Advancing Brain Tumor Analysis: Curating a High-Quality MRI Dataset for Deep Learning-Based Molecular Marker Profiling

Divya D. Reddy*[1], Niloufar Saadat*[1], James M. Holcomb*[1], Benjamin C. Wagner*[1], Nghi C. Truong[1], Jason Bowerman[1], Kimmo J. Hatanpaa[1], Toral R Patel[1], Marco C. Pinho[1], Ananth J Madhuranthakam[1], Chandan Ganesh Bangalore Yogananda[1], Joseph A. Maldjian[1]
*Authors with equal contribution
[1]University of Texas Southwestern Medical Center

Abstract

This study sought to address the critical need for high-quality datasets to advance deep learning (DL) models for non-invasive profiling of brain tumor molecular markers. We curated a comprehensive brain tumor dataset from patients diagnosed at University of Texas at Southwestern Medical Center (UTSW) between 2012 and 2020. The study curated 1740 MRI sessions from 709 subjects, emphasizing pre-operative cases. Molecular analyses were conducted on tumor tissues as part of standard clinical care to identify key genetic alterations (MGMT, IDH, 1p/19q co-deletion). Detailed demographic, histological, and molecular marker information were extracted from the clinical electronic medical record (EMR). An internal reporting system was created in XNAT to comprehensively catalogue and curate the imaging and associated clinical data. MR Images were pre-processed using the FeTS platform, and quality checked (QC) by expert neuroradiologists. Brain tumors were segmented into key components using FeTS and were manually corrected by experienced research staff and verified by expert neuroradiologists. Here we describe the critical challenges, approaches, and solutions for creating a high-quality curated medical imaging dataset for use in deep learning studies.

1. Introduction

Genetic subtyping and molecular profiling of brain tumors are transforming therapeutic strategies to enhance prognostic accuracy. Compelling evidence supporting this paradigm is the 2021 revision of the World Health Organization's (WHO) classification of gliomas [1]. The revision transformed the pathological diagnosis of gliomas from a purely histological to a multilayered integrated approach [1, 2]. Three important molecular markers have been extensively validated: O-6-methylguanine-DNA methyltransferase (MGMT), isocitrate dehydrogenase (IDH), and 1p/19q co-deletion status. The revision has underlined the importance of molecular markers, which are extensively studied for their critical role in tailoring patient-specific therapeutic approaches.

Recent advances in radiomics and deep learning (DL) have shown great success in non-invasive profiling of molecular markers using MRI [3-9]. The ability of these DL models to learn from vast datasets carries substantial importance in determining therapy and predicting prognosis. However, the efficacy of these DL models is heavily contingent upon the size and quality of the data they are trained with. As data-driven entities, the DL models thrive on large, diverse, and high-quality datasets to refine their predictive accuracy and generalize across different clinical settings.

Large datasets encompassing a wide range of tumor types, stages, and patient demographics allow DL models to learn the intricate patterns in brain tumors. This can lead to accurate predictions and diagnoses. Similarly, the quality of data, including the resolution of MR images, the accuracy of tumor annotations, and molecular profiling, are crucial for DL models to reliably identify and characterize tumors. Studies have shown that high-quality, large-scale datasets significantly improve the performance of DL models in brain tumor tasks [10, 11]. For example, the BraTS21 challenge provided a substantial dataset of brain tumor MRI with expert annotations and MGMT status[12]. This dataset, along with TCIA [13], has been instrumental in advancing the state of the art in tumor segmentation and molecular marker classification. This underscores the significance of rigorous data curation. It also highlights the need for datasets that span various imaging modalities and encapsulate rich, annotated information on molecular markers. It further emphasizes the role of data quality and size in harnessing the full potential of DL models in brain tumor analysis.

Curating a large-scale brain tumor database imposes substantial challenges. The process of collecting and annotating brain tumor MRI, necessitating medical expertise, is both resource-intensive and crucial for maintaining data quality. However, the essentials of data quality are non-negotiable, as they emphasize the generalization of DL models. To facilitate further research on using DL models for brain tumor analysis, we have meticulously curated a comprehensive brain tumor database. The curated database, including MRI, molecular marker status, and clinical information, is a testament to data quality's critical role in advancing brain tumor research. By curating a dataset that meets these stringent

criteria, we aim to enhance the predictive modeling of molecular markers and the personalization of glioma treatment strategies.

2. Materials and Methods

2.1 Patient Selection

The dataset included patients diagnosed with brain tumors at UTSW from 2012 to 2020. The electronic health records (EHR) were reviewed by a neuropathologist to identify eligible patients. Our study protocol, including the use of patient data, received approval from the Institutional Review Board (IRB), with consent requirements waived due to the retrospective nature of the patient data. We focused on patients older than 18 years and excluded patients with any history of prior brain tumor resection.

2.2 Demographic and Molecular Marker Information

Detailed patient demographics and clinical information were extracted from the EHR system, EPIC (Epic Systems Corporation, Verona, WI). An internal reporting system was created (using the imaging informatics platform XNAT) to comprehensively document the extracted information, as shown in *Figure 1*. This included demographics (age, sex, race, ethnicity), clinical details (dates of pathology, surgery, study), and results from histopathological exams.

Genetic and molecular testing were implemented on tumor tissues collected via biopsy or surgical resections, a standard part of the clinical care process. The choice between immunohistochemistry (IHC) and Next-generation Genetic Sequencing (NGS) hinged on the quantity of tissue available. Brain tumors were assessed for type, grade, and the Mindbomb Homolog-1 (MIB-1) index. IDH mutations were detected using either IHC or NGS methods. Similarly, the presence of 1p/19q co-deletion was determined through NGS or fluorescence-in-situ-hybridization (FISH) techniques. To assess the methylation status of the MGMT promoter, dual real-time polymerase chain reaction (PCR) assays were employed, utilizing both published and in-house developed primers specific to methylated and unmethylated sequences. These assays help quantify the methylation level in finely dissected, paraffin-embedded tissue samples.

NGS provided a comprehensive overview of various molecular alterations including ATRX, P53, TERT mutations, EGFR amplification, CDKN2A deletions, and alterations in chromosome numbers +7 and -10. Special attention was given to cases that deviated from the usual, such as IDH wildtype or ATRX mutated but 1p19q co-deleted or any grade 1 tumors including pituitary adenoma, astrocytic neoplasm, were subjected to additional scrutiny. These outliers were removed from the dataset to ensure the integrity and consistency.

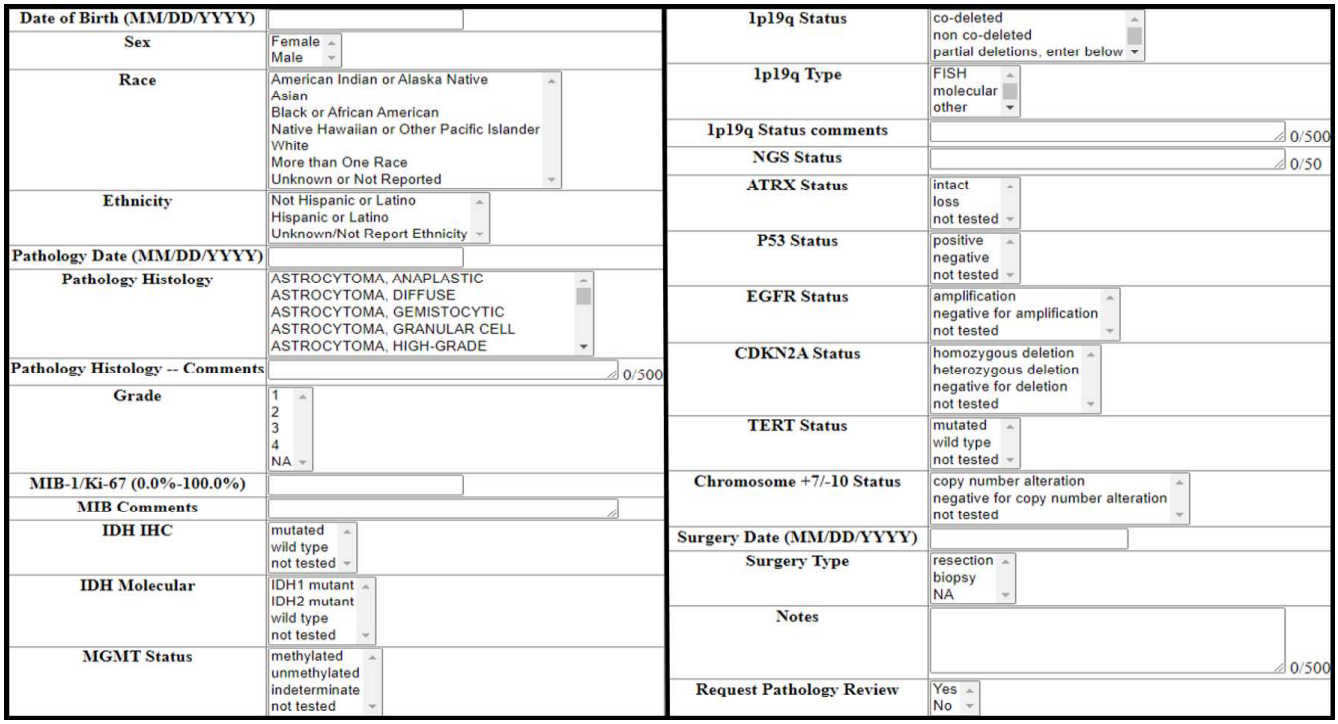

Figure 1: Tabular representation of XNAT.

2.3 Imaging Parameter and Pre-processing

MRI scans for the patient list were reviewed in the Clinical Picture Archiving and Communication System (PACS) to identify pre-operative studies. These were then sent via DICOM image transfer to the XNAT internal research platform. In-house automated pipelines launched through XNAT converted the DICOM images into the Neuroimaging Informatics Technology Initiative (NIfTI) format, enhancing their suitability for computational analysis. A meticulous selection process (*Figure 2*) was undertaken for the NIfTI images, ensuring the inclusion of pre-contrast T1-weighted, post-contrast T1-weighted, T2-weighted, and T2w-FLAIR sequences from each patient. Axial and 3D acquisitions were prioritized for their superior detail and spatial resolution.

MR images were obtained from scanners manufactured by various vendors (GE, SIEMENS, Hitachi, Toshiba, and Philips), with magnetic field strengths spanning from 0.3 Tesla to 3 Tesla. The diversity of MRI data necessitated a standardized pre-processing routine to normalize image dimensions and voxel sizes, and to address the variations in scanner types, magnetic field strengths, and acquisition protocols. A substantial subset of patients, confirmed to have a complete series of the requisite MRI sequences, was processed further using the Federated Tumor Segmentation (FeTS) platform, version 007 [14]. The FeTS platform co-registers multi-contrast MRI with the SRI24 brain atlas template (240x240x155) with isotropic resolution (1mm^3), removes non-brain structures (skull-stripping), and segments the brain tumor. The tumor is segmented into i) "enhancing tumor (ET) ii) the "tumor core" (TC) which includes the ET and the necrotic part (NCR), and iii) the "whole tumor" (WT), a union of TC and the peritumoral edematous/infiltrated tissue (ED). Additionally, all datasets were (a) N4 bias corrected to remove RF inhomogeneities, and (b) intensity normalized to zero-mean and unit variance [15, 16]. This standardization and careful curation ensures consistency, inclusion of the highest quality scans and improves the reliability of subsequent analyses.

2.4 Image Quality Assessment

All imaging data were carefully checked for quality (QC) by trained research personnel and then validated by expert neuroradiologists. The review focused on the overall image quality, the effectiveness of skull-stripping, the op-status, and biopsies. MRI with significant image artifacts (patient movement or hardware anomalies) were excluded. MRI depicting post-surgical cavities were classified as post-operative (post-op) cases. Images with small skull defects (burr holes) without substantial resection were categorized as biopsy cases, while those with ambiguous features were labeled as unknown *(Fig. 3).*

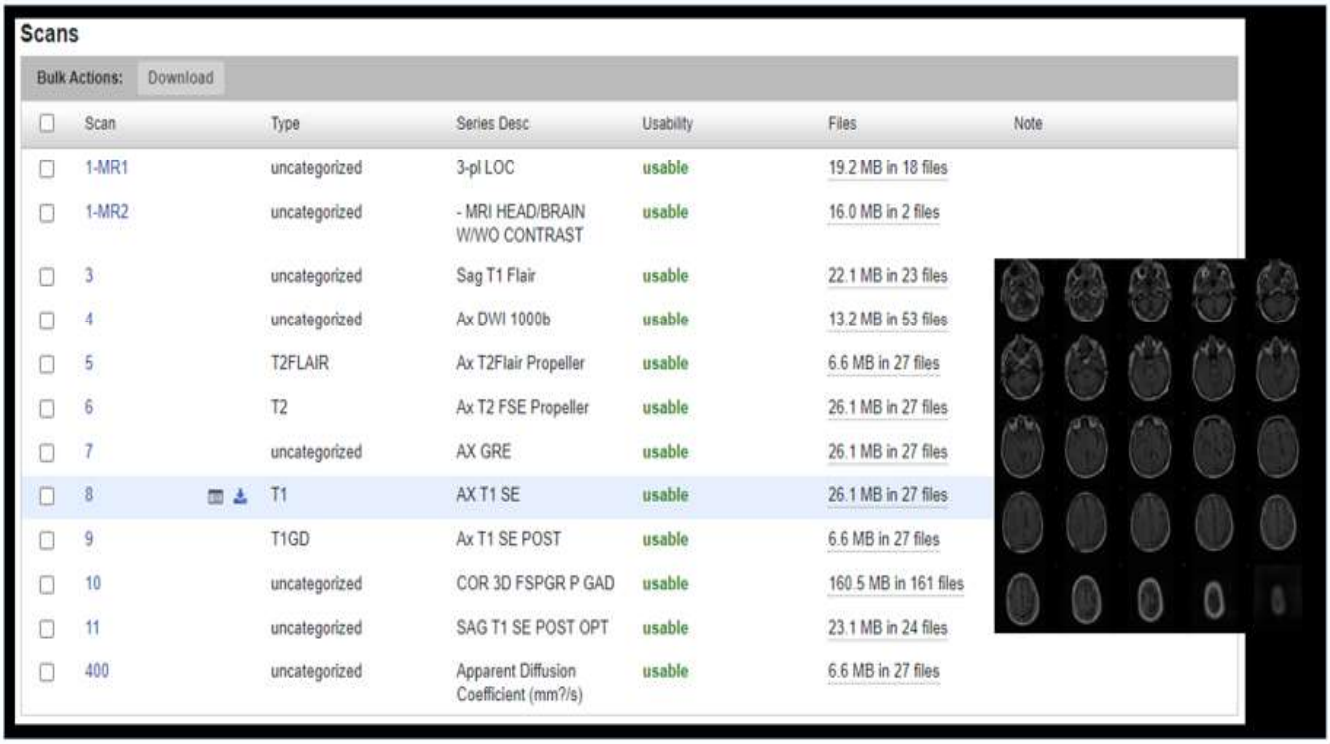

Figure 2: XNAT Interface for the MRI contrast selection process.

2.5 Tumor Segmentation and manual corrections

MR images were segmented into three key tumor areas (ET, NCR & ED) using a multi-contrast MRI approach. This approach assisted in accurately delineating complex tumor regions, including infiltrated tissues. The initial step involved leveraging an ensemble of FeTS deep-learning models for automated segmentation, creating a preliminary map of the tumor compartments. Following the automated process, an expert team of annotators undertook manual refinement of these segmentations to ensure accuracy using 3D Slicer [17]. The manual refinement was further subjected to a rigorous review by experienced neuro-radiologists, assessing each segmentation as either acceptable or in need of further correction (*Fig 4).* Annotators and expert neuroradiologists engaged in an iterative review cycle, meticulously analyzing, and refining these unsatisfactory segmentations. This approach facilitated comprehensive discussions on challenging cases, allowing the team to address and minimize subjective discrepancies effectively. The neuroradiologists conducted a final review to ensure all segmentations met the high standards required for inclusion in the dataset. This multi-stage review ensures the integrity and reliability of tumor segmentation *(Fig 5)*.

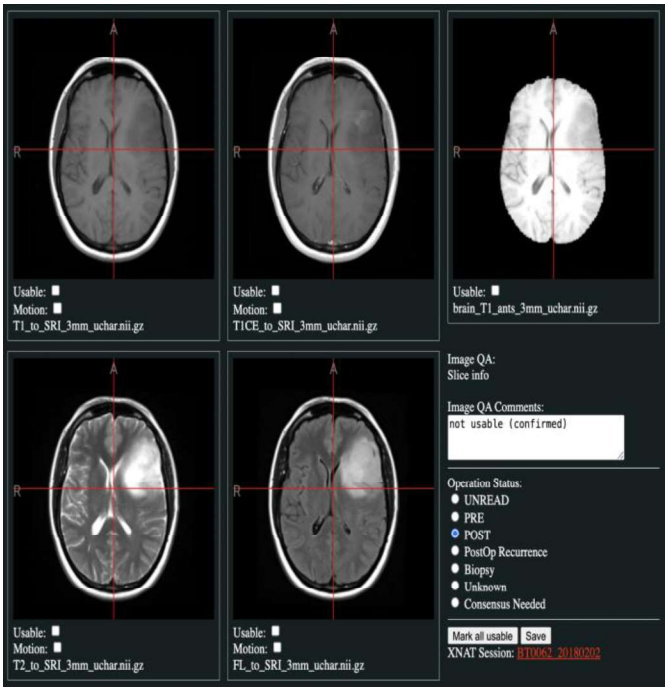

Figure 3: Interface for Image Quality Assessment – To note motion artifacts, op-status, and skull stripping quality.

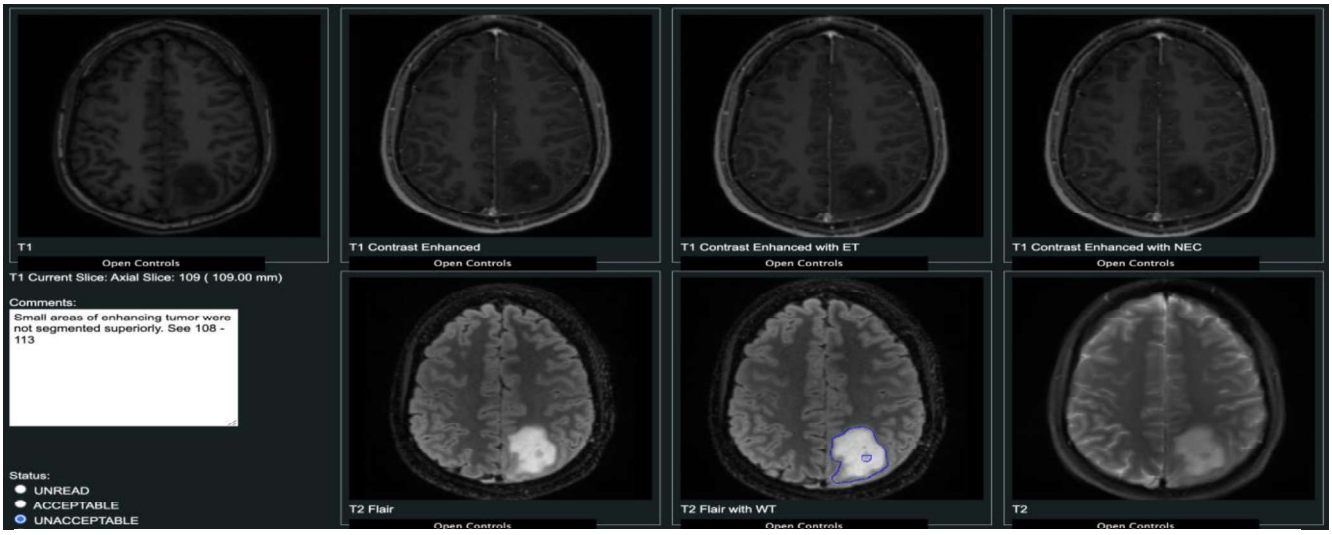

Figure 4: An XNAT QC interface to obtain Neuro-radiologist's review and comments on manual corrections.

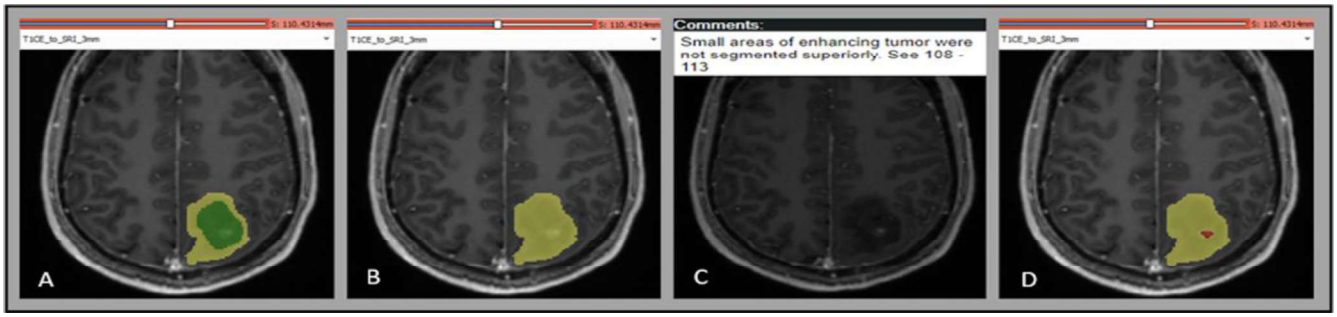

Figure 5: Tumor Segmentation Process: A. Automated Segmentation, B. Manual Correction, C. Radiologist's comments (to include ET), D. Acceptable Segmentation. Tumor Regions: Enhancing tumor (Red), Necrosis (Green) and Edema and NET (Yellow).

3. Results

Our study rigorously curated 1,740 MRI sessions from a cohort of 709 subjects, focusing primarily on the pre-operative cases. Figure 6(A) provides an overview of MRI sessions, including pre-operative (812), postoperative (766), biopsy (5), and sessions with undetermined status (157). The pre-op cases were further examined and repeated MRI sessions from the same patient were excluded. Figure 6(B) summarizes the MRI and pre-operative sessions. Only the unique pre-op sessions (642) were used for further evaluations.

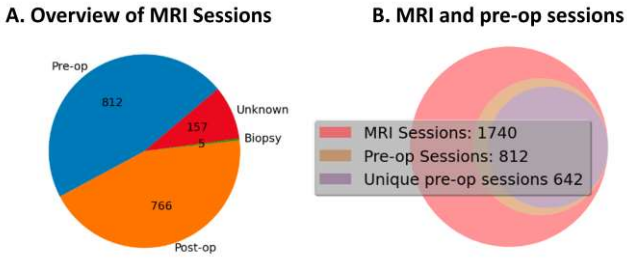

Figure 6: Overview of MRI Sessions with op-status.

3.1 Subject Demographics

A demographic analysis of the subset of pre-operative cases revealed a distribution of 348 males (54.2%), 251 females (39.1%), and 43 subjects (6.7%) with unspecified gender.

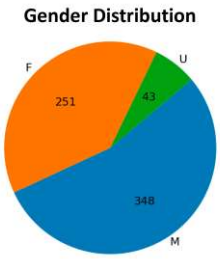

Figure 7: Gender Distribution on pre-op cases.

3.2 Histological Classification and Tumor Grade

Histological examination of pre-operative cases indicated a dominance of glioblastoma (377 subjects), underscoring the aggressive nature of brain tumors within our cohort. The prevalence of other glioma subtypes also provides a comprehensive landscape for comparison, as demonstrated in Figure 8(a). Figure 8(b) depicts the percentage of tumor grades in our dataset.

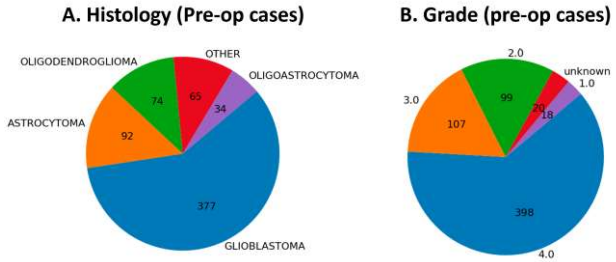

Figure 8: Tumor grade and histological distribution.

3.3 Molecular Profiling

The molecular characterization of the cohort is summarized in Figure 9. It displays the distribution of IDH mutations, 1p19q co-deletion, and MGMT methylation status. The intersection of these molecular markers is critical for understanding the composite molecular landscape of gliomas.

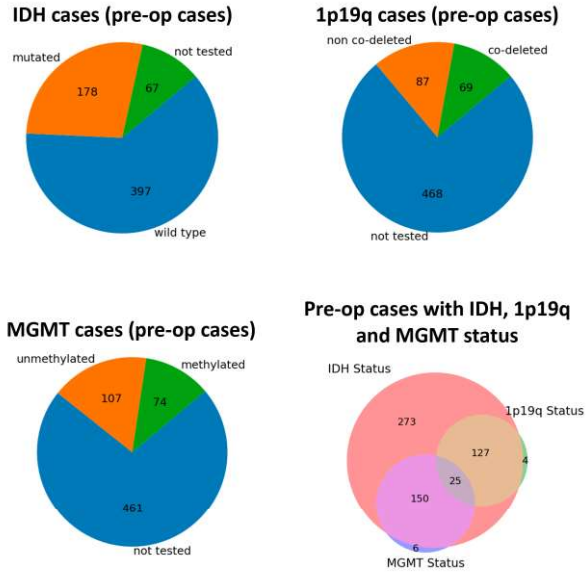

Figure 9: Summary of molecular status.

3.4 MRI Data Availability and Usability

Figure 10 showcases the outcome of the QC process applied to MRI data. This QC initiative is pivotal for ensuring that subsequent analyses are grounded on reliable and high-quality imaging data, with a clear distinction between usable and non-usable cases highlighted for each MRI sequence. The QC process of MRI data checks the image quality and marks them as usable or unusable. Figure 10(a) presents the summary of the number of usable cases for each MRI sequence, while Figure 10(b) depicts the number of cases before and after the QC process for each molecular marker.

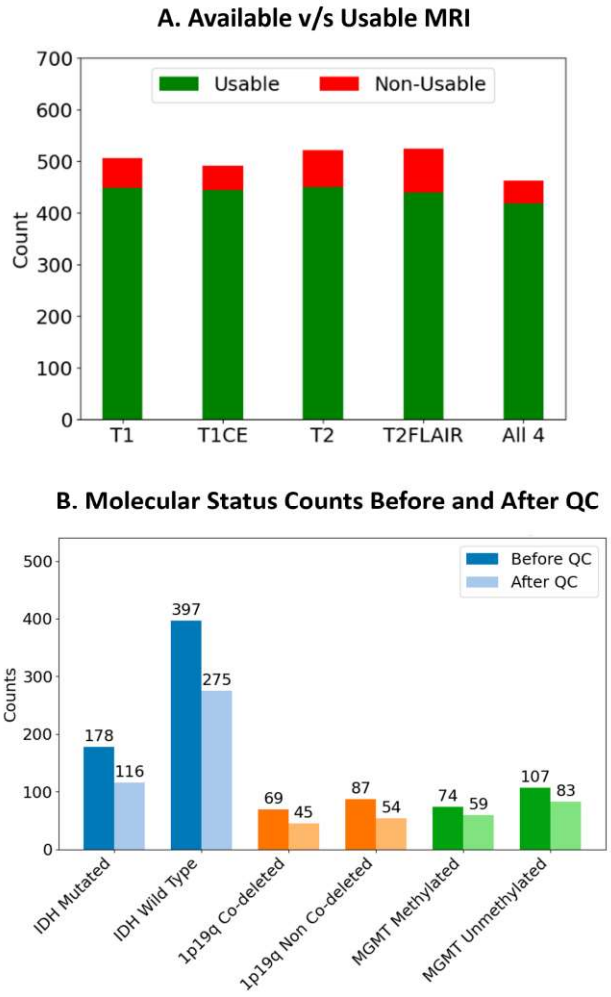

Figure 10: Summary of the data curation process.

4. Discussion

This study underscores the importance of high quality datasets in enhancing non-invasive profiling of molecular markers using MRI. It also aligns with the evolving paradigms of genetic subtyping and molecular profiling in brain tumor diagnostics and therapeutics. The 2021 WHO revision, by integrating molecular markers with traditional histological analysis, has set a new standard in glioma classification, emphasizing the necessity for high-quality, comprehensive datasets in the advancement of DL models.

This curated brain tumor dataset comprises detailed MRI, molecular marker status, and clinical information. It represents a significant step in addressing the critical challenge of data scarcity and variability in brain tumor research. This dataset encapsulates a wide range of tumor types, grades, and patient demographics to facilitate the refinement of DL models and ensures their applicability across diverse clinical settings. The meticulous data curation process, emphasizing the quality and size of data, enables the development of DL models for reliably identifying and characterizing brain tumors.

The adoption of the Federated Tumor Segmentation (FeTS) platform for image preprocessing and segmentation demonstrates the integration of advanced computational tools in managing and analyzing complex imaging data. This approach ensures consistency and reliability in the data preparation phase, which is crucial for subsequent application of DL models in tumor segmentation and molecular marker prediction.

The process of manual correction to the FeTS output adds a valuable layer of precision and accuracy to the automated segmentation. By incorporating expert neuroradiological review, these manual adjustments significantly enhance the quality of the data availale for DL algorithms. This synergy between automated processes and human expertise not only improves the reliability of the segmentation results but also provides rich, annotated datasets that are more conducive to the training and refinement of DL models. The meticulous manual corrections ensure that the DL algorithms can learn from the most accurate representations of tumor characteristics, thereby enhancing their performance in real-world diagnostic applications.

5. Limitations and Future Directions

The retrospective collection of MRI data and the inherent variability in imaging protocols across different clinical settings pose challenges to standardization and generalization procedures. Moreover, upon examination of the detailed steps outlined in this paper regarding dataset collection and curation, it becomes evident that the process is significantly time-consuming. Therefore, the potential for automating both the standardization and entire curation processes using deep-learning algorithms in future endeavors holds considerable promise and practical utility.

Future research will also focus on expanding the dataset to include a broader spectrum of both benign and malignant tumor types and their respective molecular markers, further enhancing the robustness and applicability of DL models. The exploration of federated learning approaches can address privacy concerns and facilitate multi-institutional collaborations, potentially leading to a more generalized and comprehensive understanding of brain tumors.

6. Conclusion

This study provides a valuable approach for the advancement of DL applications in brain tumor analysis, reflecting the critical importance of high-quality data curation. The curated database bridges the gap between complex molecular data and advanced imaging techniques. It paves the way for innovative approaches in tumor characterization, and treatment of gliomas, indicating new data-driven insights in neuro-oncology.

Acknowledgements

This work was funded by NIH/NCI R01CA260705 (J.A.M.) and NIH/NCI U01CA207091 (A.J.M. and J.A.M). The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of University of Texas Southwestern Medical STU-2020-1184.

References

[1] D. N. Louis *et al.*, "The 2021 WHO classification of tumors of the central nervous system: a summary," *Neuro-oncology,* vol. 23, no. 8, pp. 1231-1251, 2021.

[2] A. Sejda *et al.*, "WHO CNS5 2021 classification of gliomas: a practical review and road signs for diagnosing pathologists and proper patho-clinical and neuro-oncological cooperation," (in eng), *Folia neuropathologica,* vol. 60, no. 2, pp. 137-152, 2022, doi: 10.5114/fn.2022.118183.

[3] X. Zhang *et al.*, "Radiomics strategy for molecular subtype stratification of lower-grade glioma: detecting IDH and TP53 mutations based on multimodal MRI," *Journal of Magnetic Resonance Imaging,* vol. 48, no. 4, pp. 916-926, 2018.

[4] B. Zhang *et al.*, "Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas," *Neuro Oncol,* vol. 19, no. 1, pp. 109-117, Jan 2017, doi: 10.1093/neuonc/now121.

[5] P. Chang *et al.*, "Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations inGliomas," *AJNR Am J Neuroradiol,* vol. 39, no. 7, pp. 1201-1207, Jul 2018, doi: 10.3174/ajnr.A5667.

[6] J. Yan *et al.*, "Predicting 1p/19q co-deletion status from magnetic resonance imaging using deep learning in adult-type diffuse lower-grade gliomas: a discovery and validation study," *Laboratory Investigation,* vol. 102, no. 2, pp. 154-159, 2022/02/01 2022, doi: 10.1038/s41374-021-00692-5.

[7] K. Zhao *et al.*, "Automatic 1p/19q co-deletion identification of gliomas by MRI using deep learning U-net network," *Computers and Electrical Engineering,* vol. 105, p. 108482, 2023/01/01/ 2023, doi: https://doi.org/10.1016/j.compeleceng.2022.108482.

[8] S. Chakrabarty, P. LaMontagne, J. Shimony, D. S. Marcus, and A. Sotiras, "MRI-based classification of IDH mutation and 1p/19q codeletion status of gliomas using a 2.5D hybrid multi-task convolutional neural network," *Neuro-Oncology Advances,* vol. 5, no. 1, 2023, doi: 10.1093/noajnl/vdad023.

[9] C. G. Bangalore Yogananda *et al.*, "MRI-Based Deep Learning Method for Classification of IDH Mutation Status," *Bioengineering,* vol. 10, no. 9, p. 1045, 2023.

[10] A. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne, "Data management challenges for deep learning," in *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2019: IEEE, pp. 140-147.

[11] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain MRI segmentation: state of the art and future directions," *Journal of digital imaging,* vol. 30, pp. 449-459, 2017.

[12] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific data,* vol. 4, no. 1, pp. 1-13, 2017.

[13] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *Journal of digital imaging,* vol. 26, pp. 1045-1057, 2013.

[14] M. J. Sheller *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Sci Rep,* vol. 10, no. 1, p. 12598, Jul 28 2020, doi: 10.1038/s41598-020-69250-1.

[15] N. J. Tustison *et al.*, "Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements," *Neuroimage,* vol. 99, pp. 166-79, Oct 1 2014, doi: 10.1016/j.neuroimage.2014.05.044.

[16] N. J. Tustison *et al.*, "N4ITK: improved N3 bias correction," *IEEE Trans Med Imaging,* vol. 29, no. 6, pp. 1310-20, Jun 2010, doi: 10.1109/TMI.2010.2046908.

[17] S. Pieper, M. Halle, and R. Kikinis, "3D Slicer," in *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*, 2004: IEEE, pp. 632-635.