The Analysis of Medical Trends in Oropharyngeal Cancer

PUBLISHED
September 29, 2023

Authors

- Prathik Bharath Jain
- Bridget Litostansky
- Stephanie Sarette
- Akarsh Reddy Tatimakula
- Md Razeenuddin Mehdi
- Alex Garofalo

Introduction

According to the American Cancer Society, *"Oropharyngeal cancer is a relatively rare cancer about 53,000 people in the United States develop this cancer each year."* ***("Oropharyngeal Cancer: Symptoms, Stages & Prognosis" n.d.)***

The oropharynx is in the midsection of the throat and along with the nasopharynx and hypopharynx they make up the pharynx section of the throat.

According to the National Cancer Institute (NCI), the most common risk factors for developing this type of cancer is smoking cigarettes for more than 10 packs a year, being infected with HPV especially HPV-16, and a personal history of head and neck cancer. ***("Oropharyngeal Cancer Treatment - NCI" 2023)***

Questions of Interest

Some questions that interested the team as a whole were as follows.

- Is there a primary cancer site that is more common with smokers than non-smokers? What about with patients who do and do not have HPV?
- Does HPV, smoking status, and cancer sites hold influence over what treatment type is used to fight the cancer?
- How successful are the treatment types?
- **Overall question: How does smoking status and HPV affect Oropharyngeal cancer patients?**
  Once the question list was identified, it was broken down into two different visualization sections, one done by Python and one done by Tableau.

Setup:

We will start importing some libraries such as pandas and matplotlib for data visualization.

```
import pandas as pd
import matplotlib.pyplot as plt
import textwrap
import plotly.graph_objects as go
```

About the Dataset:

The dataset chosen for this project comes from the Cancer Imaging Archives website. The dataset is comprised of clinically collected radiation therapy treatment results. It is comprised of 3,346 patients and used 3 different CT scan brand manufacturers to conduct the imaging for these tests. The median patient age for the study is 63 years and is comprised of 80% males and 20% females. ***(Welch et al. 2023)***

Some possible sources of bias found while cleaning the data set was due to the large population of men in the data set. The conclusions that could be drawn may or may not apply to women, as medical treatments affect genders differently. As such, the conclusions drawn have been from the perspective of males. Another source of bias found is the field for how many packs smoked per year was a best guess field, which a person may or may not have been truthful when answering.

Here's a glimpse of the data before data cleaning and manipulation process. The same has been uploaded to the github.

We have added both the original dataset url and the url for the dataset after the data cleaning step.

```
orginal_data_url = 'https://github.com/prathikbafna/Data-Science-as-a-field/blob/main/data.xlsx'
cleaned_data_url = 'https://github.com/prathikbafna/Data-Science-as-a-field/blob/main/dtsc_data.xlsx'
```

Now we are importing the dataset from our local machine.

```
path = './data.xlsx'
```

```
df = pd.read_excel(path)
df.head()
```

	patient_id	Age	Sex	ECOG PS	Smoking PY	Smoking Status	Ds Site	Subsite	T	N	...	Cause of Death	Local	Date Local	Regional	Date Regional	Distant	Date Distant	2nd Ca	Date 2nd Ca	R. cl
0	RADCURE-0005	62.6	Female	ECOG 0	50	Ex-smoker	Oropharynx	post wall	T4b	N2c	...	Other Cause	NaN	NaT	NaN	NaT	NaN	NaT	NaN	NaT	0
1	RADCURE-0006	87.3	Male	ECOG 2	25	Ex-smoker	Larynx	Glottis	T1b	N0	...	Other Cause	NaN	NaT	NaN	NaT	NaN	NaT	NaN	NaT	0
2	RADCURE-0007	49.9	Male	ECOG 1	15	Ex-smoker	Oropharynx	Tonsil	T3	N2b	...	NaN	NaN	NaT	NaN	NaT	NaN	NaT	NaN	NaT	0
3	RADCURE-0009	72.3	Male	ECOG 1	30	Ex-smoker	Unknown	NaN	T0	N2c	...	NaN	NaN	NaT	NaN	NaT	NaN	NaT	S (suspecious)	2008-05-27	0
4	RADCURE-0010	59.7	Female	ECOG 0	0	Non-smoker	Oropharynx	Tonsillar Fossa	T4b	N0	...	NaN	NaN	NaT	NaN	NaT	NaN	NaT	NaN	NaT	0

5 rows × 34 columns

Data Cleaning

We can see that all the column names are not meaningful and have spaces in between.

```
df.columns
```

```
Index(['patient_id', 'Age', 'Sex', 'ECOG PS', 'Smoking PY', 'Smoking Status',
       'Ds Site', 'Subsite', 'T', 'N', 'M ', 'Stage', 'Path', 'HPV',
       'Tx Modality', 'Chemo? ', 'RT Start', 'Dose', 'Fx', 'RT Tech',
       'Last FU', 'Status', 'Length FU', 'Date of Death', 'Cause of Death',
       'Local', 'Date Local', 'Regional', 'Date Regional', 'Distant',
       'Date Distant', '2nd Ca', 'Date 2nd Ca', 'RADCURE-challenge'],
      dtype='object')
```

We will now be renaming the columns to aid the process of manipulating data.

```
df = df.rename(columns = {'T': 'tumor_size', "Smoking PY": "annual_packs_smoked", "Path": "diagnosis_type", "Ds Site": "cancer_site", "Su
```

Now lets assess the null values in the dataset

```
print(df.shape)
print(df.isnull().sum())
```

```
(3346, 34)
patient_id                0
Age                       0
Sex                       0
ECOG PS                   1
annual_packs_smoked       5
smoking_status            0
cancer_site               0
cancer_subsite          374
tumor_size               12
N                        13
M                        14
Stage                    27
diagnosis_type            0
HPV                    1629
treatment_type            0
Chemo?                    0
radio_therapy_startDt     0
Dose                      0
Fx                        0
radio_therapy_type        0
last_follow_up            0
Status                    0
Length FU                 0
date_of_death          2288
cause_of_death         2294
Local                  2966
Date Local             2966
```

```
Regional               3157
Date Regional          3157
Distant                2933
Date Distant           2933
2nd_cancer_site        2905
Date 2nd Ca            2907
RADCURE-challenge         0
dtype: int64
```

Now we try to fix the null values

```
df['smoking_status'].value_counts()
```

```
Ex-smoker      1290
Current        1139
Non-smoker      871
unknown          45
non-drinker       1
Name: smoking_status, dtype: int64
```

Assumption 1: *Since there is only one non-drinker mentioned explicitly in the dataset, we are assuming that the rest of the patients did drink.*

Moreover, as there is only a sole non-drinker among a group of around 4000 patients. We can drop this value since it is almost negligible.

```
df = df[df['smoking_status'] != 'non-drinker']
```

Assumption 2: *We are assuming that the NA values in the HPV columns are patients who didn't get tested. Hence we choose to replace them with 'Not tested'*

```
df['HPV'].fillna('No', inplace = True)
```

Now we assess the columns with different types of tumor

```
df['M '].value_counts()
```

```
M0    3327
MX       2
M1       2
Name: M , dtype: int64
```

Percentage of people with M0 (Benign Tumor)

```
p = (3327/3331)*100
print(p)
```

```
99.87991594115881
```

We can see that 99.8% of the data is about M0 (Benign Tumor). Hence, we can drop the rows that has MX and M1 and remove the entire column and add in the data description that everybody has M0 (Benign Tumor)

```
df.drop('M ', axis = 1, inplace = True)
```

Currently the number of packs smoked is 'NA' for people with smoking status 'Unknown'.

So we choose to replace them with 0.

There are 5 such values in the dataset.

```
df['annual_packs_smoked'].fillna(0, inplace = True)
df['cancer_subsite'].fillna('Unknown', inplace = True)
```

We have two columns:

- **"Dead":** *Saves the status if the patient is alive or not.*
- **"date_of_death":** *Stores the date when the patient passed away.*

We now choose to combine these two columns as an alive person's 'date_of_death' will be null

```
for i in df.index:
    if(df["Status"][i] == "Dead"):
        df["Status"][i] == df["date_of_death"][i]
```

Assumption 3: *We are assuming that patients with NA values in the column "cause_of_death" are still alive.*

```
df['cause_of_death'].fillna('Alive', inplace = True)
```

Now we replace the null values in the 'Local', 'Regional', and '2nd_cancer_site' columns with 'No'. As we would further require them in our data visualization stage.

```
df['Local'].fillna('No', inplace = True)
df['Regional'].fillna('No', inplace = True)
df['2nd_cancer_site'].fillna('No', inplace = True)

df.isnull().sum()
```

```
patient_id                0
Age                       0
Sex                       0
ECOG PS                   1
annual_packs_smoked       0
smoking_status            0
cancer_site               0
cancer_subsite            0
tumor_size               12
N                        13
Stage                    27
diagnosis_type            0
HPV                       0
treatment_type            0
Chemo?                    0
radio_therapy_startDt     0
Dose                      0
Fx                        0
radio_therapy_type        0
last_follow_up            0
Status                    0
Length FU                 0
date_of_death          2287
cause_of_death            0
Local                     0
Date Local             2966
Regional                  0
Date Regional          3156
Distant                2932
Date Distant           2932
2nd_cancer_site           0
Date 2nd Ca            2906
RADCURE-challenge         0
dtype: int64
```

We have around 3346 rows, and in few columns more than 80% of the data is empty.

Hence, we choose to drop those columns in order to further clean the dataset.

```
df.drop(columns = ['Date Local', 'Date Regional', 'Distant', 'Date Distant', 'Date 2nd Ca', 'RADCURE-challenge'])
```

	patient_id	Age	Sex	ECOG PS	annual_packs_smoked	smoking_status	cancer_site	cancer_subsite	tumor_size	N	...	Fx	radio_therapy_type	last_follow_u
0	RADCURE-0005	62.6	Female	ECOG 0	50	Ex-smoker	Oropharynx	post wall	T4b	N2c	...	25	IMRT	2003-05-12
1	RADCURE-0006	87.3	Male	ECOG 2	25	Ex-smoker	Larynx	Glottis	T1b	N0	...	20	IMRT	2007-06-28
2	RADCURE-0007	49.9	Male	ECOG 1	15	Ex-smoker	Oropharynx	Tonsil	T3	N2b	...	40	IMRT	2011-12-05
3	RADCURE-0009	72.3	Male	ECOG 1	30	Ex-smoker	Unknown	Unknown	T0	N2c	...	35	IMRT	2008-05-27
4	RADCURE-0010	59.7	Female	ECOG 0	0	Non-smoker	Oropharynx	Tonsillar Fossa	T4b	N0	...	40	IMRT	2010-05-25
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3341	RADCURE-4126	58.3	Male	ECOG 0	50	Ex-smoker	Oropharynx	Tonsil Pillar	T2	N0	...	35	IMRT	2007-07-10
3342	RADCURE-4127	52.4	Female	ECOG 0	30	Current	Oropharynx	Base of Tongue	T2	N2b	...	35	IMRT	2012-01-17

	patient_id	Age	Sex	ECOG PS	annual_packs_smoked	smoking_status	cancer_site	cancer_subsite	tumor_size	N	...	Fx	radio_therapy_type	last_follow_u
3343	RADCURE-4128	71.3	Male	ECOG 1	50	Ex-smoker	Oropharynx	Soft Palate	T2	N0	...	35	IMRT	2008-04-06
3344	RADCURE-4129	53.9	Female	ECOG 0	5	Current	Oropharynx	Tonsillar Fossa	T2	N2b	...	35	IMRT	2002-02-07
3345	RADCURE-4130	85.5	Male	ECOG 2	15	Ex-smoker	Skin	Nose	T2	N0	...	35	IMRT	2008-11-18

3345 rows × 27 columns

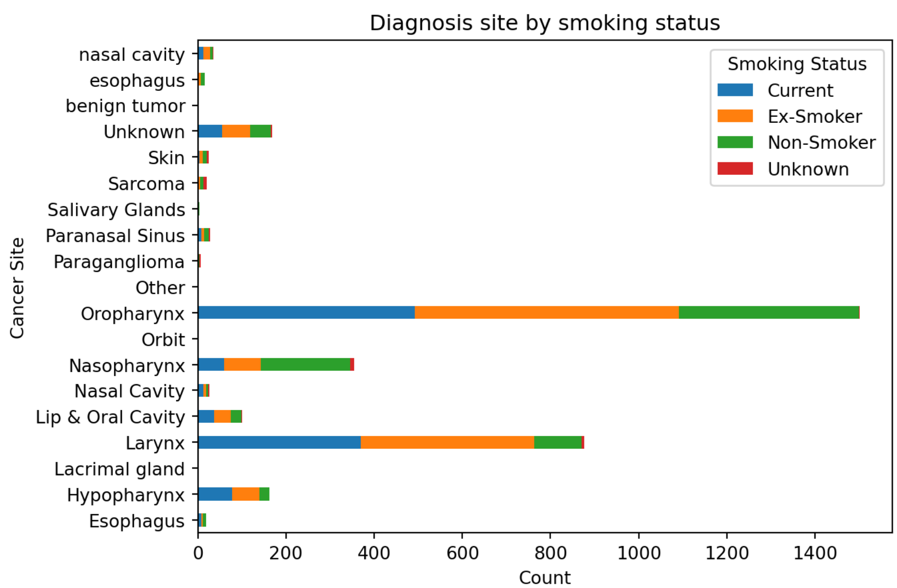
Data Visualization & Analysis:

Figure 1:

The first thing we wanted to look at was if there was a primary cancer site that is more common in smokers than non-smokers.

```
smoking_df = df.groupby('cancer_site').smoking_status.value_counts().unstack()
smokingSite =smoking_df.plot(kind='barh',
        stacked=True,
        title='Diagnosis site by smoking status',
        ylabel = "Cancer Site",
        xlabel = "Count");

smokingSite.legend(['Current', 'Ex-Smoker','Non-Smoker', 'Unknown'], title = 'Smoking Status');
```

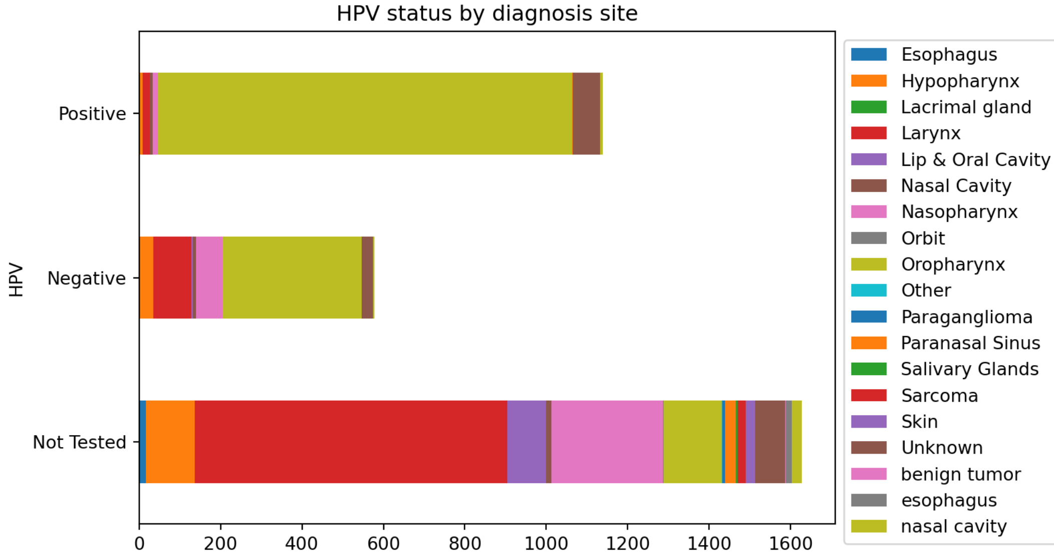


In Figure 1 we can see that the most common cancer sites for smokers and ex-smokers are the Oropharynx and the Larynx, while the Nasopharynx is more common in patients who do not smoke.

Then we looked at how likely HPV cells were present in the cancer based on the location site of the cancer.

Figure 2:

```
hpv_df = df.groupby('HPV').cancer_site.value_counts().unstack()
hpvSite = hpv_df.plot(kind='barh', stacked=True, title='HPV status by diagnosis site')
hpvSite.legend(loc = 'upper left', bbox_to_anchor=(1.0, 1.0))
plt.yticks([0, 1, 2], ['Not Tested', 'Negative', 'Positive']);
```



We also wanted to look at the distribution of treatment types among the patients.

Figure 3:

```
# Look at the treatment type column in the date frame
combined_treatment_type_counts = df['treatment_type'].value_counts()

#Assign colors to each treatment type
adjusted_color_palette = ["#00008B", "#FF4500", "#32CD32", "#FF69B4"]  # Corrected color codes
#Assign descriptions to each treatment type
descriptions = {
    "RT alone": "This represents cases where only Radiation Therapy is used as the treatment method.",
    "ChemoRT": "This represents cases where a combination of Chemotherapy and Radiation Therapy is used for treatment.",  # Corrected key
    "RT + EGFRI": "This represents cases where Radiation Therapy is combined with Epidermal Growth Factor Receptor Inhibitor for treatmen
    "Postop RT alone": "This represents cases where only Postoperative Radiation Therapy is used, typically after a surgical procedure."
}
#Calculate percentage of each treatment type and provide descriptions for each treatment type
labels_with_descriptions = []
for label, count in zip(combined_treatment_type_counts.index, combined_treatment_type_counts):
    pct = 100 * count / combined_treatment_type_counts.sum()  # Percentage
    description = descriptions.get(label, '')  # Matching with description keys
    indented_description = '\n    '.join(textwrap.wrap(description, width=30))
    labels_with_descriptions.append(f"{label} ({pct:.1f}%)\n    {indented_description}.")
```

```
#Create the plot for each treatment type
plt.figure(figsize=(10, 10))
plt.pie(combined_treatment_type_counts, textprops=dict(color="w"), startangle=140, colors=adjusted_color_palette, wedgeprops=dict(width=0
plt.gca().add_artist(plt.Circle((0, 0), 0.70, fc='white'))
plt.legend(loc='upper left', bbox_to_anchor=(1.0, 1.0), labels=labels_with_descriptions)
plt.title('Distribution of Treatment Types')
plt.axis('equal')
plt.show()
```

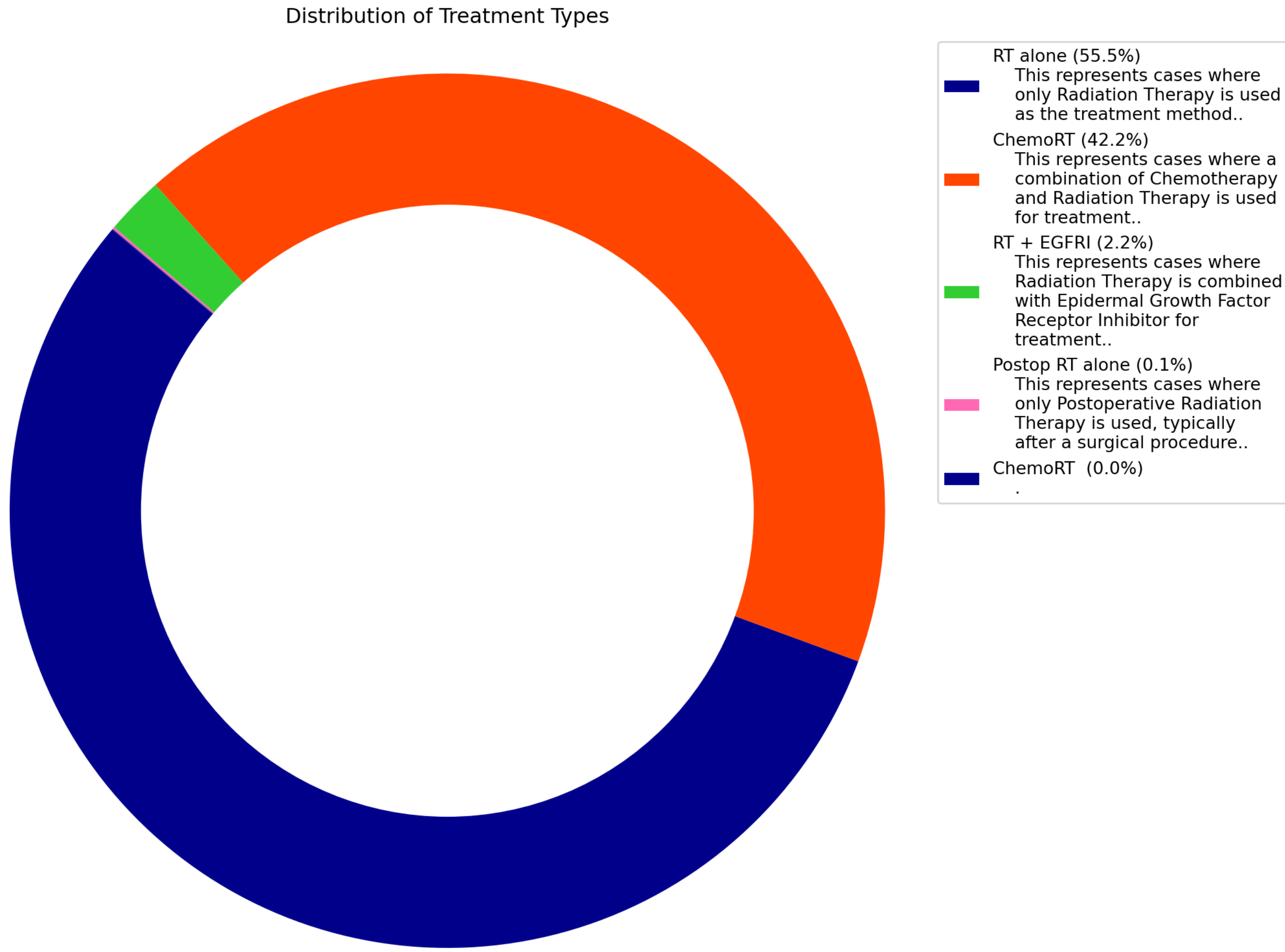


Figure 3 shows that the majority of treaments use only radiation therapy (55.5%). However, there is also a large number of patients who receive chemotherapy along with the radiation therapy (42.2%).

We then took this a step further and looked at how often each treatment type is used based on the site of the cancer.

Figure 4:

```
#Set up
df['treatment_type'] = df['treatment_type'].str.rstrip()
ct = pd.crosstab(df['cancer_site'], df['treatment_type'])
ct_filtered = ct[ct.sum(axis=1) >= 100]
ct_filtered = ct_filtered.drop(index='Unknown', errors='ignore')
ct_filtered.loc['Total'] = ct_filtered.sum(axis=0)
ct_normalized_filtered = ct_filtered.div(ct_filtered.sum(axis=1), axis=0) * 100
ct_normalized_filtered
```

treatment_type	**ChemoRT**	**Postop RT alone**	**RT + EGFRI**	**RT alone**
cancer_site				
Hypopharynx	37.037037	0.617284	4.320988	58.024691
Larynx	10.832383	0.114025	1.026226	88.027366
Lip & Oral Cavity	21.000000	0.000000	5.000000	74.000000
Nasopharynx	87.887324	0.000000	0.000000	12.112676
Oropharynx	51.565623	0.000000	3.197868	45.236509
Total	42.136895	0.066778	2.303840	55.492487

```
# colors
colors_dict = {
    'RT alone': 'black',
    'RT + EGFR': 'red',
```

```
    'RT + ICI': 'purple',
    'RT + Other': 'yellow',
}

# Create figure
fig = go.Figure()

fig = go.Figure()

for treatment_type, color in zip(ct_normalized_filtered.columns, colors_dict.values()):
    fig.add_trace(go.Bar(
        y=[f"{index} (Patients: {ct_filtered.loc[index].sum()})" for index in ct_normalized_filtered.index],
        x=ct_normalized_filtered[treatment_type],
        name=treatment_type,
        orientation='h',
        marker_color=color
    ))

fig.update_layout(
    barmode='stack',
    title_text='How Different Treatments Are Used for Each Cancer Type (Minimum of 100 patients)',
    title_font_size=14,
    xaxis_title_text='Percentage (%)',
    xaxis_title_font_size=12,
    yaxis_title_text='Cancer Site',
    yaxis_title_font_size=12,
    font_size=10,
    legend_title_text='Treatment Type',
    legend_title_font_size=10,
    legend_font_size=8,
)

fig.show()
```

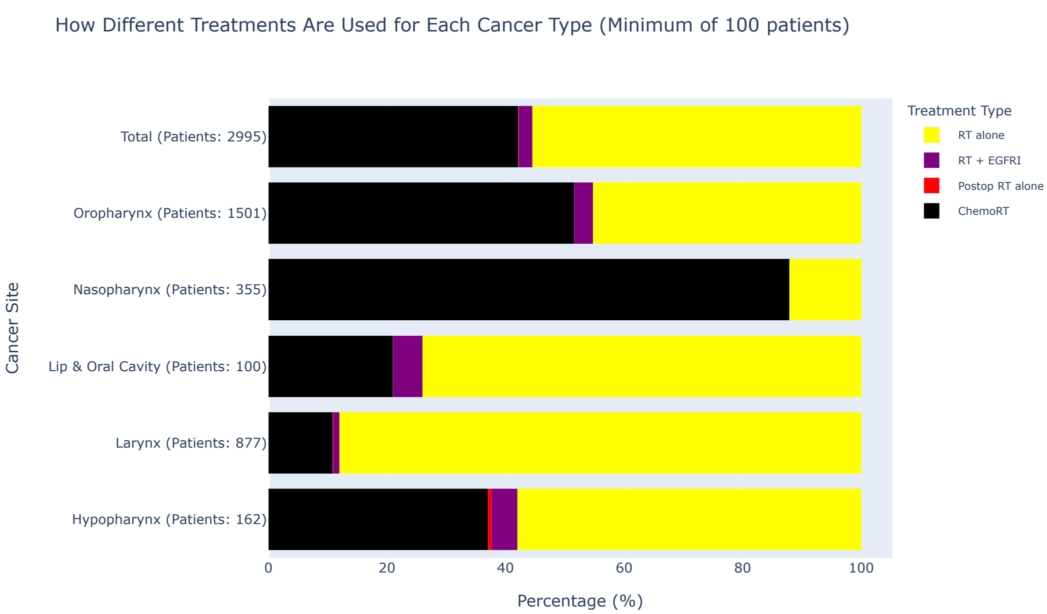


Figure 4 shows that over half the patients with cancer in the oropharynx are treated with both radiation therapy and chemotherapy, but there is also a large amount of patients with cancer in the oropharynx who receive only radiation therapy. However, in the larynx the majority of patients receive only radiation therapy.

Next we looked at the success rate of the treatment indicated by a relapse. There are two kinds of relapse that a patient can have: local and regional. A local relapse occurs when the patient has cancer in the same site as the first time they had cancer. A regional relapse occurs when the patient develops cancer in areas around the original cancer site. We decided to look into if relapse was common among patients with or without HPV for both local and regional relapse.

Figure 5:

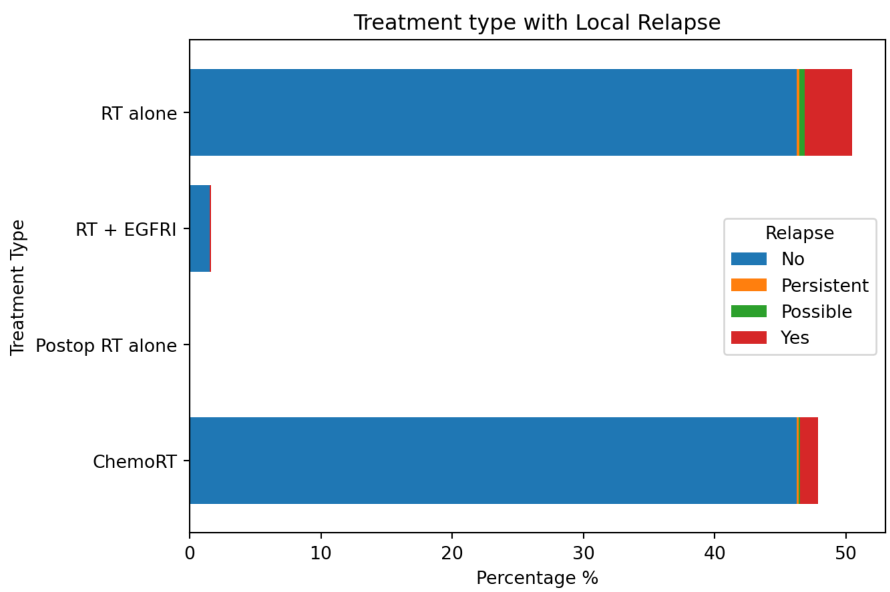
```
# Focus only on HPV testing and local relapse
alive = ['Alive', 'alive']
```

```
mask = df['Status'].isin(alive)
alive_patients = df[mask]
```

```
# Focus only on treatment type and local relapse
local = alive_patients.groupby('Local').treatment_type.value_counts().unstack()
#fill null values with 0
local = local.fillna(0)
local = local.transpose()
```

```
local = pd.DataFrame(local)
local = (local/2287)*100
```

```
# Create the bar graph
local_relapse = local.plot(kind = 'barh',
                           stacked = True,
                           width = 0.75,
                           xlabel = 'Percentage %',
                           ylabel = 'Treatment Type',
                           title = 'Treatment type with Local Relapse')
local_relapse.legend(['No', 'Persistent','Possible', 'Yes'], title = 'Relapse');
```



In Figure 5, we can see that relapse was most common among patients who were only treated with radiotherapy. However, overall local relapse did not occur the majority of the time.

Figure 6:

```
# Percentages of patients who had the cancer return Regionally
regional = alive_patients.groupby('Regional').treatment_type.value_counts().unstack()
# Fill null values with 0
regional = regional.fillna(0)
regional = regional.transpose()
```

```
# Finds the percentage of regional relapse
regional = pd.DataFrame(regional)
regional = (regional/2287)*100
```

```
reg_relapse = regional.plot(kind = 'barh',
                           stacked = True,
                           width = 0.75,
                           xlabel = 'Percentage %',
                            ylabel = 'Treatment Type',
                            title = 'Presence of HPV with Regional Relapse')
reg_relapse.legend(['No', 'Persistent','Possible', 'Yes'], title = 'Relapse')
```

```
<matplotlib.legend.Legend at 0x137178f90>
```

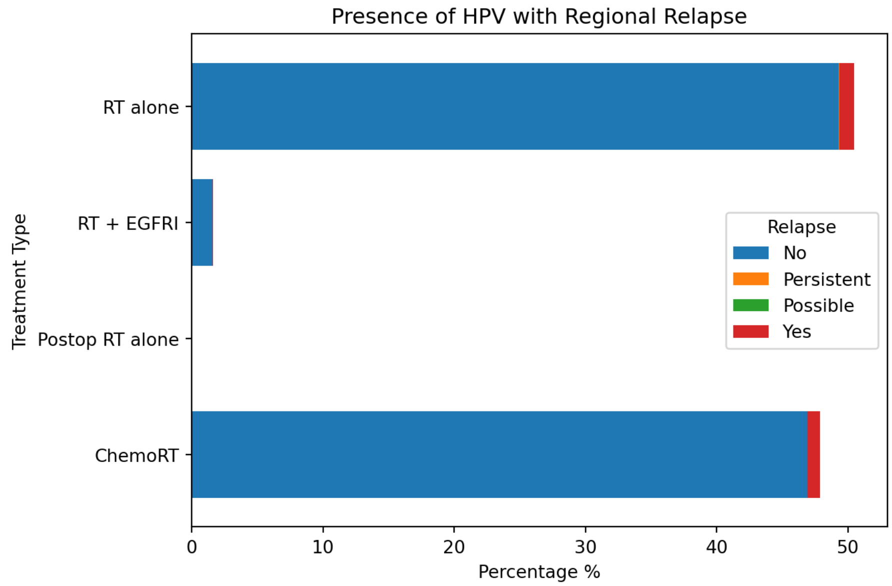


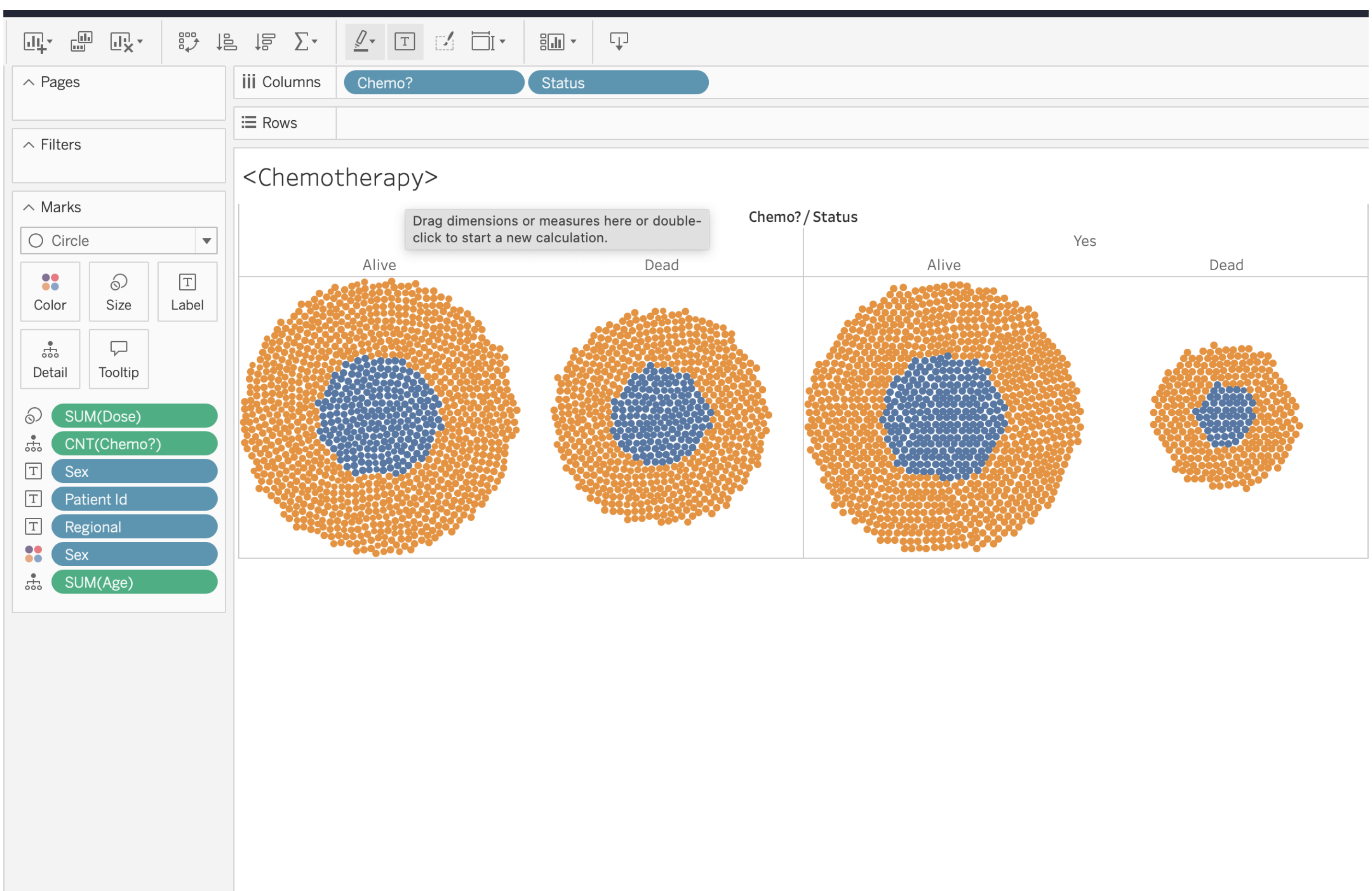
Figure 6 shows that regional relapse is not very common with any of the treatment types. Compared to Figure 5, we can observe that regional relapse is less common than local relapse.

Tableau Visualization:

Figure 1:

In the first visualization we analyze the status of the patients who went through Chemotherapy and if they are alive or have passed away.

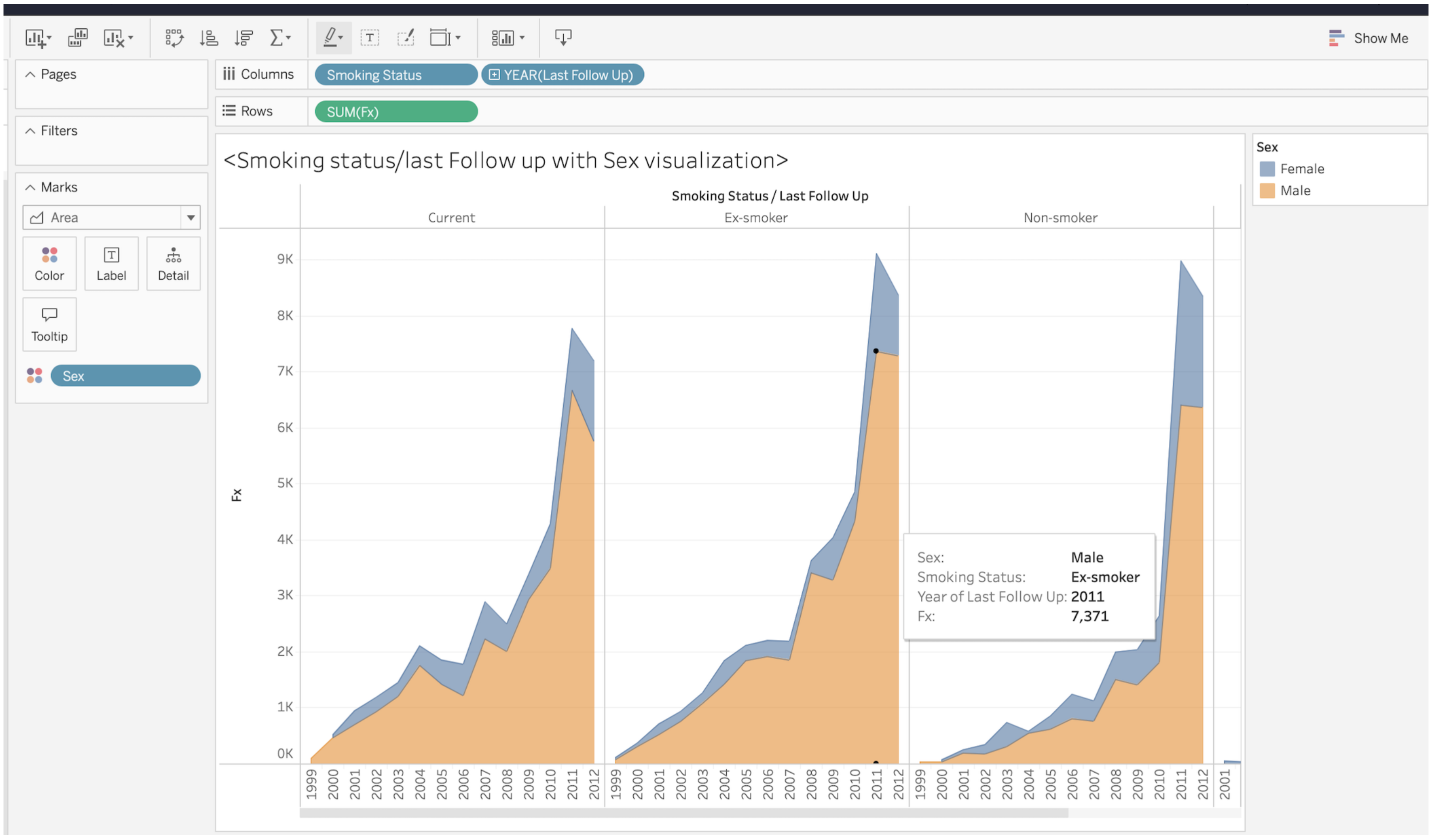
We also assess how many people went through chemotherapy as it is a rather dificult decision to make.



Chemotherapy vs Status

Figure 2:

In Figure 2 we analyze the last follow up date of patients who engaged in smoking.



Smoking Status vs Last Follow Up Date

Figure 3:

In Figure 3 we analyze what sort of treatments did the patient receive and their smoking status. We visualize this data in a tabular format.

Sheet 1

		Smoking Status			
Sex	Treatment Type	Current	Ex-smok..	Non-smo..	unknown
Female	ChemoRT	4,760	4,773	10,070	606
	RT + EGFRI	350	210	210	
	RT alone	9,295	7,922	6,564	841
Male	ChemoRT	25,102	28,714	23,866	690
	Postop RT alone	60	116		
	RT + EGFRI	840	2,164	1,252	
	RT alone	35,610	41,132	17,084	748

Treatment Type vs Smoking Status

Conclusion & Bias:

Since the patient data was 80% male, the conclusions we have gained from the analysis of the dataset will mostly apply to men. However, according to the American Cancer Society, oropharyngeal cancer is twice as likely to occur in men than in women (“Risk Factors for Oral Cavity and Oropharyngeal Cancers”). It is possible that this is due to men historically being more likely to use tobacco products ***(“Risk Factors for Oral Cavity and Oropharyngeal Cancers” n.d.)***.

The second Tableau visualization shows that among the female patients, more of them were non-smokers and ex-smokers than current smokers, while the men had more current smokers and ex-smokers than non-smokers. Oral HPV is also most common in men, as 10% of men are exposed to it during their life, while only 3.6% of women are exposed to it during their lifetime ***("HPV and Oropharyngeal Cancer CDC" 2023)***. While the conclusions we draw have a bias towards men because our patient sample is mostly male, this is understandable due to Oropharyngeal cancer being more common in men than women.

From the above analysis, we can conclude that the most common site for Oropharyngeal cancer is the oropharynx (python Figure 1). This is particularly common for patients who are current smokers or ex-smokers (python Figure 1). This was also a very common site for patients who tested for HPV (python Figure 2). It did not matter if the test result was positive or negative, the oropharynx was the dominant cancer site for both results. HPV affects the throat and mouth, and is the cause of up to 70% of Oropharyngeal cancer cases ***("HPV and Oropharyngeal Cancer CDC" 2023)***, so it is not surprising that patients with cancer in the oropharynx would be tested for HPV. Further analysis could be done for patients with cancer in the oropharynx to compare HPV status and their smoking status. The larynx was the second most common site for Oropharyngeal cancer (python Figure 1). This was also a common site for patients who have smoked in their life (python Figure 1) and it was the most common site for patients not tested for HPV (python Figure 2), indicating that this cancer site is most likely to occur because of smoking status. The third most common site was the nasopharynx, which was more common among patients who have never smoked (python Figure 1). This was also not a very common site for patients with HPV (python Figure 2). Since this site was common among patients who did not smoke and did not have HPV, the cause of the cancer here could be genetic. The common causes of Oropharyngeal cancer have influence over the cancer site.

The distribution of treatment types revealed that radiotherapy and chemo-radiotherapy were the two most utilized treatment types for Oropharyngeal cancer, with radiation therapy alone being the most common treatment type (python Figure 3). Two of the three most common cancer sites had a more common treatment type. The Larynx, which was common among smokers, was mostly treated with only radiotherapy, while the Nasopharynx, which was common among non-smokers, was mostly treated with chemo-radiotherapy (python Figure 3). Chemo-radiotherapy and only radiotherapy were both common to treat cancer in the oropharynx (python Figure 3). Since, experts decide how to treat a patient based on how quickly a cancer is spreading ***("Cancer Treatment: Radiation Therapy Versus Chemotherapy" 2021)***, we can conclude that cancer in the Larynx does not spread very quickly since typically only radiotherapy is used to treat the cancer. Since this is a common site among smokers, it is possible that if smoking is the cause of cancer, the cancer remains isolated. We also observed that for men and women, smokers and ex-smokers were treated by RT alone more than non-smokers (Tableau Figure 3). ChemoRT was used a lot more to treat non-smokers than smokers and ex-smokers (Tableau Figure 3). Since cancer in the nasopharynx was mostly treated with chemo-radiotherapy, the cancer there most likely spreads quickly. Treatment for cancer in the oropharynx, seemed to be dependent on each case since chemoradiotherapy and radiotherapy alone were both very commonly used. Further study could be done to check if smoking and HPV status had influence over the decision to use chemoradiotherapy or only radiotherapy as treatment for cancer in the oropharynx.

Our analysis indicated that treatment type had influence over the likeliness of the survival rate of the patient. Fewer patients who were treated with chemoradiotherapy passed away than patients who were not treated with chemoradiotherapy (Tableau Figure 1). However, there are many causes of death, so we are making the assumption that Oropharyngeal cancer is the cause of death when we draw this conclusion. An alternative method to looking into the success of treatment is if there was a relapse. In general, both local and regional relapses were not very common, but local relapse was more common than regional relapse (python Figures 5 and 6). Local relapse was also more common with only radiotherapy as a treatment (python Figure 5). However, these conclusions are based on the assumption that no data for the patient indicates that they have not relapsed. Patients who have passed away were removed from the data set for this analysis. It is also possible that a patient relapsed after the last patient check in, so it would not have been recorded. Treatments appear to be successful for patients who survived the cancer, and chemo-radiotherapy seems to improve these chances.

Oropharyngeal cancer has several causes including HPV and smoking. These causes can influence where the cancer occurs. Cancer in the oropharynx seems to be caused by HPV and/or smoking, cancer in the larynx seems to be caused by smoking, and cancer in the nasopharynx seems to be caused by other reasons. Radiotherapy and chemo-radiotherapy are the two most common treatments to treat Oropharyngeal cancer. Cancer in the larynx is mostly treated with radiotherapy, cancer in the nasopharynx is mostly treated with chemo-radiotherapy, and cancer in the oropharynx is treated with both. Less patients died and relapsed when chemo-radiotherapy was used to treat the Oropharyngeal cancer. However, the majority of patients who survived treatment, did not relapse by the time the study was concluded. It is important to identify these trends and biases within the dataset, as it can help improve the chances of survival and help identify treatment types for future patients who develop Oropharyngeal cancer.

Citations:

"Cancer Treatment: Radiation Therapy Versus Chemotherapy." 2021. *Lindenberg Cancer & Hematology Center Marlton, NJ 08053*. https://lindenbergcancer.com/blog/cancer-treatment-radiation-therapy-versus-chemotherapy/.

"HPV and Oropharyngeal Cancer CDC." 2023. https://www.cdc.gov/cancer/hpv/basic_info/hpv_oropharyngeal.htm.

"Oropharyngeal Cancer Treatment - NCI." 2023. {pdqCancerInfoSummary}. https://www.cancer.gov/types/head-and-neck/patient/adult/oropharyngeal-treatment-pdq.

"Oropharyngeal Cancer: Symptoms, Stages & Prognosis." n.d. *Cleveland Clinic*. Accessed September 20, 2023. https://my.clevelandclinic.org/health/diseases/12180-oropharyngeal-cancer.

"Risk Factors for Oral Cavity and Oropharyngeal Cancers." n.d. Accessed September 29, 2023. https://www.cancer.org/cancer/types/oral-cavity-and-oropharyngeal-cancer/causes-risks-prevention/risk-factors.html.

Welch, Mattea L., Sejin Kim, Andrew Hope, Shao Hui Huang, Zhibin Lu, Joseph Marsilla, Michal Kazmierski, et al. 2023. "Computed Tomography Images from Large Head and Neck Cohort (RADCURE)." The Cancer Imaging Archive. https://doi.org/10.7937/J47W-NM11.