

Assessment of intra-observer variability in measurement of high-grade brain tumors

James M. Provenzale · Michael C. Mancini

Received: 30 July 2011 / Accepted: 20 February 2012 / Published online: 10 March 2012
© Springer Science+Business Media, LLC. 2012

Abstract A 25% increase in bidimensional products (BPs) of tumor diameter has been used as a criterion for brain tumor progression. We studied intra-observer variability in measurements of BPs. Ten patients with contrast-enhancing glioblastoma multiforme underwent baseline and follow-up MR imaging. Seven observers measured BPs in various planes. Differences in BPs between scans were expressed as a percentage of baseline. This calculation was performed for both readings of the baseline and follow-up scans. Differences between change from baseline to follow-up on each reading (termed *D* values) were calculated for each reader (total of 196 *D* values). Median *D* value in each plane was calculated for each reader. Range of *D* values was 12.36–33.64% in axial plane (average 10.63%), 12.18–38.62% in coronal plane (average 26.84%) and 15.12–35.48% in sagittal plane (average

26.11%). Across all planes, 88 (45%) *D* values were >25%. When all imaging planes for any single observation were combined, in 76% of cases, at least one *D* value of >25% was seen. Based on the high degree of intra-observer variability, tumor measurements producing an increase in BP of >25% can routinely be obtained solely by chance.

Keywords Tumor · Brain · Glioblastoma multiforme · Intra-observer · Variability · Measurements · Macdonald criteria · Progression · Clinical trial

Introduction

Methods of assessment of tumor response to treatment are a topic of great interest because of their importance for development of clinical trials. The major method in use for assessing response of high-grade brain tumors to therapy for the past few decades has been the Macdonald criteria, which uses the product of two perpendicular measurements in the largest area of contrast-enhancing tumor as a metric [1]. However, a number of deficiencies in use of these criteria have been noted [2, 3]. In an attempt to take some of these limitations into account, revised criteria have been suggested by the response assessment neuro-oncology (RANO) working group, which attempt to overcome some of the limitations of the Macdonald criteria [4]. However, the new criteria continue to employ bidimensional product (BP) of measurements of enhancing tumor (in addition to other features).

Techniques that are heavily reliant on manual measurements are subject to variability within, and between, observers. A major issue in any type of tumor measurement is the issue of reproducibility. In studies evaluating inter-observer variability in brain tumor measurements, poor

Electronic supplementary material The online version of this article (doi:10.1007/s11060-012-0843-2) contains supplementary material, which is available to authorized users.

J. M. Provenzale (✉)
Department of Radiology, Duke University Medical Center,
Durham, NC 27710, USA
e-mail: Prove001@mc.duke.edu

J. M. Provenzale
Department of Radiology, Oncology and Biomedical
Engineering, Emory University School of Medicine,
Atlanta, GA 30322, USA

M. C. Mancini
Department of Biomedical Engineering, Georgia Institute
of Technology, Atlanta, GA 30332, USA

M. C. Mancini
Department of Biomedical Engineering, Emory University
School of Medicine, Atlanta, GA 30322, USA

inter-observer reproducibility has been shown [5, 6]. In a previous study, we found marker variability between observers of similar training in measuring BP measurements in a series of high-grade brain tumors [6]. Such variability resulted in marked differences between observers as to whether a tumor would have been judged as stable or showing progression.

As opposed to inter-observer variability, the topic of intra-observer variability in measuring diameter of contrast-enhancing tumors has rarely been addressed [7]. Differences within measurements within the same observer also have the potential for affecting the determination as to whether a brain tumor is stable or has undergone progression. In the present study, we evaluated intra-observer variability in contrast-enhancing primary brain tumors.

Methods

Study population and entry criteria

Ten patients were randomly selected for inclusion in this study. Entry criteria included a previous MR scan that showed contrast-enhancing tumor, biopsy or surgery-proven diagnosis of high-grade (i.e., World Health Organization grade III or grade IV) glioma, and ability to provide informed consent. All patients underwent a second MR scan using the same imaging protocol within 12 weeks. This study was approved by the institutional review board (IRB) at our hospital. All patients provided informed consent. A waiver of informed consent was provided by our IRB to allow inclusion of observers as research subjects.

Tumor imaging characteristics

All ten tumors chosen for the analysis were solitary. Three tumors had central regions of necrosis, five tumors had resection cavities, and two tumors had neither feature. Four tumors had ill-defined borders consistent with visible local infiltration and six tumors had discrete borders.

All scans were performed on a 1.5T MR scanner (GE Healthcare, Milwaukee, WI). The following scans included contrast-enhanced T1-weighted images (TR 480 ms, TE 14 ms) with 5 mm slices and 2.5 mm interslice gap and contrast-enhanced coronal T1-weighted SPGR images (TR 12 ms, TE 5 ms) with 1.2 mm slices. In eight patients, images were reconstructed in the sagittal plane at 1.6 mm contiguous slices from the coronal SPGR data set.

Image display and measurement techniques

MR scans were anonymized using eFilm (Merge Healthcare, Milwaukee, WI). Seven observers (four residents and

three radiologists within 5–10 years of specialized training in neuroradiology) performed tumor measurements. For the first reading, observers individually viewed images on a desktop computer using eFilm software (Merge Healthcare, Milwaukee, WI) during two sessions within the same week. After a period between 6 and 12 weeks elapsed (to prevent recall bias), the observers repeated the tumor measurements.

In eight patients, the tumor was measured in all three planes; in two patients, no sagittal images were produced, giving 28 bidimensional measurements for the ten patients per observer for each reading session.

Observers were provided with a set of 20 scans in a single reading, consisting of the first and second scan for all ten patients. Observers were not told that all scans were from ten patients (i.e., in pairs). For tumor measurements, the entire data set (rather than solely selected images) were available. For any single scan, observers placed electronic calipers on images in each plane and measured tumor diameters in two orthogonal directions to obtain the greatest BP. They recorded products and slice chosen on a paper form, which was then entered into an electronic database.

Determination of change in bidimensional products on first reading

Observers measured diameters in the axial, sagittal and coronal planes and obtained the maximal BP in each plane for the first scan, which was termed BP_{ax1} for the axial plane, BP_{sag1} for the sagittal plane, and BP_{cor1} for the coronal plane (Fig. 1). They repeated the procedure for the second scan to obtain BP_{ax2} , BP_{sag2} , and BP_{cor2} . Observers were not asked to make an assessment of tumor stability or progression but merely to record tumor dimensions.

The analyst then calculated the percent difference between the maximal BP on the first scan (e.g., BP_{ax1}) and that on the second scan (e.g., BP_{ax2}) in each plane, which was designated as ΔBP_{ax} for the axial plane, ΔBP_{sag} for the sagittal plane, and ΔBP_{cor} for the coronal plane (Fig. 1).

Determination of change in bidimensional products on second reading

The observers repeated the measurements after an interval of 6–12 weeks, which was deemed a time interval sufficient to prevent recall from the first measurements. They performed the measurements in the same manner as on the first reading to obtain maximum BPs in each plane for the first scan (termed BP'_{ax1} for the axial plane, BP'_{sag1} for the sagittal plane, and BP'_{cor1} for the coronal plane and the second scan) and maximum BPs for the second scan (i.e., BP'_{ax2} , BP'_{sag2} , and BP'_{cor2} (Fig. 1).

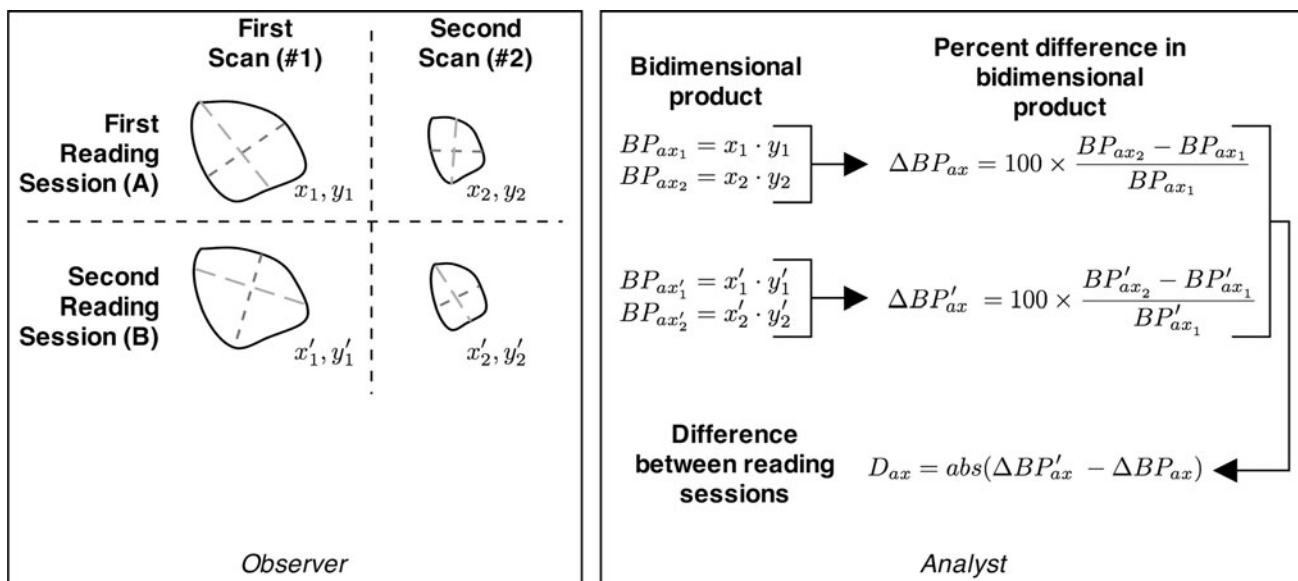


Fig. 1 Depiction of steps performed in the study. Observers measured maximum tumor diameters (*upper left*) on the initial MR scan (x_1, y_1) and follow-up MR scan (x_2, y_2) in all imaging planes. Observers repeated measurements (*lower left*) on the same scans 6–12 weeks later to obtain repeat maximum diameters on initial scans (x'_1, y'_1) and follow-up scans (x'_2, y'_2). The analyst computed percent difference between BPs for the first reading of the initial scan and the follow-up scan, to obtain ΔBP . In this diagram, solely ΔBP in the

axial plane, i.e., $\Delta BP'_{ax}$ is depicted (right, top row). However, the same procedure was performed for all three imaging planes. The analyst then performed calculations for the second reading of scans to obtain percent difference in BP for the second reading. In this example, it is performed for the axial plane, producing $\Delta BP'_{ax}$. Finally, the analysts computed the difference between reading sessions, termed D (*right, lower row*)

The analyst then calculated differences in the BPs for this second reading, which were designated as $\Delta BP'_{ax}$ for the axial plane, $\Delta BP'_{sag}$ for the sagittal plane and $\Delta BP'_{cor}$ for the coronal plane.

Differences in change in bidimensional products between readings

The analyst derived the absolute difference between the endpoints in each plane of the first reading (e.g., ΔBP_{ax}) and that on the second reading ($\Delta BP'_{ax}$). This difference we termed D , where D_{ax} is defined by the equation $D_{ax} = \text{abs}(\Delta BP'_{ax} - \Delta BP_{ax})$ and so on for each plane. This value D was the major finding of interest in this study.

Computation of percent of all readings having a D value $\geq 25\%$

For this computation, we used the total of D values in all planes as the denominator. The total number of D values was 196, based on ten axial, ten coronal and eight sagittal readings for each of seven observers recording measurements in ten tumors. We arbitrarily designated a D value $>25\%$ as of importance because the Macdonald criteria use that metric as a criterion of tumor progression.

Differences in changes in bidimensional products by imaging plane

A second finding of interest was the determination whether the rate of D values $\geq 25\%$ were the same in all three imaging planes. We recorded the number of times a D value $\geq 25\%$ was seen for each imaging plane to compute the total numbers of such measures. We also determined, for each reader, the median difference between the change in BPs between the first reading in the axial plane (termed D_{ax}), sagittal plane (D_{sag}) and coronal plane (D_{cor}). These median values for these measurements are recorded in Table 1.

Differences in changes in bidimensional products by tumor

We recorded the number of times a D value $\geq 25\%$ was seen for each tumor in order to determine whether specific tumors were associated with higher rates of D values $\geq 25\%$.

Differences in changes in bidimensional products by observer

We calculated the number of times each observer produced a $D \geq 25\%$ in any plane. In order to introduce a further

Table 1 Median D value (i.e., the percent difference between changes in tumor size on the first reading and the second reading) for each observer in each imaging plane

Observer	Median D_{ax}	Median D_{cor}	Median D_{sag}
1	17.18 (40%)	26.93 (50%)	19.44 (38%)
2	19.62 (30%)	36.19 (70%)	35.48 (63%)
3	33.64 (50%)	36.07 (70%)	25.68 (50%)
4	24.25 (40%)	38.62 (70%)	35.42 (63%)
5	16.82 (30%)	19.94 (40%)	15.12 (50%)
6	12.36 (10%)	12.18 (30%)	31.90 (63%)
7	20.56 (20%)	17.93 (30%)	19.75 (38%)
Average	20.63 (31%)	26.84 (51%)	26.11 (52%)

Percent of D values $\geq 25\%$ are shown in parentheses

measure of extent of disagreement within observers, we also measured how often D values were $\geq 50\%$.

We recorded number of D values for individual observers in each plane and also number of D values for radiology residents and practicing radiologists. We also determined whether the rate of D values $>25\%$ was substantially different for preoperative tumors as compared to tumors that had undergone partial resection.

Discordance in tumor status determinations between readings

We did not ask observers to provide a designation as to tumor status (i.e., partial response, stable or progression) according to Macdonald criteria. However, we were able to retrospectively assign tumor status based on each reading. Thus, on the first reading, if the difference in tumor size measurements was $>25\%$ in any one plane, we designated the status as progressive disease (PD). For changes in tumor size $<25\%$ or with a decrease in size of $<50\%$ in all imaging planes, the status was designated as stable disease (SD). If a decrease in tumor size exceeded 50% in any imaging plane, the tumor was designated as showing partial response (PR). If all enhancing tumor was absent on the second scan, the tumor was designated as showing complete response. Because seven observers measured ten tumors, 70 possible concordance or discordance events were possible.

We next examined the rate of discordance of progression/stable disease according to imaging plane. A total of 70 possible concordance or discordance events existed in the axial plane, 70 possible events in the coronal plane and (because sagittal measurements were solely available in eight tumors), 56 possible events in the sagittal plane.

Statistical analysis

Calculations were obtained using Microsoft Excel 2008 for Mac (Microsoft; Redmond, WA). Plots were generated

with Prism for Mac, version 5.0d (GraphPad Software; La Jolla, CA).

Results

As an estimate of the range of tumor sizes, the range of the mean BP in the axial plane on the first scan was 2.1–11.27 cm² (mean 6.8 cm²).

Differences in bidimensional products on either the first or second reading

In the axial plane, median ΔBP_1 and ΔBP_2 values ranged from -23 to 16.1% , in the coronal plane, from -29.7 to 18.1% and in the sagittal plane from -43.6 to 11.5% . Note that unlike the D values displayed later (which are absolute values), the ΔBP_1 and ΔBP_2 values can have positive or negative values depending on whether BP on the follow-up scan greater or less than the initial scan.

Differences in change in bidimensional products between readings

D values (Table 1) ranged between 12.4 and 33.6% in the axial plane (average 20.6%), 12.2 and 38.6% in the coronal plane (average 26.8%) and 15.1 and 35.5% in the sagittal plane (average 26.1%).

Differences in changes in bidimensional products by imaging plane

For all tumors, across measurements in all imaging planes, 88 of the 196 (45%) measures of D were $\geq 25\%$ (Table 2). These included 23 D values $\geq 25\%$ in the axial plane (33% of axial readings), 36 D values $\geq 25\%$ in the coronal plane (51% of coronal readings) and 29 D values $\geq 25\%$ in the sagittal plane (52% of sagittal readings). The number of D values $\geq 50\%$ were 9 (13%) in the axial plane, 17 (24%) in the coronal plane and 8 (14%) in the sagittal plane.

Median D values in each plane for each observer are shown in Table 1. The average of D values in the axial plane for all observers was 20.63% in the axial plane, 26.84% in the coronal plane, and 26.11% in the sagittal plane.

Differences in changes in bidimensional products by tumor

Figure 2 shows that in 8 tumors (i.e., all except tumors 2 and 9), more than half of observers (i.e., ≥ 4) recorded D values $\geq 25\%$ in at least one plane. In 4 tumors (i.e., tumors 4–7), more than half the observers produced

Table 2 Number of *D* values $\geq 25\%$ for each tumor in the axial (D_{ax}), coronal (D_{cor}), and sagittal (D_{sag}) imaging plane

Tumor	$D_{ax} \geq 25\%$	$D_{cor} \geq 25\%$	$D_{sag} \geq 25\%$
1	1 (1)	1 (0)	4 (0)
2	0	1 (1)	0
3	1 (0)	3 (2)	6 (2)
4	4 (2)	5 (1)	3 (2)
5	6 (2)	6 (2)	5 (2)
6	4 (1)	1 (0)	5 (1)
7	3 (2)	6 (5)	4 (1)
8	1 (1)	4 (0)	2 (0)
9	3 (0)	3 (1)	–
10	0	6 (5)	–
Total	23 (9)	36 (17)	29 (8)

Number of values $\geq 50\%$ are shown in parentheses

D values $\geq 25\%$ in two imaging planes. Tumors 1, 3, 5, 7, and 9 had resection cavities. The occurrence of *D* values $>25\%$ was very similar for those tumors as for the remaining tumors (Fig. 2). Therefore, presence or absence of a resection cavity did not substantially affect observer performance.

Differences in changes in bidimensional products by observer

Numbers of observers for whom *D* was $\geq 25\%$ in any plane are shown in Fig. 2. For all observers and for any single tumor, a *D* value $<25\%$ was produced in all imaging planes in only 24% of cases. For radiology residents, this value was attained in only 15% of tumors; for radiologists, the value was attained in 37% of cases.

Figure 3 shows the number of instances when a *D* value $>25\%$ or $\geq 50\%$ was produced in any imaging plane. A total of 28 *D* values were provided by each observer. The figure shows that the number of *D* values $<25\%$ per observer ranged from 12 (43% of observations) to 19 (68%), those >25 but $<50\%$ ranged from 6 to 9, and those $\geq 50\%$ ranged from 2 to 8.

Discordance in tumor status between readings

In 29 of the 70 events (41%), a discrepancy in tumor status was noted. These include 25 instances in which the observer’s measurements would be classified as SD on one reading and PD on the other, three instances in which measurements would be classified as SD on one reading and PR on the other, and one instance in which measurements would be classified as PR on one reading and PD on the other.

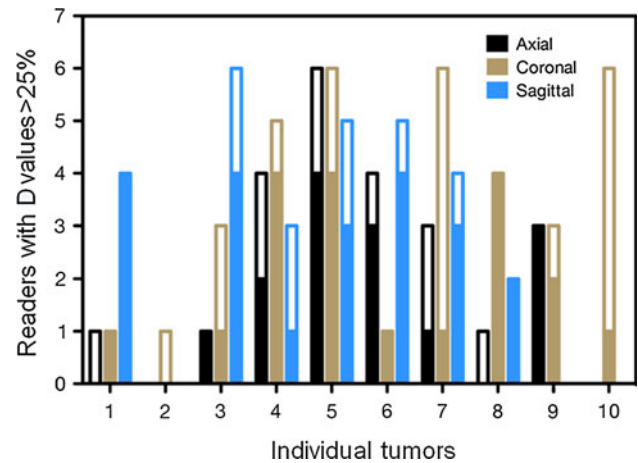


Fig. 2 Bar graph showing the number of observers for whom *D* was $>25\%$ in any imaging plane. For each tumor, one bar for each imaging plane is present. The ten individual tumors are listed on the x-axis; number of observers for whom *D* was $>25\%$ are shown on the y-axis. *D* values $\geq 25\%$ but $<50\%$ are shown in solid bars and *D* values $\geq 50\%$ are shown in open bars. In 8 of 10 tumors, more than half of the observers produced a *D* value of $\geq 25\%$ in at least one imaging plane

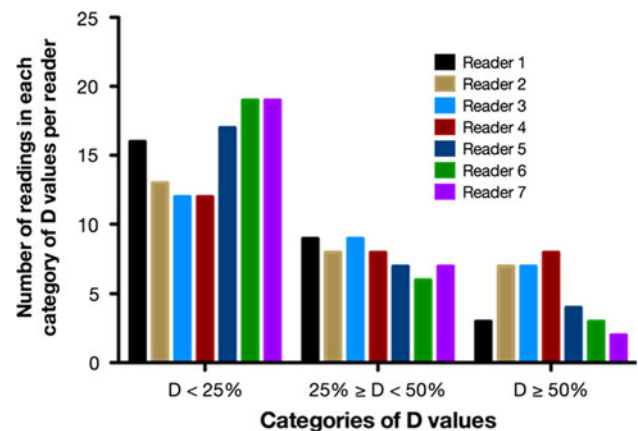


Fig. 3 Bar graph showing the number of times. a *D* was $<25\%$, b $\geq 25\%$ but $<50\%$ or c $>50\%$ for each observer ($n = 28$ for each observer). For observers 2–4, at least half of the time, *D* was $\geq 25\%$ and often greater than 50%. Number of *D* $<25\%$ per observer ranged from 12 (observers three and four) to 19 (observers six and seven), those ≥ 25 but $<50\%$ ranged from six (observer six) to nine (observers one and three), and those $\geq 50\%$ ranged from two (observer seven) to eight (observer four)

Discordant readings per imaging plane included 22 (31%) in the axial imaging plane, 32 (46%) in the coronal plane, and 24 in the sagittal plane (43%).

Discussion

Previous studies have generally have not addressed the issue of intra-observer or inter-observer variability [7, 8, 10–12].

The major finding in this study was that intra-observer variability, defined as a $\geq 25\%$ difference between measurements of change in tumor size of the same sets of MR scans, was high. In 76% of instances the difference between the change in tumor size on the two readings differed by $\geq 25\%$ in at least one imaging plane. In about one-fourth of instances, a $>25\%$ difference was encountered in two imaging planes. Even among experienced radiologists with specialty training in neuroradiology, in almost two-thirds of cases, a difference of $\geq 25\%$ between readings was found in at least one imaging plane. When taking change in tumor status into account, we found that a discrepancy between tumor status designations between readings occurred in approximately 40% of instances and that the discrepancy rate was moderately higher in the coronal and sagittal planes than in the axial plane. Thus, a high likelihood exists that a change in size of tumor that meets standard bidimensional criteria for tumor progression will be found based solely on chance. Our findings have substantial ramifications for reliability of tumor progression criteria that employ bidimensional tumor measurements, such as the RANO criteria [4].

The findings in our study substantially differed from those of one study assessing intra- and inter-observer variability in brain metastases [7]. In that study, two observers measured the single largest tumor diameter in brain metastases on thin section sagittal images obtained on 3D-SPR imaging. The intra-observer variability measurements were a mean relative difference of only 2.3% for one observer and 2.5% for the other observer. High-grade primary brain neoplasms are typically more infiltrative than brain metastases; differences in degree of tissue infiltration between these tumor types could, at least in theory, explain different findings in the two studies.

We measured intra-observer variability in the three cardinal planes in order to help determine whether any gains are made by such measurements compared to solely the axial plane. Previous studies have examined tumor measurements in these three planes [8]. For instance, in one study, the investigators developed bidimensional diameters by measuring the overall largest diameter and the tumor length in the perpendicular sagittal plane. In another study, readers were asked to measure the three largest orthogonal diameters of brain tumors [9].

In a previous study, we saw substantial inter-observer disagreement in measurements of contrast-enhancing tumor using the Macdonald criteria [6]. We found indications that the nature of the measurement method itself made it susceptible to a high rate of false-positive readings for tumor progression [6]. A large residual variance component was seen, suggesting that repeated measurements on the same image likely are variable even for the same observer, a finding that was substantiated by the results of the study presented here.

The best method for measuring tumor progression is a matter of active study. In particular, it remains unclear whether assessments based on diameter measurements or volumetric measurements are preferred relative to one another. For instance, in one study, investigators found that linear methods were very comparable to volumetric methods in that median progression-free survival and 6 months progression-free-survival did not differ significantly for the various techniques [8]. However, in another study, other investigators showed that a volume-based method generated by a technologist using a software program to draw perimeters had significantly lower intra- and inter-observer variability when compared to a method based on hand-drawn diameters [9].

Limitations

As in any study, our findings are subject to a number of limitations. First, observers were not aware that they were measuring tumors on pairs of scans. Thus, the presentation of scans did not truly simulate actual clinical practice of measurements of scans on serial studies. The unavailability of prior studies may have increased the rate of substantial disparity between readings. Second, we did not employ a homogenous population of observers and our subpopulation of practicing radiologists is rather small.

Summary

In conclusion, we found substantial intra-observer variability in measurement of tumor size in a sample of high-grade brain neoplasms. If our observers are representative of individuals measuring tumor size in clinical trials using RANO criteria, intra-observer variability may be a substantial limitation of use of these criteria.

Acknowledgments The authors thank the following individuals for participating in the study: Timothy Amrhein MD, Erica Lee Berg MD, Courtney Coursey MD, Steven Huang MD, David Ihn MD, Ronaldo Lessa MD, and Basar Sarikaya MD.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Macdonald DR, Cascino TL, Schold SC Jr, Cairncross JG (1990) Response criteria for phase II studies of supratentorial malignant glioma. *J Clin Oncol* 8:1277–1280
2. Wen PY, Norden AD, Drappatz J, Quant E (2010) Response assessment challenges in clinical trials of gliomas. *Curr Oncol Rep* 12:68–75

3. Reardon DA, Galanis E, DeGroot JF et al (2011) Clinical trial end points for high-grade glioma: the evolving landscape. *Neuro Oncol* 13:353–361
4. Wen PY, Macdonald DR, Reardon DA et al (2010) Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol* 28:1963–1972
5. Hayward RM, Patronas N, Baker EH, Vézina G, Albert PS, Warren KE (2008) Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. *J Neuro Oncol* 90:57–61
6. Provenzale JM, Ison C, DeLong D (2009) Bi-dimensional measurements in brain tumors: assessment of inter-observer variability. *AJR Am J Roentgenol* 193:W515–W522
7. Bauknecht HC, Romano VC, Rogalla P et al (2010) Intra- and interobserver variability of linear and volumetric measurements of brain metastases using contrast-enhanced magnetic resonance imaging. *Invest Radiol* 45:49–56
8. Shah GD, Kesari S, Xu R et al (2006) Comparison of linear and volumetric criteria in assessing tumor response in adult high-grade gliomas. *Neuro Oncol* 8:38–46
9. Sorensen AG, Patel S, Harmat Ch et al (2001) Comparison of diameter and perimeter methods for tumor volume calculation. *J Clin Oncol* 19:551–557
10. Weltens C, Menten J, Feron M et al (2001) Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. *Radiother Oncol* 60:49–59
11. Leunens G, Menten J, Weltens C, Verstraete J, van der Schueren E (1993) Quality assessment in medical decision making in radiation oncology: variability in target volume delineation for brain tumours. *Radiother Oncol* 29:169–175
12. Vos MJ, Uitdehaag BMJ, Barkhof F et al (2003) Interobserver variability in the radiological assessment of response to chemotherapy in glioma. *Neurology* 60:826–830