Journal of Neuroimaging in Psychiatry & Neurology

http://dx.doi.org/10.17756/jnpn.2016-008

Research Article

Open Access

Assessing the Effects of Software Platforms on Volumetric Segmentation of Glioblastoma

William D. Dunn Jr[1], Hugo J.W.L. Aerts[2,3], Lee A. Cooper[1,4,5], Chad A. Holder[6], Scott N. Hwang[7], Carle C. Jaffe[8], Daniel J. Brat[9], Rajan Jain[10], Adam E. Flanders[11], Pascal O. Zinn[12], Rivka R. Colen[13] and David A. Gutman[1,4*]

[1]Departments of Biomedical Informatics and Neurology, Emory University School of Medicine, Atlanta, GA, USA
[2]Departments of Radiation Oncology and Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[3]Department of Biostatistics & Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[4]Department Winship Cancer Institute, Emory University, Atlanta, GA, USA
[5]Department Biomedical Engineering, Georgia Institute of Technology/Emory University, Atlanta, GA, USA
[6]Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, GA, USA
[7]Department of Diagnostic Imaging Department, St. Jude Children's Research Hospital, Memphis, TN, USA
[8]Department of Radiology, Boston University School of Medicine, Boston, MA, USA
[9]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA
[10]Departments of Radiology and Neurosurgery, NYU School of Medicine, New York, NY, USA
[11]Department of Neuroradiology, Thomas Jefferson University Hospitals, Philadelphia, PA, USA
[12]Department of Neurosurgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[13]Department of Diagnostic Radiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

***Correspondence to:**
David A. Gutman, MD, PhD
Department of Biomedical Informatics Emory University School of Medicine, Atlanta, GA, USA
E-mail: DGutman@emory.edu

Received: May 17, 2016
Accepted: July 18, 2016
Published: July 20, 2016

Citation: Dunn WD Jr, Aerts HJWL, Cooper LA, Holder CA, Hwang SN, et al. 2016. Assessing the Effects of Software Platforms on Volumetric Segmentation of Glioblastoma. *J Neuroimaging Psychiatry Neurol* 1(2): 64-72.

Published by United Scientific Group

Abstract

Background: Radiological assessments of biologically relevant regions in glioblastoma have been associated with genotypic characteristics, implying a potential role in personalized medicine. Here, we assess the reproducibility and association with survival of two volumetric segmentation platforms and explore how methodology could impact subsequent interpretation and analysis.

Methods: Post-contrast T1- and T2-weighted FLAIR MR images of 67 TCGA patients were segmented into five distinct compartments (necrosis, contrast-enhancement, FLAIR, post contrast abnormal, and total abnormal tumor volumes) by two quantitative image segmentation platforms - 3D Slicer and a method based on Velocity AI and FSL. We investigated the internal consistency of each platform by correlation statistics, association with survival, and concordance with consensus neuroradiologist ratings using ordinal logistic regression.

Results: We found high correlations between the two platforms for FLAIR, post contrast abnormal, and total abnormal tumor volumes (spearman's r(67) = 0.952, 0.959, and 0.969 respectively). Only modest agreement was observed for necrosis and contrast-enhancement volumes (r(67) = 0.693 and 0.773 respectively), likely arising from differences in manual and automated segmentation methods of these regions by 3D Slicer and Velocity AI/FSL, respectively. Survival analysis based on AUC revealed significant predictive power of both platforms for the following volumes: contrast-enhancement, post contrast abnormal, and total abnormal tumor volumes. Finally, ordinal logistic regression demonstrated correspondence to manual ratings for several features.

Conclusion: Tumor volume measurements from both volumetric platforms produced highly concordant and reproducible estimates across platforms for general features. As automated or semi-automated volumetric measurements replace manual linear or area measurements, it will become increasingly important to keep in mind that measurement differences between segmentation platforms for more detailed features could influence downstream survival or radio genomic analyses.

Keywords

Imaging-genomics, GBM, MRI, Cancer, Computational science, 3D Slicer

Introduction

Glioblastoma is a highly malignant brain tumor with one of the worst survival rates of all cancers [1]. Standard-of-care treatment typically involves neurosurgical gross total resection, radiation therapy, and temozolomide chemotherapy [2]. However, despite therapeutic advances [3-6], relatively modest progress has been made in the last 20 years in terms of overall survival [7, 8].

The molecular heterogeneity among patients with GBM discovered in large genomic studies suggests that personalized treatment approaches targeting specific pathways may be beneficial [9-11]. With this increasing focus on genomic analysis, new opportunities exist for quantitative radiological imaging to aid in clinical management. However, GBMs differ not only in genomic makeup and microscopic properties, but also in their macroscopic phenotypic properties as observed on MR imaging, which likely reflect both molecular alterations and clinical course [12]. The development of non-invasive biomarkers based on neuroimaging and mining the information embedded in these images that are already routinely collected as part of standard-of-care may better capture the observed heterogeneity and offer valuable insight into cancer biology.

Radiological-genomic correlation studies provide insight into the relationship between tumor genotypes and imaging phenotypes. For example, glioblastoma imaging features have been related to a wide variety of genomic phenomena. *TP53* mutations have been correlated with multifocal gliomas [13] and *IDH1* mutations with non-enhancing tumor [14]. At the transcriptional level, VEGF expression and its prognostic value are affected by the level of edema of the tumor [15], while other imaging features such as tumor contrast enhancement and mass effect have been correlated with hypoxic and proliferation transcriptional patterns [16]. Finally, imaging features such as mixed-nodular enhancement have been correlated with tumor DNA methylation status [17].

The methodologies employed in these studies vary from simple 2D measurements of maximal tumor width on a single axial plane, qualitative assessments for the presence/absence of features of interest (e.g. necrosis, contrast enhancement, DTI signal inhomogeneity), or volumetric/pixel-based approaches. In a previous study, we assessed a set of standardized MR imaging features quantified by three board-certified neuroradiologist reviewing cases according to the VASARI (Visually Accessible Rembrandt Images) standard [18]. The VASARI feature-set consists of 30 features with specific guidelines on how to score each (i.e. 0-5%, 6-33%, 34-67%, 68-95% or > 95% contrast-enhancing), and is designed to allow for accurate and reproducible MR image scoring using a qualitative approach.

One of the challenges is that robust volumetric assessment and tumor segmentation are not a routine part of the radiologists' regular clinical workflow, nor are these capabilities routinely implemented across the various DICOM workstations. However, as technological advances continue, automated volumetric measurements may eventually supplement manual conventions and offer more accuracy [19]. One area that will likely benefit from more accurate 3D segmentations is *imaging genomics* or *radiomics*, a growing field focused on extracting useful information from radiology imaging data in order to comprehensively quantify the tumor phenotype using advanced imaging algorithms [20-23].

Many tools have been developed to define tumor boundaries and several fully-automated segmentation algorithms have been proposed, such as automated probabilistic segmentation [24, 25], unsupervised fuzzy c-means and nonfuzzy clustering [26], and morphological edge detection and region growing segmentation [27]. Brain Tumor Image Analysis (BraTumIA) is a popular tool for automated tumor segmentation that has been validated in several independent studies, suggesting its potential for patient segmentation or disease monitoring [28, 29]. Semi-automated tumor segmentation, where a computerized segmentation is carried out after an initial manual annotation is made, is often more accurate since it immediately eliminates surrounding tissue that has similar intensity to the tumor but is clearly not malignant [30]. Two important algorithms using a semi-automated segmentation approach for glioblastoma are interactive multi-scale watershed methods that divide brain from tumor based on information from a manual tracing of an initial slice [31] and balloon inflation method [32].

Within the academic and research community, a popular platform for such work is 3D Slicer, a free open-source application for medical image analysis and visualization that, due to its special functionalities, extensibility, and portability across platforms, is actively supported by the National Institutes of Health (NIH) and frequently used by its Quantitative Imaging Network (QIN) [33]. This platform has a number of segmentations algorithms, ranging in complexity and automation. One basic segmentation mechanism uses a "Grow Cut" algorithm based on assigning "seed" pixels that should be segmented together [33-36].

To determine the potential influence of image analysis methods on segmentation, we compared volumetric results obtained from two different platforms: 3D Slicer as well as an in-house method based on Velocity AI (now part of Varian) [37] and the FMRIB Software Library, an image analysis library (FSL) [38]. Using an initial cohort of 67 patients from the TCGA glioblastoma database, we compared these methods to each other through correlation statistics, survival analyses, and agreement to manual ratings obtained through a consensus group of neuroradiologists.

Materials and Methods

Patient selection

MR image sets from 67 patients from various institutions downloaded from The Cancer Imaging Archive (TCIA) were used in our analyses. These patients represent a subset of patients who had been recruited from The Cancer Genome Atlas (TCGA) project [39]. Image sets containing both post-

gadolinium (Gd) contrast-enhanced T1-weighted (T1w) and T2-FLAIR images were downloaded in DICOM-format and were individually reviewed before feature segmentation to confirm their pre-surgical and treatment-naive status, as well as to exclude images of exceptionally poor quality. As the patients had been previously de-identified by the TCGA and are available for public download, no Institutional Review Board approval was required.

Volumetric segmentation and measurement

Platform 1: 3D Slicer: We used 3D Slicer, a package developed at BWH/Harvard Medical School [40], to initially perform volumetric analysis. The volumes for the 67 patients used in the present analysis were calculated and reported from a subset of 78 volumes used in a previous report [41]. This previous report studied the correlation between edema/cellular invasion molecular patterns and MRI-phenotypes in GBM, whereas the present study investigates the reliability of the segmentation method itself. Segmentation was performed on the images by two board-certified neuroradiologists with over 25 years combined experience in image segmentation (R. Colen, F. Jolesz) to segment whole tumor, necrotic, and contrast-enhancement regions. 3D Slicer has more than 125 modules for image visualization, segmentation, registration, and 3D visualization [42]. T2-FLAIR images were used for calculating peritumoral FLAIR region and post-contrast T1w images for segmentation of the contrast-enhancement and necrosis volume. Spoiled Gradient Echo Recalled (SPGR) images were used for segmentation of the necrosis and contrast-enhancement volume if post-contrast T1w images were not present in the dataset. In patients where FLAIR images were not available, T2w and/or proton density images were used to segment the peritumoral FLAIR volume.

In short, the FLAIR and post-contrast T1w image sets were brought into spatial alignment. Afterwards, a fusion step took place to blend the data in both images into each other so that the three tumor image compartments, namely, peri-lesional edema/infiltration, tumor-enhancement, and necrosis could be delineated and quantified on the same image/slice level. Post-contrast T1w images were used as fixed images and T2-FLAIR as moving images for registration by mutual information optimization (Figure S1). The General Registration (BRAINS) and Transforms module were used for registration. BRAINS is an automatic registration module that provides linear and elastic transforms. The Transforms module requires more user interaction and was used to rigidly and manually align voxels of different volumes in space coordinates until optimal mapping was reached. After images were aligned, the 3D Slicer Editor module was used for segmentation. Abnormal peritumoral FLAIR volume (reflecting edema/invasion), tumor-enhancement, and necrosis volumes were manually color-coded and delineated beginning from the peripheral edema/invasion and going centrally to enhancing and necrotic regions in a single label map approach (Figure S2). The Draw effect module was used to manually and precisely delineate different structures to create colored label maps that were then used to build up 3D models to better visualize spatial relationships between tumor components. The volumes of the delineated tumor compartments were automatically calculated and generated by 3D Slicer Label Statistics module.

Platform 2: Velocity AI/FSL: The same 67 patients above were also analyzed using a semi-automated approach combining Velocity AI (Atlanta, GA) and FSL (Oxford, UK). Velocity AI, marketed primarily for radiation therapy treatment and image fusion, is specifically designed for the markup/masking and visualization of MR images. For post-contrast T1w images, masks were manually drawn over the tumor and the region it surrounds using a segmentation tool in Velocity AI (Figure 1A). Of note, only a single mask corresponding to the abnormal signal detected on a post-contrast T1w image was drawn; necrosis and contrast-enhancement regions were not delineated at this point.

Similarly, for T2-FLAIR sequences, masks representing the tumor's total abnormal signal detected by FLAIR were drawn in an analogous manner to the post-contrast T1w images (Figure 1B). This area included all of the areas indicative of edema, as well as other regions that showed signal abnormalities [43]. For the purposes of this study, we did not attempt to differentiate between bright FLAIR signals generated from edema versus non-contrast enhancing tumor. Because we were interested in the FLAIR signal as an estimation of overall tumor involvement, we limited the FLAIR segmentations to regions clearly contiguous with the primary tumor, and did not include for example, thin layers of extension away from the dominant mass along the epididymal surface.

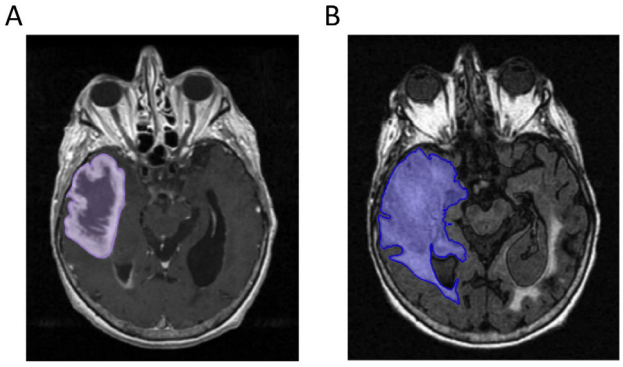

Figure 1: Binary mask segmentation for a 60 year old male patient with a right temporal GBM. Masks were manually drawn in Velocity AI using the 2D flood fill tool (purple, blue areas) to cover tumor volumes on (A) post-contrast T1w images and (B) T2-FLAIR images. Note that the area occupied by the ventricles has been omitted from the T2-FLAIR mask.

Following initial tumor markup, the masks were exported from Velocity AI as DICOM-RT objects. These image markups were subsequently converted to NIFTI and PNG formats to allow further visualization and segmentation. As a secondary quality check, a custom visualization platform was developed as part of a larger project in our lab (TumorView, Figure 2) that enabled rapid screening of the image volumes that were analyzed for this study. This tool helped eliminate images of particularly poor quality or those from post-surgical image sets (and hence not amenable to estimation of presurgical tumor volume), as well as to ensure that the

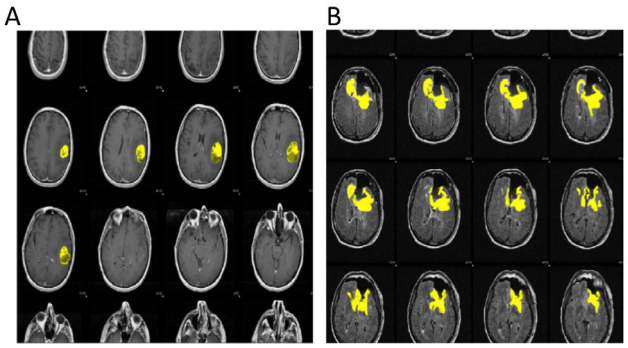

Figure 2: Tumor View Dynamic Reader. Image sets of TCGA patients with accompanying binary masks were uploaded onto an ad hoc website and analyzed. Patients with acceptable images and mask overlay (A) were cleared for analysis whereas patients with poor images or post-surgical images (B) were excluded.

masks and accompanying images were overlaid and exported properly.

Contrast-enhancement and necrosis volumes were calculated from the masked region on post-contrast T1w images using FAST (FMRIB Automated Segmentation Tool [44], a tool included in FSL). Briefly, for the masks drawn on the post-contrast T1w images, encompassing both contrast-enhancement and necrotic regions, a k-means clustering algorithm was used to identify two clusters based on relative pixel intensity. This resulted in a binary classification of tumor into either bright (e.g. contrast-enhancing) or dark (necrotic) regions (Figure 3). This voxel-based measure could then be directly converted to a volumetric measure by multiplying the number of dark voxels (class 1) or bright voxels (class 2) by the voxel size in mm^3.

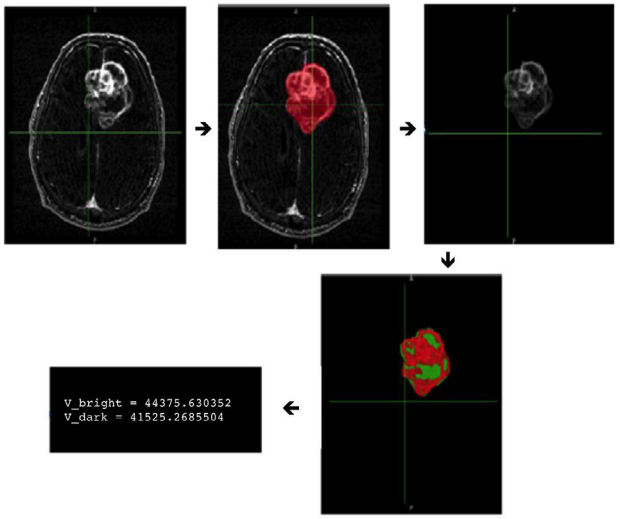

Figure 3: Semi-automated segmentation and measurement method for T1w post-contrast images. Tumor volume was identified by a manually-drawn binary mask (red, second frame) in Velocity AI. K-means clustering divides pixels covered by the mask into bright (red, contrast enhancement) and dark (green, necrosis) clusters based on relative pixel intensity. Volume of individual features can be estimated by converting voxels into mm^3.

For subsequent analysis used throughout this manuscript, we define the post contrast abnormal volume (PCAV) as the sum of the necrosis (NE) and contrast-enhancement (CE) volumes. For the FLAIR images, the entire abnormal signal is referred to as the total abnormal tumor volume (TATV). Finally, the FLAIR envelope refers to the difference between the PCAV and the TATV (Figure 4).

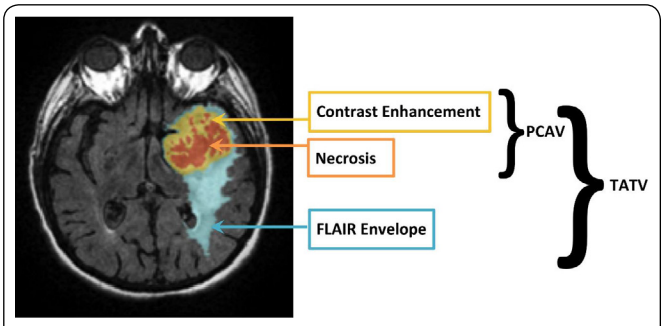

Figure 4: Naming conventions used throughout this article for various tumor compartments. TATV (Total abnormal tumor volume) corresponds to total abnormal signal from the T2-FLAIR scan. PCAV (post contrast abnormal volume) represents the sum of necrosis and contrast-enhancement volumes. FLAIR Envelope represents the difference of PCAV from TATV volume. Base MR image reprinted from [41] under open access license.

We will refer to 3D Slicer as Platform 1 and Velocity AI/FSL as Platform 2 for the purposes of this article.

Consistency of measurements between volumetric images methods: In order to measure the correlation between measurements for various tumor compartments across both platforms, Spearman correlation coefficients were calculated for each of the five volumes. Spearman method was used due to its robustness in approaching outliers as well as to the fact that according to Shapiro-Wilk normality tests, nearly all volume measurements from both platforms did not follow normal distributions. In addition, to view global trends between Platforms 1 and 2, Bland-Altman plots were generated [45]. Bland-Altman plots analyze agreement by comparing the means and differences of each value measured by two different instruments. They are used to compare clinical measurement techniques and are designed to provide more information than simple correlation coefficients.

Survival analysis

In order to explore the accuracy of our segmenting techniques, we performed several analyses to investigate whether the volumes obtained through both platforms could predict survival. We hypothesized that if our survival correlations were significant using imaging features previously implicated as survival imaging markers in glioblastoma, it would be more probable that these imaging features were accurately segmented. We first assessed whether the mean volumes for various tumor compartments were significantly different from each other when patients are grouped into short-term and long-term survivors. Given the lack of a true gold standard for the volume of any particular tumor sub-region, we assessed the ability of each methodology to predict a clinically meaningful endpoint, survival at one year. Patients in the initial 67 patient cohort with survival data were split into groups surviving more than one year (N = 37) and those surviving less than one year (N = 23) and the means of these groups were compared using a Wilcoxon rank-sum test for each imaging feature and several derivative ratios.

After measuring the association between imaging features and survival, we next investigated whether various imaging features were predictive of survival. To quantify prognostic performance of the features for survival prediction, the area under the curve (AUC) of the receiver-operating characteristic (ROC) was assessed using the same 60 patients as the analysis above. For the survival analysis, we predicted one-year survival after date of initial diagnosis.

Concordance of volumetric segmentations from both platforms with consensus ratings from neuroradiologists: Finally, we also explored the accuracy of our segmentation techniques by measuring the agreement between the volumes obtained through each platform and those estimated by a consensus group of neuroradiologists rated through the VASARI project. VASARI is a vetted, tested, and validated controlled terminology that aims to comprehensively and reproducibly describe MR imaging [46]. Of the original 67 patient cohort, 59 were also analyzed by neuroradiologists using the VASARI feature-set. Tumors from these 59 patients were analyzed by at least three different neuroradiologists and final scores were based on a consensus between the raters.

In order to compare these resulting categorical scores to our volumetric results, we first identified equivalent variables between the VASARI feature set and our volumetric features: proportion necrosis, proportion contrast-enhancement, and proportion edema. We converted our volumetric features to these VASARI features by dividing necrosis volume, contrast enhancement volume, and FLAIR volume, respectively, by TATV. Next, for each volumetric platform, we measured the association between these quantitative measurements with the categorical measurements based on VASARI (0-5%, 6-33%, 34-67%, 68-95% or >95%) using ordinal logistic regression. Each categorical range was converted to 1, 2, 3, 4, and 5 respectively.

The proportional-odds assumption of our models was verified by visually inspecting that the empirical odds ratios between larger versus smaller VASARI scores and predictor variables was relatively constant over the levels, in a number of cases checked, and thus considered valid to the extent possible given our data. For significance, we used the Wald Chi-Square test to test the null hypotheses that all the fitted coefficients in the model are zero, i.e. the quantitative volumetric measures do not predict the trend.

Results

Consistency of measurements between volumetric images methods

First, we investigated the correlation between both volumetric techniques for estimation of total abnormal tumor volume (TATV), post contrast abnormal volume (PCAV), FLAIR envelope volume, necrosis volume, and contrast-enhancement volume.

We found strong Spearman's rank correlation coefficients between both platforms for measuring TATV and PCAV (r(67) = 0.97 and 0.96 respectively) (Table 1). Strong correlations were also observed for FLAIR envelope volume (r(67) = 0.95). Fair to good correlations of 0.693 and 0.773 were observed for both platforms for estimating necrosis volume and contrast-enhancement volume respectively.

Table 1: Correlation results from volumes measured between platform 1 and 2 (N = 67).

Feature	Spearman rho between platform 1 and 2	p-value rho ≠ 0
Necrosis (mm^3)	0.693	**8.398E-11***
Contrast Enhancement (mm^3)	0.773	**1.807E-14***
FLAIR Envelope (mm^3)	0.952	**2.2E-16***
PCAV (mm^3)	0.959	**2.2E-16***
TATV (mm^3)	0.969	**2.2E-16***

Significant P values (< 0.05) indicated by *

Bland-Altman plots were also generated to evaluate these patterns more in depth (Figure 5). In general, the differences in measurements obtained by the two platforms increased as their resulting volumes increased. Near zero mean differences between both platforms for TATV, PACV and FLAIR envelope volume suggests that no systematic platform bias exists. However, the delineations between contrast-enhancing volume and tumor necrotic volume were relatively less congruent. For example, Platform 1 consistently identified more contrast enhancement and less necrosis volume than did Platform 2.

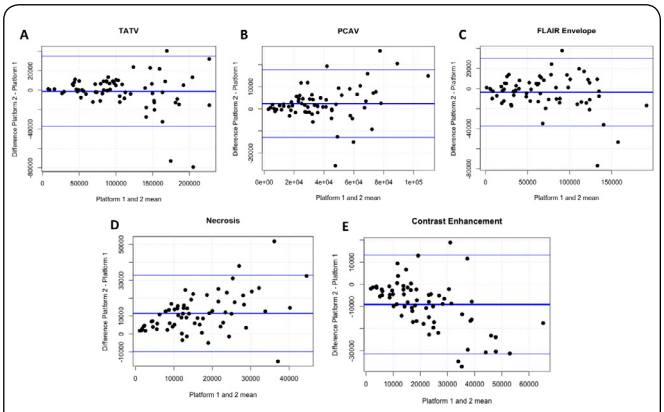

Figure 5: Bland-Altman plots showing the measurement trends for each imaging feature. In each graph, the middle horizontal bar represents the average difference of the measurements obtained from Platform 1 subtracted from those of Platform 2 for all patients measured (N = 67). A mean of 0 suggests no bias between the two platforms. The upper and lower horizontal bars represent 1 standard deviation from the average difference. According to literature consensus, points that fall outside of these lines significantly differ between measurements. Total abnormal volume (A), post contrast abnormal volume (B), FLAIR Envelope (C), necrosis volume (D), and contrast-enhancing volume (E).

Survival analysis

As no gold standard exists for tumor volume, we chose a secondary endpoint to assess the relative robustness of these two volumetric methods. For simplicity, we first generated a cohort of patients with survival greater than one year (N = 37) or less than one year (N = 23). Using volumes obtained

from Platform 2 (Velocity AI/FSL), patients surviving less than a year were found to have on average significantly greater volumes of necrosis (P = 0.0027), contrast-enhancement (P = 0.0094), post-contrast abnormal volume (P = 0.0041), and total abnormal tumor volume (P = 0.0267) than those surviving more than a year (Table 2). Measurements obtained by Platform 1 showed similar trends and significances, except with regards to necrosis volume (P = 0.0975).

Table 2: Comparison of means of volumes and ratios between patients surviving less than one year (N = 23) and those surviving more than one year (N = 37) for various imaging features. Same patients were used across both volumetric platforms. Significance values reflect Wilcoxon rank-sum test.

	Platform 1: 3D Slicer			Platform 2: Velocity A1		
Imaging Feature	**$\bar{x}$ < 1 yr**	**$\bar{x}$ > 1 yr**	**Sig.**	**$\bar{x}$ < 1 yr**	**$\bar{x}$ > 1 yr**	**Sig.**
Necrosis (mm^3)	1.33E+04	8.50E+03	9.75E-02	2.78E+04	1.76E+04	**2.70E-03***
Contrast Enhancement (mm^3)	3.34E+04	2.04E+04	**2.06E-03***	2.18E+04	1.39E+04	**9.42E-03***
FLAIR Envelope (mm^3)	7.74E+04	6.10E+04	3.57E-01	6.91E+04	5.95E+04	3.49E-01
PCAV (mm^3)	4.67E+04	2.89E+04	**2.06E-03***	4.96E+04	3.15E+04	**4.11E-03***
TATV (mm^3)	1.24E+05	8.99E+04	**2.36E-02***	1.19E+05	9.10E+04	**2.67E-02***
Necrosis / TATV	1.31E-01	1.05E+01	8.43E-01	2.46E-01	2.14E+01	5.16E-01
Contrast Enhancement / TATV	2.81E-01	2.47E+01	4.23E-01	1.96E-01	1.68E+01	6.51E-01
FLAIR Envelope / TATV	5.88E-01	6.48E+01	5.66E-01	5.58E-01	6.18E+01	6.08E-01
PCAV / TATV	4.12E-01	3.52E-01	5.66E-01	4.42E-01	3.82E-01	6.08E-01
Necrosis / Contrast Enhancement	4.68E-01	4.88E-01	8.55E-01	1.35E+00	1.39E+00	7.40E-01
Contrast Enhancement / PCAV	7.38E-01	7.27E-01	8.55E-01	4.34E-01	4.29E-01	7.40E-01

Significant P values (P < 0.05) are indicated by *

The prognostic performance of the volumes was assessed using the area under the curve (AUC) of the receiver operating characteristic (ROC). For both platforms, contrast-enhancement, PCAV, and TATV, were measured to be significantly prognostic (Table 3). Furthermore, similar to above, Platform 2 additionally quantified necrosis as a strong prognostic parameter, with an AUC > 0.7 (P = 0.0008). For both platforms, no significance was found for the FLAIR envelope volume, Necrosis/TATV, Contrast-Enhancement / TATV, and FLAIR Envelope/TATV, indicating the limited prognostic value of these features.

Concordance of volumetric segmentations from both platforms with consensus ratings from neuroradiologist: Finally, ordinal logistic regression results are displayed in Table 4. P values for Wald Chi-Square tests show that Platform 1 showed consistent agreement with all three scores derived using the VASARI criteria, while Platform 2 showed agreement with only proportion of edema.

Discussion

In this work, we compared two volumetric segmentation approaches based on Velocity AI/FSL and 3D Slicer and demonstrated that both platforms produced very consistent estimates of TATV and PCAV (r(67) = 0.97 and 0.96 respectively). While estimates of sub-compartment volumes (necrosis and contrast enhancement) were also significantly correlated, these correlations were much less robust (r(67) = 0.69 and 0.77 respectively). In the case of 3D Slicer, these compartments were estimated by manual contouring, and the sum of those two sub compartments would then equate to the PCAV. For estimates of NE and CE based on the Velocity AI/FSL approach, we instead relied on automated k-means clustering to semi-automatically segment the PCAV into the two compartments (bright/dark → CE/ NE).

Technical explanation for segmentation discrepancies

Given the relationship between NE and CE, we generated Bland-Altman plots to look for any systematic differences between these results to assess for a bias. Bland-Altman plots indicated Platform 1 consistently labeled more voxels as contrast-enhancement (and fewer as necrosis) than Platform 2. These results suggest that not only is there a loss of reproducibility between platforms for smaller sub-regions compared to larger regions, but also that there is a systematic difference rather than a difference attributable to random chance. It is important to note these systematic differences as they could translate to systematic differences in downstream survival or biological analyses.

The differences in segmentation results for these smaller sub-regions were likely a result of the inherent k-means clustering algorithm used in Platform 2 that clusters pixels based on intensities using a cutoff derived from that specific intensity volume histogram. The threshold between what determines whether a voxel is labeled as contrast-enhancement or necrosis is also fairly subjective in the case of manual delineation in Platform 1, whereas a k-means method is based purely on the raw values fed into the cluster algorithm, and therefore not subject to the brightness/contrast and windowing settings of the workstation that could impact human segmentation.

It is important to take into consideration that the data set we used was collected from a consortium of institutions within the TCGA network using different MR scanners, protocols,

and pulse sequences which all add significant variability in signal-to-noise ratios (SNR) and what appears to be contrast-enhancement/necrosis regions [9]. This would prevent any reasonable attempts to try and normalize the window level used for threshold in Platform 2 that could potentially obviate that effect.

Table 3: Area Under the Curve (AUC) statistic calculated for each volumetric compartment and its derivative ratios measured for both platforms to predict one year survival.

	Platform 1: 3D Slicer		Platform 2: Velocity A1	
Imaging Feature	**AUC**	**Sig.**	**AUC**	**Sig.**
Necrosis (mm^3)	0.6287	6.67E-002	0.7286	**8.01E-004***
Contrast Enhancement (mm^3)	0.7344	**2.51E-004***	0.6992	**5.30E-003***
FLAIR Envelope (mm^3)	0.5723	3.49E-001	0.5734	3.25E-001
PCAV (mm^3)	0.7344	**6.05E-004***	0.7192	**1.91E-003***
TATV (mm^3)	0.6745	**9.06E-003***	0.6710	**1.12E-002***
Necrosis/TATV	0.5159	8.31E-001	0.5511	4.95E-001
Contrast Enhancement /TATV	0.5629	3.80E-001	0.5358	6.39E-001
FLAIR Envelope/ TATV	0.5452	5.55E-001	0.5405	5.96E-001

Significant P values (P < 0.05) are indicated by *

Table 4: Ordinal logistic regression measuring agreement between each Platform 1 and Platform 2 and gold standard rating using features derived from the VASARI scale. 'Prop' refers to the proportion of contrast-enhancement (CE), necrosis (NE), or Edema over TATV. Coef = coefficient of regression, SE = standard error.

		Coef	SE	P
Platform 1	Prop CE	1.08E-01	2.97E-02	**2.91E-04***
	Prop NE	1.66E-01	4.10E-02	**5.20E-05***
	Prop Edema	7.79E-02	1.82E-02	**1.87E-05***
Platform 2	Prop CE	-8.09E-03	2.74E-02	7.68E-01
	Prop NE	4.61E-02	2.33E-02	4.76E-02
	Prop Edema	5.19E-02	1.41E-02	**2.43E-04***

Significant P values corrected by bonferroni (P < 0.015) are indicated by *

Survival analysis

Using one-year survival as an endpoint, both platforms demonstrated that patients with more contrast-enhancement, PCAV, and TATV showed significantly shorter survival times. A more sophisticated model using AUC analyses also demonstrated that these same measures were associated with survival. Interestingly, for both types of analyses investigated, Platform 2 additionally measured necrosis volume as being a significant predictor of survival.

The results from our survival analysis suggest that both platforms accurately identify gross overall features such as TATV and PCAV as previous research has demonstrated association between metrics that measure overall tumor and survival [47, 48]. In addition, volume of contrast-enhancement, representing neovascularity and angiogenesis, has also been implicated as being significantly associated with overall survival through qualitative [49, 50] and quantitative [48] studies.

While necrosis has also typically been implicated in survival [50, 51], other studies have not found significant associations. For example, in a study, based in part on the patients used in the present analysis, necrosis was not found to be associated with survival. Colen et al. [52] suggested that this discrepancy may be related to patient sex: necrosis was only significantly associated with survival for females, likely due to differences in MYC and TP53 in cell death between males and females [52].

We should note previous studies have suggested that feature to survival correlation likely does not follow a linear model [50]. In addition, although the quantified features showed strong performance (AUC~0.70), these results have to be validated in larger cohorts to determine true prognostic performance. However, the main goal of these analyses, rather than to thoroughly assess the relationship between specific imaging features and survival, was to investigate whether both platforms produced similar results.

Segmentation accuracy

Our work evaluates the reproducibility of the two segmentation platforms in regards to their ability to consistently segment the same regions. Our results suggest that the segmentation technique can lead to different measurements in certain cases that may ultimately influence survival prediction and radio-genomic correlations.

A related but different concept to reproducibility is accuracy, or how close the measured value is to the "true" number within the specific region of interest (ROI). Indeed, the task of choosing a gold standard to objectively measure accuracy was a challenge throughout this project and a repeating issue in the field in general. To explore this question, we used various markers at our disposal to measure accuracy. For instance, a number of imaging features have been shown to be significantly associated with overall survival, which is both an unambiguous endpoint and completely independent of the imaging metrics. Thus, we could consider our measurements "accurate" if we could reproduce these associations with the measurements using our two platforms. However, this approach is only valid if literature already consistently demonstrates the features have been associated with survival; in the case of many potential radiological features this is not the case. We also explored accuracy through measuring agreement with radiologists' measurements of related features using the qualitative VASARI method. This is also not a perfect marker of accuracy since the VASARI results were influenced by inter-rater variability and while experts performed the ratings, visual estimations of tumor volume from a set of 2D images is an inherently difficult task.

The implications of accurate segmentations vary depending on the specific use case and in many cases necessitate that the investigators weigh the relative advantages/disadvantages of

the various approaches (e.g. manual segmentation takes longer, but may result in segmentations with improved accuracy). The ability to obtain accurate volumetric measurements is important in a variety of clinical or research applications. For example, in a direct clinical setting, accurate volumetric measurements are important for surgical and radiotherapy guidance and planning, and to a lesser extent, for surveillance during clinical trials. In regards to research with downstream clinical implications, accurate segmentations are important when the volumes and masks serve as the basis for subsequent radiogenomic analysis. It is likely that more accurate segmentations will more precisely target feature of interest that have common molecular or genetic bases from confounding features and lead to more reproducible associations. Imaging genomics is a growing field especially in the cancer domain [41, 53, 54] and has the potential to expand into different areas such as neurology and public health [55]. Our results suggest that, in order for volumetric MRI analysis programs such as 3D Slicer to continue to play a strong role in the future of imaging genomics, as well as more traditional roles of prognosis, staging, and response assessment, special focus is required to ensure accurate segmentation measurements. Therefore, future work will more strongly establish survival-associated imaging markers through consistent results based on large-sample studies as well as incorporate other potential markers for accuracy such as treatment response or relapse of the tumor during surgery.

Conclusion

As quantitative volumetric image analysis gains an ever-increasing foothold in clinical and research domains, we are likely to benefit by gaining a stronger insight into the biology of glioblastoma, as well as other oncological or neurological diseases. Many volumetric imaging platforms already exist - open source, licensed, automated, semi-automated, etc. – and are based on a wide range of underlying techniques. Our results suggest that certain features are robust across platforms, particularly those related to the total area of abnormal signal. Likewise, since measurements for more detailed sub volumes varied more between platforms, results from downstream radiogenomic analyses should be interpreted more carefully until these volumetric techniques are thoroughly validated.

Acknowledgements

This work was supported by funding from the National Cancer Institute (NCI U24 CA194362 and NCI U24 CA194354).

References

1. Ahmed R, Oborski MJ, Hwang M, Lieberman FS, Mountz JM. 2014. Malignant gliomas: current perspectives in diagnosis, treatment, and early response assessment using advanced quantitative imaging methods. *Cancer Manag Res* 6: 149-170. doi: 10.2147/CMAR.S54726
2. White ML, Zhang Y, Kirby P, Ryken TC. 2005. Can tumor contrast enhancement be used as a criterion for differentiating tumor grades of oligodendrogliomas? *AJNR Am J Neuroradiol* 26(4): 784-790.
3. Zhao S, Wu J, Wang C, Liu H, Dong X, et al. 2013. Intraoperative fluorescence-guided resection of high-grade malignant gliomas using 5-aminolevulinic acid-induced porphyrins: a systematic review and meta-analysis of prospective studies. *PloS One* 8(5): e63682. doi: 10.1371/journal.pone.0063682
4. Westphal M, Hilt DC, Bortey E, Delavault P, Olivares R, et al. 2003. A phase 3 trial of local chemotherapy with biodegradable carmustine (BCNU) wafers (Gliadel wafers) in patients with primary malignant glioma. *Neuro Oncol* 5(2): 79-88. doi: 10.1093/neuonc/5.2.79
5. Stieber VW, Mehta MP. 2007. Advances in radiation therapy for brain tumors. *Neurol Clin* 25(4): 1005-1033. doi: 10.1016/j.ncl.2007.07.005
6. Norden AD, Young GS, Setayesh K, Muzikansky A, Klufas R, et al. 2008. Bevacizumab for recurrent malignant gliomas: efficacy, toxicity, and patterns of recurrence. *Neurology* 70(10): 779-787. doi: 10.1212/01.wnl.0000304121.57857.38
7. Yamada S, Takai Y, Nemoto K, Ogawa Y, Kakuto Y, et al. 1992. Radioresponse and prognosis of malignant glioma. *Tohoku J Exp Med* 167(1): 79-87. doi: 10.1620/tjem.167.79
8. Mao H, Lebrun DG, Yang J, Zhu VF, Min Li. 2012. Deregulated signaling pathways in glioblastoma multiforme: molecular mechanisms and therapeutic targets. *Cancer Invest* 30(1): 48-56. doi: 10.3109/07357907.2011.630050
9. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, et al. 2013. The cancer genome atlas pan-cancer analysis project. *Nat Genet* 45(10): 1113-1120. doi: 10.1038/ng.2764
10. Arimappamagan A, Somasundaram K, Thennarasu K, Peddagangannagari S, Srinivasan H, et al. 2013. A fourteen gene GBM prognostic signature identifies association of immune response pathway and mesenchymal subtype with high risk group. *PloS One* 8(4): e62042. doi: 10.1371/journal.pone.0062042
11. Brennan CW, Verhaak GW, McKenna A, Campos B, Noushmehr H, et al. 2013. The somatic genomic landscape of glioblastoma. *Cell* 155(2): 462-477. doi: 10.1016/j.cell.2013.09.034
12. Wang MY, Cheng JL, Han YH, Li YH, Dai JP, et al. 2012. Measurement of tumor size in adult glioblastoma: classical cross-sectional criteria on 2D MRI or volumetric criteria on high resolution 3D MRI? *Eur J Radiol* 81(9): 2370-2374. doi: 10.1016/j.ejrad.2011.05.017
13. Kyritsis AP, Bondy ML, Xiao M, Berman EL, Cunningham JE, et al. 1994. Germline p53 gene mutations in subsets of glioma patients. *J Natl Cancer Inst* 86(5): 344-349. doi: 10.1093/jnci/86.5.344
14. Carrillo JA, Lai A, Nghiemphu PL, Kim HJ, Phillips HS, et al. 2012. Relationship between tumor enhancement, edema, IDH1 mutational status, MGMT promoter methylation, and survival in glioblastoma. *AJNR Am J Neuroradiol* 33(7): 1349-1355. doi: 10.3174/ajnr.A2950
15. Carlson MR, Pope BW, Horvath S, Braunstein JG, Nghiemphu P, et al. 2007. Relationship between survival and edema in malignant gliomas: role of vascular endothelial growth factor and neuronal pentraxin 2. *Clin Cancer Res* 13(9): 2592-2598. doi: 1078-0432.CCR-06-2772
16. Diehn M, Nardini C, Wang DS, McGovern S, Jayaraman M, et al. 2008. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc Natl Acad Sci U S A* 105(13): 5213-5218. doi: 10.1073/pnas.0801279105
17. Eoli M, Menghi F, Bruzzone MG, De Simone T, Valletta L, et al. 2007. Methylation of O6-methylguanine DNA methyltransferase and loss of heterozygosity on 19q and/or 17p are overlapping features of secondary glioblastomas with prolonged survival. *Clin Cancer Res* 13(9): 2606-2613. doi: 1078-0432.CCR-06-2184
18. Gutman DA, Cooper LA, Hwang SN, Holder CA, Gao J, et al. 2013. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology* 267(2): 560-569. doi: 10.1148/radiol.13120118
19. Chow DS, Qi J, Guo X, Miloushev VZ, Iwamoto FM, et al. 2014. Semiautomated volumetric measurement on postcontrast MR imaging for analysis of recurrent and residual disease in glioblastoma multiforme. *AJNR Am J Neuroradiol* 35(3): 498-503. doi: 10.3174/ajnr.A3724
20. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout

RG, et al. 2012. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 48(4): 441-446. doi: 10.1016/j.ejca.2011.11.036

21. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, et al. 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 5: 4006. doi: 10.1038/ncomms5006
22. Gutman DA, Dunn WD Jr, Grossmann P, Cooper LA, Holder CA, et al. 2015. Somatic mutations associated with MRI-derived volumetric features in glioblastoma. *Neuroradiology* 57(12): 1227-1237. doi: 10.1007/s00234-015-1576-7
23. ElBanan MG, Amer AM, Zinn PO, Colen RR. 2015. Imaging genomics of glioblastoma: state of the art bridge between genomics and neuroradiology. *Neuroimaging Clin N Am* 25(1): 141-153. doi: 10.1016/j.nic.2014.09.010
24. Egger J, Kapur T, Fedorov A, Pieper S, Miller JV, et al. 2013. GBM volumetry using the 3D Slicer medical image computing platform. *Sci Rep* 3: 1364. doi: 10.1038/srep01364
25. Zou KH, Wells WM III, Kaus MR, Kikinis R, Jolesz FA, et al. 2002. Statistical validation of automated probabilistic segmentation against composite latent expert ground truth in MR imaging of brain tumors. In: Dohi T, Kikinis R (eds) Medical Image Computing and Computer-Assisted Intervention-MICCAI 2002 Lecture Notes in Computer Science. Springer Berlin Heidelberg, Germany, pp 315-322.
26. Velthuizen RP, Clarke LP, Phuphanich S, Hall LO, Bensaid AM, et al. 1995. Unsupervised measurement of brain tumor volume on MR images. *J Magn Reson Imaging* 5(5): 594-605. doi: 10.1002/jmri.1880050520
27. Gibbs P, Buckley DL, Blackband SJ, Horsman A. 1996. Tumour volume determination from MR images by morphological segmentation. *Phys Med Biol* 41(11): 2437-2446.
28. Meier R, Knecht U, Loosli T, Bauer S, Slotboom J, et al. 2016. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Sci Rep* 6: 23376. doi: 10.1038/srep23376.
29. Rios Velazquez E, Meier R, Dunn WD Jr, Alexander B, Wiest R, et al. 2015. Fully automatic GBM segmentation in the TCGA-GBM dataset: prognosis and correlation with VASARI features. *Sci Rep* 5: 16822. doi: 10.1038/srep16822
30. Prastawa M, Bullitt E, Ho S, Gerig G. 2004. A brain tumor segmentation framework based on outlier detection. *Med Image Anal* 8(3): 275-283. doi: 10.1016/j.media.2004.06.007
31. Letteboer MM, Olsen OF, Dam EB, Willems PW, Viergever MA, et al. 2004. Segmentation of tumors in magnetic resonance brain images using an interactive multiscale watershed algorithm. *Acad Radiol* 11(10): 1125-1138. doi: 10.1016/j.acra.2004.05.020
32. Zukić D, Egger J, Bauer MA, Kuhnt D, Carl B, et al. 2011. Glioblastoma multiforme segmentation in MRI data with a balloon inflation approach. *arXiv*:1102.0634.
33. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, et al. 2012. 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* 30(9): 1323-1341. doi: 10.1016/j.mri.2012.05.001
34. Vezhnevets V, Konouchine V. 2005. "GrowCut" - interactive multi-label N-D image segmentation by cellular automata. presented at the Graph-icon, Novosibirsk Akademgorodok, Russia.
35. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, et al. 2014. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PloS One* 9(7): e102107. doi: 10.1371/journal.pone.0102107
36. Rios Velazquez E, Parmar C, Jermoumi M, Mak RH, van Baardwijk A, et al. 2013. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci Rep* 3: 3529. doi: 10.1038/srep03529
37. VelocityTM Varian Medical Systems. 2016.
38. FSL - FslWiki. 2016.
39. Clark K, Vendt B, Smith K, Freymann J, Kirby J, et al. 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 26(6): 1045-1057. doi: 10.1007/s10278-013-9622-7
40. Gering DT, Nabavi A, Kikinis R, Hata N, O'Donnell LJ, et al. 2001. An integrated visualization system for surgical planning and guidance using image fusion and an open MR. *J Magn Reson Imaging* 13(6): 967-975. doi: 10.1002/jmri.1139
41. Zinn PO, Mahajan B, Sathyan P, Singh SK, Majumder S, et al. 2011. Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. *PloS One* 6(10): e25451. doi: 10.1371/journal.pone.0025451
42. Documentation/4.3 - SlicerWiki. 2016.
43. Pavlisa G, Rados M, Pavlisa G, Pavic L, Potocki K, et al. 2009. The differences of water diffusion between brain tissue infiltrated by tumor and peritumoral vasogenic edema. *Clin Imaging* 33(2): 96-101. doi: 10.1016/j.clinimag.2008.06.035
44. Zhang Y, Brady M, Smith S. 2001. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 20(1): 45-57. doi: 10.1109/42.906424
45. Bland JM, Altman DG. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1(8476): 307-310. doi: 10.1016/S0140-6736(86)90837-8
46. VASARI Research Project - The Cancer Imaging Archive (TCIA) public access - cancer imaging archive Wiki. 2016.
47. Zhang Z, Jiang H, Chen X, Bai J, Cui Y, et al. 2014. Identifying the survival subtypes of glioblastoma by quantitative volumetric analysis of MRI. *J Neurooncol* 119(1): 207-214. doi: 10.1007/s11060-014-1478-2
48. Mazurowski MA, Zhang J, Peters KB, Hobbs H. 2014. Computer-extracted MR imaging features are associated with survival in glioblastoma patients. *J Neurooncol* 120(3): 483-488. doi: 10.1007/s11060-014-1580-5
49. Li WB, Tang K, Chen Q, Li S, Qiu XG, et al. 2012. MRI manifestations correlate with survival of glioblastoma multiforme patients. *Cancer Biol Med* 9(2): 120-123. doi: 10.3969/j.issn.2095-3941.2012.02.007
50. Lacroix M, Abi-Said D, Fourney DR, Gokaslan ZL, Shi W, et al. 2001. A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J Neurosurg* 95(2): 190-198.
51. Ekici MA, Bulut T, Tucer B, Kurtsoy A. 2011. Analysis of the mortality probability of preoperative MRI features in malignant astrocytomas. *Turk Neurosurg* 21(3): 271-279. doi: 10.5137/1019-5149.JTN.3321-10.3
52. Colen RR, Wang J, Singh SK, Gutman DA, Zinn PO. 2015. Glioblastoma: imaging genomic mapping reveals sex-specific oncogenic associations of cell death. *Radiology* 275(1): 215-227. doi: 10.1148/radiol.14141800.
53. Rutman AM, Kuo MD. 2009. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *Eur J Radiol* 70(2): 232-241. doi: 10.1016/j.ejrad.2009.01.050
54. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, et al. 2012. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data--methods and preliminary results. *Radiology* 264(2): 387-396. doi: 10.1148/radiol.12111607
55. Thompson PM, Martin NG, Wright MJ. 2010. Imaging genomics. *Curr Opin Neurol* 23(4): 368-373. doi: 10.1097/WCO.0b013e32833b764c