SCIENTIFIC DATA

OPEN

Data Descriptor: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features

Spyridon Bakas[1,2], Hamed Akbari[1,2], Aristeidis Sotiras[1,2], Michel Bilello[1,2], Martin Rozycki[1,2], Justin S. Kirby[3], John B. Freymann[3], Keyvan Farahani[4] & Christos Davatzikos[1,2]

Gliomas belong to a group of central nervous system tumors, and consist of various sub-regions. Gold standard labeling of these sub-regions in radiographic imaging is essential for both clinical and computational studies, including radiomic and radiogenomic analyses. Towards this end, we release segmentation labels and radiomic features for all pre-operative multimodal magnetic resonance imaging (MRI) ($n = 243$) of the multi-institutional glioma collections of The Cancer Genome Atlas (TCGA), publicly available in The Cancer Imaging Archive (TCIA). Pre-operative scans were identified in both glioblastoma (TCGA-GBM, $n = 135$) and low-grade-glioma (TCGA-LGG, $n = 108$) collections via radiological assessment. The glioma sub-region labels were produced by an automated state-of-the-art method and manually revised by an expert board-certified neuroradiologist. An extensive panel of radiomic features was extracted based on the manually-revised labels. This set of labels and features should enable i) direct utilization of the TCGA/TCIA glioma collections towards repeatable, reproducible and comparative quantitative studies leading to new predictive, prognostic, and diagnostic assessments, as well as ii) performance evaluation of computer-aided segmentation methods, and comparison to our state-of-the-art method.

Design Type(s)	parallel group design • data integration objective
Measurement Type(s)	nuclear magnetic resonance assay
Technology Type(s)	MRI Scanner
Factor Type(s)	diagnosis
Sample Characteristic(s)	Homo sapiens • glioma cell

[1]Center for Biomedical Image Computing and Analytics (CBICA), Perelman School of Medicine, University of Pennsylvania, Richards Medical Research Laboratories, Floor 7, 3700 Hamilton Walk, Philadelphia, Pennsylvania 19104, USA. [2]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Richards Medical Research Laboratories, Floor 7, 3700 Hamilton Walk, Philadelphia, Pennsylvania 19104, USA. [3]Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research (FNLCR), Cancer Imaging Program (CIP), 8560 Progress Drive, Frederick, Maryland 21701, USA. [4]Cancer Imaging Program (CIP), National Cancer Institute (NCI), 9609 Medical Center Drive, Bethesda, Maryland 20892, USA. Correspondence and requests for materials should be addressed to S.B. (email: s.bakas@uphs.upenn.edu) or to C.D. (email: Christos.Davatzikos@uphs.upenn.edu).

Background & Summary

Gliomas are the most common primary central nervous system malignancies. These tumors, which exhibit highly variable clinical prognosis, usually contain various heterogeneous sub-regions (i.e., edema, enhancing and non-enhancing core), with variable histologic and genomic phenotypes. This intrinsic heterogeneity of gliomas is also portrayed in their radiographic phenotypes, as their sub-regions are depicted by different intensity profiles disseminated across multimodal MRI (mMRI) scans, reflecting differences in tumor biology. There is increasing evidence that quantitative analysis of imaging features[1–3] extracted from mMRI (i.e., radiomic features), beyond traditionally used clinical measurements (e.g., the largest anterior-posterior, transverse, and inferior-superior tumor dimensions, measured on a subjectively-/arbitrarily-chosen slice), through advanced computational algorithms, leads to advanced image-based tumor phenotyping[4]. Such phenotyping may enable assessment of reflected biological processes and assist in surgical and treatment planning. Furthermore, its correlation with molecular characteristics established radiogenomic research[5–12], leading to improved predictive, prognostic and diagnostic imaging biomarkers[9,12–32], hence yielding the potential benefit towards non-invasive precision medicine[33]. However, it is clear from current literature[26,34–38] that such advanced image-based phenotyping requires accurate annotations of the various tumor sub-regions.

Both clinical and computational studies focusing on such research require the availability of ample data to yield significant associations. Considering the value of big data and the potential of publicly available datasets for increased reproducibility of scientific findings, the National Cancer Institute (NCI) of the National Institutes of Health (NIH) created TCGA (cancergenome.nih.gov) and TCIA[39] (www.cancerimagingarchive.net). TCGA is a multi-institutional comprehensive collection of various molecularly characterized tumor types, and its data are available in NCI's Genomic Data Commons portal (gdc-portal.nci.nih.gov). Building upon NIH's investment in TCGA, the NCI's Cancer Imaging Program approached sites that contributed tissue samples, to obtain corresponding de-identified routine clinically-acquired radiological data and store them in TCIA. These repositories make available multi-institutional, high-dimensional, multi-parametric data of cancer patients, allowing for radiogenomic analysis. However, the data available in TCIA lack accompanying annotations allowing to fully exploit their potential in clinical and computational studies.

Towards addressing this limitation, this study provides segmentation labels and a panel of radiomic features for the glioma datasets included in the TCGA/TCIA repositories. The main goal is to enable imaging and non-imaging researchers to conduct their analyses and extract measurements in a reproducible and repeatable manner, while eventually allowing for comparison across studies. Specifically, the resources of this study provide i) imaging experts with benchmarks to debate their algorithms, and ii) non-imaging experts (e.g., bioinformaticians, clinicians), who do not have the background to interpret and/or appropriately process the raw images, with data helpful to conduct correlative genomic/clinical studies. Following radiological assessment of both the Glioblastoma Multiforme (TCGA-GBM[39], $n = 262$ [Data Citation 1]) and the Low-Grade-Glioma (TCGA-LGG[39], $n = 199$ [Data Citation 2]) collections, we identified 135 and 108 pre-operative mMRI scans, respectively. These scans include at least pre- and post-contrast T1-weighted, T2-weighted, and T2 Fluid-Attenuated Inversion Recovery (FLAIR) volumes. The segmentation labels provided for these scans are divided into two categories: a) computer-aided segmentation labels that could be mainly used for computational comparative studies, and b) manually corrected segmentation labels (approved by an expert board-certified neuroradiologist—M.B.) for use in clinically-oriented analyses, as well as for performance evaluation and training of computational models. The method employed to produce the computer-aided labels is named GLISTRboost[36,38], which was awarded the 1st prize during the International Multimodal Brain Tumor Image Segmentation challenge 2015 (BraTS'15)[36,38,40–56].

The generated data describe two independent datasets [Data Citation 3 and Data Citation 4], one for each glioma collection, and include the computer-aided and manually-revised segmentation labels, coupled with the corresponding co-registered and skull-stripped TCIA scans, in the Neuroimaging Informatics Technology Initiative (NIfTI[57]) format, allowing for direct analysis. Furthermore, a panel of radiomic features is included entailing intensity, volumetric, morphologic, histogram-based, and textural parameters, as well as spatial information and parameters extracted from glioma growth models[58–60]. In consistency with the FAIR (Findable, Accessible, Interoperable, Re-usable) principle, these data are made available through TCIA and should enable both clinical and computational quantitative analyses, as well as serve as a resource for i) educational training of neuroradiology and neurosurgery residents, and ii) performance evaluation of segmentation methods. Furthermore, it could potentially lead to predictive, prognostic, and diagnostic imaging markers suitable for enabling oncological treatment models customized on an individual patient basis (precision medicine), through non-invasive quantification of disease processes.

Methods

Data collection

The complete radiological data of the TCGA-GBM and TCGA-LGG collections consist of 262 [Data Citation 1] and 199 [Data Citation 2] mMRI scans provided from 8 and 5 institutions, respectively (Table 1). The data included in this study describe the subset of the pre-operative baseline scans of these collections, with available MRI modalities of at least T1-weighted pre-contrast (T1), T1-weighted post-

contrast (T1-Gd), T2, and T2-FLAIR (Fig. 1a). Specifically, we considered 135 and 108 pre-operative baseline scans from the TCIA-GBM and TCIA-LGG collections, respectively. Further detailed information on the diversity of the imaging sequences used for this study is included in Table 2 (available online only). This table covers the TCIA institutional identifier, patient information (i.e., age, sex, weight), scanner information (i.e., manufacturer, model, magnetic field strength, station name), as well as specific imaging volume information extracted from the dicom headers (i.e., modality name, series number, accession number, acquisition/study/series date, scan sequence, type, slice thickness, slice spacing, repetition time, echo time, inversion time, imaging frequency, flip angle, specific absorption rate, numbers of slices, pixel dimensions, acquisition matrix rows/columns).

It should be noted that the diversity of the available scans in NCI/NIH/TCIA is driven by the fact that TCIA collected all available scans for subjects whose tissue specimens had passed the quality evaluation of the NCI/NIH/TCGA program. Due to this collection being retrospective all the MRI scans are considered 'standard-of-care', without following any uniform imaging protocol.

Pre-processing

All pre-operative mMRI volumes were re-oriented to the LPS (left-posterior-superior) coordinate system (which is a requirement for GLISTRboost), co-registered to the same T1 anatomic template[61] using affine registration through the Oxford center for Functional MRI of the Brain (FMRIB) Linear Image Registration Tool (FLIRT)[62,63] of FMRIB Software Library (FSL)[64–66], and resampled to 1 mm^3 voxel resolution (Fig. 1b). The volumes of all the modalities for each patient were then skull-stripped using the Brain Extraction Tool (BET)[67,68] from the FSL[64–66] (Fig. 1c). Subsequent skull-stripping, on cases that BET produced insufficient results, was performed using a novel automated method based on a multi atlas registration and label fusion framework[69]. The template library for this task consisted of 216 MRI scans and their brain masks. This library was then used for target specific template selection and subsequent registrations using an existing strategy of MUlti-atlas Segmentation utilizing Ensembles (MUSE)[70]. A final region-growing based processing step, guided by T2, was applied to obtain a brain mask that includes the intra-cranial CSF. The resulted volumes are the ones provided in [Data Citation 3 and Data Citation 4].

For producing the computer-aided segmentation labels, further preprocessing steps included the smoothing of all volumes using a low-level image processing method, namely Smallest Univalue Segment Assimilating Nucleus (SUSAN)[71], in order to reduce high frequency intensity variations (i.e., noise) in regions of uniform intensity profile while preserving the underlying structure (Fig. 1d). The intensity histograms of all modalities of all patients were then matched[72] to the corresponding modality of a single reference patient, using the implemented version in ITK (HistogramMatchingImageFilter).

It should be noted that we did not use any non-parametric, non-uniform intensity normalization algorithm[73–75] to correct for intensity non-uniformities caused by the inhomogeneity of the scanner's magnetic field during image acquisition, as we observed that application of such algorithm obliterated the T2-FLAIR signal (Fig. 1e).

Collection	n	Institutions contributed data—(n)	TCGA ID	Scanner (strength in T)
TCGA-GBM	262	Henry Ford Hospital, Detroit, MI—(74)	TCGA-06	GE (1.5, 3): Genesis Signa, Signa Excite
		CWRU School of Medicine, Cleveland, OH—(38)	TCGA-19	Siemens (1.5, 3): Avanto, Symphony, Verio
		University of California, San Francisco, CA—(32)	TCGA-08	GE (1.5, 3): Genesis Signa, Signa Excite
		Emory University, Atlanta, GA—(31)	TCGA-14	Philips (1.5): Intera Siemens (1.5, 3): Avanto, Trio
		MD Anderson Cancer Center, Houston, TX—(25)	TCGA-02	GE: Genesis Signa, Signa Excite
		Duke University School of Medicine, Durham, NC—(24)	TCGA-12	GE (1.5): Genesis Signa, Signa HDx,Signa Excite Siemens (1.5, 3): Avanto, Trio, Symphony
		Thomas Jefferson University, Philadelphia, PA—(22)	TCGA-76	Philips (1.5, 3): Achieva Siemens (1.5): Magnetom Vision
		Fondazione IRCCS Instituto Neuroligico C. Besta, Milan, Italy—(16)	TCGA-27	Philips (0.5): Intera Siemens (1.5): Avanto
TCGA-LGG	199	St Joseph Hospital/Medical Center, Phoenix, AZ—(98)	TCGA-HT	GE (1.5, 3): Signa Excite, Signa HDx, Signa HDxt
		Henry Ford Hospital, Detroit, MI—(57)	TCGA-DU	Hitachi (1.16): Oasis GE (1.5, 3): Genesis, Signa Excite, Signa HDxt, Philips (1.5, 3): Intera, Ingenia
		Case Western Reserve University, Cleveland, OH—(22)	TCGA-FG	Siemens (1.5, 3): Avanto, Symphony, Skyra, Verio
		Thomas Jefferson University, Philadelphia, PA—(20)	TCGA-CS	GE (1.5): Genesis Signa, Signa HDxt Philips (1.5, 3): Achieva Siemens (1.5): Magnetom Vision
		University of North Carolina, Chapel Hill, NC—(2)	TCGA-EZ	Siemens (3): TrioTim

Table 1. **Source of radiographic data for patients (n) provided in TCIA.**
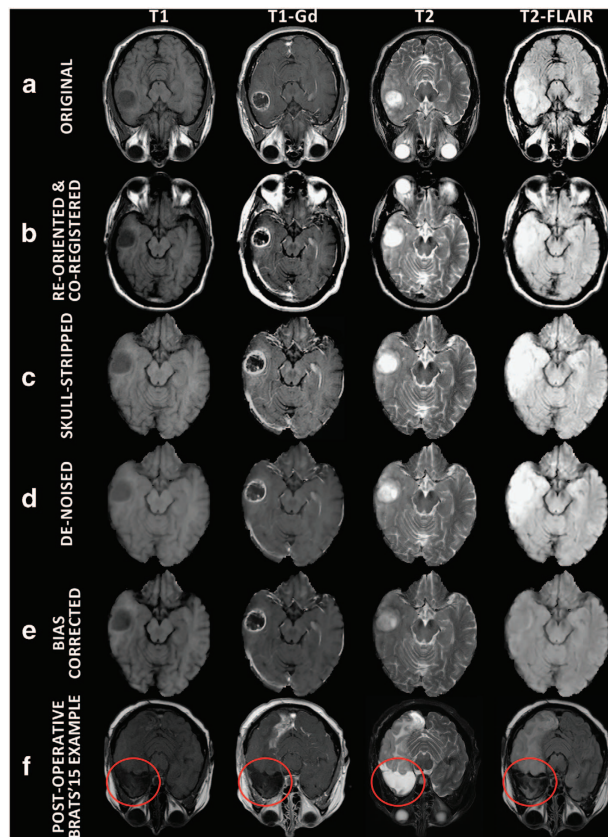
Figure 1. Single slice multimodal (T1, T1-Gd, T2, T2-FLAIR) MRI scans of example subjects. Examples are shown (**a**) in the original TCIA volume; (**b**–**e**) after application of various pre-processing steps; (**f**) for post-operative volumes in the BraTS '15 data. Note that the step shown in (**e**), which is usually used to correct for intensity non-uniformities caused by the inhomogeneity of the scanner's magnetic field during image acquisition, was not applied in the current data as it obliterated the T2-FLAIR signal.

Segmentation labels of glioma sub-regions

Consistent with the BraTS challenge[56] the segmentation labels that we consider in the present study, and make available through TCIA [Data Citation 3 and Data Citation 4], delineate the enhancing part of the tumor core (ET), the non-enhancing part of the tumor core (NET), and the peritumoral edema (ED) (Fig. 2). The ET is described by areas that show hyper-intensity in T1-Gd when compared to T1, but also when compared to normal/healthy white matter (WM) in T1-Gd. Biologically, ET is felt to represent regions where there is leakage of contrast through a disrupted blood-brain barrier that is commonly seen in high grade gliomas. The NET represents non-enhancing tumor regions, as well as transitional/pre-necrotic and necrotic regions that belong to the non-enhancing part of the tumor core (TC), and are typically resected in addition to the ET. The appearance of the NET is typically hypo-intense in T1-Gd when compared to T1, but also when compared to normal/healthy WM in T1-Gd. Finally, the ED is described by hyper-intense signal on the T2-FLAIR volumes.

Computer-aided segmentation approach

The method used in this study to produce the computer-aided segmentation labels for all pre-operative scans of both TCGA-GBM and TCGA-LGG collections is named GLISTRboost[36,38] and it is based on a hybrid generative-discriminative model. The generative part incorporates a glioma growth model[58–60], and is based on an Expectation-Maximization (EM) framework to segment the brain scans into tumor (i.e., ET, NET and ED), as well as healthy tissue labels (i.e., WM, gray matter, cerebrospinal fluid, vessels and cerebellum). The discriminative part is based on a gradient boosting[76,77] multi-class classification scheme, which was trained on BraTS'15 data (www.virtualskeleton.ch/BRATS/Start2015), to refine tumor labels based on information from multiple patients. Lastly, a Bayesian strategy[78] is employed to further refine and finalize the tumor segmentation based on patient-specific intensity statistics from the multiple modalities available. Example segmentation labels are illustrated in Fig. 2.

GLISTRboost[36,38] is based on a modified version of the GLioma Image SegmenTation and Registration (GLISTR)[79] software. GLISTR jointly performs a) the registration of a healthy population probabilistic atlas to brain scans of patients with gliomas using a tumor growth model to account for mass effects, and
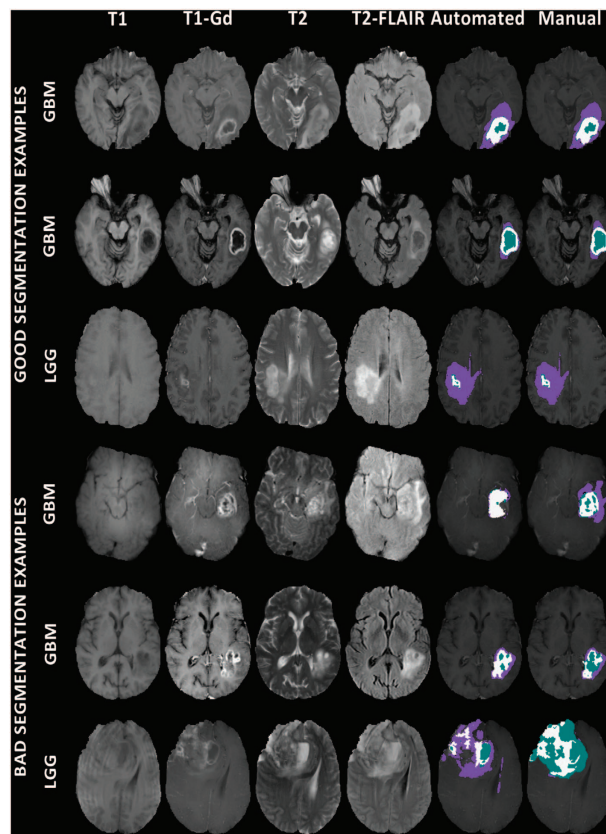
Figure 2. Single slice multimodal MRI scans of example subjects, illustrating all modalities used in GLISTRboost[36,38] and example segmentation labels. The first three rows depict good segmentation examples, whereas the following three depict bad segmentation examples, produced by GLISTRboost[36,38].

b) the segmentation of such scans into healthy and tumor tissues. The whole framework of GLISTR is based on a probabilistic generative model that relies on EM, to recursively refine the estimates of the posteriors for all tissue labels, the deformable mapping to the atlas, and the parameters of the incorporated brain tumor growth model[58–60]. GLISTR was originally designed to tackle cases with solitary GBMs[79–81], and subsequently extended to handle multifocal masses and tumors of complex shapes with heterogeneous texture[82]. Furthermore, the original version of GLISTR[79–82] was based on a single seed-point for each brain tissue label to represent its mean intensity value, while the variance was described by a fixed value for all labels. On the contrary, GLISTRboost incorporates multiple tissue seed-points for each label, to model more accurately the intensity distribution, i.e., mean and variance, for each tissue class. Note that both GLISTR and GLISTRboost take into account only the intensity value of the initialization tissue seed-points on each modality, while they discard spatial information regarding the coordinate position of the respective points. As a consequence, even if the initialized tissue seed-points during two independent segmentation attempts have different coordinates, the output sets of segmentation labels should be identical, given that the modeled intensity distributions during these attempts are the same. In addition to the tissue seed-points, GLISTR and on that account GLISTRboost, requires the definition of a single seed-point and a radius for approximating the center and the bulk volume of each apparent tumor by a sphere. All these seed-points are initialized using the 'Cancer Imaging Phenomics Toolkit' (CaPTk)[83] (www.med.upenn.edu/sbia/captk.html), which has been primarily developed for this purpose, by the Center for Biomedical Image Computing and Analytics (CBICA) of the University of Pennsylvania. Given the tumor seed-point and radius for a tumor, a growth model is initiated by the parametric model of a sphere. This growth model is used to deform a healthy atlas into one with tumor and edema tissues matching the input scans, while approximating the deformations occurred to all brain tissues due to the mass effect of the tumors. A tumor shape prior is estimated by a random-walk-based generative model, which uses the tumor seed-points as initialization cues. This shape prior is systemically incorporated into the EM framework via an empirical Bayes model[82]. Furthermore, a minimum of three initialization seed-points is needed for each brain tissue label, in order to capture the intensity variation and model the intensity distribution across all modalities. Use of multiple seed-points improves the initialization of the EM framework, leading to more accurate segmentation labels, when compared to the single seed-point approach[82]. The output of GLISTR is a

posterior probability map for each tissue label, as well as an integrative label map, which describes a very good 'initial' segmentation of all different tissues within a patient's brain.

This 'initial' segmentation is then refined by taking into account information from multiple patients via a discriminative machine-learning algorithm. Specifically, we used the gradient boosting algorithm[76] to perform voxel-level multi-label classification. Gradient boosting produces a prediction model by combining weak learners in an ensemble. We used decision trees of maximum depth 3 as 'weak learners', which were trained in a sub-sample of the training set, in order to introduce randomness[77]. The sampling rate was set equal to 0.6, while additional randomness was introduced by sampling stochastically a subset of imaging (i.e., radiomic) features at each node. The number of sampled features was set equal to the square root of the total number of features. The algorithm was terminated after 100 iterations.

The set of features used for training our model was extracted volumetrically and consists of i) intensity information, ii) image derivative, iii) geodesic information, iv) texture features, and v) the GLISTR posterior probability maps. The intensity information is summarized by the raw intensity value, I, of each image voxel, v_i, at each modality, m, (i.e., $I(v_i^m)$), as well as by the respective differences among all four modalities, i.e., $I(v_i^{T1})$- $I(v_i^{T1Gd})$, $I(v_i^{T1})$- $I(v_i^{T2})$, $I(v_i^{T1})$- $I(v_i^{T2FLAIR})$, $I(v_i^{T1Gd})$- $I(v_i^{T1})$, $I(v_i^{T1Gd})$- $I(v_i^{T2})$, $I(v_i^{T1Gd})$- $I(v_i^{T2FLAIR})$, $I(v_i^{T2})$- $I(v_i^{T1})$, $I(v_i^{T2})$- $I(v_i^{T1Gd})$, $I(v_i^{T2})$- $I(v_i^{T2FLAIR})$, $I(v_i^{T2FLAIR})$- $I(v_i^{T1})$, $I(v_i^{T2FLAIR})$- $I(v_i^{T1Gd})$, $I(v_i^{T2FLAIR})$- $I(v_i^{T2})$. The image derivative component consists of the Laplacian of Gaussians and the image gradient magnitude. Note that in order to ensure that the intensity-based features are comparable, intensity normalization was performed across subjects based on the median intensity value of the cerebrospinal fluid label, as provided by GLISTR. Geodesic information was used to introduce spatial context information. At any voxel v_i we calculated the geodesic distance from the seed-point at voxel v_s, which was used in GLISTR as the tumor center. The geodesic distance between v_i and v_s was estimated using the fast marching method[84,85] and by taking into account local image gradient magnitude[86]. Furthermore, we used texture features computed from a gray-level co-occurrence matrix (GLCM)[87]. Specifically, these texture features describe first-order statistics (i.e., mean and variance of each modality's intensities within a radius of 2 voxels for each voxel), as well as second-order statistics. To obtain the latter, the image volumes were firstly normalized to 64 different gray levels, and then a bounding box of 5-by-5-by-5 voxels was used for all the voxels of each image as a sliding window. Then, a GLCM was populated by taking into account the intensity values within a radius of 2 pixels and for the 26 main 3D directions to extract the energy, entropy, dissimilarity, homogeneity (i.e., inverse difference moment of order 2), and inverse difference moment of order 1. These features were computed for each direction and their average was used. To avoid overfitting, the gradient boosting machine was trained using simultaneously both LGG and GBM training data of BraTS'15, in a 54-fold cross-validation setting (allowing for using a one out of the 54 available LGGs of the BraTS'15 training data, within each fold).

Finally, the segmentation results were further refined for each patient separately, by assessing the local intensity distribution of the segmentation labels and updating their spatial configuration based on a probabilistic model[78]. The intensity distributions of the WM, ED, NET and ET, were populated separately using the corresponding voxels of posterior probability equal to 1, as given by GLISTR. Histogram normalization was then performed for the 3 pair-wise distributions considered; ED versus WM in T2-FLAIR, ET versus ED in T1-Gd, and ET versus NET in T1-Gd. Maximum likelihood estimation was used to model the class-conditional probability densities ($Pr(I(v_i)|Class)$) by a distinct Gaussian model for each class. In all pair-wise comparisons described before, the former tissue is expected to be brighter than the latter. Voxels of each class with spatial proximity smaller than 4 voxels to the voxels of the paired class, were evaluated by assessing their intensity $I(v_i)$ and comparing the ('$Pr(I(v_i)|Class_1)$ with $Pr(I(v_i)|Class_2)$). The voxel v_i was then classified into the tissue class with the larger conditional probability. This is equivalent to a classification based on Bayes' Theorem with equal priors for the two classes, i.e., $Pr(Class_1) = Pr(Class_2) = 0.5$.

Manual revision

The output of GLISTRboost segmentation is expected to yield labels for ET, NET, and ED. However, some gliomas, especially LGG, do not exhibit much contrast enhancement, or ED. Biologically, LGGs may have less blood-brain barrier disruption (leading to less leak of contrast during the scan), and may grow at a rate slow enough to avoid significant edema formation, which results from rapid disruption, irritation, and infiltration of normal brain parenchyma by tumor cells. As such, manual revision of the segmentation labels was performed, particularly for LGG cases lacking ET or ED regions. Specifically, after taking all the above into consideration, in scans of LGGs without an apparent ET area we consider only the NET and ED labels (Fig. 3a,d), whereas in LGG scans without ET and without obvious texture differences across modalities we consider only the NET label, allowing for distinguishing between normal and abnormal brain tissue (Fig. 3e). The difficulty in calculating the accurate boundaries between tumor and healthy tissue in the operating room is reflected in the segmentation labels as well; there is high uncertainty among neurosurgeons, neuroradiologists, and imaging scientists in delineating these boundaries. Therefore, small regions within the segmented labels that were ambiguous of their exact classification, were left as segmented by GLISTRboost.

Manual revisions/corrections applied in the computer-aided segmentation labels include: i) obvious under- or over-segmented ED/ET/NET regions (Fig. 3d–g), ii) voxels classified as ED within the tumor core (Fig. 3b,c,g), iii) unclassified voxels within the tumor core (Fig. 3c–g), iv) voxels classified as NET
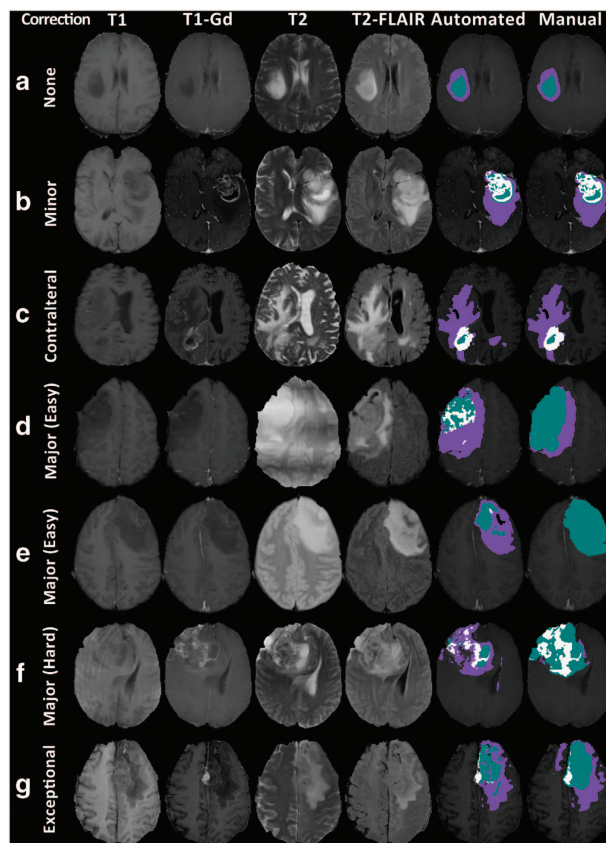
Figure 3. Single slice multimodal MRI scans of example subjects, illustrating all modalities used in GLISTRboost[36,38] and examples of the computer-aided (automated) and the manually-revised (manual) segmentation labels. The type of corrections applied during the manual-revision of the segmentation labels is also shown in the left side of each row; (**a**) no correction, (**b**) minor corrections, (**c**) corrections in the contralateral edema, (**d**) major (easy) correction of LGG without ET, (**e**) major (easy) correction of LGG without ET or ED, (**f**) major (hard) corrections, (**g**) exceptional subject (TCGA-DU-7304) that could have a meningioma in the midline as the apparent lesion seems to raise from the dura.

outside the tumor core. Contralateral and periventricular regions of T2-FLAIR hyper-intensity were excluded from the ED region (Fig. 3c,f), unless they were contiguous with peritumoral ED (Fig. 3g—addition of apparent contralateral ED), as these areas are generally considered to represent chronic microvascular changes, or age-associated demyelination, rather than tumor infiltration[88].

Radiomic features panel

An extensive panel of more than 700 radiomic features is extracted volumetrically (in 3D), based on the manually-revised labels of each tumor sub-region that comprised i) intensity, ii) volumetric[89], iii) morphologic[90–93], iv) histogram-based[31], and v) textural parameters, including features based on wavelets[94], GLCM[87], Gray-Level Run-Length Matrix (GLRLM)[93,95–98], Gray-Level Size Zone Matrix (GLSZM)[95–97,99], and Neighborhood Gray-Tone Difference Matrix (NGTDM)[100], as well as vi) spatial information[101], and vii) glioma diffusion properties extracted from glioma growth models[58–60], that are already evaluated as having predictive and prognostic value[30–32,102,103]. The specific features provided are all shown in Table 3 (available online only).

These radiomic features are provided on an 'as-is' basis, and are distinct from the panel of features used in GLISTRboost. The biological significance of these individual radiomic features remains unknown, but we include them here to facilitate research on their association with molecular markers, clinical outcomes, treatment responses, and other endpoints, by researchers without sufficient computational background to extract such features. Although researchers can derive their own radiomic features from our segmentation labels, and the corresponding images we included a collection of features that have been shown in various studies to relate to clinical outcome[31] and underlying tumor molecular characteristics[30,32]. Note that the radiomic features we provide are extracted from the denoised images, and the users might also want to consider extracting features from the unsmoothed images provided in [Data Citation 3 and Data Citation 4].

Code availability

All software tools used for pre-processing, initialization, and generation of the hereby described segmentation labels are based on publicly available tools. Specifically, the tools used for the pre-processing steps of skull-stripping (BET)[67,68] and co-registration (FLIRT)[62,63] are publicly available from the FMRIB Software Library (FSL)[64–66], in: fsl.fmrib.ox.ac.uk. The software used for the further skull-stripping approaches, i.e., Multi-Atlas Skull-Stripping (MASS)[69] and MUSE[70], are publicly available in www.med.upenn.edu/sbia/mass.html and www.med.upenn.edu/sbia/muse.html, respectively.

We developed CaPTk[83] as a toolkit to facilitate translation of complex research algorithms into clinical practice, by enabling operators to conduct quantitative analyses without requiring substantial computational background. Towards this end CaPTk is a dynamically growing software platform, with various integrated applications, allowing 1) interactive definition of coordinates and regions, 2) generic image analysis (e.g., registration, feature extraction), and 3) specialized analysis algorithms (e.g., identification of genetic mutation imaging markers[12]). Specifically for this study, CaPTk was used to 1) manually initialize seed-points required for the initialization of GLISTRboost[36,38], 2) apply the de-noising approach (SUSAN)[71] used for smoothing images before their input to GLISTRboost, as well as 3) to extract the radiomic features released in TCIA [Data Citation 3 and Data Citation 4]. The exact version used for initializing the required seed-points in this study was released on the 14th of October 2016 and the code source, as well as executable installers, are available in: www.med.upenn.edu/sbia/captk.html.

Finally, our segmentation approach, GLISTRboost[36,38], has been made available for public use through the Online Image Processing Portal (IPP—ipp.cbica.upenn.edu) of the CBICA. CBICA's IPP allows users to perform their data analysis using integrated algorithms, without any software installation, whilst also using CBICA's High Performance Computing resources. It should be noted that we used the Python package scikit-learn[104] for the implementation of the gradient boosting algorithm.

Data Records

We selected only the pre-operative multimodal scans of the TCGA-GBM [Data Citation 1] and TCGA-LGG [Data Citation 2] glioma collections, from the publicly available TCIA repository. The generated data, which is made publicly available through TCIA's Analysis Results Directory (wiki.cancerima-gingarchive.net/x/sgH1) [Data Citation 3 and Data Citation 4], comprise pre-operative baseline re-oriented, co-registered and skull-stripped mMRI scans together with their corresponding computer-aided and manually-revised segmentation labels in NIfTI[57] format. We have further enriched the file containers to include an extensive panel of radiomic features, which we hope may facilitate radiogenomic research using the TCGA portal, as well as comparison of segmentation methods, even among those scientists without image analysis resources.

A subset of the pre-operative scans included in the generated data [Data Citation 3 and Data Citation 4] was also part of the BraTS'15 dataset (Table 4 (available online only)), which were skull-stripped, co-registered to the same anatomical template and resampled to 1 mm³ voxel resolution by the challenge organizers. For this subset, we provide the identical MRI volumes as provided by the BraTS'15 challenge, allowing other researchers to compare their segmentation labels to the leaderboard of the BraTS'15 challenge. Furthermore, the manually-revised segmentation labels provided in [Data Citation 3 and Data Citation 4] are included in the datasets of the BraTS'17 challenge, for benchmarking computational segmentation algorithms against tumor delineation validated by expert neuroradiologists, allowing for repeatable research.

Technical Validation

Data collection

Our expert board-certified neuroradiologist (M.B.) identified 135 and 108 pre-operative baseline scans of the TCGA-GBM and the TCGA-LGG glioma collections, via radiological assessment and while blinded to the glioma grade. Since it is not always easy to determine if a scan is pre-operative or post-operative only by visually assessing MRI volumes, and the radiological reports were not available through the TCGA/TCIA repositories, whenever we mention 'pre-operative scans' in this study, we refer to those that radiographically do not have clear evidence of prior instrumentation. Specifically, the main evaluation criterion for classifying scans as pre-operative, was absence of obvious skull defect and of operative cavity through either biopsy or resection.

We note that a mixed (pre- and post-operative) subset of 223 and 59 scans from the TCIA-GBM and TCIA-LGG datasets, respectively, were included in the BraTS'15 challenge, as part of their training ($n_{GBM} = 200$, $n_{LGG} = 44$) and testing ($n_{GBM} = 23$, $n_{LGG} = 15$) datasets, via the Virtual Skeleton Database (VSD) platform[56,105] (www.virtualskeleton.ch). Since an explicit distinction as pre- or post-operative was not provided for the BraTS'15 dataset, we conducted the radiological assessment of the complete TCIA collections, blind to whether a scan was part of the BraTS challenge, and only included the BraTS'15 volumes identified as pre-operative (Fig. 1f) (Table 4 (available online only)).

Segmentation labels

The segmentation method we developed to produce the segmentation labels, GLISTRboost[36,38], was ranked as the best performing method and awarded the 1st prize during the International Multimodal Brain Tumor Image Segmentation challenge 2015 (BraTS'15)[36,38,40–56]. Specifically, the performance of

the computer-aided segmentation labels was assessed during the challenge for the test data, through the VSD platform, by comparing the voxel-level overlap between the segmentation labels produced by GLISTRboost and the ground truth labels provided by the BraTS organizers in three regions, i.e., the whole tumor (WT), the tumor core (TC) and the ET. The WT describes the union of the ET, NET and ED, whereas the TC describes the union of the ET and NET. The performance was quantitatively validated by the per-voxel overlap between respective regions, using the DICE coefficient and the robust Hausdorff distance (95% quantile), as suggested by the BraTS'15 organizers[56]. The former metric takes values between 0 and 1, with higher values corresponding to increased overlap, whereas lower values in the latter correspond to segmentation labels closer to the gold standard labels. Note that the quantitative results for the test data were not provided to the participants, until a manuscript summarizing the results of BraTS'14 and BraTS'15 is published. However, for reporting the performance of our method, we report here the cross-validated results of the same metrics used in BraTS'15 for the subset of GBM subjects included in the training set of BraTS'15 and identified as pre-operative in this study (Table 4 (available online only)). The median DICE values with their corresponding inter-quartile ranges (IQR) for the three evaluated regions, i.e., WT, TC, ET, were equal to 0.92 (IQR: 0.88–0.94), 0.88 (IQR: 0.81–0.93) and 0.88 (IQR: 0.81–0.91), respectively. Equivalently, the 95th percentile of the Hausdorff distance for WT, TC and ET were equal to 3.61 (IQR: 2.39–8.15), 4.06 (IQR: 2.39–7.29), and 2 (IQR: 1.41–2.83), respectively.

Furthermore, we used the Jaccard coefficient, in order to quantify the difference between the computer-aided segmentation labels produced for all the scans identified as pre-operative and all the manually-corrected labels that we provide in [Data Citation 3 and Data Citation 4]. The median (mean ± std.dev) Jaccard values for the three regions of interest i.e., WT, TC, ET, were equal to 0.96 (0.93 ± 0.1), 0.87 (0.78 ± 0.23), and 0.86 (0.73 ± 0.29), respectively.

Manual correction

The classification scheme of segmentation labels considered for the manual corrections of the GBM and LGG cases describe all three segmentation labels (i.e., ET, NET, and ED) for both GBMs and LGGs with an apparent ET area. However, whenever we note LGG scans without an apparent ET area and not obvious texture differences, we considered only the NET label, allowing for distinguishing between normal and abnormal brain tissue, as slowly growing tumors are not expected to induce ED. Furthermore, due to high uncertainty (reported by neurosurgeons, neuroradiologists, and imaging scientists) on the exact boundaries between the various tumor labels, particularly between NET and ED, small regions that visual assessment was ambiguous of their exact classification, were left as segmented by GLISTRboost.

Manual revisions/corrections applied in the computer-aided segmentation labels comprise: i) obvious under- or over-segmented ED/ET/NET regions (Fig. 3d–g), ii) voxels classified as ED within the tumor core (Fig. 3b,c,g), iii) unclassified voxels within the tumor core (Fig. 3c–g), iv) voxels classified as NET outside the tumor core. Note that during the manual corrections only peritumoral ED was considered, and both contralateral, and periventricular ED was deleted (Fig. 3c,f), unless it was a clear continuation of the peritumoral ED, in which cases was added (Fig. 3g). The rationale for this is that contralateral and periventricular white matter hyper-intensities regions might be considered pre-existing conditions, related to small vessel ischemic disease, especially in older patients.

The scheme followed for the manual correction included two computational imaging scientists (S.B., A.S.) and a medical doctor (H.A.) working in medical image computing and analysis for 10, 12 and 8 years, respectively. These operators corrected mislabeled voxels following the rules set by our expert board-certified neuroradiologist (M.B.) with 14 years of experience. The corrected labels were then iteratively re-evaluated by the latter and re-iterated until they were satisfactory segmented.

References

1. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magnetic Resonance Imaging* **30**, 1234–1248 (2012).
2. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* **48**, 441–446 (2012).
3. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **278**, 563–577 (2015).
4. Hu, L. S. *et al.* Multi-Parametric MRI and Texture Analysis to Visualize Spatial Histologic Heterogeneity and Tumor Extent in Glioblastoma. *PLOS ONE* **10**, e0141506 (2015).
5. Rutman, A. M. & Kuo, M. D. Radiogenomics: Creating a link between molecular diagnostics and diagnostic imaging. *European Journal of Radiology* **70**, 232–241 (2009).
6. Jaffe, C. C. Imaging and Genomics: Is There a Synergy? *Radiology* **264**, 329–331 (2012).
7. Proud, C. Radiogenomics: The Promise of Personalized Treatment in Radiation Oncology? *Clinical Journal of Oncology Nursing* **18**, 185 (2014).
8. Rosenstein, B. S. *et al.* Radiogenomics: Radiobiology Enters the Era of Big Data and Team Science. *International Journal of Radiation Oncology • Biology • Physics* **89**, 709–713 (2014).
9. Ellingson, B. M. Radiogenomics and imaging phenotypes in glioblastoma: novel observations and correlation with molecular characteristics. *Curr Neurol Neurosci Rep* **15**, 506 (2015).
10. Mahajan, A., Moiyadi, A. V., Jalali, R. & Sridhar, E. Radiogenomics of glioblastoma: a window into its imaging and molecular variability. *Cancer Imaging* **15**, P14 (2015).
11. Mazurowski, M. A. Radiogenomics: What It Is and Why It Is Important. *Journal of the American College of Radiology* **12**, 862–866 (2015).

12. Bakas, S. *et al.* In vivo detection of EGFRvIII in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep peritumoral infiltration: the φ-index. *Clinical Cancer Research*, doi:10.1158/1078-0432.ccr-16-1871 (2017).

13. Tykocinski, E. S. *et al.* Use of magnetic perfusion-weighted imaging to determine epidermal growth factor receptor variant III expression in glioblastoma. *Neuro-oncology* **14,** 613–623 (2012).

14. Gevaert, O. *et al.* in Proceedings of the 103rd Annual Meeting of the American Association for Cancer Research.

15. Gutman, D. A. *et al.* MR Imaging Predictors of Molecular Profile and Survival: Multi-institutional Study of the TCGA Glioblastoma Data Set. *Radiology* **267,** 560–569 (2013).

16. Mazurowski, M. A., Desjardins, A. & Malof, J. M. Imaging descriptors improve the predictive power of survival models for glioblastoma patients. *Neuro-Oncology* **15,** 1389–1394 (2013).

17. Akbari, H. *et al.* Pattern Analysis of Dynamic Susceptibility Contrast MRI Reveals Peritumoral Tissue Heterogeneity. *Radiology* **273,** 502–510 (2014).

18. Gevaert, O. *et al.* Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology* **273,** 168–174 (2014).

19. Gill, B. J. *et al.* MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proc Natl Acad Sci USA* **111,** 12550–12555 (2014).

20. Jain, R. *et al.* Outcome Prediction in Patients with Glioblastoma by Using Imaging, Clinical, and Genomic Biomarkers: Focus on the Nonenhancing Component of the Tumor. *Radiology* **272,** 484–493 (2014).

21. Arevalo-Perez, J. *et al.* T1-Weighted Dynamic Contrast-Enhanced MRI as a Noninvasive Biomarker of Epidermal Growth Factor Receptor vIII Status. *AJNR Am J Neuroradiol* **36,** 2256–2261 (2015).

22. Bakas, S. *et al.* Identification of Imaging Signatures of the Epidermal Growth Factor Receptor Variant III (EGFRvIII) in Glioblastoma. *Neuro-oncology* **17,** 154–154 (2015).

23. Bonekamp, D. *et al.* Association of overall survival in patients with newly diagnosed glioblastoma with contrast-enhanced perfusion MRI: Comparison of intraindividually matched T1- and T2*-based bolus techniques. *Journal of Magnetic Resonance Imaging* **42,** 87–96 (2015).

24. Itakura, H. *et al.* Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. *Sci Transl Med* **7,** 303ra138 (2015).

25. Nicolasjilwan, M. *et al.* Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. *Journal of Neuroradiology* **42,** 212–221 (2015).

26. Rios Velazquez, E. *et al.* Fully automatic GBM segmentation in the TCGA-GBM dataset: Prognosis and correlation with VASARI features. *Scientific Reports* **5,** 16822 (2015).

27. Akbari, H. *et al.* Imaging Surrogates of Infiltration Obtained Via Multiparametric Imaging Pattern Analysis Predict Subsequent Location of Recurrence of Glioblastoma. *Neurosurgery* **78,** 572–580 (2016).

28. Bakas, S. *et al.* Highly-expressed wild-type EGFR and EGFRvIII mutant glioblastomas have similar MRI signature, consistent with deep peritumoral infiltration. *Neuro-Oncology* **18,** vi125 (2016).

29. Batmanghelich, N., Dalca, A., Quon, G., Sabuncu, M. & Golland, P. Probabilistic Modeling of Imaging, Genetics and Diagnosis. *IEEE Trans Med Imaging* **35,** 1765–1779 (2016).

30. Binder, Z. A. *et al.* Extracellular EGFR289 activating mutations confer poorer survival and exhibit radiographic signature of enhanced motility in primary glioblastoma. *Neuro-Oncology* **18,** vi105–vi106 (2016).

31. Macyszyn, L. *et al.* Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-oncology* **18,** 417–425 (2016).

32. Rathore, S., Akbari, H., Rozycki, M., Bakas, S. & Davatzikos, C. Imaging pattern analysis reveals three distinct phenotypic subtypes of GBM with different survival rates. *Neuro-Oncology* **18,** vi128 (2016).

33. Aerts, H. J. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. *JAMA oncology* **2,** 1636–1642 (2016).

34. Kaus, M. R. *et al. Medical Image Computing and Computer-Assisted Intervention—MICCAI'99: Second International Conference, Cambridge, UK, September 19-22, 1999. Proceedings* (eds Taylor C. & Colchester A.)1–10 (Springer Berlin Heidelberg, 1999).

35. Akkus, Z. *et al.* Semi-automated segmentation of pre-operative low grade gliomas in magnetic resonance imaging. *Cancer Imaging* **15,** 12 (2015).

36. Bakas, S. *et al. Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 5–12 (Technische Universität München (T.U.M.), 2015).

37. Simi, V. R. & Joseph, J. Segmentation of Glioblastoma Multiforme from MR Images—A comprehensive review. *The Egyptian Journal of Radiology and Nuclear Medicine* **46,** 1105–1110 (2015).

38. Bakas, S. *et al.* GLISTRboost: Combining Multimodal MRI Segmentation, Registration, and Biophysical Tumor Growth Modeling with Gradient Boosting Machines for Glioma Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* **9556,** 144–155 (2016).

39. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging* **26,** 1045–1057 (2013).

40. Agn, M., Puonti, O., Law, I., Rosenschold, P. M. A. & Leemput, K. V. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 1–4 (Technische Universität München (T.U.M.), 2015).

41. Dvorak, P. & Menze, B. H. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 13–24 (Technische Universität München (T.U.M.), 2015).

42. Haeck, T., Maes, F. & Suetens, P. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*)25–28 (Technische Universität München (T.U.M.), 2015).

43. Havaei, M., Dutil, F., Pal, C., Larochelle, H. & Jodoin, P.-M. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 29–33 (Technische Universität München (T.U.M.), 2015).

44. Hoogi, A., Lee, A., Bharadwaj, V. & Rubin, D. L. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 34–37 (Technische Universität München (T.U.M.), 2015).

45. Maier, O., Wilms, M. & Handels, H. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015),* (eds Menze B. H. *et al.*) 38–41 (Technische Universität München (T.U.M.), 2015).

46. Malmi, E., Parambath, S., Peyrat, J.-M., Abinahed, J. & Chawla, S. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 42–47 (Technische Universität München (T.U.M.), 2015).

47. Meier, R., Karamitsou, V., Habegger, S., Wiest, R. & Reyes, M. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 48–51 (Technische Universität München (T.U.M.), 2015).

48. Pereira, S., Pinto, A., Alves, V. & Silva, C. A. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 52–55 (Technische Universität München (T.U.M.), 2015).

49. Rao, V., Sarabi, M. S. & Jaiswal, A. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 56–59 (Technische Universität München (T.U.M.), 2015).

50. Vaidhya, K., Santhosh, R., Thirunavukkarasu, S., Alex, V. & Krishnamurthi, G. *In Proceedings of the Multimodal Brain Tumor Image Segmentation Challenge held in conjunction with MICCAI 2015 (MICCAI-BRATS 2015)* (eds Menze B. H. *et al.*) 60–64 (Technische Universität München (T.U.M.), 2015).

51. Agn, M., Puonti, O., Rosenschöld, P. M. A., Law, I. & Leemput, K. V. Brain Tumor Segmentation Using a Generative Model with an RBM Prior on Tumor Shape. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* **9556,** 168–180 (2016).

52. Havaei, M., Dutil, F., Pal, C., Larochelle, H. & Jodoin, P.-M. A Convolutional Neural Network Approach to Brain Tumor Segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* **9556,** 195–208 (2016).

53. Maier, O., Wilms, M. & Handels, H. Image Features for Brain Lesion Segmentation Using Random Forests. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* **9556,** 119–130 (2016).

54. Pereira, S., Pinto, A., Alves, V. & Silva, C. A. Deep Convolutional Neural Networks for the Segmentation of Gliomas in Multi-sequence MRI. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* **9556,** 131–143 (2016).

55. Pereira, S., Pinto, A., Alves, V. & Silva, C. A. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging* **35,** 1240–1251 (2016).

56. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34,** 1993–2024 (2015).

57. Cox, R. W. *et al.* in Tenth Annual Meeting of the Organization for Human Brain Mapping.

58. Hogea, C., Biros, G., Abraham, F. & Davatzikos, C. A robust framework for soft tissue simulations with application to modeling brain tumor mass effect in 3D MR images. *Physics in medicine and biology* **52,** 6893–6908 (2007).

59. Hogea, C., Davatzikos, C. & Biros, G. Brain-tumor interaction biophysical models for medical image registration. *SIAM Journal on Scientific Computing* **30,** 3050–3072 (2008).

60. Hogea, C. S., Biros, G. & Davatzikos, C. An image-driven parameter estimation problem for a reaction-diffusion glioma growth model with mass effects. *Journal of Mathematical Biology* **56,** 793–825 (2008).

61. Rohlfing, T., Zahr, N. M., Sullivan, E. V. & Pfefferbaum, A. The SRI24 Multi-Channel Atlas of Normal Adult Human Brain Structure. *Human brain mapping* **31,** 798–819 (2010).

62. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Medical image analysis* **5,** 143–156 (2001).

63. Jenkinson, M., Bannister, P. R., Brady, J. M. & Smith, S. M. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17,** 825–841 (2002).

64. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23,** 208–219 (2004).

65. Woolrich, M. W. *et al.* Bayesian analysis of neuroimaging data in FSL. *NeuroImage* **45,** S173–S186 (2009).

66. Jenkinson, M., Beckmann, C. F., Behrens, T. E. & Woolrich, M. W. FSL. *NeuroImage* **62,** 782–790 (2012).

67. Smith, S. M. Fast robust automated brain extraction. *Hum Brain Mapp* **17,** 143–155 (2002).

68. Jenkinson, M., Pechaud, M. & Smith, S. in Eleventh Annual Meeting of the Organization for Human Brain Mapping.

69. Doshi, J., Erus, G., Ou, Y., Gaonkar, B. & Davatzikos, C. Multi-Atlas Skull-Stripping. *Acad Radiol* **20,** 1566–1576 (2013).

70. Doshi, J. *et al.* MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage* **127,** 186–195 (2016).

71. Smith, S. M. & Brady, J. M. SUSAN—a new approach to low level image processing. *International Journal of Computer Vision* **23,** 45–78 (1997).

72. Nyul, L. G., Udupa, J. K. & Zhang, X. New Variants of a Method of MRI Scale Standardization. *IEEE transactions on medical imaging* **19,** 143–150 (2000).

73. Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging* **17,** 87–97 (1998).

74. Tustison, N. J. *et al.* N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging* **29,** 1310–1320 (2010).

75. Larsen, C. T., Iglesias, J. E. & Van Leemput, K. *Bayesian and grAphical Models for Biomedical Imaging: First International Workshop, BAMBI 2014, Cambridge, MA, USA, September 18, 2014, Revised Selected Papers* (eds Jorge Cardoso M. *et al.*) 1–12 (Springer International Publishing, 2014).

76. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232 (2001).

77. Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38,** 367–378 (2002).

78. Bakas, S. *et al.* Fast Semi-Automatic Segmentation of Focal Liver Lesions in Contrast-Enhanced Ultrasound, Based on a Probabilistic Model. *TCIV Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **5,** 329–338 (2017).

79. Gooya, A. *et al.* GLISTR: Glioma Image Segmentation and Registration. *IEEE transactions on medical imaging* **31,** 1941–1954 (2012).

80. Gooya, A., Pohl, K., Billelo, M., Biros, G. & Davatzikos, C. in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* Vol. 14 532-540 (Canada, 2011).

81. Gooya, A., Biros, G. & Davatzikos, C. Deformable Registration of Glioma Images Using EM Algorithm and Diffusion Reaction Modeling. *IEEE transactions on medical imaging* **30,** 375–390 (2011).

82. Kwon, D., Shinohara, R. T., Akbari, H. & Davatzikos, C. in *Medical Image Computing and Computer-Assisted Intervention-—MICCAI 2014.* 763-770 (Lecture Notes in Computer Science).

83. Pati, S. *et al.* in *Radiological Society of North America Scientific Assembly and Annual Meeting (RSNA)* (Chicago, IL, 2016).

84. Sethian, J. A. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences of the United States of America* **93,** 1591–1595 (1996).

85. Deschamps, T. & Cohen, L. D. Fast extraction of minimal paths in 3D images and applications to virtual endoscopy1. *Medical Image Analysis* **5,** 281–299 (2001).

86. Gaonkar, B. *et al.* Automated tumor volumetry using computer-aided image segmentation. *Acad Radiol* **22,** 653–661 (2015).

87. Haralick, R. M., Shanmugam, K. & Dinstein, I. H. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **3,** 610–621 (1973).

88. Haller, S. *et al.* Do brain T2/FLAIR white matter hyperintensities correspond to myelin loss in normal aging? A radiologic-neuropathologic correlation study. *Acta Neuropathologica Communications* **1,** 14–14 (2013).

89. Vallières, M., Freeman, C., Skamene, S. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in medicine and biology* **60,** 5471 (2015).

90. Max, J. Quantizing for minimum distortion. *IRE Transactions on Information Theory* **6,** 7–12 (1960).

91. Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory* **28,** 129–137 (1982).
92. Li, Q. & Griffiths, J. G. *in Geometric modeling and processing*, proceedings 335–340 (IEEE, 2004).
93. Thibault, G. *et al.* Shape and texture indexes application to cell nuclei classification. *International Journal of Pattern Recognition and Artificial Intelligence* **27,** 1357002 (2013).
94. Collewet, G., Strzelecki, M. & Mariette, F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magnetic Resonance Imaging* **22,** 81–91 (2004).
95. Galloway, M. M. Texture analysis using grey level run lengths. *Computer Graphics and Image Processing* **4,** 172–179 (1975).
96. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11,** 415–419 (1990).
97. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level—run length distributions. *Pattern Recognition Letters* **12,** 497–502 (1991).
98. Tang, X. Texture information in run-length matrices. *IEEE transactions on image processing* **7,** 1602–1609 (1998).
99. Thibault, G. *Indices de forme et de texture: de la 2D vers la 3D: application au classement de noyaux de cellules.* Aix Marseille2 (2009).
100. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics* **19,** 1264–1274 (1989).
101. Bilello, M. *et al.* Population-based MRI atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma. *NeuroImage. Clinical* **12,** 34–40 (2016).
102. Assefa, D. *et al.* Robust texture features for response monitoring of glioblastoma multiforme on T1-weighted and T2-FLAIR MR images: A preliminary investigation in terms of identification and segmentation. *Medical physics* **37,** 1722–1736 (2010).
103. Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **5,** 4006 (2014).
104. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).
105. Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R. & Büchler, P. The Virtual Skeleton Database: An Open Access Repository for Biomedical Research and Collaboration. *Journal of Medical Internet Research* **15,** e245 (2013).

Data Citations

1. Scarpace, L. *et al. The Cancer Imaging Archive* http://doi.org/10.7937/K9/TCIA.2016.RNYFUYE9 (2016).
2. Pedano, N. *et al. The Cancer Imaging Archive* http://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK (2016).
3. Bakas, S. *et al. The Cancer Imaging Archive* https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q (2017).
4. Bakas, S. *et al. The Cancer Imaging Archive* https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF (2017).

Acknowledgements

Author Contributions

S.B. conceived the study, provided initializations for the automated segmentation method, corrected segmentation labels, extracted radiomic features and wrote the manuscript. H.A. provided initializations for the automated segmentation method, corrected segmentation labels, extracted radiomic features and revised the manuscript. A.S. provided initializations for the automated segmentation method, corrected segmentation labels, and revised the manuscript. M.B. evaluated iteratively with S.B., H.A., and A.S. all the manually-revised segmentation labels until confirmation of their correctness, and revised the manuscript. M.R. pre-processed all images, created the metadata tables, quality checked the final provided data and revised the manuscript. J.K. assisted with the collection of metadata, the requirements for sharing the data in TCIA, and revised the manuscript. J.F. assisted with the collection of metadata, the requirements for sharing the data in TCIA, and revised the manuscript. K.F. assisted with the collection of metadata, sharing the data in TCIA, benchmarking of the segmentation algorithm, and revised the manuscript. C.D. supervised the study and revised the manuscript.

Additional Information

Tables 2, 3 and 4 are only available in the online version of this paper.

Competing interests: The authors declare no competing financial interests.

How to cite this article: Bakas, S. *et al.* Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117 doi: 10.1038/sdata.2017.117 (2017).