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

Runqi Zhao, Masanao Yajima

12/5/2021

Abstract

In this study, we explore oropharynx cancer patient population from the SEER database to investigate whether there is disparities in receiving standard treatment defined by NCCN Guidelines. We find that even after accounting for patients' AJCC Stages, age and whether they have an insurance, we see gender, race, and income Level will affect whether the given therapy follows guidelines. Since standard treatments have significant influence on the survival of a patient, this result suggests concerning systematic issues that require further investigation.

Introduction

[Talk about the disparty of patient treatment elsewhere.] [_____] Please add more details. The Surveillance, Epidemiology, and End Results (SEER) Program is an authoritative source for cancer statistics in the United States. The SEER registries collect data on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment, and they follow up with patients for vital status.

[_____] Please add more details about oropharynx cancer.

The oropharynx includes the base of the tongue, tonsils, soft palate, and posterior pharyngeal wall. Here we combined three sites together for analysis.

Oropharyngeal cancer that is p16-positive (ie, HPV-mediated) is different disease than p16-negative cancer and the suggested therapies are different for AJCC Stage III when T Stage is 1 and N Stage is 1.

Data Processing

This SEER data set cited from Anand Devaiah, Pratima Agarwal and Jacob Bloom contains cancer care statistics for people in 4 different states with 7 different head and neck cancer types. We have the cancer type, the stage, the patient's social background, socioeconomic status, and their types of treatments.

This report focus on Oropharynx cancer that constitute of three sub-sites: Tongue, Tonsil, Oropharynx. AJCC Stage is one of the most important index to identify the standard therapy. In this step, we refer to CS Extension code to assign AJCC Stage for patients without AJCC Stage information. According the NCCN Clinical Practice Guidelines(Version 1.2020), we defined the standard therapy for each stage. Processing details are attached in appendix.

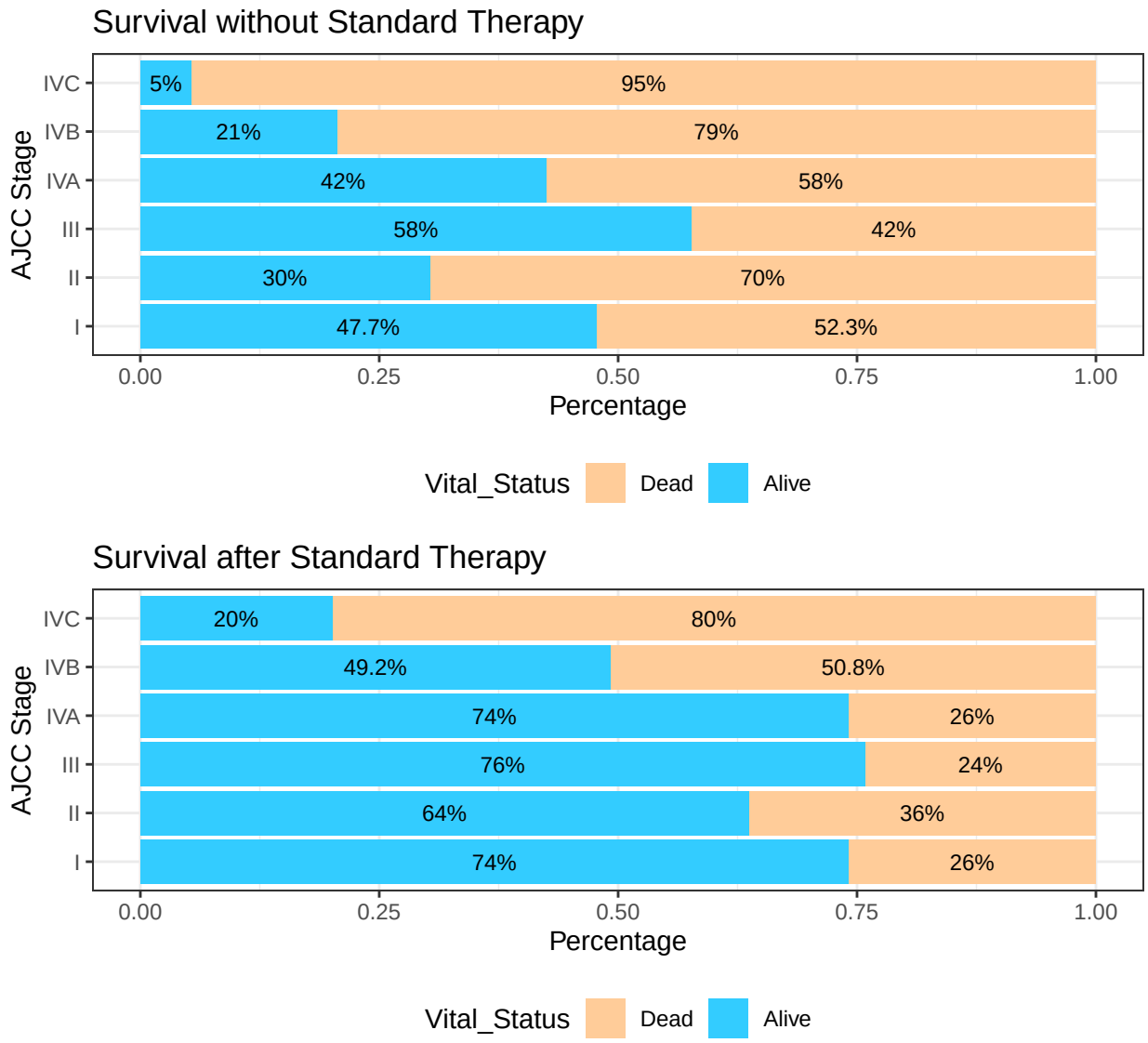
In this data set, approximate 12.2% patients don't receive standard therapy.

Exploratory data analysis

In EDA part, we looked at the AJCC Stage, Gender, Race, Region, Insurance and Age. Here we only look at the AJCC Stage, Race and Region. Other EDA are attached in appendix.

AJCC Stage

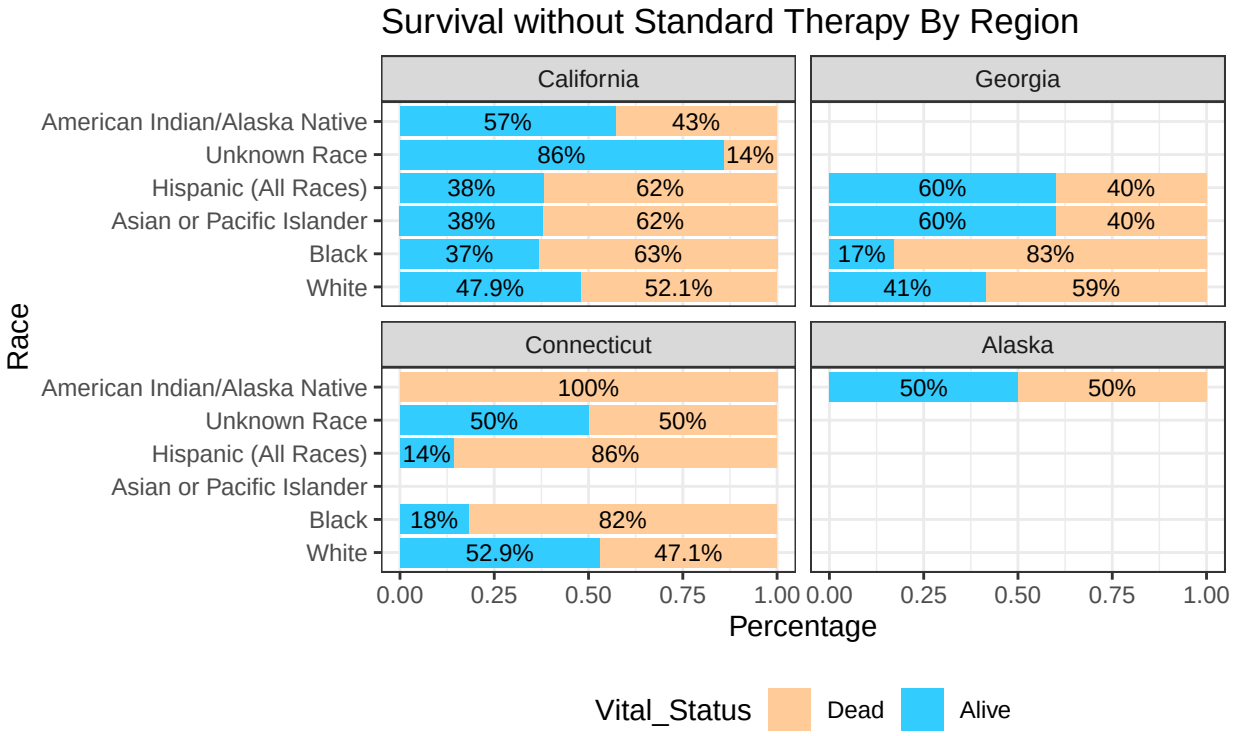
By looking at the AJCC Stage, we can find that survival rate could be highly affected by standard therapy, and shows difference between stages.

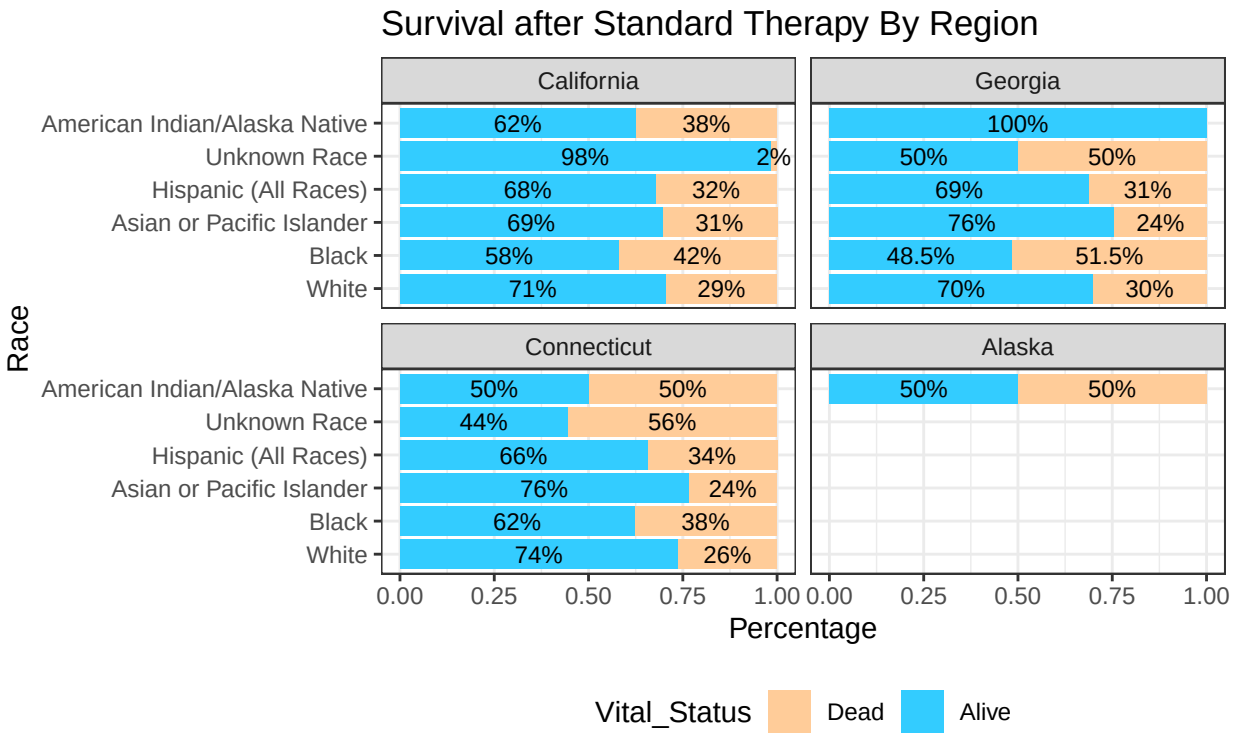


Race and Region

The data used in this report contains 4 different states and 5 races. However, the sample number is highly unbalanced between states and races. Most of the patients are White people, and high percent of them come from California.

When we look at the survival rate, we can find much difference between four states, but with consistence that Black people get lower survival rate.


Survival without Standard Therapy By Region
California
Georgia
Connecticut
Alaska
American Indian/Alaska Native
Unknown Race
Hispanic (All Races)
Asian or Pacific Islander
Black
White
57%
43%
86%
14%
38%
62%
38%
62%
37%
63%
47.9%
52.1%
60%
40%
60%
40%
17%
83%
41%
59%
100%
50%
50%
14%
86%
18%
82%
52.9%
47.1%
50%
50%
0.00
0.25
0.50
0.75
1.00
Race
Percentage
Vital_Status
Dead
Alive

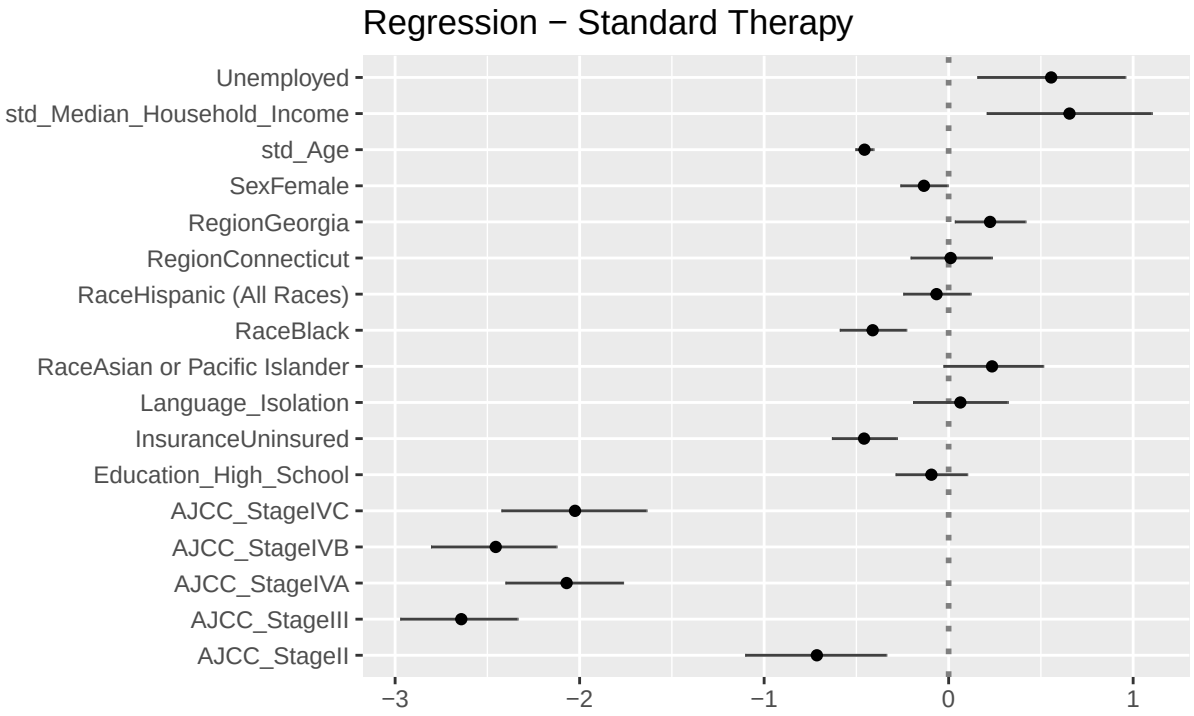

Survival after Standard Therapy By Region
California
Georgia
Connecticut
Alaska
American Indian/Alaska Native
Unknown Race
Hispanic (All Races)
Asian or Pacific Islander
Black
White
62%
38%
98%
2%
68%
32%
69%
31%
58%
42%
71%
29%
100%
50%
50%
69%
31%
76%
24%
48.5%
51.5%
70%
30%
50%
50%
44%
56%
66%
34%
76%
24%
62%
38%
74%
26%
50%
50%
0.00
0.25
0.50
0.75
1.00
Race
Percentage
Vital_Status
Dead
Alive

Regression

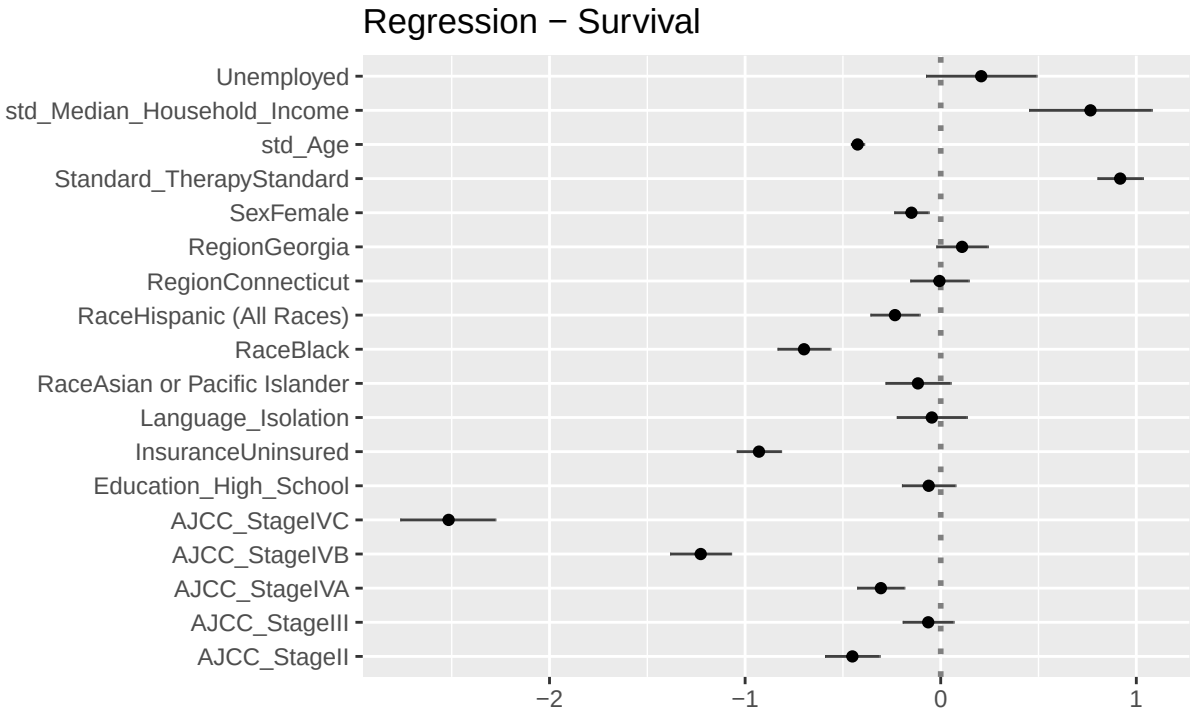
Based on the EDA, we removed Race as Unknown Race and American Indian/Alaska Native, Region as Alaska.

We fit logistic regression for the Standard Therapy and Survival Status, with baseline set at Insured white male with age of 60 from California, take the AJCC Stage, Gender, States, Race, Age, Insurance, Percentage of receiving High School Education, Percentage of Unemployed, Income situation, Language Isolation and Standard Therapy(for SurvivaL Status) as our predictors. The Percentage of receiving High School Education, Unemployed and Language Isolation are at 10% scale, Income is at log scale.

The regression results show that Black People have less chance to get standard therapy, compared to White people, while Asian or Pacific Islander have larger chance. Higher income level indicates higher probability to take standard therapy. AJCC Stages also have a large impact on therapy. Compared with Stage I, other stages have less chance to get standard therapy, especially for stage III and IV.



Standard therapy suggestion has large impact on survival results of patients. In this report, the standard therapy we used for analysis is based on suggestion, however, patients might refuse suggested therapies. This means, compared with other therapies, standard suggestion gives patients larger survival probability. Gender shows some difference here. Compared with male, female tell lower chance to survive. Black People also show lower survival probability than White people. Income level gives positive impact on survival rate. Higher income level shows higher survival chance. Compared with insuranced people, patients without insurance has less chance to survive. AJCC Stages show some regular here. Except stage II, the higher the stage is, the lower the survival probability will be.



Conclusion and Discussion

From EDA and regression results, we think that for oropharynx, discrimination exist in Gender, Race, Insurance and Income. Black people, female and uninsured patients show significant lower chance of taking standard therapy and getting survived, while higher income level shows significant higher chance of taking standard therapy and getting survived.

However, we have a lot of limitations come with the data and for the analysis. Oropharyngeal cancer require p16 index to tell the standard therapy at Stage III. Without these information, the only choice we have is to exclude these samples. This exclusion will add bias to our results.

The interaction between states and some predictors, like race and income, might give better explanation to the impact source. But given the missing factors in some states, we can not dig into the interactions. In addition, the unbalanced sample size also add some uncertainty to our results, since most of the samples are collects from California. Its doubtful that whether California has a good representative for all four states.

Appendix

AJCC Stages - TMN Stages

By removing the Blank(s) in both AJCC Stage and Tumor size, we get 20157 observations, with 8637 Blanks in AJCC Stage.

According CS code, we try to determine the TNM Stage for each patient. By this step, we aim to:

(1) Involve the patients with blanks in AJCC Stage information.

(2) Exclude the patients at Stage III while T Stage is 1 and N Stage is 1. (We can not tell whether the patient was given a standard therapy because we don't have p16-test result)

The dataset provides lymph nodes code but no nodes size, we can not distinguish N0-N3, we can only tell `N0`: No Nodal Involvement, `NX` or `Other`: Nodal Involvement (N1-N3).

According to the mets code we determine the M Stage as:
`M0`: No distant metastasis and,
`M1`: Distant metastasis .

With limited lymph nodes information, the AJCC Stage is difficult to tell between III and IV for some patients. After assign AJCC Stage for the patients, we exclude the Unknown Stages, IVNOS and unsure stages. Now we have 13696 observations for further analysis.

Standard Therapy

`Radiation`: "None/Unknown" or "Recommended" (All other recording besides None/Unknown) `Surgery`: "Not recommended" or "Recommended" (All other recording besides Not recommended or Not recommended, contraindicated due to other cond; autopsy only (1973-2002)) `Chemotherapy`: "No/Unknown" or "Yes"

According to NCCN Clinical Practice Guidelines(Version 1.2020), we defined the standard therapy for each stage:

- Stage I: Either Radiation or Surgery
- Stage II: Either Radiation or Surgery
- Stage III:
  `T1`: (removed, need hpv test information)
  `T2`: Radiation and Chemotherapy
  `T3`: Radiation and Chemotherapy, or Surgery
- Stage IV: Radiation and Chemotherapy, or Surgery

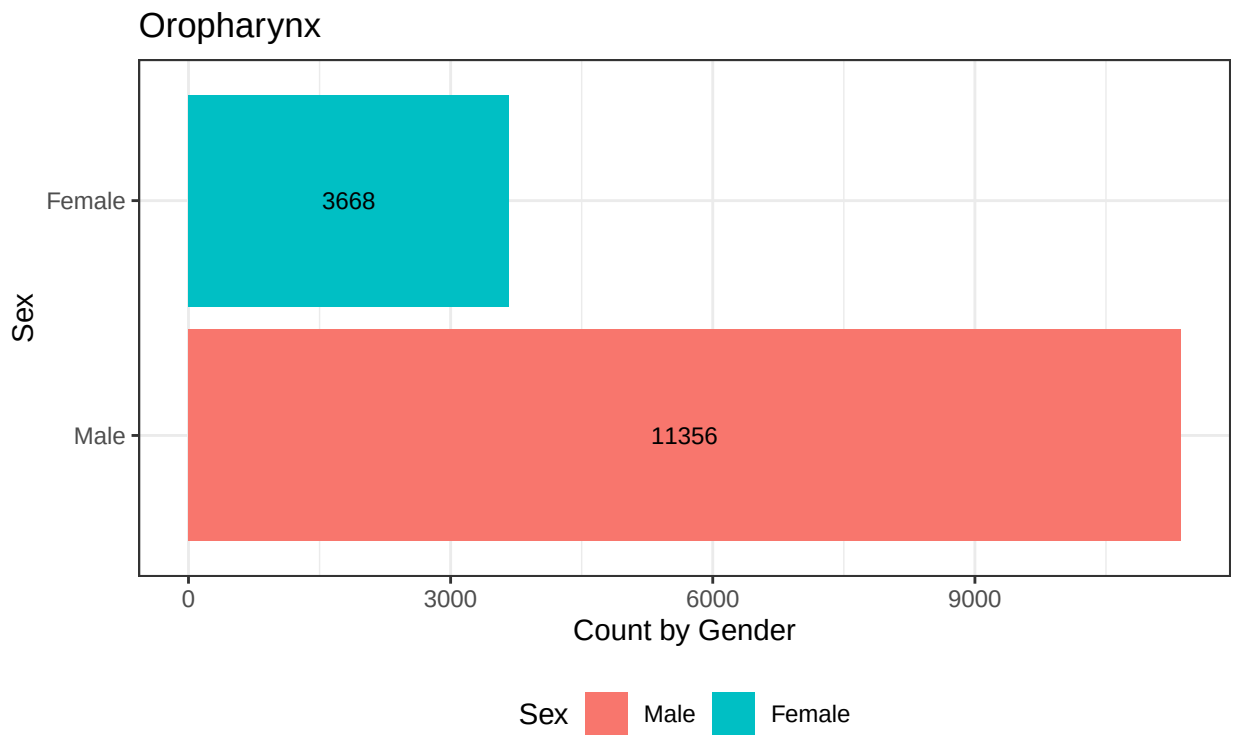
This Standard therapy is based on suggestion. (Patients might refuse some therapies)

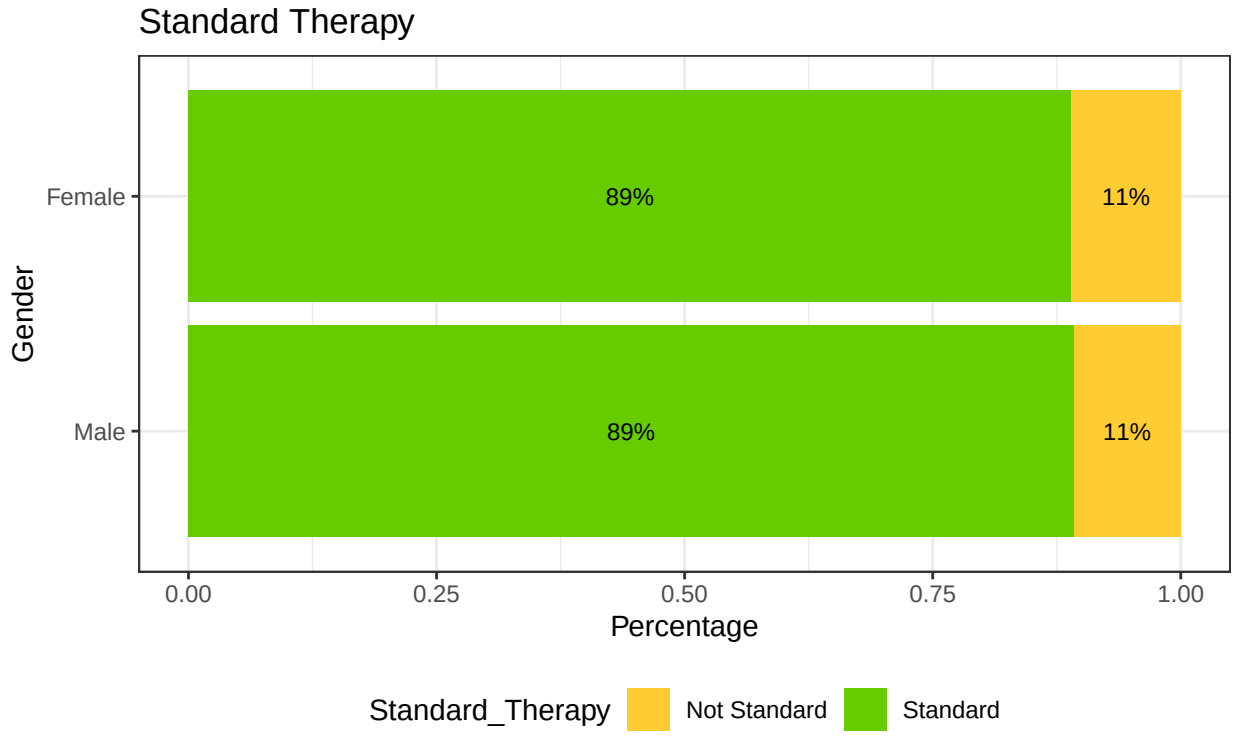
The standard therapy is a binary variable where `0` indicates `Not Get Standard Therapy` while `1` indicates the patient `Get Standard Therapy`.

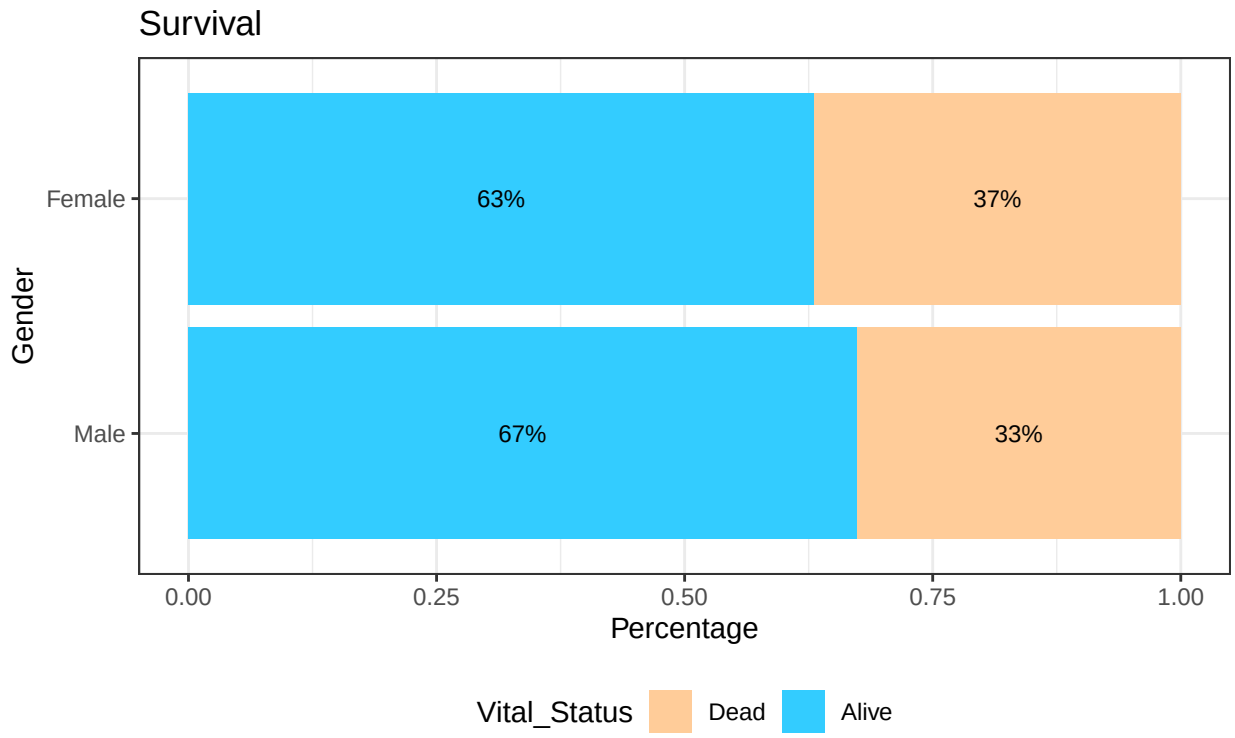
EDA

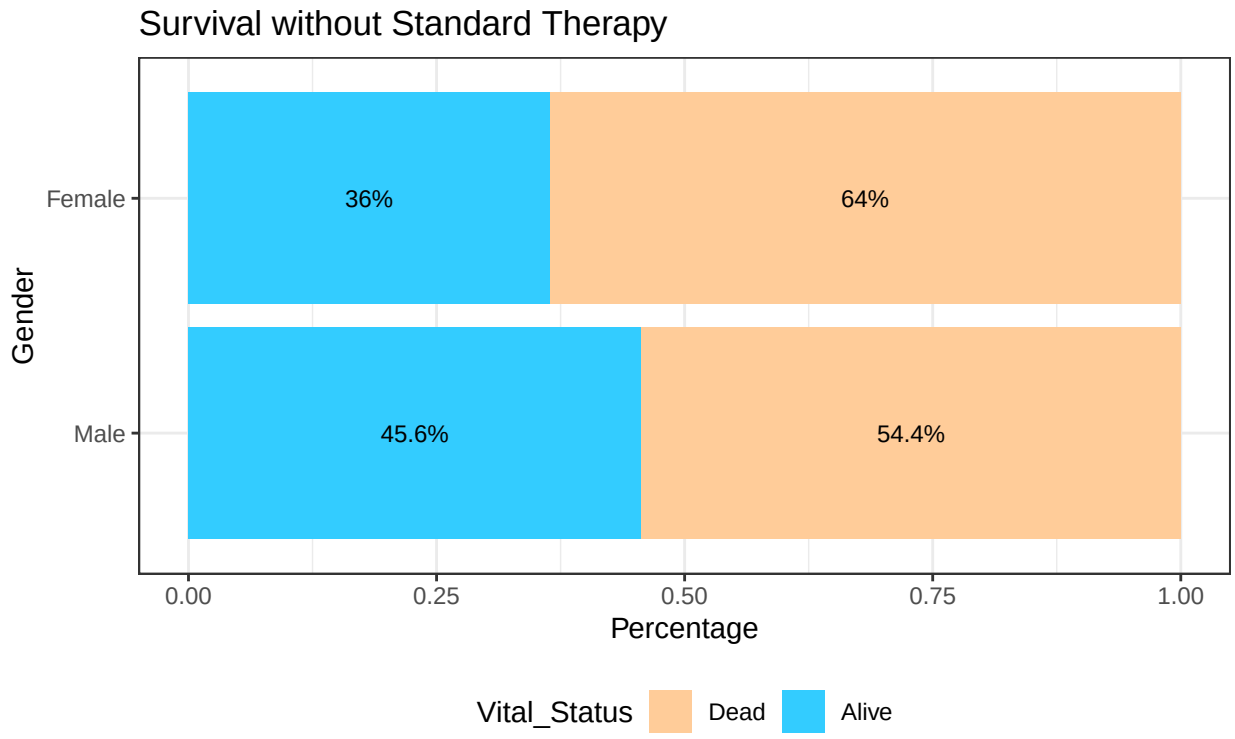
In this part I will include more EDA plots.

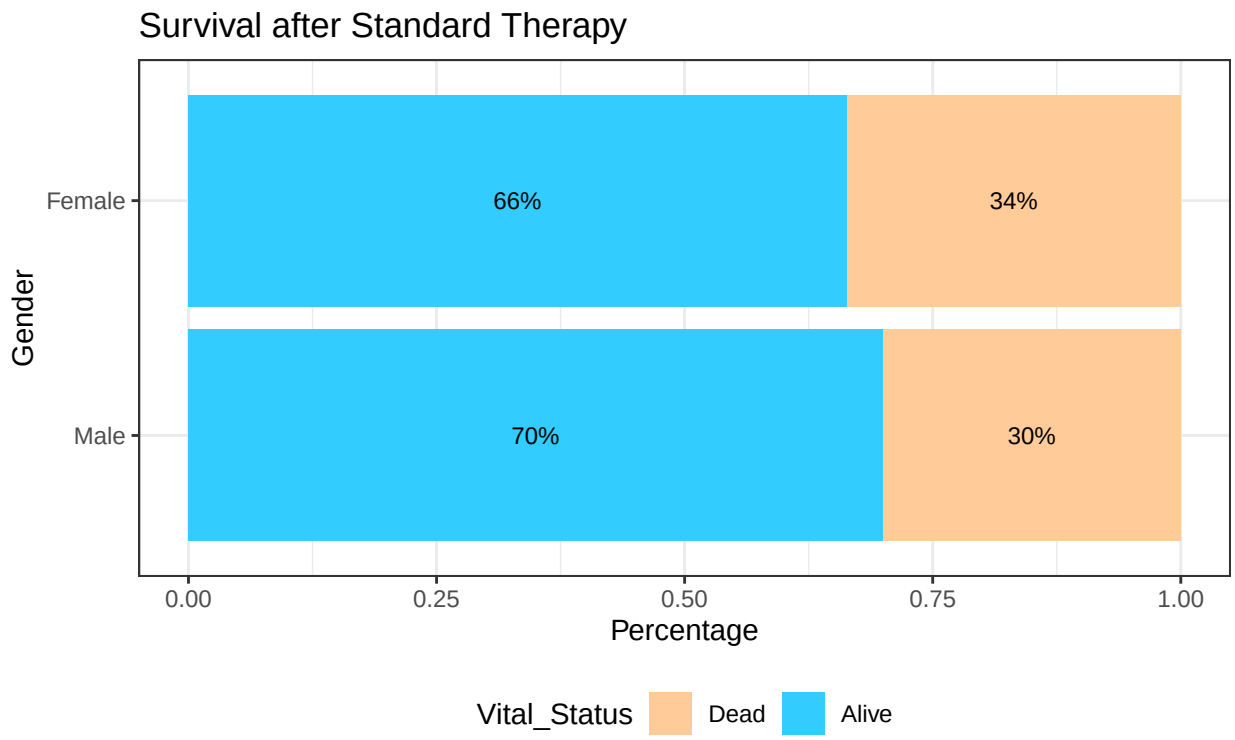
- **Gender**


Oropharynx
Female
Male
Sex
3668
11356
0
3000
6000
9000
Count by Gender
Sex
Male
Female


Standard Therapy
Female
Male
Gender
89%
11%
89%
11%
0.00
0.25
0.50
0.75
1.00
Percentage
Standard_Therapy
Not Standard
Standard

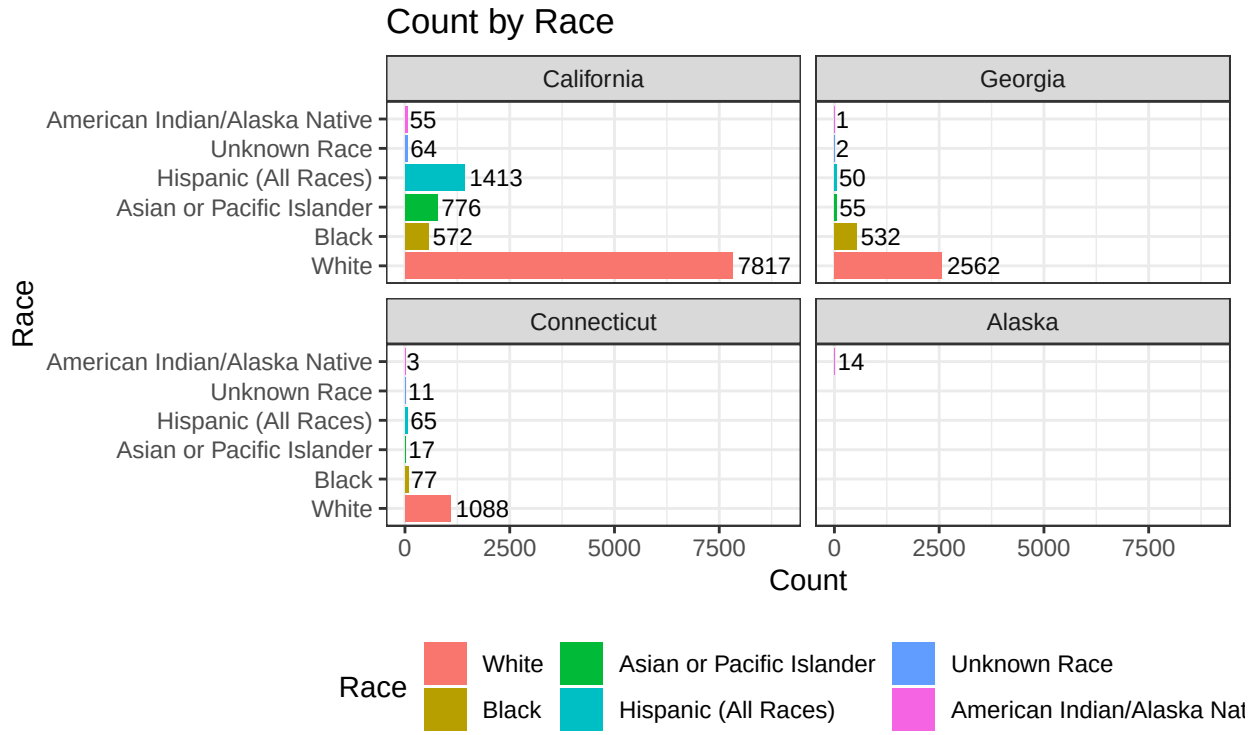

Survival
Female
Male
Gender
63%
37%
67%
33%
0.00
0.25
0.50
0.75
1.00
Percentage
Vital_Status
Dead
Alive

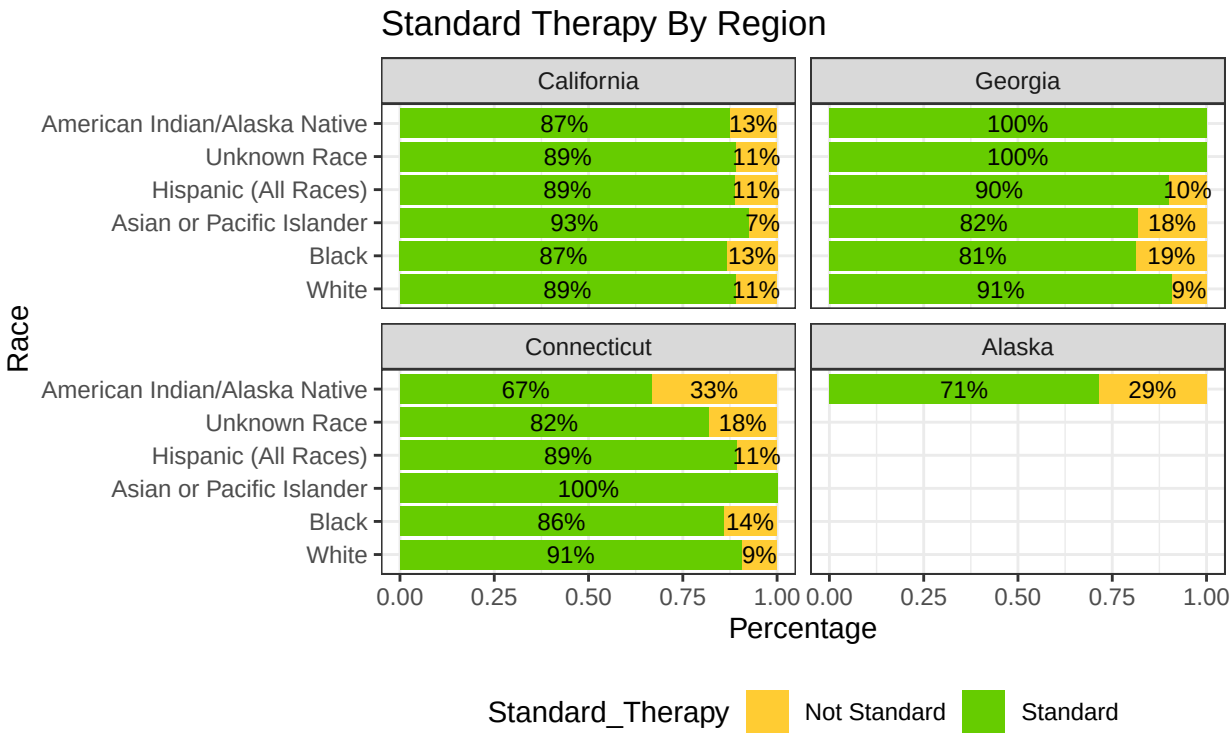

Survival without Standard Therapy
Female
Male
Gender
36%
64%
45.6%
54.4%
0.00
0.25
0.50
0.75
1.00
Percentage
Vital_Status
Dead
Alive

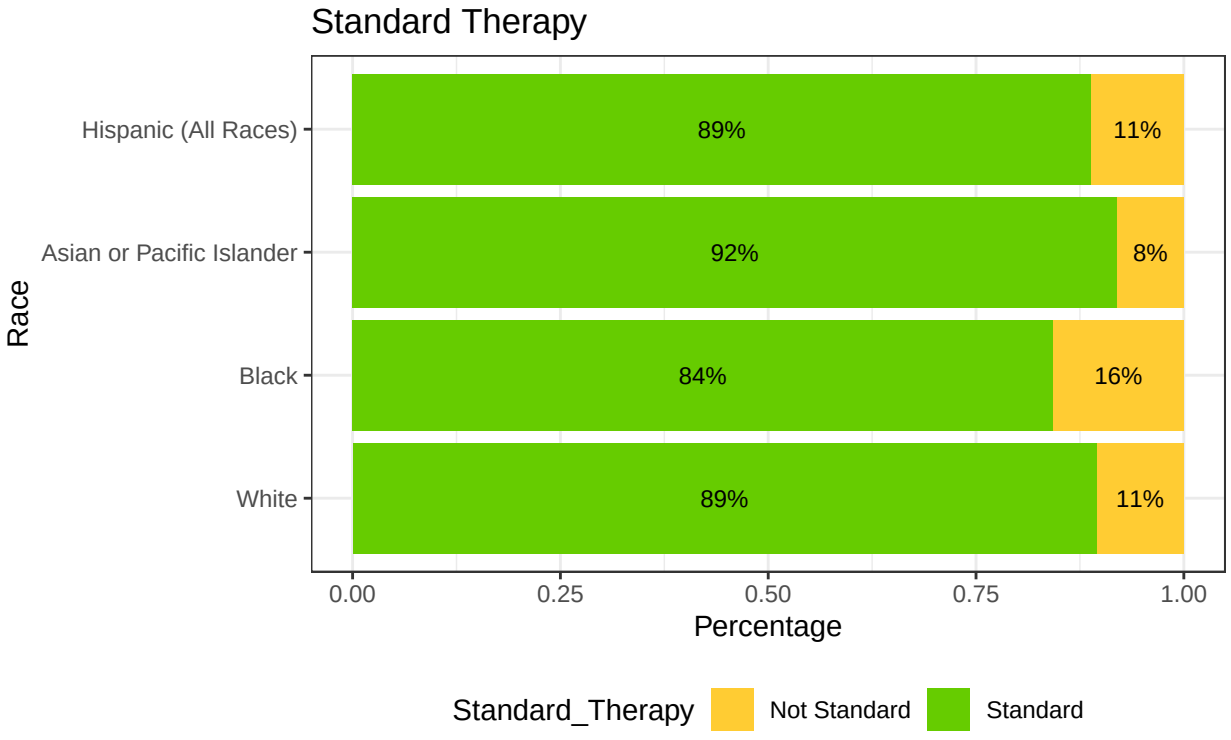


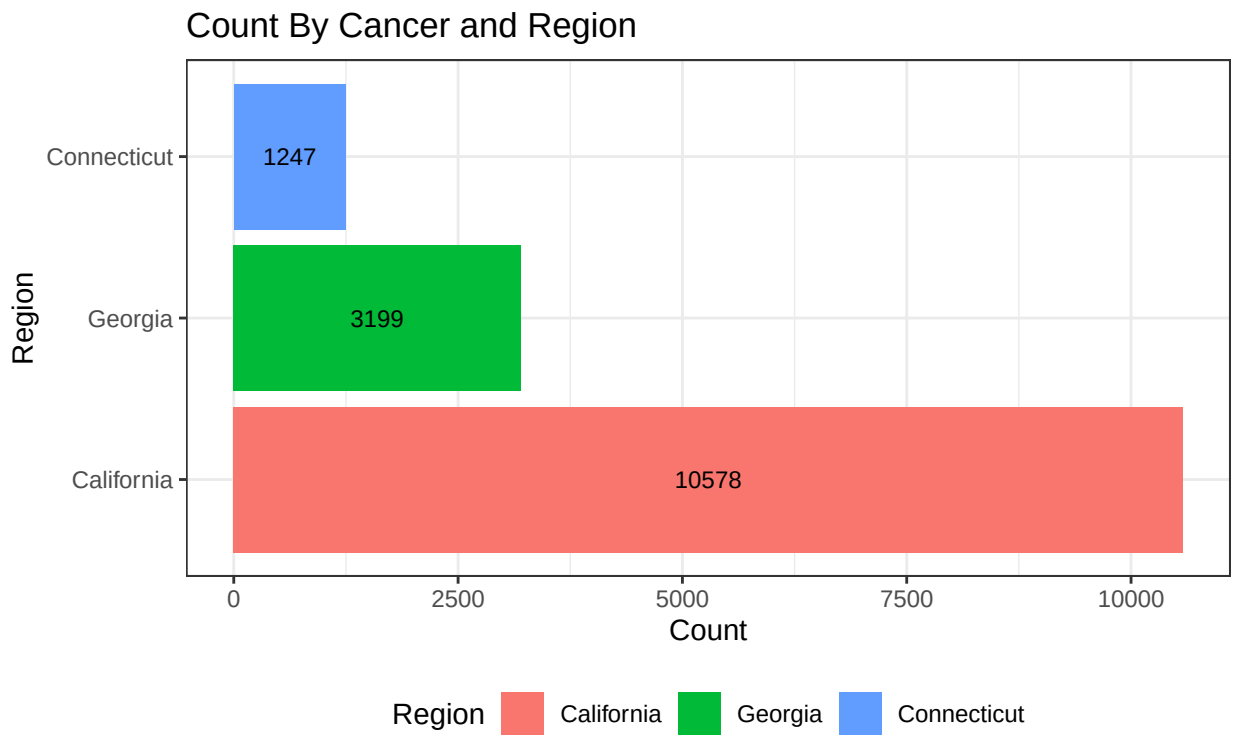
- **Race and Region**

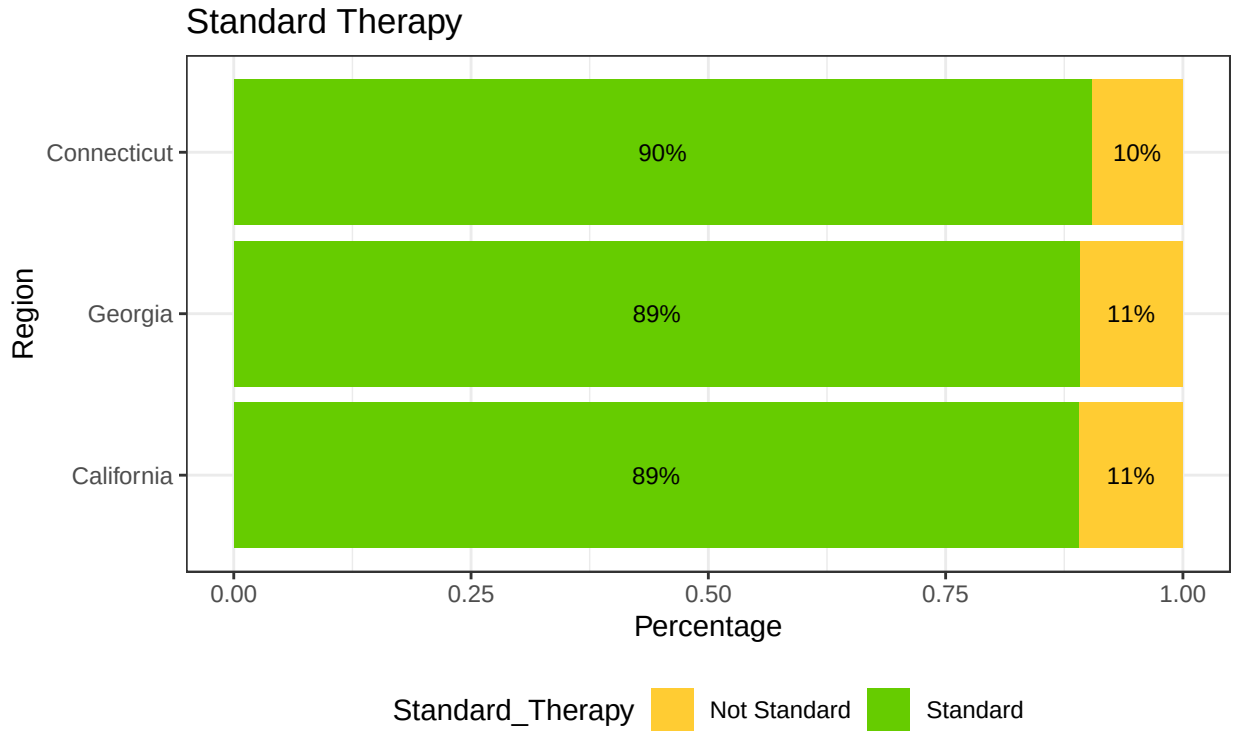
The rates of giving standard therapy are similar between states for White, Black and Hispanic people, but much different for American Indian/Alaska Native and Asian or Pacific Islander.



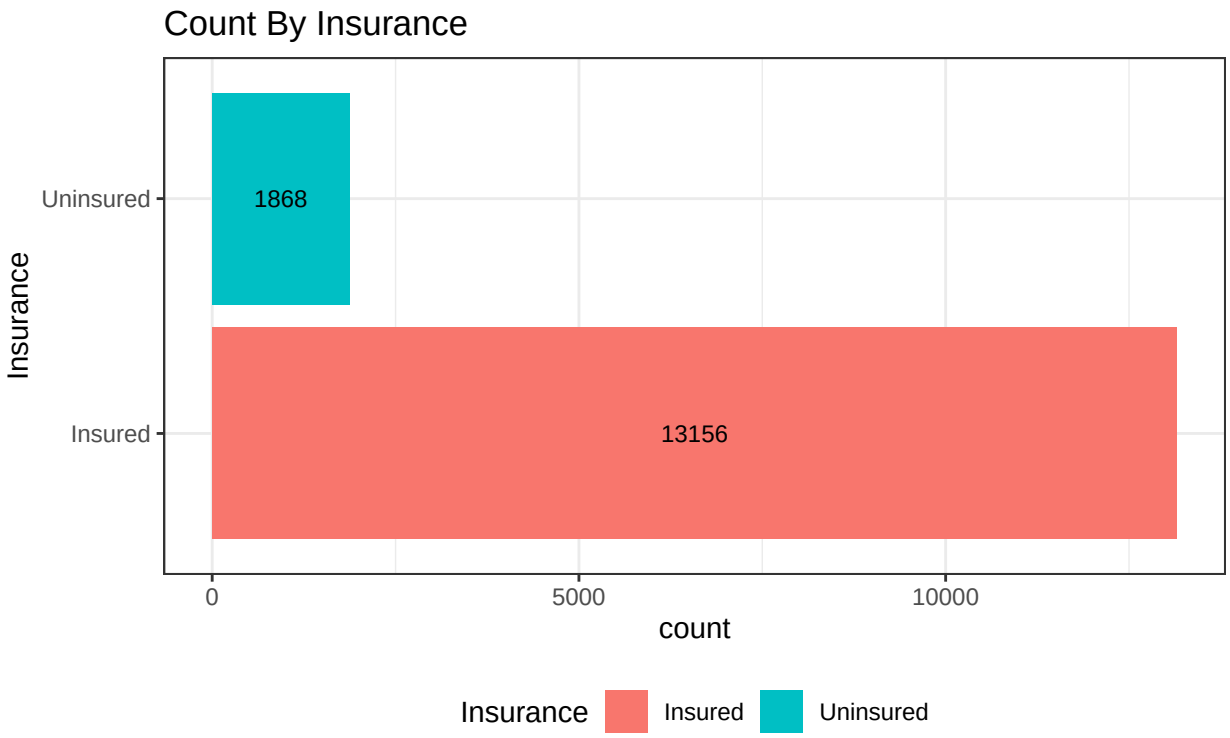

Standard Therapy By Region
California
Georgia
Connecticut
Alaska
American Indian/Alaska Native
Unknown Race
Hispanic (All Races)
Asian or Pacific Islander
Black
White
87%
13%
89%
11%
89%
11%
93%
7%
87%
13%
89%
11%
100%
100%
90%
10%
82%
18%
81%
19%
91%
9%
67%
33%
82%
18%
89%
11%
100%
86%
14%
91%
9%
71%
29%
Race
0.00
0.25
0.50
0.75
1.00
Percentage
Standard_Therapy
Not Standard
Standard

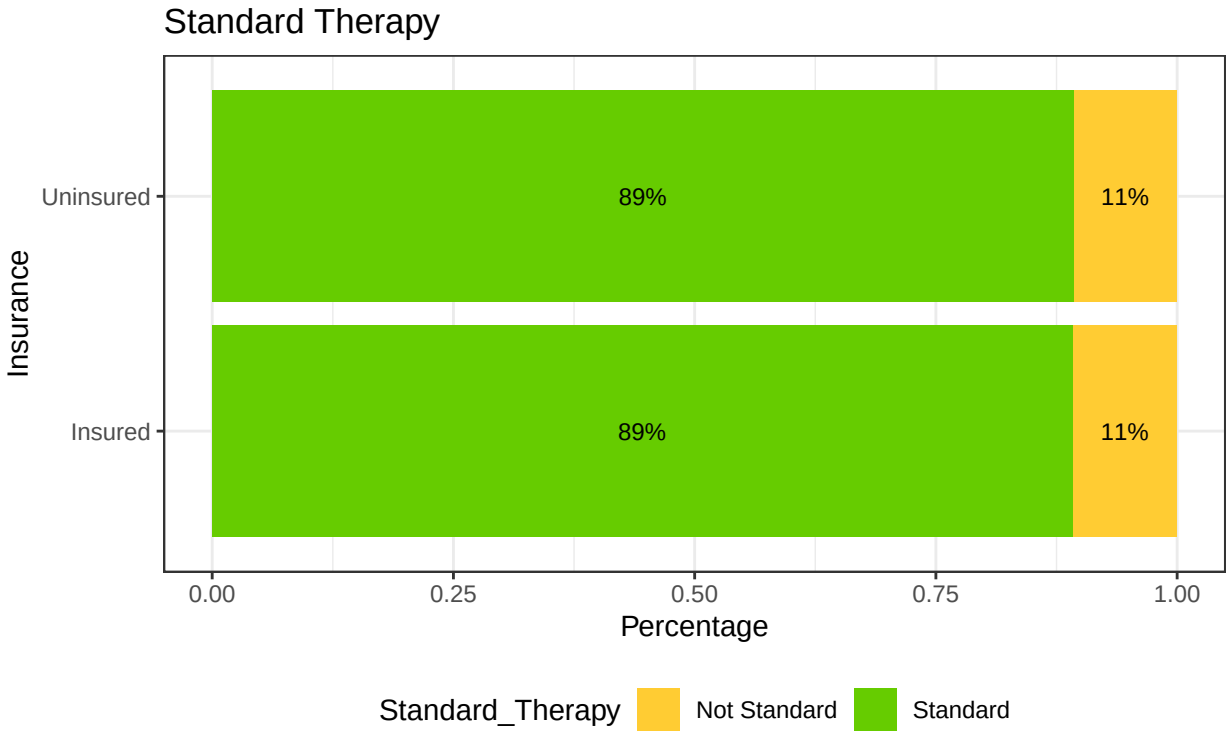

Standard Therapy
Hispanic (All Races)
Asian or Pacific Islander
Black
White
89%
11%
92%
8%
84%
16%
89%
11%
Race
0.00
0.25
0.50
0.75
1.00
Percentage
Standard_Therapy
Not Standard
Standard

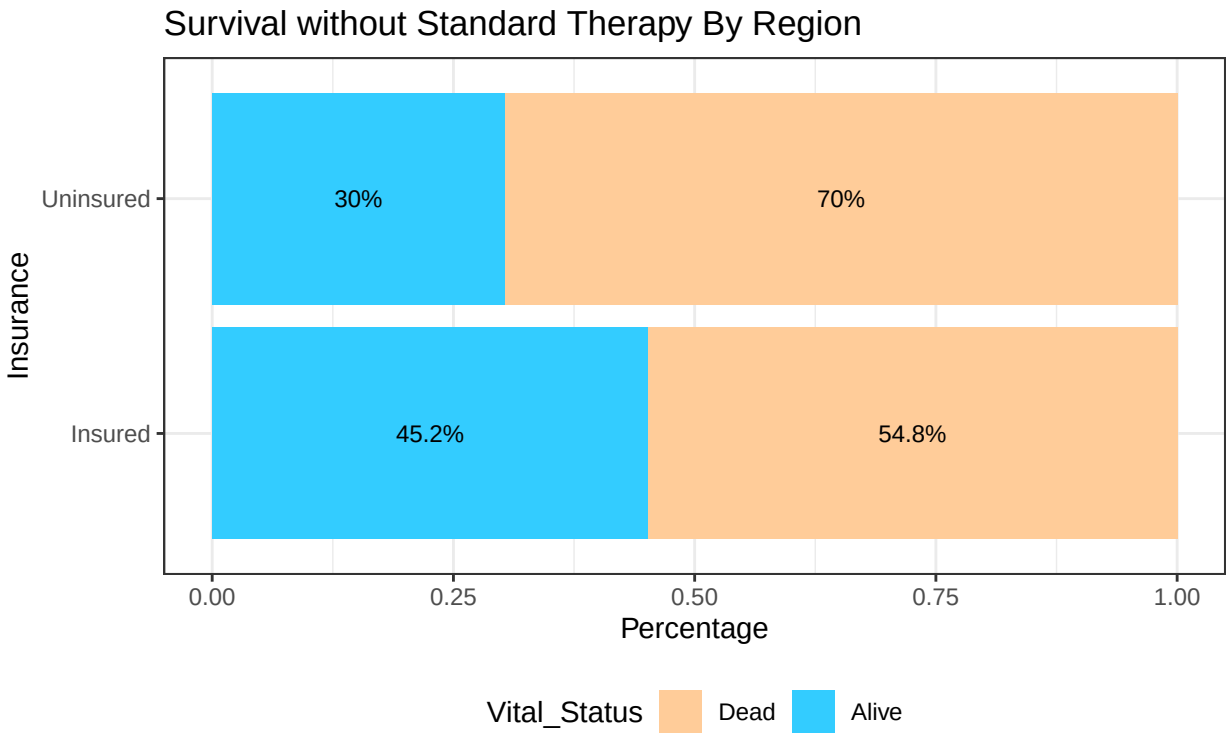

Count By Cancer and Region
Connecticut
Georgia
California
Region
1247
3199
10578
0
2500
5000
7500
10000
Count
Region
California
Georgia
Connecticut

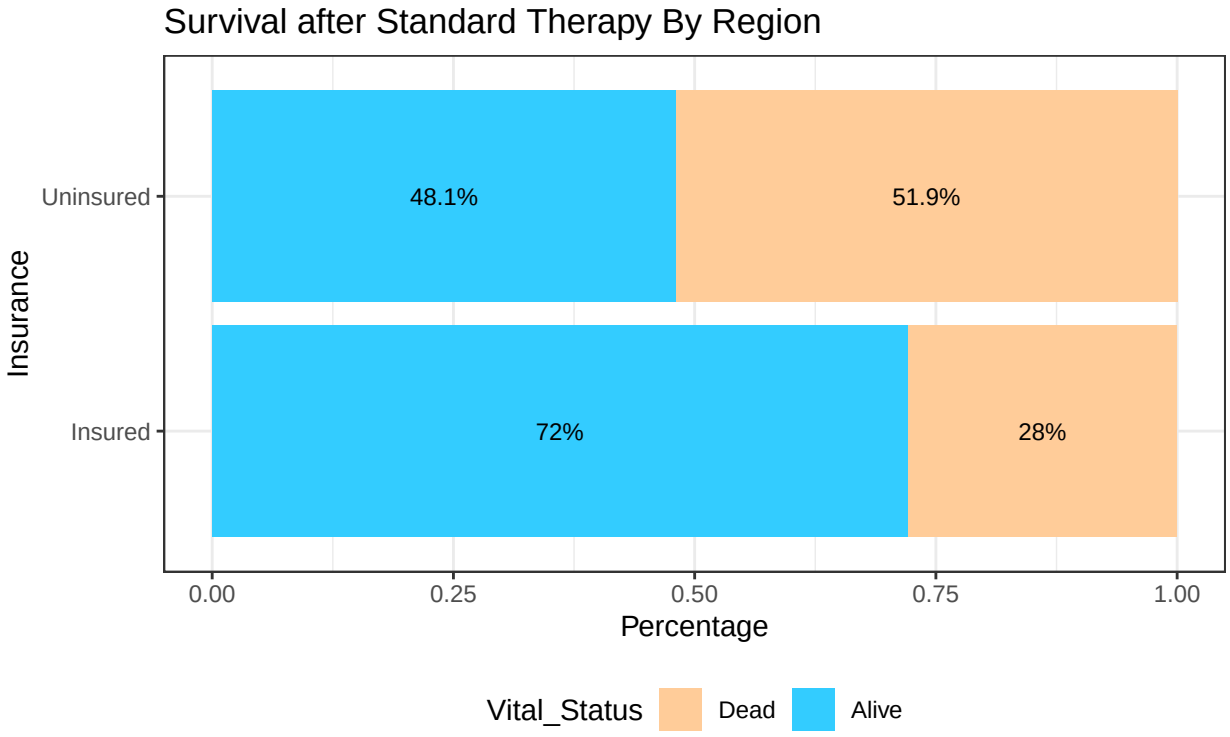

Standard Therapy
Connecticut
Georgia
California
Region
90%
10%
89%
11%
89%
11%
0.00
0.25
0.50
0.75
1.00
Percentage
Standard_Therapy
Not Standard
Standard

- **Insurance**


Count By Insurance
Uninsured
Insured
Insurance
1868
13156
0
5000
10000
count
Insurance
Insured
Uninsured

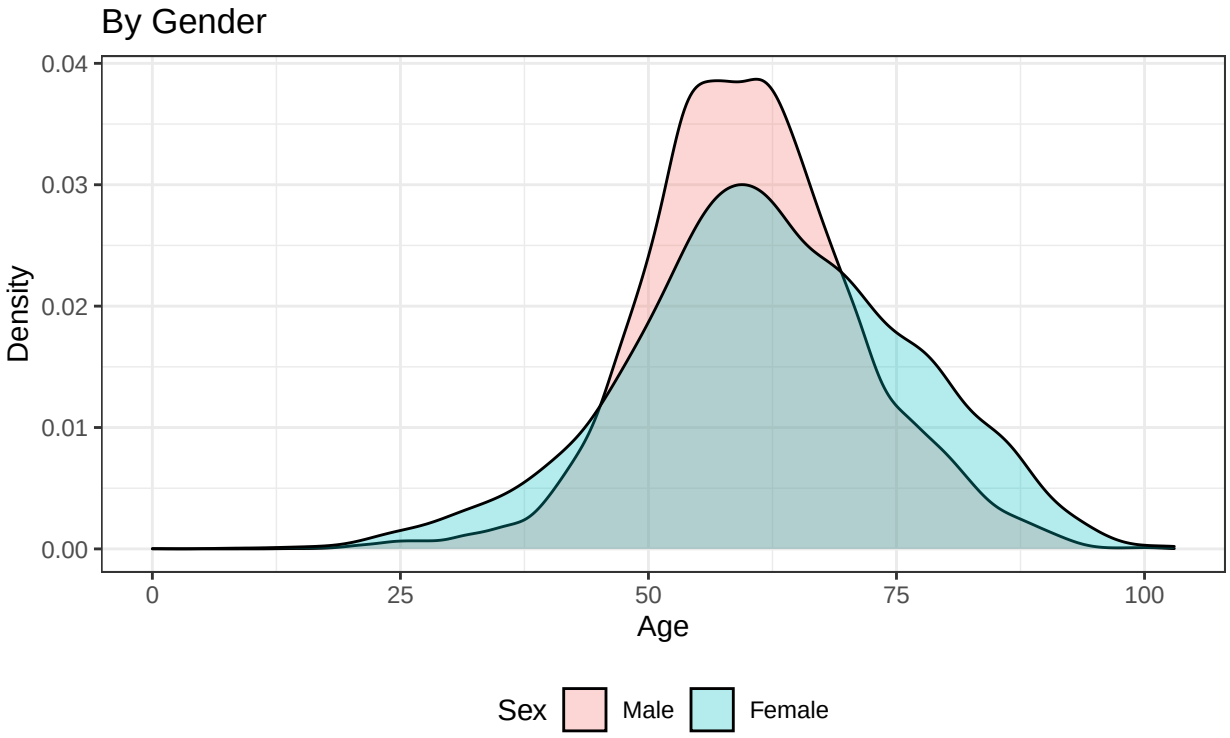

Standard Therapy
Uninsured
Insured
Insurance
89%
11%
89%
11%
0.00
0.25
0.50
0.75
1.00
Percentage
Standard_Therapy
Not Standard
Standard

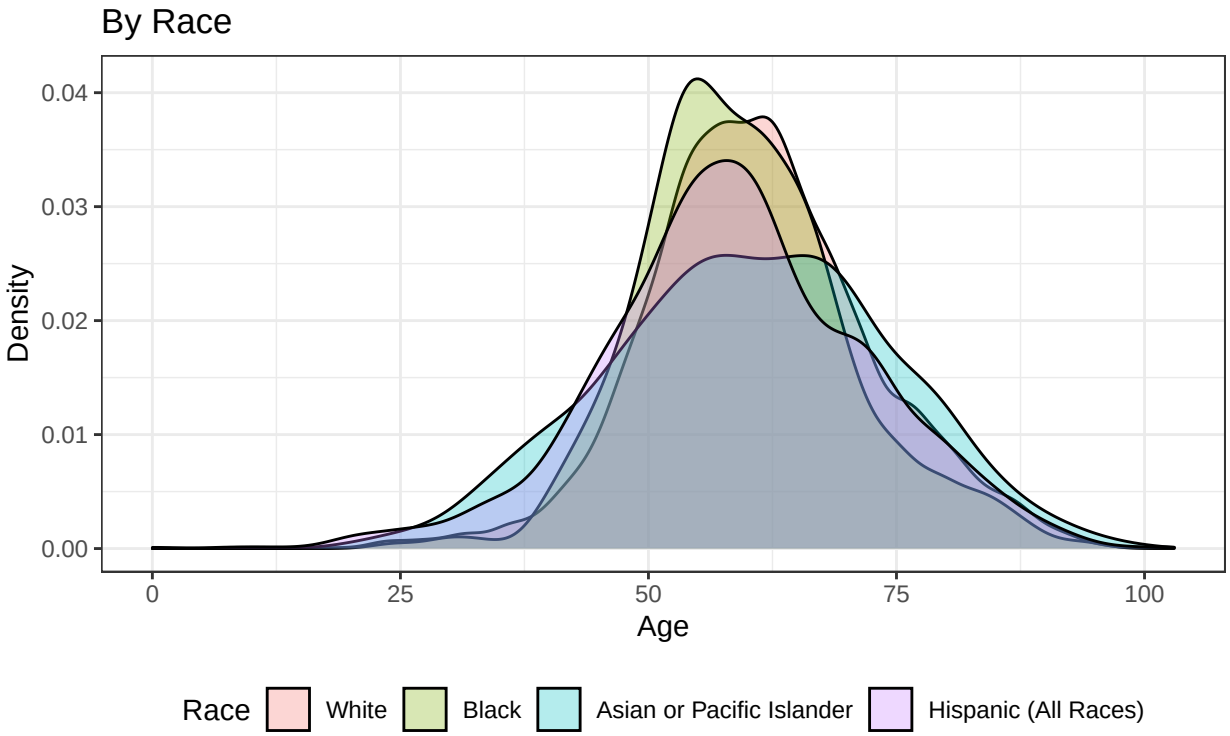


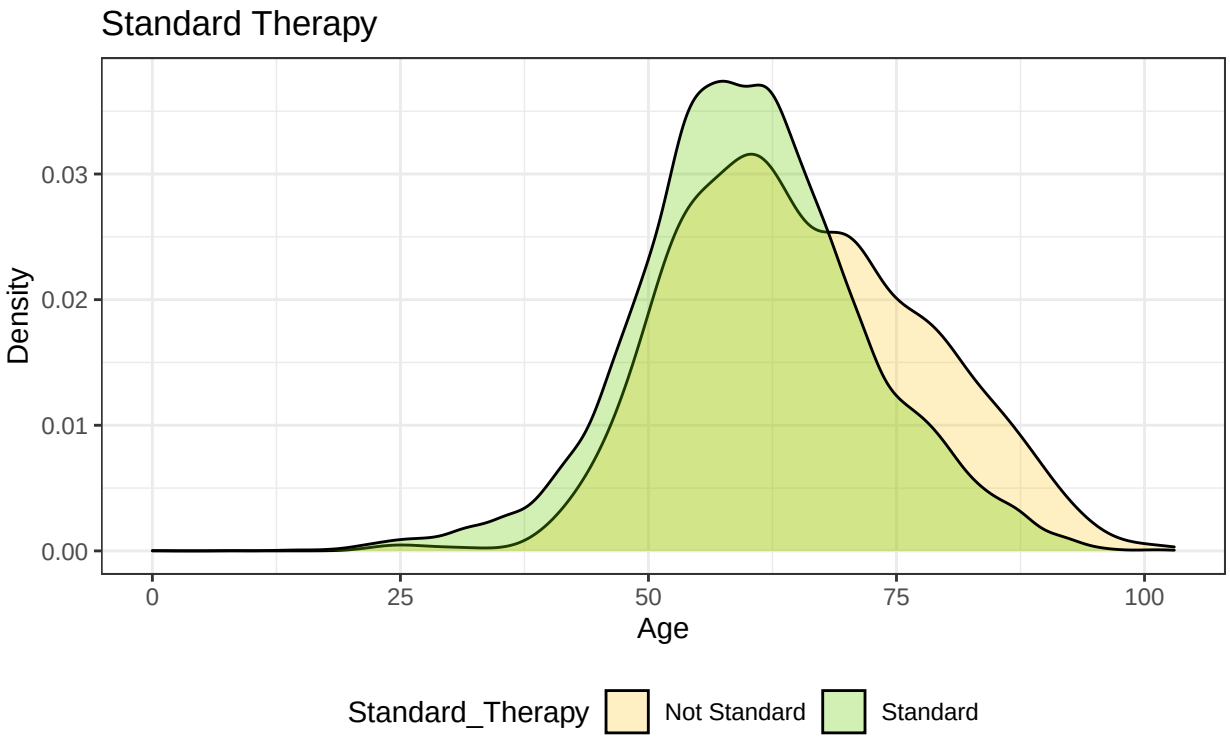


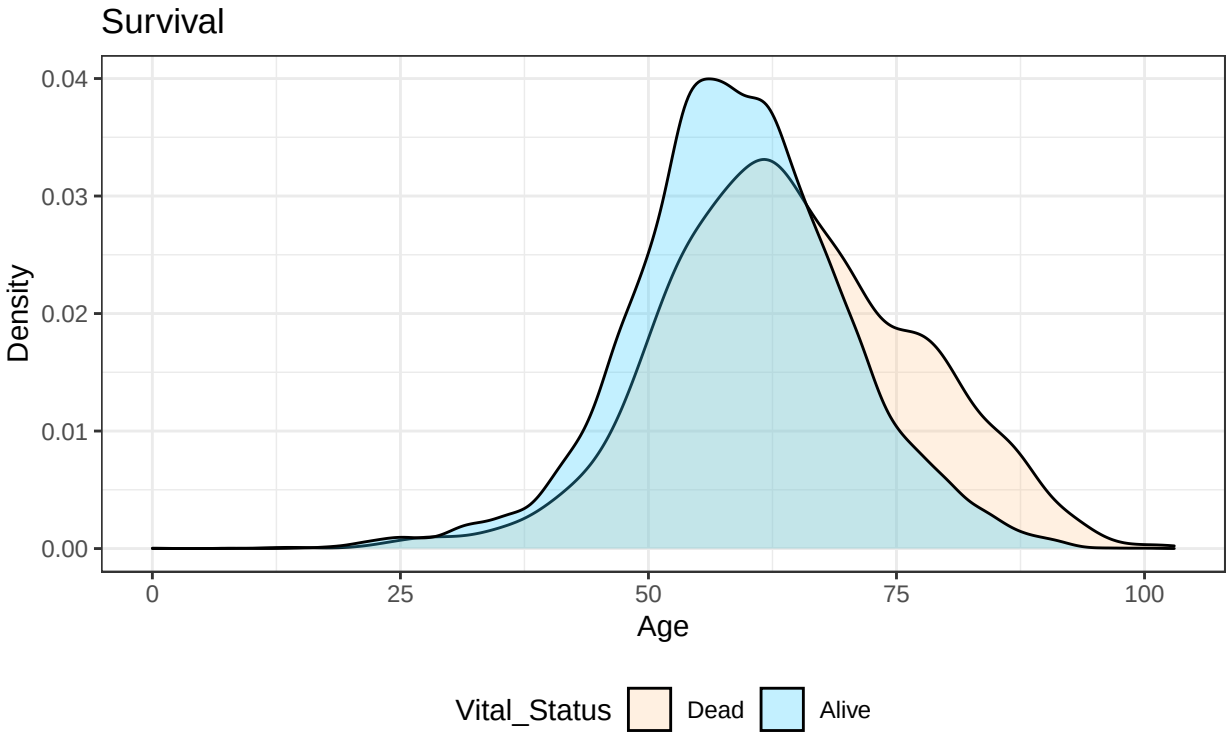
- **Age**

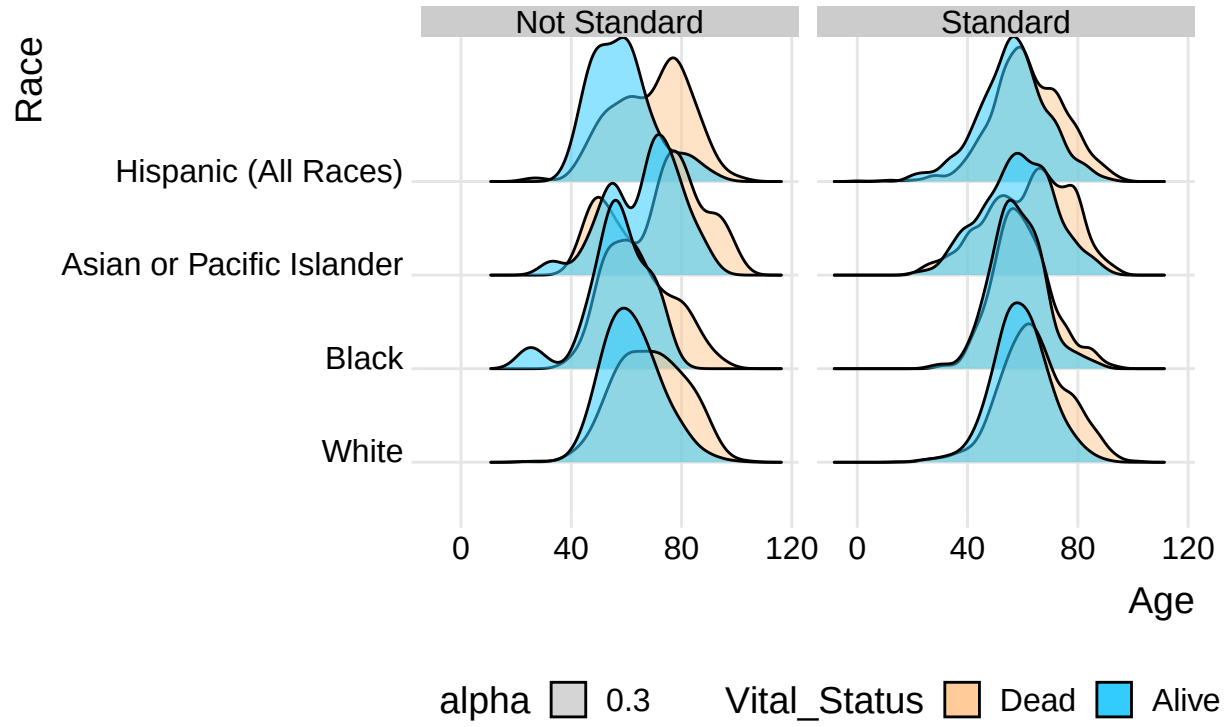
From the Age plots, we can find that patients who are getting standard therapy and survival are younger than patients who are not getting standard therapy, and not survived.


By Gender
Density
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
Age
Sex
Male
Female


By Race
Density
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
Age
Race
White
Black
Asian or Pacific Islander
Hispanic (All Races)

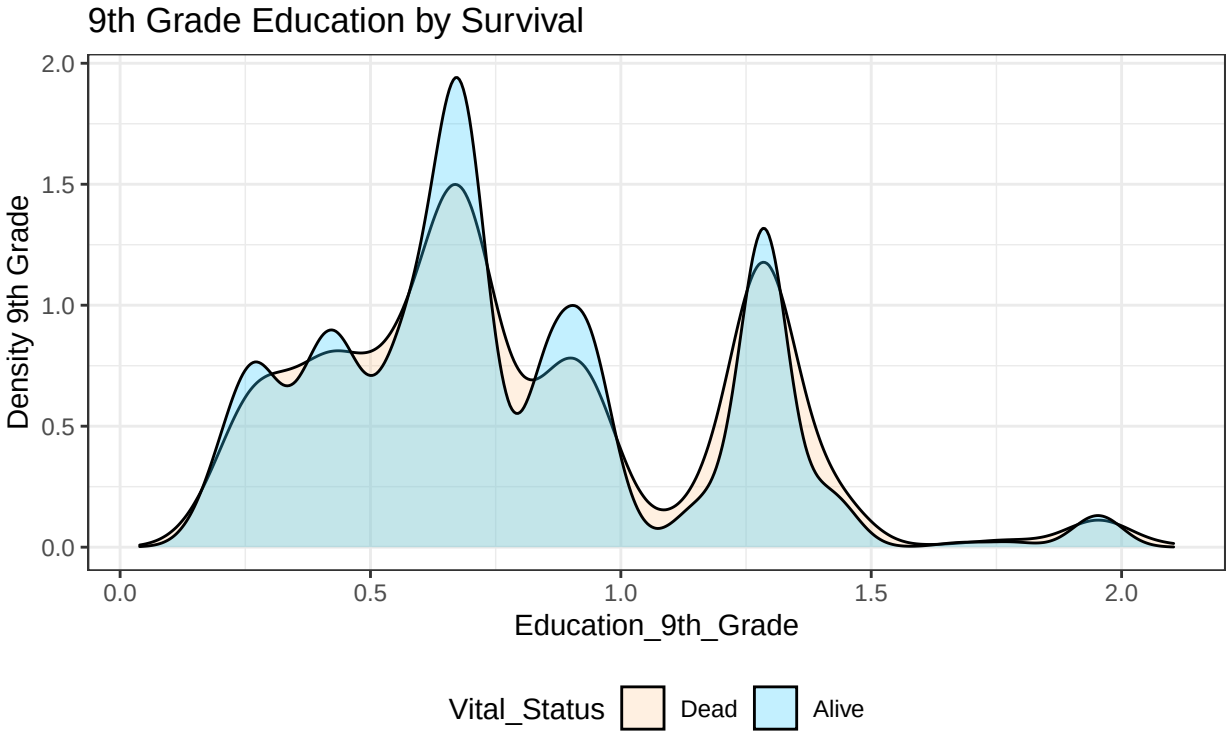

Standard Therapy
Density
0.03
0.02
0.01
0.00
0
25
50
75
100
Age
Standard_Therapy
Not Standard
Standard

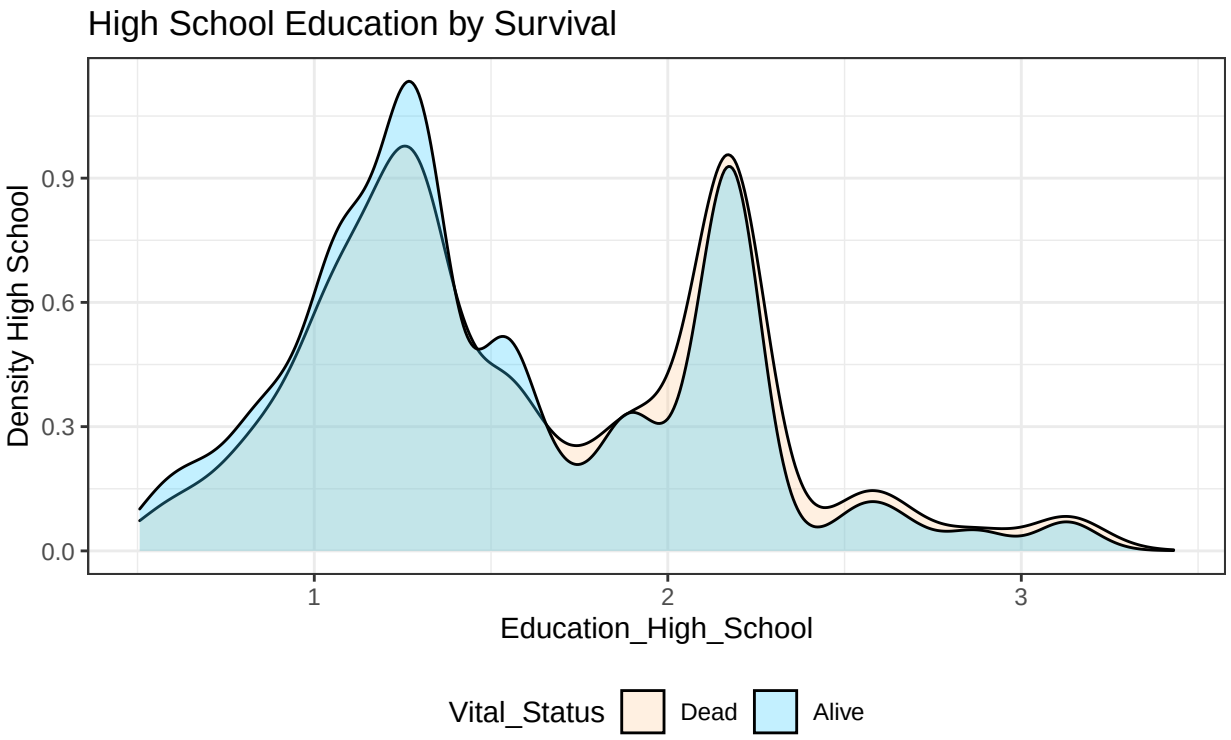

Survival
Density
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
Age
Vital_Status
Dead
Alive

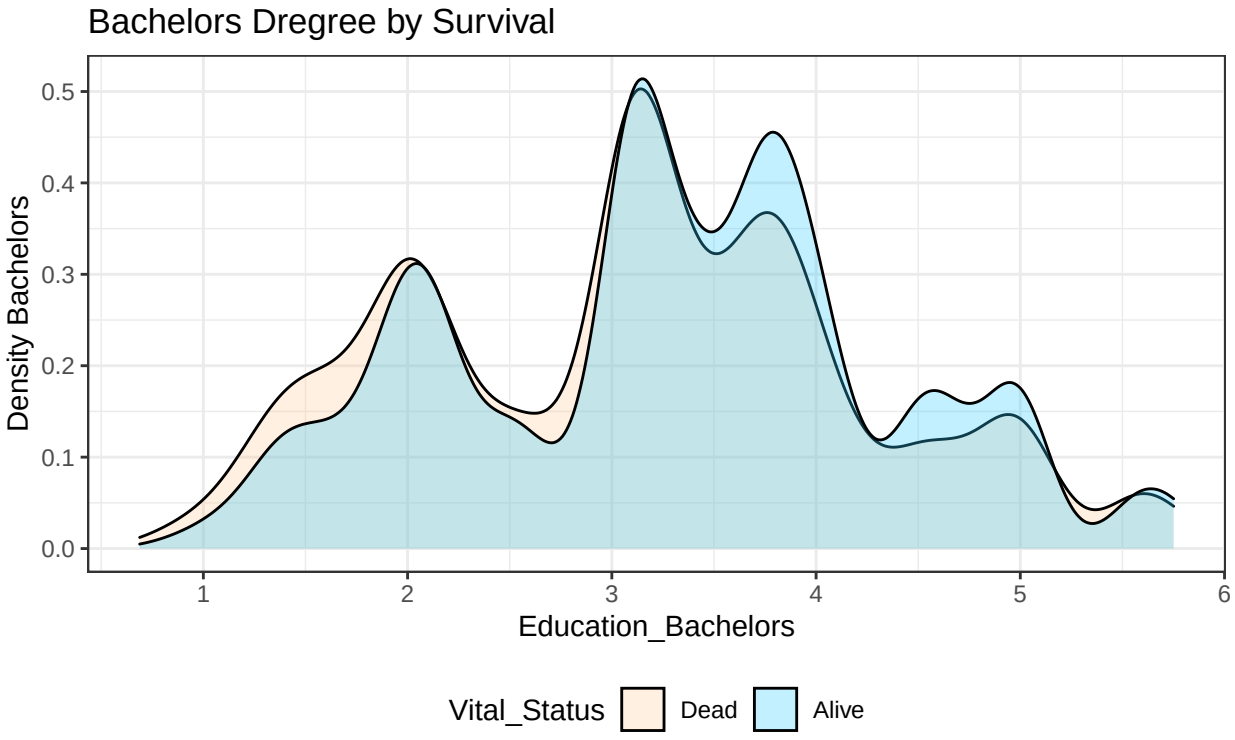


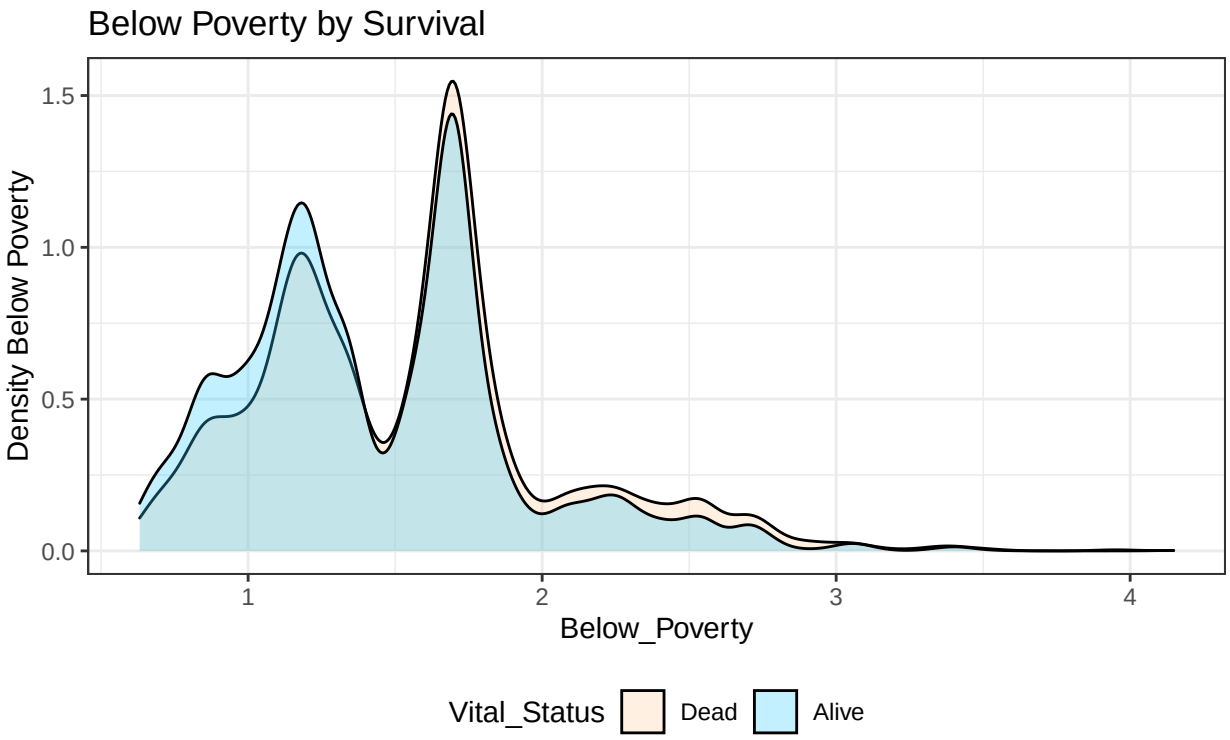
- **Education, Income and Language**

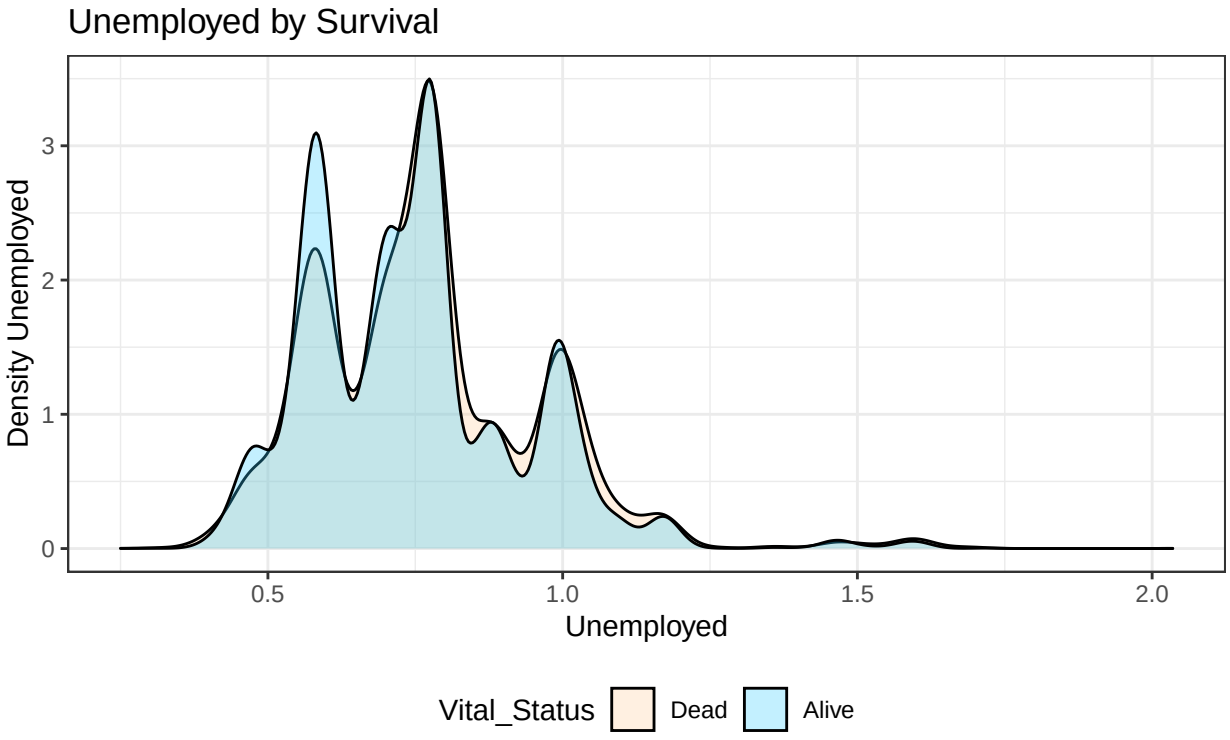
In these plots, education, language and unemployment are at 10% scale, income are at log scale. From these plots, we didn't find obvious difference between different treatment and survival status.

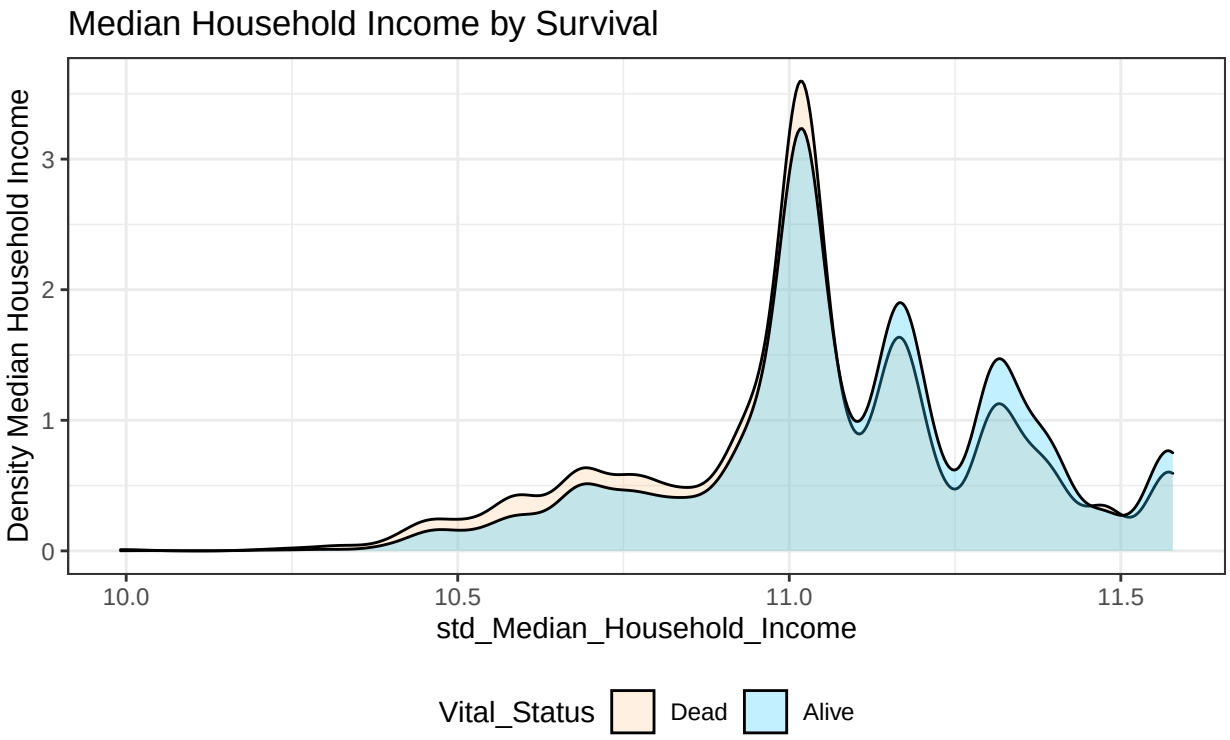


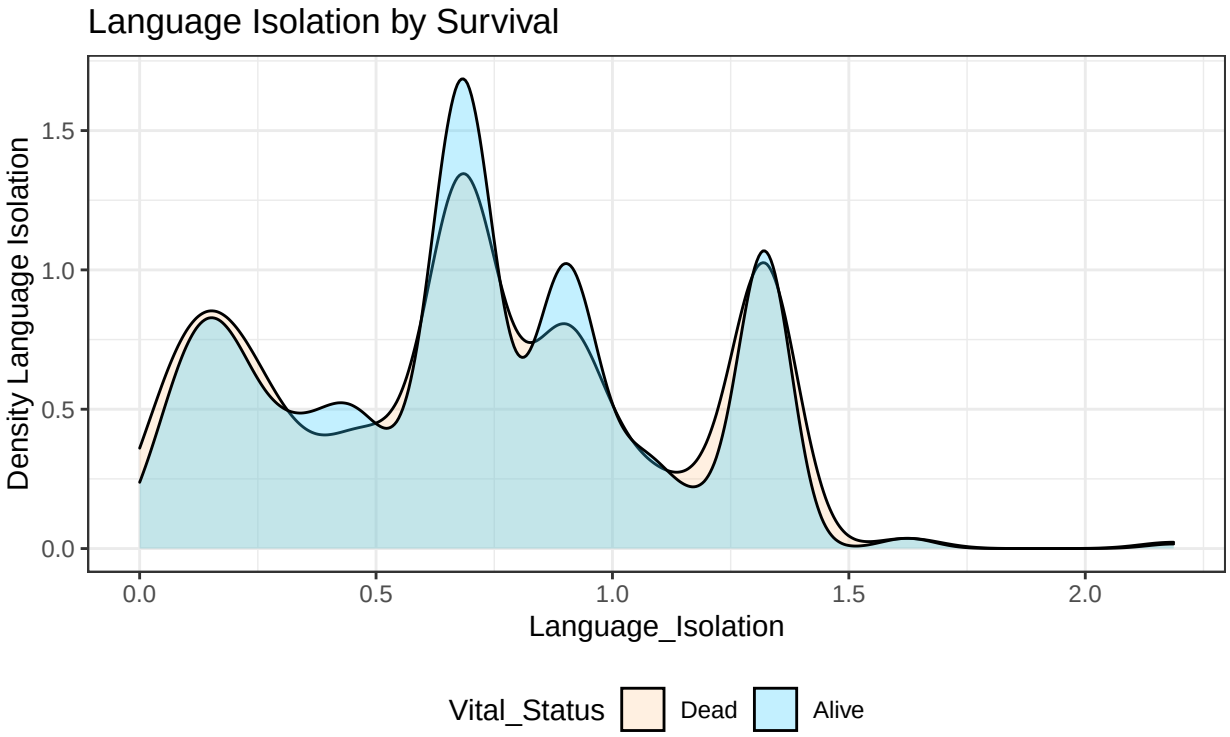

High School Education by Survival
Density High School
0.0
0.3
0.6
0.9
1
2
3
Education_High_School
Vital_Status
Dead
Alive

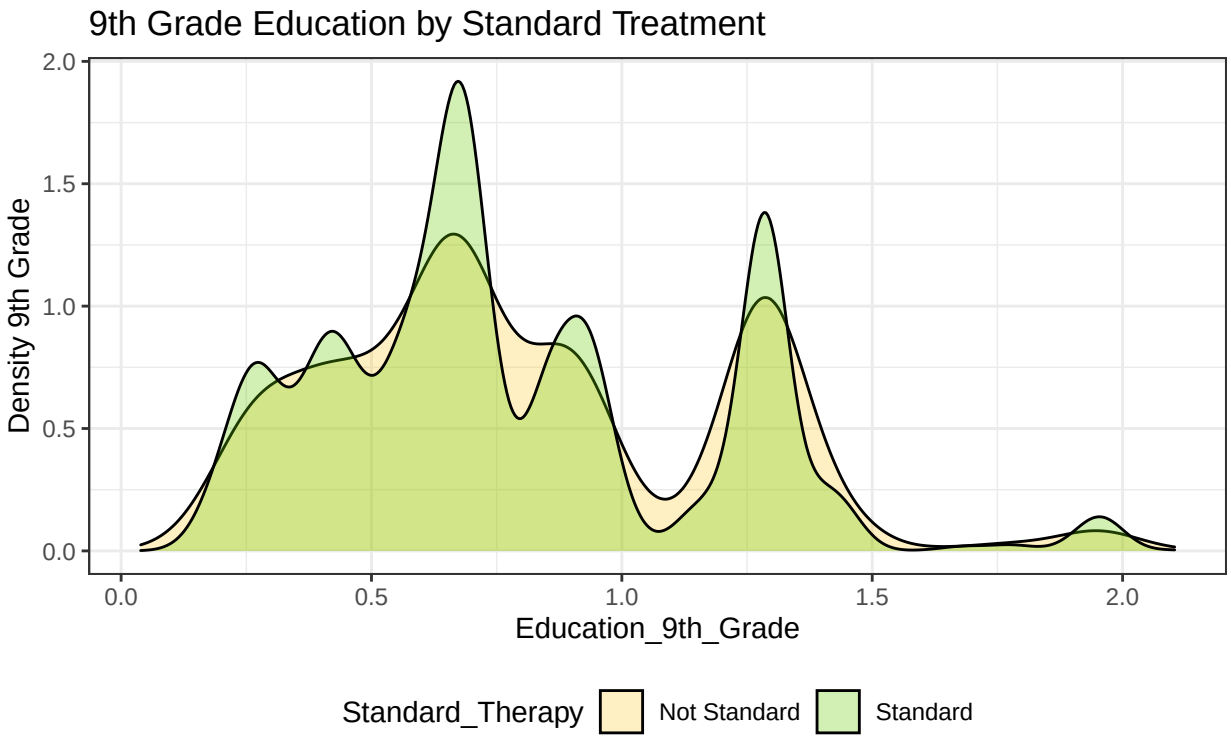

Bachelors Dregree by Survival
Density Bachelors
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
Education_Bachelors
Vital_Status
Dead
Alive

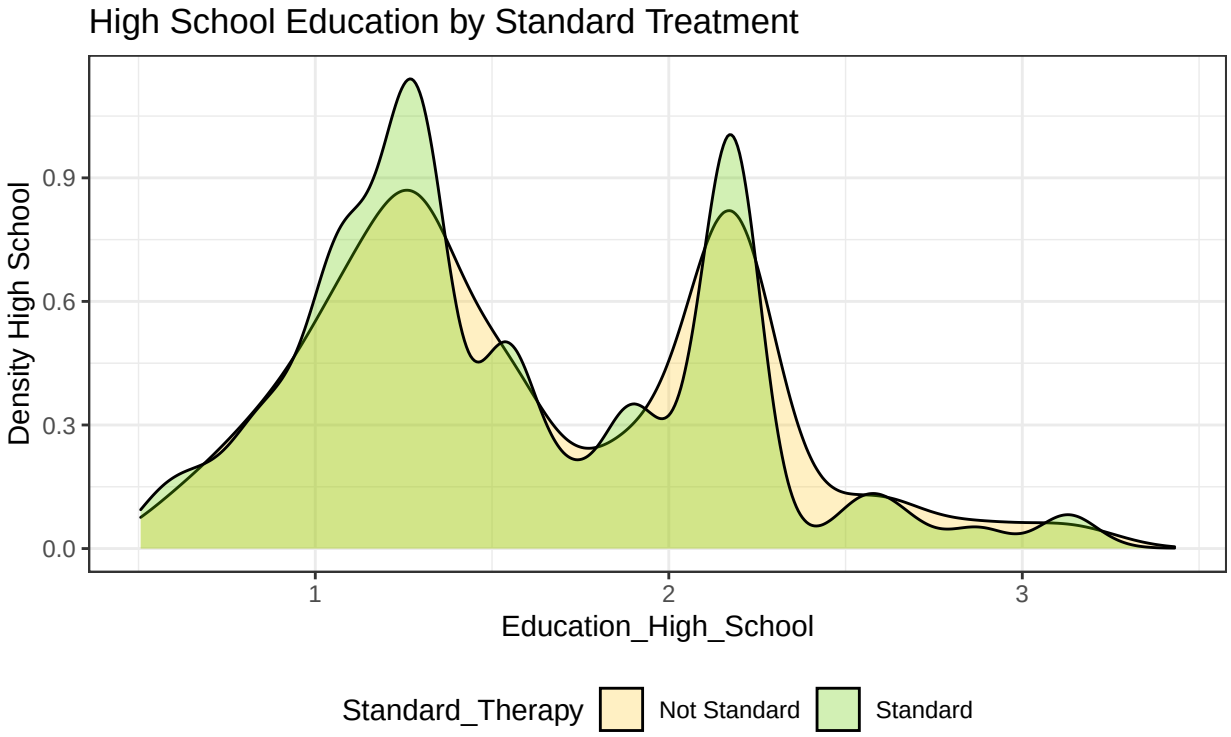

Below Poverty by Survival
Density Below Poverty
1.5
1.0
0.5
0.0
1
2
3
4
Below_Poverty
Vital_Status
Dead
Alive

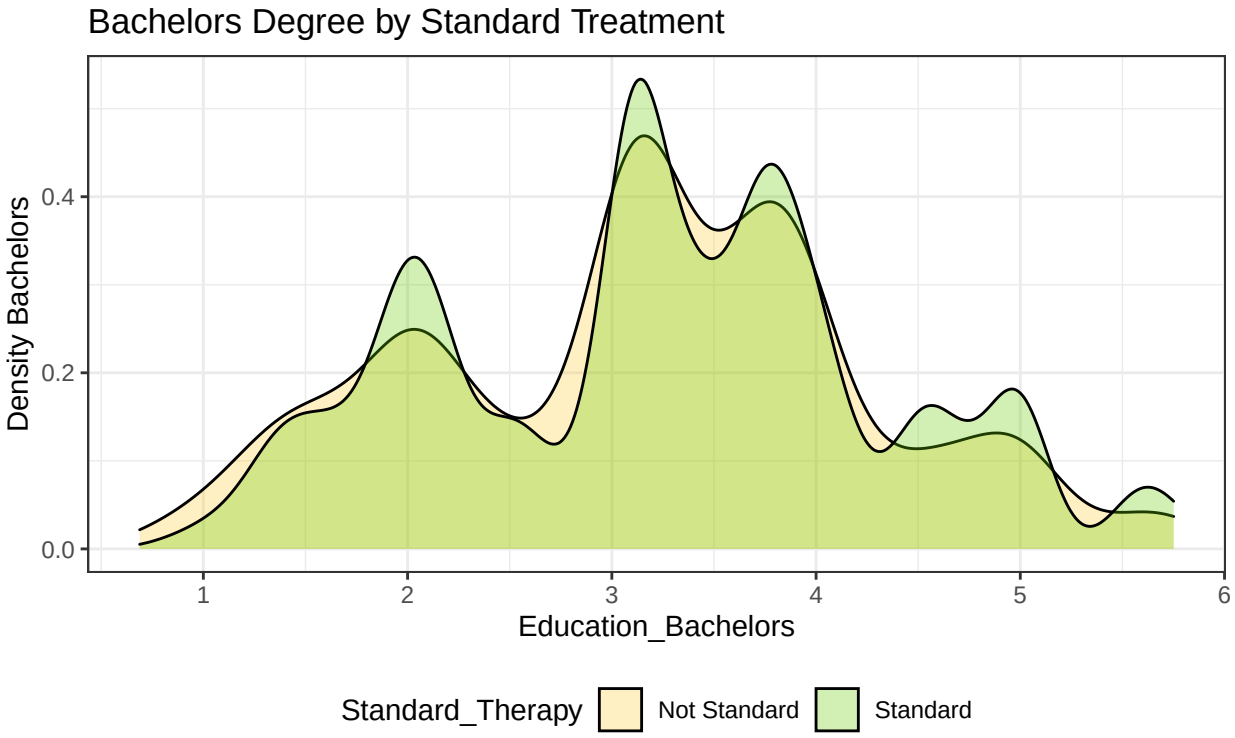

Unemployed by Survival
Density Unemployed
3
2
1
0
0.5
1.0
1.5
2.0
Unemployed
Vital_Status
Dead
Alive

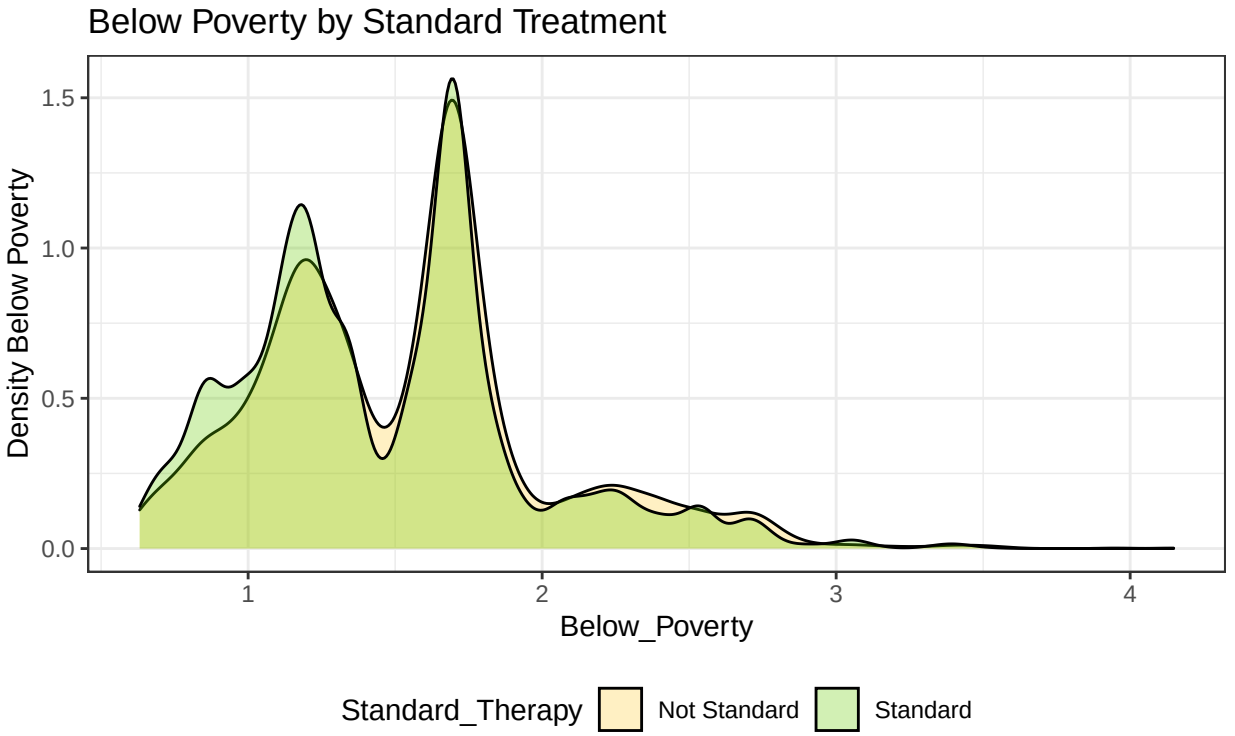

Median Household Income by Survival
Density Median Household Income
0
1
2
3
10.0
10.5
11.0
11.5
std_Median_Household_Income
Vital_Status
Dead
Alive

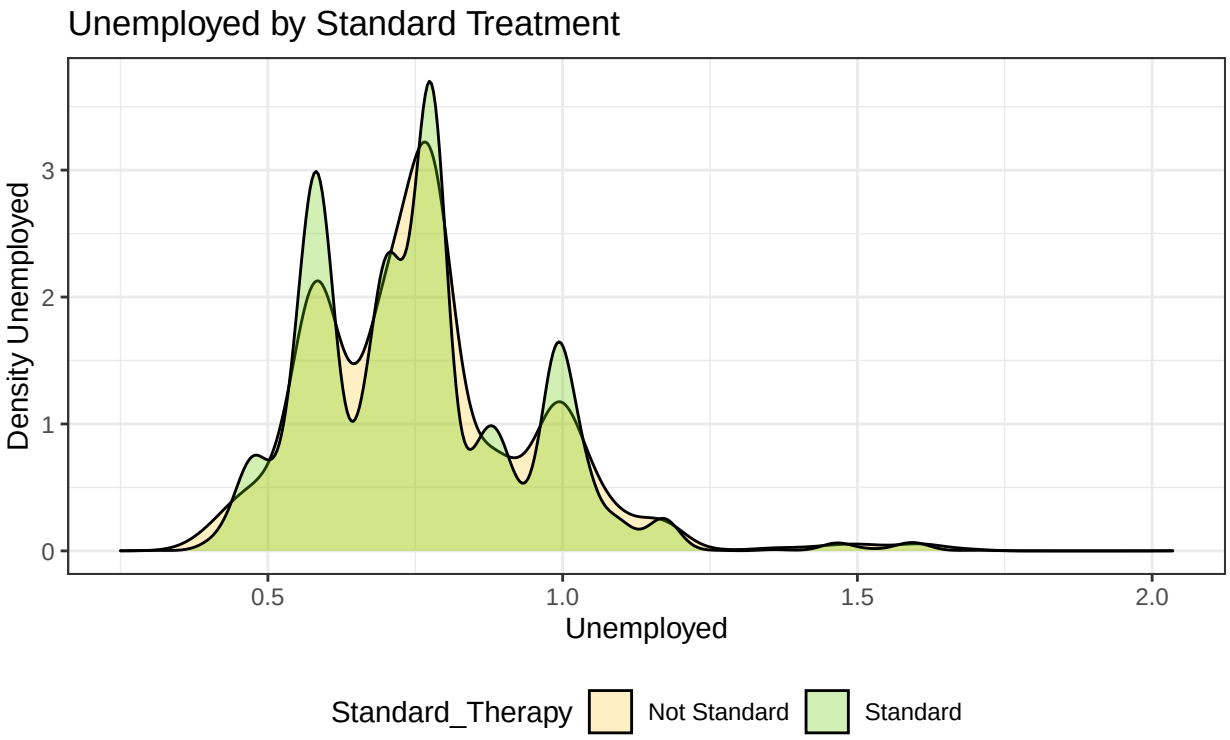

Language Isolation by Survival
Density Language Isolation
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Language_Isolation
Vital_Status
Dead
Alive

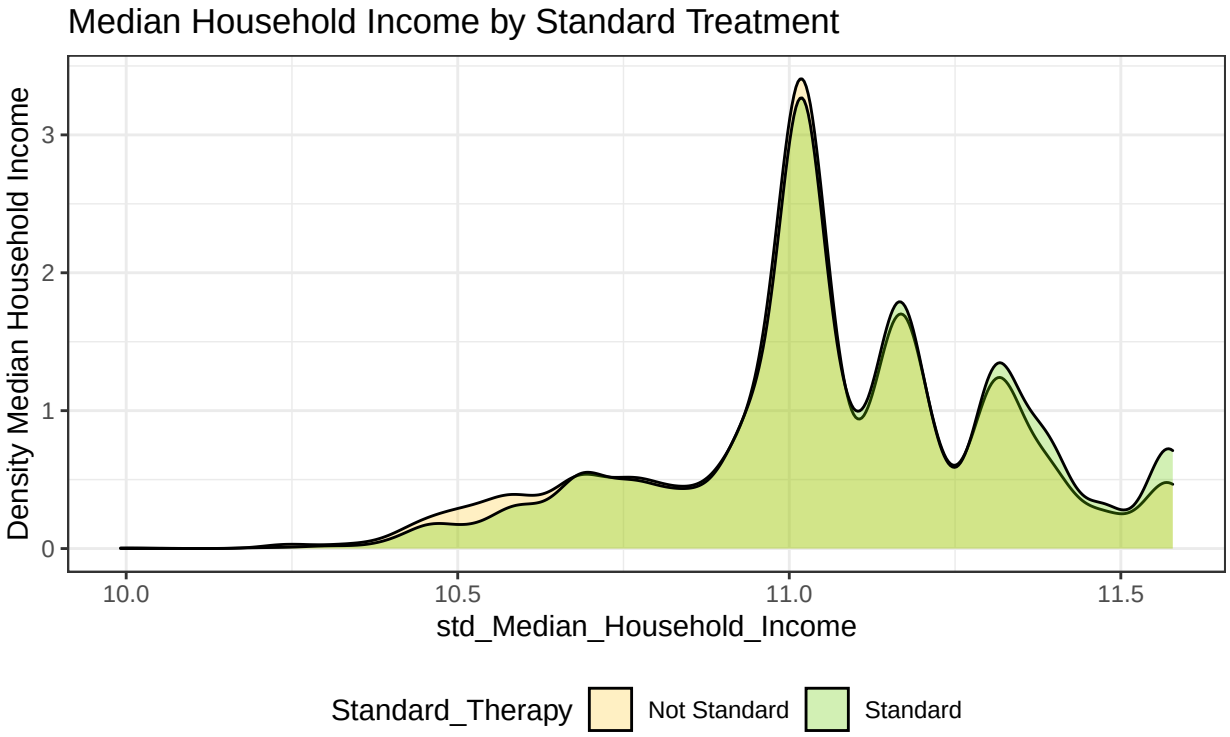

9th Grade Education by Standard Treatment
Density 9th Grade
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
Education_9th_Grade
Standard_Therapy
Not Standard
Standard

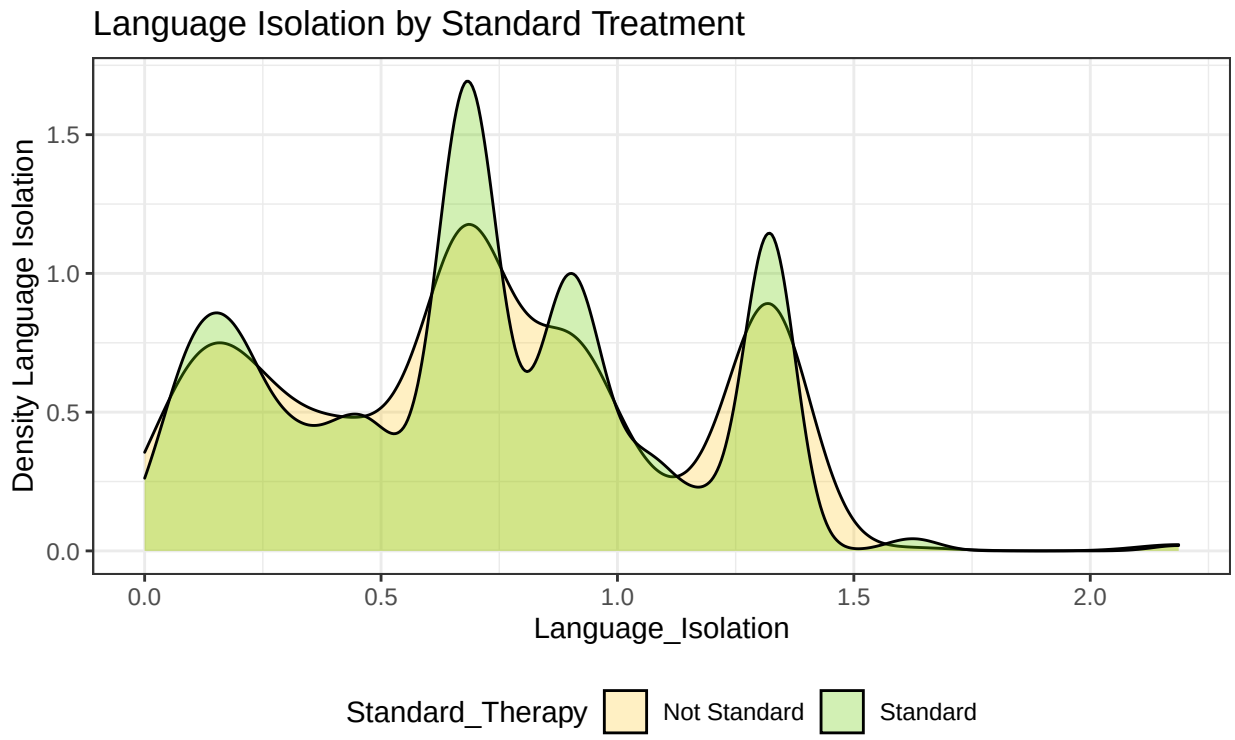

High School Education by Standard Treatment
Density High School
0.9
0.6
0.3
0.0
1
2
3
Education_High_School
Standard_Therapy
Not Standard
Standard


Bachelors Degree by Standard Treatment
Density Bachelors
0.0
0.2
0.4
1
2
3
4
5
6
Education_Bachelors
Standard_Therapy
Not Standard
Standard


Below Poverty by Standard Treatment
Density Below Poverty
0.0
0.5
1.0
1.5
1
2
3
4
Below_Poverty
Standard_Therapy
Not Standard
Standard


Unemployed by Standard Treatment
Density Unemployed
0
1
2
3
0.5
1.0
1.5
2.0
Unemployed
Standard_Therapy
Not Standard
Standard


Median Household Income by Standard Treatment
Density Median Household Income
0
1
2
3
10.0
10.5
11.0
11.5
std_Median_Household_Income
Standard_Therapy
Not Standard
Standard

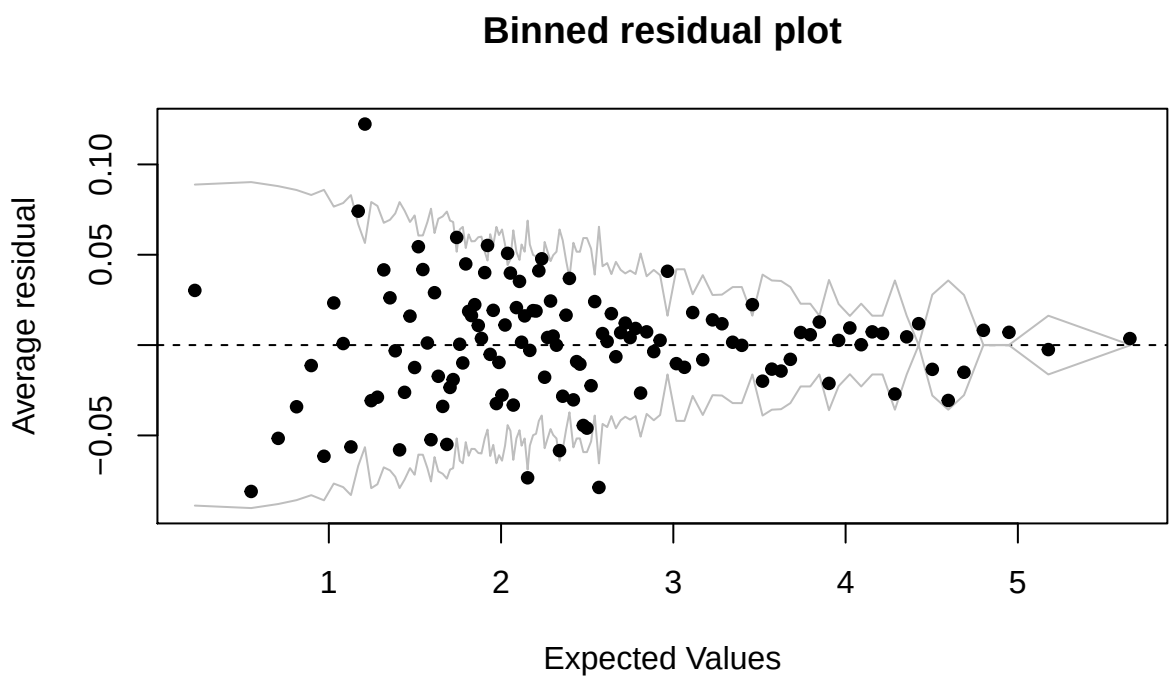


Model Check

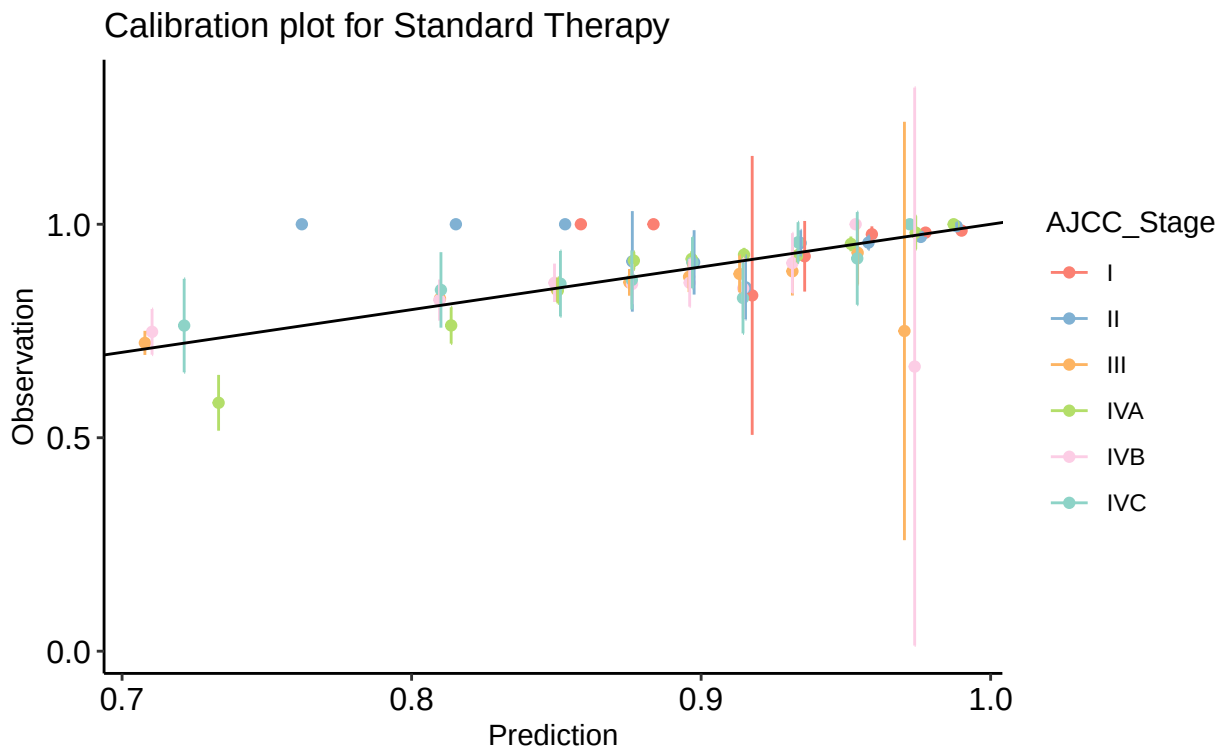
- **Standard Therapy**

```
##
## Call:
## glm(formula = Standard_Therapy ~ AJCC_Stage + Sex + Region +
##     Race + std_Age + Insurance + Education_High_School + Unemployed +
##     std_Median_Household_Income + Language_Isolation, family = binomial(link = "logit"),
##     data = oropharynx)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2573   0.2112   0.4023   0.5315   1.4396
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.25573    2.67326  -1.218 0.223268
## AJCC_StageII                   -0.71420    0.19450  -3.672 0.000241 ***
## AJCC_StageIII                  -2.64131    0.16173 -16.331  < 2e-16 ***
## AJCC_StageIVA                  -2.07044    0.16223 -12.762  < 2e-16 ***
## AJCC_StageIVB                  -2.45458    0.17308 -14.182  < 2e-16 ***
## AJCC_StageIVC                  -2.02495    0.19999 -10.125  < 2e-16 ***
## SexFemale                      -0.13392    0.06533  -2.050 0.040371 *
## RegionGeorgia                   0.22455    0.09764   2.300 0.021455 *
## RegionConnecticut               0.01100    0.11221   0.098 0.921902
## RaceBlack                      -0.41163    0.09193  -4.478 7.55e-06 ***
## RaceAsian or Pacific Islander   0.23526    0.13767   1.709 0.087486 .
## RaceHispanic (All Races)       -0.06564    0.09298  -0.706 0.480174
```

```
## std_Age                       -0.45591    0.02500 -18.239  < 2e-16 ***
## InsuranceUninsured            -0.45824    0.08982  -5.101 3.37e-07 ***
## Education_High_School         -0.09346    0.09900  -0.944 0.345186
## Unemployed                     0.55579    0.20455   2.717 0.006584 **
## std_Median_Household_Income    0.65512    0.22818   2.871 0.004090 **
## Language_Isolation             0.06373    0.13096   0.487 0.626519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10312.9  on 15023  degrees of freedom
## Residual deviance:  9281.3  on 15006  degrees of freedom
## AIC: 9317.3
##
## Number of Fisher Scoring iterations: 6
```




```
## $calibration_plot
```

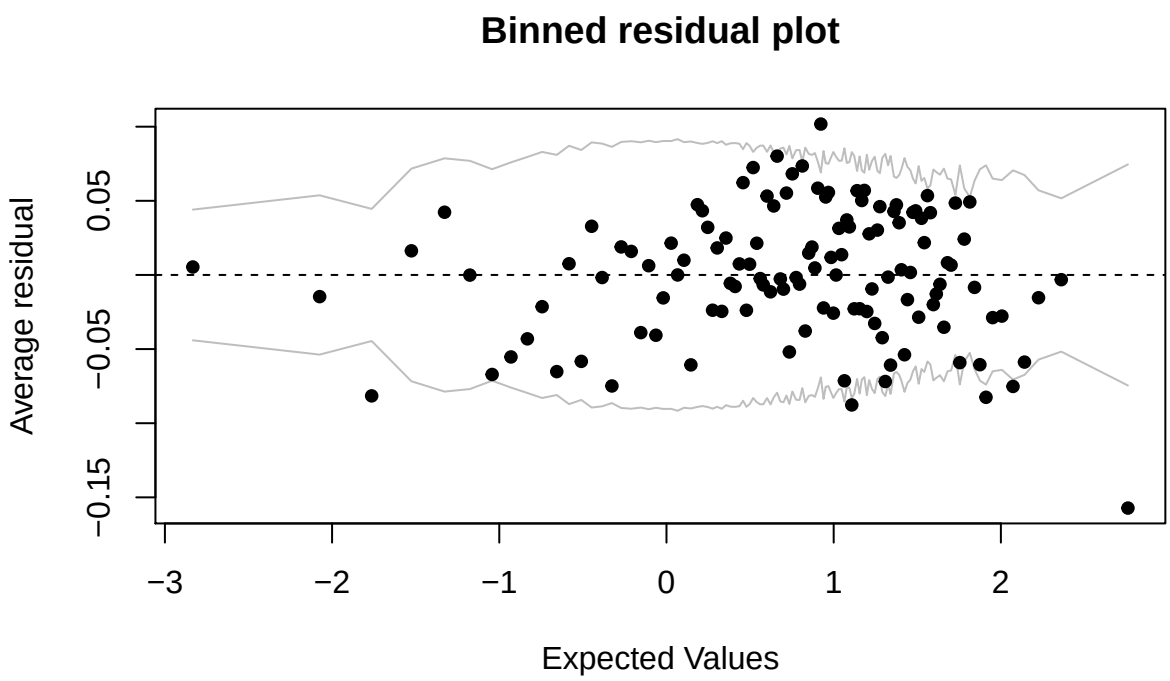



```
## C Statistic =  0.7287779
```

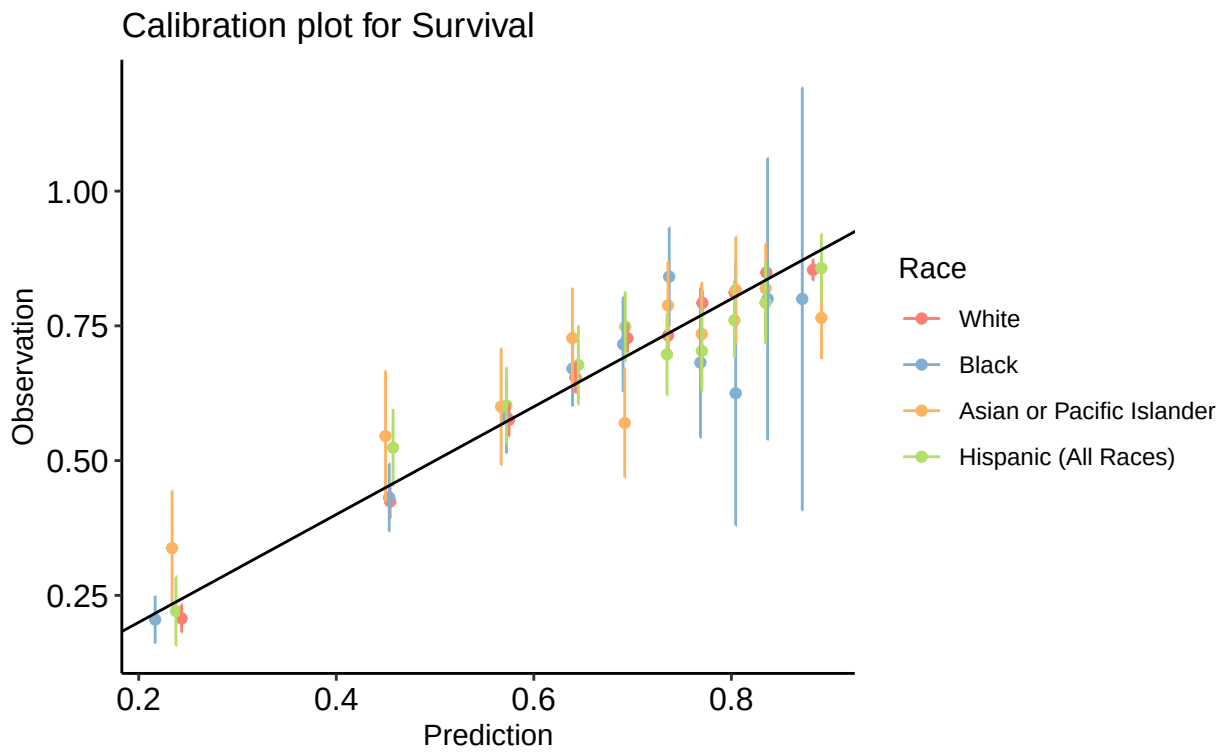
- **Survival**

```
##
## Call:
## glm(formula = Vital_Status ~ AJCC_Stage + Sex + Region + Race +
##     std_Age + Insurance + Standard_Therapy + Education_High_School +
##     Unemployed + std_Median_Household_Income + Language_Isolation,
##     family = binomial, data = oropharynx)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5947  -1.0024   0.6276   0.8259   2.4219
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -7.946383   1.873553  -4.241 2.22e-05 ***
## AJCC_StageII                    -0.451589   0.071296  -6.334 2.39e-10 ***
## AJCC_StageIII                   -0.063830   0.066371  -0.962 0.336193
## AJCC_StageIVA                   -0.305894   0.061512  -4.973 6.59e-07 ***
## AJCC_StageIVB                   -1.226920   0.079530 -15.427  < 2e-16 ***
## AJCC_StageIVC                   -2.515955   0.123896 -20.307  < 2e-16 ***
## SexFemale                       -0.149545   0.044884  -3.332 0.000863 ***
## RegionGeorgia                    0.109174   0.067165   1.625 0.104067
## RegionConnecticut               -0.006648   0.076209  -0.087 0.930485
## RaceBlack                       -0.698853   0.068936 -10.138  < 2e-16 ***
## RaceAsian or Pacific Islander -0.116553   0.084928  -1.372 0.169945
## RaceHispanic (All Races)        -0.233814   0.064104  -3.647 0.000265 ***
## std_Age                         -0.425031   0.017377 -24.460  < 2e-16 ***
```

```
## InsuranceUninsured            -0.929135   0.057785 -16.079  < 2e-16 ***
## Standard_TherapyStandard       0.917601   0.059168  15.508  < 2e-16 ***
## Education_High_School         -0.061368   0.069855  -0.879 0.379669
## Unemployed                     0.207099   0.144191   1.436 0.150921
## std_Median_Household_Income    0.765732   0.159986   4.786 1.70e-06 ***
## Language_Isolation            -0.045389   0.091394  -0.497 0.619454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19203  on 15023  degrees of freedom
## Residual deviance: 16731  on 15005  degrees of freedom
## AIC: 16769
##
## Number of Fisher Scoring iterations: 4
```




```
## $calibration_plot
```




```
## C Statistic =  0.7258336
```

- **Matching**

```
##
## Call:
## glm(formula = Vital_Status ~ AJCC_Stage + Sex + Region + Race +
##     std_Age + Insurance + Standard_Therapy + Education_High_School +
##     Unemployed + std_Median_Household_Income + Language_Isolation,
##     family = binomial, data = matched.data, weights = weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9224  -0.3909   0.0173   0.7002   1.7916
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -37.36925    2.95129 -12.662  < 2e-16 ***
## AJCC_StageII                    -1.76151    0.11829 -14.892  < 2e-16 ***
## AJCC_StageIII                   -0.70167    0.09323  -7.526 5.23e-14 ***
## AJCC_StageIVA                   -1.36226    0.08969 -15.188  < 2e-16 ***
## AJCC_StageIVB                   -5.77303    0.31628 -18.253  < 2e-16 ***
## AJCC_StageIVC                  -18.61364  141.78694  -0.131  0.89555
## SexFemale                       -0.54474    0.07270  -7.493 6.75e-14 ***
## RegionGeorgia                    0.40307    0.10008   4.028 5.63e-05 ***
## RegionConnecticut               -0.05566    0.10684  -0.521  0.60240
## RaceBlack                       -3.15616    0.19157 -16.475  < 2e-16 ***
## RaceAsian or Pacific Islander   -0.75458    0.12743  -5.922 3.19e-09 ***
## RaceHispanic (All Races)        -1.09418    0.10072 -10.864  < 2e-16 ***
## std_Age                         -1.49014    0.03861 -38.597  < 2e-16 ***
```

```
## InsuranceUninsured             -3.72711    0.16615 -22.432  < 2e-16 ***
## Standard_TherapyStandard        0.65021    0.09655   6.734 1.65e-11 ***
## Education_High_School          -0.16357    0.11192  -1.461  0.14389
## Unemployed                      1.22755    0.24329   5.046 4.52e-07 ***
## std_Median_Household_Income     3.41267    0.25228  13.528  < 2e-16 ***
## Language_Isolation             -0.42249    0.14375  -2.939  0.00329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14040.4  on 10127  degrees of freedom
## Residual deviance:  7662.7  on 10109  degrees of freedom
## AIC: 7700.7
##
## Number of Fisher Scoring iterations: 16
```