Vishnu Nair

Report: An Analytical Approach to Understanding the GBM Treatment Market

1. Introduction

The purpose of this report is to analyze the structure of the Glioblastoma Multiforme (GBM) market and identify clinically meaningful patient sub-segments that are currently treated differently. This analysis was conducted in hopes of presenting distinct opportunities for our client based on meaningful patterns and associations we find between patient information and treatment plans. Through exploratory data analysis (EDA), plotting, and statistical techniques, we aim to provide insights that will guide our management team's discussion in leveraging the data effectively.

2. Data Overview

This GBM dataset under analysis consists of patient demographic and health status information and features 750 rows and 42 columns. It includes meaningful information regarding the comorbidities of the GBM patient, their age at diagnosis, genes that are overexpressed/mutated, their ECOG score (measuring disease progression) and relates it to the first and second line treatment regimens. Through effective EDA and statistical techniques, we can elucidate underlying associations between these features and the aforementioned treatment plans. Data quality checks were run before any analysis was conducted. It was found that there are no null values or duplicate patient IDs in the dataset. Additionally, all unique values were closely checked in each column to ensure the presence of valid entries. It was concluded that the quality of the data was sound overall, except for the presence of “999” values in the “% of tumor mass surgically resected“ column.

3. Identifying Meaningful Features for Modeling

Categorical Variable Analysis

As an initial step, we gain a better understanding of underlying segments by exploring the categorical variables that comprise a majority of the dataset. In Figure 2, we use the Cramer’s V statistic to get a snapshot of all associations between the categorical variables in our dataset, finding meaningful covariates. The aim here was to note any patient demographic information (such as race or comorbidities) that are strongly associated with the Regimen in 1st and 2nd Line. Surprisingly, there seems to be very little association between Gender, Race, and comorbidities in relation to the 1st line of treatment. However, there seems to be a moderate correlation between Race and the Regimen in the 2nd Line. We note that all strong associations (correlation coefficient above 0.5) that are occurring in the dataset are between the gene mutations of the GBM patient as well as the first line of treatment. This elucidates an important point: to consider what treatment a patient needs (hence better understanding the GBM treatment market), it is important to consider the patient’s race and specific gene mutations they have. Here, we discovered the specific variables that have moderate to strong associations with the Regimen in 1st and 2nd Lines for the patients. These discovered variables will be used in the clustering and modeling we will conduct to create the patient sub-segments. To confirm the associations of categorical variables to the “Regimen in 1st Line” in Figure 2, a Chi-Square test for association was run. The p-values obtained were below the threshold of 0.05 for the following variables: “ECOG at 1st Line”, “ECOG at 2nd Line”, “MGMT methylated”, “EGFR mutated”, “TP53 mutated”, “IDH1/IDH2 mutated”, and “PD-L1 overexpressed”. Hence, there is evidence to suggest that there are associations between the degree of

disease progression/gene mutations to the treatment the patient receives as a first response.

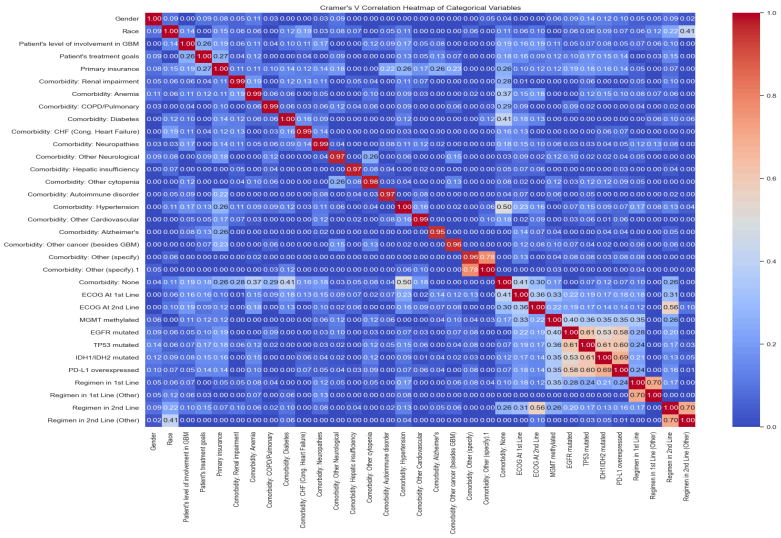

Figure 2. Cramer's V correlation heatmap of categorical variables

Numerical Variable Analysis

We now shift our focus to testing and quantifying associations for the two primary numerical variables in relation to different treatment types: "Age at Diagnosis" and "% of tumor mass surgically resected" and finding valuable patient sub-segments. The aim here is to test how age groups vary in treatment plans. First, an ANOVA test between these variables and the first line of treatment was first conducted. We obtained a p-value of 0.0043 for the former variable and 3.892e-24 for the latter. From this, we conclude that the means of these numerical variables differ significantly across different treatments, indicating that these are important variables that should in fact be used when segmenting. Now we take a deep dive into how the age at diagnosis for a patient specifically ties into a treatment regimen in Figure 3. This will serve to roughly begin identifying patient subsets..

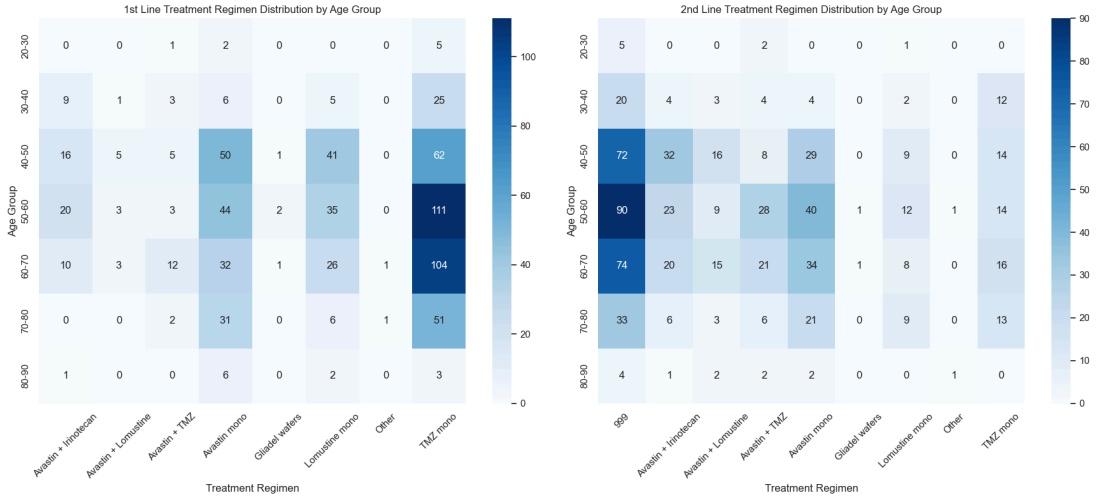

Figure 3. 1st and 2nd line treatment regimen distribution by age group

The mean age at diagnosis is between 56-57 years old for a patient. Looking at Figure 3, we see that Avastin mono, Lomustine mono, and TMZ mono are the top three choices of treatment for a patient. Out of the three, TMZ mono is used the most by a wide margin for the 50-60 age group. TMZ mono seems to be used the most for age groups at both tail ends of the spectrum as well, outcompeting other treatments, particularly in the 30-40 and 70-80 age groups. If one were to invest in one treatment plan, it seems this treatment would have been the most robust option based on age alone. For second line treatment, we note that Avastin + Irinotecan, Avastin + TMZ, and Avastin mono are the top three most used. The '999' category, is hypothesized to be null values but needs to be confirmed with the stakeholder/database engineer. If our hypothesis is correct, then a large proportion of the patients do not need a 2nd line of treatment, meaning that the TMZ mono, Avastin mono, and Lomustine mono treatments are sufficient as the first line of defense to treat GBM. From this, we can conclude that TMZ mono is a drug that can be used across virtually all age ranges and for patient sub-segments whose age range from 30-80, Avastin mono, Lomustine mono, and TMZ mono are all great options.

4. Deep Dive of Significant Features Found

Figure 4 below shows plots of all variables we found significant. First, let us discuss how patient segment mutations relate to their treatment plan. Our analysis previously showed that mutations are significantly correlated to treatment type as expected in accordance with GBM research, so we do a deep dive here to confirm this [1]. As a side note, we observed that there are '9's in these columns and currently assume it signifies a high level of mutation/expression/methylation in these columns. Firstly, EGFR mutations are known to be very common for GBM patients and decrease the quality of living [2]. Our data upholds this notion as we note a rapid spike in ECOG values equalling 4 where EGFR mutations are high. In instances where there is a medium level of EGFR mutation, Lomustine mono and Avastin mono are used the most. Research shows that the overexpression of PD-L1 is also strongly associated with poor quality of life for a patient as well [3]. Our data upholds this notion as well as the same trend observed in ECOG was noted. In these instances, TMZ mono and Avastin mono are widely used as seen in Figure 4.

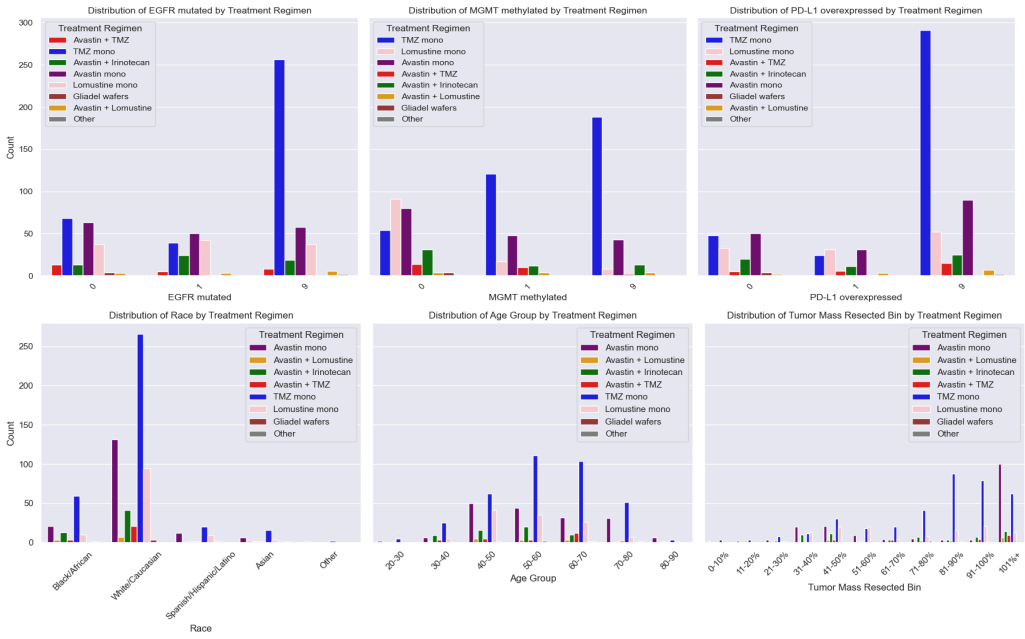

Figure 4. Treatment regimen distribution of all statistically significant variables

Next, MGMT methylation is known to be correlated with higher overall survival for a patient [4]. We found that none of the patients who had MGMT methylation had an “ECOG at 1st line” of the value 4 and comparable amounts of values 0 and 1 to the no-methylation group, supporting the aforementioned statement. The prevalence of a normal value of expression for MGMT methylation (equalling 1) is higher than that of the other two gene mutations based on Figure 4 as well. If we wished to bolster the probability of survival for this major patient population, TMZ mono or Avastin mono would be a good treatment option based on the plot. If we wish to treat the non-methylation and the over-expressed methylation population (assumed to be the “9” group), the same treatments are used widely.

Now, let us discuss the other significant variables and how they relate to treatment plans. In the three bottom plots in Figure 4, we explore treatment distributions across race, age group, and percentage of tumor masses resected. Across all races, TMZ mono and Avastin mono are used at a greater rate. We note that Avastin + TMZ is used predominantly for the White/Caucasian group that are 60-70 years old. For smaller percentages of the tumor mass being resected in a patient, the treatments used are primarily Avastin mono and Lomustine mono. As this percentage increases, it seems TMZ mono becomes the better option by a wide margin. As previously mentioned, there seems to be a group that has a percentage greater than 100 in the data analyzed. This group is treated using Avastin mono and TMZ mono in most cases and needs to be further studied after consultation with database engineers.

Overall, Lomustine mono and Avastin mono are the most-used drugs when the patient possesses mutations occurring at a medium-frequency regardless of race. When the patient possesses no mutation or a high-frequency mutation, TMZ mono and Avastin mono are administered most frequently as seen in Figure 4. These are important patient sub-segments to note when conducting feature selection prior to modeling/clustering.

5. Statistical Analysis / Clustering

Finally, we conducted clustering using the K-Prototypes algorithm to segment the patients using the categorical and numerical variables we deemed important from the previous sections. Doing so, we were able to group the patients efficiently using the elbow method (note the distinct clusters formed in Figure 5). The clusters were obtained, summary statistics were found for the clusters, then further aggregations were conducted in order to obtain concrete patient segments (see Figure 6).

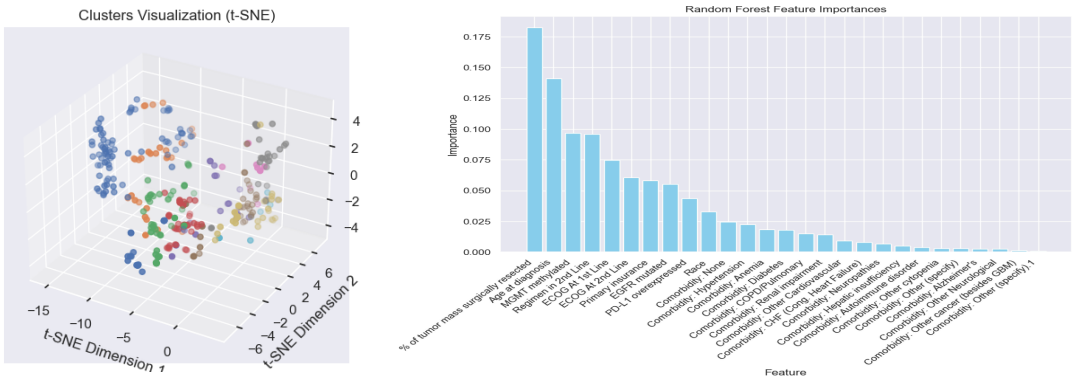

Figure 5. Clustering using K-Prototypes and visualizing using t-SNE and random forest feature importances

	Race	Age Range	Age at Diagnosis Mean	% of tumor resected range	% of Tumor Mass Surgically Resected Mean	ECOG At 1st Line	MGMT methylated	EGFR mutated	PD-L1 overexpressed	Regimen in 1st Line
0	Asian	52-52	52.000000	80-80	80.000000	1	1	9	0	Avastin + Irinotecan
1	Black/African	31-70	53.625000	34-100	59.000000	1	0	0	9	Avastin + Irinotecan
2	Spanish/Hispanic/Latino	58-58	58.000000	38-38	38.000000	1	0	1	0	Avastin + Irinotecan
3	White/Caucasian	33-70	49.656250	34-100	63.562500	1	0	1	0	Avastin + Irinotecan
4	Asian	59-63	61.000000	25-50	37.500000	1	9	9	9	Avastin + Lomustine
5	Black/African	34-34	34.000000	100-100	100.000000	2	0	1	1	Avastin + Lomustine
6	White/Caucasian	44-47	45.000000	30-46	40.666667	0	0	1	1	Avastin + Lomustine
7	Asian	75-75	75.000000	25-25	25.000000	1	1	1	1	Avastin + TMZ
8	Black/African	66-66	66.000000	50-50	50.000000	1	0	9	9	Avastin + TMZ
9	Spanish/Hispanic/Latino	55-55	55.000000	50-50	50.000000	1	1	1	9	Avastin + TMZ
10	White/Caucasian	24-75	56.000000	30-100	77.357143	1	0	0	9	Avastin + TMZ
11	Asian	49-49	49.000000	35-35	35.000000	1	0	1	0	Avastin mono
12	Black/African	37-67	50.666667	10-99	46.222222	1	0	0	9	Avastin mono
13	Spanish/Hispanic/Latino	45-52	48.666667	21-80	45.833333	1	0	1	0	Avastin mono
14	White/Caucasian	34-72	49.690909	14-96	49.400000	1	0	1	1	Avastin mono
15	Spanish/Hispanic/Latino	48-48	48.000000	75-75	75.000000	0	0	0	0	Gliadel wafers
16	White/Caucasian	53-61	57.666667	30-100	70.000000	0	0	0	0	Gliadel wafers
17	Asian	33-69	50.000000	65-95	81.666667	1	0	9	9	Lomustine mono
18	Black/African	44-62	53.714286	33-95	66.714286	1	0	0	9	Lomustine mono
19	Spanish/Hispanic/Latino	41-86	60.777778	32-100	58.888889	1	0	1	0	Lomustine mono
20	White/Caucasian	19-83	53.238095	23-100	67.369048	1	0	1	9	Lomustine mono
21	Asian	72-72	72.000000	90-90	90.000000	1	0	9	9	Other
22	Black/African	69-69	69.000000	75-75	75.000000	2	9	9	9	Other
23	Asian	26-78	57.333333	50-100	87.750000	1	9	9	9	TMZ mono
24	Black/African	40-80	59.640000	20-100	81.920000	1	9	9	9	TMZ mono
25	Other	60-64	62.000000	75-95	85.000000	0	0	9	9	TMZ mono
26	Spanish/Hispanic/Latino	19-72	55.466667	20-100	69.466667	1	9	9	9	TMZ mono
27	White/Caucasian	30-84	57.162162	10-100	76.531532	1	9	9	9	TMZ mono

Figure 6. Patient sub-segments found using K-Prototypes + aggregation by Regimen in 1st Line and Race

In Figure 6, we see the 27 patient segments created from the K-Prototypes algorithm. We aggregated by Race and 1st line treatment in order to get age ranges, age at diagnosis means and ranges of the tumor percentages respected for each group. There are many takeaways we can observe from this, but we will focus on the White/Caucasian group as it comprises a majority of the dataset. We note the group using Avastin + Irinotecan is composed of 33-70 year olds at diagnosis with mean age at diagnosis of 49. Most of these individuals have an ECOG value of 1, no MGMT methylation or PD-L1 overexpression and mutation of EGFR. The White/Caucasian group that has TMZ mono as their treatment have a mean age of 57 at diagnosis and on average 76% of their tumor resected, possessing high levels of MGMT methylation, EGFR mutation, and PD-L1 overexpression. The White/Caucasian group that gets treated with Lomustine mono also have a mean age of 53 at diagnosis, varying from 19-83 years old at diagnosis and having on average 67% of their tumor mass resected, no MGMT methylation, moderate EGFR mutation, and high PD-L1 overexpression. As a validation to ensure the validity in these features for predicting/classifying treatment options, a random forest model was built on a subset of features and feature importances were deduced (see right hand plot in Figure 5). A 66.67% accuracy was obtained, indicating that these features are decent predictors of 1st line treatment options.

6. Key Findings

In summary, our analysis began with thorough data quality checks, EDA and statistical tests of both numerical and categorical variables at hand. From our numerical analysis, we ascertained that Avastin mono, Lomustine mono, and TMZ mono are the top three choices of treatment for a patient if they are middle-aged. From exploring the effects of mutations on ECOG and treatment options, we found that Lomustine mono and Avastin mono are the most-used drugs when the patient possesses mutations occurring at a moderate frequency regardless of race. After carefully selecting the most important candidate features, we created a K-Prototypes aggregation system in order to create 27 patient sub-segments that are treated differently, presenting opportunities for us to invest in said treatments for different patient groups. In the future, supervised learning techniques may be implemented to predict treatment plans for patients using the identified features as well.

Appendix/References:

1. Montemurro, N. (2020). Glioblastoma Multiforme and Genetic Mutations: The Issue Is Not Over Yet. An Overview of the Current Literature. J Neurol Surg A Cent Eur Neurosurg, 81(01), 064-070. doi: 10.1055/s-0039-1688911

2. Pan, P. C., & Magge, R. S. (2020). Mechanisms of EGFR Resistance in Glioblastoma. International Journal of Molecular Sciences, 21(22), 8471. https://doi.org/10.3390/ijms21228471

3. Hao, C., Chen, G., Zhao, H., Li, Y., Chen, J., Zhang, H., Li, S., Zhao, Y., Chen, F., Li, W., & Jiang, W. G. (2020). PD-L1 Expression in Glioblastoma, the Clinical and Prognostic Significance: A Systematic Literature Review and Meta-Analysis. Frontiers in Oncology, 10, 1015. https://doi.org/10.3389/fonc.2020.01015

4. Moradi Binabaj, M., Bahrami, A., ShahidSales, S., Joodi, M., Joudi Mashhad, M., Hassanian, S. M., Anvari, K., & Avan, A. (2017). The prognostic value of MGMT promoter methylation in glioblastoma: A meta-analysis of clinical trials. Journal of Cellular Physiology, 232(1), 68-81. https://doi.org/10.1002/jcp.25896

5. https://oncologypro.esmo.org/oncology-in-practice/practice-tools/performance-scales