Relevance of two manual tumour volume estimation methods for diffuse low-grade gliomas

Meriem Ben Abdallah[1,*] ✉, *Marie Blonski*[1,2], *Sophie Wantz-Mézières*[3], *Yann. Gaudeau*[1,4],
Luc Taillandier[1,2], *Jean-Marie Moureaux*[1]

[1]*Centre de Recherche en Automatique de Nancy (CRAN), Nancy, France*
[2]*Neuro-Oncology Unit, Nancy University Hospital, Nancy, France*
[3]*Institut Elie Cartan de Lorraine and INRIA BIGS, Nancy, France*
[4]*Université de Strasbourg, Strasbourg, France*
[*]*Current address: Laboratoire de recherche en Imagerie et Orthopédie (LIO), École de Technologie Supérieure, Centre de recherche du CHUM, Montréal, Canada*
✉ *E-mail: Meriem.Ben-Abdallah@etsmtl.ca*

Published in Healthcare Technology Letters; Received on 29th January 2017; Revised on 27th July 2017; Accepted on 4th October 2017

Management of diffuse low-grade glioma (DLGG) relies extensively on tumour volume estimation from MRI datasets. Two methods are currently clinically used to define this volume: the commonly used three-diameters solution and the more rarely used software-based volume reconstruction from the manual segmentations approach. The authors conducted an initial study of inter-practitioners' variability of software-based manual segmentations on DLGGs MRI datasets. A panel of 13 experts from various specialties and years of experience delineated 12 DLGGs' MRI scans. A statistical analysis on the segmented tumour volumes and pixels indicated that the individual practitioner, the years of experience and the specialty seem to have no significant impact on the segmentation of DLGGs. This is an interesting result as it had not yet been demonstrated and as it encourages cross-disciplinary collaboration. Their second study was with the three-diameters method, investigating its impact and that of the software-based volume reconstruction from manual segmentations method on tumour volume. They relied on the same dataset and on a participant from the first study. They compared the average of tumour volumes acquired by software reconstruction from manual segmentations method with tumour volumes obtained with the three-diameters method. The authors found that there is no statistically significant difference between the volumes estimated with the two approaches. These results correspond to non-operated and easily delineable DLGGs and are particularly interesting for time-consuming CUBE MRIs. Nonetheless, the three-diameters method has limitations in estimating tumour volumes for resected DLGGs, for which case the software-based manual segmentation method becomes more appropriate.

1. Introduction: Diffuse low-grade glioma (DLGG) is a rare primitive cerebral tumour of adults. In [1], Mandonnet *et al.* showed that tumour diameter proves to be a good predictor of the evolution of DLGGs. Consequently, nowadays, patients' monitoring in specialised clinical centres relies heavily on a longitudinal supervision of tumours' diameter evolution. This monitoring uses two consecutive, at least 3 months spaced, axial MRI sequences that are either T2 weighted or, more commonly, FLAIR weighted. DLGG's volume is estimated from these MRI sequences either through a segmentation followed by a software reconstruction or through the three diameters method. Whereas the three diameters method uses a simple formula to compute tumour volume, software solutions [1], make it possible to reconstruct tumour volume from the manual segmentations of the practitioner on the MRI slices where the tumour lesion appears. Moreover, the three-diameters method is fast compared to the software-based volume reconstruction solution. Therefore, an automatic segmentation algorithm could make the segmentation task time-efficient and improve the therapeutic management of DLGG patients.

The medical imaging community has been aware of the importance of segmentation algorithms and has been organising MICCAI (International Conference on Medical Computing and Computer Assisted Intervention) conference challenges since 2007. These challenges include the brain tumour image segmentation challenge [2, 3], which focuses on brain tumours and which has enabled different research teams to evaluate the performance of their automatic segmentation algorithms [4–7]. Several solutions were proposed for the segmentation of brain tumours, including support vector machines [4], the level set method [8], the k-nearest neighbour algorithm [9] and, more recently, deep learning approaches using convolutional neural networks [5–7]. However, the major automatic segmentation algorithms proposed so far are more generalised for different brain tumours [6–9], and hence, neglect particular segmentation difficulties that are specific to DLGG (ill-defined boundaries, heterogeneity of the tumour). Moreover, manual segmentation is still the ground truth in automatic segmentation studies of brain tumours and no algorithm has yet been proved capable of replacing human expertise in a clinical routine. Therefore, at present, software-based manual segmentation should be preferred for follow-up in the treatment of DLGG. To allow quick and timely therapeutical decisions, this procedure needs to be performed by different clinicians. However, for a distribution of segmentation tasks among various independent practitioners, reproducibility and inter-practitioner inconsistency in segmentation is a major challenge. To our knowledge, no studies have been conducted to assess the reproducibility of DLGG's software-based manual segmentation on MRI datasets [10]. A first purpose of this work is to address this topic by conducting a subjective study on the impact of the practitioner factor on segmented tumour volume. We also investigate the influence of the specialty and of the years of experience on the obtained tumour volume.

Furthermore, the use of the three-diameters method remains fairly widespread in clinical practice for DLGG volume estimation. Nevertheless, this approach offers an ellipsoidal approximation of the tumour volume [1, 11, 12], which could be assumed to be less precise than the software-based volume reconstruction solution. In fact, the three-diameters method consists in defining the two largest diameters in the axial plane of an MRI exam at a given date and in drawing the largest diameter in the sagittal or in the

doi: 10.1049/htl.2017.0013

coronal plane of an MRI exam carried out on the same date as the axial MRI scan. However, its use is complicated and, sometimes, almost impossible following surgery or radiotherapy treatments because it becomes difficult to define the diameters to be selected (whether or not to include the post-treatment residuals and ill-defined tumour boundaries after treatment). It is also complicated to apply in the case of highly infiltrative DLGG. In spite of these limitations, it was important to study this method of volume estimation because of its wide use within the medical community. To our knowledge, there has been no formal study comparing the results obtained by the three-diameters method with those acquired with the software-based volume reconstruction solution based on manual segmentations. A second purpose of this work is to address this topic by conducting a subjective study of the impact of the volume estimation method on tumour volume estimate. In the conclusion, we will present a series of recommendations, that are based on our results, regarding the relevance of the two tumour volume estimation methods.

It should be noted that part of this work has been published in the EMBC's international conference proceedings [13].

2. Subjective manual segmentation reproducibility's study: materials and methods: The subjective manual segmentation study of reproducibility was conducted in PROMETEE (http://prometee.telecomnancy.eu.), the healthcare Living Lab in TELECOM Nancy, France. PROMETEE provides a standard environment that complies with the ITU-BT.500-13 recommendations for subjective tests to evaluate the quality of medical images and videos. The 32-bit free version of OsiriX software was adopted for the segmentation, as OsiriX is one of the best medical imaging softwares including segmentation tools and it is widely used among the neuro-oncology community. An expert neuroradiologist, who does not belong to the study panel, selected 12 longitudinal MRI scans in the axial plane from 9 DLGG patients. The MRI images in the study had 512 × 512 pixels in the axial plane and a number of slices ranging between 29 and 512. All exams were FLAIR-weighted but for one T2-weighted MRI exam. Moreover, there were three cube MRI exams and nine regular exams. CUBE MRI sequences are recent GE MRI exams that replace the conventional slice by slice, plane by plane, 2D MRI acquisition of a volume, enabling a single 3D volume scan. The CUBE MRI volume is isotropic, allowing a reconstruction with a similar resolution to that of the native plane. Moreover, sub-millimetre and ultra-thin slices help to better visualise the details of the lesions in the images. Raw, unformatted CUBE MRI dataset is longer to process, as in the case of our test dataset, where the number of slices for these sequences ranged between 256 and 512, with tumoural lesions present in around 100 slices sometimes. Nonetheless, it was important to integrate CUBE MRI exams to compare the results on these new sequences with those of conventional MRI sequences.

A panel of 14 experts was selected to perform the reproducibility test. This panel included six neurologists (neurologists include neuro-oncologists and neurosurgeons.), four radiologists and three radiotherapists. Moreover, eight participants had <10 years of experience, whereas five participants had >10 years of experience. In order to be consistent with daily clinical practice, there was no specification on the radiological windowing and on the slices to be segmented. The only provided instruction was to delineate DLGG for each slice containing this tumour. The participants started by completing a visual test on a tablet to detect participants with vision problems. Then, they performed a segmentation on a training dataset whose results were excluded from the final study results. They went on delineating half the exams, taking a 5 min. break and then completing the delineation of the rest of the exams. At the end, they provided information about their specialty and about their years of experience since residency. Afterwards, for each MRI exam, we saved the manual tracings and we reconstructed tumour volumes using OsiriX based on the Delaunay triangulation reconstruction method. An example of the manual segmentation of an MRI's slice during the test is displayed in Fig. 1. Each coloured curve corresponds to the segmentation performed by each participant.

The first tests of consistency showed the incoherence of one participant's results. Thus, the following statistical results are based on the segmentations of the 13 consistent participants.

3. Statistical analysis tools: Let $(x_{i,j})_{i=1\ldots13,j=1\ldots12}$ be the variable corresponding to the 12-tumour volumes for each of the 13 participants. The first aim of this study is to assess the variability introduced by the practitioner factor on tumour volumes. For this purpose, a one-way analysis of variance (ANOVA) is applied on tumour volume $x_{i,j}$. The second aim of this study is to analyse the relationship between the participants' medical specialty as well as their years of experience and tumour volumes. Several objective metrics are applied to achieve this. Among these metrics, the coefficient of variation (COV) [9] quantifies the change in the segmented tumour volumes:

$$\mathrm{COV}_j = \frac{\sigma_j}{\bar{x}_j} \tag{1}$$

$\bar{x}_j$ is the mean volume and σ_j is the standard deviation by volume.

We also use the agreement index (AI) [8] which provides, for each volume, the inter-participants agreement, in pairs of participants:

$$\mathrm{AI}_{(i,i'),j} = 1 - \frac{2|x_{i,j} - x_{i',j}|}{x_{i,j} + x_{i',j}} \tag{2}$$

for all pair of participants (i, i'); $i \neq i'$; $i, i' \in \{1, \ldots, 13\}$. AI values are upper bounded by 1 (perfect agreement between participants).

In order to estimate the inter-participant variability on a pixel level, the interoperator variance (IV) [8] is applied. This metric, computed for each commonly segmented slice of each exam, measures the overlap of two segmented regions by each pair of participants. It is defined by

$$\mathrm{IV} = 1 - \frac{A_{M_i} \cap A_{M_{i'}}}{A_{M_i} \cup A_{M_{i'}}} \tag{3}$$

A_{M_i} is the segmented area by participant i and $A_{M_{i'}}$ is the segmented area by participant i'. IV values vary from 0 (perfect matching of pixel values) to 1 (no matching of pixel values).

Finally, the results of the metrics above are confirmed by a Fisher's exact test which is a classical statistical test to use on a

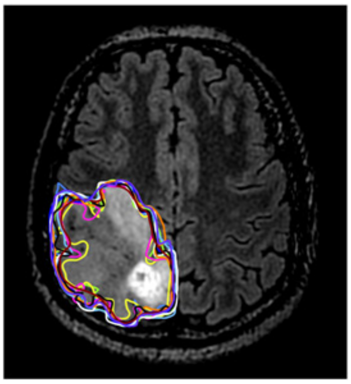

Fig. 1 *Example of the manual segmentation of an MRI's slice with OsiriX. Each coloured curve corresponds to the segmentation performed by each participant*

small sample of data as in our case [14]. We applied Fisher's exact test on the standard deviation σ_{y_i} of the standard volume $y_{i,j}$, which is used instead of $x_{i,j}$ to account for the difficulty of segmentation. The standard volume $y_{i,j}$ is computed as follows:

$$y_{i,j} = \left(\frac{x_{i,j} - \overline{x}_j}{\sigma_j} \right) \tag{4}$$

4. Statistical analysis results: An ANOVA was performed on the segmented tumour volumes and, with a significance level of 5%, we concluded that the practitioner factor has no significant impact on the average values of the volume variable.

Regarding the variability introduced by the medical specialisation and the years of experience on the tumour volume variable, we computed the mean and standard deviation of the COV, AI and IV metrics for the different categories of medical specialty and of years of experience. The results are detailed in Tables 1 and 2. We can see that the different values are quite close. This was confirmed with a Kolmogorov–Smirnov test which is a classical statistical test to use on a small sample of data as in our case [15]. We applied the Kolmogorov–Smirnov test on the COV metric between pairs of groups (with a significance level of 5%).

It should be noted that tumour volume values are between 1.67 cm^3 and 117.35 cm^3 along the different MRI exams. This large variation in volume size makes COV values more sensitive to small volumes.

These results are further confirmed thanks to Fisher's exact test. With a p-value equal to 0.604 for a significance level of 5%, Fisher's exact test could not prove that the medical specialisation has a significant impact on the tumour volume estimation. As for the variability generated by the years of experience on the tumour volume variable, Fisher's exact test released a p-value of 0.8961, indicating, clearly, that the number of years of experience could not be shown to have a significant influence on the segmented volume.

5. Comparison between the software-based manual segmentation method and the three-diameters method: materials and methods: We conducted a test in the Living Lab PROMETEE to compare the software-based manual segmentation method and the three-diameters method. For this study, we used the 12 MRI scans that we described in the previous section. In order to compute tumour volume with the three-diameters method, we also included for each MRI exam in the axial plane, the MRI scan in the coronal or in the sagittal plane. There were 11 exams in the sagittal plane and one exam in the coronal plane.

For this test, we selected a participant from our previous study on the reproducibility of the manual segmentation of DLGG. This participant had tumour volume results in the previous study which were close to the average of the tumour volumes segmented by all the participants. This allows us to fairly compare tumour volumes obtained through manual segmentation and through the three-diameter method for DLGG dataset. The instruction given to the participant consisted in drawing, for each MRI dataset, the three largest diameters in the axial and in the sagittal/coronal planes. Similarly to our previous study, there was no specification on the radiological windowing and on the slices to be outlined. Fig. 2 shows an example of an MRI dataset in the study. The figure shows the three largest diameters as defined in the axial plane and in the sagittal plane, respectively.

Once the participant had finished drawing the three diameters for all the datasets, we saved the ROIs and then we calculated tumour volume V using the approximative formula:

$$V = D1 \times D2 \times D3/2$$

where $D1$, $D2$ and $D3$ are the three largest diameters in the three spatial planes [1, 11, 12].

Table 1 Mean and standard deviation of the COV, AI and IV by medical specialty

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

Table 2 Mean and standard deviation of the COV, AI and IV by years of experience

Years of experience	]0; 10]	]10; + ∞[
COV (mean ± SD)	16.58 ± 11.09	14.86 ± 11.88
AI (mean ± SD)	0.75 ± 0.28	0.73 ± 0.27
IV (mean ± SD)	0.25 ± 0.05	0.3 ± 0.09

6. Experimental results and discussion: Tumour volume results are detailed in Table 3. This table presents the values of tumour volumes obtained with the three-diameters method ($3d$) as well as the average of the volumes acquired by software reconstruction from manual segmentations (Av) and the volumes obtained from the manual segmentations of this study's participant (Vp). We can already notice that tumour volumes are very close. Table 3 also reports the max difference between $3d$ and volumes obtained from manual segmentations. This difference does not seem to be affected by the type of sequence as it is least important for exam 4, a CUBE FLAIR MRI, and most important for exam 12, another CUBE FLAIR MRI.

We started by plotting tumour volume variation curves with the two methods of volume estimation for all datasets. For the software-based manual segmentation method, we included the results of our participant in the previously described test in Fig. 3 (green) as well as the average of the volumes obtained by all the participants (blue). We can thus see that the results of the participant in this study are very close to those of the average of all the participants. Moreover, as displayed in Fig. 3, the values of the volumes with the three diameters method (red) and with the software-based manual segmentation method are close, especially for MRI datasets from 8 to 11.

In addition, results for CUBE FLAIR-weighted MRI scans (4, 9 and 12) do not appear to be affected by the choice of the volume estimation method. Indeed, two out of three CUBE FLAIR-weighted MRI datasets show very close results with both methods.

In order to objectively assess the results, we applied the Wilcoxon signed-rank test on all tumour volumes obtained by the

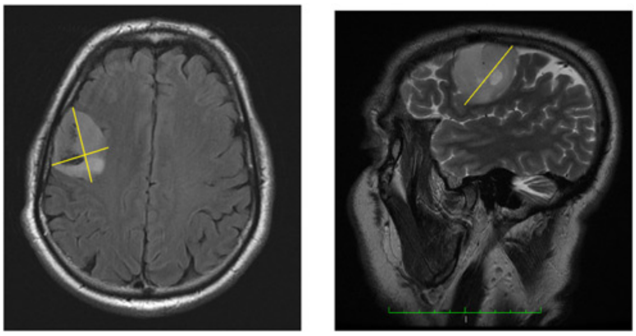

Fig. 2 *Example of the two largest diameters as defined on an axial FLAIR-weighted MRI (left) and of the third largest diameter as defined on a sagittal T2-weighted MRI scan (right) in the study with the three-diameters method*

Table 3 Volumes obtained with the three-diameters method 3*d* and with volume reconstruction from manual segmentations *Vp* and *Av* and max difference between 3*d* and volumes from manual segmentations

Exam number	1	2	3	4	5	6	7	8	9	10	11	12
3*d*, cm^3	69.31	34.03	20.94	2.60	41.24	14.34	33.13	12.43	8.95	32.22	45.39	111.43
Vp, cm^3	77.59	36.22	24.42	2.15	32.62	19.58	29.23	11.48	9.18	31.40	44.35	105.08
Av, cm^3	77.02	39.51	28.04	2.64	32.48	17.91	30.07	11.85	9.36	29.95	44.74	99.19
max differences, cm^3	25.70	29.15	24.76	3.05	12.80	7.80	5.34	3.74	3.32	6.44	11.99	41.67

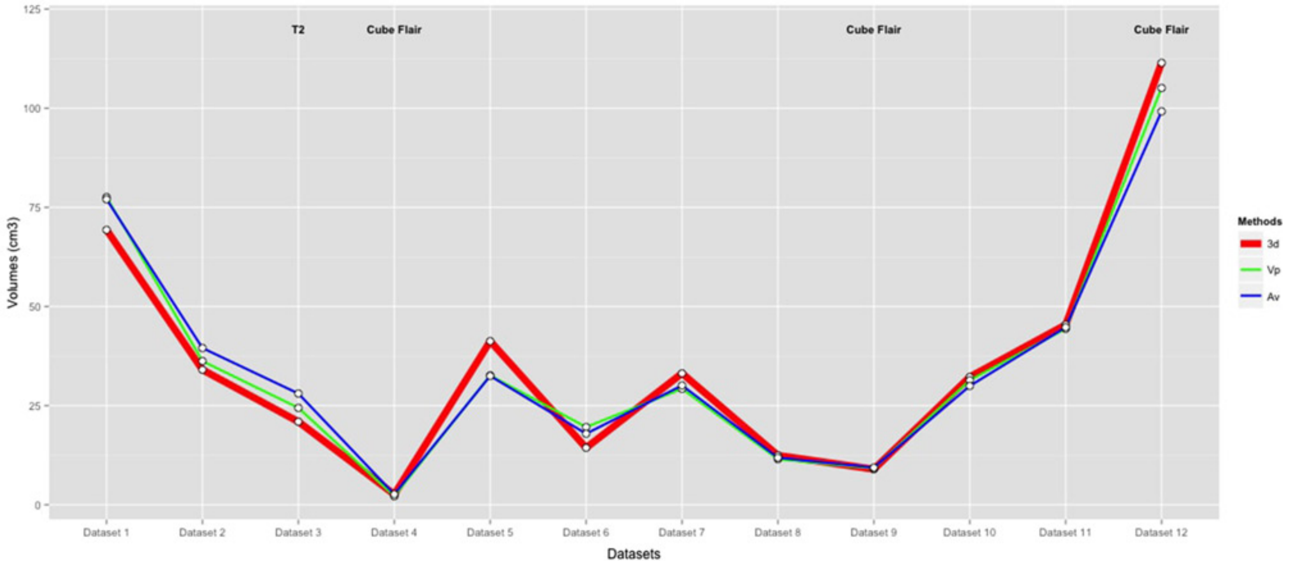

Fig. 3 *Change in tumour volume based on MRI datasets with the three-diameters method (red) and with a software-based manual segmentation method for the average of the volumes obtained by all the participants (blue) and for the results of our participant in the previously described test (green)*

two volume estimation methods (3*d* and *Av*) with a significance level of 5%. This test released a *p*-value equal to 0.9097. Consequently, we do not reject the null hypothesis considering that the volume distributions obtained by the two methods are the same.

Consequently, DLGG tumour volumes estimated with the three-diameters method are quite similar to those determined with a software-based manual segmentation approach. It should be noted that the MRI datasets represent relatively simple DLGG cases, with easily delineable tumour contours. Another test which will include more complex cases (post-surgery or post-radiotherapy MRI scans, highly infiltrating DLGG that are difficult to delineate) is in perspective for this work.

7. Conclusion: We evaluated the reproducibility of DLGG manual segmentation on MRI datasets with regard to practitioners, their years of experience and their specialty. Based on several commonly used criteria in the literature dedicated to inter-variability assessment, we could not prove that the practitioner factor, the medical speciality factor or the years of experience factor had a significant impact on the estimate of tumour volume, regardless of the type of MRI sequence (cube sequence, versus classical sequence). As automatic segmentation algorithms do not yet offer a reliable solution for DLGG, our study confirms the inter-observer reproducibility of manual contouring. This is absolutely essential considering that management of this type of tumour is necessarily multidisciplinary (including, among other practitioners, neurosurgeons, neuro-oncologists, radiotherapists and neuro-radiologists) and that, to date, this point had not yet been demonstrated. For the future, we continue this project by working on semi-automatic algorithms which, in case of correlation with the manual techniques, would save time for clinicians. Moreover, they would make patients' monitoring and, thus, therapeutic evaluations and decisions more reliable.

Furthermore, we conducted a study to compare two tumour volumes' estimation methods in the case of DLGG, namely the three diameters method and the software-based manual segmentation method. The initial results presented on 12 DLGG MRI datasets reveal that there is no statistically significant difference between the volumes estimated with the two approaches. We can therefore use the three diameters method for non-operated patients in the case of cube FLAIR MRI scans which are time-consuming to segment manually. If we consider the case of the example in Fig. 4, which shows an FLAIR-weighted MRI slice in the axial plane of a diffuse low-grade glioma after a surgery, we can already see the limits of the three-diameters method. Indeed, in order to estimate the tumour volume using the three-diameters method, we need to draw the largest diameter in each of the three planes in space. If we limit ourselves to the two drawings on the axial plane in Fig. 4, we quickly realise that we are facing difficult choices: Shall the post-operative cavity be included in the drawing or not? Which tumour residual should be included in the tracing knowing that it is impossible to join all the residuals together with a segment?

This example showcases the limitations of the three diameters method in estimating tumours' volumes after surgery. Since most patients diagnosed with DLGG are operated on some point during their follow-up, this method cannot answer the problem of calculating the volume for similar cases. The technique based on manual contouring becomes, therefore, more appropriate. As DLGG in this study represent simple cases that are easy to delimit, another study with more complex examples is in perspective, particularly for post-surgery cases. Our ultimate aim is to assist clinicians in defining DLGG volumes as accurately as possible.

doi: 10.1049/htl.2017.0013

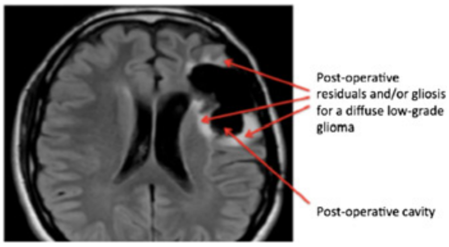
Post-operative residuals and/or gliosis for a diffuse low-grade glioma
Post-operative cavity

Fig. 4 *Example of an FLAIR-weighted MRI slice in the axial plane of a diffuse low-grade glioma after a surgery*

Finally, we aim to compare these results to semi-automatic segmentations to propose adaptive medical recommendations. Selecting the most appropriate volume estimation method is key to achieve this, enabling a better follow-up of DLGG patients. This work is the starting point for proposing recommendations on which volume estimation method should be used for volume calculation in the case of DLGG. This Letter details our progress towards this goal.

8. Acknowledgment: Great recognition and thanks are due to all the clinicians from CHRU Nancy and CHR Metz who participated in this study. Namely, we 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.

9. Funding and Declaration of Interests: None declared.

10 References

[1] Mandonnet E., Pallud J., Clatz O., ET AL.: 'Computational modeling of the WHO grade II glioma dynamics: principles and applications to management paradigm', *Neurosurg. Rev.*, 2008, **31**, (3), pp. 263–269

[2] MICCAI: 'Miccai-BRATS 2016', 2017, http://braintumorsegmentation.org/

[3] Menze B.H., Jakab A., Bauer S., ET AL.: 'The multimodal brain tumor image segmentation benchmark (BRATS)', *IEEE Trans. Med. Imaging*, 2015, **34**, (10), pp. 1993–2024

[4] Bauer S., Nolte L.-P., Reyes M.: 'Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization'. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, 2011, pp. 354–361

[5] Zikic D., Ioannou Y., Brown M., ET AL.: 'Segmentation of brain tumor tissues with convolutional neural networks'. Proc. MICCAI-BRATS, 2014, pp. 36–39

[6] Havaei M., Dutil F., Pal C., ET AL.: 'A convolutional neural network approach to brain tumor segmentation'. Int. Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 2015, pp. 195–208

[7] Kamnitsas K., Ferrante E., Parisot S., ET AL.: 'DeepMedic for brain tumor segmentation'. Int. Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 2016, pp. 138–149

[8] Xie K., Yang J., Zhang Z.G., ET AL.: 'Semi-automated brain tumor and edema segmentation using MRI', *Eur. J. Radiol.*, 2005, **56**, (1), pp. 12–19

[9] Kaus M.R., Warfield S.K., Nabavi A., ET AL.: 'Automated segmentation of MR images of brain tumor', *Radiology*, 2001, **218**, (2), pp. 586–591

[10] Chamberlain M.C.: 'Is the volume of low-grade glioma measurable and is it clinically relevant?', *Neuro-Oncology*, 2014, **16**, (8), pp. 1027–1028

[11] Pallud J., Blonski M., Mandonnet E., ET AL.: 'Velocity of tumor spontaneous expansion predicts long-term outcomes for diffuse low-grade gliomas', *Neuro-Oncology*, 2013, **15**, (5), pp. 595–606

[12] Sorensen A.G., Patel S., Harmath C., ET AL.: 'Comparison of diameter and perimeter methods for tumor volume calculation', *J. Clin. Oncol.*, 2001, **19**, (2), pp. 551–557

[13] Ben Abdallah M., Blonski M., Wantz-Mézières S., ET AL.: 'Statistical evaluation of manual segmentation of a diffuse low-grade glioma MRI dataset'. 38th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society, Orlando, FL, 2016, pp. 4403–4406, doi: 10.1109/EMBC.2016.7591703

[14] Altman D.G.: 'Practical statistics for medical research' (CRC Press, Boca Raton, FL, USA, 1990)

[15] Pett M.A.: 'Nonparametric statistics for health care research: statistics for small samples and unusual distributions' (Sage Publications, 2015)