MA679 SEER Project Final Report

Anna Cook, Bruce Mallory, Alison Pedraza, & Ryan O'Dea

5/5/2021

Introduction

The purpose of this assignment was to analyze the SEER head and neck cancer dataset in order to determine whether there is any racial bias that arises in how patients are treated. More specifically, our group was interested to know whether there was any evidence of racial or gender bias in the rates at which patients are given surgery as part of their cancer treatment after controlling for insurance and cancer stage.

Data

The SEER dataset we worked with included patients with cancer in one of 7 different sites in the head and neck. We chose to focus our analysis on just one site, the oral cavity, as different sites may have different treatment protocols and therefore may show differences in patients receiving surgery or not. In the original dataset, the cancer stages were broken down into stages I, II, III, IVA, IVB, IVC, and IVNOS (stage 4 not otherwise specified). Because these IVNOS patients cannot accurately be placed into a stage IV category and there are too few observations to analyze these patients as a stage of their own, we removed these 147 observations. Additionally, the original dataset contained both "insured" and "insured/no specifics" which we combined into a single category called "insured." Lastly, there were not very many observations for the American Indian/Alaska Native or the Asian or Pacific Islander racial groups, so these two groups were combined into a single group called "Native Am. or Asian."

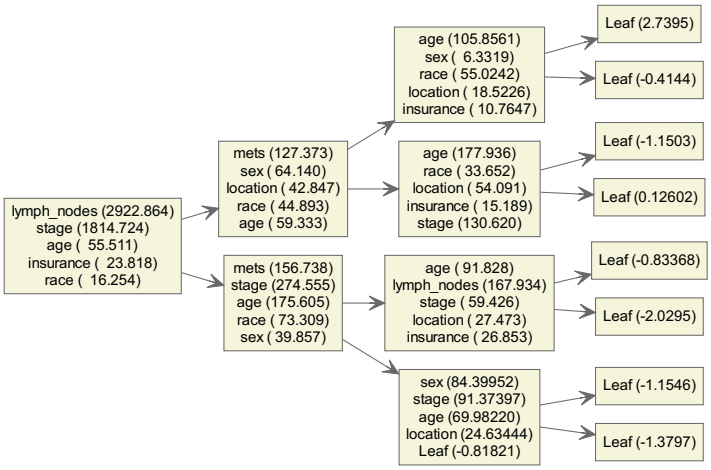
EDA/Methods

As part of a preliminary exploratory analysis, we build a machine learning model using xgboost in R in order to identify the most important variables for predicting whether a patient received surgery or not. This model showed the relative imbalance of the provided SEER data set. After the hyperparameter search, if we trained the model on the full data (without consideration for accuracy testing) the relative importance matrix would show a much different story compared to random sampling into train and test sets. Two different matrices are shown in the appendix. Additionally, each time a different seed was forced into the random sampling, the order of importance would drastically shift between the different features. See total data set importance matrix in Appendix figure 2 and seeded sample importance in figure 3. As conjecture, it could be considered that the random sampling has the potential of completely avoiding certain groups due to the imbalance of the data set within each feature. Hypothetically, it is entirely possible that the random sampling for the training data could entirely miss black people who are uninsured. Because of this imbalance, it would likely be much more advisable to continue with descriptive statistics over the entire dataset compared to attempting to model the set.

When looking at tuning the model on the entire set (below), we see the relative importance gain at each stage of the decision tree proceed by the log odds of case 1- Surgery Completed. We can also look into an Rpart decision tree which is generated given a complexity metric. This tree is generally a more interpretable basic decision tree - one without gradient booster to improve accuracy of the fit. Using a complexity parameter of 0.0015 yielded results similar to the gradient boosted tree- dividing by lymph node involvement, by stage, then location age and race (Appendix, Figure 4). These results led us to conclude that while there are a

number of important factors contributing to whether a patient receives surgery or not, race seems to play a role. The rest of our analysis is aimed at exploring this relationship further.

In order to identify whether there is a difference in the likelihood of patients of different races being treated with surgery, we fit a series of logistic regression models with surgery as the outcome variable, with "Yes" or 1 indicating the patient received surgery as part of their cancer treatment, and "No" or 0 indicating no surgery. The results are discussed in the next section.



Results

First, we fit a logistic regression model with Race, Age, Sex, Insurance, and Cancer Stage as predictors, and surgery as the outcome variable. Initially, we did not include any interaction terms. The results are displayed below. The intercept represents insured white males with stage I cancer. From the output, we can see that almost all of the variable levels are statistically significant at alpha = 0.05. Importantly, we noticed that as the cancer stage progresses from I to IVC, patients are less likely to receive surgery as part of their cancer treatment. Additionally, patients who have Medicaid are less likely to receive surgery than patients who are privately insured, and patients who do not have insurance are even less likely than that to receive surgery. It also appears that females are slightly more likely to receive surgery than males. When looking at the racial groups, the model summary shows significant coefficients (at alpha = 0.05) for both Black and Native American or Asian patients, where Black patients are less likely to receive surgery than white patients, and Native American or Asian patients are more likely. Although these results are interesting, it is unclear whether this result is due to bias from the doctors giving different recommendations, or whether there may be some kind of interaction going on. To answer this question, we fit several additional logistic regression models with various interaction terms.

```
##
## Call:
## glm(formula = 1 * (`Surgery Performed?` == "Yes") ~ Sex + Insurance +
##     Race + `AJCC 7 Stage` + `Age at Diagnosis`, family = binomial(link = "logit"),
##     data = oral_cavity)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.0076   0.2456   0.3613   0.5680   1.9239
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               5.459677   0.226940  24.058  < 2e-16 ***
## SexFemale                 0.157295   0.079324   1.983  0.04737 *
## InsuranceAny Medicaid    -0.564323   0.096327  -5.858 4.67e-09 ***
## InsuranceUninsured       -1.009403   0.170136  -5.933 2.98e-09 ***
## RaceBlack                -0.522797   0.120971  -4.322 1.55e-05 ***
## RaceHispanic              0.199732   0.119618   1.670  0.09497 .
## RaceNative Am. or Asian   0.395589   0.145235   2.724  0.00645 **
## `AJCC 7 Stage`II         -0.693689   0.135908  -5.104 3.32e-07 ***
## `AJCC 7 Stage`III        -1.438069   0.126871 -11.335  < 2e-16 ***
## `AJCC 7 Stage`IVA        -1.704394   0.108709 -15.679  < 2e-16 ***
## `AJCC 7 Stage`IVB        -2.718996   0.183482 -14.819  < 2e-16 ***
## `AJCC 7 Stage`IVC        -3.590998   0.212656 -16.886  < 2e-16 ***
## `Age at Diagnosis`       -0.036840   0.003004 -12.264  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5575.3  on 6940  degrees of freedom
## Residual deviance: 4710.0  on 6928  degrees of freedom
## AIC: 4736
##
## Number of Fisher Scoring iterations: 5
```

The first interaction we were interested in exploring was that between race and cancer stage. We fit a logistic regression model identical to the previous one, with the only difference being that we also included the interaction term for race and cancer stage. The model summary is displayed below. From this output, we no longer see any significant coefficients for the race or the interaction between race and cancer stage. However, it is worth noting that the interaction between the Black racial group and stage IVB cancer was borderline, although not quite significant at alpha = 0.05 (p-value = .07). This aligns with a pattern we noticed in our EDA, where it appeared that Black patients with stage IVB cancer received surgery less often than the other racial groups at the same stage (see Figure 1)

```
##
## Call:
## glm(formula = 1 * (`Surgery Performed?` == "Yes") ~ Sex + Insurance +
##     Race * `AJCC 7 Stage` + `Age at Diagnosis`, family = binomial(link = "logit"),
##     data = oral_cavity)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9965   0.2421   0.3640   0.5636   1.9767
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              5.429971   0.230989  23.508  < 2e-16 ***
## SexFemale                                0.153047   0.079730   1.920   0.0549 .
## InsuranceAny Medicaid                   -0.559746   0.097047  -5.768 8.03e-09 ***
## InsuranceUninsured                      -1.006666   0.170700  -5.897 3.70e-09 ***
## RaceBlack                               -0.469722   0.408682  -1.149   0.2504
## RaceHispanic                             0.267607   0.325626   0.822   0.4112
## RaceNative Am. or Asian                  0.841054   0.465037   1.809   0.0705 .
## `AJCC 7 Stage`II                        -0.729452   0.153833  -4.742 2.12e-06 ***
## `AJCC 7 Stage`III                       -1.416693   0.147189  -9.625  < 2e-16 ***
## `AJCC 7 Stage`IVA                       -1.648615   0.123324 -13.368  < 2e-16 ***
## `AJCC 7 Stage`IVB                       -2.451117   0.232709 -10.533  < 2e-16 ***
## `AJCC 7 Stage`IVC                       -3.688023   0.263805 -13.980  < 2e-16 ***
## `Age at Diagnosis`                      -0.036827   0.003016 -12.211  < 2e-16 ***
## RaceBlack:`AJCC 7 Stage`II               0.551522   0.583129   0.946   0.3443
## RaceHispanic:`AJCC 7 Stage`II            0.311593   0.478330   0.651   0.5148
## RaceNative Am. or Asian:`AJCC 7 Stage`II -0.719843   0.578216  -1.245   0.2132
## RaceBlack:`AJCC 7 Stage`III              0.101105   0.498590   0.203   0.8393
## RaceHispanic:`AJCC 7 Stage`III           0.063842   0.419280   0.152   0.8790
## RaceNative Am. or Asian:`AJCC 7 Stage`III -0.775209   0.557043  -1.392   0.1640
## RaceBlack:`AJCC 7 Stage`IVA             -0.109389   0.441811  -0.248   0.8045
## RaceHispanic:`AJCC 7 Stage`IVA          -0.304777   0.371011  -0.821   0.4114
## RaceNative Am. or Asian:`AJCC 7 Stage`IVA -0.307824   0.517615  -0.595   0.5520
## RaceBlack:`AJCC 7 Stage`IVB             -1.146373   0.641441  -1.787   0.0739 .
## RaceHispanic:`AJCC 7 Stage`IVB          -0.525678   0.553937  -0.949   0.3426
## RaceNative Am. or Asian:`AJCC 7 Stage`IVB -0.463291   0.741007  -0.625   0.5318
## RaceBlack:`AJCC 7 Stage`IVC             -0.023539   0.746136  -0.032   0.9748
## RaceHispanic:`AJCC 7 Stage`IVC           0.645310   0.604746   1.067   0.2859
## RaceNative Am. or Asian:`AJCC 7 Stage`IVC -0.551777   0.910623  -0.606   0.5446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 5575.3  on 6940  degrees of freedom
## Residual deviance: 4692.3  on 6913  degrees of freedom
## AIC: 4748.3
##
## Number of Fisher Scoring iterations: 6
```

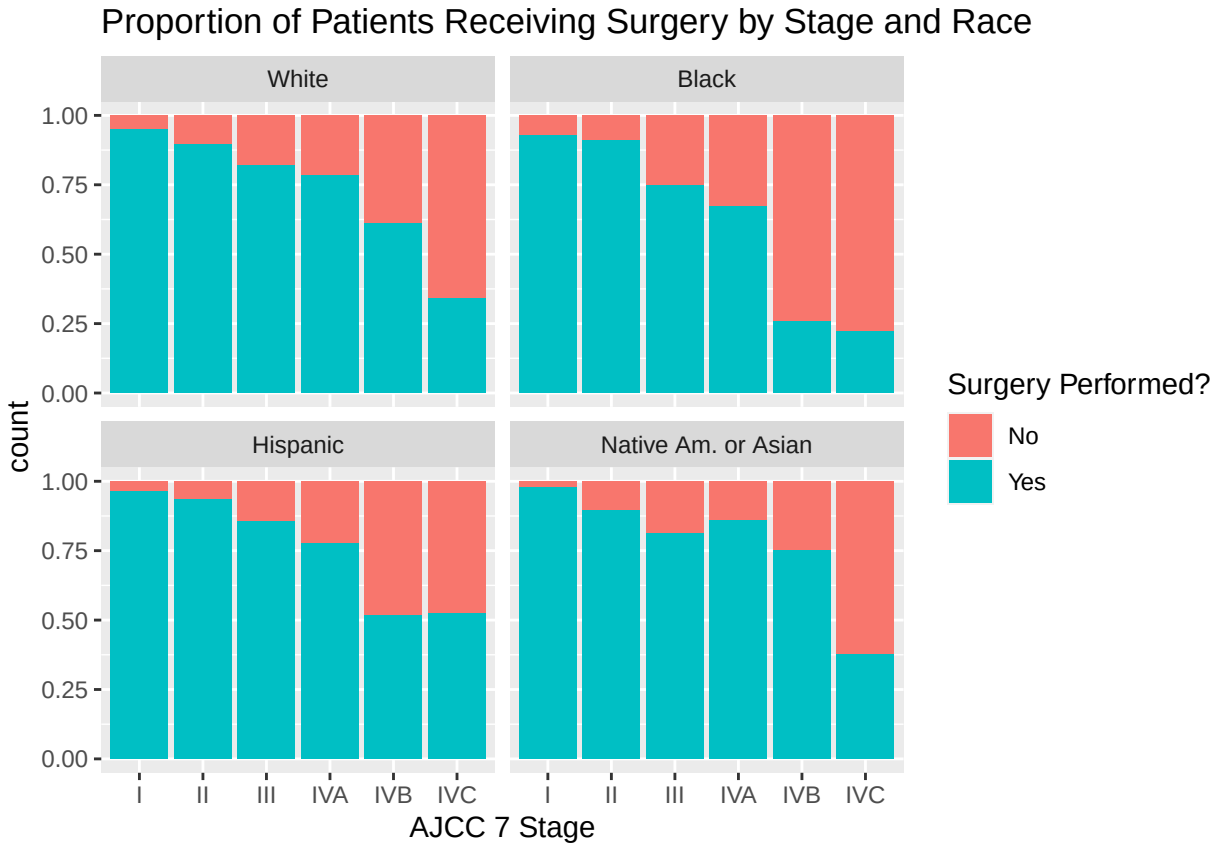


Figure 1: Black patients at stage IVB appear to receive surgery less frequently than patients from other racial groups at the same stage. This result was nonsignificant at alpha = 0.05, although not far from the threshold with a p-value of 0.07.

Another interaction we considered was that between race and insurance. Because insurance is an important factor in determining a patient's treatment plan, and insurance may be more easily available for some demographic groups than others, we wanted to explore this interaction further. We fit another logistic regression model with surgery as the outcome, and all of the predictors that were included in the first model, with the addition of the race and insurance interaction. The results are displayed below. In this output, we significant coefficients (alpha = 0.05) for the Black racial group as well as the interaction between the Hispanic racial group and the uninsured group. These patients are less likely to receive surgery as part of their cancer treatment according to the model.

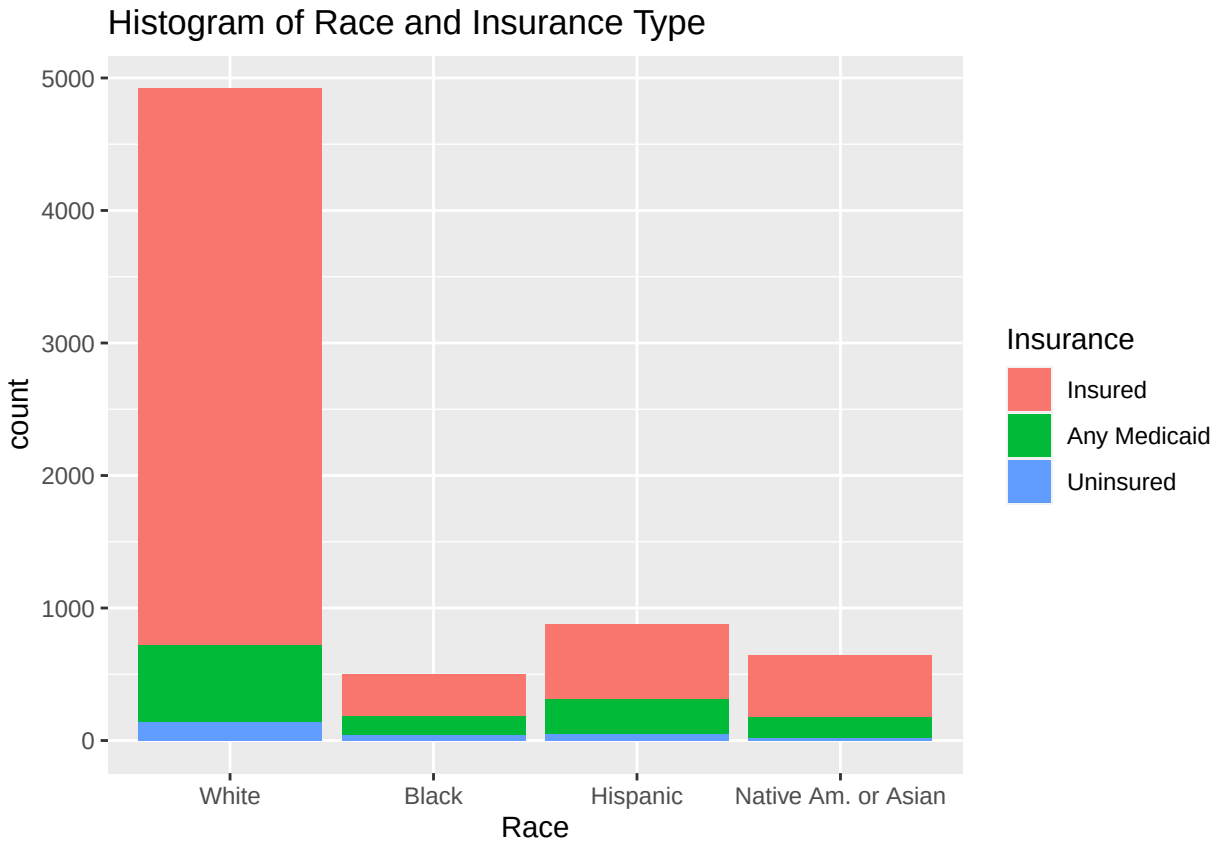
```
##
## Call:
## glm(formula = 1 * (`Surgery Performed?` == "Yes") ~ Sex + Race *
##     Insurance + `AJCC 7 Stage` + `Age at Diagnosis`, family = binomial(link = "logit"),
##     data = oral_cavity)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.013   0.245   0.360   0.566   1.957
##
## Coefficients:
##                                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                    5.490275   0.229599  23.912  < 2e-16 ***
## SexFemale                                      0.153445   0.079462   1.931  0.05348 .
## RaceBlack                                     -0.595328   0.152437  -3.905 9.41e-05 ***
## RaceHispanic                                   0.217396   0.157421   1.381  0.16728
## RaceNative Am. or Asian                        0.357903   0.185033   1.934  0.05308 .
## InsuranceAny Medicaid                         -0.650365   0.126450  -5.143 2.70e-07 ***
## InsuranceUninsured                            -0.873876   0.231456  -3.776  0.00016 ***
## `AJCC 7 Stage`II                              -0.693444   0.135968  -5.100 3.40e-07 ***
## `AJCC 7 Stage`III                             -1.438409   0.127020 -11.324  < 2e-16 ***
## `AJCC 7 Stage`IVA                             -1.705839   0.108755 -15.685  < 2e-16 ***
## `AJCC 7 Stage`IVB                             -2.709452   0.183940 -14.730  < 2e-16 ***
## `AJCC 7 Stage`IVC                             -3.586631   0.213218 -16.821  < 2e-16 ***
## `Age at Diagnosis`                            -0.037152   0.003025 -12.280  < 2e-16 ***
## RaceBlack:InsuranceAny Medicaid                0.159740   0.268503   0.595  0.55189
## RaceHispanic:InsuranceAny Medicaid             0.219706   0.266222   0.825  0.40922
## RaceNative Am. or Asian:InsuranceAny Medicaid  0.140598   0.313728   0.448  0.65404
## RaceBlack:InsuranceUninsured                   0.265022   0.448831   0.590  0.55487
## RaceHispanic:InsuranceUninsured               -0.957513   0.434307  -2.205  0.02748 *
## RaceNative Am. or Asian:InsuranceUninsured     0.061299   0.710291   0.086  0.93123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5575.3  on 6940  degrees of freedom
## Residual deviance: 4701.8  on 6922  degrees of freedom
## AIC: 4739.8
##
## Number of Fisher Scoring iterations: 5
```

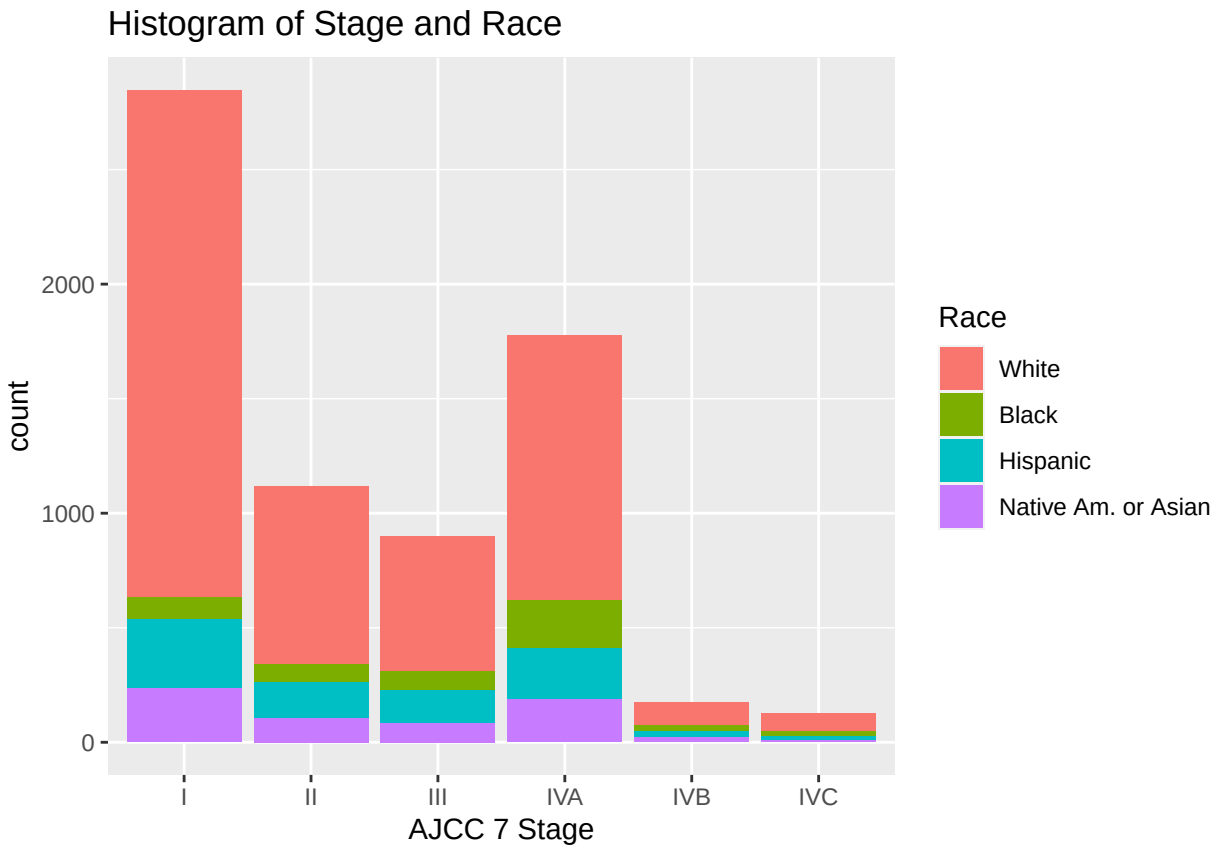
Conclusion/Discussion

The results described above hint at a couple differences (and potentially some biases) in the way that oral cavity cancer patients are treated with or without surgery. First, in each of the logistic regression models

above, each level of the cancer stages were highly significant at alpha = 0.05. The models showed that the later the patient's cancer stage (with IVC being the latest), the less likely the patient is to receive surgery as part of their treatment. This finding was consistent even when adding various interaction terms with different combinations of predictors. However, our primary interest in this analysis was in biases that may arise due to race. In our initial logistic regression model, we found that the Black and Native American or Asian racial groups showed statistically significant coefficients against white patients as the baseline. However, when including interactions between race and insurance or race and cancer stage, these relationships because less pronouced, or disappeared. Importantly, we noticed a slight trend in the interaction between race and cancer stage, where it appears that Black patients with stage IVB cancer may be treated with surgery less often than stage IVB cancer patients of other races. However, this relationship was non-significant at alpha = 0.05, so this cannot be definitely concluded based on the results presented here.

There are some important limitations to consider for this analysis. First, we were working with a fairly unbalanced dataset. Specifically, the vast majority of the patients were white and insured, and there are very few patients with stage IVB or IVC cancer (see Figures in Appendix). Second, we did not take into account the other forms of cancer treatment in this analysis. For example, many patients receive various combinations of radiation and/or chemotherapy in addition or in lieu of surgery. Here our analysis focused only on whether surgery was performed or not, so we cannot draw any conclusions about the overall quality/rigor of care based on these analyses.




Histogram of Stage and Race
count
2000
1000
0
I
II
III
IVA
IVB
IVC
AJCC 7 Stage
Race
White
Black
Hispanic
Native Am. or Asian

Appendix

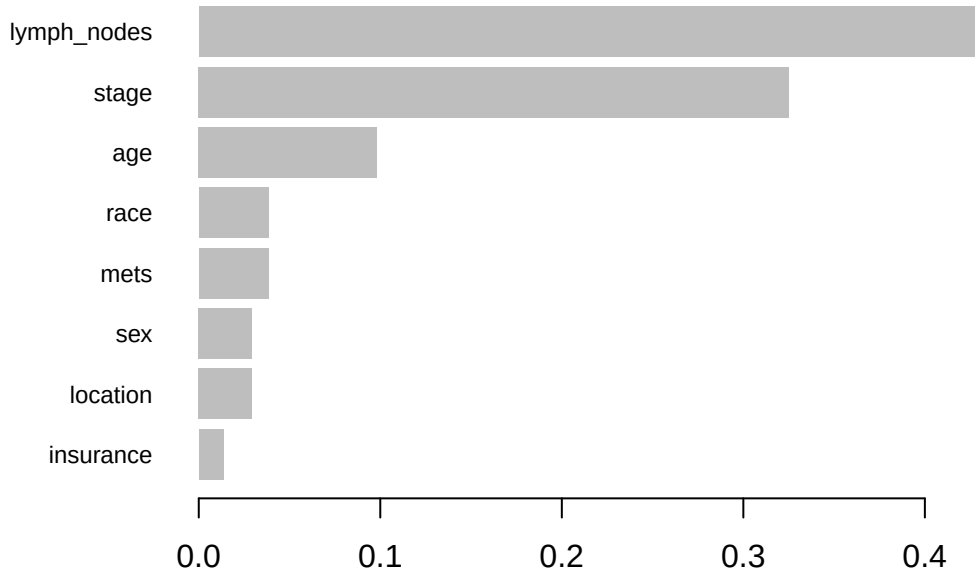


Figure 2: Visualization of the importance matrix for the entire data set

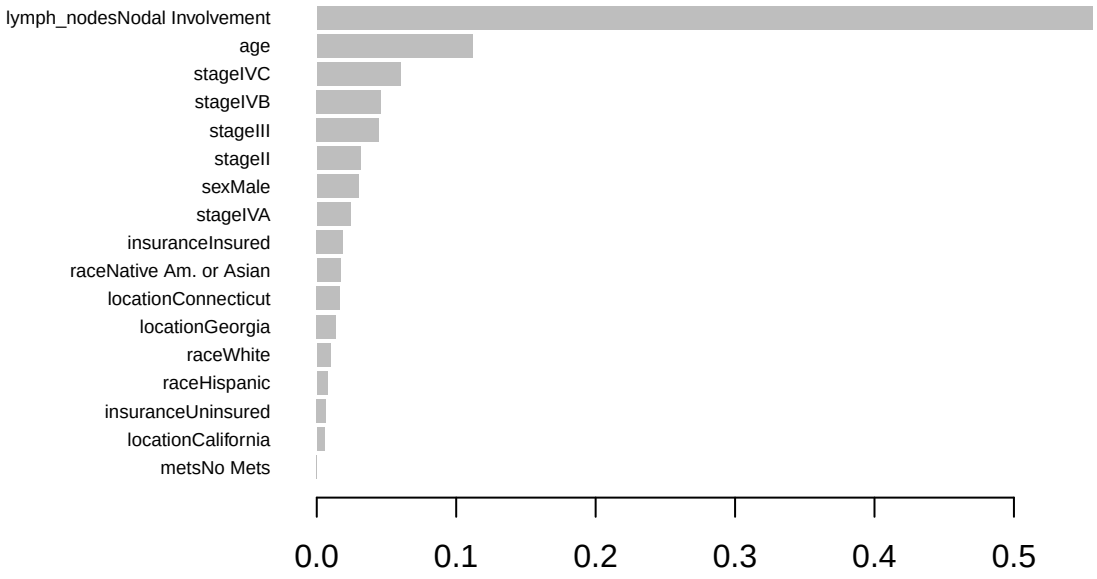


Figure 3: Visualization of the importance matrix using a seeded (2021) subset

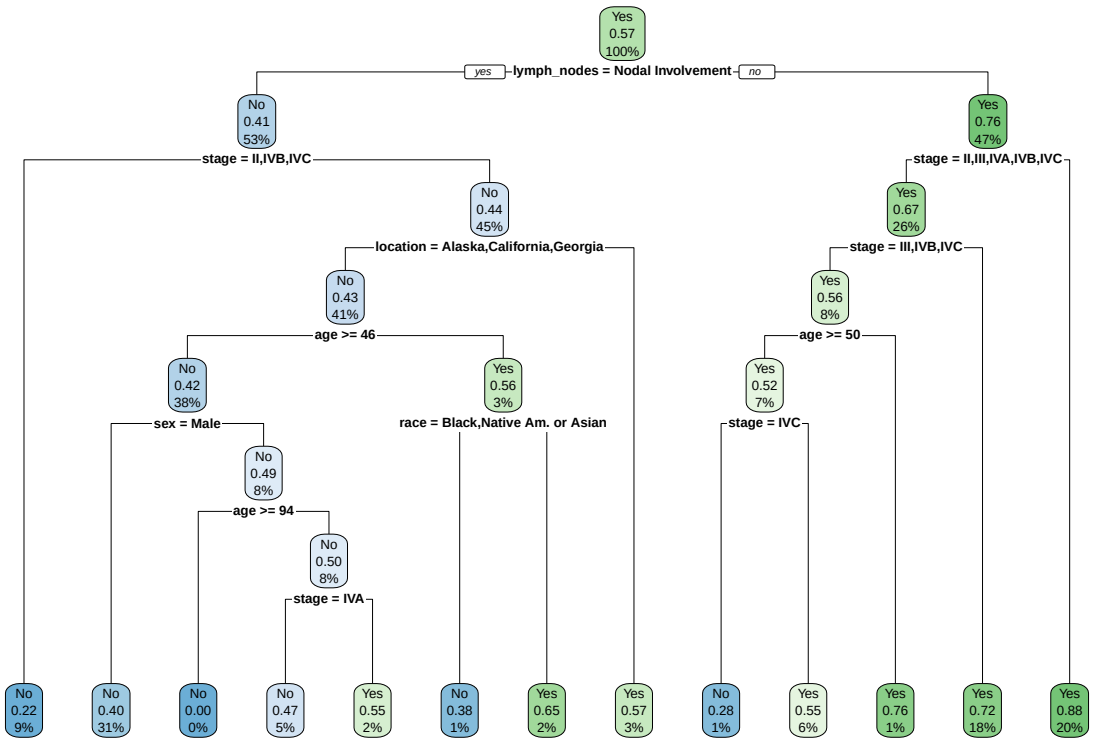


Figure 4: Rpart plot with complexity metric 0.0015 divided the trees similar to the XGBoost algorithm splicing