SEER Analysis on Medical Discrimination

Group12: Zhiwei Liang, Fan Feng, Xiaozhou Lu, Yinfeng Zhou

May 2021

Introduction

The Surveillance, Epidemiology, and End Results (SEER) Program is an authoritative source for cancer statistics in the United States. It provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. SEER is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS).

In this project, we applied statistical analysis on the SEER dataset, in order to identify if there are any discremination on races, genders, ages or income levels during the treatment of head and neck cancers. To achieve our primary goal, we decompose the primary goal into several secondary goals.

The first step of our analysis should be detecting if there is a significant difference between standard therapy and actual given therapy. Then, we will analyse if deviation from standard therapy will cause worse outcomes (such as more deaths). By statistical modelling, we are able to find what kind of people are more likely to be given different therapy from standard, and eventually have worse outcomes.

Dataset Overview and Data Cleaning

In this project, we used two datasets : the SEER dataset and the NCCN guidelines for head and neck cancers.

In the SEER dataset, there are 36 columns with three main aspects: patients' demographic information, average education and income level at the patient's registry region, and information about the tumor and doctor's treatment. In the NCCN guidelines, it provides standard therapy of tumors found in the head and neck, given their T stage and N stage. By referencing the Collaborative Stage Data Collection System Coding Instructions (CS Manual), we transformed the code of 'tumor size' , 'cs extension' into its corresponding T stage, and 'lymph node' into its corresponding N stage.

Primary Site	T Stage	N Stage	Therapy 1	Therapy 2	Therapy 3	Therapy 4	Therapy 5
Oral Cavity	T1	N0	Surgery	Radiation			
Oral Cavity	T1	N1	Surgery	Clinical Trials			
Oral Cavity	T1	N2	Surgery	Clinical Trials			
Oral Cavity	T1	N3	Surgery	Clinical Trials			
Oral Cavity	T2	N0	Surgery	Radiation			
Oral Cavity	T2	N1	Surgery	Clinical Trials			
Oral Cavity	T2	N2	Surgery	Clinical Trials			
Oral Cavity	T2	N3	Surgery	Clinical Trials			
Oral Cavity	T3	N0	Surgery	Clinical Trials			
Oral Cavity	T3	N1	Surgery	Clinical Trials			
Oral Cavity	T3	N2	Surgery	Clinical Trials			
Oral Cavity	T3	N3	Surgery	Clinical Trials			
Oral Cavity	T4a	N0	Surgery	Clinical Trials			
Oral Cavity	T4a	N1	Surgery	Clinical Trials			
Oral Cavity	T4a	N2	Surgery	Clinical Trials			
Oral Cavity	T4a	N3	Surgery	Clinical Trials			

Table 1 : NCCN Guidelines Dataset

Code	Description	TNM 7 Map
000	In situ, intraepithelial, noninvasive	Tis
100	OBSOLETE DATA RETAINED V0200 Invasive tumor confined to one of the following subsites: Inferior wall (superior surface of soft palate) One lateral wall Posterior superior wall (vault)	ERROR
105	Invasive tumor confined to one of the following subsites: Inferior wall (superior surface of soft palate) One lateral wall Posterior superior wall (vault)	T1
200	OBSOLETE DATA RETAINED V0200 Involvement of two or more subsites: Lateral wall extending into eustachian tube/middle ear Posterior, inferior, or lateral wall(s)	ERROR
205	Involvement of two or more subsites: Lateral wall extending into eustachian tube/middle ear Posterior, inferior, or lateral wall(s)	T1
300	OBSOLETE DATA RETAINED V0200 Confined to nasopharynx Localized, NOS	ERROR
305	Confined to nasopharynx Localized, NOS	T1
400	Oropharynx Soft palate, inferior surface including uvula WITHOUT parapharyngeal extension	T1
500	Nasal cavity WITHOUT parapharyngeal extension	T1
505	Extension to soft tissue, NOS (excluding soft tissue of neck)	T1
510	Stated as T1 with no other information on extension	T1

Table 2 : CS Extension Code for Nasopharynx

After transformation, we joined the NCCN guidelines dataset with SEER, so that we had information about standard therapy for each patient. I should be specified that, after joining the two datasets, there is only information in the primary site of Oral Cavity, Oropharynx, Hypopharynx, Nasopharynx and Salivary Gland.

In the dataset, there are columns describing reasons for surgery and radiation. For the reason of surgery, the values are :
[1] "Surgery performed"
[2] "Not recommended"
[3] "Recommended, unknown if performed"
[4] "Not recommended, contraindicated due to other cond; autopsy only (1973-2002)"
[5] "Recommended but not performed, unknown reason"
[6] "Recommended but not performed, patient refused"
[7] "Not performed, patient died prior to recommended surgery"

We decided that the value "Not recommended" and "Not recommended, contraindicated due to other cond; autopsy only (1973-2002)" suggests the doctor didn't recommend surgery, while the others suggest the doctor did. For the reason of radiation, the values are:
[1] "Beam radiation"

[2] "None/Unknown"
[3] "Radiation, NOS method or source not specified"
[4] "Refused (1988+)"
[5] "Recommended, unknown if administered"
[6] "Radioactive implants (includes brachytherapy) (1988+)"
[7] "Combination of beam with implants or isotopes"
[8] "Radioisotopes (1988+)"

We decided that if the value is "None/Unknown" , it suggests that the doctor didn't recommend radiation.

Furthermore, since the sequence of radiation, surgery and chemotherapy is ambiguous in the seer dataset, we simplified the comparison between standard therapy and recommended therapy, by only considering if surgery, radiation and chemotherapy are contained in the standard therapy. For example, if there is surgery in standard therapy, no matter which one it is in, we would denote the column "surgery_standard" as 1. If the doctor also recommended surgery, then the column "*Diff_Reco_Sur*" will be 0, indicating there is no difference between recommended therapy and standard therapy.

Therapy_1	Therapy_2	Therapy_3	Therapy_4	Therapy.5	surgery_standard	surgery_recommended	Diff_Reco_Sur
Radiation	Surgery	Clinical Trials	*NA*	*NA*	1	0	1
Radiation	Surgery	Clinical Trials	*NA*	*NA*	1	0	1
Radiation	Surgery	Clinical Trials	*NA*	*NA*	1	1	0
Radiation	Surgery	Clinical Trials	*NA*	*NA*	1	0	1

Table 3: Comparison between standard therapy and recommended therapy

In conclusion, we managed to join NCCN guidelines standard therapy to the seer dataset, and compared them with recommended therapy. From these information, we were able to create several binary response variables:

1. *Diff_Reco_Sur* : 1 indicates difference in surgery, 0 indicates not.
2. *Diff_Reco_Rad*:1 indicates difference in radiation, 0 indicates not.
3. *Diff_Reco_Chem*:1 indicates difference in chemotherapy, 0 indicates not.
4. *surgery_refused*: 1 indicates the patient refused recommended surgery
5. *radiation_refused*: 1 indicates the patient refused recommended radiation

Exploratory Data Analysis

Demographic Analysis

For our projects focus on exploring if there is any discrimination in the process of treatment for head and neck cancer, we first visualized the relationship between Races and other demographic information containing countywide education level, household income and language isolation level.

First we plotted the distribution between categorical variables: *Sex*, *Race* and *Insurance*. The boxplot below shows that, in different races, it seems that there are some imbalances of gender. There are more male diagnosed head and neck cancers than female in white people.

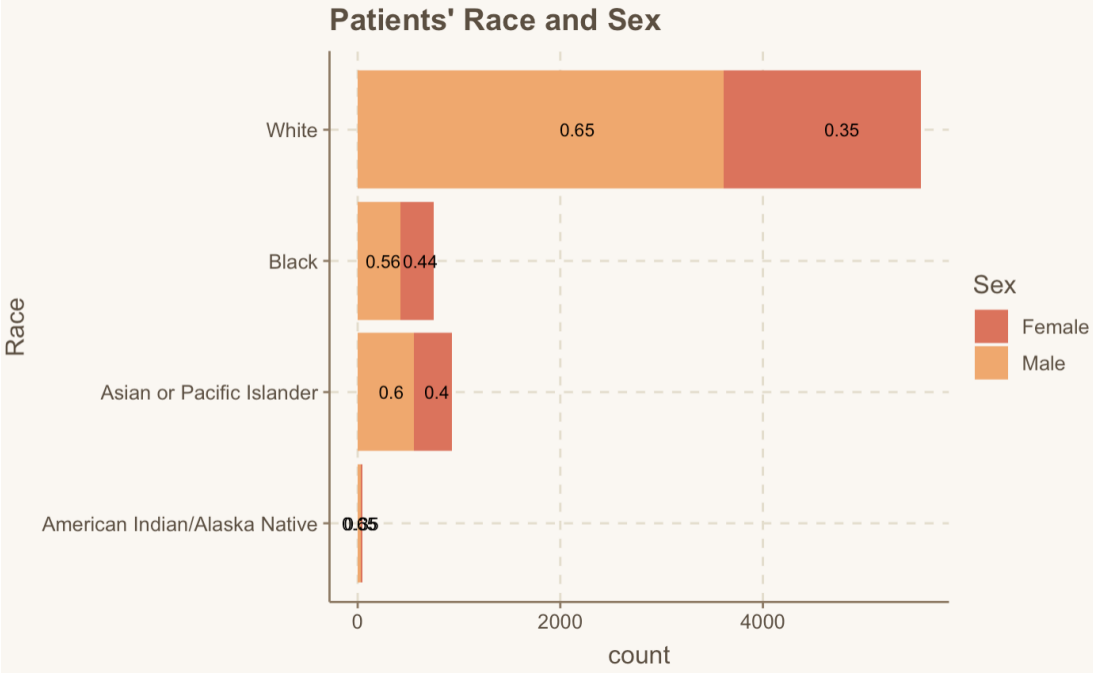

Figure 1: Distribution of gender from different races

In the boxplot below, we can also tell that different races have different distributions of Insurance. For example, white people generally have less Medicaid.

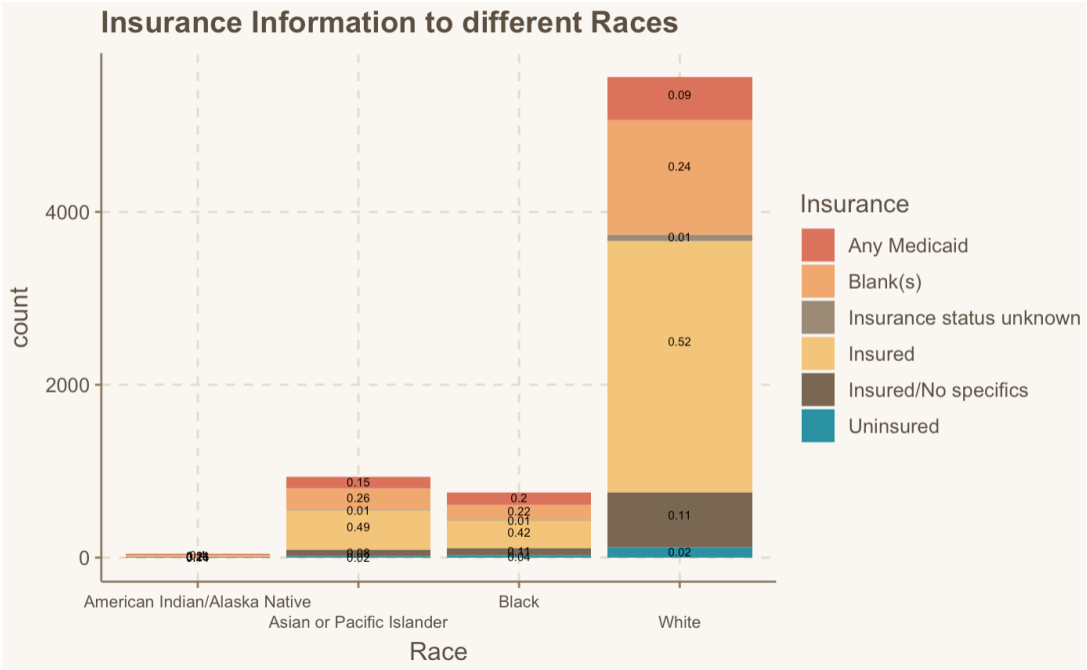

Figure 2: Distribution of Insurance from different races

The density plot below, on the other hand, shows that among head and neck patients, different races have different distributions. In general, patients from white people are older, while Asian or Pacific Islander have the youngest patients.

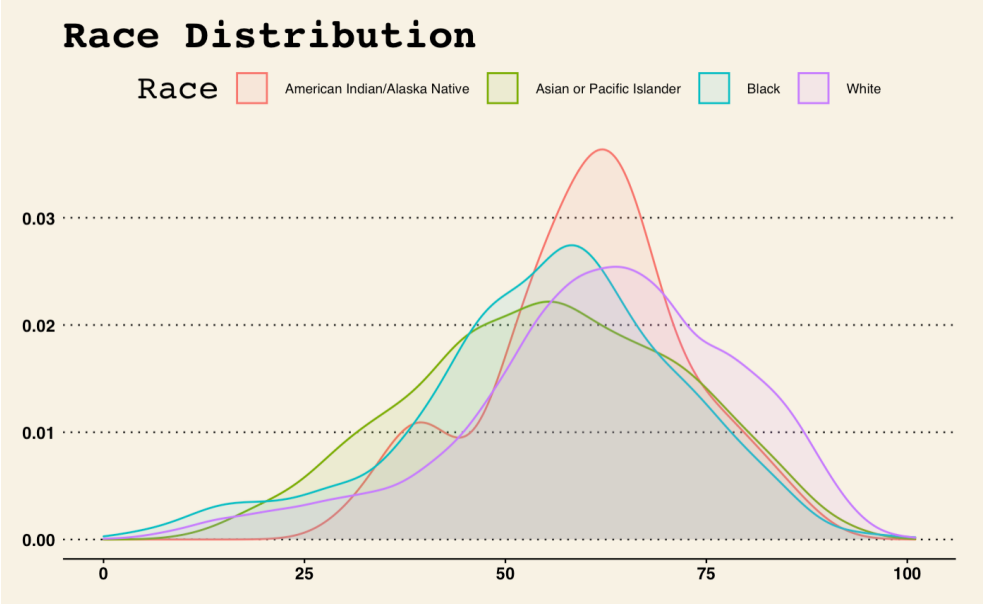

Figure 3: Distribution of Age from different races

In the line chart below, NineGrade stands for average countywide percentage of residents with only ninth grade of education level and the same is true for HighSchool, and Bachelor stands for average countywide percentage of residents with at least a bachelor's degree of education level. BelowPoverty and Unemployed stands for the average countywide percentage of residents' poverty rate and unemployment rate. LanguageIsolation stands for the average countrywide percentage of residents to be linguistically isolated when a household with all members above the age of 14 speak non-English language or speak English less than "very well".

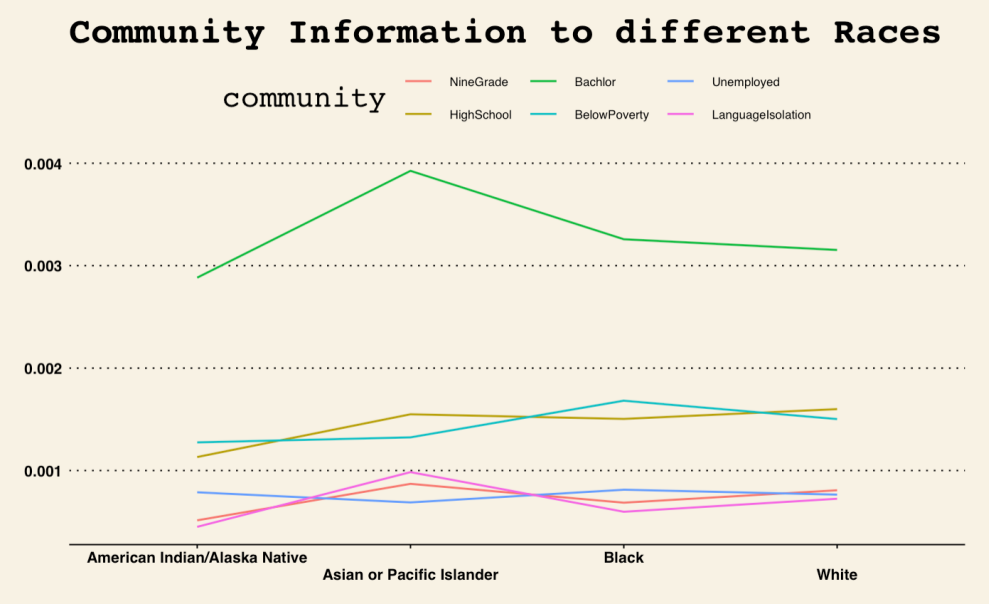

Figure 4 : Community Information Comparison

From the line chart above we can see that *NineGrade*, *HighSchool*, *Bachelor* and *LanguageIsolation* shows a similar pattern towards *Race*. It is a vague impression on how these features vary between different races, so we call this pattern with a relatively higher value for Asian or Pacific Islander and a relatively lower value for some other races like American Indian/Alaska Native or Black pattern 1 for better illustration. While BelowPoverty and Unemployed shows a similar pattern which is opposite from pattern 1, and we call this pattern 2. The income information is also similar to pattern 1 as you can see below, where Income stands for average countywide household income.

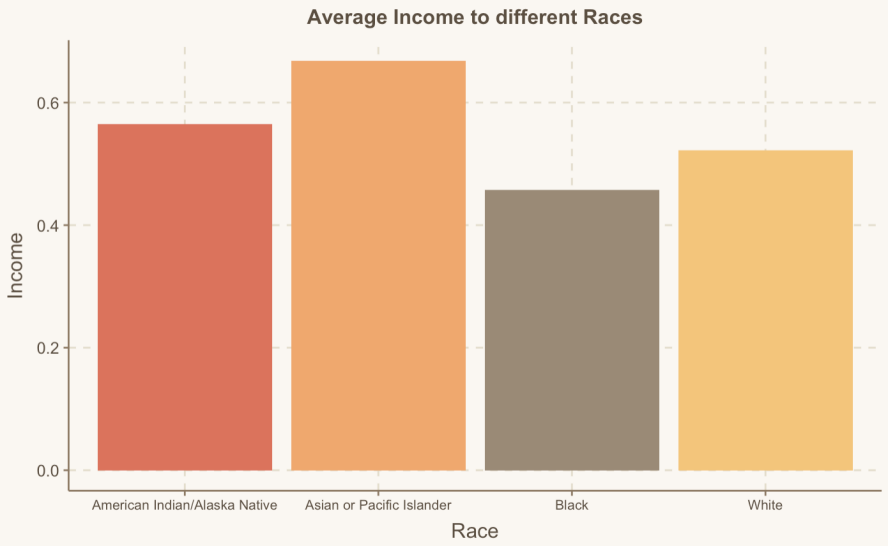

Figure 5 : Household Income Information Comparison

Clinical Information Analysis

To combine the demographic information together with clinical information, we also plotted a box plot comparing the average survival months after being diagnosed between different races. And the result also indicates to be similar to pattern 1 as it is shown in the following box plot.

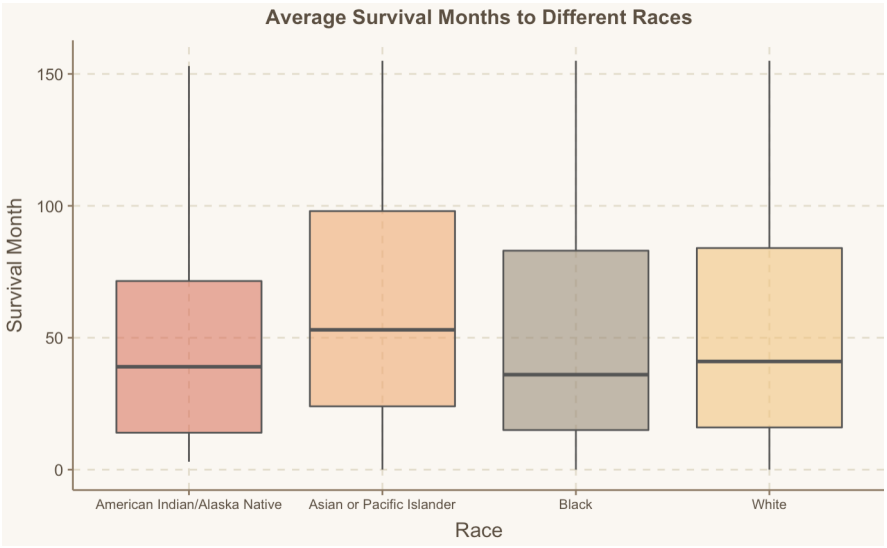

Figure 6: Average Survival Months Comparison

Apart from survival months, we also took the death rate directly attributable to this cancer into consideration. The *Death_Rate* in the next histogram stands for the proportion of death that is directly caused by this cancer over all patients within a race.

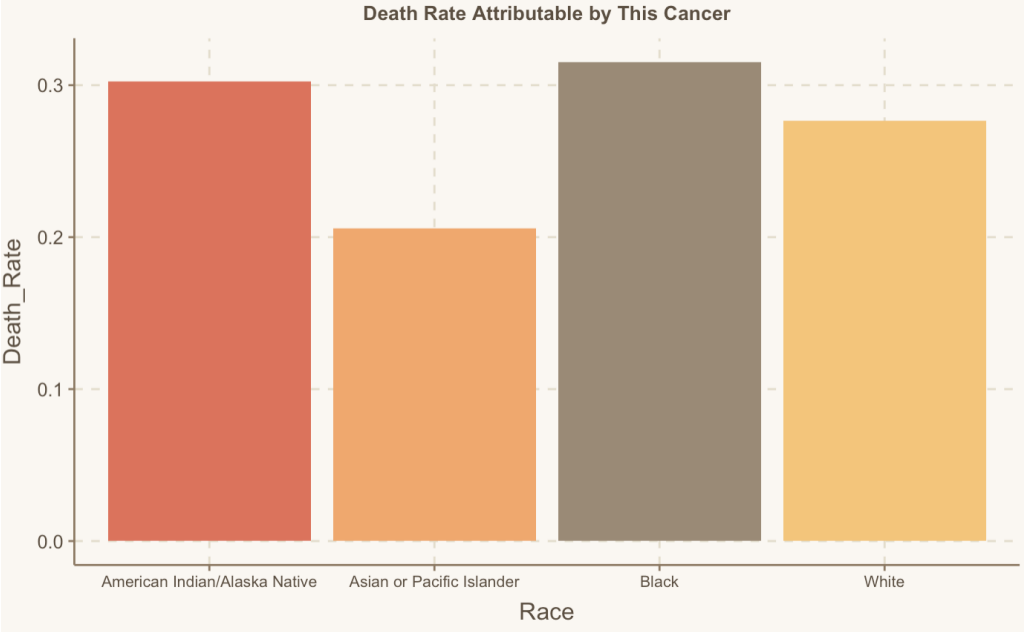

Figure 7: Average Survival Months Comparison

The histogram above shows a similar pattern to pattern 2.

Insights From EDA

In the analysis, we found that there are some correlations between sex, race and insurance. It also indicates some common characteristics, that nearly all the emotionally positive features like education level, household income and survival months shows pattern 1 with Asian or Pacific Islander getting relatively higher values and some other races getting relatively lower value. On the contrary, all the emotionally negative features like unemployment rate, poverty rate and death rate shows pattern 2. These findings give us an indication that it is significative and valuable for us to conduct a deeper research to the assumption that there could be bias or discrimination during clinical treatment for head and neck cancer.

Modeling

Feature Extraction and Model Selection

We expected the predictors represent the patients' socioeconomic status, so we chose *Registry, County, Sex,Year, Age, Insurance,Nine_Grade, High_School,A tleast_Bachlor, Person_Below_Poverty, Unemployed, Median_Household_Income, Language_Isolation.*

However, we worried that these variables were somehow relatively correlated. For example, the variables which represented the education level and income level were the same in one County. To detect and decrease the multicollinearity, we did Correlation Analysis to categorical and continuous variables separately.

For the categorical variables, it seems that there are some correlations among Sex, Race and Insurance. Therefore, we processed the Chi-Squared Test and found that the p values were all <0.05. It seemed like each two of *Race*, *Sex* and *Insurance* were not independent from each other. So, we only chose *Race* as the predictor of our model.

For the numeric variables, we tried several methods like PCA, Stepwise Regression, Ridge Regression, Lasso Regression. As a result, we deleted variable *High_School* and gained the value Kappa 75, which indicated the low multicollinearity.

In summary, we choose *Race, Age, Nine_Grade ,Atleast_Bachlor, Person_Below_Poverty, Unemployed, Median_Household_Income, Language_Isolation* as the predictors.

Considering our dataset has both numeric and nominal values as predictors and binary response, so a dichotomic classifier is a good fit for our analysis. Among all the common classifiers, we choose logistic regression for better interpretation of the coefficients. While other methods might not perfectly cater to our needs. For instance, linear discriminant analysis requires that the observations are drawn from a Gaussian distribution with a common covariance matrix in each class, which is difficult to affirm with nominal variables as predictors. Since our goal is to analysis if there is any discrimination or bias during treamtment process, so it is crucial for us to have a quantitive understanding of the relationship between predictors and the response. Overall, logistic regression is an ideal model for our research.

Topic 1: difference between the recommended and standard therapy

The first step, we wondered whether the doctors had bias towards patients in different socioeconomic status. In particular, our goal was to find which kind of patients are likely given different therapy from NCCN guidelines standard by doctors.

Below are our logistics regression results. The coefficients of *Age* and *Langauge_Isolation* are statistically significant. On the one hand, doctors are more likely to make different surgery decisions from the NCCN guidelines standard on older patients. On the other hand, the lower percent of households in language isolation is, the higher possibility for doctors to make biased surgery decisions. (The Census Bureau considers a household to be linguistically isolated when all members above the age of 14 speak a non-English language and speaks English less than “very well”.)

Similarly, in the *Diff_Reco_Rad* and *Diff_Reco_Chem* model, the coefficient of *Age* is statistically significant. Doctors are more likely to make different radiation and chemotherapy decisions from the NCCN guidelines standard on older patients.

	Sur	p_value	Rad	p_value	Chem	p_value
(Intercept)	-2.3437895	0.0000050	-1.3419360	0.0028096	-1.8754938	0.0002094
RaceAsian or Pacific Islander	-0.4734351	0.2116178	-0.2322991	0.4772652	-0.0291568	0.9378949
RaceBlack	0.3659326	0.3303074	0.1129409	0.7305674	0.3886179	0.2990802
RaceWhite	0.0773989	0.8328599	0.2122223	0.5058669	0.2267101	0.5346286
Age	0.0217951	0.0000000	0.0083921	0.0000000	0.0128869	0.0000000
Nine_Grade	2.2807464	0.3075702	0.9406653	0.6345783	-2.0606709	0.3488137
Atleast_Bachlor	0.9921700	0.1141006	0.3341335	0.5545548	-1.0887807	0.0841187
Person_Below_Poverty	0.9169140	0.4697463	-0.1778469	0.8761546	1.0317778	0.4106067
Unemployed	-3.3908430	0.1101392	1.5730059	0.4096959	-0.7222602	0.7280076
Median_Household_Income	-0.4276413	0.2772385	0.0936843	0.7831731	0.2970716	0.4357298
Language_Isolation	-4.2223617	0.0294704	0.3167227	0.8526933	1.7403540	0.3595659

Table 4 : summary output of three logistic regression model in topic 1

Below is the validation for logistic regression. According to AUC and ROC, the model is not a perfect classifier, but since we are aiming at inference instead of prediction. We focus more on if the coefficients are positive or negative and if it is significant or not. So the result is acceptable for our analysis. The residual plots show some kind of patterns, which will be discussed later in Discussion.

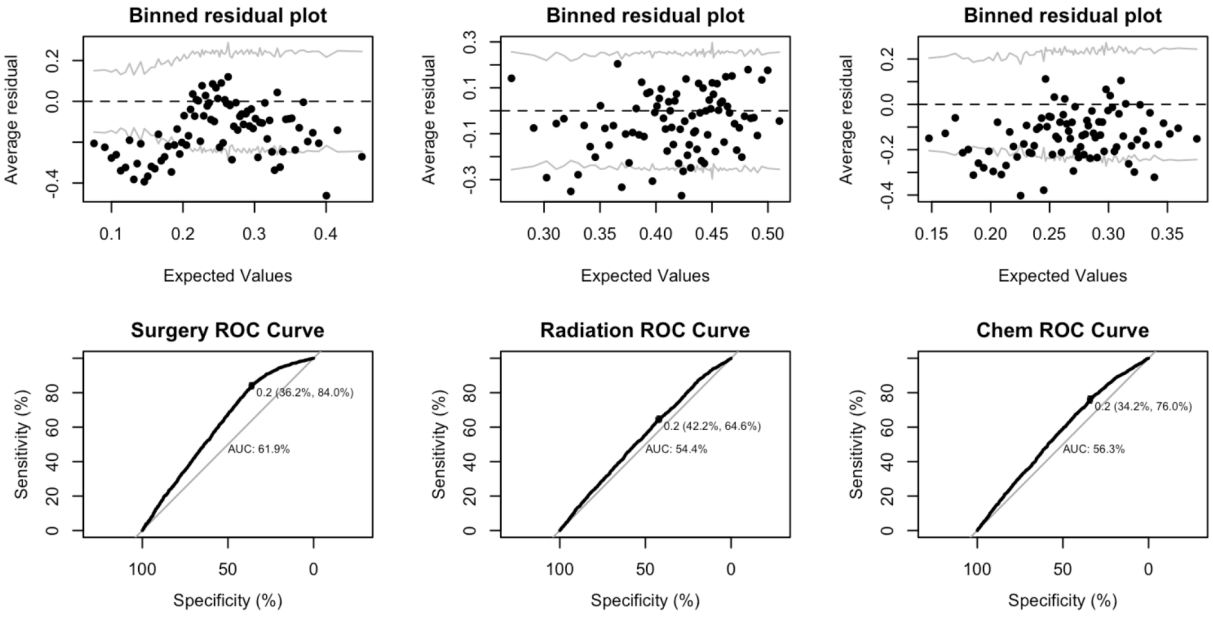

Figure 8: Validation for three logistic regression model in topic 1

Topic 2: difference between the recommended and patient's attitude

Why were there differences between the actual treatment and guideline standard? It not only related to the doctors' bias, but also to the patients' attitude and behaviour. After researching doctors' therapy decisions, we would like to research the patients' attitude towards the doctors' decisions. Specifically, our goal was to define which kind of patients tend to refuse doctors' recommendation of surgery and radiation.

We filtered the subset of SEER data, in which *surgery_recommend* and *radiation_recommend* were TRUE. In other words, we chose the patients who were recommended to take surgery and radiation.

Below are our regression results. The coefficient of *Age* is statistically significant. Older patients are more likely to refuse the doctors' recommendations of surgery and radiation.

	Surgery_refused	p_value	Radiation_refused	p_value
(Intercept)	-6.3678216	0.0020538	-17.3112107	0.9633599
RaceAsian or Pacific Islander	-0.8008826	0.4772449	12.7004285	0.9731143
RaceBlack	-0.1756207	0.8738721	12.8079116	0.9728869
RaceWhite	-1.5072715	0.1577008	12.4981728	0.9735423
Age	0.0551113	0.0000000	0.0227095	0.0016752
Nine_Grade	-17.8953528	0.1451499	2.5317572	0.7513202
Atleast_Bachlor	-2.6047746	0.3865191	1.5839369	0.4750757
Person_Below_Poverty	6.7547999	0.2393254	-3.0322867	0.5210557
Unemployed	-8.9762348	0.3710856	7.1654695	0.3204015
Median_Household_Income	1.4521546	0.4245910	-2.3159989	0.1337215
Language_Isolation	9.1234180	0.3874945	-2.3665228	0.7315947

Table 5 : summary output of two logistic regression model in topic 2

Below is the validation for logistic regression. It's similar to the validation in topic 1. We will discuss more about this problem in the next part of our report.

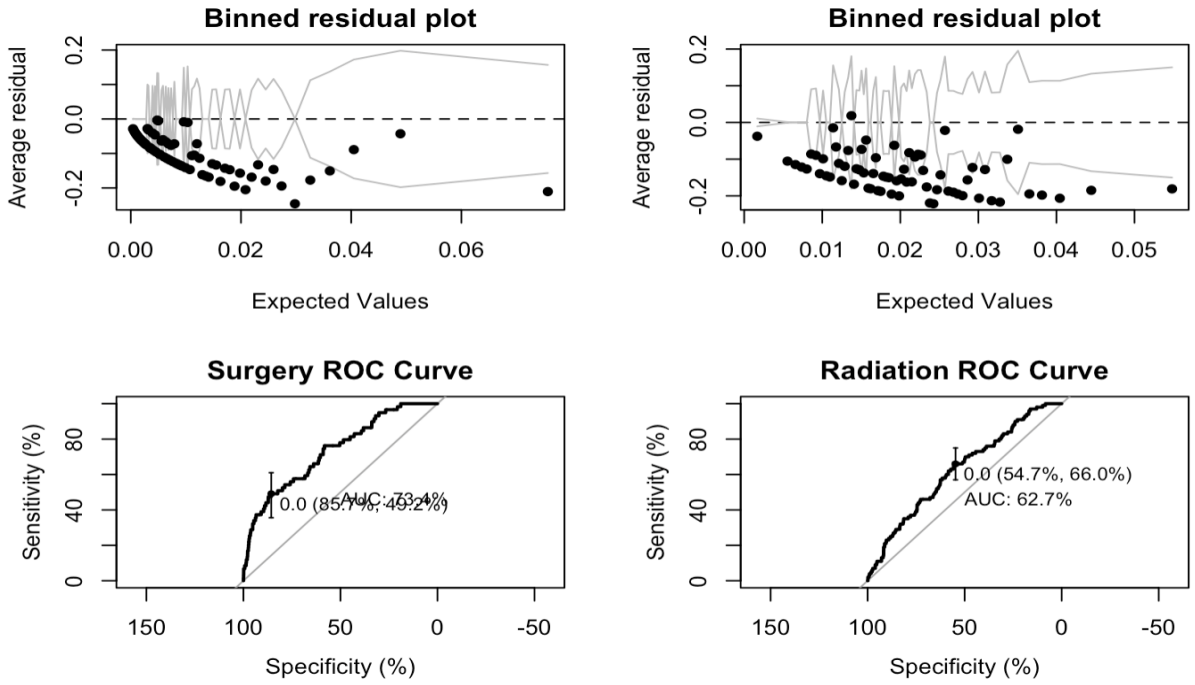

Figure 9: Validation for two logistic regression model in topic 2

Topic 3: Vital Status

The next step is to analyse if deviation from standard therapy will cause worse outcomes. In the dataset, there is a numeric variable called "Survival_Months", indicating how long the patient lived after diagnosis, and another binary variable called "Vital_Status", indicating if the patient died from cancer. In this scenario, we converged on causal mediation analysis.

Causal mediation analysis is able to explain the underlying mechanism between the treatment and the outcome via a mediator. In our case, difference of surgery/ difference of radiation/ difference of chemotherapy should be the treatment, and we are interested in how deviation from standard therapy is related to vital status, through the mediator set as survival months.

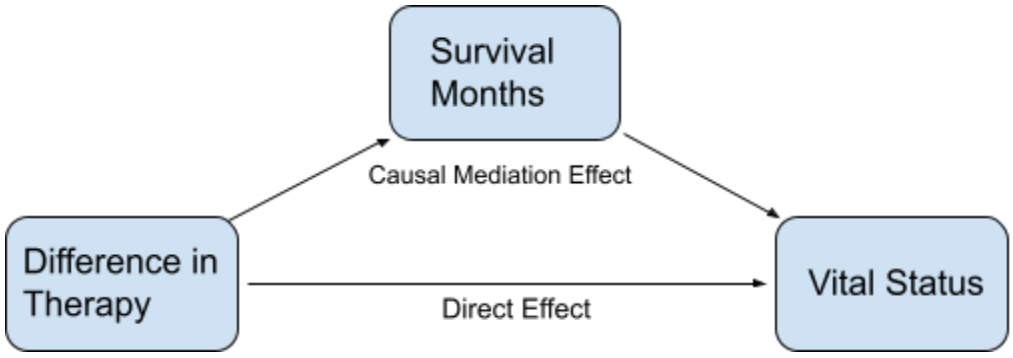

Figure 10: Causal Mediation mechanism diagram

We ran causal mediation analysis with “Diff_Reco_Sur”, “Diff_Reco_Rad”, “Diff_Reco_Chem” as treatment separately. Direct effect is measured by ADE at 5% significance level, and the causal mediation effect (indirect effect) is measured by ACME at 5% significance level. Here is a table showing the results.

Treatment	ACME (Indirect Effect)	ADE(Direct Effect)	Total Effect
Diff_Reco_Sur	significant	significant	significant
Diff_Reco_Rad	significant	non-significant	significant
Diff_Reco_Chem	significant	significant	significant

Table 4: Causal mediation model results

From the results, we can tell that deviation from standard therapy does have some causal effects on the patient’s survival and vital status. Nevertheless, since Age is strongly related to the three treatment variables we used, the real cause might be the Age. For example, when treating an older person, the doctor might be less likely to follow the standard therapy. At the same time, older people in general have a higher death rate, which affects the survival months and vital status.

To validate our inference, we applied a random forest model on the data, to identify the importance of variables to survival months and vital status. From the variable importance plots, we can tell that Age is a dominated important variable, which confirms our hypothesis.

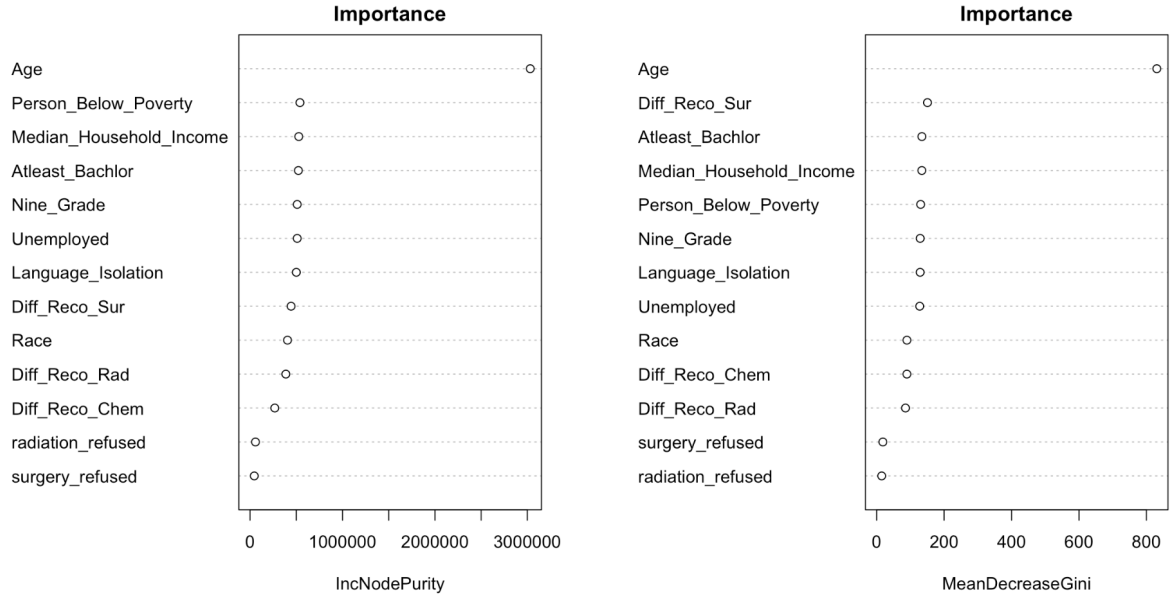

Figure 11: Variable Importance Plot on Survival_Months & Vital_Status

Conclusion

In conclusion, there is no obvious discrimination on races, genders, ages or income levels during the treatment of head and neck cancers.

Specifically, the bias of therapy decisions from standard is more likely related to the patients' age, instead of socioeconomic status like income and education level. Older people are more likely to be given different therapy from standard, and eventually have worse outcomes.

Discussion

In the topic 1 part, we found that language isolation and the different surgery recommendation from NCCN guidelines standard were negatively related. But we cannot say the higher language isolation directly represents the lower socioeconomic status, thus we cannot say that the lower socioeconomic status, the lower the bias of therapy decisions.

There are also some limitations in our analysis. First, the sequence of therapy is not considered in our analysis. In the seer dataset, the sequence of radiation and surgery is way more complicated than that in NCCN guidelines. In NCCN guidelines, there are only two sequences: radiation after surgery and surgery after radiation. However, in the SEER dataset, radiation might be used before and after surgery, or even during surgery. Therefore, it requires more medical knowledge to determine if the actual therapy given is different from the standard guidelines. In addition, since we are not familiar with the TNM stage, there might be some mistakes when identifying the T stage and N stage of the patient. Furthermore, in the modeling part, although we have reduced multicollinearity to an acceptable level, there are still some correlations between variables. It might have affected the model accuracy to some extent.

Moreover, there might have been some patterns that the current model didn't explain. In figure 8 there are binned residuals plots for the three logistic regression models. Overall the points are within the boundaries. Nevertheless, the first plot shows that there might be some higher-order relationships between the response variables and the predictors. The third plot shows the average residuals tend to be negative. Therefore, it requires further research on the dataset and the modeling to figure the relationship between responses and predictors. In figure 9, the residual plot shows a clear pattern, indicating that the link function might be inappropriate. Therefore, the validation of models in topic 2 is questionable.

Apart from residual plot, we also plotted ROC curves and calculated the AUC. According to AUC and ROC, the model is not a perfect classifier with 95% confidence interval lies between 0.6 and 0.7. But since we are aiming at inference instead of prediction. We focus more on if the coefficients are positive or negative and if it is significant or not. So the result is acceptable for our analysis.

Reference

[Multicollinearity Essentials and VIF in R - Articles]
[Multicollinearity in R]
[Correlations Analysis]