Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset

Meriem Ben Abdallah, Marie Blonski, Sophie Wantz-Mézières, Yann Gaudeau, Luc Taillandier, Jean-Marie Moureaux

Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset

Mériem Ben Abdallah[*], Marie Blonski[*†], Sophie Wantz-Mézières[‡], Yann Gaudeau[*§], Luc Taillandier[*†] and Jean-Marie Moureaux[*]

[*] Université de Lorraine, Centre de Recherche en Automatique de Nancy (CRAN), CNRS UMR 7039, Faculté de Médecine - Bât D - BP 184, Vandoeuvre-lès-Nancy, 54505, France

[†] Neuro-Oncology Unit, Nancy University Hospital, Avenue du Maréchal de Lattre de Tassigny, 54035 Nancy, France

[‡] Université de Lorraine, Institut de Mathématiques Elie Cartan, INRIA BIGS CNRS UMR 7502, BP 239, F-54506 Vandoeuvre-lès-Nancy Cedex, France

[§] Université de Strasbourg, 30 Rue du Maire André Traband, Haguenau, 67500, France

***Abstract*— Software-based manual segmentation is critical to the supervision of diffuse low-grade glioma patients and to the optimal treatment's choice. However, manual segmentation being time-consuming, it is difficult to include it in the clinical routine. An alternative to circumvent the time cost of manual segmentation could be to share the task among different practitioners, providing it can be reproduced. The goal of our work is to assess diffuse low-grade gliomas' manual segmentation's reproducibility on MRI scans, with regard to practitioners, their experience and field of expertise. A panel of 13 experts manually segmented 12 diffuse low-grade glioma clinical MRI datasets using the OSIRIX software. A statistical analysis gave promising results, as the practitioner factor, the medical specialty and the years of experience seem to have no significant impact on the average values of the tumor volume variable.**

I. INTRODUCTION

Diffuse Low-Grade Gliomas (DLGG) are rare primitive cerebral tumours of adults. These tumours progress continually over time and then turn to a higher grade of malignancy associated with neurological disability and consequentially become fatal. Tumour size is one of the most important static prognostic factors [1]. Linear Regression analysis using mixed models have reported an average increase rate of 4.1 mm per year in tumor diameter [2]. The therapeutic strategy is based on a personalized and long-term multistage approach with online adaptation over the years related to volumetric and clinical changes. Early functional surgery is usually the first therapy when possible. Chemotherapy can be used as an adjuvant treatment but sometimes, also, in a neoadjuvant position before surgical resection. Radiotherapy is usually reserved for cases of progression after chemotherapy for unresectable tumours or at the time of anaplastic transformation [3].

For patient monitoring, it is essential to apprehend the volumetric evolution under usual clinical conditions (during consultations) in order to optimally adapt the treatment in real time [4]. The simple qualitative comparison of two separate MRI examinations at 4 to 6 months intervals does not usually objectify the growth. It was originally proposed to measure the largest diameters in the 3 spatial planes, $D1$, $D2$, $D3$, and then to extrapolate the volume with the following formula: $D1 * D2 * D3/2$ [5]. A software-based manual segmentation was developed in [6] and has, since, become the standard technique for the majority of experts. This method is time-consuming, thus a massive segmentation by many different clinicians would improve the therapeutic treatment of patients. However, to our knowledge, DLGG manual segmentation's reproducibility on MRI images has not yet been assessed [7]. Indeed, the main up-to-date studies include several brain tumor types in the same study and focus rather on a comparison between automatic and manual segmentation performance. If automatic segmentation can be of great interest, we claim that in the case of DLGG, manual segmentation remains not only the ground truth but the current best way to determine the volume of such tumors for the majority of specialized teams. Indeed, automatic segmentation does not yet seem to be reliable for distinguishing tumor signal abnormalities and other causes of signal abnormalities (post-surgical or post-radiotherapy modifications, leucoencephalopathy from various aetiologies, etc.). But as manual segmentation is time-consuming, the less accurate 3 diameters method is mostly prefered for assessing the volume in daily hospital practice.

The work we propose here addresses the question of manual segmentation's reproducibility by studying the impact of the practitioner on the DLGG's volume estimation. Indeed, the latter can strongly influence the choice of a therapy and the time to start or to stop it. Such repercussions motivate the conduct of a subjective study of manual segmentation consistency among a group of DLGG experts. Such a consistency is key to the reliability and reproducibility of clinical diagnosis and, consequently, to the selection of the appropriate therapy.

The rest of the paper is organized as follows. In section 2, the methodology and the tools applied in this study are described. In section 3, the statistical techniques of evaluation are detailed. Section 4 presents the results of the statistical

study. Section 5 summarizes these results and discusses their consequences for medical practice.

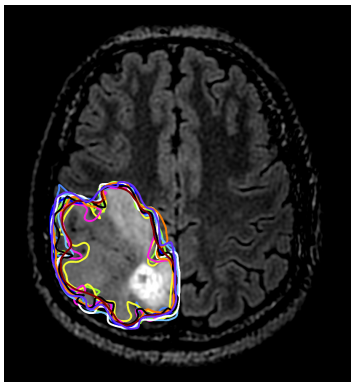

Fig. 1: Example of the manual segmentation of an MRI's slice with OSIRIX. Each colored curve corresponds to the segmentation performed by one participant.

II. MATERIALS AND METHODS

Subjective tests are often conducted to evaluate the quality of images and videos, as for example in the context of data compression studies[8], [9]. These tests take into consideration the most current medical practices and are performed in a strictly controlled environment. The results of subjective tests are quantified by objective metrics and rely on a pre-defined ground truth for their interpretation. In this study, the average of the volumes was selected as a ground truth due to the absence of an absolute ground truth. A neuroradiology expert, who was excluded from the experts panel, selected 12 longitudinal MRI scans of 9 patients diagnosed with DLGG without any previous treatment. All patients were informed and provided proof of their written agreement to participate to the present study. Moreover, the patient information was anonymized and de-identified prior to analysis. The datasets were all FLAIR-weighted axial scans except for one T2-weighted axial scan. There were 3 Cube MRI scans and 9 regular MRI scans. The reproducibility of manual segmentation study was carried out within the Living Lab PROMETEE.[1] This platform is an innovation platform allowing the study and management of videos and images' technical quality with respect to medical usage. It is well-equipped and arranged, provides a highly efficient environment to comply with the general visualization conditions for these kind of tests, as fixed by the ITU-BT.500-13 recommendation [10]. The room lighting was controlled so as not to produce reflections on the screen. The surrounding environment was all in white in order to avoid visual distractions. A panel of 14 experts performed manual segmentation on the dataset with OSIRIX as illustrated in the example of fig.1. OSIRIX is an open source Dicom Viewer software for Apple Macintosh [11] [12].The 32-bit OSIRIX version was adopted in this study. The participants started by performing a visual test on a tablet with the purpose of detecting the participants with vision problems. Then, they proceeded with a learning dataset, which was not included in the study results, so as to get familiar with the OSIRIX segmentation tool. The instruction was to manually delineate tumor contours on slices containing contrast enhancement related to a DLGG. In order to be consistent with medical practice, the radiological windowing and the number of slices to be segmented were not specified. The participants started the test by segmenting half the dataset, taking a 5-minutes break, and then completing the segmentation of the other half. At the end, they completed a questionnaire about their medical specialty and their years of experience since residency.

[1]PeRceptiOn utilisateur pour les usages du MultimÉdia dans les applications mÉdicalEs: User perception for multimedia usages in medical applications. PROMETEE is located in TELECOM Nancy engineering School, Nancy France. http://telecomnancy.univ-lorraine.fr/fr/recherche/living-lab.

TABLE I: The distribution of participants by medical specialty.

Medical specialty	Neurology	Radiology	Radiotherapy
Number of participants	6	4	3

TABLE II: The distribution of participants by years of experience.

Years of experience	$]0;10]$	$]10;+\infty[$
Number of participants	8	5

Following the first tests of consistency, it turned out that one of the 14 participants had inconsistent results. Thus, all the results described hereafter are based on the ratings made by 13 consistent participants. For the study of the variability introduced by the medical specialty on the tumor, three categories were defined : neurologists, radiotherapists and radiologists. As for the years of experience, two groups were set: $]0;10]$ and $]10;+\infty[$.The distribution of medical specialties and years of experience is listed, respectively, in Tables I and II.

After the test was complete, the manual tracings were saved and the tumor volume, for each dataset, was computed under OSIRIX based on the Delaunay triangulation reconstruction method.

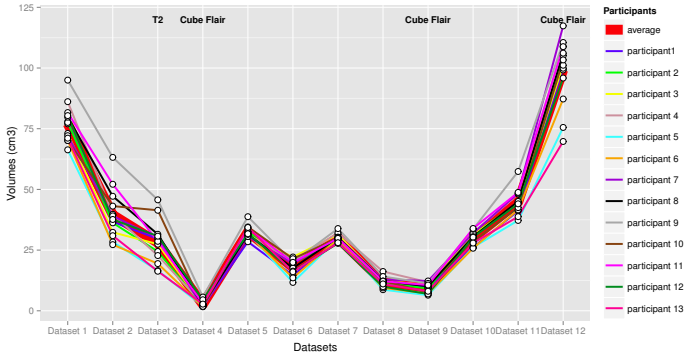

Fig. 2: Change in tumor volume based on MRI datasets and compared to the average volumes for all participants.

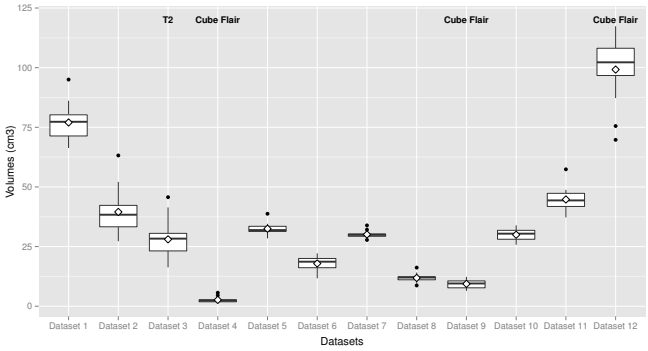

Fig. 3: Boxplot of the tumor volumes for all datasets.

III. STATISTICAL ANALYSIS

The final study dataset consists of 12 tumor volumes for each of the 13 participants: $(x_{i,j})_{i=1...13,j=1...12}$. The purpose of the statistical analysis is to investigate the variability introduced by the practitioner factor on the tumor volume variable in order to examine the influence of the practitioner on the acquired tumor volumes. Other purposes of this study include the analysis of the relationship between the participant's medical specialty as well as their years of experience and the tumor volumes. For the study of the variability introduced by the practitioner, a one-way analysis of variance (ANOVA) [13] was applied to tumor volumes.

In order to statistically quantify the variability introduced by the years of experience and the medical specialty on the tumor volumes, a standard volume, $y_{i,j}$, was first calculated as follows:

$$y_{i,j} = \left(\frac{x_{i,j} - \bar{x}_j}{\sigma_j} \right) \tag{1}$$

where $\bar{x}_j$ is the mean volume, and σ_j is the standard deviation by volume. Centering $x_{i,j}$ around the mean values of volumes for a given dataset and dividing by its standard deviation accounts for the difficulty of segmentation. The standard deviation, σ_{y_i}, of $y_{i,j}$ was then calculated and an exact Fisher test [14] [15] was applied on σ_{y_i} for both studies. In order to assess the inter-observer variability, the coefficient of variation (COV) [16] [17] by volume was used. This coefficient measures the change in volume of the segmented objects, and is defined by:

$$COV_j = \frac{\sigma_j}{\bar{x}_j} \tag{2}$$

Another metric that is used to assess the inter-participant variability is the agreement index (AI) [18]. This metric gives the inter-participants agreement , in pairs of participants, for each volume $j = 1, ..., 12$:

$$AI_{(i,i'),j} = 1 - \frac{2\left|x_{i,j} - x_{i',j}\right|}{x_{i,j} + x_{i',j}} \tag{3}$$

for all pair of participants $(i,i'); i \neq i'; i,i' \in \{1, ..., 13\}$.

AI values vary from 0 (no agreement between participants) to 1 (perfect agreement between participants).
Finally, to estimate the inter-participant variability on a pixel level, the interoperator variance (IV) [18] was applied. This metric, computed for each commonly segmented slice of each MRI exam, quantifies the overlap of two segmented regions by each pair of participants. It is defined by:

$$IV = 1 - \frac{A_{M1} \cap A_{M2}}{A_{M1} \cup A_{M2}} \tag{4}$$

A_{M1} is the segmented area by participant 1 and A_{M2} is the segmented area by participant 2. IV values vary from 0 (perfect matching of pixel values) to 1 (no matching of pixel values).

TABLE III: COV, AI and IV by medical specialty.

Medical specialty	Neurology	Radiology	Radiotherapy
COV (mean±S.D.)	17.99 ±12.44	16.56 ±10.11	14.48 ±12.32
AI (mean±S.D.)	0.74 ±0.28	0.73 ±0.27	0.74 ±0.27
IV (mean±S.D.)	0.27 ±0.07	0.3 ±0.08	0.29 ±0.09

IV. RESULTS

The statistical analysis was implemented with R software and Matlab. First of all, the volumes of the different participants for all datasets were plotted against the average of the volumes, the selected ground truth. As displayed in fig.2, the practicians have the same volume variations, even for Cube FLAIR datasets, which are supposed to be harder to segment than regular FLAIR datasets. The set of curves merges well with the curve of the ground truth. This first result is confirmed by the boxplot in fig.3, where the dispersion of tumor volumes per dataset around the mean and the median is low. This would suggest that practitioners tend to segment DLGG tumors similarly. In order to confirm this visual result, an ANOVA was performed on the dataset. With a significance level of 5%, it can be concluded that the practitioner factor has no significant influence on the average values of the volume variable.

Regarding the variability introduced by the medical specialization on the tumor volume variable, Fisher's exact test was performed with significance levels of 5%. With p-value equal to 0.604, we could not prove that the medical speciality has a significant impact on the assessed tumor volume.
Table III confirms this assertion for medical specialty. Moreover, Kolmogorov-Smirnov tests on the COV between pairs of groups (with level of significance 5%) have confirmed this assertion. Note that the tumor volumes vary from 1.67 cm^3 to 117.35 cm^3 through the different exams. So we have a huge variety of volume size. COV is obviously more sensitive to small volumes.
As for the variability generated by the years of experience on the tumor volume variable, Fisher's exact test released a p-value of 0.8961, indicating, clearly, that the years of experience could not be proved to have a significant influence on the segmented volume. This result is confirmed as well by Table IV.

V. DISCUSSION AND CONCLUSIONS

In this work, the reproducibility of manual MRI volume segmentation with regard to practitioners, their experience

TABLE IV: COV, AI and IV by years of experience.

Years of experience	$]0;10]$	$]10;+\infty[$
COV (mean±S.D.)	16.58 ±11.09	14.86 ±11.88
AI (mean±S.D.)	0.75 ±0.28	0.73 ±0.27
IV (mean±S.D.)	0.25 ±0.05	0.3 ±0.09

and their field of expertise, was assessed. This study shows that, on average, neither the practitioner nor the medical speciality or experience seem to have a significant impact on the tumor volume. The latter result is rather surprising as one would expect that experience would be a discriminatory factor.

Another surprising observation was that the largest differences in volumes were noticed on the 3 first datasets, which are supposed to be easy to delineate. Their boxplot's (see Fig. 3) spread around the median is large compared to, say, datasets 4 and 11, Cube Flair datasets, that are supposed to be more complicated to segment. This might be explained by the novelty of the used delineation tool, OSIRIX, to some participants. So, on the first datasets, the participants still hadn't mastered this tool. Across Cube Flair datasets, dataset 14 seemed to be harder to delineate than datasets 4 and 11. This might be explained by the effect of tiredness by the end of the test. But it doesn't seem to affect the overall study result.

On the basis of several commonly used criteria of the literature dedicated to inter-variability assessment, the statistical analysis achieved on this study did not prove that the medical specialty or the years of experience had any impact on the segmented tumor volumes, regardless of the dataset difficulty (Cube vs classical MRI scan). This is an encouraging result that promotes cross-disciplinary collaboration among clinicians, especially for very frequent alternating consultations between neurosurgeons and medical neuro-oncologists. And even if this result should be confirmed by additional larger studies, it opens the door to interesting perspectives in the difficult context of DLGG, where automatic segmentation does not yet seem to be able to offer a fully reliable solution. As a consequence of this finding, the manual segmentation process could be speeded up, as many clinicians would be able to delineate DLGG of different patients, even the ones they don't follow. The monitoring of tumor evolution would be improved as less time would be devoted by practitioners to manual segmentation and more time could be spent on clinical decisions regarding the appropriate treatment to prescribe at different stages of the disease.

ACKNOWLEDGMENT

This work has been supported by the Celtic-Plus European Project E3.

Great recognition and thanks are due to all the clinicians from CHRU Nancy and CHR Metz who participated in this study. Namely, we would like to thank: Pr Serge Bracard, Pr Luc Taillandier, Dr Marie Blonski, Dr Basile Wittwer, Dr Guillaume Vogin, Dr Christian Delgoffe, Dr Claire Griffaton-Taillandier, Dr Marie-Alexia Ottenin, Dr Sophie Planel, Dr Fabien Rech, Dr Valérie Bernier, Camille Dahan, Dr Emanuelle Schmitt, Dr Lavinia Jager Simon and Dr Philippe Quetin.

REFERENCES

[1] L. Capelle, D. Fontaine, E. Mandonnet, L. Taillandier, J. L. Golmard, *et al.*, "Spontaneous and therapeutic prognostic factors in adult hemispheric world health organization grade II gliomas: a series of 1097 cases: clinical article," *Journal of Neurosurgery*, vol. 118, No. 6, pp. 1157–1168, June 2013.

[2] E. Mandonnet, J.-Y. Delattre, M. L. Tanguy, K. R. Swanson, A. F. Carpentier, *et al.*, "Continuous growth of mean tumor diameter in a subset of grade II gliomas," *Annals of Neurology*, vol. 53, No. 4, pp. 524–528, April 2003.

[3] H. Duffau and L. Taillandier, "New concepts in the management of diffuse low-grade glioma: Proposal of a multistage and individualized therapeutic approach," *Neuro-Oncology*, vol. 17, No. 3, pp. 332–342, March 2015.

[4] J. Pallud, L. Taillandier, L. Capelle, D. Fontaine, M. Peyre, *et al.*, "Quantitative morphological magnetic resonance imaging follow-up of low-grade glioma: a plea for systematic measurement of growth rates," *Neurosurgery*, vol. 71, No. 3, pp. 729–739, Sep 2012.

[5] J. Pallud, M. Blonski, E. Mandonnet, E. Audureau, D. Fontaine, *et al.*, "Velocity of tumor spontaneous expansion predicts long-term outcomes for diffuse low-grade gliomas," *Neuro-Oncology*, vol. 15, No. 5, pp. 595–606, May 2013.

[6] E. Mandonnet, J. Pallud, D. Fontaine, L. Taillandier, L. Bauchet, *et al.*, "Inter- and intrapatients comparison of who grade II glioma kinetics before and after surgical resection," *Neurosurgical Review*, vol. 33, No. 1, pp. 91–96, Jan. 2010.

[7] M. C. Chamberlain, "Is the volume of low-grade glioma measurable and is it clinically relevant?" *Neuro-Oncology*, vol. 16, No. 8, pp. 1027–1028, Aug 2014.

[8] Y. Gaudeau, J. Lambert, N. Labonne, and J.-M. Moureaux, "Compressed image quality assessment: application to an interactive upper limb radioloy atlas," in *IEEE International Conference on Image Processing, ICIP 2014*, Paris, France, Oct. 2014.

[9] A. Chaabouni, Y. Gaudeau, J. Lambert, J.-M. Moureaux, and P. Gallet, "Subjective and objective quality assessment for H264 compressed medical video sequences," in *4th International Conference on Image Processing Theory, Tools and Applications, IPTA'14*, Paris, France, Oct. 2014.

[10] I.-R. Rec-BT.500, "Recommandation 500-13, methodology for the subjective assessment of the quality of television pictures," 2012.

[11] A. Rosset, L. Spadola, and O. Ratib, "Osirix: An open-source software for navigating in multidimensional DICOM images," *Journal of Digital Imaging*, vol. 17, No. 3, pp. 205–216, Sept. 2004.

[12] Osirix downloads. [Online]. Available: http://www.osirix-viewer.com/Downloads.html

[13] G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters: an introduction to Design, Data Analysis,and Model Building*. Wiley Series in Probability and Mathematical Statistics, 1978.

[14] D. G. Altman, *Practical statistics for medical research*. Chapman & Hall, 1991.

[15] R Core Team, "fisher.test: stats package," https://www.r-project.org/.

[16] C. Weltens, J. Menten, M. Feron, E. Bellon, P. Demaerel, *et al.*, "Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging," *Radiotherapy & Oncology*, vol. 60, No. 1, pp. 49–59, July 2001.

[17] M. R. Kaus, S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, *et al.*, "Automated segmentation of MR images of brain tumors," *Radiology*, vol. 218, No. 2, pp. 586–591, Feb 2001.

[18] K. Xie, J. Yang, Z. G. Zhang, and Y. M. Zhu, "Semi-automated brain tumor and edema segmentation using MRI," *European Journal of Radiology*, vol. 56, No. 1, pp. 12–19, Oct 2005.