Galina Glousker
Senior Scientist, EPFL
galina.glousker@epfl.ch

Data Science Project

The impact of molecular profiling of gliomas on treatment and prognosis

Conceptual Design Report

1 October 2024

Abstract

Gliomas are a diverse group of primary brain tumors originating from glial cells (astrocytes, oligodendrocytes, and ependymal cells) in the central nervous system. Starting from 2016 WHO classification of CNS tumors and its revisions shifted the focus from purely histopathological analysis to an integrated diagnosis incorporating molecular markers. By integrating molecular markers into the diagnostic framework, clinicians can better stratify patients, tailor treatment plans, and predict clinical outcomes. This shift towards personalized medicine, based on well-established molecular data, is improving the management and treatment of gliomas, particularly in identifying patients who may benefit from specific therapies or have better prognoses.

In my project I will use genomic data from more than 900 glioma patients from the study published in Clinical Cancer Research journal in 2019 and apply supervised and unsupervised machine learning algorithms to identify relationships between genomic alterations and disease progression and treatment response.

Table of Contents

1 Project Objectives

Malignant primary brain tumors remain among the most difficult cancers to treat, with a 5 year overall survival no greater than 35%. The most common malignant primary brain tumors in adults are gliomas. They include: astrocytomas, which arise from astrocytes; oligodendrogliomas, originating from myelin producing cells; and ependymomas, arising from ependymal cells (lining of brain ventricles). Gliomas are graded by the WHO classification into low grade gliomas (grade I-II, least aggressive) and high grade gliomas (grade III- IV, most aggressive).

Central nervous system (CNS) tumors have long been classified based on histological findings supported by tissue-based tests (eg, immunohistochemical, ultrastructural). More recently, molecular biomarkers have gained importance in providing both ancillary and defining diagnostic information. For instance, tumors previously classified as “glioblastomas” may now be identified as IDH-mutant gliomas with a better prognosis. Genomic markers also provide prognostic information, with certain mutations (e.g., IDH mutations, MGMT promoter methylation) correlating with better survival, while others (e.g., EGFR amplification, H3 K27M mutation) indicating a worse prognosis. Genomic profiling can also identify potential targets for therapy, such as EGFR inhibitors for tumors with EGFR amplification and IDH inhibitors for IDH-mutant gliomas. (WHO classification).

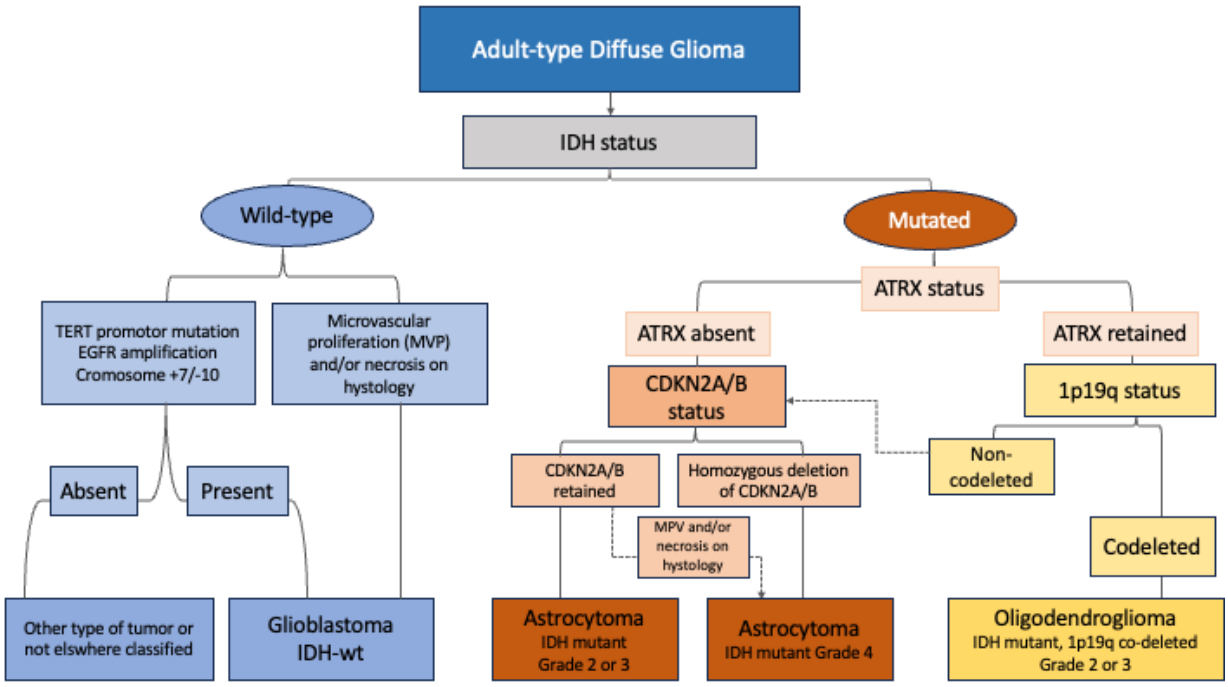

Figure 1. Decision tree of glioma diagnostics based on WHO classification

The goal of my project is to analyze genomic data from a large cohort of adult patients with glioma to identify genomic alterations associated with clinical behavior and therapy outcome, and use machine learning methods to be able to predict glioma outcomes in adult patients depending on their tumor molecular profile.

2 Methods

2.1 Infrastructure

I am going to use Google Colab and local Python installations (Visual Studio Code, Anaconda Distribution for Python) on my personal computer. Depending on the further analyses required I might need to use the High-Performance Computing facility at EPFL.

2.2 Software libraries and tools

First data evaluation and some plots generation will be done using tools on cBioportal.org.

Project will be done in Python. I am going to use the following libraries for the analysis:

- Bravado: allows to easily interact with REST APIs by generating a client for an API based on its Swagger (OpenAPI) specification (https://bravado.readthedocs.io/en/stable/).

- Pandas: data manipulation (https://pandas.pydata.org/)

- Numpy: mathematical operations on arrays (https://numpy.org/)

- Matplotlib: plots and visualization (https://matplotlib.org/)

- Geneview: making genomics graphics in Python (https://github.com/ShujiaHuang/geneview)

- g:Profiler: functional profiling of gene list from large-scale experiments (https://biit.cs.ut.ee/gprofiler/gost)

- statsmodels.api: statistical test and tools (https://www.statsmodels.org/stable/api.html)

- SciKitLearn or sklearn: machine learning models (https://scikit-learn.org/stable/)

2.3 Modeling, algorithms and statistical methods

For exploring and modeling of the relationships between patients' molecular profiles and clinical outcomes I will start with supervised (linear regression, random forest) and unsupervised (UMAP) machine learning algorithms. I am going to use 80:20 split and 5-fold cross-validation to train and validate the supervised ML models.

To validate accuracy of my classifications I am going to use Precision and Recall metrics, and to estimate how good my predictions are I am going to use Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared.

At more advanced stage of the project I will try to use neural network based algorithms that are often used in cancer genomics: Multilayer perceptron (MLP), recurrent neural network (RNN) and convolutional neural network (CNN) [4] .

3 Data

3.1 Data collection and acquisition

I am going to use genomic data from more than 800 glioma patients from the study published in Clinical Cancer Research in 2019. The chromosomal rearrangements, methylation and somatic mutations data are anonymized and publicly available from cBioportal.org (www.cbioportal.org/study?id ¼ glioma_mskcc_2019).

Table 1. Cohort characteristics

	All patients (*N* = 923)
Age at diagnosis (years)	
Median	52
Range	11–90
Sex (%)	
Male	557 (60)
Female	366 (40)
WHO class at diagnosis (%) [a]	
IDH-WT	596
Glioblastoma	468 (47)
Anaplastic astrocytoma	65 (7)
Diffuse astrocytoma	23 (2)
Gliosarcoma	15 (2)
Diffuse midline glioma, H3 K27M-mutant	5 (<1)
Other	20 (2)
IDH-mutant	327
Oligodendroglioma	93 (9)
Diffuse astrocytoma	86 (9)
Anaplastic astrocytoma	80 (8)
Anaplastic oligodendroglioma	43 (4)
Glioblastoma	24 (2)
Other	1 (<1)

Table 1. Study cohort

I will use three data frames for my project. From the “genes” I will have to extract mutated genes that then will be mapped to Mutated Pathways to create a new feature. In “mutations” I will use tumor_sample_barcode column to map each mutated gene and pathway to the patient in “clinical” data frame that contains all the information about patents and that will be used for further analysis. Unfortunately, since the dataset contains unmatched Patient and Sample IDs, I had to omit almost 2000 entries out of more than 9000, into my final data frame “merged_df_clean”

```
<class 'pandas.core.frame.DataFrame'>
Index: 7674 entries, 8 to 9670
Data columns (total 31 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Hugo_Symbol                            7674 non-null   object
 1   Entrez_Gene_Id                         7612 non-null   float64
 2   Consequence                            7674 non-null   object
 3   Tumor_Sample_Barcode                   7674 non-null   object
 4   HGVSp_Short                            7175 non-null   object
 5   Hotspot                                7674 non-null   int64
 6   Pathway                                7674 non-null   object
 7   Study ID                               7674 non-null   object
 8   Patient ID                             7674 non-null   object
 9   Sample ID                              7674 non-null   object
 10  Actionable Lesion1                     7667 non-null   object
 11  Diagnosis Age                          7667 non-null   float64
 12  Cancer Type                            7674 non-null   object
 13  Cancer Type Detailed                   7674 non-null   object
 14  Enhancing                              2630 non-null   object
 15  Gene Panel                             7674 non-null   object
 16  Histology                              7674 non-null   object
 17  MGMT Status                            5822 non-null   object
 18  Mutation Count                         7674 non-null   float64
 19  Oncotree Code                          7674 non-null   object
 20  Overall Survival (Months)              7667 non-null   float64
 21  Overall Survival Status                7667 non-null   object
 22  Patient Display Name                   7667 non-null   object
 23  Progress Free Survival (Months)        4259 non-null   float64
 24  Progression Free Status                4273 non-null   object
 25  Number of Samples Per Patient          7674 non-null   float64
 26  Sample Type                            7674 non-null   object
 27  Sex                                    7667 non-null   object
 28  TMB (nonsynonymous)                    7674 non-null   float64
 29  WHO Classification of Diagnostic Tumor 7667 non-null   object
 30  WHO Grade                              7674 non-null   object
dtypes: float64(7), int64(1), object(23)
```

Table 2. Final data frame after cleaning and formatting

For the initial analyses I am going to use the following parameters:

Hugo_symbol, Patient ID, Sample ID, Diagnosis Age, Cancer Type, Pathway, Sex, WHO Grade, TMB (nonsynonimous), Overall survival, Progress free survival.

To validate my ML models, I am going to test them on another dataset from glioma patients [3] (42 patients, study is done by the same laboratory which should decrease variability).

4 Metadata

Metadata for this study is available on cBioportal.org and is also stored on my personal computer. It includes among others the description of the type of cancer, cancer study identifier (glioma_mskcc_2019), study description, pmid (PubMed ID) and citation information. Each data table has a corresponding meta data file.

Gene panels for this study can be found here:
https://github.com/cBioPortal/datahub/tree/master/reference_data/gene_panels

5 Data Quality

This dataset is curated and I can also filter the data according to my selection criteria on cBioportal, and I am going to use samples with no missing values for the features that I need.

The problem can happen when merging different tables, because some patient will be lacking some samples, and the missing pateint ID will lead to filtering these samples out decreasing the sample size.

6 Data Flow

Data Extraction: Pull data from cBioPortal using the **Glioma MSKCC 2019** study.

Preprocessing: The data are curated, so I will not need to clean anything, but preprocessing and normalization will be needed (e.g., pathway enrichment analysis, encoding categorical features). The proper names of the genes need to be checked for the further analysis. Cleaning entries without fully matching information is needed after merging data from different experiments.

Feature Engineering: Focus on mutation burden, specific mutations, mutated pathways and methylation profiles related to clinical outcomes.

Modeling: Apply machine learning models (linear regression, random forest and random forest on the most important principal component) to predict survival.

Evaluation: Assess the model's performance.

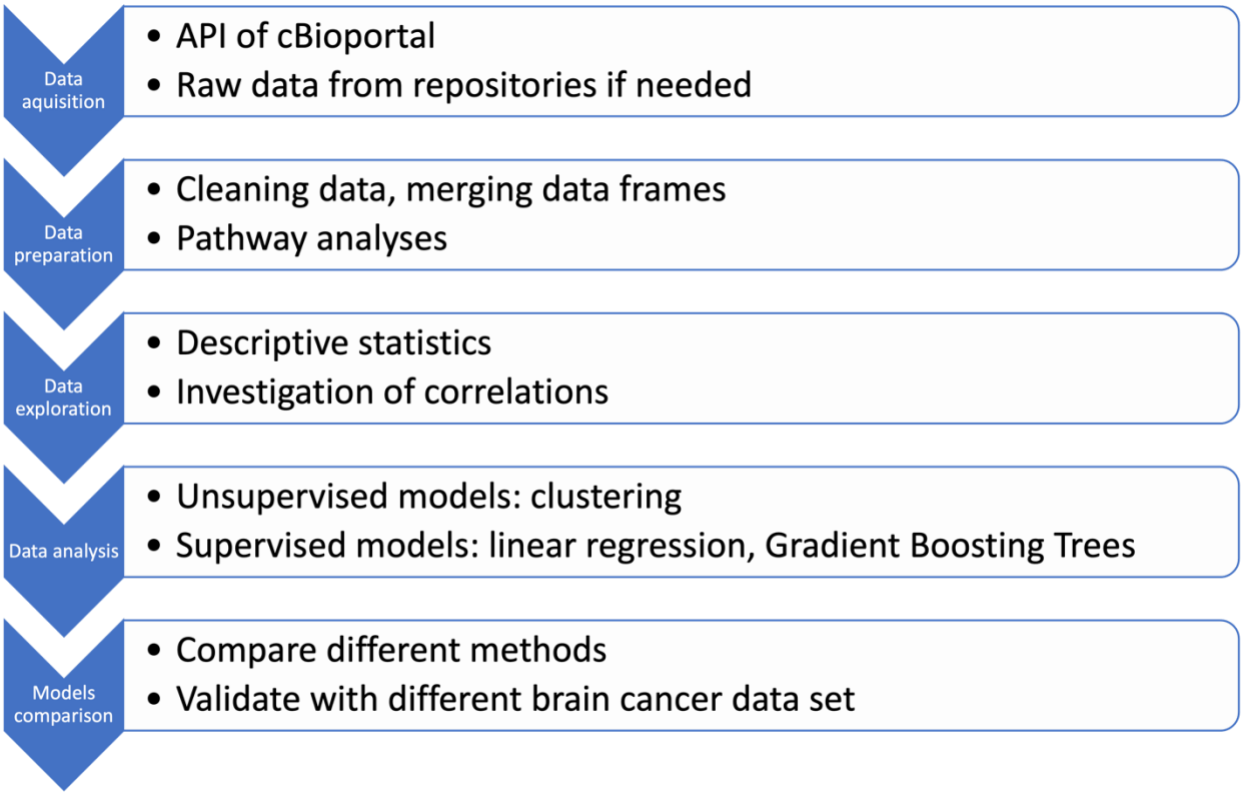

Figure 2: Data flow diagram.

7 Data Model

Draw and explain your data model at the conceptual level, the logical level and the physical level.

At the conceptual level, I aim to define the relationships between molecular profiles of glioma patients, treatment they receive and clinical outcomes.

At the logical level, I specify the attributes and detail how entities relate to each other:

Entities and Attributes:

- **Patients**: Patient ID, age, sex, WGO Grade
- **Molecular Pathways**: Patient ID, Sample ID, Hugo_Symbol, TMB, Pathway
- **Clinical Outcomes**: Overall survival, Overall survival time, Progression free survival, Progression free survival time

Relationships:

- **One-to-Many** between **Patients** and **Genes**.

- **One-to-Many** between **Patients** and **Pathways**.
- **One-to-One or One-to-Many** between **Pathways** and **Clinical Outcomes**.

At the **physical level**, there are no specific requirements ans the dataset I am using is relatively small and I don't need any specific infrastructure.

8 Documentation

I am going to use Jupiter notebook on Google Colab, saved both on the Google Drive and on my computer, with comments on the code so that my study can be reproduced.

9 Risks

Glioma is not a very frequent cancer type so my main concern is not to have the sample size large enough to create a powerful predictive model.

In order to test if my models are generalizable, I will use another dataset from glioma patients on cBioportal.com to validate the models (42 patients, study is done by the same laboratory which should decrease variability). If I see that my dataset is too small to provide reliable predictions, I can switch to larger datasets coming from another cancer type available at cBioportal.com.

10 Preliminary Studies

After matching data frames from different parts of studies I ended up with 841 patients instead of 923 I had in the fist table. Out of 841 patients 499 are males, 341 are females and for one patient gender is not assigned.

One of the first diagnostic features in glioma is IDH status. Depending on the study and the dataset, 30-70% of glioma patients carry mutations in IDH1 or IDH2 gene. In our dataset we can see that a little less than 40% of the patients contain IDH mutations. IDH mutations are associated with better prognosis, but also with higher recurrence, and it seems to be represented on the Fig. 3: we see higher fraction of secondary tumor samples that have IDH mutation that those that don't.

Another characteristic feature is tumor mutational burden (TMB). It is an emerging biomarker for the prediction of immunotherapy success for solid tumors. In the same time it is associated with poor prognosis in glioma. We can see that TMB is significantly higher in more advanced tumors (Fig. 4).

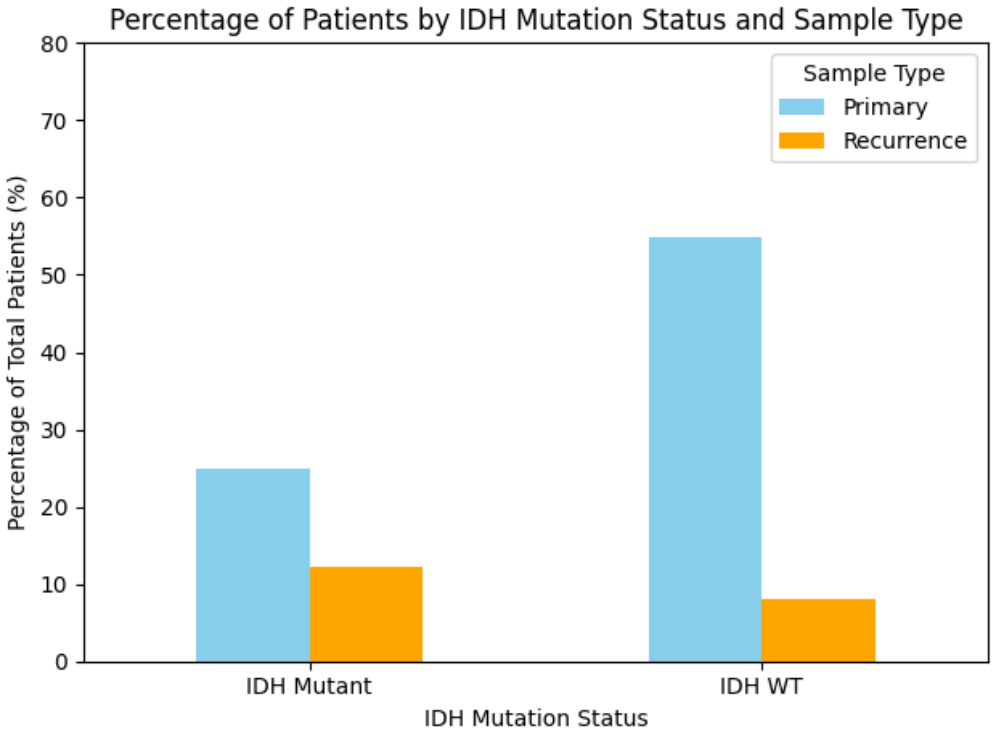

Figure 3: Percentage of total patients by IDH mutation and sample type. Mann-Whithney test, P-values: {'Primary': 1.659953987802262e-147, 'Recurrence': 7.581727399256121e-39}

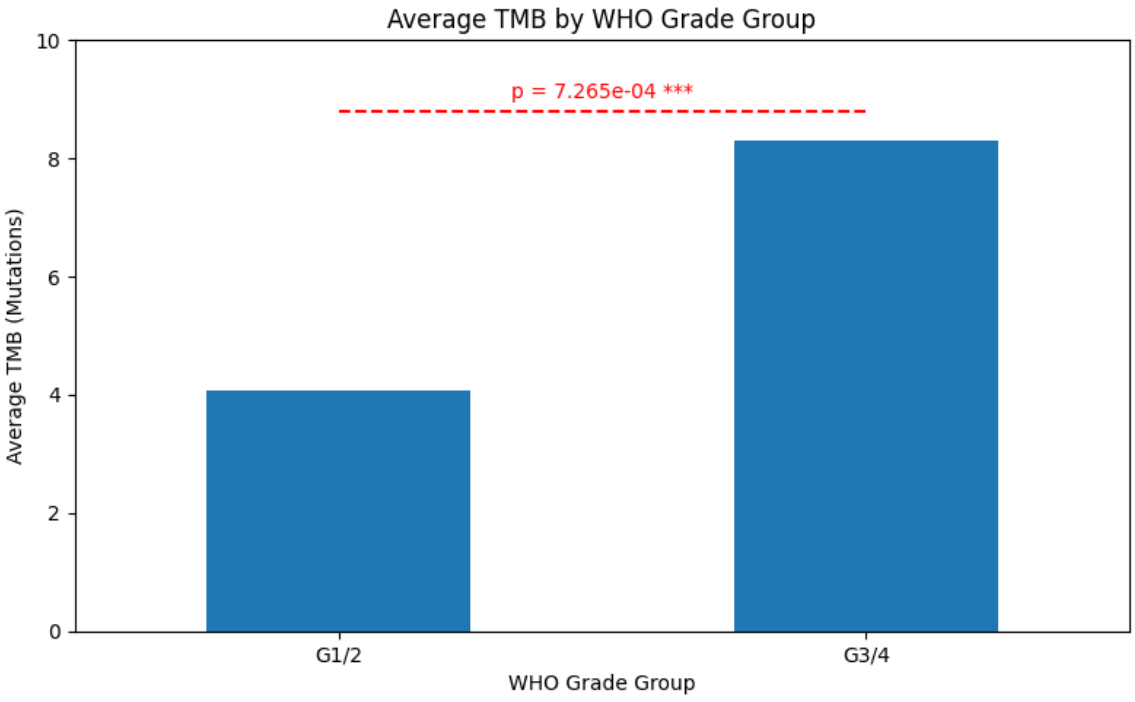

Figure 4: Relationship between tumor mutation burden and WHO Grade Group, t-test.

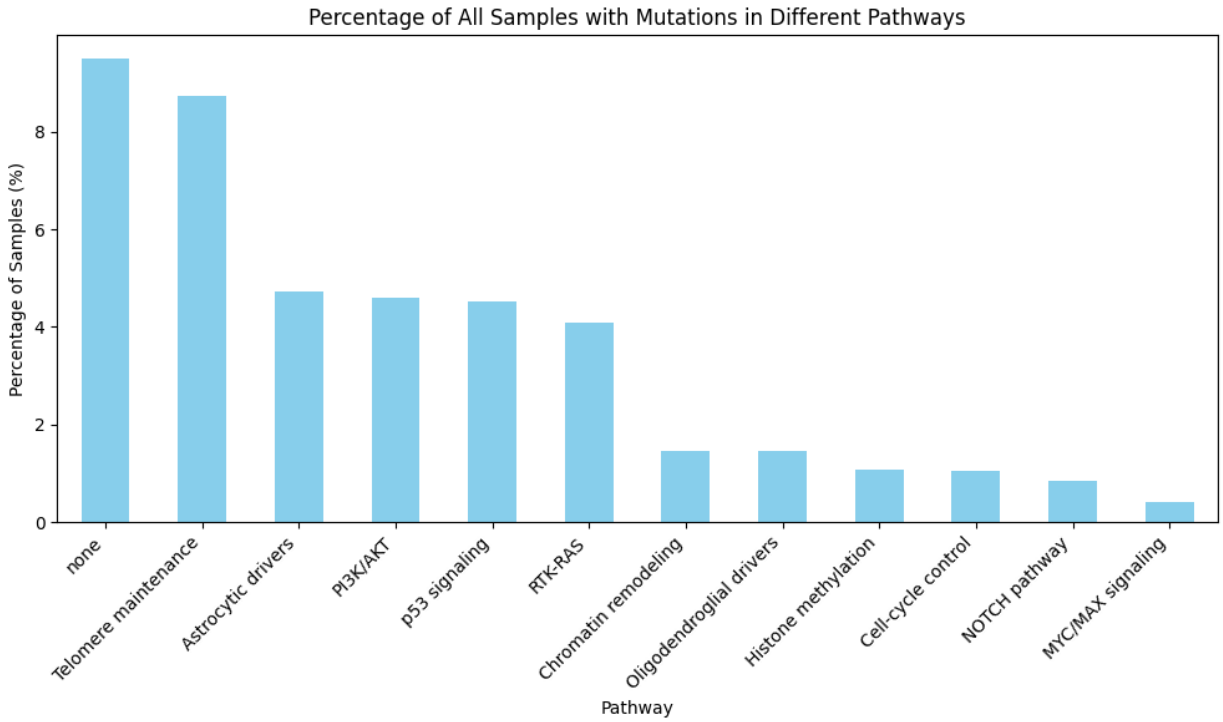

Figure 5: Percentage of mutations in the subset of pre-defined signaling pathways.

Another way to look at the biology underlying certain cancer progression processes is to map the mutated genes to different cellular processes. We mapped the mutated genes to a list of predefined signaling pathways. We can see that the majority of mutations were not assigned a pathway, and the most present pathways were telomere maintenance (which is one of the first stages of carcinogenesis), then there is a number of proliferative signaling pathways that are typical for malignant transformations are targeted (Fig. 5). Most importantly, we can see some differences in the pathways affected depending on WHO grade (Fig. 6), which can teach us something about brain cancer progression.

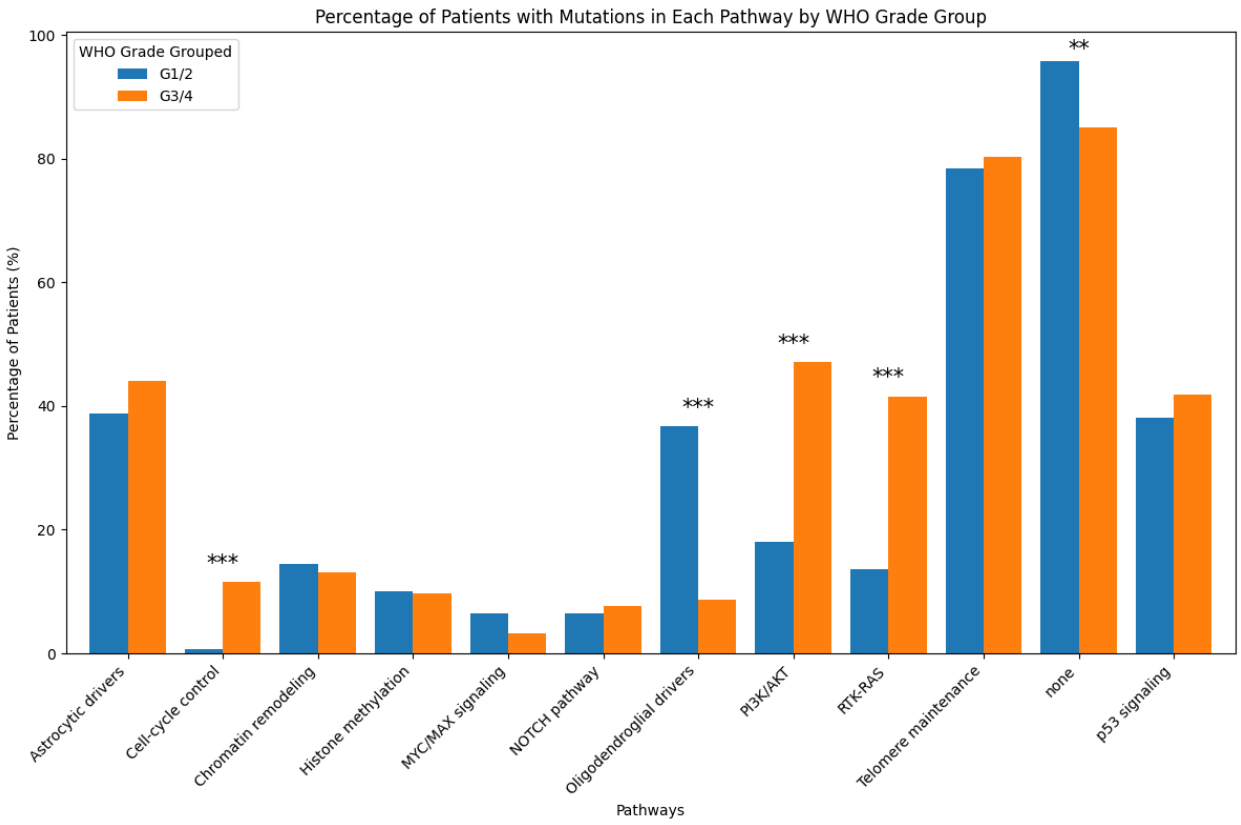

Figure 6: Distribution of mutations in the subset of pre-defined signaling pathways depending on the WHO grade. Chi-2 test. *** $p<0.001$

11 Conclusions

My project will provide insights into how specific mutations and molecular profiles influence clinical outcomes in glioma patients. The machine learning models will offer predictive power for patient survival based on genomic and clinical data.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 01.10.2024 Signature(s): 

References and Bibliography

[1] The 2021 WHO Classification of Tumors of the Central Nervous System. Neuro-Oncology 23(8), 1231–1251, 2021 | doi:10.1093/neuonc/noab106

[2] Genomic Correlates of Disease Progression and Treatment Response in Prospectively Characterized Gliomas. Clin Cancer Res. 2019 Sep 15;25(18):5537-5547. doi: 10.1158/1078-0432.CCR-19-0032.

[3] Tracking tumour evolution in glioma through liquid biopsies of cerebrospinal fluid. Nature. 2019 Jan;565(7741):654-658. doi: 10.1038/s41586-019-0882-3.

[4] Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* 13, 152 (2021). https://doi.org/10.1186/s13073-021-00968-x