MR WOUTER LEENDERT LODDER (Orcid ID : 0000-0002-4364-3853)

PROFESSOR BERNARD F.A.M. VAN DER LAAN (Orcid ID : 0000-0002-5016-2871)

Semi-automatic tumour volume measurements on MR-imaging using Smartbrush® in oropharyngeal carcinomas; our experience in 5 patients.

Lodder WL[1,2]*, Dorgelo B[3], Lancaster J[4], Hanlon R[5], Wieshmann H[5], de Bock GH[6], van der Laan BFAM[1]

1) University of Groningen, Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, Groningen, the Netherlands.

2) Medical Centre Leeuwarden, Department of Otorhinolaryngology/Head and Neck Surgery, Leeuwarden, The Netherlands.

3) University of Groningen, Department of Radiology, University Medical Center Groningen, Groningen, the Netherlands.

4) Aintree University Hospital, Department of Otorhinolaryngology/Head and Neck Surgery, Liverpool, England.

5) Aintree University Hospital, Department of Radiology, Liverpool, England.

6) University of Groningen, Department of Epidemiology, University Medical Center Groningen, Groningen, the Netherlands.

Conflict of interest: None.

*CORRESPONDING AUTHOR

Wouter Lodder, MD, PhD

Medical Centre Leeuwarden, Department of Otorhinolaryngology/Head and Neck Surgery, Leeuwarden, The Netherlands.

Henri Dunantweg 2, 8934 AD Leeuwarden, the Netherlands

Tel: +31582866200

Fax: +31582863331

E-mail:wouter.lodder@znb.nl

SHORT TITLE: Uncontrolled case serie of observer variation MRI tumour volume measurements.

KEY WORDS: imaging, tumour volume, observer variation, semi-automatic measurement, MRI, head and neck cancer, Smartbrush.

KEY POINTS

- The authors describe the results of their initial experience in attempting to evaluate (Smartbrush®), a semi-automated computer programme developed by Brainlab Munich, to estimate tumour volumes from standard MRI images of oropharyngeal malignant tumours.

- Interobserver variability measured with Intraclass Coefficient Correlation was 0.98 (95% confidence interval 0.80 – 0.998, SD 34.16). The standard error of the measurement was 5.07 cm^3.

- Intraobserver variability measured with Intraclass Coefficient Correlation for the first observer was 0.88 (95% confidence interval -0.143 – 0.998, SD 27.45). The standard error of the measurement was 9.51 cm^3. For the second observer Intraclass Coefficient Correlation was 0.984 (95% confidence interval 0.842 – 0.998, SD 34.19). The standard error of the measurement was 4.32 cm^3.

- At a first glance the ICC seemed reasonable high, unfortunately these measured volumes did not represent a true or similar tumour outline and thus true volume.

INTRODUCTION

Head and neck cancer is worldwide the sixth most common cancer.[1] Prognosis is linked to tumour (primary tumours, nodal stage and distant metastasis) and general patients' characteristics. Currently, the TNM-staging system is the most used staging system in the world.[2] Beside this TNM-staging system, other prognostic factors are used to select the best treatment for each individual patient. The TNM-staging system uses only 2-dimensional tumour sizes. Therefore, in more advanced tumours selection of the optimal treatment (such as surgery, chemotherapy or radiation) for the individual patient can be challenging in more advanced tumours.

In hypopharyngeal and laryngeal cancer, primary tumour volume emerged as an independent significant factor for predicting local control and overall survival.[3,4] These primary tumour volumes are often determined with help of computed tomography or magnetic resonance imaging.[3-6]

Mukherji et al, have shown (already in 2005), that the interobserver reliability of manual delineated tumour volumes is excellent with an intraclass correlation coefficient of 0.81.[6] Until today however tumour volumes are not regularly determined, probably as the current methods are time-consuming. It seems therefore preferable to develop an (semi-)automated reproducible method for the tumour volume measurements of both the primary and nodal tumour volume.[3]

In 2013 Lodder et al published the results of semi-automated tumour volume measurements on dynamic contrast enhanced images.[7] Although preliminary results were promising it was not possible to reproduce the measurements in another hospital.

New software (Smartbrush®) developed by Brainlab Munich shows a reasonable agreement between semi-automated and manually-derived tumour volumes in maxillary sinus tumours.[8] The lead authors' intention is to evaluate the programme in other head & neck tumour sites and the aim of this study is to determine inter- and intraobserver variability for Smartbrush® software on MR-images for patients with oropharyngeal carcinomas.

METHODS

Ethical considerations

As patient data were coded, patient consent was not required for the retrospective review of records and images. Protection of a patient's identity was guaranteed by assigning study specific unique patient numbers. The investigator of this study had only access to this coded database including clinical data and images. Therefore, according to the Dutch law no further Institutional Review Board approval was needed.

Participants

Only patients, treated in University Medical Center Groningen, the Netherlands, with primary untreated squamous cell carcinoma were selected. Patients with pre-treatment MR-images were selected. Only five patients (mean age 63 years, range 52-69) were included. Further patient demographics can be found in Table 1.

Imaging protocol

Available imaging

The MR-images were acquired equivalent to the clinical standard acquisition protocol on a 1.5 Tesla system (MAGNETOM Area, Siemens, Erlangen, Germany). All the measurements were performed on T2-weighted sequences.

MR-Technique

The measurements with Smartbrush can be performed on the currently standard MR-images.

Measurement tumour volumes

Technique for semi-automated derived gross tumour volume

Automatic segmentations were performed using Smartbrush® using MR-images. According to Medical Device Directive, Council Directive 93/42/EEC, Smartbrush 2.0 is a Class IIb product.

The Smartbrush® software works by semi-automatically measures tumour volumes. After selection of a region-of-interest by the user in at least 2 orthogonal directions, Smartbrush® segments voxels (3D pixels) with similar signal intensities. To achieve this, the observer selects a region-of-interest within the bulk tumour in 2 orthogonal directions (axial and coronal or sagittal view) to create a 3D-image. This volume is expended using a voxel-based algorithm (i.e. it selects an area of equivalent adjacent signal intensity). Subsequently, the observer can briefly evaluate the automatically outlined volume in 9 slices selected from the most cranial and caudal slice with even intervals. When the outline is satisfactory to the user, a tumour volume will be calculated and described in a volumetric report, which can also be uploaded in DICOM.

Measurements with Smartbrush® were performed by 3 observers (WL, BD & RH), an otorhinolaryngology specialist, and 2 experienced head & neck radiologists working in a high volume UK head & Neck unit and in the University Medical Centre Groningen. Two observers (WL and BD) evaluated the MR-images twice. The observers were blinded to the results of the volume measurements from the other observer, and for their own earlier measurement.

Observer analysis

Inter and intra observer variation was studied using Intraclass Coefficient Correlations (ICC).

Statistical analysis

Power analysis revealed that 25 scans would be necessary to study observer variation between three observers. This will achieve a power of 80%, with an alpha of 0.05, using Two-Way random ICC (proving an ICC of 0.8 compared to the zero-hypothesis with ICC of 0.6). Following the interpretations of Fleiss[9] a value of 0.50-0.74 will represent a good agreement and 0.74-1.00 will represent an excellent correlation (clinical measures). Additional the standard error of the measurement (SEM) was measured (SEM = SD x $(\sqrt{1 - ICC})$). The standard error of measurement estimates how repeated measures of a person on the same instrument tend to be distributed around his or her "true" score, 68% of the measurements will be ± SEM around the measured value.

RESULTS

Instead of the intended 25 patients, only 5 patients were used for the calculations of the interobserver variability due to the difficulties experienced and the non-accurate tumour delineations within these cases.

Interobserver variability

Despite these difficulties, the performed measurements of the tumour volumes (in cm^3) ranged between 5.85 and 47.70 (mean 23.54, SD 18.34) for the first observer, 8.11 and 46.70 (mean 29.10, SD 17.8) for the second observer and ranged between 6.44 and 46.10 (mean 24.25, SD 15.93) for the third observer showing very close correlation and confirming that to the experienced eye, the radiological tumour boundaries were unequivocal.

Interobserver variability measured with Intraclass Coefficient Correlation (ICC) was 0.98 (95% confidence interval 0.873 – 0.997, SD 50.85). The standard error of the measurement was 8.04 cm^3.

Intraobserver variability

The measurement of intraobserver variability was performed on the same 5 patients, by two observers.

For the first observer these tumour volumes (in cm^3) ranged between 6.44 and 46.10 (mean 24.25, SD 15.93) for the first session and ranged between 5.67 and 38.80 (mean 20.47, SD 13.0) for the second session. Intraobserver variability measured with ICC was 0.88 (95% confidence interval -0.143– 0.998, SD 27.45). The standard error of the measurement was 9.51 cm^3.

For the second observer these tumour volumes (in cm^3) ranged between 8.11 and 46.70 (mean 29.8, SD 17.76) for the first session and ranged between 7.01 and 41.60 (mean 25.66, SD 16.70) for the second session. Intraobserver variability measured with ICC was 0.984 (95% confidence interval 0.842– 0.998, SD 34.19). The standard error of the measurement was 4.32 cm^3.

Figure 1 shows the distribution of tumour volumes for both the interobserver and intraobserver variation. An example of the semi-automated tumour volume measurement with Smartbrush is shown in Figure 2. Figure 3 shows an example of an exact similar tumour volume result, however the region-of-interest is different for the majority of the lesions.

DISCUSSION

Synopsis of key/new findings

The preliminary results presented reveal both technical and programme-related problems using the software program in the oropharynx. At a first glance the ICC seemed reasonable high, unfortunately these measured volumes did not represent a true or similar tumour outline and thus true volume. Instead of 25 patients, only 5 patients were measured. The main reasons were that it was difficult to create a 3D-image, as due to the applied scan technique MR-images. Secondly, the outlines suggested by Smartbrush were not anatomically correct or tumour specific. Outlines had to be adjusted as air, bone, tumour, but also non-tumour tissues were selected automatically.

Strengths of the study

This is the first paper reporting the observer variation for Smartbrush software and the first paper reporting on Smartbrush software used for tumour volume measurements in the oropharyngeal space.

Comparisons with other studies

Rana et al[8] evaluated in 2015 the automatic tumour segmentation of mandibular odontogenic cysts and tumours with Smartbrush. Two-sample t-test showed no significant difference for manual segmentation compared with threshold segmentation (p=0.971) or segmentation via Smartbrush (p=0.838). Also Smartbrush segmentation resulted in volumes not significantly different from threshold segmentations (p=0.810).

However, the authors did not report correlations measured with the Intraclass Coefficient Correlation. Nor did the authors supply us with information about the observer experience, the used MR-scanning protocol or the order in which measurements were conducted. Measurements with Smartbrush were significantly faster compared with manual or threshold segmentation (time in minutes 4.01 min, 6.11 min and 7.31 min respectively). However, maybe this data consists of bias as it is unknown if measurements on Smartbrush were performed with knowledge about the tumour outlines from the other methods. Without measuring the exact times we needed for our measurements, the mean time was around 20 minutes for every patient. Also Rana et al[8] reported the volumes in mm^3, with an average tumour size of 5.5 mm^3. It is likely these measurements should have been given in cm^3, as the figure in their article shows a tumour which fills the entire maxillary sinus. A possible explanation for the fact we were not able to confirm their good results could be the difference in primary tumour site. We believe the software program will be able to measure volumes in more spherical and well-defined tumours such as the maxillary sinus better than superficial spreading tumours such as in the oropharyngeal region.

Limitations

Despite the efforts made to deliver all measurements to meet up with the power analysis, due to different problems we were not able to perform all measurements with 3 independent observers. The outlines suggested by Smartbrush were not anatomically correct or tumour specific, and therefore the software program did not give an advantage compared to fully manually derived tumour volume measurements.

Despite these limitations, the findings here add important new information to the current available literature as the results are different from the only paper published until now. The interobserver variation reached an excellent value with ICC of 0.98 and the intraobserver variation reached 0.88 for the first and 0.984 for the second observer respectively. However, it is questionable if this represents a clinical useful method.

Clinical applicability of the study

Hopefully, results can be declined or confirmed in a second larger project. Further measurements are currently performed on CT-images, where accuracy of Smartbrush will be compared to histopathology and manually derived tumour volume in laryngeal carcinomas.

CONCLUSIONS

The use of Smartbrush for the measurements of primary tumour volumes in the oropharyngeal space seems reasonable high. However, the outlines suggested by Smartbrush were not anatomically correct or tumour specific, and therefore the software program did not give an advantage compared to fully manually derived tumour volume measurements.

REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. 2010;127(12):2893-2917.

2. Sobin LH, Gospodarowicz M, Wittekind C. TNM classification of malignant tumors, wiley-blackwell, chichester, west sussex, UK. . 2009;7th edition.

3. Lodder WL, Pameijer FA, Rasch CR, van den Brekel MW, Balm AJ. Prognostic significance of radiologically determined neck node volume in head and neck cancer: A systematic review. *Oral Oncol*. 2012;48(4):298-302.

4. van den Broek GB, Rasch CR, Pameijer FA, et al. Pretreatment probability model for predicting outcome after intraarterial chemoradiation for advanced head and neck carcinoma. *Cancer*. 2004;101(8):1809-1817.

5. Flint PW. Cummings otolaryngology - head and neck surgery, 3-volume set: Expert consult: Online and print. Mosby Elsevier. 2010.

6. Mukherji SK, Toledano AY, Beldon C, et al. Interobserver reliability of computed tomography-derived primary tumor volume measurement in patients with supraglottic carcinoma. *Cancer*. 2005;103(12):2616-2622.

7. Lodder WL, Gilhuijs KG, Lange CA, Pameijer FA, Balm AJ, van den Brekel MW. Semi-automated primary tumor volume measurements by dynamic contrast-enhanced MRI in patients with head and neck cancer. *Head Neck*. 2013;35(4):521-526.

8. Rana M, Modrow D, Keuchel J, et al. Development and evaluation of an automatic tumor segmentation tool: A comparison between automatic, semi-automatic and manual segmentation of mandibular odontogenic cysts and tumors. *J Craniomaxillofac Surg*. 2015;43(3):355-359.

9. Fleiss J. Design and analysis of clinical experiments. John wiley & sons: Toronto. 1986.

Table 1. Patient demographics

Patient details		N = 5
Age	Mean (range)	63 years (52-69)
Gender	Male	3 (60%)
	Female	2 (40%)
T-stage	T1	0 (0%)
	T2	1 (20%)
	T3	0 (0%)
	T4a	3 (60%)
	T4b	1 (20%)
Site	Oropharynx	5 (100%)

**Figure 1. Interobserver and intraobserver variation (n=5)**
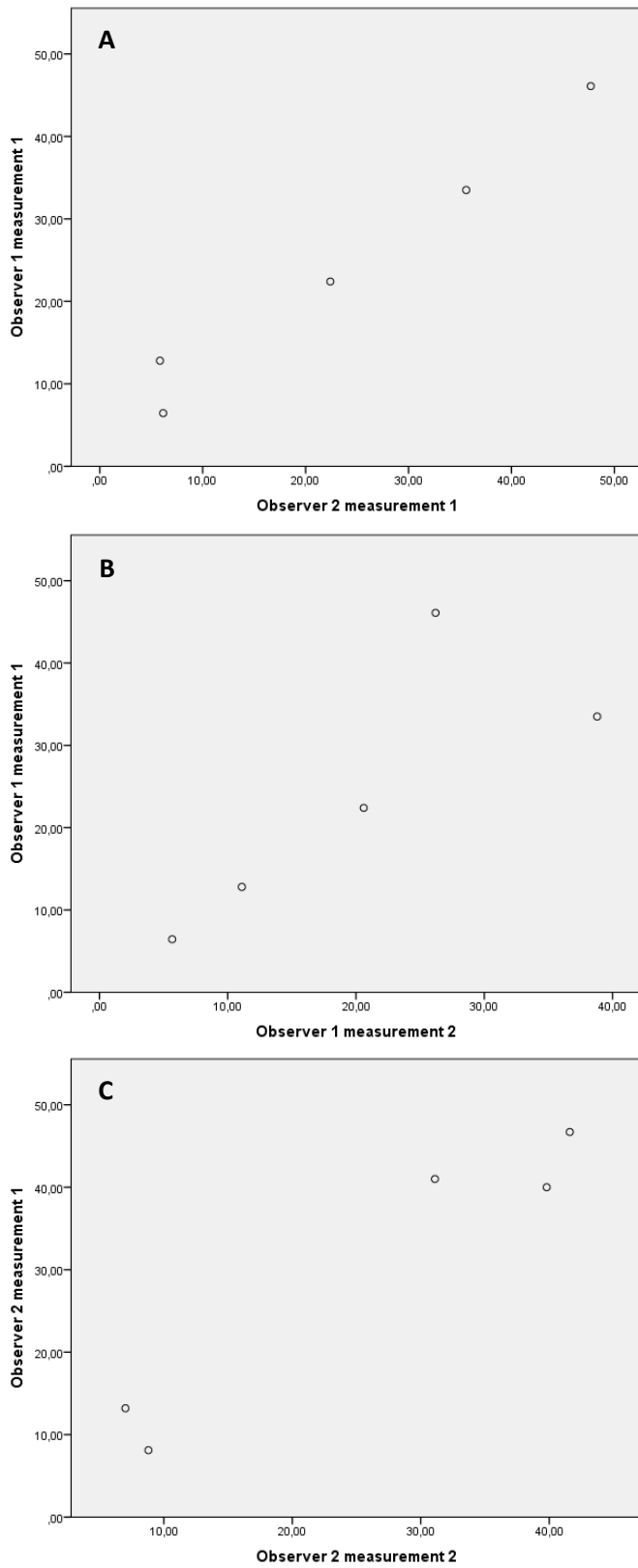
This figure shows the distribution of tumour volume measurements in cm^3. A shows the Intraclass Coefficient Correlation (ICC) for the interobserver variation (shown for the first and second observer. ICC was 0.98 (95% confidence interval 0.873 – 0.997, SD 50.85). The standard error of the measurement was 8.04 cm^3. B shows the ICC for the intraobserver variation for the first observer. ICC was 0.88 (95% confidence interval -0.143– 0.998, SD 27.45). The standard error of the measurement was 9.51 cm^3. C shows the ICC for the intraobserver variation for the second observer. ICC was 0.984 (95% confidence interval 0.842– 0.998, SD 34.19). The standard error of the measurement was 4.32 cm^3.

Figure 2. Example of tumour volume measurements with successful segmentation
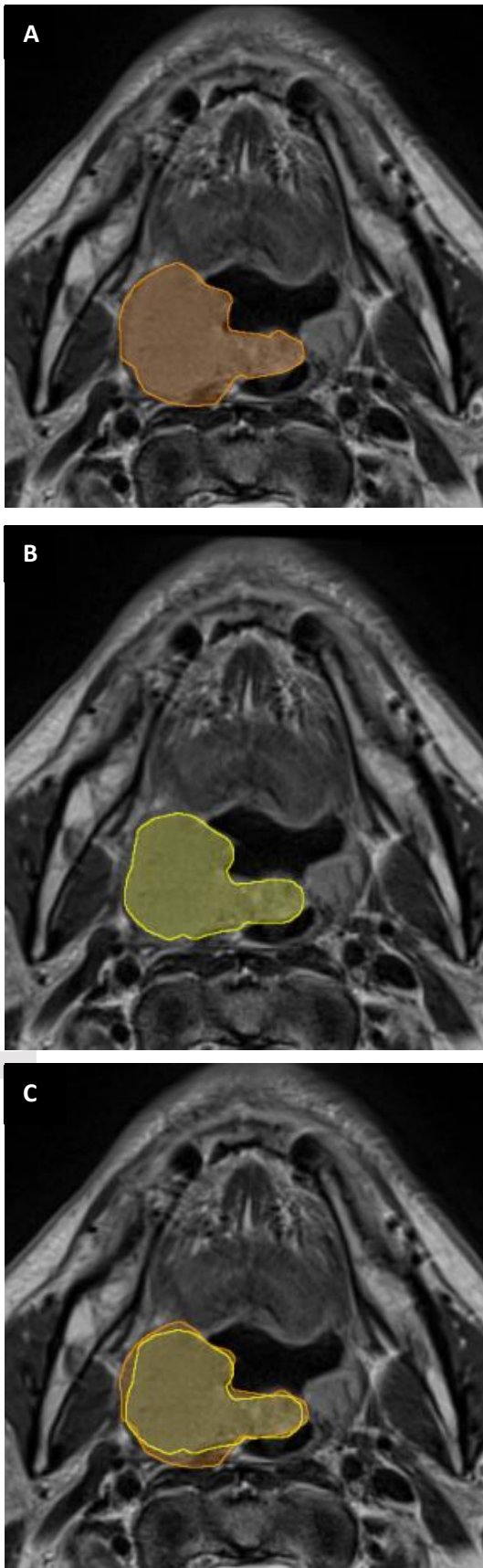
The figure shows the 1st and 2nd tumour volume measurement on a patient with a T4 oropharyngeal carcinoma at the right side. A shows the first measurement, giving a volume of 23.4 cm^3, B shows the second session resulting in a volume of 22.5 cm^3 as C shows the merged images of A and B. This is an example of the best result achieved.

Figure 3. Example of tumour volume measurements with different region-of-interest
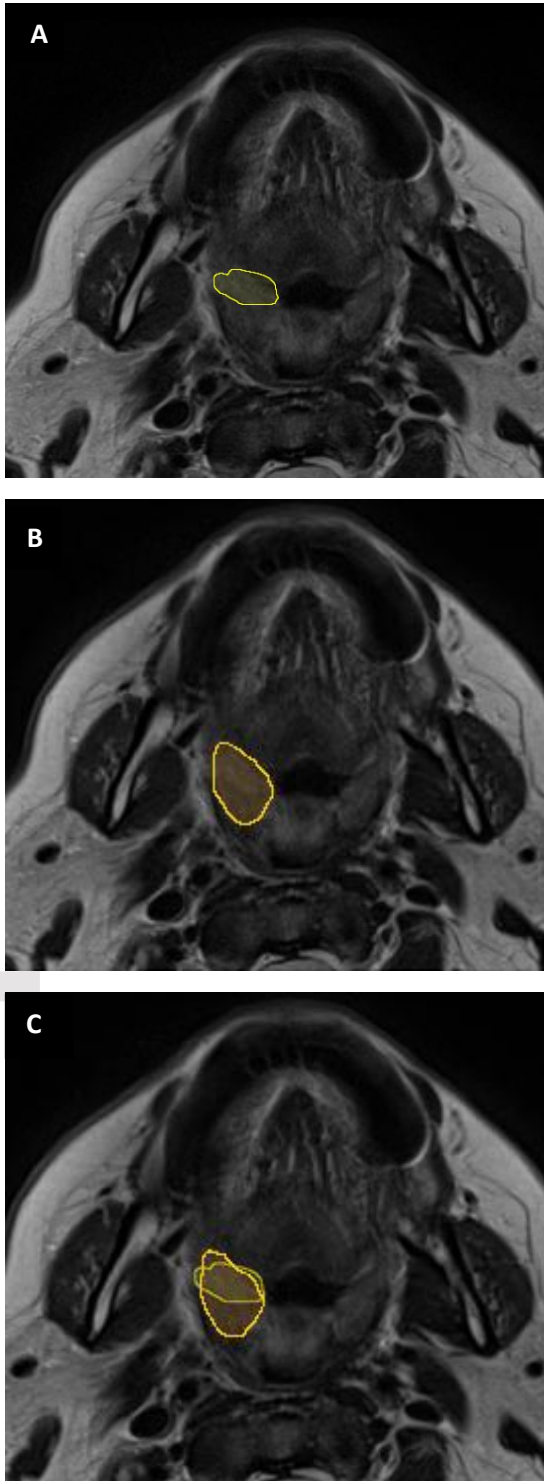
The figure shows the tumour volume measurement on a patient with a T4 oropharyngeal carcinoma at the right side. A shows the measurement of the first observer, giving a volume of 22.4 cm^3, B shows the measurement of the second observer also resulting in a volume of 22.4 cm^3. C shows the merged images of A and B. Note that the volume is exactly the same; however the majority of the drawn region-of-interest is at different place.