Published in final edited form as:
J Neurooncol. 2008 October ; 90(1): 57–61. doi:10.1007/s11060-008-9631-4.

Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas

Robert M. Hayward, **Nicolas Patronas**, **Eva H. Baker**, **Gilbert Vézina**, **Paul S. Albert**, and **Katherine E. Warren**

Abstract

Diffuse intrinsic pontine glioma (DIPG) is an invasive pediatric brainstem tumor with a poor prognosis. Patients commonly enter investigational trials, many of which use radiographic response as an endpoint for assessing drug efficacy. However, DIPGs are difficult to measure on magnetic resonance imaging (MRI). In this study, we characterized the reproducibility of these commonly performed measurements. Each of four readers measured 50 MRI scans from DIPG patients and inter-observer variability was estimated with descriptive statistics. Results confirmed that there is wide variability in DIPG tumor measurements between readers for all image types. Measurements on FLAIR imaging were most consistent. For patients on clinical trials, measurement of DIPG should be performed by a single reader while comparing prior images side-by-side. Endpoints for clinical trials determining efficacy in this population should also include more objective measures, such as survival, and additional endpoints need to be investigated.

Keywords

brainstem; chemotherapy; childhood malignant gliomas; diffuse intrinsic pontine gliomas; magnetic resonance imaging; tumor measurements

INTRODUCTION

Diffuse intrinsic pontine gliomas (DIPG) account for approximately 15 percent of childhood central nervous system (CNS) tumors [1-3]. Conventional local field radiation therapy is the standard treatment [4], after which a significant number of patients improve clinically. However, the progression-free survival interval is short and median survival is less than one year from diagnosis [3]. Due to the poor prognosis of DIPG, many patients enter investigational trials before, during, or after radiation therapy. In many of these trials, endpoints for assessing drug efficacy are determined by measuring the change in tumor size on magnetic resonance imaging relative to a pre-treatment or best-response scan. Although unidimensional methods (e.g. RECIST criteria) have been proposed, national pediatric consortia generally utilize two-dimensional tumor measurements (WHO criteria) as measures of efficacy in investigational trials [5,6]. For some CNS tumors, including DIPG, the use of change in tumor size on magnetic resonance imaging is problematic due to the difficulty in delineating the boundaries of these invasive lesions. For the measurement of DIPG tumors in particular, significant inter-reader variability is suspected. The objective of this study is to quantify the variability in DIPG tumor measurements among 4 independent readers.

R.M. Hayward · K.E. Warren, Pediatric Oncology Branch, National Cancer Institute, National Institutes of Health, 9000 Rockville Pike, Building 10 CRC, Rm 1-3940, Bethesda, MD 20892, Fax: 301-451-7052, Email: warrenk@mail.nih.gov.

MATERIALS AND METHODS

Patients and MRI Exam Selection

A total of 50 MRI exams were selected from 16 DIPG patients (6 M, 10 F) enrolled in a phase II study of pegylated interferon alfa-2b (PEG-Intron™) administered after radiation therapy. The median age of the patients was 6 (range 1.8 to 12) years. To be included in this study, subjects were required to have a non-exophytic DIPG involving more than 50% of the pons with the center of the lesion in the pons. In addition, each lesion had to be hypointense on T1-weighted MR images and hyperintense on T2-weighted MR images. Subjects greater than 21 years of age were excluded from the study. Scans included those after radiation therapy and before starting PEG-Intron as well as those during the course of PEG-Intron treatment. All imaging examinations had been obtained and analyzed in the context of an institutional review board approved protocol at the National Cancer Institute.

MR Imaging

MRI was performed using a 1.5 tesla scanner with a standard quadrature head coil. Prior to contrast infusion, axial FLAIR, axial T2-weighted, and axial T1-weighted sequences were obtained. After contrast administration, axial T1-weighted and axial FLAIR sequences were acquired (Fig. 1). Most patients were sedated with propofol and received supplemental oxygen during the examination.

Tumor Measurements

Three neuroradiologists and one pediatric neuro-oncologist each independently measured tumors on 50 MRI exams on PACS workstations. Readers were blinded to the identity and clinical status of the study subjects. Exams were presented to the readers in a random order that was the same for all readers. All measurements were made on axial images. For each imaging sequence in each MR examination, the reader identified the slice in which the tumor had the largest diameter in any one direction and measured that diameter (d_1, see Figure 2) as well the corresponding largest perpendicular diameter on the same slice (d_2). The third diameter (d_3) was determined by counting the number of axial slices showing contiguous evidence of tumor and multiplying by the slice thickness. Tumor bidimensional product was calculated as $d_1 \times d_2$, while tri-dimensional product was as $d_1 \times d_2 \times d_3$. If multiple discrete lesions were present, readers were instructed to measure only the largest lesion. Lesions estimated to be less than 10 mm in any of the three diameters were considered to be not measurable.

Each reader made a total of 450 measurements, as 3 measurements were made on 3 imaging sequences (T2-weighted, FLAIR, and post-contrast T1-weighted images) for each of 50 exams. Fifty-four of the 450 measurements were not included in the statistical analysis for the following reasons: lesions were not measurable (less than 10 mm in any of the dimensions of interest) (36), technical issues related to the uploading of deidentified scans onto the system (13), failure of any one reader to record a given measurement (2), and poor scan quality (3).

Statistical Analysis

Analysis was carried out for 1, 2, and 3 dimensional measurements for each of the 3 imaging sequences (T2-weighted, FLAIR, and post-contrast T1-weighted images), for a total of nine variables. For each variable, the median % difference between each of the 6 reader pairs was computed for each examination. The upper 95th percentile of these measurements was presented for each imaging variable. The median (and ranges) examination-specific coefficient of variation was also presented. Specifically, the coefficient of variation (defined as the standard deviation divided by the mean) was computed for each examination, and then the median was taken over all examinations. Also, the % of cases for which two raters disagree by

more than 25% was estimated by enumerating all pairwise differences and computing the proportion that differ more than 25%.

RESULTS

Patient and Tumor Characteristics

The number of scans included in the analysis, the mean tumor size, and the range of tumor sizes for each measurement method are presented in Table 1.

Measures of inter-observer variability

The median relative difference and median coefficient of variation for each measurement strategy are presented in Table 2. These statistics show that there is a significant amount of inter-observer variation for each measurement strategy. For example, for 2D measurement on post-contrast T1-weighted images, the median absolute difference between any two raters is 12.7% and the median CV across individuals is 13.9 (range 2.8-66.2)%. Two-dimensional measurements on FLAIR images had the lowest median CV of 8.5%.

Inter-observer variability is further shown in Table 3, where the upper 95th percentile for the difference between any two raters and the % of cases for which two or more raters disagree by more than 25% are presented. For example, for 2D measurement on FLAIR images, any two raters agree within 62% in 95% of cases. In other words, any two raters disagree by more than 62% in 5% of cases. Further, Table 3 shows that for 2D measurement on FLAIR images, in 19% of cases, raters disagree by more than 25%.

DISCUSSION

In this study, we determined the inter-observer variability in tumor measurements for diffuse intrinsic pontine gliomas using 3 MR imaging sequences (post-contrast T1-weighted, T2-weighted, and FLAIR images). Tumor size was analyzed in 1, 2, and 3 dimensions by 4 independent readers. For each measurement strategy, inter-observer variability was examined by estimating the median relative differences between any two raters, the median coefficient of variation across patient-visits, and the percent of cases for which two raters disagreed by more than 25%. We also estimated the 95th percentile in the distribution of differences between any two raters. All methods showed significant inter-observer variation.

Increased agreement between raters appears to be a benefit of unidimensional measurements. As seen in Table 3, the 95th percentiles are narrower in percentage terms for one-dimensional measurements than for two-dimensional measurements. This is not wholly unexpected, as inter-rater reproducibility is expected to improve as the number of dimensions that must be measured decreases.

The results of this study suggest that FLAIR images yield more consistent results between raters than T2-weighted images. As seen in Table 3, the 95th percentiles were narrower for FLAIR than for T2-weighted sequences for one, two, and three dimensional measurement methods. This effect is likely due to an intrinsic advantage in measuring the boundaries of these tumors on FLAIR images, but it could also be due to the fact that the tumors had larger measured sizes on FLAIR than T2 (Table 1). This would tend to decrease percent disagreement for the larger tumor if the disagreement is the same in absolute terms.

Drug efficacy is frequently evaluated by change in tumor size on imaging studies in response to therapy. Standard World Health Organization (WHO) response criteria utilize 2-dimensional (2D) tumor measurements (the product of the longest diameter and its longest perpendicular diameter for each tumor). A complete response is defined as disappearance of all known disease

for a minimum of 4 weeks, partial response is a ≥50% decrease in the sum of the products of perpendicular diameters of all measured tumors, and progressive disease is a ≥25% increase in the product of perpendicular diameters of any measurable lesion or the appearance of new lesions. Minor response criteria using a decrease in tumor size of ≥25% but <50% is sometimes used. Tumor measurements that do not fulfill the criteria for an objective response or progressive disease are considered to be stable disease. However, which MR sequence and plane on which to perform measurements is not frequently defined in investigational studies, and is therefore determined by reader preference.

For two-dimensional measurements, readers disagreed by more than 25% in their measurements a substantial portion of the time (29% for post-contrast T1, 21% for T2, 19% for FLAIR, Table 3). Since a 25% change in tumor size over time is enough to make the difference between an objective response or stable disease, the observed disagreement between readers is likely sizable enough to affect the radiographic determination of response to treatment if different radiologists are measuring the tumors at different time points. We therefore strongly recommend that a single reader measure each patient's DIPG with images from different time points side-by-side. This study also stresses the importance of central radiographic review for patients on investigational studies.

Additional measurement methods have been investigated in an effort to improve precision and be less user-dependent. RECIST criteria were introduced as a method of unidimensional measurement to determine response criteria. This method was simple, relatively quick to perform, and potentially reduce additional error associated with measuring lesions in multiple dimensions. However, diameter measurements have difficulty with irregular lesions, as well as lesions with cystic and necrotic regions [7,8]. Pediatric and brain tumor trials have continued to rely on two-dimensional measurements more so than trials involving other tumor types [7, 9]. Recently, interest has also focused on three-dimensional tumor measurements performed with or without the assistance of a computer. Studies of pediatric brain tumors have shown that concordance in determining partial response is high comparing 1D, 2D and 3D measurements. However, concordance was lower for defining partial response and disease progression, which are more often observed in the pontine glioma setting.[8]. Computer-aided volumetric methods for determining tumor size have also been studied. In these techniques, a computer determines the border between tumor and normal tissue on each slice containing tumor, then calculates tumor volume with a variable amount of input from a radiologist (depending on the specific method). Automated segmentation of adult brain tumors has been demonstrated to be a rapid method with accuracy comparable to manual segmentation methods [10]. One advantage of computer-assisted volumetric assessment is that the computer can segment out non-enhancing or cystic components [7]. A disadvantage is that these methods still rely on determination of a threshold defining tumor or nontumor tissue.

One of the difficulties in managing patients with diffuse intrinsic pontine gliomas is that sustained radiographic responses are rare and rapid clinical progression, with or without radiographic progression, is common. As a result, progression is often determined clinically both in the management of patients and in clinical trials. Future studies should therefore focus on additional, more objective endpoints, such as survival. Radiographic response does remain an important component of investigational trials therefore reproducibility and standard procedures are necessary to optimize their use.

One potential weakness of this study was the fact that several MRI examinations had to be removed from the analysis for reasons listed previously. These reasons do not appear to be related to the nature of the scans that were excluded, so bias is unlikely although possible. Another potential source of variability is that readers had the option of performing FLAIR measurements on pre-contrast or post-contrast FLAIR series. We do, however, expect the

differences between pre-contrast and post-contrast FLAIR series to be small, especially when measuring solitary pontine glioma lesions. A final weakness of the study is that we have not examined reproducibility of measurements by a single reader. It is possible that some readers systematically measure all tumors as larger or smaller than other readers and this could account for some of the inter-reader variability.

In conclusion, there is significant potential for disagreement among measurers of diffuse intrinsic pontine gliomas on MRI and these differences are likely significant in terms of their effect in determining tumor response or progression. Measurements on FLAIR imaging are most consistent. Given these results, measurements for patients with DIPG should be performed by a single reader using side-by-side images from each patient in order to determine response. In addition, endpoints for clinical trials determining efficacy should include more objective measures, e.g. survival, and additional endpoints need to be investigated.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research. The views expressed do not necessarily represent the views of the National Institutes of Health or the United States Government.

REFERENCES

1. Smith MA, et al. Trends in reported incidence of primary malignant brain tumors in children in the United States. J Natl Cancer Inst 1998;90(17):1269–77. [PubMed: 9731733]
2. Freeman CR, Farmer JP. Pediatric brain stem gliomas: a review. Int J Radiat Oncol Biol Phys 1998;40 (2):265–71. [PubMed: 9457808]
3. Hargrave D, Bartels U, Bouffet E. Diffuse brainstem glioma in children: critical review of clinical trials. Lancet Oncol 2006;7(3):241–8. [PubMed: 16510333]
4. Finlay JL, Zacharoulis S. The treatment of high grade gliomas and diffuse intrinsic pontine tumors of childhood and adolescence: a historical - and futuristic - perspective. J Neurooncol 2005;75(3):253–66. [PubMed: 16195805]
5. Miller AB, et al. Reporting results of cancer treatment. Cancer 1981;47(1):207–14. [PubMed: 7459811]
6. Therasse P, et al. European Organization for Research and Treatment of Cancer; National Cancer Institute of the United States; National Cancer Institute of Canada. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 2000;92(3):205–16. [PubMed: 10655437]
7. Henson JW, Ulmer S, Harris GJ. Brain Tumor Imaging in Clinical Trials. AJNR Am J Neuroradiol 2008;29(3):419–424. [PubMed: 18272557]
8. Warren KE, et al. Comparison of one-, two-, and three-dimensional measurements of childhood brain tumors. J Natl Cancer Inst 2001;93(18):1401–5. [PubMed: 11562391]
9. Therasse P, Eisenhauer EA, Verweij J. RECIST revisited: a review of validation studies on tumour assessment. Eur J Cancer 2006;42(8):1031–9. [PubMed: 16616487]
10. Kaus MR, et al. Automated segmentation of MR images of brain tumors. Radiology 2001;218(2): 586–91. [PubMed: 11161183]

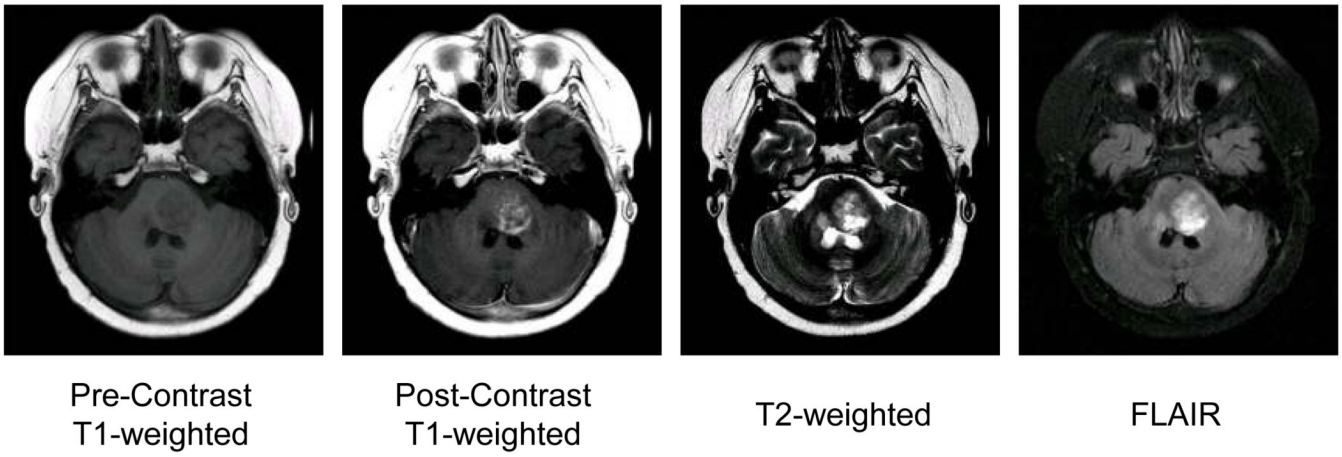

Figure 1. MRI Examples

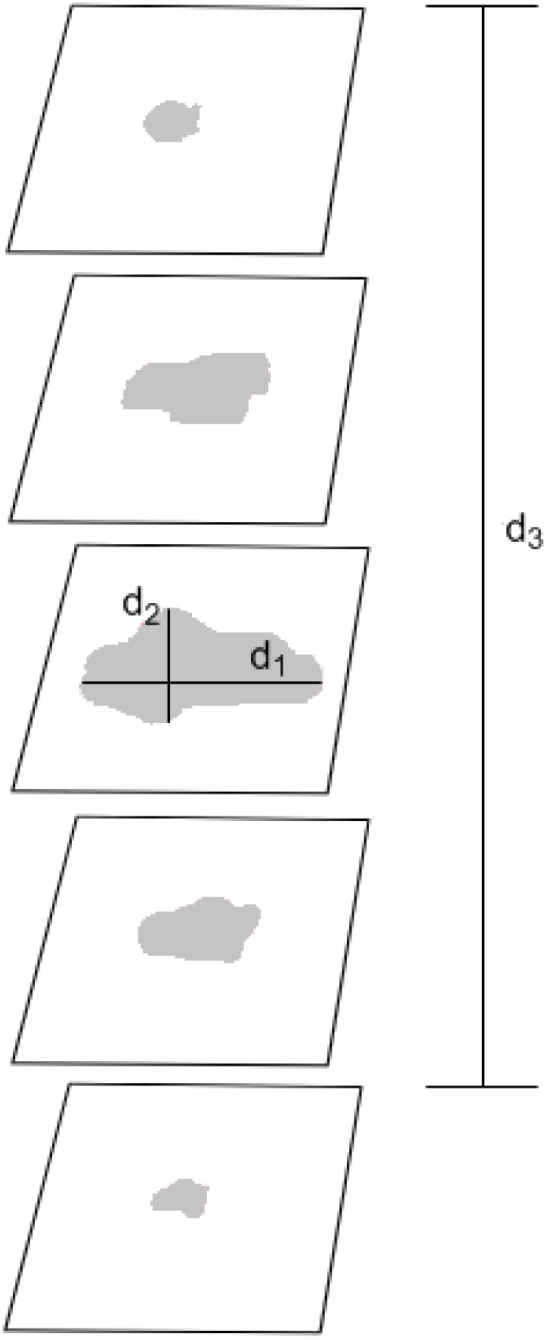

Figure 2. Measurement Procedure

The diagram depicts contiguous axial slices with the tumor in gray. d_1 and d_2 are the longest perpendicular diameters for the largest slice. d_3 is determined by multiplying the number of slices involved by the slice thickness.

Table 1

The number of scans included in the analysis, the mean tumor size, and the range of tumor sizes for each measurement method

	Scans Evaluated	Mean Tumor Size	Tumor Size Range*
1 Dimensional			
Post-contrast T1-weighted	38	3.0 cm	1.0-5.3 cm
T2-weighted	49	4.2 cm	2.4-8.2 cm
FLAIR	50	4.4 cm	2.2-8.6 cm
2 Dimensional			
Post-contrast T1-weighted	33	8.4 cm^2	2.4-20.9 cm^2
T2-weighted	49	14.2 cm^2	4.0-57.3 cm^2
FLAIR	50	15.4 cm^2	3.8-61.4 cm^2
3 Dimensional			
Post-contrast T1-weighted	31	30.1 cm^3	4.3-109.1 cm^3
T2-weighted	48	61.6 cm^3	6.9-405.5 cm^3
FLAIR	48	68.7 cm^3	6.4-459.3 cm^3

* Each tumor measurement was averaged across the four raters before the range of alll measurements was computed

Table 2

The median relative difference and median coefficient of variation(CV) for each measurement method

	Scans Evaluated	Median % Difference between Rater Pairs	Median CV (%)	CV Range (%)
1 Dimensional				
Post-contrast T1-weighted	38	6.1	7.8	0.0 - 62.8
T2-weighted	49	5.2	5.0	0.9 - 40.5
FLAIR	50	5.9	5.2	0.0 -38.9
2 Dimensional				
Post-contrast T1-weighted	33	12.7	13.9	2.8 - 66.2
T2-weighted	49	9.6	10.1	1.4 - 59.9
FLAIR	50	10.2	8.5	2.4 - 58.1
3 Dimensional				
Post-contrast T1-weighted	31	16.9	16.1	1.9 - 83.9
T2-weighted	48	17.4	15.0	1.4 - 82.8
FLAIR	48	14.6	13.6	3.0 - 76.2

Table 3

The 95th percentile for the difference between two raters and the % of cases for which two or more raters disagreed by more than 25% for each measurement method

	95th Percentile for Difference Between Rater Pairs	% of Pairings for Which Two Raters Disagreed by More Than 25%
1 Dimensional		
Post-contrast T1-weighted	± 127%	20%
T2-weighted	± 47%	11%
FLAIR	± 40%	8%
2 Dimensional		
Post-contrast T1-weighted	± 216%	29%
T2-weighted	± 100%	21%
FLAIR	± 62%	19%
3 Dimensional		
Post-contrast T1-weighted	± 277%	37%
T2-weighted	± 191%	32%
FLAIR	± 119%	25%