A novel magnetic resonance imaging segmentation technique for determining diffuse intrinsic pontine glioma tumor volume

Ranjodh Singh, BPhil, BSc; Zhiping Zhou, MD, PhD; Jamie Tisnado, MD; Sofia Haque, MD; Kyung K. Peck, PhD; Robert J. Young, MD; Apostolos John Tsiouris, MD; Sunitha B. Thakur, PhD; and Mark M. Souweidane, MD

OBJECTIVE Accurately determining diffuse intrinsic pontine glioma (DIPG) tumor volume is clinically important. The aims of the current study were to 1) measure DIPG volumes using methods that require different degrees of subjective judgment; and 2) evaluate interobserver agreement of measurements made using these methods.

METHODS Eight patients from a Phase I clinical trial testing convection-enhanced delivery (CED) of a therapeutic antibody were included in the study. Pre-CED, post–radiation therapy axial T2-weighted images were analyzed using 2 methods requiring high degrees of subjective judgment (picture archiving and communication system [PACS] polygon and Volume Viewer auto-contour methods) and 1 method requiring a low degree of subjective judgment (k-means clustering segmentation) to determine tumor volumes. Lin’s concordance correlation coefficients (CCCs) were calculated to assess interobserver agreement.

RESULTS The CCCs of measurements made by 2 observers with the PACS polygon and the Volume Viewer auto-contour methods were 0.9465 (lower 1-sided 95% confidence limit 0.8472) and 0.7514 (lower 1-sided 95% confidence limit 0.3143), respectively. Both were considered poor agreement. The CCC of measurements made using k-means clustering segmentation was 0.9938 (lower 1-sided 95% confidence limit 0.9772), which was considered substantial strength of agreement.

CONCLUSIONS The poor interobserver agreement of PACS polygon and Volume Viewer auto-contour methods highlighted the difficulty in consistently measuring DIPG tumor volumes using methods requiring high degrees of subjective judgment. k-means clustering segmentation, which requires a low degree of subjective judgment, showed better interobserver agreement and produced tumor volumes with delineated borders.

http://thejns.org/doi/abs/10.3171/2016.4.PEDS16132

KEY WORDS diffuse intrinsic pontine glioma; k-means clustering; magnetic resonance imaging; tumor volume determination; oncology

DIFFUSE intrinsic pontine gliomas (DIPGs) are inoperable and lethal brainstem gliomas that are found in children, with a median age of diagnosis of 6–7 years.9 DIPGs account for approximately 80% of pediatric brainstem tumors. Despite numerous efforts to improve treatment, the median survival of patients with these tumors is less than 1 year.6,10

Accurately determining DIPG tumor volume is important for therapeutic planning, monitoring disease progression, and measuring tumor response to therapy.3,18 Previous studies using MRI and semiautomated image segmentation methods to determine DIPG tumor volume showed poor interobserver agreement.3,9,17 The image segmentation methods used in these studies depend highly on subjective judgment of the observers. We hypothesized that a method requiring a low degree of an observer’s subjective judgment would produce better interobserver agreement and more consistent measurements. The aims
of the current study were to 1) measure DIPG tumor volumes using methods that require different degrees of subjective judgment, and 2) evaluate interobserver agreement of measurements made using these methods.

**Methods**

**Patients and Image Selection**

Eight patients with radiographically confirmed diagnoses of DIPG were the subjects of this analysis. Patients participated in a Phase I clinical trial testing convection-enhanced delivery (CED) of therapeutic murine monoclonal antibody 124I-8H9 (clinical trial registration no. NCT01502917; http://clinicaltrials.gov), approved by the Memorial Sloan Kettering Cancer Center (MSKCC) Institutional Review Board. Informed consent was obtained from all patients or their parents/legal guardians. All patients were enrolled in this trial within 4–14 weeks after standard conventional external-beam radiation therapy had been administered, and in the absence of clinical and radiographic disease progression. Pre-CED, post–radiation therapy axial T2-weighted images were acquired at 3T (GE Discovery MR750; TE 99–104 msec, TR > 3000 msec, acquisition matrix size 256 × 256 [2 scans] or 512 × 512 [6 scans], voxel size 0.555–2.651 mm³).

Axial T2-weighted images were analyzed by 2 independent observers trained to identify DIPG on T2-weighted images using 3 methods, including 2 image segmentation methods that require high degrees of subjective judgment (picture archiving and communication system [PACS] polygon method and Volume Viewer auto-contour method) and a parameter-dependent method that requires a low degree of subjective judgment (k-means clustering segmentation). One independent observer (R.S.) had 2 years of experience in measuring DIPG volumes and 2 years of experience in image segmentation software. The other observer (J.T.) had 1 year of experience in measuring DIPG volumes, 2 years of experience in neuroradiology, and 1 year of experience in image segmentation software. Tumor volume assessment was overseen by a board-certified radiologist (S.H.) holding a certificate of added qualification in neuroradiology (11 years of neuroradiology experience).

**Image Segmentation Methods Requiring High Degrees of Subjective Judgment**

Tumor volumes were measured using the PACS polygon method (Centricity PACS 4.0 SP3, GE Healthcare) and the Volume Viewer auto-contour method (Advantage Workstation V11.3–4.2, GE Healthcare). For the PACS polygon method, observers manually outlined the tumor and the software determined the tumor area per slice. Observers then added the areas from consecutive slices and multiplied the sum by slice spacing to derive total tumor volume (Fig. 1).

The Volume Viewer auto-contour method is based on a region-growing method. Observers selected a seed region in the tumor on an axial T2-weighted slice and allowed it to grow to the border of the tumor, after which the software automatically highlighted the tumor in other slices and represented the tumor in reconstructed sagittal and coronal planes. A 3D representation of the tumor was constructed and tumor volume was calculated by the software (Fig. 2). The observers were allowed to select and unselect highlighted areas to improve tumor approximation.

**Parameter-Dependent Image Segmentation Method**

Tumor volumes were also determined using k-means clustering segmentation in ImageJ (National Institutes of Health). k-means clustering segmentation is a parameter-dependent segmentation technique in which the parameters are the number of clusters assigned by the observers, where clusters are generated based on the signal intensity and location of the voxels of the images. For each patient, to apply k-means clustering segmentation to the site of tumor, the observers selected a region of interest (ROI) that included the entire tumor with a safe margin on the slice containing the largest cross-sectional area of the tumor. The ROI was thereafter extended orthogonally to the remaining tumor-containing slices. The observers then assigned the number of clusters (k), which remained unchanged for all slices, and applied k-means clustering segmentation (plugin developed by Jarosław Sacha, Procter & Gamble Company, Miami Valley Innovation Center; http://sourceforge.net/projects/ij-plugins/) to the slices. The algorithm partitions the voxels into the designated number of clusters by iteratively calculating and minimizing intracluster dissimilarity until further reduction is not possible, i.e., each cluster contains voxels that are similar to each other in signal intensity and location and dissimilar to voxels in other clusters (Fig. 3). In this study, all images were successfully clustered using 4–6 clusters.

**Statistical Analysis**

Interobserver agreement was evaluated using Lin’s concordance correlation coefficient (CCC).11 Modified Bland-Altman plots2 were used to visually demonstrate the presence or absence of agreement between measurements made by the 2 observers.

**Results**

The median age of patients was 6 years (range 3–8 years), and the mean (± SD) time between completion of radiation therapy and the analyzed MR images was 65 ± 16 days. The average tumor volumes measured by the 2 observers using the PACS polygon method were 11.48 ± 7.71 cm³ and 13.61 ± 7.76 cm³ (Table 1). The CCC of the 2 observers’ measurements was 0.9465 with a lower 1-sided 95% confidence limit of 0.8472, which was considered poor agreement.5,16 Precision (0.9872, lower 1-sided 95% confidence limit 0.9453) and accuracy (0.9588, lower 1-sided 95% confidence limit 0.8756) suggested that the poor agreement resulted mainly from differences between the judgments of the 2 observers. One observer consistently measured greater tumor volumes than the other (Table 1). The modified Bland-Altman plot showed that the difference between the observers’ measurements was more than 25% in 3 of the 8 tumors (Fig. 4A).

The average tumor volumes measured by the 2 observers using the Volume Viewer auto-contour method were 13.64 ± 6.40 cm³ and 14.77 ± 7.26 cm³. The CCC of
FIG. 1. PACS polygon method. Observers manually outline the tumor in each tumor-containing slice and the software determines the tumor area per slice. The areas from consecutive slices are then summed and multiplied by slice spacing to produce the total tumor volume \((\text{Area 1} + \text{Area 2} + \text{Area 3} + \text{Area 4} + \text{Area 5}) \times \text{slice spacing} = \text{tumor volume})\). Figure is available in color online only.

FIG. 2. Volume Viewer auto-contour method. Green lines represent the tumor border that the software identifies. A: Observers select a seed region in the tumor in an axial T2-weighted MRI slice and allow it to grow to the border of the tumor. B and C: The software automatically highlights the tumor in other slices and represents the tumor in reconstructed sagittal and coronal planes. D: A 3D representation of the tumor is constructed and tumor volume is calculated. Figure is available in color online only.
the 2 observers’ measurements was 0.7514 with a lower 1-sided 95% confidence limit of 0.3143, which was considered poor agreement.15,16 Precision (0.7691, lower 1-sided 95% confidence limit 0.2753) and accuracy (0.9769, lower 1-sided 95% confidence limit 0.3873) suggested that the poor agreement was a result of both poor consistency in the judgments of each individual observer and differences between the judgments of the 2 observers. The modified Bland-Altman plot highlighted that the difference between the 2 observers’ measurements was more than 25% (positive or negative) in 5 of the 8 tumors (Fig. 4B).

The average tumor volumes measured by the 2 observers using k-means clustering segmentation were 8.97 ± 5.55 cm³ and 8.88 ± 5.69 cm³ (Table 1). The CCC of the 2 observers was 0.9938 with a lower 1-sided 95% confidence limit of 0.9772, which was considered substantial strength of agreement.15,16 Precision (0.9942, lower 1-sided 95% confidence limit 0.9752) and accuracy (0.9996, lower 1-sided 95% confidence limit 0.9432) suggested that both judgments between the 2 observers and consistency within each individual observer were good. The modified Bland-Altman plot showed that none of the measurements of the 2 observers differed by more than 25% (Fig. 4C).

k-means clustering segmentation delineated the tu-
Determining diffuse intrinsic pontine glioma tumor volume

Discussion

Accurate and reproducible measurements of DIPG tumor volumes are clinically important. CED, one of the major available techniques to overcome the blood-brain barrier in treating DIPG, requires an accurate determination of DIPG tumor volume to determine optimal dosage. Furthermore, tumor volume serves as a marker for tumor response and progression. In our study we assessed the utility of using MR images and image segmentation techniques requiring different degrees of an observer's subjective judgment to determine DIPG tumor volume.

In both PACS polygon and Volume Viewer auto-contour methods, the 2 observers showed poor agreement, a reflection of the high degrees of subjective judgment involved. Poor interobserver agreement undermines the believability of the methods. In the clinical setting, poor interobserver agreement also creates difficulty in continuous care from multiple providers. Previous studies using a method similar to the PACS polygon method obtained tumor area per slice by using the 2 longest diameters rather than outlining the tumor perimeter. While outlining the tumor perimeter, as performed in the current study, produces less variability between observers compared with using the 2 longest diameters, especially for irregular, nonspherical tumors. Despite this, the interobserver agreement was still poor using the PACS polygon method in our study. These data underscore the difficulty of achieving interobserver agreement in measuring DIPG tumor volumes using methods that require high degrees of subjective judgment, a difficulty that has also been demonstrated in other studies. Furthermore, many of the tumor volumes determined using PACS polygon and Volume Viewer auto-contour methods differed by more than 25%, which could potentially lead to significantly under- or overdosing patients receiving CED therapy for DIPG treatment.

Due to the poor interobserver agreement of methods requiring high degrees of subjective judgment, we investigated the use of k-means clustering segmentation, a parameter-dependent segmentation technique that requires a low degree of subjective judgment. k-means clustering segmentation demonstrated a higher interobserver agreement than the PACS Polygon or Volume Viewer auto-contour methods. None of the measurements of the 2 observers using k-means clustering differed by more than 25%, which is another demonstration that the data showed better interobserver agreement. Average tumor volumes measured using k-means clustering segmentation were lower than those from the other 2 methods. The likely reason for these outcomes is that k-means clustering segmentation delineated the tumor, which allowed the observer to select only clusters making up the tumor. k-means clustering segmentation also delineated components of the tumor with different signal intensities.

While k-means clustering segmentation delineated the tumor and showed substantial agreement between observers, this methodology needs improvement before it could be used routinely by clinicians. k-means clustering segmentation was the most time-consuming of the 3 methods because observers had to individually cluster each slice containing the tumor. The ImageJ k-means clustering plugin has the capability to perform segmentation on 3D images. However, it was difficult to clearly visualize the clusters and to identify clusters containing tumor on the segmented 3D images. Therefore we elected to perform k-means clustering segmentation on individual slices. Segmentation of 3D images would be ideal if the segmented clusters could be easily inspected. However, to our knowledge, there is currently no user-friendly medical image processing software for performing and visualizing k-means clustering segmentation on 3D images.

By design, k-means clustering segmentation is param-

<table>
<thead>
<tr>
<th>Case No.</th>
<th>PACS Polygon (cm³)†</th>
<th>Volume Viewer Auto-Contour (cm³)‡</th>
<th>k-means Clustering Segmentation (cm³)§</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observer 1</td>
<td>Observer 2</td>
<td>Difference (%)</td>
</tr>
<tr>
<td>1</td>
<td>6.95</td>
<td>10.51</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>23.86</td>
<td>25.18</td>
<td>5.4</td>
</tr>
<tr>
<td>3</td>
<td>9.64</td>
<td>13.65</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>13.97</td>
<td>14.85</td>
<td>6.1</td>
</tr>
<tr>
<td>5</td>
<td>21.23</td>
<td>24.25</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>8.99</td>
<td>10.40</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>5.64</td>
<td>6.50</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>1.59</td>
<td>3.50</td>
<td>75</td>
</tr>
</tbody>
</table>

Mean ± SD: 11.48 ± 7.71 13.61 ± 7.76 13.64 ± 7.26 8.97 ± 5.55 8.88 ± 5.69

* Difference was calculated as the absolute difference between measurements made by the 2 observers divided by the average of the measurements made by the 2 observers.
† Poor agreement between observers, CCC = 0.9465 (lower 1-sided 95% confidence limit 0.8472).
‡ Poor agreement between observers, CCC = 0.7514 (lower 1-sided 95% confidence limit 0.3143).
§ Substantial agreement between observers, CCC = 0.9938 (lower 1-sided 95% confidence limit 0.9772).
eter dependent, and requires the observer to assign the value of $k$ (number of clusters), which could introduce subjectivity into tumor volume determination. To identify an adequate $k$ value in the current study, as in all applications of $k$-means clustering, the observers had to experiment with a few values until a clustered image that delineated the tumor was generated. All images were successfully clustered with 4–6 clusters. Choosing the optimal value of $k$ is important in generating the most medically relevant results, which also requires further study. One approach to choosing an optimal $k$ value is to generate a histogram of voxel signal intensity values of the images. The $k$ value should be equal to or greater than the number of peaks in the histogram. If a peak consists of groups of voxels at

![Modified Bland-Altman plots of measurements made by the 2 observers. In these plots, the percentage difference between measurements made by the 2 observers is plotted against the average of the measurements made by the 2 observers. The 2 dashed lines denote 25% difference (positive or negative). A: When using the PACS polygon method, the measurements of the observers differed by more than 25% in 3 of 8 tumors. One observer consistently obtained larger tumor volumes than the other observer. B: When using the Volume Viewer auto-contour method, the measurements of the observers differed by more than 25% in 5 of 8 tumors. C: When using $k$-means clustering segmentation, none of the measurements of the 2 observers differed by more than 25%.](image-url)
Determining diffuse intrinsic pontine glioma tumor volume

more than 1 spatial location, then this peak needs to be represented by more than 1 cluster. The other question in producing the most medically relevant clusters is which clustering algorithms (both the initialization method and the iterative refinement method) to use. The ImageJ \( k \)-means clustering segmentation plugin uses the \( k \)-means++ \(^1\) initialization and MacQueen\(^{13}\) iterative refinement (called the standard \( k \)-means method by many), which are the more robust ones among available algorithms. While the most commonly used algorithm in recent years is the Hartigan-Wong\(^7\) method, the Forgy,\(^4\) Lloyd,\(^12\) and MacQueen\(^{13}\) methods all have their advantages and disadvantages. Even though there has been considerable discussion in the literature, there does not seem to be a universally advantageous algorithm.

While a few other sequences—such as T1-weighted imaging with and without Gd contrast and T2-FLAIR—are also included in MRI studies of patients with DIPG, T2-weighted images were used in this study for tumor volume determination because T2-FLAIR imaging generates more artifacts in the posterior fossa. Furthermore, T2-FLAIR imaging may miss white matter lesions in the posterior fossa, which is noteworthy because DIPGs classically grow along white matter tracts.\(^{22}\) We did not use T1-weighted imaging in the analysis because in DIPGs, T1-weighted imaging without contrast does not show tumor well, and contrast enhancement with Gd is variable, with no enhancement being common.\(^{14}\)

One limitation of this study was that the actual DIPG tumor volume was not known, as resection is not a treatment option for these tumors. However, because PACS polygon and Volume Viewer auto-contour methods produced poor interobserver agreement, they may be less likely to represent the actual tumor volume compared with \( k \)-means, which produced substantial interobserver agreement. It was also not possible to accurately differentiate the tumor boundary from peritumoral edema using standard MRI techniques. This is an inherent shortcoming of all studies attempting to determine preoperative brain tumor volumes. We attempted to minimize the extent of peritumoral edema by using scans from approximately 2 months after radiation therapy. Another limitation was that the current study used only 2 of the numerous available image segmentation methods that require a significant degree of subjective judgment. The PACS polygon and Volume Viewer auto-contour methods were chosen because they were readily available and widely used by clinicians.

Conclusions

Image segmentation methods that require high degrees of subjective judgment show poor interobserver agreement in determining DIPG tumor volumes. \( k \)-means clustering segmentation, which is parameter-dependent and requires a low degree of subjective judgment, shows good interobserver agreement and delineates DIPG tumors. To make \( k \)-means clustering segmentation easily usable in clinical settings, further studies should be conducted to optimize this image segmentation method, including developing algorithms for optimal parameter setting, and 3D image visualization for easy inspection of clusters on clustered images. The aim would be to streamline the processing steps into a more automated method.

Acknowledgments

We would like to acknowledge the following organizations...
and programs for supporting the study: The Alex Lemonade Stand Foundation Pediatric Oncology Student Training Program, The Matthew Larson Foundation, The Cure Starts Now Foundation, The Dana Foundation, The Solving Kids’ Cancer Foundation, The Lyla Nsofui Foundation, The Perry’s Promise Fund, Witmer Fund-Pediatric Brain Tumor, Cookies For Kids’ Cancer, MSKCC internal funding, The St. Baldrick’s Foundation, The Pediatric Brain Tumor Foundation, The McKenna Claire Foundation, The Cristian Rivera Foundation, and The Ian’s Friends Foundation. The authors are thankful to Ms. Maria Donzelli, Pediatric Neurosurgery Nurse Practitioner, MSKCC, for providing the necessary patient information. The authors are also thankful for the expert editorial advice provided by Ms. Ada Mueller, Editor, Department of Radiology, MSKCC.

References


Disclosures

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions

Conception and design: Zhou, Singh, Haque, Peck, Young, Tsiorinis, Thakur, Souweidane. Acquisition of data: Singh, Tisnado, Haque. Analysis and interpretation of data: all authors. Drafting the article: Zhou, Singh, Tisnado, Haque, Peck, Young, Souweidane. Critically revising the article: Zhou, Singh, Tisnado, Haque, Tsiorinis, Souweidane. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Zhou. Statistical analysis: Zhou. Study supervision: Zhou, Haque, Souweidane.

Correspondence

Zhiping Zhou, 1300 York Ave., Box 99, New York, NY 10065. Email: zhipingzhou@gmail.com.