Performance of five research-domain automated WM lesion segmentation methods in a multi-center MS study

Alexandra de Sitter[a], Martijn D. Steenwijk[b], Aurélie Ruet[c,d,e], Adriaan Versteeg[a], Yaou Liu[a], Ronald A. van Schijndel[a], Petra J.W. Pouwels[a], Iris D. Kilsdonk[a], Keith S. Cover[a], Bob W. van Dijk[b], Stefan Ropele[f], Maria A. Rocca[g], Marios Yiannakas[h], Mike P. Wattjes[a], Soheil Damangir[i], Giovanni B. Frisoni[j,k], Jaume Sastre-Garriga[l], Alex Rovira[m], Christian Enzinger[f,n], Massimo Filippi[g], Jette Frederiksen[o], Olga Ciccarelli[p], Ludwig Kappos[q], Frederik Barkhof[a,r], Hugo Vrenken[a,b] for the MAGNIMS study group and for neuGRID.

[a] Department of Radiology and Nuclear Medicine, VUmc, Amsterdam, Netherlands

[b] Department of Anatomy and Neuroscience, VUmc, Amsterdam, Netherlands

[c] Department of Neurology, CHU-Bordeaux, Bordeaux, France

[d] Inserm-CHU CIC-P0005, CHU-Bordeaux, Bordeaux, France

[e] Inserm U-1215 Magendie Neurocenter-Pathophysiology of neural plasticity, CHU-Bordeaux, Bordeaux, France

[f] Department of Neurology, Medical University of Graz, Graz, Austria

[g] Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, San Raffaele Scientific Institute, UniSR, Milan, Italy.

[h] Department of Neuroinflammation, Institute of Neurology, UCL, London, UK

[i] Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden

[j] Laboratory of Epidemiology, Neuroimaging and Telemedicine, IRCCS Centro "S. Giovanni di Dio-F.B.F.", Brescia, Italy

[k] Memory Clinic and LANVIE - Laboratory of Neuroimaging of Aging, HUG, Geneva, Switzerland

[l] Centre d'Esclerosi Múltiple de Catalunya (Cemcat), Department of Neurology/Neuroimmunology, VHIR, Barcelona, Spain.

[m] Magnetic Resonance Unit. Department of Radiology (IDI), VHIR, Barcelona, Spain.

[n] Division of Neuroradiology, Vascular and Interventional Radiology, Department of Radiology, Medical University of Graz, Graz, Austria

[o] Department of Neurology, Glostrup University Hospital, Copenhagen, Denmark

[p] UK/NIHR UCL-UCLH Biomedical Research Centre, Institute of Neurology, UCL, London, UK

[q] Neurologic Clinic and Policlinic, University Hospital, University of Basel, Switzerland

[r] Institutes of Neurology & Healthcare Engineering, UCL, London, UK

Keywords: multiple sclerosis; white matter lesion; automated methods segmentation; MRI

Author contact information:

Name: Alexandra de Sitter

Address: VUmc, PO Box 7057, 1007 MB Amsterdam, The Netherlands

Telephone: +31 204444596

Fax: +31 204440397

Email: a.desitter@vumc.nl

Word count: 6455

Acknowledgements:

Aurélie Ruet was supported by an ECTRIMS research fellowship. Iris D. Kilsdonk was supported by a grant provided by the Noaber Foundation (Lunteren, The Netherlands). Adriaan Versteeg, Ronald A. van Schijndel, Keith S. Cover, Soheil Damangir and Giovanni B. Frisoni were partly funded by neuGRID4you (www.neuGRID4you.eu), an European Community FP7 project (grant agreement 283562). Olga Ciccarelli and Frederol Barkhof were supported by the National Institute for Health Research (NIHR) University College London Hospitals (UCLH) Biomedical Research Centre (BRC). The MS Center Amsterdam is supported by the Dutch MS Research Foundation through program grants (09-358d and 14-358e). The authors thank M. Jonker (from the Image Analyse Center, VUmc) for his work on the manual segmentations. MAGIMS (Magnetic Resonance Imaging in MS) study group is an European study group of academics that share the interest in MS and imaging.

Abstract

Background and Purpose: *In vivo* identification of white matter lesions plays a key-role in evaluation of patients with multiple sclerosis (MS). Automated lesion segmentation methods have been developed to substitute manual outlining, but evidence of their performance in multi-center investigations is lacking. In this work, five research-domain automated segmentation methods were evaluated using a multi-center MS dataset.

Methods: 70 MS patients (median EDSS of 2.0 [range 0.0-6.5]) were included from a six-center dataset of the MAGNIMS Study Group (www.magnims.eu) which included 2D FLAIR and 3D T1 images with manual lesion segmentation as a reference. Automated lesion segmentations were produced using five algorithms: Cascade; Lesion Segmentation Toolbox (LST) with both the Lesion growth algorithm (LGA) and the Lesion prediction algorithm (LPA); Lesion-Topology preserving Anatomical Segmentation (Lesion-TOADS); and k-Nearest Neighbor with Tissue Type Priors (kNN-TTP). Main software parameters were optimized using a training set (N=18), and formal testing was performed on the remaining patients (N=52). To evaluate volumetric agreement with the reference segmentations, intraclass correlation coefficient (ICC) as well as mean difference in lesion volumes between the automated and reference segmentations were calculated. The Similarity Index (SI), False Positive (FP) volumes and False Negative (FN) volumes were used to examine spatial agreement. All analyses were repeated using a leave-one-center-out design to exclude the center of interest from the training phase to evaluate the performance of the method on 'unseen' center.

Results: Compared to the reference mean lesion volume (4.85±7.29 mL), the methods displayed a mean difference of 1.60±4.83 (Cascade), 2.31±7.66 (LGA), 0.44±4.68 (LPA), 1.76±4.17 (Lesion-TOADS) and -1.39±4.10 mL (kNN-TTP). The ICCs were 0.755, 0.713, 0.851, 0.806 and 0.723, respectively. Spatial agreement with reference segmentations was higher for LPA (SI=0.37±0.23),

Lesion-TOADS (SI=0.35±0.18) and kNN-TTP (SI=0.44±0.14) than for Cascade (SI=0.26±0.17) or LGA (SI=0.31±0.23). All methods showed highly similar results when used on data from a center not used in software parameter optimization.

Conclusion: The performance of the methods in this multi-center MS dataset was moderate, but appeared to be robust even with new datasets from centers not included in training the automated methods.

1. Introduction

Multiple sclerosis (MS) is an inflammatory and neurodegenerative disease of the central nervous system, with inflammatory white matter (WM) lesions as prominent pathological hallmark.[1,2] *In vivo* visualization of lesions by means of MRI plays a crucial role in the diagnosis and study of MS. Moreover, several clinical trials have used WM lesion volume as a (secondary) study outcome.[3-6]

For clinical and research purposes, delineation of WM lesions in MS is either performed manually or with a semiautomatic tool. These approaches, however, are labor-intensive and suffer from considerable inter- and intra-rater variability.[7,8] To overcome these problems, automated WM lesion segmentation methods have been developed in the last decade.[9] However, these methods are not routinely applied in research, clinical trials or individual patient care. One important hurdle is the lack of comparative data reporting the accuracy and robustness of these methods when using data obtained from different centers. Evidence of their performance in multi-center investigations is lacking.

The aim of this study was, firstly, to evaluate the performance of research-domain automated WM lesion segmentation methods in a multi-center MS dataset with diverging scanners and protocols. And secondly, to investigate how these methods perform on data from a new center (using other centers for training). We selected five algorithms for automated segmentation: Cascade[10,11]; Lesion growth algorithm (LGA)[12] and Lesion prediction algorithm (LPA)[13] both from the Lesion Segmentation Toolbox (LST)[12]; Lesion-Topology-preserving Anatomical Segmentation (Lesion-TOADS)[14]; and k-Nearest Neighbor with Tissue Type Priors (kNN-TTP)[15].

2. **Methods and materials**

2.1. Subjects

The data for this study were drawn from a multi-center MS dataset that was collected by the MAGNIMS Study Group (www.magnims.eu) as described previously.[16] For the analyses described in the current paper, we selected the patients with a 2D FLAIR acquisition, and we excluded three patients with co-morbidity (vascular disease, glioblastoma, surgical removal of part of the brain) that could interfere with the automated lesion segmentation and one patients whose data were acquired after contrast agent administration.

This resulted in a dataset of 70 patients (67% female) that were scanned in six different MAGNIMS centers. Of the 70 patients, 1 was diagnosed with clinically isolated syndrome, 65 had relapsing-remitting MS, and 4 had secondary-progressive MS. The average age was 34.9±8.5 years and mean disease duration was 7.6±6.0 years from onset. The disease severity was measured using the Expanded Disability Status Scale (EDSS)[17] on the day of scanning and scores a median of 2 (range 0.0-6.5), see Table 1. We additionally collected data of 12 healthy controls (2 per center) from the larger sample (58% females; age: 31.2±7.2 years) for kNN-TTP training (see 2.4.4). Written informed consent had been obtained from all subjects and the institutional review boards of each participating center approved the study.

2.2. MR imaging

MR imaging was performed on 3.0 Tesla whole-body MR systems. Each protocol contained a 3D T1-weighted sequence with a (near) isotropic voxel size of approximately $1mm^3$ and a 2D FLAIR sequence with a 3.0-mm slice thickness and an in plane resolution of 0.75-1.0 mm. More details are listed in Table 2.

2.3. Manual segmentations

The manual reference segmentations were constructed as follows: an experienced neuro-radiologist (YL) manually marked the lesions, after which a trained rater outlined the lesions. To assess intra-observer variability, the rater constructed a second set of segmentations for five subjects based on the same markings. Furthermore, for these same five patients, the neuro-radiologist marked the lesions a second time after which the rater created segmentations based on these new markings, resulting in a third set of segmentations used for quantification of inter-rater variability.

2.4. Automated WM lesion segmentation methods

Methods were included if: the method 1) is able to segment WM lesions in a fully automated manner from FLAIR (and T1-weighted) images; 2) is freely available for academic research work; and 3) can be installed on the NCA grid cluster (www.amsterdamresearch.org/web/neuroscience). Based on these criteria, we selected Cascade; LGA and LPA, both from the LST; Lesion-TOADS; and kNN-TTP. More information on the inclusion is listed in Table 9 of supplementary data. The post-processing used is listed in Table 3 and a short description of each method is listed below. A short introduction on each of the methods is provided below.

2.4.1. Cascade

Cascade[10,11] is an unsupervised method based on a proposed statistical definition of WML which was not specifically developed for MS WM lesions. The method applies a single node support vector machine (SVM) to preselect tissues that shows WM changes (areas with low and high intensities on respectively, T1 and T2 weighted images) compared to normal grey matter (GM) and WM. The WM

changes are tested against a statistical definition of WM lesions, and by applying a threshold (α), a selection of lesions is made. From this selection, a WM lesion p-value (LPV) map is created by applying a morphological filter and removing small lesions (lesion size < 27 mm3). Then the LPV map is converted to a binary mask by applying a threshold (LPV threshold).

2.4.2. LGA

LGA[12], implemented in the LST, is based on a lesion growth algorithm. For pre-processing, FLAIR images are linearly registered to T1-weighted images, as well as the generated partial volume estimated (PVE) label and inverse warping of the WM tissue probability map (TPM_{WM}). Next, WM, GM and cerebrospinal fluid (CSF) lesion belief maps (B_{WM}, B_{GM}, B_{CSF}) are calculated and a liberal lesion map is created as the sum of the belief maps. By binarizing B_{GM} with a threshold (κ) the initial lesion map is constructed. Lesion growing is then performed on the initial lesion map toward the liberal lesion map, and stopped when no more voxels are added to the lesions. Finally, the lesion-probability (LP) map is converted to a binary lesion mask by applying an user specified lesion probability threshold (LPT).

2.4.3. LPA

LPA[13], implemented in the Lesion Segmentation Toolbox (LST), is based on a lesion prediction algorithm. For pre-processing, FLAIR images are linearly registered to T1-weighted images and are normalized to MNI space. WM, GM and CSF are segmented and the FLAIR intensities are standardized by dividing each voxel by the mean of the GM segmentation followed with a subtraction on each voxel with the mean of the standardized GM intensities. The remaining positive voxels on the FLAIR images are multiplied with a WM tissue probability map and negative voxels

set to zero to create the LP map. Lastly, the LP map is converted to a binary lesion mask by applying a LPT.

2.4.4. Lesion-TOADS

Lesion-TOADS[14] is a topology preserving atlas-based segmentation method. In short, Lesion-TOADS performs fuzzy segmentation and defines topologically consistent regions with an additional lesion class. A relationship function is used for reducing the false positive (FP) lesions in areas where they commonly occur, and an intensity-weighting scheme is incorporated for optimizing the effect of each channel of the multichannel inputs.

2.4.5. kNN-TTP

kNN-TTP[15] segments WM lesions using kNN classification extended with anatomical TTP derived from healthy controls. The segmentation is done by comparing the voxels of new data with a training set of labelled examples, which means that, in contrast to the three other methods, kNN-TTP needs a labelled reference data set. A leave-one-out procedure was used to correct for the use of labelled examples to make the method not biased. Finally, the LP map is converted to a binary lesion mask by applying a LPT.

2.5. Training

The patients were divided into a separate training and test set. For each center, patients were ranked according to reference lesion volume and three patients were then selected for the training set: the patient with the median lesion volume, the patient with the second lowest volume and the patient

with the second highest lesion volume. This resulted in a training set of 18 patients, while the remaining 52 patients formed the test set.

2.6. Pre- and post-processing

For LGA, LPA and kNN-TTP the data was pre-processed as based on the reports in the literature[12,13,15] The publications on Cascade and Lesion-TOADS do not prescribe specific pre-processing steps; therefore, the same pre-processing steps as in kNN-TTP were applied.

Because LGA and LPA produced the LP-maps in T1-space, the LP maps found by LGA and LPA were first co-registered to FLAIR space before thresholding was applied (FSL-FLIRT, 6 DOF and trilinear interpolation) to create the binary lesions maps. Cascade and kNN-TTP needed post-processing to construct binary lesion maps by applying a threshold, which was optimized during the training phase. See Table 3 for an overview.

2.7. Parameter tuning

The performance of Cascade, LGA, LPA and kNN-TTP can be improved by optimizing method-specific configuration parameters. For Cascade, optimal performance was assured by running the segmentation algorithm on the training data while sweeping the two main parameters (α={0.05, 0.10, …, 1.00} and LPV threshold={α, α-0.05, …, 0}).[10,11] The parameter combination that provided the group-wise highest average spatial overlap was considered as the optimal configuration. A similar approach was used to derive the optimal configuration for LPA, sweeping κ={0.05, 0.10, …, 1.00} and LPT={0, 0.05, …, 1.00}.[12] LGA and kNN-TTP were optimized by sweeping LPT={0, 0.05, …, 1.00}. Because in kNN-TTP, the reference data are actively used in every run of the algorithm, a leave-one-out cross-validation was used to optimize kNN-TTP parameters to ensure independence of

the evaluation.[15] The optimal parameter settings were selected on the highest mean spatial overlap (similarity coefficient (SI), see section Evaluation) compared with the manual references.

2.8. Experiments

The parameter tuning was performed several times in the following experiments, see Figure 1.

Experiment 1) To evaluate the performance of the automated methods while tuning the parameters on the data of all centers. In this experiment, all parameter tuning was done using the full training set and subsequently tested using the full test set.

Experiment 2) To evaluate the performance of the automated methods on new (unseen) data, six partial experiments were performed, treating each of the six centers in turn as the (unseen) center-of-interest. Each time, the data from the center-of-interest was excluded during the parameter tuning, and the optimized methods were then tested on the center-of-interest data.

Three additional analyses were performed with the data of Experiment 1.

1) Impact of lesion volume:

   the data was re-analyzed for subject groups with low (<5 mL), intermediate (5-15 mL) and high (>15 mL) lesion volume.

2) Impact of the training set selection:

   the data was re-analyzed with use of two different training sets, selecting for training from each center first the three patients with lowest lesion volumes, and second the three patients with the highest lesion volumes. The impact of the training set composition was subsequently evaluated by comparing the subset of patients that were not part of any of the three training sets (N = 28). The agreement of the segmentations with the manual references was evaluated using SI.

3) Comparison of automated methods against data-driven reference:

we assessed the performance of the algorithm against a data driven reference using the STAPLE algorithm.[19] The STAPLE algorithm estimates the sensitivity and specificity for each automated segmentation compared to an estimated true segmentation derived by STAPLE.

2.9. Evaluation

The method-wise performance was evaluated by examining volumetric and spatial agreement of the automated methods with the manual reference segmentation. The segmentations of Experiment 2 were evaluated as one group, despite differences in optimized parameters depending on the centers included in the parameter tuning in order to provide a complete assessment of the performance on unseen data that can be directly compared with the results of Experiment 1.

For volumetric agreement, the mean difference in lesion volume between the automated segmentation and the manual reference segmentation was evaluated. Furthermore, the intra-class correlation coefficient (ICC, two-way mixed model with absolute agreement)[20] was calculated. Spatial agreement between the methods and the manual reference was evaluated using the Dice similarity index[21]:

$$SI = (2 \times TP)/(2 \times TP + FP + FN),$$

where TP, FP and FN are, respectively, true positive, false positive and false negative volume. To determine the origin of spatial disagreement, we also computed the detection error rate (DER; the detection error volume divided by mean total volume) to quantify the effect of missed and false positive lesions on spatial disagreement, and the outline error rate (OER; the outline error volume divided by mean total volume) to quantify the effect of altered boundaries of correctly detected

lesions (see supplementary data for equations of DER and OER).[15,22] Finally, FP and FN volumes were determined for the results of Experiment 1.

3. Results

An overview of the optimal configurations derived in the training phase of both experiments is provided in Table 4. Figure 2 displays a typical example of FLAIR image, the corresponding manual reference segmentation and the corresponding automated segmentation results.

The voxelwise intra-rater variability was SI=0.73±0.11 (mean±SD) when comparing the first and second segmentations and 0.75±0.11 when comparing the segmentations on the first and second marking of the lesions.

3.1. Volumetric agreement

The average lesion volume measured by the manual rater was 4.85±7.29 mL, which is in the expected range for an MS cohort with the given demographic and clinical features (see also Table 1). Scatterplots of the lesion volumes with correlation coefficients found by the automated methods versus the volumes of the manual outlined lesions are shown in Figure 3. The mean volumetric difference and ICC between the automated segmentation results and the manual reference segmentations are presented in Table 5.

In Experiment 1, Cascade, LPA and kNN-TTP showed the smallest deviation from the manual reference in terms of lesion volume (mean difference: 0.60±4.83, 0.44±4.68 and-1.39±4.10 mL, respectively), compared to LGA and Lesion-TOADS (mean difference: 2.31±7.66 mL and 1.76±4.17, respectively). Figure 4 shows boxplots of this volume differences for both experiments and each center.

In Experiment 2, with testing on new centers, LPA showed on average the largest differences: larger deviations and lower ICC were observed (see Table 5) compared to Experiment 1. The volumetric performance of Cascade, LGA and kNN-TTP was less affected by the different training. Because Lesion-TOADS does not use training or optimization, results for Experiment 2 were identical to those for Experiment 1.

3.2. Spatial agreement

The average SI, DER, OER and FN and FP volumes of the automated segmentation methods compared to the manual reference are presented in Table 5. Figure 5 displays boxplots of the SI for each experiment and center.

In Experiment 1, higher average SIs were found for LPA (0.37±0.23), Lesion-TOADS (0.35±0.189) and kNN-TTP (0.44±0.14) than for Cascade (0.26±0.17) and LGA (0.31±0.23). The OER was relatively constant between the methods, whereas the DER showed more variation (see Table 5).

Compared to Experiment 1, Experiment 2 did –on average– not reveal differences in spatial performance of any method. The small overall reduction in average spatial performance of Cascade, LGA, LPA and kNN-TTP in Experiment 2 does not imply that the performance always degrades when segmenting 'unseen' data. As can be seen from Figure 5, depending on the method, the performance on data from individual centers may actually also improve when using only data from other centers in the training set, see also Figure 4.

3.3. Additional analyses

Table 6 displays the SI for each method for different ranges of the reference lesion volume. In general, independently of the method, an increase in average SI can be observed with increasing lesion volume.

The spatial agreement for the different training set selection schemes is shown in Table 7. Independently of the training set selection scheme, all methods showed comparable performance.

Table 8 presents the estimated sensitivity and specificity of the methods relative to the hidden truth computed by the STAPLE algorithm. On average, LPA, Lesion-TOADS and kNN-TTP showed higher estimated sensitivities and specificities than the two other methods.

4. Discussion

In this study we directly compared five research-domain automated WM lesion segmentation methods in a multi-center MS dataset, to obtain quantitative results on their volumetric and spatial performance in a multi-center dataset. Accurate and robust segmentation of WM lesions would be beneficial for clinical trials in which lesion volumes are used as a (secondary) study outcome and studies on accurately measuring the GM atrophy.[23,24] Our results show performance differences between methods, with LPA, Lesion-TOADS and kNN-TTP overall exhibiting best performance, followed by Cascade and LGA. However, all methods were robust when applied to data of a new ('unseen') center, exhibiting very similar performance with and without inclusion of that center in the training and only using data of other centers. An interesting follow up study to improve the applications of the automated WM lesion segmentation methods in clinical setting, could be finding the optimal amount of centers or number of patients per center that should be used for the training.

The systematic comparison in this study showed a good performance of LPA, Lesion-TOADS and kNN-TTP in terms of both volumetric and spatial performance. Cascade was comparable to these

for volumetric but not spatial agreement, while LGA showed lower volumetric and spatial agreement compared to LPA, Lesion-TOADS and kNN-TTP. Nevertheless, the spatial agreement of the methods with the manual reference segmentation in the current study (mean SI range [0.27-0.45]) was lower than commonly reported by earlier studies.[15,25-27] To be able to compare the results of those study to the data of our study, we also report means and standard deviations. Because our data were mostly not normally distributed, the median, 1st and 3rd quartile are reported as Supplementary Table 10. As the boxplots do show that the performance of the methods stays comparable when reporting the median instead of the mean, we decided for better comparison of this study to other studies to report the mean.

The lower spatial agreement found in the current work could result firstly reside from the fact that this study used a multi-center data set, which is by definition more heterogeneous than data from a single center and acquisition protocol. This is confirmed by another recent study on lesion segmentation in a multi-center data set, reporting similar (low) performances (range of SI=0.11-0.45) compared to the current study.[28] The use of a multi-center data set is a strength of this study, providing a better approximation of the clinical setting, and moreover allowing us to show that the automated segmentation methods are robust to new data of an 'unseen' center. Summarizing, the four methods are robust, however, not accurate enough on spatial agreement. The relatively low spatial agreement may secondly be explained by the use of 2D FLAIR images instead of 3D FLAIR, because other studies that reported a higher SI used 3D FLAIR. [15,29] A higher SI with 3D FLAIR is probably due to the (normally) better resolution of the 3D images compared to 2D sequences, which results in more precise distinction of lesion boundaries and can also improve image co-registration. The use of 2D sequences may be considered a limitation of this study, because of the recent expert opinion guidelines and recommendation [30-32] and also because it is not possible to investigate cortical

lesions from this type of sequence. However, given our research question, aiming to accommodate the lack of multi-center validations of automated MS WM lesion segmentation techniques, this study was performed with data that is representative for the clinical practice. A follow up study could benefit from using data that is even more representative for the current clinical practice, so not only use 2D and 3D FLAIR images, 2 scanners and one field strength, but also include more different scanner or even scanners with both 1.5T and 3T field strengths.

Another explanation for the relatively low spatial agreement in our study may be sought in the composition of the data set. The patients participating in the current study had on average a relatively low lesion burden (according to the manual rater: 4.85±7.29 mL), with even a few patients with extremely low burden. In two of those cases, none of the automated methods agreed with the manual reference on the location of the lesions, leading to zero overlap. Moreover, especially LGA produced more segmentations without agreement with the manual reference, resulting in a lower average SI and varying results for specific centers (i.e., center F).

In addition, the training set selection scheme still can play a crucial role in the low spatial agreement. In this study we constructed the training set by selecting the subjects with the median, second lowest and second highest lesion loads for each center reasoning that these subjects properly reflect the data contributed by each center. This also ensured that each center was equally represented in the parameter tuning irrespective of the total number of subjects per center. However, this selection scheme may have had consequences for the final segmentation depending on the classifier within each method. To rule out that this had an effect we did an additional experiment using the patients with the three lowest and three highest lesion volumes for parameter tuning. This did not reveal large differences in performance for the methods evaluated in this study, but future studies should investigate the influence of training dataset composition more exhaustively. For example.

future studies could have a larger test set, because in our additional experiment the test set had only 25 subjects. Such future work could also study optimum numbers and compositions for the training set, whether or not this should depend on the number of subjects for each center. Moreover, the lesion volumes could have more variation .

The generally low spatial agreement was driven by both high DER and high OER, so both error in the detections of lesions as in outlining of the lesions, although the substantially lower performance of Cascade and LGA can be attributed in particular to a higher DER. Cascade and LGA had higher FP volumes than the other two methods.

Concerning false negative voxels, all automated methods had around 2 mL average FN lesion volume. This may derive from the limitation that this study only used one single expert rater to construct the manual reference dataset, the rater possibly being too conservative in detecting and outlining the WM lesions. To address this issue, we took an alternative approach and compared the performance of the automated segmentations with the data-driven STAPLE algorithm. This algorithm does not require manual data, but estimates the sensitivity and specificity of the different methods by comparing to the 'hidden true segmentation' that is constructed by the STAPLE algorithm using the four automated segmentations. This analysis showed similar results as when comparing to the manual reference, with LPA, Lesion-TOADS and kNN-TTP showing better performance than Cascade and LGA. This confirms the observation from the manual reference comparison that Lesion-TOADS and kNN-TTP had best performance in terms of spatial agreement. Future studies should try to improve construction of the manual segmentations, possibly by using intelligent data-driven approaches or a larger number of manual raters, of which the agreement in terms of lesion detection and outline error can be quantified.

The results of this study also shows that improvements in terms of volumetric agreement are necessary, as the average volumetric difference between the manual reference and automated methods ranged from 0.44±4.68 (kNN-TTP) to 2.31±7.66 mL (LGA). When visualily inspecting the outlines and MR images of these outliers we concluded that this was mainly due to GM atrophy in these subjects. By comparison, the average lesion load changes occurring in MS patients over a 1-year period in clinical studies are in ranges of 0.38±1.81,[3] 0.11±2.55,[5] and 0.08±0.97 mL.[6] The large inter-subject variability of the volumetric differences within each method indirectly suggests that the errors are not entirely systematic, and therefore these errors are a concern for detecting the lesion volume changes over time. Future studies should investigate performance at quantifying lesion volume change over time using suitable datasets and with high-quality reference measurements. Further studies could help optimize performance by investigating the dependence of performance on lesion location and on total lesion load. Moreover, a follow up study could look with more detail to different lesions types and locations.

In conclusion, the five research-domain automated WM lesion segmentation methods exhibited moderate accuracy in this multi-center dataset consisting of MS patients with moderate lesion burden. All methods exhibited relatively low performance compared to other studies. LPA, Lesion-TOADS and kNN-TTP outperformed Cascade and LGA in terms of both volumetric and spatial agreement. The performance of the method was not affected by using ‘unseen’ data for training, which is important for the acceptance of automated methods in a clinical setting. However, the methods should perform better before their use in a clinical setting can be realistically considered. Furthermore, better understanding of which parameters influence the performance of the segmentation methods is required for better interpreting their performance.

References:

[1] Benedict R, and Bobholz J. Multiple sclerosis. Semin Neurol 2007;27:78-85

[2] Lucchinetti C, Brück W, Parisi J, Scheithauer B, Rodriguez M, Lassmann H. Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. Ann Neurol 2000;47:707-17

[3] Calabresi PA, Radue EW, Goodin D et al. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. Lancet Neurol 2014;13:545-56

[4] Kappos L, Radue EW, O'Connor P, et al. A Placebo-Controlled Trial of Oral Fingolimod in Relapsing Multiple Sclerosis. N Engl J Med 2010;362:387-401

[5] Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol 2011;69:292-302

[6] Radue EW, Barkhof F, Kappos L, et al. Correlation between brain volume loss and clinical and MRI outcomes in multiple sclerosis. Neurology. 2015;84:784-93

[7] Grimaud J, Lai M, Thorpe J, et al. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn Reson Imaging 1996;14:495-505

[8] Paty DW, Issac CD., Grochowslci E, et al. Magnetic resonance imaging (MRI) in multiple sclerosis(MS): A serial study in relapsing and remitting patients with quantitative measurements of lesions size abstract. Neurology 1986;36:177

[9] Mortazavi D, Kouzani AZ, Soltanian-Zadeh H. Segmentation of multiple sclerosis lesions in MR images: a review. Neuroradiology 2012;54:299-320

[10] Damangir S, Manzouri A, Oppedal K, et al. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. J Neurol Sci 2012;322:211-6

[11] Damangir, S., Westman, E., Simmons, A. et al. Magnetic Resonance Materials in Physics, Biology and Medicine 2016;doi:10.1007/s10334-016-0599-3

[12] Schmidt P, Gaser C, Arsic M, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. Neuroimage 2012;59:3774-83

[13] Schmidt P. . Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. PhD thesis, LudwigMaximilians-Universität München, Januar 2017. URL http://nbn-resolving. de/urn:nbn:de:bvb:19-203731.

[14] Shiee N, Bazin PL, Ozturk A, Reich DS, Calabresi PA, Pham DL. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage 2010;49:1524-35

[15] Steenwijk MD, Pouwels PJ, Daams M, et al. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). Neuroimage Clinical 2013;3:462-9

[16] Ropele S, Kilsdonk ID, Wattjes MP, et al. Determinants of iron accumulation in deep grey matter of multiple sclerosis patients. Multiple Sclerosis J. 2014;20:16928.

[17] Kurtzke J.K. Rating Neurologic Impairment in Multiple Sclerosis: An Expanded Disability Status Scale (EDSS). Neurology 1983;33:1444-52

[18] Plummer D. DispImage: a display and analysis tool for medical images. Riv Neuroradiol 1992;5:488–95

[19] Warfield SK, Zou KH, Wells WM, et al. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2014;23:903-21

[20] Koch GG, Intraclass correlation coefficient. In: Samuel K. and Norman L.J. Encyclopedia of Statistical Sciences 4. New York, John Wiley & Sons; 1982;213–17

[21] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297-302

[22] Wack DS, Dwyer MG, Bergsland N, et al. Improved assessment of multiple sclerosis lesion segmentation agreement via detection error and outline error estimates. BMC Med Imaging 2012;12:12-7

[23] Amiri H, de Sitter A, Bendfeldt K, et al. Urgent challenges in quantification and interpretation of grey matter atrophy in multiple sclerosis. *Under review*

[24] Rocca MA, Battaglini M, Benedict RH, et al. Brain MRI atrophy quantification in MS: From methods to clinical application. Neurology 2017; 88:403-13

[25] Admiraal-Behloul F, Van Den Heuvel DMJ, Olofsen H, et al. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. Neuroimage 2005;28:607-17.

[26] Anbeek P, Vincken KL, van Osch MJ, Bisschops RH, Van der Grond J. Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage 2004;21:1037-44

[27] Khayati R, Vafadust M, Towhidkhah F, Nabavi M. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model. Comput Biol Med 2008;38:379-90

[28] Roura E, Oliver A, Cabezas M, et al. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology 2015;57:1-13

[29] Anbeek, P., Vincken, K.L., van Bochove, G.S., van Osch, M.J.P., van der Grond, J. Probabilistic segmentation of brain tissue in MR imaging. Neuroimage 2005;27:795–804.

[30] Rovira À, Wattjes MP, Tintoré M, et al. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis [mdash] clinical implementation in the diagnostic process. Nat Rev Neurol 2015;11:471-82

[31] Vrenken H, Jenkinson M, Horsfield MA, et al. Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis. J Neurol 2013;260:2458-71.

[32] Wattjes MP, Rovira À, Miller D, et al. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis [mdash] establishing disease prognosis and monitoring patients. Nat Rev Neurol 2015;11;597-606

Tables

Table 1 Demographics of MS patients[a]

Center	N[b]	Age years	Disease types	DD years	EDSS[c]
A	14 [86%]	37.6±7.2	1 CIS, 13 RR	11.5±8.6	2.5 (0.0-6.5)
B	11 [73%]	28.1±8.6	11 RR	7.6±4.8	2.5 (1.0-4.0)
C	10 [80%]	35.2±10.5	10 RR	3.6±1.3	2.0 (0.0-6.5)
D	14 [64%]	35.6±10.0	14 RR	5.3±3.3	1.5 (0.0-3.0)
E	11 [80%]	34.2±3.6	11 RR	7.5±4.4	2.5 (1.0-3.5)
F	10 [10%]	38.4±9.4	6 RR, 4 SP	9.2±7.2	4.0 (1.5-6.5)
Total	70 [67%]	34.9±8.5	1 CIS, 65 RR, 4 SP	7.6±6.0	2.0 (0.0-6.5)

[a]Mean±standard deviation
[b]Number of subjects [%female]
[c]Median (range)
Abbreviations: DD=disease duration, EDSS=expanded disability status scale, CIS = clinically isolated syndrome RR=relapsing-remitting, SP=secondary-progressive

Table 2 An overview of the acquisition parameters for each center

Center	Scanner brand, scanner type		Sequence parameters 2D FLAIR/3DT1 TR (ms)	TE (ms)	TI (ms)	FA (^{0})	Voxel size (mm^3)
A	Siemens, Trio	2D FLAIR	9000	93	2500	-	256x256 (0.94x0.94x3)
		3DT1	2300	2.98	900	9	182x240 (1x1x1)
B	Siemens, Trio	2D FLAIR	9000	136	2500	-	320x320 (0.75x0.75x3)
		3DT1	1570	2.76	900	9	160x256 (1x1x1)
C	Philips, Achieva	2D FLAIR	11000	125	2600	-	256x256(0.94x0.94x3)
		3DT1	6.9	2.78.	831	9	160x240(1x1x1)
D	Siemens, Trio	2D FLAIR	10000	69	2500	-	256x256 (0.94x0.94x3)
		3DT1	1900	2.19	900	9	176x448 (1x1x1)
E	Philips, Achieva	2D FLAIR	8000	125	2400	-	240x180 (1x1x3)
		3DT1	6.9	3.1	831	8	256x256 (1x1x1)
F	Philips, Achieva	2D FLAIR	11000	120	2800	-	256x256 (0.94x0.94x3)
		3DT1	8.3	3.7	815	8	182x240 (1x1x1)

Abbreviations: FLAIR=fluid-attenuated inversion recovery, TR=repetition time, TE=echo time, TI=inversion time, FA=flip angle;

Table 3 Description of the methods, parameter tuning and pre- and post-processing steps of automated segmentation methods

Method	Reference	Underlying method	Parameter tuning	Pre-processing	Post-processing
Cascade	Damangir, 2012	SVM; Statistical definition of WML	Yes; α and LPV threshold	Brain extraction on 3DT1 3DT1 to FLAIR registration Brain mask applied on FLAIR	Not needed
LGA	Schmidt, 2012	Algorithm based	Yes; κ and LPT	Generate PVE label image coregistration FLAIR to 3DT1 registration TPMWM to 3DT1	Needed; LP maps to FLAIR registration
LPA	Schmidt; 2017	Algorithm based	Yes; LPT	Generate PVE label image coregistration FLAIR to 3DT1 registration TPMWM to 3DT1	Needed; see LGA
Lesion-TOADS	Shiee, 2010	Topological and Statistical atlas based	None	Brain extraction on 3DT1 3DT1 to FLAIR registration Brain mask applied on FLAIR	Not needed
kNN-TTP	Steenwijk, 2013	kNN classification	Yes; LPT	Brain extraction on 3DT1 3DT1 to FLAIR registration Brain mask applied on FLAIR	Not needed

Abbreviations: SVM=support vector machine, PVE=partial volume estimate, TPMWM=white matter tissue probability map, FLAIR=fluid-attenuated inversions recovery, LPV=lesion pvalue, LPT=lesion probability threshold, LP=lesion probability

Table 4 Optimal setting for threshold

Method	Parameter	Training on all centers	Center excluded from training A	B	C	D	E	F
Cascade	α	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	Lesion-pvalue	0.05	0.05	0.05	0.05	0.05	0.15	0.05
LGA	κ	0.15	0.20	0.20	0.15	0.15	0.20	0.20
	Lesion-probability	0.85	0.75	0.65	0.85	0.75	0.75	0.75
LPA	Lesion-probability	0.55	0.55	0.55	0.60	0.55	0.55	0.55
Lesion-TOADS	None							
kNN-TTP	Lesion-probability	0.55	0.55	0.50	0.55	0.65	0.55	0.55

Table 5 Volumetric and Spatial agreement of the automated methods with the manual reference[a]

	Test set	Method	Volume[b]	Volume difference[b]	ICC	SI	DER	OER	FN Volume[b]	FP Volume[b]
		Manual reference	4.85±7.29							
Experiment 1[c]	**N=52**	Cascade	5.45±6.47	0.60±4.83	0.755	0.26±0.17	0.98±0.55	0.50±0.35	2.28±3.04	3.64±3.44
		LGA	7.16±12.70	2.31±7.66	0.713	0.31±0.23	0.83±0.70	0.56±0.34	2.70±4.73	2.93±10.03
		LPA	5.30±9.63	0.44±4.68	0.851	0.37±0.23	0.59±0.45	0.68±0.27	1.71±2.44	2.92±1.82
		Lesion-TOADS	7.45±7.97	1.76±4.17	0.806	0.35±0.18	0.65±0.50	0.64±0.34	1.36±1.75	4.73±4.11
		kNN-TTP	5.92 ±6.42	-1.39±4.10	0.723	0.44±0.14	0.49±0.38	0.62±0.24	2.30±4.01	1.51±1.79
Experiment 2[c]	**N=52**	Cascade	5.44±6.48	0.75±4.56	0.774	0.23±0.18	0.96±0.57	0.58±0.47	2.26±3.05	4.11±8.32
		LGA	8.15±17.06	3.45±12.12	0.550	0.32±0.23	0.82±0.70	0.55±0.33	2.41±3.56	5.67±14.21
		LPA	5.16±9.65	0.46±5.25	0.809	0.36±0.17	0.59±0.45	0.68±0.32	1.72±2.51	2.89±4.83
		Lesion-TOADS	7.45±7.97	2.60±4.18	0.806	0.35±0.18	0.66±0.50	0.64±0.34	1.36±1.75	4.73±4.11
		kNN-TTP	3.32±4.17	-1.53±4.26	0.738	0.43±0.14	0.49±0.38	0.62±0.24	2.28±4.00	1.15±1.68

[a]Mean±standard deviation

[b]Volume in mL

[c]Centers were combined in experiment 2

Abbreviations: N=number of subjects, ICC=intra-class correlation coefficient, SI=similarity index, DER=detection error rate, OER=outline error rate, FN =false negative, FP=false positive

Table 6: SI for different lesion volumes[a]

Volume	Test set	Method	SI
<5mL	N=34	Cascade	0.21±0.14
		LGA	0.28±0.22
		LPA	0.32±0.16
		Lesion-TOADS	0.29±0.15
		kNN-TTP	0.42±0.14
5-15mL	N=12	Cascade	0.36±0.17
		LGA	0.33±0.27
		LPA	0.49±0.07
		Lesion-TOADS	0.48±0.13
		kNN-TTP	0.49±0.11
>15mL	N=6	Cascade	0.50±0.14
		LGA	0.45±0.23
		LPA	0.46±0.19
		Lesion-TOADS	0.63±0.08
		kNN-TTP	0.50±0.18

[a]Mean±standard deviation
Abbreviations: N=number of subjects, SI=similarity index.

Table 7: SI of the original test set and obtained in only the patients present in the test sets for different used training sets[a]

Training set	Original (N=52)	A (N=28)	B (N=28)	C (N=28)
Method				
Cascade	0.26±0.17	0.33±0.12	0.32±0.15	0.31±0.16
LGA	0.31±0.23	0.36±0.22	0.35±0.22	0.35±0.22
LPA	0.37±0.17	0.44±0.13	0.44±0.14	0.43±0.13
Lesion-TOADS	0.35±0.18	0.44±0.16	0.44±0.16	0.44±0.16
kNN-TTP	0.42±0.17	0.50±0.11	0.51±0.11	0.49±0.13

[a]Mean±standard deviation
Abbreviations: N=number of subjects, SI=similarity index.
Training sets: for each center for original and A the patients with the median, second minimum and second maximum lesion volumes, for B the three patients with lowest lesion volume and for C the three patients with highest lesion leas.

Table 8: Estimated sensitivity and specificity compared to data-driven reference

Method	**Test set (N=52)**	**Estimated sensitivity**	**Estimated specificity**
Cascade		0.55±0.27	0.38±0.31
LGA		0.69±0.22	0.56±0.37
LPA		0.81±0.13	0.78±0.23
Lesion-TOADS		0.81±0.24	0.61±0.25
KNN-TTP		0.75±0.13	0.72±0.24

[a]Mean±standard deviation
Abbreviations: N=number of subjects

Figures

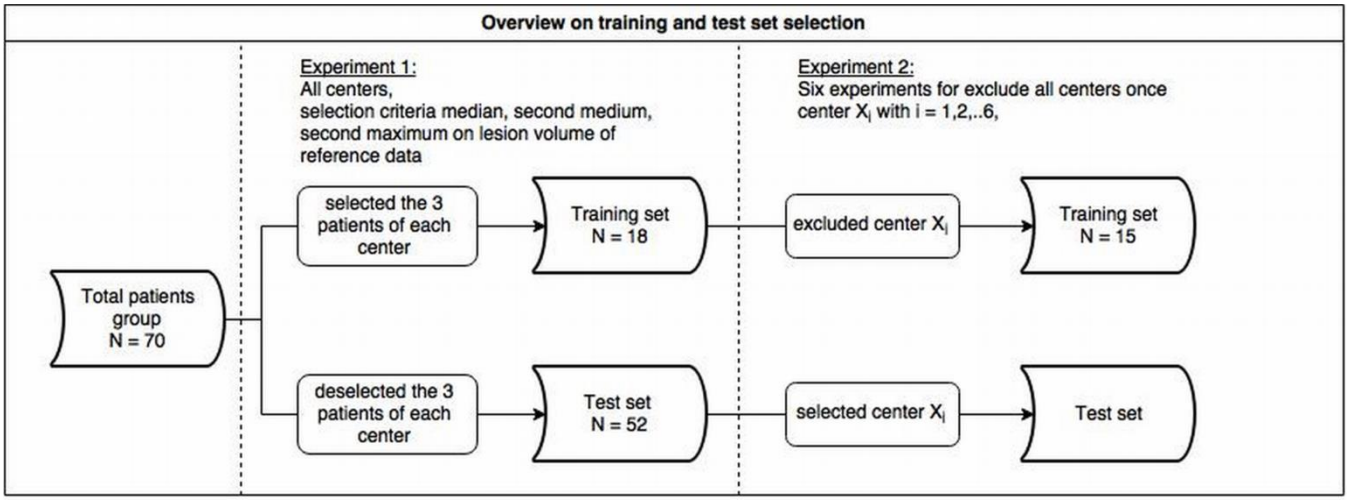

Figure 1: Flowchart of selection method for training and test set. The number of patients in test set when select center Xi, depends on the selected centers. Abbreviations: N=number of subjects.

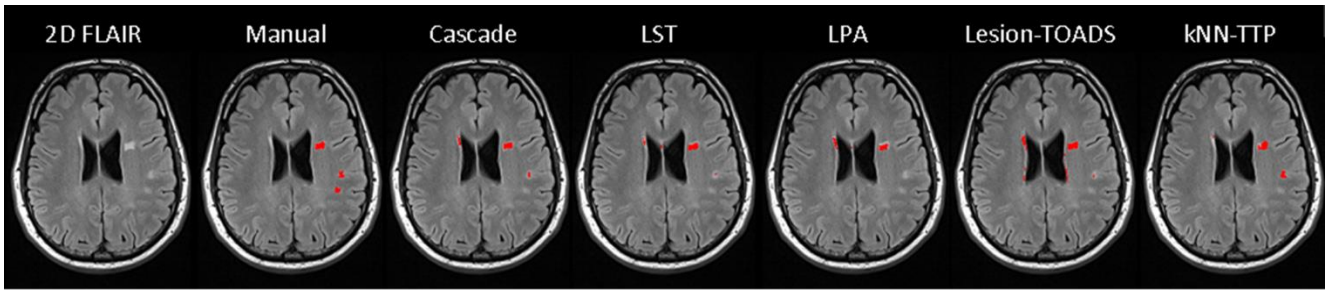

Figure 2: A FLAIR images is shown (first from left) with the lesion masks of the manual reference (second from left) and the 4 segmentation methods (images 3 to 6 from left). Abbreviations: FLAIR=FLuid-Attenuated Inversion Recovery

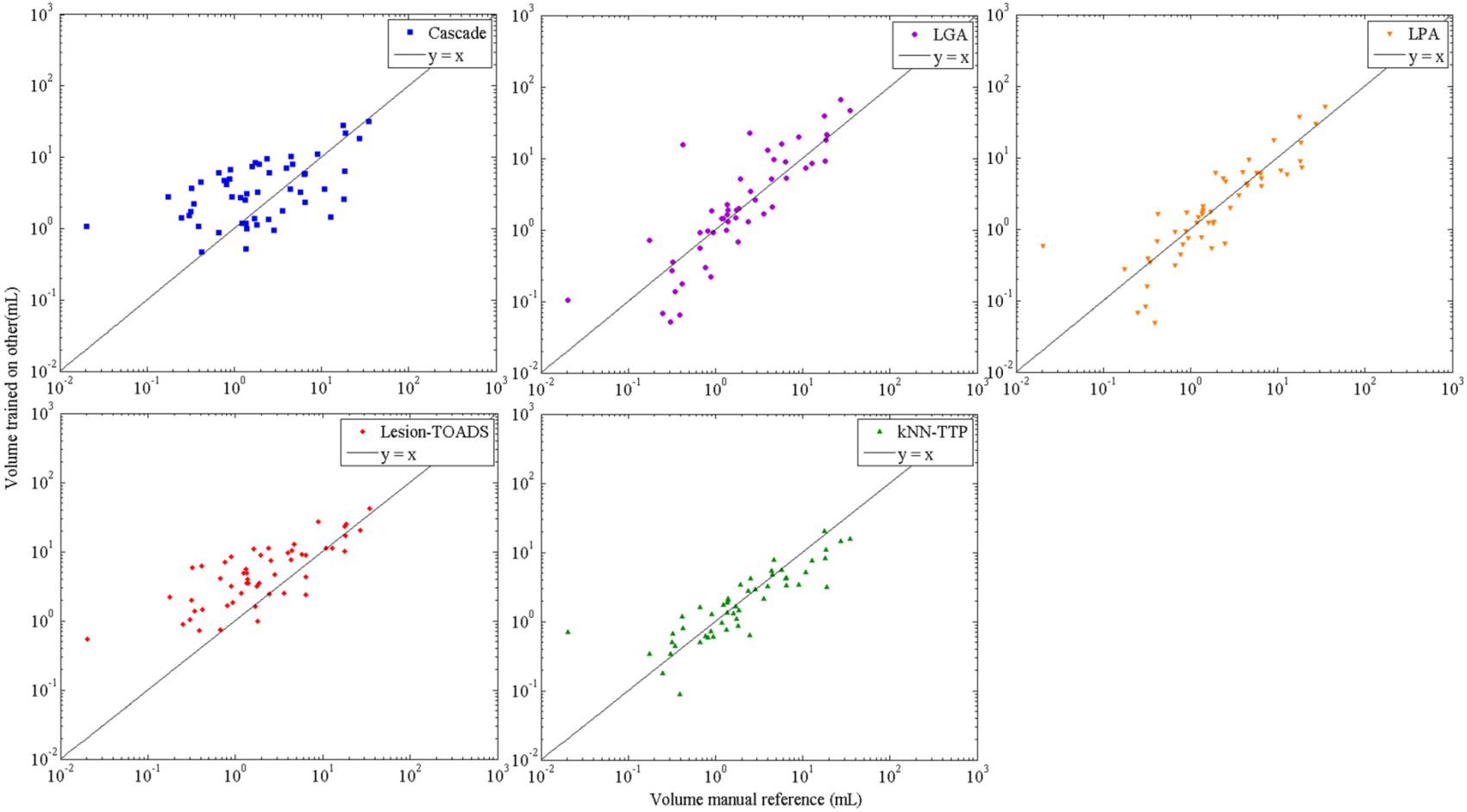

Figure 3: Lesion volume between automated segmentations (left to right, up to down; Cascade, LGA, LPA, Lesion-TOADS and kNN-TTP) against the lesion volume of the manual reference.

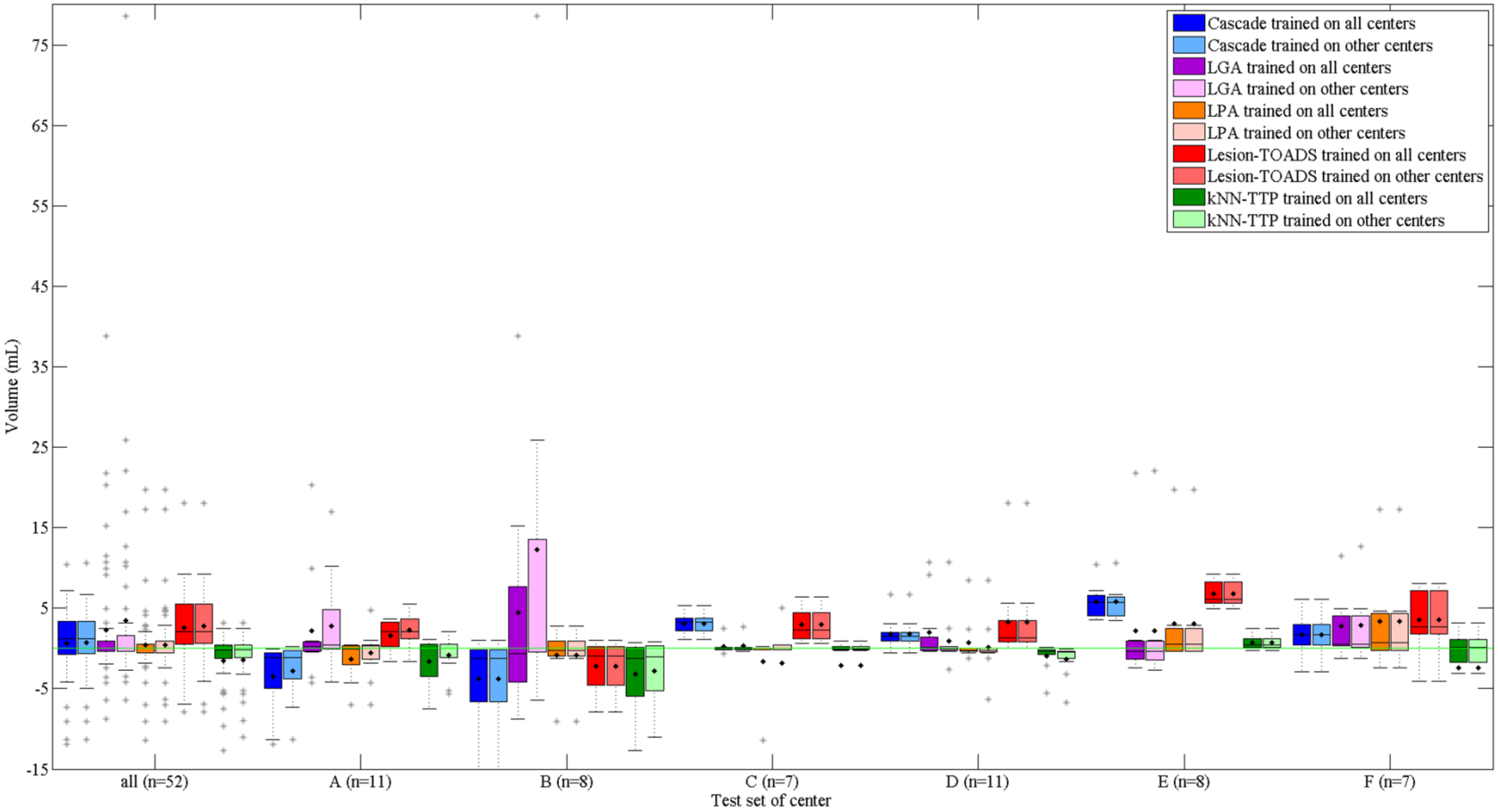

Figure 4: Boxplot of difference in lesion volume between automated segmentations (left to right; Cascade trained on all centers and Cascade trained on other centers, LGA trained on all centers and LGA trained on other centers, LPA trained on all centers and LPA trained on other centers, Lesion-TOADS trained on all centers and Lesion-TOADS trained on other centers and kNN-TTP trained on all centers and kNN-TTP trained on other centers) and the manual reference for the total group and separate centers (left to right: all, center A, B .. F), with star = outliner, diamond = mean of difference, green line is zero.

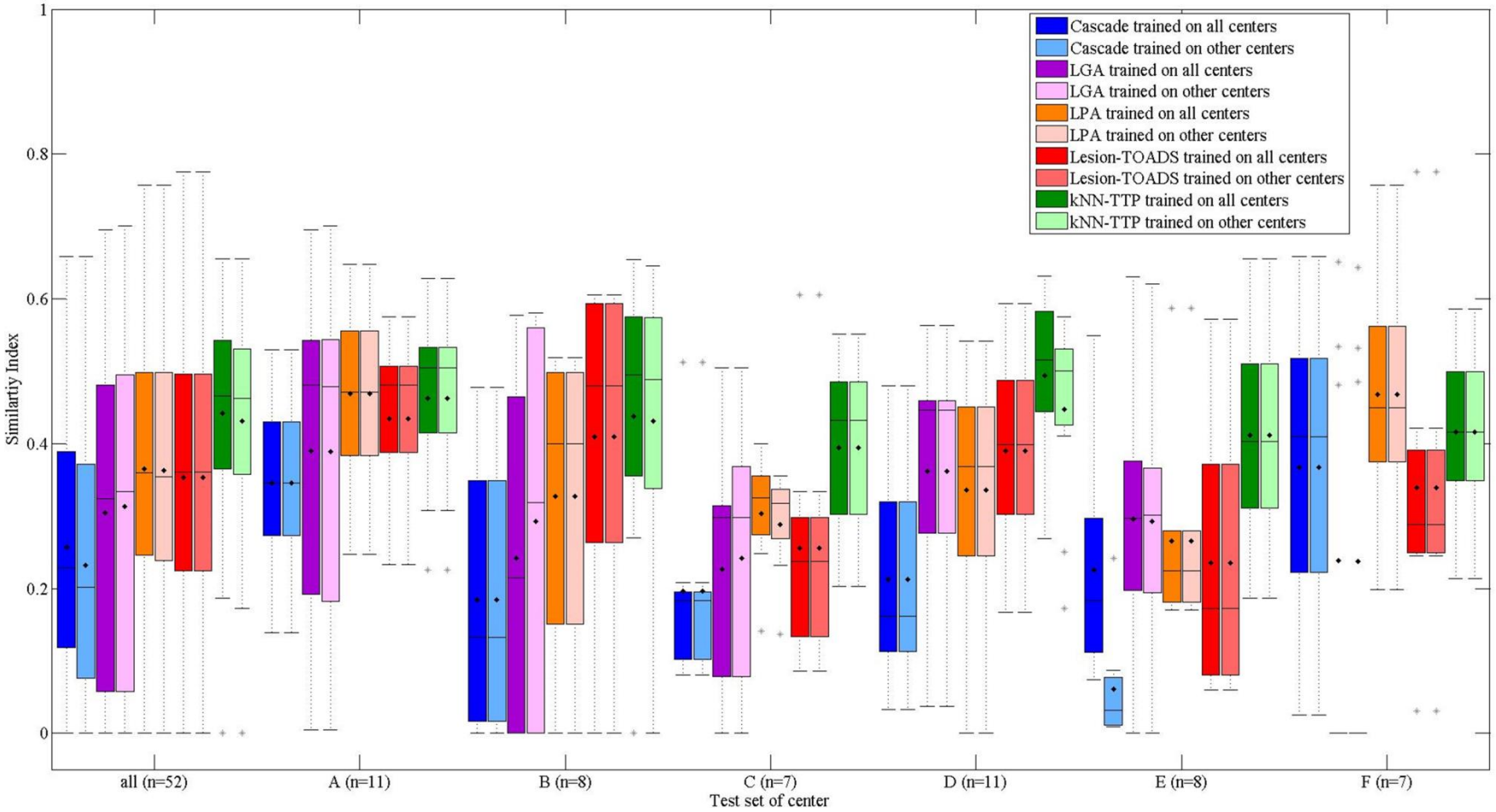

Figure 5: Boxplot of the Similarity Index (left to right; Cascade trained on all centers and Cascade trained on other centers, LGA trained on all centers and LGA trained on other centers, LPA trained on all centers and LPA trained on other centers, Lesion-TOADS trained on all centers and Lesion-TOADS trained on other centers and kNN-TTP trained on all centers and kNN-TTP trained on other centers) and the manual reference for the total group and separate centers (left to right: all, center A, B .. F), with star=outliner, diamond=mean.