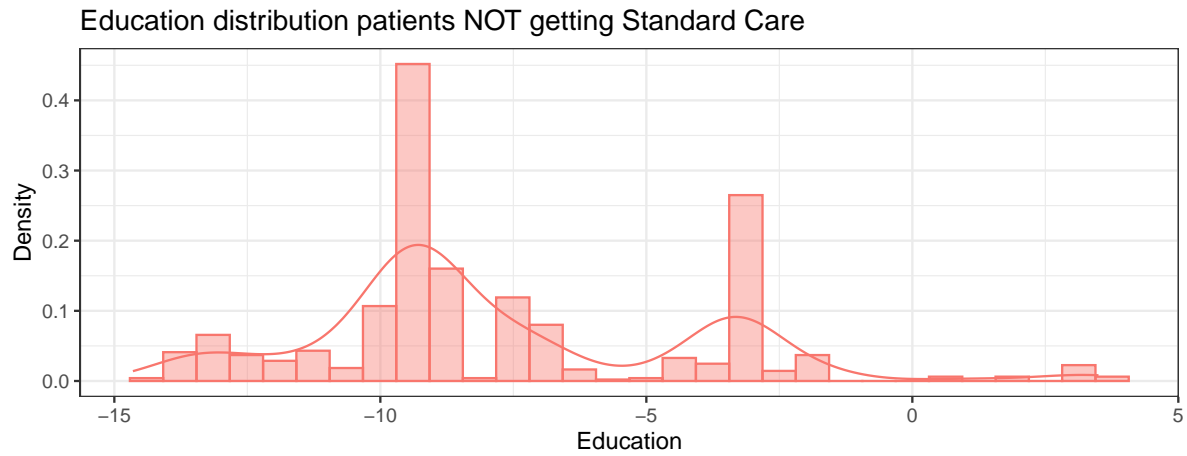
Standard Care by SEER.Registry
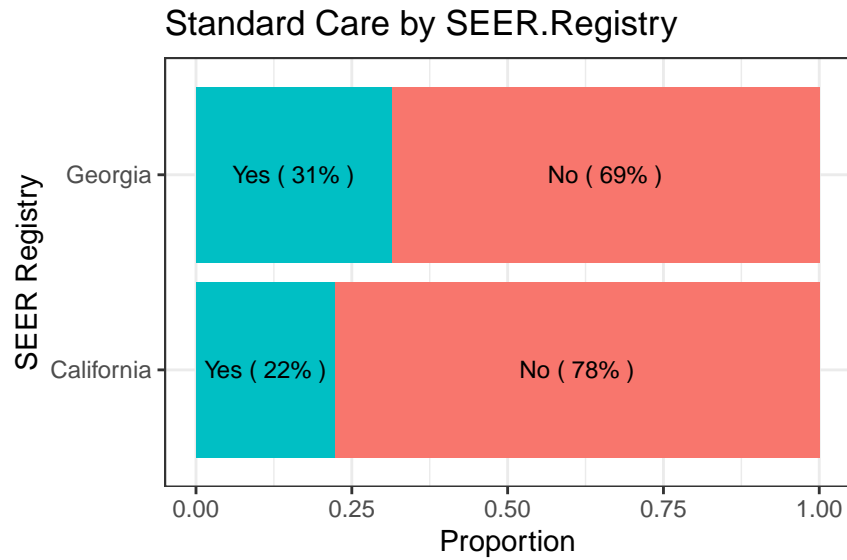
There are also differences between regions. Georgia has the highest rate of receiving standard treatment.

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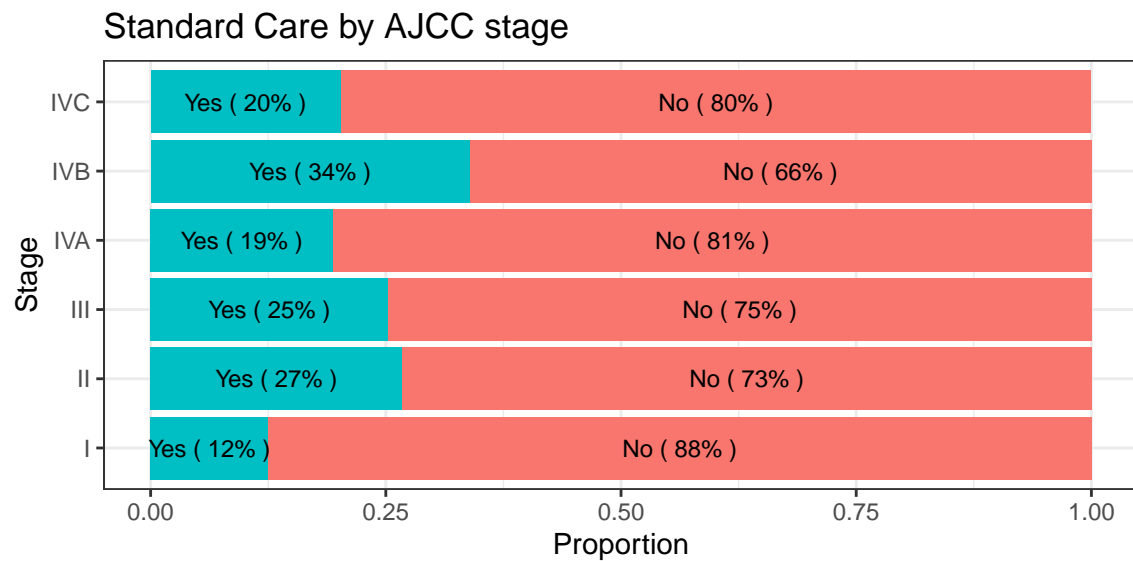
From the plot above, AJCC stage is also an important factor. Patients in stage II, III, IVB are more likely to be given standardized treatment.

Median household income distribution on patients NOT getting Standard Care

Median household income distribution on patients getting Standard Care

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

Education distribution patients NOT getting Standard Care

Education distribution on patients getting Standard Care

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 distribution on patients NOT getting Standard Care

Language isolation distribution on patients getting Standard Care
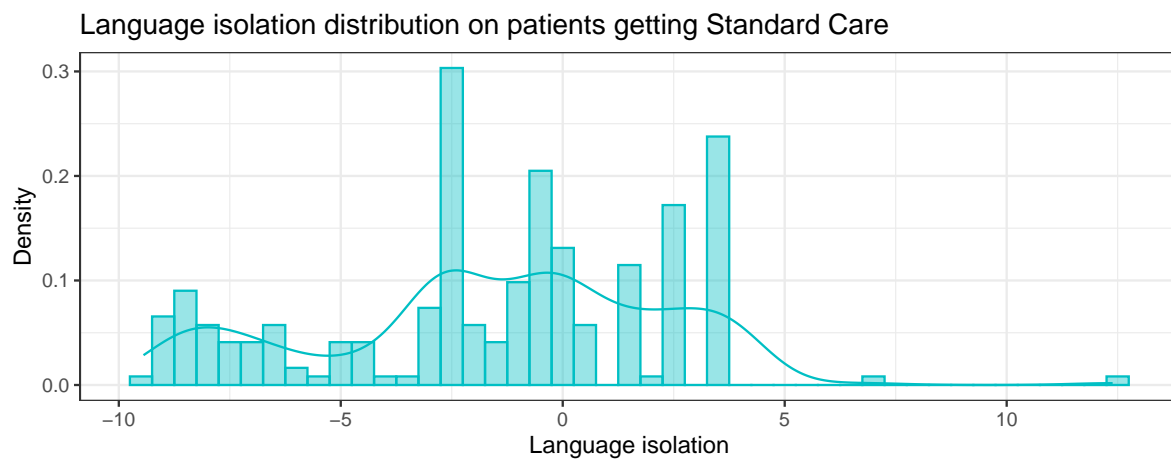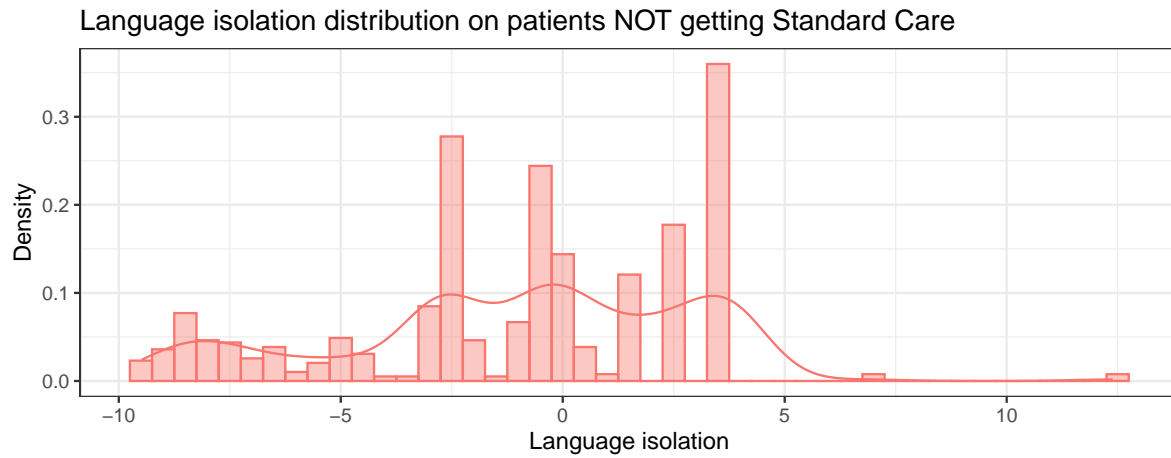
Language isolation has no significant impact on whether standardized treatment is given.

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.
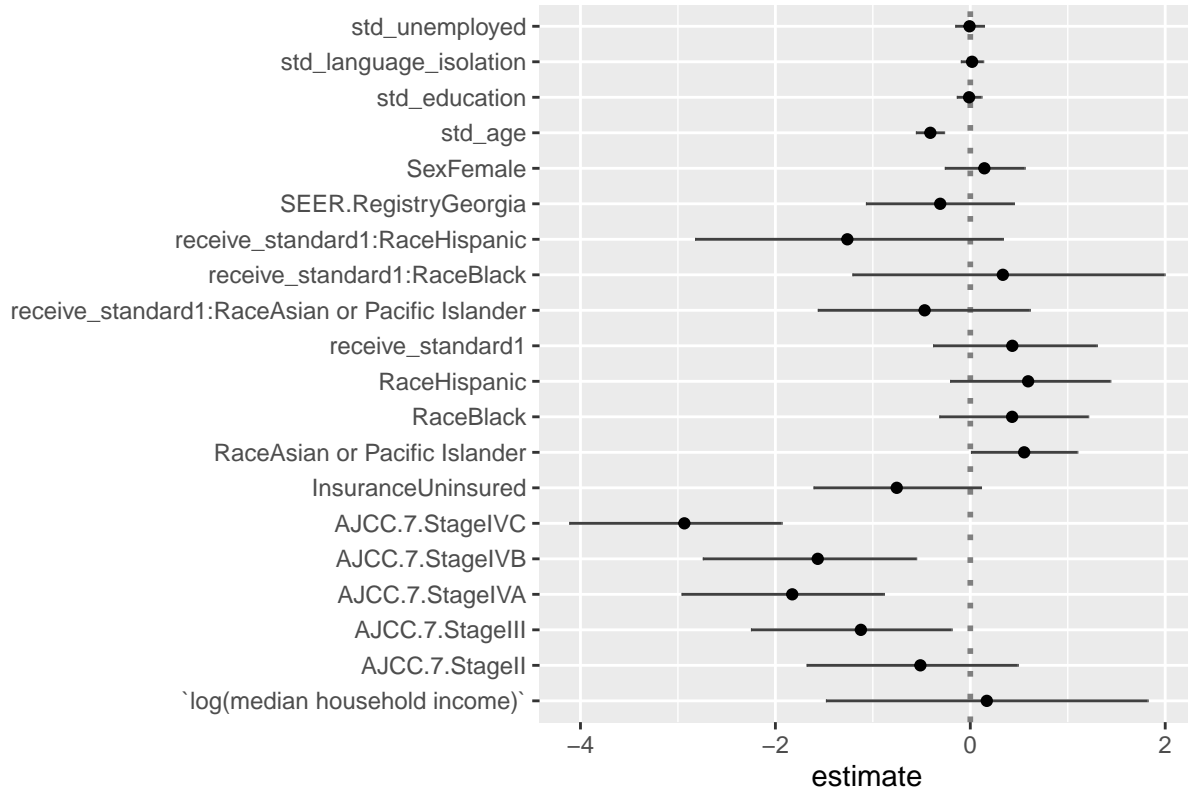

Coefficient Plot of Standard Treatment
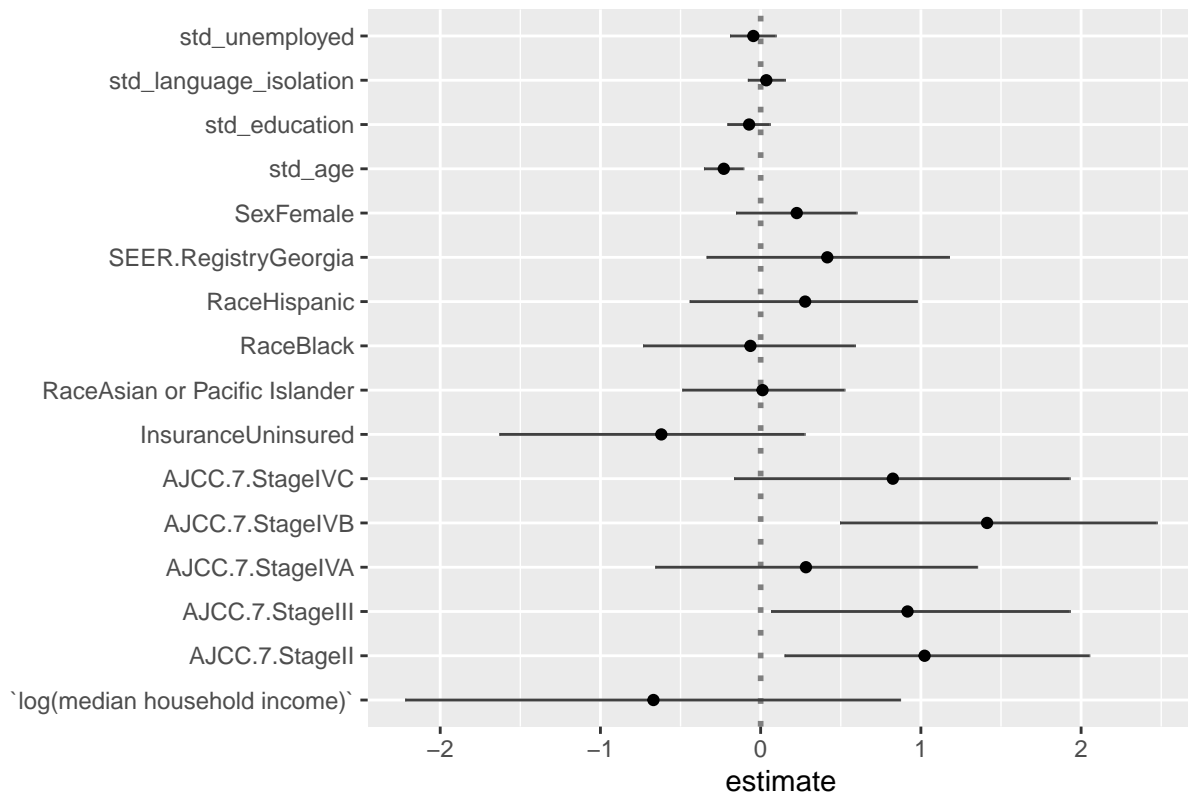
Survival Situation

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.

Coefficient Plot of Survival Situation

Discussion

Standard Therapy

Based on the coefficient plot above, we can find that the race, stage of cancer, the area where the patient is located, age at diagnosis, median household income and whether or not have insurance, impact whether the patient receives standard treatment that meets the NCCN guidelines. Compared with whites, Hispanic is more likely to get standard treatment. When a patient's cancer is in stage II, III, IVB or IVC, they are more probably to receive standard treatment. Uninsured patients will have difficulty getting standardized treatment. Besides, patients in Georgia are more likely to receive standard treatment compared with patients in California. The older the patient, the less likely it is to be given standardized treatment. The education, has negative impact on whether patients receive standard therapy. Unemployed and language isolation situation in the patients community have no impact on patients receive standard treatment or not.

Meanwhile, compared to patients in other stages, patients in the Stage IVA are less likely to receive standard therapy. We can find that people in this stage 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.

Survival Situation

Based on the coefficient plot of the survival situation, we can find that cancer stage and whether they have insurance have impact on the patient's survival status. The later the cancer stage is, the more likely it is that they will die of the cancer. After receiving standardized treatment, the survival situation becomes better.

Meanwhile, there is an interaction between whether patients receive standard therapy and race. Compared with white people who receive standard treatment, the survival situation of Hispanic and Asian or Pacific Islander who received standard treatment are worse, and the survival situation of blacks who received standard treatment are better.

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
## -2.5465   0.2049  0.4826   0.7088   1.6652
##
## Coefficients:
##                                               Estimate Std. Error z value
## (Intercept)                                   0.036084   9.092396   0.004
## receive_standard1                             0.431049   0.425832   1.012
## RaceHispanic                                  0.593386   0.417171   1.422
## RaceBlack                                     0.429258   0.388185   1.106
## RaceAsian or Pacific Islander                 0.552858   0.278064   1.988
## SexFemale                                     0.144399   0.208814   0.692
## AJCC.7.StageII                               -0.512180   0.542422  -0.944
## AJCC.7.StageIII                              -1.122878   0.515783  -2.177
## AJCC.7.StageIVA                              -1.827081   0.519906  -3.514
## AJCC.7.StageIVB                              -1.565271   0.549113  -2.851
## AJCC.7.StageIVC                              -2.933921   0.547302  -5.361
## std_age                                      -0.409909   0.073811  -5.554
## SEER.RegistryGeorgia                         -0.308810   0.386604  -0.799
## `log(median household income)`               0.169312   0.840262   0.201
## InsuranceUninsured                           -0.754876   0.435515  -1.733
## std_education                                -0.012358   0.064567  -0.191
## std_unemployed                               -0.007889   0.074805  -0.105
## std_language_isolation                        0.018215   0.058350   0.312
## receive_standard1:RaceHispanic               -1.261538   0.798975  -1.579
## receive_standard1:RaceBlack                   0.333937   0.807389   0.414
## receive_standard1:RaceAsian or Pacific Islander -0.468363 0.552426  -0.848
##                                               Pr(>|z|)
## (Intercept)                                   0.996834
## receive_standard1                             0.311418
## RaceHispanic                                  0.154909
## RaceBlack                                     0.268810
## RaceAsian or Pacific Islander                 0.046785 *
## SexFemale                                     0.489239
## AJCC.7.StageII                               0.345045
## AJCC.7.StageIII                               0.029478 *
## AJCC.7.StageIVA                               0.000441 ***
## AJCC.7.StageIVB                               0.004364 **
## AJCC.7.StageIVC                               8.29e-08 ***
## std_age                                       2.80e-08 ***
## SEER.RegistryGeorgia                          0.424421
## `log(median household income)`               0.840308
## InsuranceUninsured                            0.083043 .
## std_education                                 0.848218
## std_unemployed                                0.916009
## std_language_isolation                        0.754912
## receive_standard1:RaceHispanic               0.114349
## receive_standard1:RaceBlack                   0.679166
## receive_standard1:RaceAsian or Pacific Islander 0.396533
```

```
## ---
## 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: 673.37  on 694  degrees of freedom
## AIC: 715.37
##
## 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.
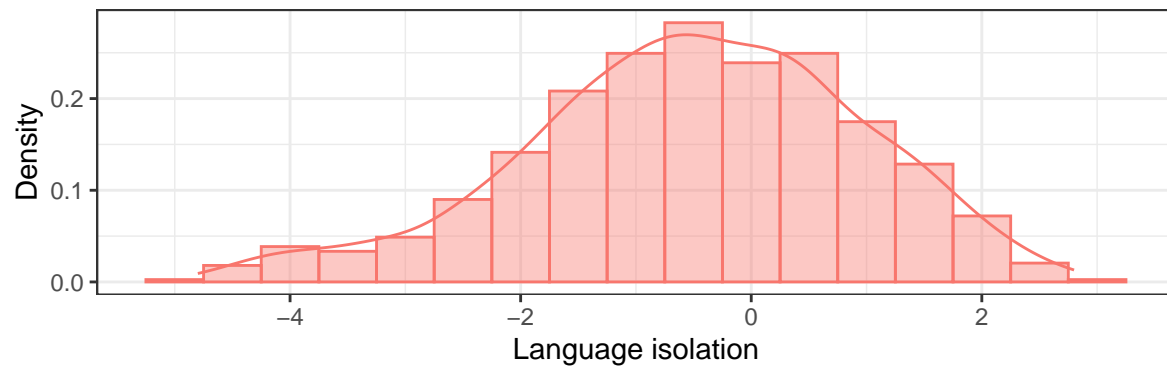
```
## $calibration_plot
```
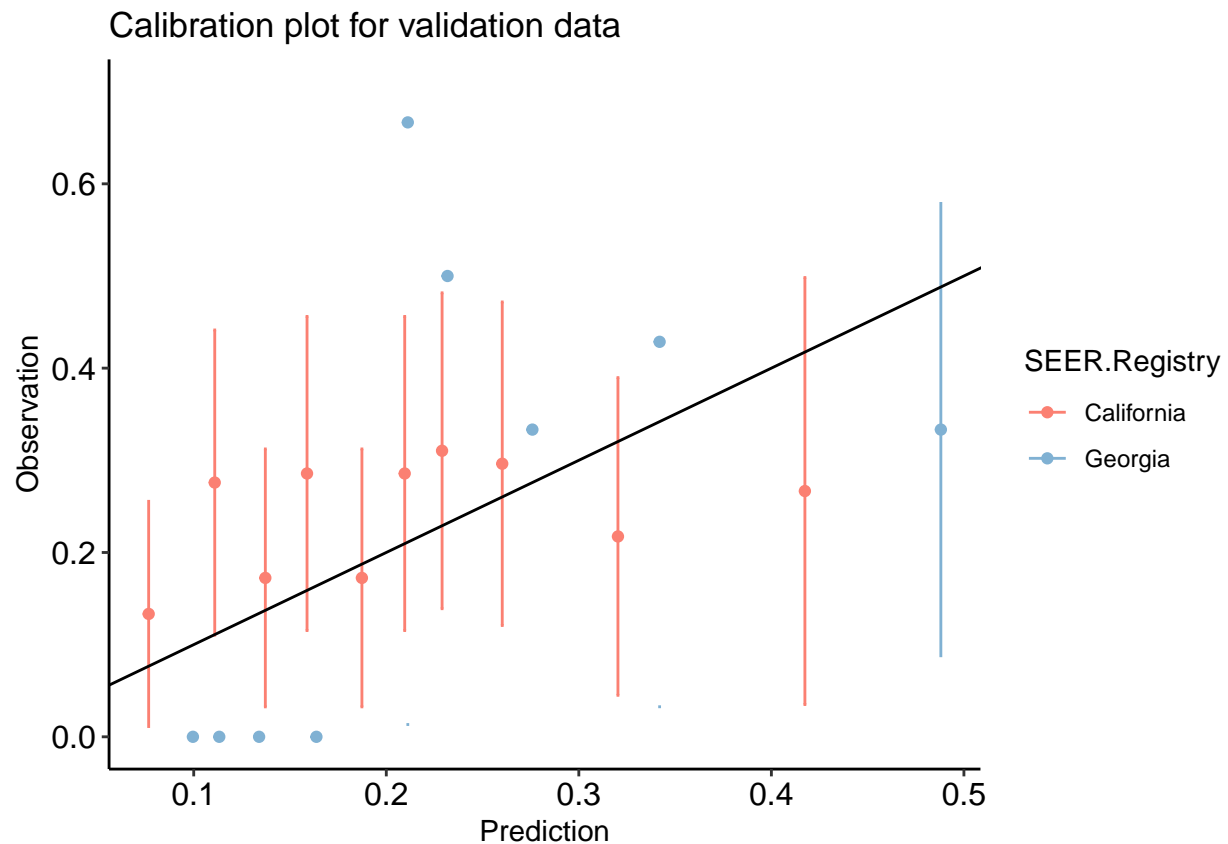


```
## The c-stat score here is 0.6610979
```

EDA

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

Language isolation distribution on patients NOT getting Standard Care

Language isolation distribution on patients getting Standard Care

Standard Care by Sex

Standard Care by Insurance