Impact of the Covid-19 Pandemic on Cancer Patients

Kari Lauro
SI 618, Data Gathering and Processing Report

Motivation

After losing a family member to cancer this past year due to cancer after a mild Covid-19 infection, which led to a delay in treatment, unfortunately allowing for the advancement of the disease, I wanted to explore any potential relationships between Covid-19 and cancer. What happened to patients battling cancer when the pandemic was in full swing, and our healthcare system was overloaded, causing delays in many chronic treatments put on hold so staff could focus on those suffering from this new virus? How were outcomes shifted for patients who unfortunately contracted Covid-19 as they received cancer treatment?
To further uncover an answer to my question, I created eight subquestions but found that the following three gave the most insight into my overarching question:

1. Was there any relationship between age and gender and mortality rate pre and post-pandemic?
2. When diagnosed, was there a relationship between tumor size post-pandemic with limited healthcare resources?
3. How was the timing between the decision to treat and the first treatment change from pre-pandemic to post-pandemic? What about the time from the initial diagnosis to the first treatment?

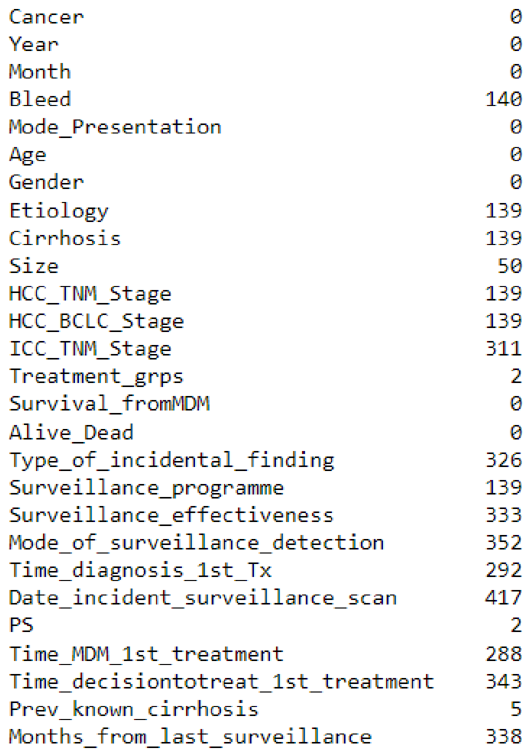
Data Source

COVID-19 effect on Liver Cancer Prediction Dataset
Retrieved from Kaggle and can be accessed here https://www.kaggle.com/datasets/fedesoriano/covid19-effect-on-liver-cancer-prediction-dataset.
Exported in a .csv format and houses 27 columns with 451 rows, this dataset contains retrospective data for liver cancer patients who presented 12 months before the pandemic (March 2019-February 2020) and prospective data for patients who presented in the first 12 months of the pandemic (March 2020-February 2021). Attributes of this dataset include age, gender, year [pre-pandmic(March 2019-February 2020), post-pandemic (March 2020-February 2021)], month, mode presentation (incidental, surveillance, symptomatic) tumor node metastasis stage (I, II, IIIA+IIIB, IV), etiology, tumor size, mortality status, staging, etc. Of note, this dataset does not contain any data regarding whether or not the patient contracted the Covid-19 virus during the measured pandemic year.

Methods

Manipulation
This data set was read into python using pandas.read_csv and df.head() to view the first five rows of the dataset. From here, I could see that there were columns that appeared to be missing

data before I could simply drop them using df.dropna(), I would need to determine how many missing values there were to decide which effect dropping them would have on my calculations. To do this, I ran the code `df.isna().sum()`, which provided me with the following results.



```
Cancer                              0
Year                                0
Month                               0
Bleed                             140
Mode_Presentation                   0
Age                                 0
Gender                              0
Etiology                          139
Cirrhosis                         139
Size                               50
HCC_TNM_Stage                     139
HCC_BCLC_Stage                    139
ICC_TNM_Stage                     311
Treatment_grps                      2
Survival_fromMDM                    0
Alive_Dead                          0
Type_of_incidental_finding        326
Surveillance_programme            139
Surveillance_effectiveness        333
Mode_of_surveillance_detection    352
Time_diagnosis_1st_Tx             292
Date_incident_surveillance_scan   417
PS                                  2
Time_MDM_1st_treatment            288
Time_decisiontotreat_1st_treatment  343
Prev_known_cirrhosis                5
Months_from_last_surveillance     338
```

Missing, incomplete, noisy data

This dataset contains a significant portion of missing data, which required careful attention to ensure that I was not including these values or skewing the results by simply deleting them all. To control for these missing values, I had to create multiple databases that I could use to complete my calculation and visualizations.

Challenges

A significant portion of methods that I had initially planned to use, noted in my proposal, ended up not working out as intended. I had to adjust some of my calculations and visualization methods to ensure things were being calculated and displayed correctly and in a readable way. Additionally, without having any data regarding Covid-19 diagnosis status, correlations drawn from these analyses must also be mindful of the potential confounding factors of the immunocompromised patients contracting the disease.

Analysis and Results

Q1. Was there any relationship between age and gender and mortality rate pre and post-pandemic?

Performing the analysis

To conduct this analysis, I first created a dataset with NaN values filtered out for the *Surival_fromMDM* attribute. Using this filtered dataset, I made a bar graph showing the overall survival rate before and after the pandemic. I then broke it down by age using a linear modeling plot and gender using another bar plot.

Results

As I had initially expected, there was an overall decrease in the survival rate, as seen in *Figure 1.* The survival rate was almost cut in half during the pandemic. Breaking it down by age revealed the same pattern, with the survival rate cut just about in half and decreasing with patient age (*Figure 2)*. Interestingly, however, when analyzing the survival rate by gender, the survival rate did fall during the pandemic but dropped much lower for females than males (*Figure 3)*.

Visualization

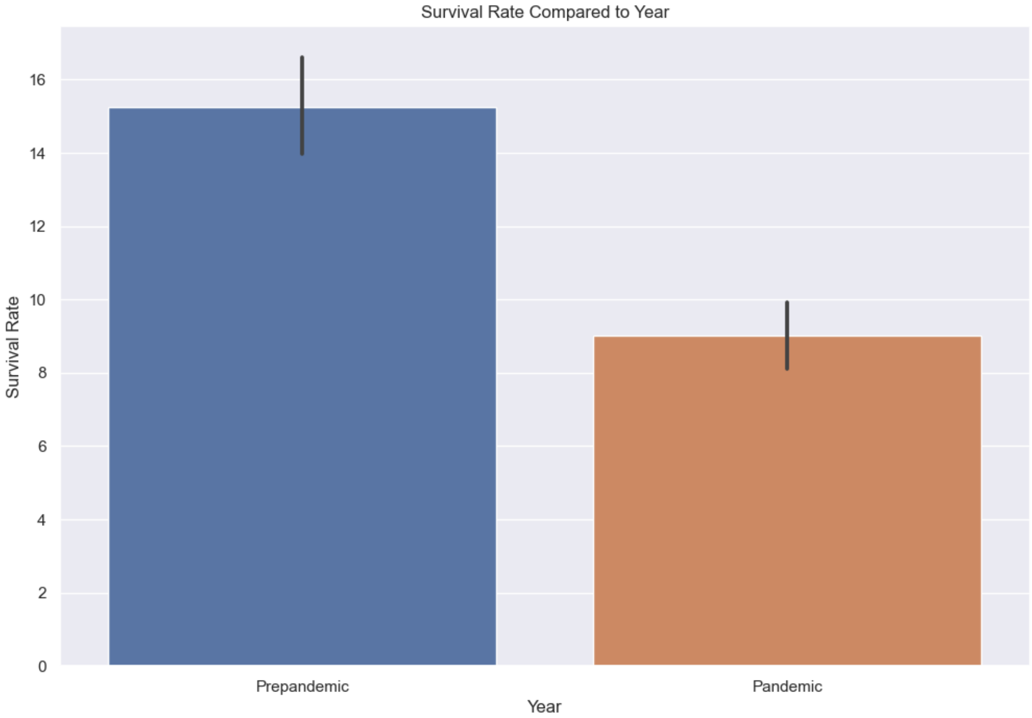


Figure 1. Bar graph showing the survival rates of liver cancer pre-pandemic and during the pandemic

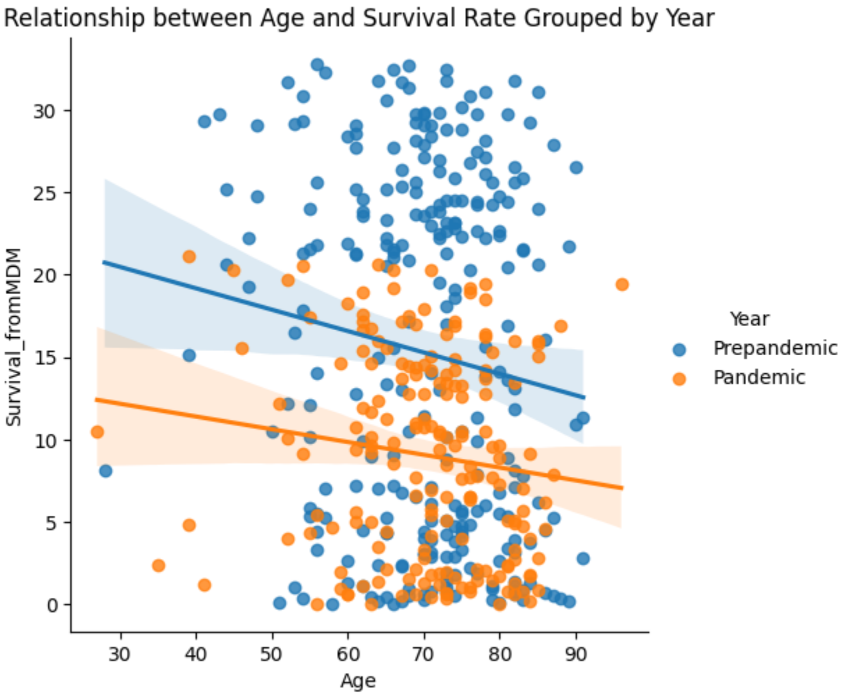


Figure 2. Relationship between age and liver cancer survival rate.

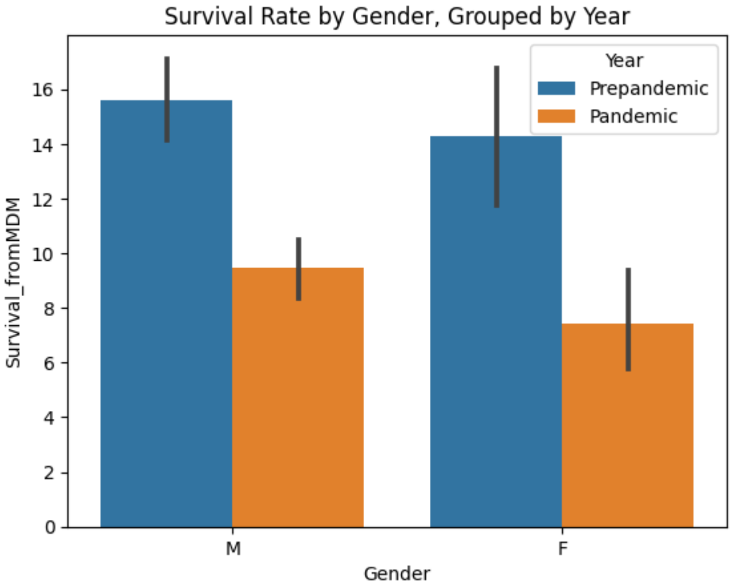


Figure 3. Liver cancer survival rate by gender

Q2. How was the timing between the decision to treat and the first treatment change from pre-pandemic to post-pandemic? What about the time from the initial diagnosis to the first treatment?

Performing the analysis

To conduct this analysis, I created two new datasets, one with NaN values filtered out for the *Time_decisiontotreat_1st_treatment* attribute and one with NaN values filtered out for the *Time_diagnosis_1st_Tx* attribute. I used boxplots for each of these visualizations to view any outliers. Of note, there was an outlier of -1400 in the time from the initial diagnosis to the initial treatment data, which skewed the plot very heavily and required me to set showfliers=False to be able to visualize the data fully.

Results

Oddly enough, even though resources were stretched during the pandemic and the survival rate of patients decreased, as seen in *Q1*. The time between the initial diagnosis, the decision to treat, and the first treatment was shorter during the pandemic than before (*Figure 4 & Figure 5*). This shows the opportunity for further analysis with additional datasets to examine why initial treatment started sooner during the pandemic.

Visualization

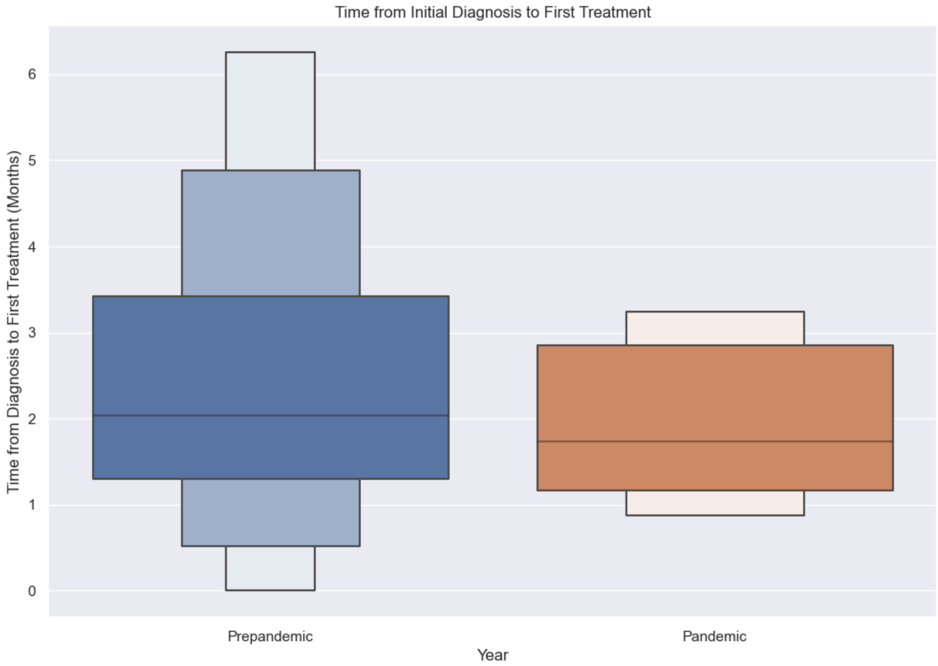


Figure 4. Boxplot showing time from the initial diagnosis to the first treatment

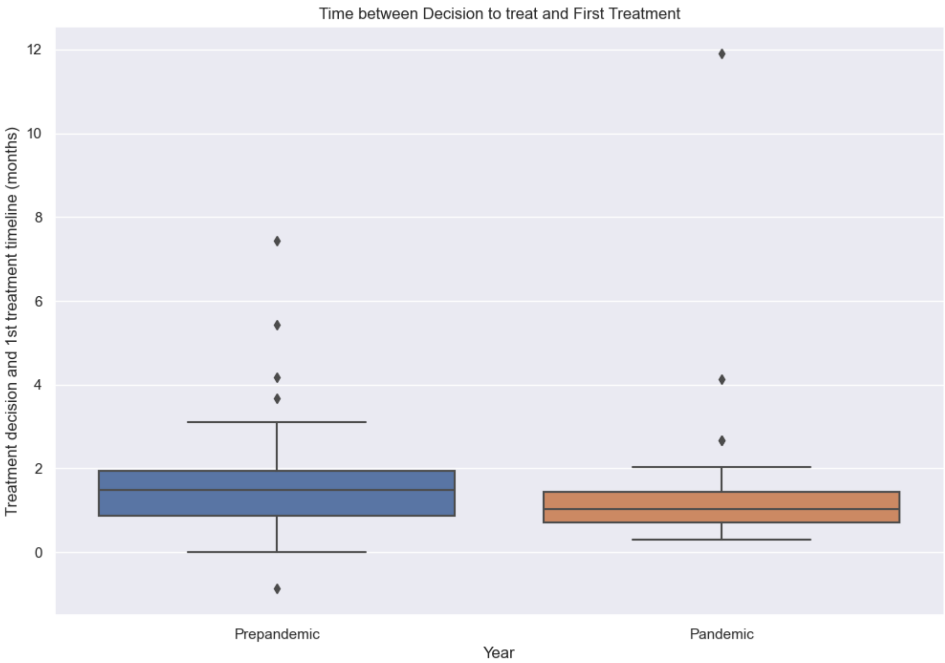
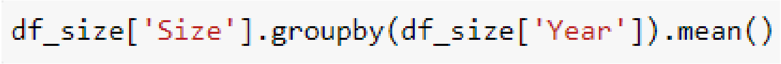


Figure 5. Boxplot showing time from decision to treat and first treatment.

Q3. When diagnosed, was there a relationship between tumor size post-pandemic with limited healthcare resources?

Performing the analysis

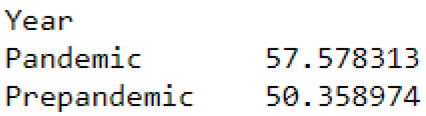
To conduct this analysis, I first created a dataset with NaN values filtered out for the *Size* attribute. While I intended to do a line plot time series to view this data, the style of the plot was tough to understand. Instead, I saw that my best visualization would come from a bar chart with the x-axis being month and the y-axis being the tumor size. Based on the graph (*Figure 6*), it was difficult to tell if there was an actual difference in tumor size during the pre-pandemic 12 months and the pandemic 12 months. To further determine the outcomes, I calculated the average tumor size for patients grouped by what *Year* category they fell under.



```
df_size['Size'].groupby(df_size['Year']).mean()
```

Results

I uncovered an increase in tumor size during the pandemic though it was only an average of about 7.6mm. While this is still a significant increase concerning cancer, it is not as substantial as I had expected. It did spike during August, November, and December of 2020, and further analysis into the rises and falls of tumor size during the pandemic would also be interesting.



```
Year
Pandemic       57.578313
Prepandemic    50.358974
```

Visualization

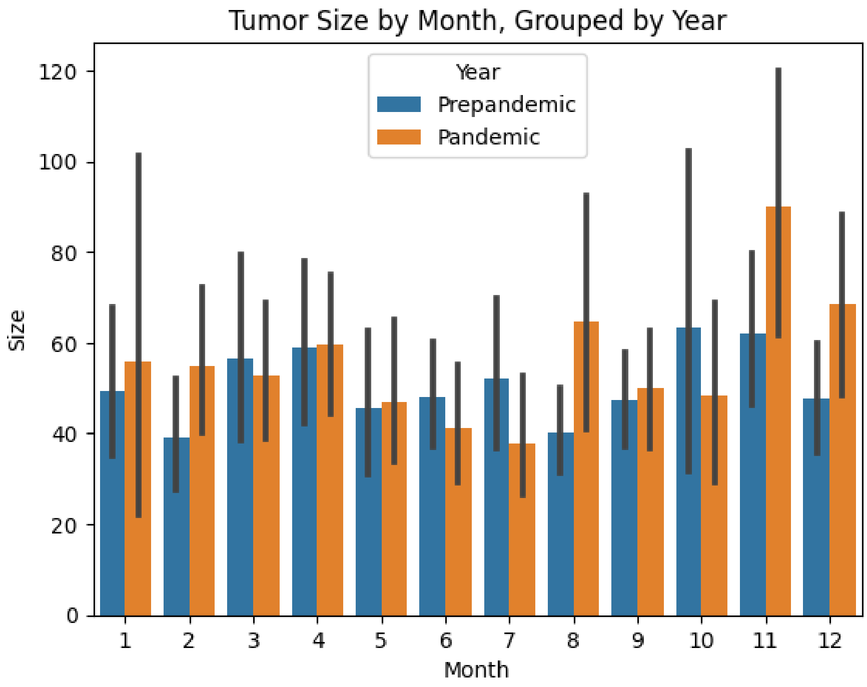


Figure 6. Bar plot showing tumor size by month

Conclusion

While my analyses did uncover the negative impact that the Covid-19 pandemic had on patients with cancer, specifically liver cancer, it also left several questions that could be explored with additional datasets and further analyses. Most of my findings aligned with my initial expectations, having seen the strain that the pandemic caused on the healthcare system. Still, some, such as the reduction in time to first treatment during the pandemic, were surprising. In the future, combining this dataset with a dataset that also contains an attribute noting whether or not the patient did contract the virus might add more insight into the reason behind some of the findings presented in this report.