

# Ground Truth in MS Lesion Volumetry – A Phantom Study

Jan Rexilius, Horst K. Hahn, Holger Bourquain, and Heinz-Otto Peitgen

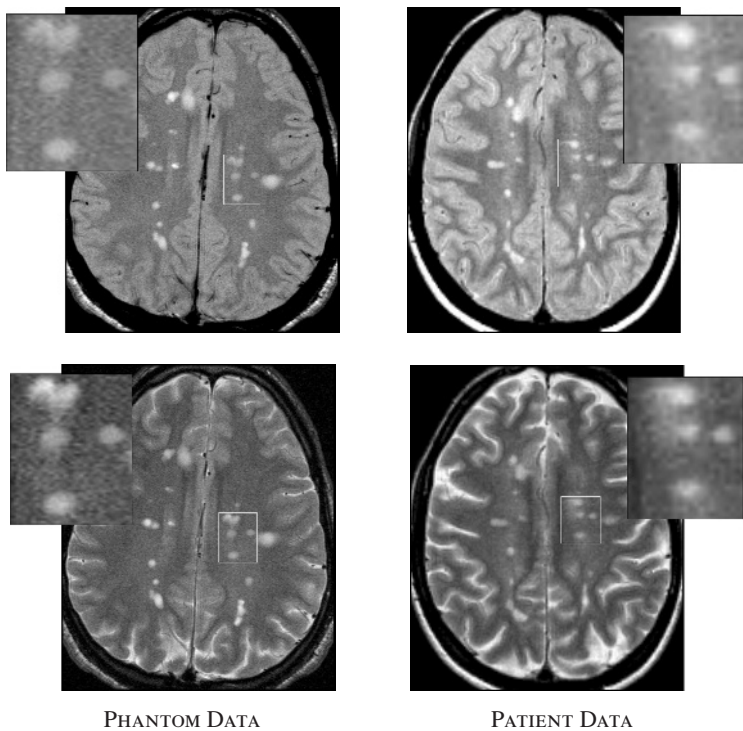
MeVis – Center for Medical Diagnostic Systems and Visualization,  
Universitaetsallee 29, 28359 Bremen, Germany, [rexilius@mevis.de](mailto:rexilius@mevis.de)

**Abstract.** A quantitative analysis of small structures such as focal lesions in patients suffering from multiple sclerosis (MS) is an important issue in both diagnosis and therapy monitoring. In order to reach clinical relevance, the reproducibility and especially the accuracy of a proposed method has to be validated. We propose a framework for the generation of realistic digital phantoms of MS lesions of known volumes and their incorporation into an MR dataset of a healthy volunteer. Due to the absence of a “ground truth” for lesions in general and MS lesions in particular, phantom data are a commonly used validation method for quantitative image analysis methods. However, currently available lesion phantoms suffer from the fact that the embedding structures are only simplifications of the real organs. We generated 54 datasets from a multispectral MR scan with incorporated MS lesion phantoms. The lesion phantoms were created using various shapes (3), sizes (6) and orientations (3). Since the common gold standard in clinical lesion volumetry is based on manual volume tracing, an evaluation is carried out from both a manual analysis of three human experts and a semi-automated approach based on regional histogram analysis. Additionally, an intra-observer study is performed. Our results clearly demonstrate the importance of an improved gold standard in lesion volumetry beyond manual tracing and voxel counting.

## 1 Introduction

Magnetic resonance imaging (MRI) has become an important imaging modality for understanding and managing several aspects of multiple sclerosis (MS) [1]. However, informations from MRI findings are often poorly correlated with the clinical manifestations of the disease. Therefore, any quantitative assessment of MS related parameters used in clinical routine and in multi-center studies has to be carefully evaluated. Today, the quantification of the volumetric lesion load is often used as an objective parameter. A fundamental issue is the accuracy of the calculated lesion volume since it can have a direct impact on diagnosis and therapy monitoring of the disease.

Several methods have been proposed in order to quantify the lesion burden ranging from manual tracing of each lesion by experts to semi-automated and fully automated methods, respectively [2,3,4,5,6]. However, due to the absence of a “ground truth” for MS lesions, computation of the exact volume of a lesion is still a challenging problem. Typical multispectral image acquisition protocols include proton density (PD) and T2-weighted sequences as well as T1-weighted sequences pre and post contrast with the tradeoff of a large slice thickness (about 3-5mm). Since MS lesions usually have a size



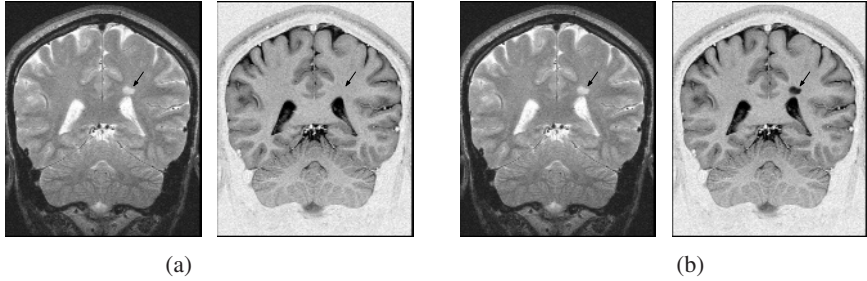
**Fig. 1.** Examples of different generated phantom MS lesions showing the potential of our approach on axial PD- and T2-weighted axial images. (Left row) MR scan of a healthy volunteer used in this work with 19 different incorporated MS lesion phantoms; (Right row) a patient's MR scan with several MS lesions.

in the order of the slice thickness, partial volume artifacts become a severe problem, affecting up to 100% of the object voxels. Nevertheless, the common gold standard for validation so far is based on manual tracing done by neurologists and neuroradiologists. Guidelines for manual outlining were proposed in recent studies [7,8].

A common approach for the validation of quantitative image analysis methods are phantom datasets. However, currently available lesion phantoms suffer from the fact that the embedding structures are only approximations of the real organs. A digital phantom for the brain was proposed in [9]. A phantom with a single lesion type and a fixed number of MS lesions is available from the BrainWeb project. A physical phantom of cylindrical shape with various known dimensions that was placed in an MR scanner is proposed in [10]. Both approaches are important steps towards accurate lesion volumetry but cover only a very limited range of MS lesions. Distributions of shape, size, and contrast behavior (e.g. black holes) that occur among different patients and within patients are not modeled. Especially a physical phantom can only be thought of as a rough estimate of a real lesion.

The aim of this paper is to introduce a new approach for the validation of MS lesion volumetry. We propose a framework for the generation of realistic phantoms of

MS lesions of known volumes incorporated in an MR scan. We carried out extensive experimental studies over a broad range of different lesions manually by domain experts as well as using a new robust semi-automated volumetry approach. The results indicate the importance of an improved gold standard in lesion volumetry.



**Fig. 2.** Modeling of different lesion types. (a) Standard T2-hyperintense lesion (left), hardly visible on the T1-weighted image (right); (b) same exact volume and positioning of lesion phantom as in (a), but now hypointense on T1-weighted image (black hole).

## 2 Method

An issue of specific clinical relevance for a quantitative analysis of MS lesions is the proof of validity of a proposed method. Due to the absence of a “ground truth” we generate datasets with realistic phantoms of MS lesions. Essential requirements for the development of our phantoms are (i) a known volume of each lesion to test the accuracy of an investigated method, (ii) the possibility to generate lesions with arbitrary shapes, sizes and contrast behavior, and (iii) the plausibility with respect to clinical image data.

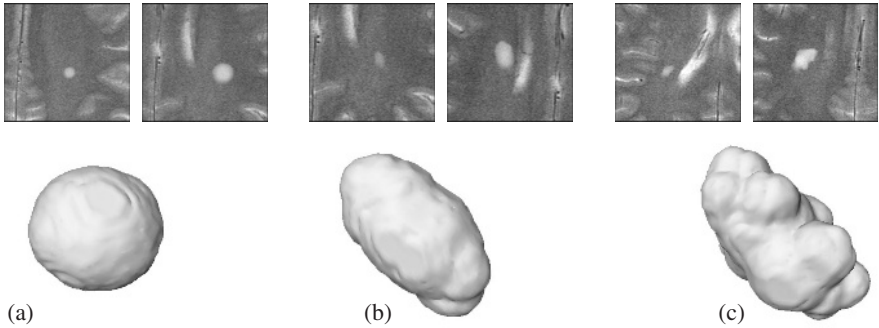
The generation of a phantom dataset with MS lesions consists of two basic steps. In a first step, a high-resolution and arbitrary shaped lesion with known volume is defined. Then, this lesion phantom is incorporated into an MR scan of the brain of a normal volunteer. Figure 1 and 2 illustrate the potential of our approach, showing a comparison to a patient’s MR scan with several MS lesion and a modeled dataset with different lesion types.

### 2.1 Generation of MS Lesion Phantoms

**Approximation of a continuous volume model.** In a first step, we generate a high-resolution binary lesion phantom volume  $I_p : \Theta \rightarrow \{0, 1\}$  with signal intensity values  $i_p(\mathbf{x}) \in \{0, 1\}$  at voxel positions  $\mathbf{x} = (x, y, z)^\top$ ,  $\mathbf{x} \in \Theta$ . A small voxel size is used in order to provide an appropriate approximation of a continuous volume

$$V_{i_p} = \int_{\Theta} i_p(\mathbf{x}) \, d\mathbf{x} .$$

Each lesion phantom is drawn on a  $512^3$  grid with a constant of voxels set to 1. Different volumes can then be easily generated by specifying a different voxel size. In order to obtain a phantom with realistic shaped MS lesions, we generated a sphere-like lesion (S1), a cylindrical lesion (S2) and a lesion of longish shape containing several deformations (S3). We did not use regular shaped cylinders or spheres since this wouldn't lead to a realistic lesion appearance. A 3D surface rendering of these shapes is shown in Figure 3.



**Fig. 3.** MS lesion phantoms used in this paper (top: incorporated in brain dataset, left: 0.1ml, right: 0.7ml; bottom: surface rendering). lesion phantom with (a) sphere like, (b) cylindrical, and (c) longish shape containing several deformations.

Besides various shapes, lesions of different volumes should also be tested by a validation study. In this work we choose volumes of size 0.05ml, 0.1ml, 0.2ml, 0.4ml, 0.7ml, and 1.0ml. The voxel size for the largest volume (1.0ml) was set to  $0.05^3$ mm resulting in 8.000.000 voxels. The voxel sizes for the remaining volumes are calculated accordingly. The maximum diameter varies from 5.4mm (S2=7.6mm, S3=7.2mm) for the smallest lesions (0.05ml) up to 13.5mm (S2=18.4mm, S3=19.3mm) for the largest lesions (1.0ml).

**Partial volume averaging.** After generating a lesion phantom of a desired shape and volume, it is incorporated into our multispectral MR scan of the brain  $I_b : \Omega \rightarrow \mathbb{R}$  with intensity values  $i_b(\mathbf{x})$ . Therefore, the lesion phantom  $I_p$  is downsampled to the same voxel size as the MR scan, using trilinear interpolation and then reformatted into the coordinate system of the MR scan. The exact volume of this object was computed as well and compared to the original volume to ensure the accuracy of this step. This results in a probability map  $\tilde{I}_p : \Omega \rightarrow \mathbb{R}$  with intensity values  $\nu := \tilde{i}_p(\mathbf{x}) \in [0, \dots, 1]$ .

**Available MR scans and generation of lesion gray values.** The volumetric MR images used in this study were acquired from a healthy volunteer (M, 28y) on a Siemens Magnetom Vision 1.5T scanner. The data acquisition protocol contained axial as well as coronal PD-, T2-, and T1-weighted images with an in-plane resolution of  $0.449 \times 0.449 \text{mm}^2$  and a slice thickness of 3mm, matrix  $512 \times 512$ , 34 axial and 51 coronal slices respectively. All images were acquired in one session with head fixation and without table movement such that all datasets are perfectly aligned without visible motion artifacts.

Furthermore, a volume  $I_l$  with intensity values  $i_l(\mathbf{x})$  containing reasonable gray values for MS lesions for each available sequence is created, and Gaussian noise with zero-mean and standard deviation  $\sigma = 1$  is added

$$\tilde{i}_l = i_l + \mathcal{N}_{\mu=0, \sigma=1}(i_l) \cdot 100 .$$

The lesion gray values as well as noise noise are adjusted based on inspections of patient datasets with MS lesions.

**Incorporation of a lesion phantom into an MR scan.** In a final step, the phantom dataset is generated as a linear combination of the MR scan  $I_b$  and the volume  $\tilde{I}_l$  containing appropriate lesion gray values. The new signal intensity value is modeled as the convex combination

$$\tilde{i}_b = \nu \cdot \tilde{i}_l + (1 - \nu) \cdot i_b$$

and assigned to each voxel of the MR dataset. This allows the construction of datasets with a large range of possible lesions as shown in Figure 1.

Since the aim of this paper is not the identification of MS lesions, but rather the volumetric analysis, each lesion is placed in the brain so that it is totally surrounded by white matter.

## 2.2 Setup for Manual Analysis by Human Experts

Each dataset was analyzed by three experts. In order to provide an intuitive but still powerful tool for the manual analysis of the provided phantom datasets, an application with a graphical user interface based on the research and development platform ILAB4 [11] was developed. Therein, the expert is able to draw the contours of a lesion, shown as overlay on the original slices. In addition to basic drawing functionalities, the user has the option to adjust the lookup-table and view several neighboring slices simultaneously. Furthermore it is possible to change between available sequences during outlining a lesion on one slice. Since we want to analyze volumetric results and not the lesion detection task of different experts, only one lesion is incorporated per dataset.

## 2.3 Robust Semi-automated Volumetry with Explicit Partial Volume Modelling

The proposed semi-automated volumetry method combines a 3d marker based segmentation and a bimodal histogram analysis with an explicit model for partial volume effects. Only the T2-weighted images were considered due to their high lesion contrast. In a first step, a cuboid subvolume that contains the entire lesion is selected and resampled in z-direction to an isotropic voxel size, using a Lanczos 3-lobed filter. Then, an Interactive Watershed Transformation is applied to generate an over-inclusive segmentation [12]. Two different marker types are used. One include marker is placed inside the lesion and another three to ten exclude markers are used to separate the lesion from other hyperintense structures. The volume is computed by an automatic bimodal histogram analysis assuming two Gaussians  $\varphi_{obj}$ ,  $\varphi_{bg}$  with equal variance, and symmetric, equally distributed partial voluming ( $p_{pv}(i)$ ).

$$p_{pv}(i) = \frac{\Phi_{bg}(i) - \Phi_{obj}(i)}{\mu_{obj} - \mu_{bg}} \quad , \quad \Phi_{\alpha}(x) = \int_{x'=-\infty}^x \varphi_{\alpha}(x') dx'$$

The model function then contains six independent parameters

$$f(i) = \sum_{\alpha} A_{\alpha} \cdot \varphi_{\alpha}(i) + A_{pv} \cdot p_{pv}(i)$$

and is fitted to the histogram  $h(i)$  using a least squared error minimization. The volume is computed by

$$V = \left[ A_{obj} + \frac{1}{2} A_{pv} + \sum_{i=\mu'}^{i_{max}} (h(i) - f(i)) \right] \cdot V_{vox}, \quad \mu' = \frac{\mu_{bg} + \mu_{obj}}{2}.$$

where the variables are given in terms of the intensity values,  $i$ , the center of the distributions,  $\mu_{\alpha}$ , the voxel volume,  $V_{vox}$ , and the area und the curve,  $A_{\alpha}$ .

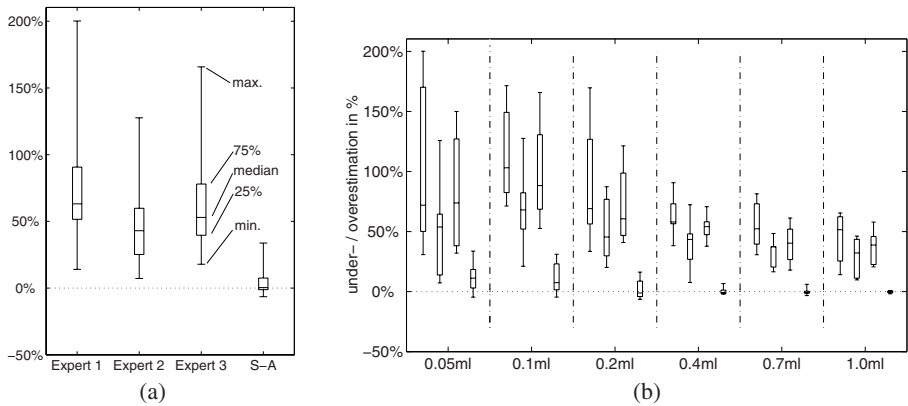
### 3 Results

We evaluated 54 phantom MR datasets generated from a brain scan of a normal volunteer with exactly one MS lesion phantom per dataset. The lesion phantoms were placed at typical paraventricular positions in the brain. A total amount of 18 different lesions was generated for this study consisting of six different volumes (0.05ml, 0.1ml, 0.2ml, 0.4ml, 0.7ml, and 1.0ml) for each of three generated lesion shapes (cf. Sect. 2.1). Each lesion phantom was placed at the exact same position in the available axial as well as coronal MR scans of a normal volunteer, which results in two different outlining tasks for the raters. A third set of phantom datasets was generated by repositioning each lesion in the axial scan by shifting it by the half slice thickness in z-direction. Due to partial volume artifacts, this can have a great influence on the apparent lesion size. The MS lesion phantoms were manually traced by three experienced raters (cf. Sect. 2.2). Furthermore, a new semi-automated volumetry approach was used (cf. Sect. 2.3) for comparison.

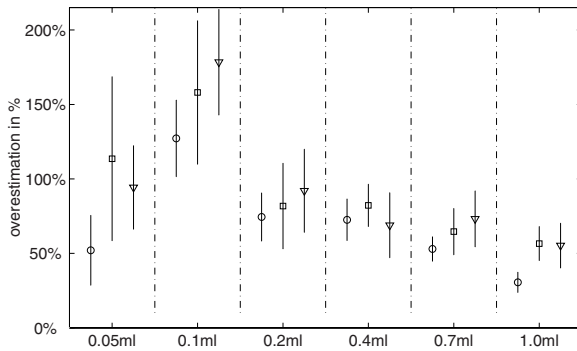
Figure 4 illustrates the error for each rater and the semi-automatic method, in percentage of the true volume in a box and whisker plot. It can be clearly observed, that all experts overestimated the real lesion volume, all to a comparable amount. The overall median overestimation for manual evaluation of the three experts ranges between 43% and 63.2%. The variability decreases with increasing volume size, since small changes already cause a significant relative error for small volumes. For small lesions (<0.3ml) the median overestimation over all experts is 73.2% (min: 45.5%, max: 103.1%), and 45% (min: 32.2%, max: 58%) for intermediate lesions (>0.3ml). No significant shape effect has been observed.

The semi-automatic partial volume analysis on the other hand has a far lower error margin with an overall median overestimation of 0.4%. For small lesions it is 7.5% and -0.65% for intermediate lesions, respectively. Especially the deviation from the median is much smaller than for the manual analysis. However, the proposed approach yet tends to overestimate the true lesion volume, due to partial volume artifacts.

The intra-observer variability was further inspected in a repeatability analysis (10x) of the axial phantom datasets by expert three. Figure 5 shows the mean and variance of the measured volumes. Here, the mean overestimation over all analyses and shapes is 108% (SD 40.7%) for small lesions and 61.9% (SD 15.2%) for intermediate lesions, respectively.



**Fig. 4.** Results of manual and semi-automated volume measurements. (a) Overall results calculated for each rater separately (S-A refers to semi-automatic partial volume analysis); (b) results for each available volume (expert 1–3 and semi-automatic approach).



**Fig. 5.** Results of intra-observer study showing a systematic overestimation along with a large variation in volume measurements (shapes: ○–S1; □–S2; ▽–S3).

## 4 Discussion and Conclusion

Our main objective in this work has been to develop and validate a framework for the generation of realistic digital phantoms of MS lesions with exactly known volumes. Therefore, an arbitrarily shaped high-resolution lesion phantom is incorporated into an MR scan of a normal volunteer.

Although a correlation between MRI and clinical findings remains difficult, the volumetric analysis of the lesion load has become an important issue and an active research field. However, an analysis of the common gold standard in MS lesion volumetry between three experts shows a median overestimation between 32% and 103%. The overall maximum was approximately 200% for one expert. An intra-observer study also showed a large variability even for a single rater. No manual tracing underestimated or met the true volume. Since accuracy is an important factor for the clinical relevance of

a method, the results clearly indicate the importance of an improved gold standard in lesion volumetry beyond manual tracing and voxel counting. New measurements that accurately address partial volume artifacts are likely to correspond better to clinical findings. Therefore, our phantoms can provide a realistic basis for the validation of current and new approaches.

Future work will investigate the accuracy and reproducibility of different lesion identification and classification methods. In order to provide a new tool for comparison in MS lesion volumetry, the MR scan used in this work as well as the generated datasets with MS lesion phantoms will be made available for other researchers, who are encouraged to contact the authors.

**Acknowledgements.** We thank Prof. Dr. B. Terwey (Center for Magnetic Resonance, Bremen) for providing the MR datasets used in this work.

## References

1. D.H. Miller, R.I. Grossman, S.C. Reingold, H.F. McFarland. The role of magnetic resonance techniques in understanding and managing multiple sclerosis. *Brain*, 121, pp. 3–24, Jan. 1998
2. M. Filippi, M.A. Horsfield, S. Bressi, et al. Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis. A comparison of techniques. *Brain*, 118 (6):1593–600, Dec. 1995.
3. J.K. Udupa, L. Wei, S. Samarasekera, et al. Multiple Sclerosis Lesion Quantification Using Fuzzy-Connectedness Principles. *IEEE TMI*, vol. 16, no. 5, pp. 598–609, Oct. 1997.
4. K. Van Leemput, F. Maes, D. Vandermeulen, et al. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE TMI*, vol. 20, no. 8, pp. 677–688, Aug. 2001.
5. A. Zijdenbos, R. Forghani, A. Evans. Automatic Quantification of MS Lesions in 3D MRI Brain Data Sets: Validation of INSECT. In *MICCAI 1998*, pp. 439–448, Oct. 1998.
6. S. Al-Zubi, K.D. Tönnies, N. Bodammer, H. Hinrichs. Fusing markov random fields with anatomical knowledge and shape based analysis to segment multiple sclerosis white matter lesions in magnetic resonance images of the brain. In *Proceedings of SPIE (Medical Imaging 2002)*, vol. 4684, pp. 206–215, San Diego, 23.–28. Feb. 2002.
7. F. Barkhof, M. Filippi, et al. Improving interobserver variation in reporting gadolinium-enhanced MRI lesions in multiple sclerosis. *Neurology*, 49:1682–1688, Dec. 1997.
8. P.S. Tofts, G.J. Barker, M. Filippi, M. Gawne-Cain, M. Lai. An oblique cylinder contrast-adjusted (OCCA) phantom to measure the accuracy of MRI brain lesion volume estimation schemes in multiple sclerosis. *J. Magn Reson Imaging*, 15(2):183–192, 1997.
9. D. Collins, A. Zijdenbos, V. Kollokian, et al. Design and Construction of a Realistic Digital Brain Phantom. *IEEE TMI*, vol. 17, no. 5, pp. 463–468, June 1998.
10. M. Filippi, M.L. Gawne-Cain, C. Gasperini, et al. Effect on training and different measurement strategies on the reproducibility of brain MRI lesion load measurements in multiple sclerosis. *Neurology*, pp. 238–244, Jan. 1998.
11. H.K. Hahn, F. Link, H.-O. Peitgen. Concepts for a Rapid Prototyping Platform in Medical Image Analysis and Visualization. *Proc. SimVis*, pp. 283–298, SCS, Mar. 2003.
12. H.K. Hahn, H.-O. Peitgen. IWT – Interactive Watershed Transform: A Hierarchical Method for Efficient Interactive and Automated Segmentation of Multidimensional Gray-Scale Images. *Medical Imaging: Image Processing; Proc. SPIE 5032*, Feb. 2003, pp. 643–653.