The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset

Evan Calabrese, MD, PhD • *Javier E. Villanueva-Meyer, MD* • *Jeffrey D. Rudie, MD, PhD* • *Andreas M. Rauschecker, MD, PhD* • *Ujjwal Baid, PhD* • *Spyridon Bakas, PhD* • *Soonmee Cha, MD* • *John T. Mongan, MD, PhD* • *Christopher P. Hess, MD, PhD*

From the Center for Intelligent Imaging (Ci^2), Department of Radiology & Biomedical Imaging, University of California San Francisco, 505 Parnassus Ave, San Francisco, CA 94143 (E.C., J.E.V.M., J.D.R., A.M.R., S.C., J.T.M., C.P.H.); and Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, Pa (U.B., S.B.). Received March 23, 2022; revision requested April 25; revision received July 5; accepted August 2. **Address correspondence to** E.C. (email: *evan.calabrese@ucsf.edu*).

Supported in part by the National Institutes of Health under awards number NCI:U01CA242871 and T32EB001631, as well as by the Radiological Society of North America Research & Education Foundation under award number RR2011.

Conflicts of interest are listed at the end of this article.

Supplemental material is available for this article.

Radiology: Artificial Intelligence 2022; 4(6):e220058 • https://doi.org/10.1148/ryai.220058 • Content codes: AI NR OI MR • ©RSNA, 2022

MRI-based artificial intelligence (AI) research in patients with brain gliomas has been rapidly increasing in popularity, in part due to a growing number of publicly available MRI datasets. Notable examples include The Cancer Genome Atlas' glioblastoma dataset (TCGA-GBM) available at The Cancer Imaging Archive, consisting of 262 patients, and the Multimodal Brain Tumor Segmentation (BraTS) challenge dataset consisting of 542 patients (including 243 preoperative cases from TCGA-GBM) (1–4). The public availability of these glioma MRI datasets has fostered the growth of numerous emerging AI techniques, including automated tumor segmentation, radiogenomics, and survival prediction. Despite these advances, existing publicly available glioma MRI datasets have been largely limited to only four MRI sequences (T2-weighted, T2-weighted fluid-attenuated inversion recovery [FLAIR], and pre- and postcontrast T1-weighted), and imaging protocols vary substantially in terms of field strength and acquisition parameters.

Here, we present the University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM) dataset, which includes 501 patients with histopathologically proven diffuse gliomas who were imaged with a standardized 3-T preoperative brain tumor MRI protocol featuring predominantly three-dimensional (3D) imaging, including diffusion and perfusion imaging. The dataset also includes isocitrate dehydrogenase *(IDH)* mutation status for all patients and O^6-methylguanine DNA methyltransferase *(MGMT)* promotor methylation status for World Health Organization (WHO) grade 3 and 4 gliomas. Finally, we have also included treatment details, including extent of resection and overall survival. The UCSF-PDGM dataset has been made publicly available in hopes that researchers around the world will use these data to continue to push the boundaries of AI applications for diffuse gliomas.

Materials and Methods

Patient Population

Data collection was performed in accordance with relevant guidelines and regulations and was approved by the UCSF institutional review board with a waiver for consent. The dataset population consisted of 501 adult patients with histopathologically confirmed WHO grade 2–4 diffuse gliomas (following the 2021 WHO Classification of Central Nervous System Tumors) (5) who underwent preoperative MRI, initial tumor resection, and tumor genetic testing at a single medical center between 2015 and 2021. Patients with any prior history of brain tumor treatment were excluded; however, prior tumor biopsy was allowed (n = 69 of 501 or 14%). Some patients included in this dataset were included in previously published studies, including 199 in reference 6, 400 in reference 7, 387 in reference 8, and 400 in reference 9.

Surgical Treatment and Survival Data

Extent of resection and overall survival were determined by review of patient electronic medical records. When available, the determination of extent of resection was based on the operative report and/or immediate postoperative MRI report. Overall survival was recorded in days from initial diagnosis to the date of death or last clinical follow-up.

Genetic Biomarker Testing

All patients' tumors were tested for *IDH* mutations by either conventional (Sanger) or next-generation genetic sequencing (10). A majority (410 of 501 or 82%) were tested for 1p/19q codeletion by fluorescence in situ hybridization. All grade 3 and 4 tumors were tested for *MGMT* methylation status using an in-house–developed sensitive, quantitative methylation polymerase chain reaction assay based on prior work (11), which yields a number of methylated promoter sites (0–17), with values of two or greater being considered positive. All molecular data were determined using tissue acquired at open gross total or subtotal resection (ie, not from burr hole biopsy).

Image Acquisition

All preoperative MRI was performed with a 3-T scanner (Discovery 750; GE Healthcare) with a dedicated eight-channel head coil (Invivo). The imaging protocol included 3D T2-weighted, T2-weighted FLAIR, suscep-

Abbreviations

AI = artificial intelligence, BraTS = The Multimodal Brain Tumor Segmentation, FLAIR = fluid-attenuated inversion recovery, HARDI = high-angular-resolution diffusion imaging, PDGM = preoperative diffuse glioma MRI, TCGA = The Cancer Genome Atlas, UCSF = University of California San Francisco, WHO = World Health Organization, 3D = three-dimensional

Summary

The newly publicly available University of California San Francisco Preoperative Diffuse Glioma MRI dataset, consisting of 501 patients with grade 2–4 diffuse gliomas, includes standardized 3-T three-dimensional preoperative MRI protocol, diffusion MRI, and perfusion MRI, multicompartment tumor segmentations, tumor genetic data, and treatment and survival data. Data are available at *https://doi.org/10.7937/tcia.bdgf-8v37*.

Keywords

Informatics, MR Diffusion Tensor Imaging, MR Perfusion, MR Imaging, Neuro-Oncology, CNS, Brain/Brain Stem, Oncology, Radiogenomics, Radiology-Pathology Integration

tibility-weighted, diffusion-weighted, and pre- and postcontrast T1-weighted images, 3D arterial spin labeling perfusion images, and two-dimensional 55-direction high-angular-resolution diffusion imaging (HARDI). Acquisition parameters for each sequence are included as supplementary material and are further described in prior publications (7). Over the study period, two gadolinium-based contrast agents were used: gadobutrol (Gadovist; Bayer), at a dose of 0.1 mL per kilogram of body weight, and gadoterate (Dotarem; Guerbet), at a dose of 0.2 mL per kilogram of body weight.

Image Quality Assessment and Exclusion

All image data were manually reviewed for completeness and quality by a panel of reviewers with varying years of experience. In total, 544 cases were reviewed, 44 were excluded, and 501 were included. Eleven cases were excluded because of severe image artifacts (patient motion or hardware related), and 33 cases were excluded due to one or more missing series. Seventy-seven percent (387 of 501) of cases were also independently manually reviewed for quality as part of the 2021 BraTS challenge (8).

Image Preprocessing

HARDI data were eddy current corrected and processed using the eddy and DTIFIT modules from FSL version 6.0.2 (FMRIB, *https://fsl.fmrib.ox.ac.uk/fsl/fslwiki*), yielding isotropic diffusion-weighted images and quantitative maps: mean diffusivity, axial diffusivity, radial diffusivity, and fractional anisotropy (12,13). Eddy correction was performed with outlier replacement. DTIFIT was performed with simple least squares regression. Each sequence was registered and resampled to the 1-mm isotropic resolution 3D space defined by the T2-FLAIR image using automated nonlinear registration (Advanced Normalization Tools) with previously published parameters (6,7). Resampled coregistered data were then skull stripped using a publicly available method (6,7), which can be found at *https://www.github.com/ecalabr/brain_mask/*.

Table 1: Study Population Demographics and Tumor Genetic Characteristics by World Health Organization Tumor Grade

Parameter	All Grades	Grade 2	Grade 3	Grade 4
Total no. of patients	501	56	43	402
No. of men	298/501 (59)	31/56 (55)	26/43 (60)	241/402 (60)
No. of women	203/501 (41)	25/56 (45)	17/43 (40)	161/402 (40)
Mean age (y)*	57 ± 15	42 ± 14	47 ± 14	60 ± 13
IDH mutant	104/501 (21)	46/55 (83)	29/43 (67)	29/402 (7)
MGMT methylated†	255/412 (62)	5/8 (63)	14/22 (64)	236/382 (62)
1p/19q codeletion†	15/410 (4)	11/56 (20)	2/43 (5)	2/311 (<1)

Note.—Unless otherwise noted, data are presented as numbers of patients, with percentages in parentheses. Race and ethnicity data were not available for the study population. *IDH* = isocitrate dehydrogenase, *MGMT* = O^6-methylguanine-DNA methyltransferase.

* Data are presented as means ± SDs.

† Not all patients were tested for MGMT methylation and 1p/19q codeletion. Denominators reflect numbers of patients who were tested for each biomarker.

Tumor Segmentation

Multicompartment tumor segmentation of study data was undertaken as part of the 2021 BraTS challenge as previously described (1). Briefly, image data first underwent automated segmentation using an ensemble model consisting of prior BraTS challenge algorithms. Images were then manually corrected by a group of annotators with varying experience and approved by one of two neuroradiologists with more than 15 years of attending experience each. Segmentation included three major tumor compartments: enhancing tumor, central nonenhancing and/or necrotic tumor, and surrounding FLAIR abnormality (consisting of nonenhancing tumor and associated edema).

Results

Patient Demographic Data

Basic demographic data for all study patients are presented in Table 1. The 501 cases included in the UCSF-PDGM dataset include 56 of 501 (11%) grade 2, 43 of 501 (9%) grade 3, and 402 of 501 (80%) grade 4 tumors. All tumor grade groups consisted of predominantly men: 31 of 56 (55%), 26 of 43 (60%), and 241 of 402 (60%), respectively, for grades 2–4. *IDH* mutations were identified in a majority of grade 2 (46 of 56 [82%]) and grade 3 (29 of 43 [67%]) tumors and a small minority of grade 4 tumors (29 of 402 [7%]), corresponding to a diagnosis of astrocytoma, *IDH*-mutant, WHO grade 4. *MGMT* promoter hypermethylation was detected in 236 of 382 (62%) grade 4 gliomas. 1p/19q codeletion was detected in 11 of 56 (20%) grade 2 tumors and a small minority of grade 3 tumors (two of 43 [5%]), both corresponding to a diagnosis of oligodendroglioma, 1p/19q-codeleted.

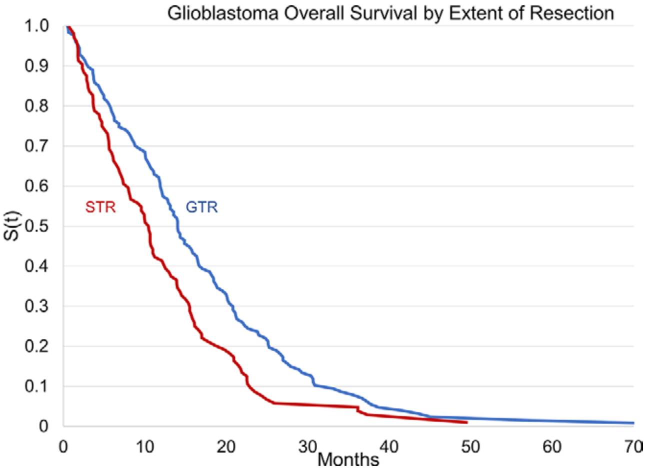

Figure 1: Overall survival of University of California San Francisco Preoperative Diffuse Glioma MRI dataset patients with isocitrate dehydrogenase–wildtype glioblastoma, as a function of time (S[t]), stratified by extent of resection: gross total resection (GTR, blue) versus subtotal resection (STR, red).

Surgical Treatment and Survival Data

Surgical treatment and survival data are included for the entire study cohort. Figure 1 shows overall survival for patients with IDH-wildtype glioblastoma in the cohort, stratified by the extent of resection.

MRI Data

A representative set of images from a single UCSF-PDGM patient is presented in Figure 2. Each patient has skull-stripped coregistered 3D images in 11 different MRI sequences, as well as multicompartment tumor segmentations.

Comparison to Related Datasets

Comparison of the UCSF-PDGM with similar existing resources is presented in Table 2. Comparison datasets include BraTS and TCGA-GBM, as well as the Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme Discovery Study (ie, CPTAC-GBM), the Quantitative Imaging Network Glioblastoma (ie, QIN-GBM) dataset, the American College of Radiology Imaging Network Assessment of Tumor Hypoxia in Glioblastoma using FMISO with PET and MRI (ie, ACRIN-FMISO-Brain) study, and Ivy Glioblastoma Atlas Project (ie, Ivy GAP) datasets (2,8,14–18). Notable differences include a higher number of cases, consistent 3-T MRI protocol, and/or increased number of sequences.

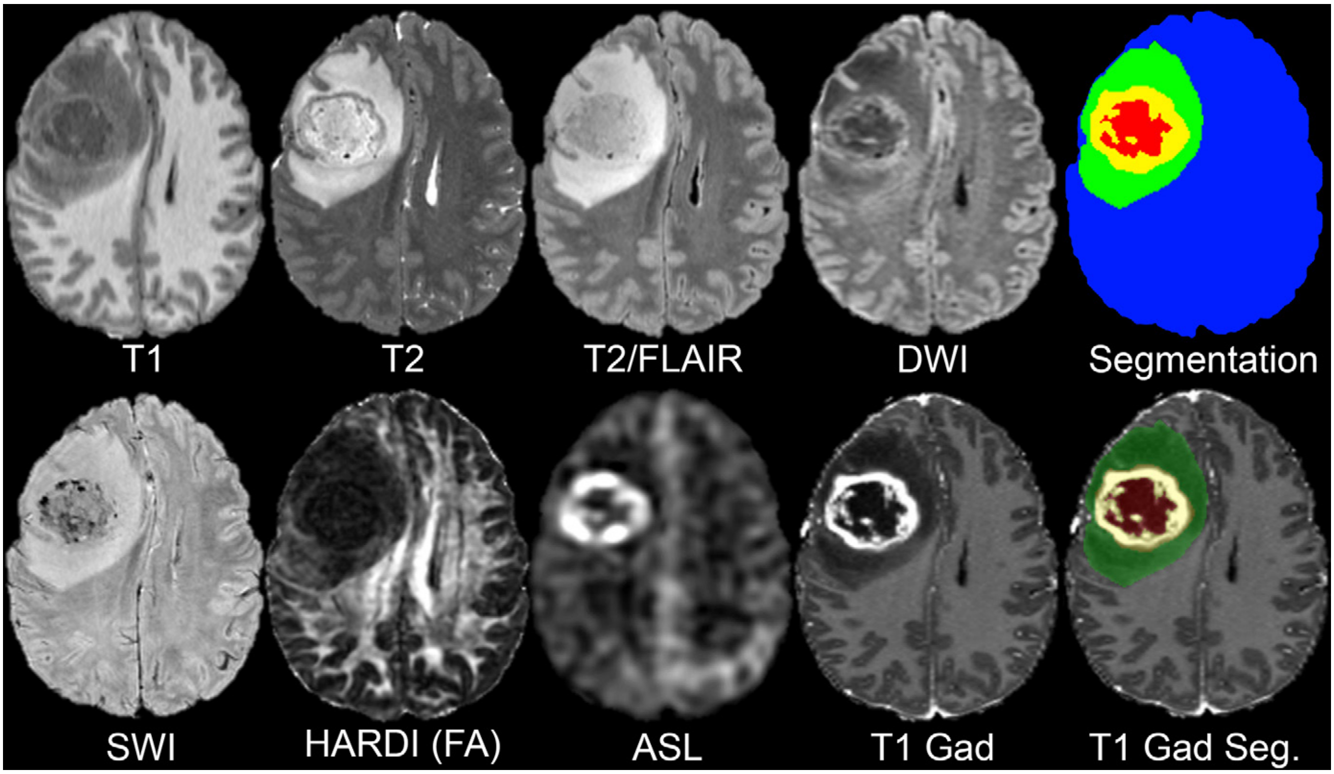

Figure 2: Representative multimodal MRI studies in a 37-year-old man with glioblastoma in the University of California San Francisco Preoperative Diffuse Glioma MRI dataset. ASL = arterial spin labeling perfusion, DWI = isotropic (trace) diffusion-weighted imaging, HARDI (FA) = fractional anisotropy derived from high-angular-resolution diffusion imaging data, Segmentation = multicompartment tumor segmentation (blue = brain, green = fluid-attenuated inversion recovery [FLAIR] abnormality, yellow = enhancing tumor, red = necrotic core), SWI = susceptibility-weighted imaging, T1 = T1-weighted precontrast, T1 Gad = T1-weighted postgadolinium, T1 Gad Seg. = tumor segmentation semitransparent overlay on T1-weighted postgadolinium image, T2 = T2-weighted, T2/FLAIR = T2-weighted FLAIR.

Table 2: Comparison of Selected Publicly Available Preoperative Diffuse Glioma MRI Datasets

Dataset	No. of Cases	Tumor Grade	MRI Sequences	Field Strength	Acquisition Dimension	Segmentation Data	Genetic Data
UCSF-PDGM	501	2–4	T1, T1c, T2, FLAIR, DWI, SWI, HARDI, ASL	3 T	3D[*]	Included	Included[†]
BraTS 2021	1470[‡§]	4	T1, T1c, T2, FLAIR	1.5 T, 3 T	2D and 3D	Included	Included[†]
BraTS 2020	494[‡]	2–4	T1, T1c, T2, FLAIR	1.5 T, 3 T	2D and 3D	Included	Not included
TCGA-GBM	262	4	T1, T1c, T2, FLAIR	1.5 T, 3 T	2D and 3D	Not included	Included[†]
TCGA-LGG	199	2–3	T1, T1c, T2, FLAIR	1.5 T, 3 T	2D and 3D	Not included	Included[†]
CPTAC-GBM	66	4	Variable	1.5 T, 3 T	2D and 3D	Not included	Included[†]
QIN GBM	54	4	T1, T2, FLAIR, MEMPRAGE, DWI, DCE	3 T	2D and 3D	Not included	Not included
ACRIN-FMISO-Brain	45	4	T1, T1c, T2, FLAIR, DWI, DCE, DSC	1.5 T, 3 T	2D and 3D	Not included	Included[†]
Ivy GAP	39	4	Variable	1.5 T, 3 T	2D and 3D	Not included	Included[†]

Note.—Genetic data not available for all patients. ACRIN-FMISO-Brain = American College of Radiology Imaging Network Assessment of Tumor Hypoxia in Glioblastoma using FMISO with PET and MRI study, ASL = arterial spin labeling, BraTS = Multimodal Brain Tumor Segmentation challenge, CPTAC-GMB = Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme Discovery study, DCE = dynamic contrast-enhanced perfusion imaging, DSC = dynamic susceptibility contrast perfusion imaging, DWI = diffusion-weighted imaging, FLAIR = T2-weighted fluid-attenuated inversion recovery, HARDI = high-angular-resolution diffusion imaging, Ivy GAP = Ivy Glioblastoma Atlas Project, MEMPRAGE = multiecho magnetization-prepared rapid gradient echo, QIN GBM = Quantitative Imaging Network Glioblastoma dataset, SWI = susceptibility-weighted imaging, T1 = T1-weighted (no contrast), T1c = T1-weighted postcontrast, TCGA-GBM = The Cancer Genome Atlas' glioblastoma dataset, TCGA-LGG = The Cancer Genome Atlas' Lower Grade Glioma dataset, UCSF-PDGM = University of California San Francisco Preoperative Diffuse Glioma MRI, 3D = three-dimensional, 2D = two-dimensional.

[*] Excludes DWI and HARDI sequences, which are two-dimensional.

[†] At least one genetic biomarker is provided.

[‡] Training and validation cases only. Includes cases from TCGA.

[§] Includes cases from UCSF-PDGM.

Data Availability

As of July 2, 2021, a portion of the UCSF-PDGM dataset is available through the 2021 BraTS challenge dataset *(http://braintumorsegmentation.org/)*. The entire UCSF-PDGM dataset is publicly available via The Cancer Imaging Archive *(https://doi.org/10.7937/tcia.bdgf-8v37)*.

Discussion

The UCSF-PDGM adds to an existing body of publicly available diffuse glioma MRI datasets that can be used in AI research. As MRI-based AI research applications continue to grow, new data are needed to foster the development of new techniques and increase the generalizability of existing algorithms. The UCSF-PDGM not only substantially increases the total number of publicly available diffuse glioma MRI cases, but also provides a unique contribution in terms of MRI technique. The inclusion of 3D sequences and advanced MRI techniques like arterial spin labeling and HARDI provides a new opportunity for researchers to explore the potential use of cutting-edge imaging for AI applications.

In summary, the UCSF-PDGM dataset, particularly when combined with existing publicly available datasets, has the potential to fuel the next phase of radiologic AI research on diffuse gliomas. However, the UCSF-PDGM dataset's potential will only be realized if the radiology AI research community takes advantage of this new data resource for the development of new techniques and discoveries.

Author contributions: Guarantors of integrity of entire study, **E.C., C.P.H.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **E.C., J.E.V.M., U.B., S.B., C.P.H.**; clinical studies, **E.C., J.E.V.M., J.D.R., U.B., S.C.**; experimental studies, **A.M.R., U.B., S.B.**; statistical analysis, **E.C., A.M.R., U.B., S.B.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **E.C.** Research was supported by the National Institutes of Health (NIH) Ruth L. Kirschstein Institutional National Research Service Award under award number T32EB001631 and by the RSNA Research & Education Foundation under grant number RR2011 to author; royalties from Elsevier Academic Press for MRI/DTI Atlas of the Rat Brain, 1st ed. **J.E.V.M.** No relevant relationships. **J.D.R.** Consulting fees from Cortechs.ai; stock or stock options in Cortechs.ai; *Radiology: Artificial Intelligence* trainee editorial board alum. **A.M.R.** *Radiology: Artificial Intelligence* trainee editorial board alum. **U.B.** No relevant relationships. **S.B.** Research reported in this publication was partly supported by the National Cancer Institute (NCI) of the NIH under award number NIH/NCI:U01CA242871. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH. **S.C.** No relevant relationships. **J.T.M.** Grants to author's institution from GE Healthcare and Siemens; royalties or licenses from GE Healthcare; chair of Radiological Society of North America Machine Learning Steering Subcommittee; stock or stock options in Annexon Biosciences; spouse employed by Annexion Biosciences and AbbVie; associate editor for *Radiology: Artificial Intelligence*. **C.P.H.** Consulting fees from GE Healthcare; research travel support from

Siemens Healthineers; member of DataSafety Monitoring Boards for Focused Ultrasound Foundation and uniQure NV.

References

1. Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS Challenge. arXiv:1811.02629 [preprint] http://arxiv.org/abs/1811.02629. Posted November 5, 2018. Accessed February 1, 2019.
2. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26(6):1045–1057.
3. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993–2024.
4. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 2017;4(1):170117.
5. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol 2021;23(8):1231–1251.
6. Calabrese E, Villanueva-Meyer JE, Cha S. A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas. Sci Rep 2020;10(1):11852.
7. Calabrese E, Rudie JD, Rauschecker AM, Villanueva-Meyer JE, Cha S. Feasibility of simulated postcontrast mri of glioblastomas and lower-grade gliomas by using three-dimensional fully convolutional neural networks. Radiol Artif Intell 2021;3(5):e200276.
8. Baid U, Ghodasara S, Bilello M, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv:2107.02314 [preprint] http://arxiv.org/abs/2107.02314. Posted July 5, 2021. Accessed July 9, 2021.
9. Calabrese E, Rudie JD, Rauschecker AM, et al. Combining radiomics and deep convolutional neural network features from preoperative MRI for predicting clinically relevant genetic biomarkers in glioblastoma. Neurooncol Adv 2022;4(1):vdac060.
10. Kline CN, Joseph NM, Grenert JP, et al. Targeted next-generation sequencing of pediatric neuro-oncology patients improves diagnosis, identifies pathogenic germline mutations, and directs targeted therapy. Neuro Oncol 2017;19(5):699–709. [Published correction appears in Neuro Oncol 2017;19(4):601.]
11. Kitange GJ, Carlson BL, Mladek AC, et al. Evaluation of MGMT promoter methylation status and correlation with temozolomide response in orthotopic glioblastoma xenograft model. J Neurooncol 2009;92(1):23–31.
12. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 2011;54(3):2033–2044.
13. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. Neuroimage 2012;62(2):782–790.
14. Gerstner ER, Zhang Z, Fink JR, et al. ACRIN 6684: assessment of tumor hypoxia in newly diagnosed glioblastoma using 18F-FMISO PET and MRI. Clin Cancer Res 2016;22(20):5079–5086.
15. Jafari-Khouzani K, Emblem KE, Kalpathy-Cramer J, et al. Repeatability of cerebral perfusion using dynamic susceptibility contrast MRI in glioblastoma patients. Transl Oncol 2015;8(3):137–146.
16. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. Science 2018;360(6389):660–663.
17. Ratai E-M, Zhang Z, Fink J, et al. ACRIN 6684: Multicenter, phase II assessment of tumor hypoxia in newly diagnosed glioblastoma using magnetic resonance spectroscopy. PLoS One 2018;13(6):e0198548.
18. Prah MA, Stufflebeam SM, Paulson ES, et al. Repeatability of Standardized and Normalized Relative CBV in Patients with Newly Diagnosed Glioblastoma. AJNR Am J Neuroradiol 2015;36(9):1654–1661.