NeuroImage 244 (2021) 118589

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset

Olivier Commowick[a,*], Michaël Kain[a], Romain Casey[b], Roxana Ameli[b], Jean-Christophe Ferré[a,c], Anne Kerbrat[d], Thomas Tourdias[e], Frédéric Cervenansky[f], Sorina Camarasu-Pop[f], Tristan Glatard[g], Sandra Vukusic[b], Gilles Edan[a,d], Christian Barillot[a], Michel Dojat[h], Francois Cotton[b]

[a] Univ Rennes, Inria, CNRS, Inserm - IRISA UMR 6074, Empenn ERL U1228, Rennes F-35000, France
[b] Department of Radiology, Lyon Sud Hospital, Hospices Civils de Lyon, Lyon, France
[c] Department of Neuroradiology, CHU Rennes, Rennes F-35033, France
[d] Department of Neurology, CHU Rennes, Rennes F-35033, France
[e] CHU de Bordeaux, Service de Neuro-Imagerie, Bordeaux, France
[f] Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, Lyon U1206, F-69621, France
[g] Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
[h] Inserm U1216, University Grenoble Alpes, CHU Grenoble, GIN, Grenoble, France

A B S T R A C T

MRI plays a crucial role in multiple sclerosis diagnostic and patient follow-up. In particular, the delineation of T2-FLAIR hyperintense lesions is crucial although mostly performed manually - a tedious task. Many methods have thus been proposed to automate this task. However, sufficiently large datasets with a thorough expert manual segmentation are still lacking to evaluate these methods. We present a unique dataset for MS lesions segmentation evaluation. It consists of 53 patients acquired on 4 different scanners with a harmonized protocol. Hyperintense lesions on FLAIR were manually delineated on each patient by 7 experts with control on T2 sequence, and gathered in a consensus segmentation for evaluation. We provide raw and preprocessed data and a split of the dataset into training and testing data, the latter including data from a scanner not present in the training dataset. We strongly believe that this dataset will become a reference in MS lesions segmentation evaluation, allowing to evaluate many aspects: evaluation of performance on unseen scanner, comparison to individual experts performance, comparison to other challengers who already used this dataset, etc.

1. Introduction

Multiple Sclerosis (MS) is a neuroinflammatory disease of the central nervous system. It affects 2.5 million persons worldwide, particularly in the northern hemisphere or highly developed countries. Its prevalence average rate is around 83 per 100,000, with higher rates in countries of the northern hemisphere. MS affects preferentially women (woman:man ratio of around 2.0) (Pugliatti et al., 2006). The disease is characterized by a widespread inflammation, focal demyelination, and a variable degree of axonal loss both in the brain and spinal cord, and as such has become one of the major causes of acquired handicap.

New disease modifying drugs have recently appeared that are able to slow down the disease evolution (Giovannoni et al., 2010; Hauser et al., 2017; Kappos et al., 2018; Polman et al., 2006). One of the major challenges in treating MS is now to obtain sensitive and specific criteria to obtain an early diagnosis, prognosis and prediction of the pathology status for a patient, earlier at least than classical clinical criteria, such as the expanded disability status scale (EDSS) (Kurtzke, 1983). Magnetic Resonance Imaging (MRI) plays an important role for the diagnosis (Polman et al., 2011; Thompson et al., 2018) and its role for the evaluation of disease evolution is growing. It even allows to provide insights in the highly variable nature of the MS disease course (Leray et al., 2010), thus allowing in the long term to adapt the treatment to each individual.

Among the criteria available from MRI, the number and spread of lesions in the brain, as well as their evolution, have become crucial markers of the patient's disease status (Polman et al., 2011; Thompson et al., 2018). Counting these lesions and measuring their sizes is however requiring a very tedious and time consuming task for the neuroradiologist i.e. the manual delineation of MS lesions. In addition, manual delineation is also prone to inter-expert variability. This is especially true when comparing radiologists from different centers (as practices towards delineation may vary between centers or formation centers - a problem sometimes refered to as "segmentation schools") or when analyzing images from different centers where the protocols, scanners

* Corresponding author.
E-mail address: olivier.commowick@inria.fr (O. Commowick).
URL: https://olivier.commowick.org (O. Commowick)

https://doi.org/10.1016/j.neuroimage.2021.118589.
Received 27 April 2021; Received in revised form 3 September 2021; Accepted 16 September 2021
Available online 24 September 2021.

Table 1
Demographics data of multiple sclerosis patients collected and their division among training and testing datasets.

Center	Scanner model and site	Training cases	Testing cases	Age (y.o.)	Gender ratio W:M
01	Siemens Verio 3T (University Hospital of Rennes)	5	10	43.6 ± 12.6	2.75
03	General Electrics Discovery 3T (University Hospital of Bordeaux)	0	8	48.9 ± 11.5	7
07	Siemens Aera 1.5T (University Hospital of Lyon)	5	10	45.3 ± 9.8	4
08	Philips Ingenia 3T (University Hospital of Lyon)	5	10	45.5 ± 7.8	1.14

or field strengths differ (giving rise to different intrinsic MRI quality and signal to noise ratios). Performing manual segmentation on large databases of patients is therefore almost impossible, although required to analyze the disease variants in the population. Automatic segmentation algorithms have therefore become a crucial need for the clinical community to simplify the clinician's task. A large literature of automatic segmentation methods has been devised (García-Lorenzo et al., 2013; Lladó et al., 2012; Mortazavi et al., 2012) for which a common ground for their evaluation is more and more required.

This evaluation is performed using databases where manual delineation was performed by one or several expert radiologists. A vast majority of the published approaches are however evaluated on in-house datasets with different image characteristics. The results obtained by different methods are then not directly comparable, making it difficult to choose a method adapted to a specific clinical context. This adds up to the fact that inter-expert variability in the manual segmentations used in the validation datasets is poorly known and may bias the results of a given evaluation. To overcome one or both of these issues, competitions (also called challenges) have been organised over the years in the MS lesion segmentation community. The first one was organized at the MICCAI 2008 conference (Styner et al., 2008). It came with a database of 45 patient images (from two different centers: 20 for training and 25 for testing), with a ground truth composed of two expert segmentations for each case. Having only two raters did not allow getting a real idea of inter-expert variability, which was actually potentially high since the protocols were not harmonized between the two sites of acquisition. The second challenge was held at the 2015 IEEE ISBI international conference (Carass et al., 2017). The database was more focused on longitudinal evolution of the lesions and the database was thus composed of five patients images each with an average of 4.4 time points, each time point being manually delineated by two experts. Again, the inter-expert variability could not really be evaluated with only two experts.

There is therefore a crucial need for an evaluation dataset with the following properties: 1- composed of a large number of patient images from different centers to evaluate image quality variability, 2- with as many as possible expert manual segmentations to characterize inter-expert variability, and 3- fully open to the community so that new methods can be evaluated on it and compared to other approaches easily. We propose in this paper such a dataset consisting of 53 patient images with each 7 manual delineations from experts. This database was used for the MICCAI 2016 challenge (Commowick et al., 2018). We present here an extended description and analysis of this database. We also make it available to the community so that validation of new methods may be conducted on the whole dataset, as well as studies on expert manual segmentation variability.

2. Materials and methods

To constitute the proposed database, three steps were conducted described in the following sub-sections. We first gathered MS patients data from different centers and scanners. Then MS lesions were manually delineated from the MRI data by seven experts, and an automatic consensus was computed from these segmentations for each patient. Finally, to provide challengers with a common ground to compare their segmentation algorithms, a common preprocessing was designed and applied to the image dataset.

2.1. MS patients database

The database of images acquired is composed of 53 multiple sclerosis patients. Data were generated by participating neurologists in the framework of Observatoire Français de la Sclérose en Plaques (OFSEP), the French MS registry (Vukusic et al., 2020). They collect clinical data prospectively in the European Database for Multiple Sclerosis (EDMUS) software (Confavreux et al., 1992). MRI of patients were provided as part of a care protocol and informed consent was provided to OFSEP by patients enrolled. Nominative data are deleted from MRI before transfer and storage on the Shanoir platform (Sharing NeurOImaging Resources).

The patient scans were performed following a harmonized protocol (Brisset et al., 2020; Cotton et al., 2015) applied in France for the constitution of OFSEP. Thanks to this protocol, the evaluation is representative of the current standards in terms of acquisition, especially in France where most centers now follow this protocol for clinical routine. More precisely, the 53 images came from three different sites in France and a total of four different MRI scanners from different manufacturers (Siemens, Philips and GE). MRI scanners included three 3T and one 1.5T magnets. The division of the patients and scanners is displayed in Table 1.

Patients in the study were aged from 24 to 61 years old, with an average age of 45.4 years old (standard deviation: 10.3 years old. The gender ratio was about 2.53 women for one man (a total of 38 women were included and 15 men). Detailed demographic details are provided as supplementary material (https://doi.org/10.5281/zenodo.5189179). No significant difference of age was present between the different centers or scanners. Gender ratio differences vary between centers as it can be seen in Table 1.

Patients in the database were selected to have variable lesion loads both in terms of volume and number. We used images from different centers and scanners to represent the variability that may be encountered across sites and manufacturers, even though a harmonized protocol is used. For each patient, MR images were acquired as detailed in Table 2. Mainly, the following images were acquired: a 3D FLAIR sequence, a 3D T1 weighted sequence pre and post-Gadolinium injection, and an axial dual PD-T2 weighted sequence, all with similar image resolutions.

Finally, to provide training data for machine learning algorithms, patients were split into two groups (see Table 1): a training and a testing datasets. We purposely excluded center 03 from the training dataset to enable the evaluation of the robustness of segmentation algorithms to cases not encountered in the design of the algorithm, and with different acquisition settings. Illustration of the contrast differences between center 03 and center 01 (similar for other centers) are shown in Fig. 1.

Table 2
Acquisition details for each sequence and each scanner for the training and testing MS patients databases.

Center	Sequence	Matrix	Slices	Voxel size (mm)	Echo time range (ms)	Repetition time range (ms)	Flip angle (degrees)	Inversion time range (ms)
01	Sag. 3D FLAIR	512x512	144	0.5x0.5x1.1	400	5000	120	1800
	Sag. 3D T1	256x256	176	1x1x1	2.26	1900	9	NA
	Ax. 2D PD-T2	240x320	44	0.69x0.69x3	PD: 9.4 T2: 84	6530	150	NA
03	Sag. 3D FLAIR	512x512	224	0.47x0.47x0.9	[140, 145]	9000	90	[2355, 2362]
	Sag. 3D T1	512x512	248	0.47x0.47x0.6	3.2	[7.5, 8]	10	NA
	Ax. 2D PD-T2	512x512	From 28 to 44	0.43x0.43x3 Gap: 0.5	PD: [8, 8.5] T2: [118, 123]	[5765, 7071]	111	NA
07	Sag. 3D FLAIR	256x224	128	1.03x1.03x1.25	336	5000	120	1800
	Sag. 3D T1	256x256	176	1.08x1.08x0.9	3.37	1860	15	NA
	Ax. 2D PD-T2	320x320	25	0.72x0.72x4 Gap: 1.2	PD: 11 T2: 100	3050	150	NA
08	Sag. 3D FLAIR	336x336	261	0.74x0.74x0.7	360	5400	90	1800
	Sag. 3D T1	336x336	200	0.74x0.74x0.85	4.3	9.4	8	NA
	Ax. 2D PD-T2	512x512	46	0.45x0.45x3	PD: 10.53 T2: 100	[5049, 5488]	90	NA

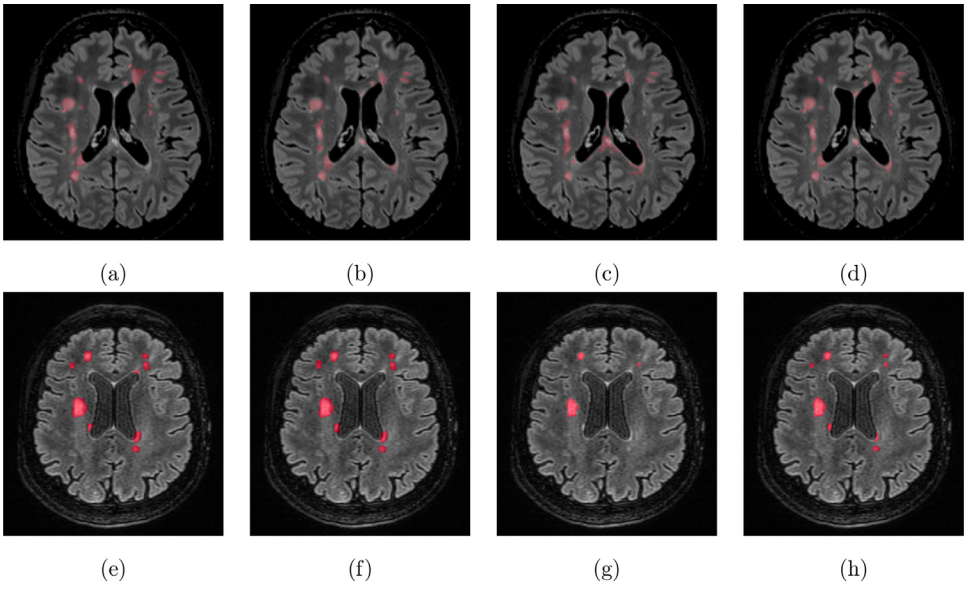

Fig. 1. Illustration of the manual delineations for two MS patients, overlaid on the 3D FLAIR image. First line: patient from center 01, second line: patient from center 03. (a-c, e-f): three out of the seven expert manual delineations of MS lesions, (d,h): consensus segmentation computed from the manual segmentations.

2.2. Lesions delineation

The dataset was then manually delineated to get a ground truth of T2/FLAIR hyperintense lesions for each patient. This task is difficult since variability exists between experts, even when they follow a common protocol both for image acquisition and delineation. This variability depends on many factors including image quality, level of expertise, sequences used for segmentation, “school” of segmentation. To obtain a reliable ground truth, we have therefore asked seven trained junior radiologists to delineate lesions in the patient scans. These experts were coming from the acquisition centers: 4 experts from Lyon, 2 experts from Rennes and one expert from Bordeaux.

The manual segmentations were performed on the 3D FLAIR image (in the following, we denote by image a full 3D volume of data) with further control on the T2 weighted image. Each manual segmentation was performed by a junior radiologist, trained under the supervision of senior neuroradiologists with a long experience in MS. More specifically, two meetings between senior radiologists and the 2016 MSSEG challenge organizers took place to determine the segmentation strategy and adopt a common tool (ITK-Snap) to perform manual segmentation. The junior radiologists were then trained by the expert radiologists on selected independent cases. After at least 5 training meetings, and when the senior expert radiologists judged that the quality of the segmentation was sufficient on the selected independent cases, the junior radiologists were allowed to delineate MS lesions on the 53 patient cases.

The following detailed rules were additionally given for segmentation. The manual segmentation had to be performed on the raw FLAIR image (no interpolation or smoothing) with control on the T2 weighted image. The most peripheral region of the lesion had to be delineated rather than an internal bound or the “heart” of the lesion. For confluent or touching lesions, lesions had to be delineated by a single contour on each slice. Lesions smaller than a threshold of 3 mm^3 were removed as they are not reliable and not included in the diagnostic criteria of the disease (Thompson et al., 2018). Some specific lesions: punctiform lesions and periventricular lesions suggestive of leukoaraiosis, were excluded of the manual segmentation process. Each case was segmented in isolation of the other cases and raters to limit possible bias. An example of the obtained manual segmentations by the individual experts is illustrated for two patients in Fig. 1.

Since even with these common guidelines experts may differ from each other, we have then constructed, from the manual segmentations of each MS patient, a consensus to be used for algorithms evaluation. This was performed using the Logarithmic Opinion Pool Based Simultaneous Truth And Performance Level Estimation (LOP STAPLE) algorithm (Akhondi-Asl et al., 2014). This method computes iteratively, using an Expectation-Maximization algorithm, a consensus segmentation based on penalties for individual deviations from agreement between manual experts segmentations. Such an approach has several advantages: 1- it is robust to differences between manual expert segmentations, and 2- it allows the computation of agreement scores of each ex-

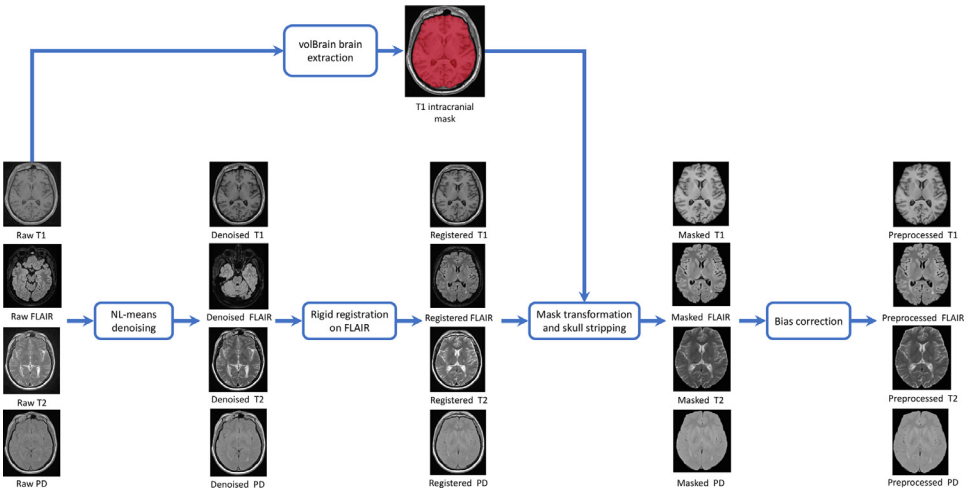

Fig. 2. Pipeline for obtaining the pre-processed dataset for each patient.

pert with respect to the consensus segmentation considered then as the ground truth.

2.3. Image preprocessing

Two versions of the dataset are provided: one with no pre-processing and one with standard pre-processing. The first dataset (raw data) includes the images in Nifti format as they are when converted from Dicom files. While no reference image is set in the raw dataset, it was advised in the 2016 challenge to use FLAIR as the reference frame as the ground truth is provided on this image.

In addition to the raw dataset, a second dataset with standard pre-processing is made available. The goal of this second dataset is to provide images with a standard pre-processing so that challengers wishing to compare only a segmentation method to the literature may do so, without having to perform all the pre-processing part again. In this dataset, several pre-processing steps are performed, illustrated in Fig. 2:

- Each MRI sequence is first denoised using the non-local means algorithm (Coupé et al., 2008) with a local patch neighborhood of 3 pixels
- Each MRI sequence is then rigidly registered onto the FLAIR image using a block-matching registration approach (Commowick et al., 2012). Resampling to the FLAIR image geometry was performed using sinc interpolation.
- Brain extraction is performed on the T1-w image using the volBrain platform (Manjón and Coupé, 2016). The mask obtained on the T1-w image is then applied to all other modalities registered on the FLAIR image.
- Finally, bias field correction is performed using the N4 algorithm (Tustison et al., 2010) using its ITK implementation with default parameters as implemented in Anima.

All individual tools for generating the preprocessed data apart from volBrain are available in our open-source code Anima[1]. The script preprocessing the data is made available in Anima scripts[2] (animaMSExamPreparationMSSEG2016). All tools, apart from mentioned parameters, were used with default parameters.

[1] Anima: Open source software for medical image processing from the Empenn team. https://anima.irisa.fr - RRID:SCR_017017

[2] Anima-Scripts: Open source scripts using Anima software for medical image processing from the Empenn team. https://anima.irisa.fr - RRID:SCR_017072

3. Data availability and access

A total of four datasets are available: pre-processed and raw data for each subset of patients for the training and testing sets used for the MICCAI 2016 challenge. These datasets are available from the Shanoir platform (Barillot et al., 2016). They are made available under a specific license to conform to the European GDPR (General Data Protection Regulation) rules. To follow these rules, we derived a data usage agreement form, to be agreed by the persons interested in downloading the dataset, inspired by the open brain consent article (Bannier et al., 2021). We provide this data usage agreement as a supplementary material and recall here its main points:

- The OFSEP publication chart (OFSEP acknowledgment and citing this data paper plus if necessary the challenge paper (Commowick et al., 2018) in any publication using a part or all of this dataset) should be adhered to
- Downloaders agree to provide their email address and research team information so that OFSEP may keep track of the use of the datasets
- Downloaders agree to make no commercial use, no redistribution of the data, and are informed that they should not use the data more than three years after download without informing OFSEP first

The link to the datasets on Shanoir is the following (Commowick, 2021): https://shanoir.irisa.fr/shanoir-ng/challenge-request. On this website, the downloader will have to read and approve the data usage agreement form and select the MSSEG 2016 study. Download of the datasets is restricted to users agreeing with the aforementioned data usage agreement (DUA) so that the OFSEP can keep track of the use of this data for its future reports.

The datasets are available as two zip files, containing respectively the training and testing patients subsets, split from the overall patients set as shown in Table 1. Each zip file contains the unprocessed and images pre-processed following the pipeline in Fig. 2. Each subset is organized by center first (as in Table 1) and then by patient. For each patient, the unprocessed data are located in the Raw_Data folder, the preprocessed data in Preprocessed_Data, the ground truth, brain mask and individual manual segmentations are located in Masks folder. Files for each patient are named after the MRI sequence of the image, plus a “preprocessed” suffix for dat apreprocessed. Manual segmentations are named “ManualSegmentation_1,...,7”. Their numbers are consistent throughout the patients: experts from Lyon correspond to manual segmentations 1, 2, 4, 5; experts from Rennes to segmentations 3 and 6; and the expert from Bordeaux to segmentation 7. The consensus segmentation is provided as a binary mask entitled “Consensus.nii.gz”.

The datasets are provided as is and can be used right away thanks to the presence of raw and pre-processed data for each patient. It was

advised for challengers in 2016 to use only the training data for machine learning algorithms, but this was not mandatory. Apart from one team which used their own training data as well, all challengers have used only the training data provided, so for better comparability it is advised to do so as well. For example of pipelines that have used this data, one may have a look at Anima scripts[3] (segmentation scripts) which contain two challenger pipelines (Beaumont et al., 2016a; 2016b).

4. Dataset validation

The dataset proposed in this paper was quality controlled for the purpose of the MSSEG challenge in 2016. It included visual checking for artifacts and a control of the acquisition parameters so that they fall in the OFSEP protocol. In itself, the fact that this dataset served as a basis for a challenge comparing 13 teams (Commowick et al., 2018) is a form of technical validation of the dataset. In addition, we propose two analyses for 1- providing database characteristics on the MS lesions, and 2- characterizing the experts performance in more depth. We show with these two analyses that the number and volume of lesions is sufficiently variable to represent what can be encountered in real life cases, and that the experts and ground truth may be trusted for evaluation.

4.1. Evaluation of lesions characteristics

To evaluate the representativeness of the database in a lesion delineation task, we computed for each patient the number of lesions and the total lesion load (lesions volume). It is indeed desirable to have a database sufficiently variable as one may want to study his/her algorithm capacity in various disease progression scenarios (many lesions or small total lesion load for example). We also wanted to check how much the training dataset was representative of the cases encountered in the testing dataset. Computing these lesions characteristics was done on the consensus segmentation by first computing connected components from the consensus binary masks using a six connectivity element. These connected components provided then the number of lesions in the patient. The total volume of these lesions (number of voxels in the binary mask multiplied by the voxel volume) provided the total lesion load. From these figures, we have computed distribution plots of the number of lesions and total lesion load, as well as a scatter plot of those two characteristics. We report these plots in Fig. 3.

This study shows that the number of lesions within a patient can be either small or quite large, ranging from 4 to 153 (median: 28). One exception that may be noted is one patient in the testing dataset that has no consensus lesion. This fact may be used, as it was done in the MSSEG 2016 challenge, to see how algorithms behave in such a situation, for which they were not designed. It may be a further important test point for new teams exploiting this dataset. The total lesion load per patient is also quite variable from one patient to another, going from 0.12 to 71.62 cm^3 (median: 9.09). The ages and genders of patients are well spread across datasets, lesion volumes and numbers. Both the training and test datasets have a variety of cases in terms of lesion load and number of lesions. With all this variety, further illustrated in Fig. 3 with respect to age, gender and dataset, a large spectrum of the cases encountered in clinical MS cases may be tested with this database.

4.2. Evaluation of experts segmentation reliability

The second technical validation of the datasaet concerns the validity of the experts segmentation. While no ground truth is available, it is important to know how much the experts vary with respect to each other and to the consensus. A too large variability would mean that some expert failed in their segmentation task. To perform this validation, we have computed the Dice segmentation and F1 detection scores as well as the average surface distance, as explained in Commowick et al. (2018), of each expert with respect to the consensus, for each patient case of the dataset. As doing so on the provided consensus from all experts would lead to a circular evaluation where the evaluated segmentation is included in the set used to generate the consensus, we have applied a leave-one-out strategy. We have thus, for each expert and each patient, computed the consensus without the evaluated segmentation and then evaluated the metrics against the obtained consensus. The results are presented as box plots in Fig. 4.

These plots show that experts are indeed variable in their assessments of the ground truth. Their median Dice score varies between 0.66 and 0.76, their detection F1 score between 0.64 and 0.84, and surface distance between 0 and 0.11, which represents good scores and above what the automatic methods achieved in the MICCAI 2016 challenge. While these figures show variations with respect to the consensus, all experts obtain good to very good scores overall. As such, this dataset provides us with ways to characterize differences between experts and could help new users find if some lesions delineated by just one or two experts could not be a miss by the others. In that case, one could further show how automatic methods may also help detect lesions that are not always obvious to the experts.

5. Discussion

The proposed dataset is the most comprehensive dataset for MS lesions segmentation validation at a single timepoint to date. However, the proposed dataset still suffers from several drawbacks. First, the size of the dataset (53 patients) is limited. This comes from the design of the dataset: we have preferred having a rather small dataset but with many manual lesions segmentations for each patient, rather than the reverse. This allows to get a very good idea of the inter-rater variability for segmentation among experts, but at the cost of a very long process to actually perform the manual segmentations. As a drawback from this choice, the dataset power for statistical analysis may be limited: all possible cases of MS lesions cannot be present in this relatively small dataset. That being said, this dataset is still the largest of its kind to date and will be very valuable for algorithms validation and comparison purposes.

Related to this size problem, the training of machine learning algorithms from this dataset, especially deep learning methods, may be limited by the small number of lesions in the training set. However, this is counterbalanced by two points: 1- machine learning algorithms often work on patches of lesions and in that case, the number of patches with lesions is large and represents a good part of patterns encountered in a clinical setting ; 2- other datasets can be added to this training set to enrich the learning phase and provide better results: some participants to the MICCAI 2016 challenge actually chose this approach (McKinley et al., 2016) and this appeared to be a good move as their results were better than many other methods.

A limitation of this dataset resides in the level of expertise of the human raters providing the manual segmentations. Due to the difficulty to recruit so many raters with a very high level of expertise, we have preferred designing carefully a training scheme for junior neuroradiologists. Resorting to asking junior neuroradiologists is often done as this task is very time consuming, usually with less well defined segmentation training and guidelines (or not exposed in the data description). Our scheme, as shown in the analysis provided in this article, has proven to reach a good level of agreement between the human raters. Combined with the computation of a robust consensus using LOP-STAPLE, the consensus can be trusted for evaluation and allows even to learn about the inter-rater variability. However, it will never reach the level that could be reached by senior neuroradiologists, especially after consensus meetings. The provided manual segmentations might therefore miss small lesions that are difficult to detect even for the human eye. There is thus

[3] Anima-Scripts: Open source scripts using Anima software for medical image processing from the Empenn team. https://anima.irisa.fr - RRID:SCR_017072

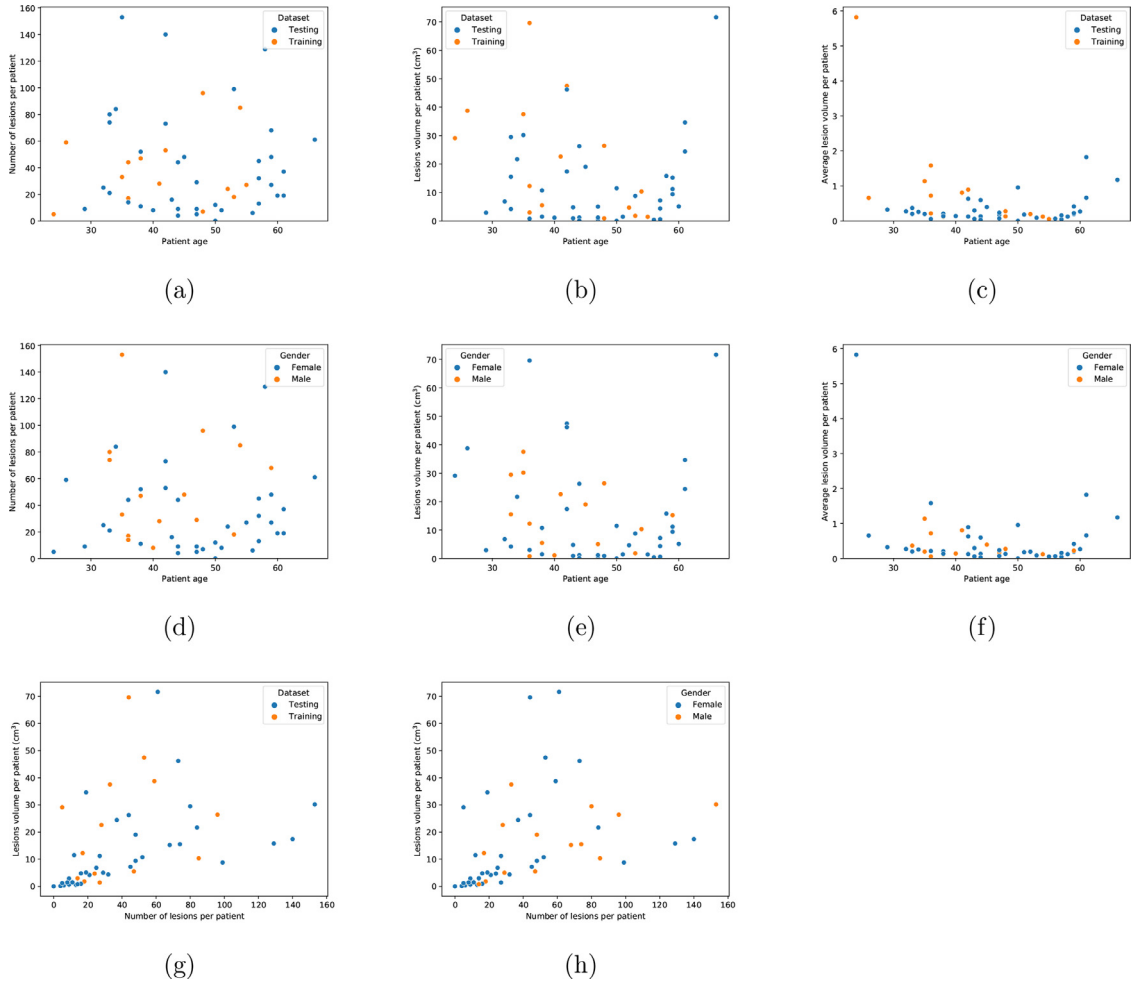

Fig. 3. Scatter plots of the population characteristics: number of lesions, lesional volume (in cm^3), ages and genders for the two datasets (training and testing). (a,d) scatter plots of the numbers of lesions with respect to age, (b,e) scatter plots of the total lesion load per patient with respect to age, (c,f) scatter plots of the average lesion volume per patient with respect to age. First line: colors indicate dataset, second line: colors indicate gender. Third line: scatter plots of number of lesions per patient with respect to lesion volume per patient colored by (g): dataset, (h): gender.

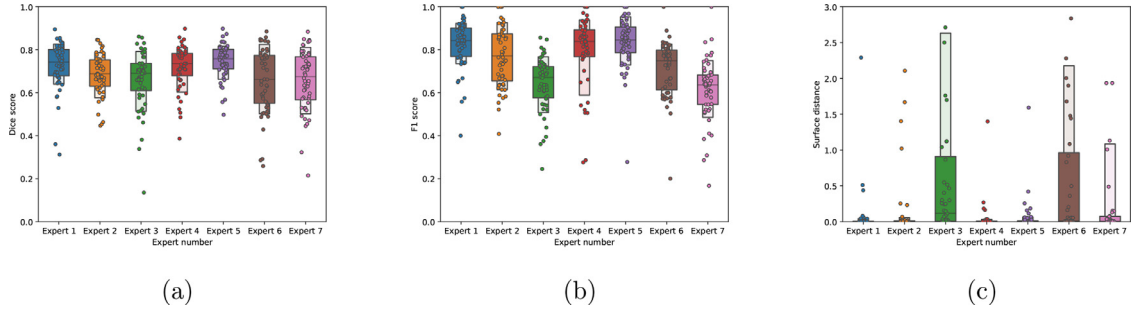

Fig. 4. Box plots of the experts Dice lesion segmentation scores (a), F1 lesion detection scores (b), and surface distance scores (c) over the whole MSSEG challenge database. These plots are boxen plots showing the first and second quartiles, the median and individual score points for each expert. Please note that a few values are outside the graph bounds and were kept out of the graph for readability.

room for future work in the definition of manual "ground truth", and its analysis. One could also imagine performing additional studies in the future using this dataset to see if automatic algorithms could help the neuroradiologist detect small lesions that would have been missed otherwise.

While constructed to have a relatively homogeneous age range overall and among centers, we have seen from our analysis that the gender (woman to man) ratio varies between centers and scanners (from 1.14 to 7). Some centers or scanners in the dataset thus do not reflect the world woman to man ratio of the disease prevalence. This drawback was however not seen to have a major impact when running the MICCAI 2016 challenge. Changes in contrasts due to this gender ratio variation are indeed small compared to those due to scanner change. Moreover, the largest gender ratio (7) is seen for the center that is present only in the testing set, thus providing a set that further tests the adaptability of the segmentation approaches. Future work with different databases could however study in depth this gender influence on segmentation to see if it has an impact on segmentation performance.

Finally, we have included with this dataset as much demographic and pathology data as we could include while respecting our constraints and rules from the European GDPR. While this data is already very comprehensive, it would be very good to have more biological and pathophysiological data to potentially see impacts of some lesion types on the ability of automatic algorithms to properly delineate / detect lesions. While beyond reach with this dataset, constructing a new database for studying these aspects would be of great importance to the field.

6. Conclusion

We have presented a detailed description of an open database of 53 MS patients designed for the evaluation of automatic lesion segmentation methods. It is notably comprising segmentations from 7 different experts and data coming from four scanners located in three centers. We make with this paper the whole dataset available for new challengers to evaluate their methods.

We believe that this database will have a great impact on lesion segmentation evaluation, especially since it as already been used for the MICCAI 2016 challenge where thirteen participants evaluated their methods. Thus newcomers will also be able to compare their results to these methods. Finally, we also believe that this database will have a great interest for the community of label fusion, where the seven different segmentations will allow for the comparison and developement of such algorithms.

Credit authorship contribution statement

Olivier Commowick: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Michaël Kain:** Resources, Software, Writing – original draft. **Romain Casey:** Resources, Data curation, Funding acquisition, Writing – original draft. **Roxana Ameli:** Resources, Methodology, Data curation, Writing – original draft. **Jean-Christophe Ferré:** Methodology, Data curation, Resources, Writing – original draft. **Anne Kerbrat:** Methodology, Resources, Data curation, Writing – original draft. **Thomas Tourdias:** Methodology, Resources, Data curation, Writing – original draft. **Frédéric Cervenansky:** Software, Writing – original draft. **Sorina Camarasu-Pop:** Software, Writing – original draft. **Tristan Glatard:** Conceptualization, Software, Writing – original draft. **Sandra Vukusic:** Resources, Funding acquisition. **Gilles Edan:** Methodology, Resources, Writing – original draft. **Christian Barillot:** Conceptualization, Methodology, Funding acquisition. **Michel Dojat:** Software, Writing – original draft. **Francois Cotton:** Conceptualization, Resources, Data curation, Funding acquisition, Writing – original draft.

Acknowledgments

This work was carried out in collaboration with The Observatoire Franąis de la Sclérose en Plaques (OFSEP), who is supported by a grant provided by the French State and handled by the “Agence Nationale de la Recherche”, within the framework of the “Investments for the Future” program, under the reference ANR-10-COHO-002, by the Eugéne Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation. This work was also partly funded and sponsored by France Life Imaging (grant ANR-11-INBS-0006 from the French “Investissements d’Avenir” program).

Finally, the authors are particularly thankful to Christian Barillot for his constant commitment in the OFSEP and FLI-IAM projects and his implication in the constitution of this dataset and the associated challenge. All this would not have existed without his help.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118589.

References

Akhondi-Asl, A., Hoyte, L., Lockhart, M.E., Warfield, S.K., 2014. A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights. IEEE Trans. Med. Imaging 33 (10), 1997–2009.

Bannier, E., Barker, G., Borghesani, V., Broeckx, N., Clement, P., Emblem, K.E., Ghosh, S., Glerean, E., Gorgolewski, K.J., Havu, M., Halchenko, Y.O., Herholz, P., Hespel, A., Heunis, S., Hu, Y., Hu, C.-P., Huijser, D., de la Iglesia Vayá, M., Jancalek, R., Katsaros, V.K., Kieseler, M.-L., Maumet, C., Moreau, C.A., Mutsaerts, H.-J., Oostenveld, R., Ozturk-Isik, E., Pascual Leone Espinosa, N., Pellman, J., Pernet, C.R., Pizzini, F.B., Trbalić, A.v.S., Toussaint, P.-J., Visconti di Oleggio Castello, M., Wang, F., Wang, C., Zhu, H., 2021. The open brain consent: informing research participants and obtaining consent to share brain imaging data. Hum. Brain Mapp. 42 (7), 1945–1951. doi:10.1002/hbm.25351.

Barillot, C., Bannier, E., Commowick, O., Corouge, I., Baire, A., Fackfack, I., Guillaumont, J., Yao, Y., Kain, M., 2016. Shanoir: applying the software as a service distribution model to manage brain imaging research repositories. Front. Inf. Commun. Technol. doi:10.3389/fict.2016.00025.

Beaumont, J., Commowick, O., Barillot, C., 2016a. Automatic multiple sclerosis lesion segmentation from intensity-normalized multi-channel MRI. In: Proceedings of the 1st MICCAI Challenge on Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure - MICCAI-MSSEG, pp. 8–15.

Beaumont, J., Commowick, O., Barillot, C., 2016b. Multiple sclerosis lesion segmentation using an automated multimodal graph cut. In: Proceedings of the 1st MICCAI Challenge on Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure - MICCAI-MSSEG, pp. 1–7.

Brisset, J.-C., Kremer, S., Hannoun, S., Bonneville, F., Durand-Dubief, F., Tourdias, T., Barillot, C., Guttmann, C., Vukusic, S., Dousset, V., Cotton, F., imaging group, O., 2020. New OFSEP recommendations for MRI assessment of multiple sclerosis patients: special consideration for gadolinium deposition and frequent acquisitions. J. Neuroradiol. 47 (4), 250–258.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., Cardoso, M.J., Cawley, N., Ciccarelli, O., Wheeler-Kingshott, C.A.M., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick, O., Barillot, C., Tomas-Fernandez, X., Warfield, S.K., Vaidya, S., Chunduru, A., Muthuganapathy, R., Krishnamurthi, G., Jesson, A., Arbel, T., Maier, O., Handels, H., Iheme, L.O., Unay, D., Jain, S., Sima, D.M., Smeets, D., Ghafoorian, M., Platel, B., Birenbaum, A., Greenspan, H., Bazin, P.-L., Calabresi, P.A., Crainiceanu, C.M., Ellingsen, L.M., Reich, D.S., Prince, J.L., Pham, D.L., 2017. Longitudinal multiple sclerosis lesion segmentation: resource & challenge. Neuroimage 148, 77–102.

Commowick, O., 2021. Multiple sclerosis lesions segmentation from multiple experts: the MICCAI 2016 challenge dataset. *Shanoir* https://shanoir.irisa.fr/shanoir-ng/challenge-request.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Camarasu-Pop, S., Girard, P., Améli, R., Ferré, J.-C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., Mahbod, A., Wang, C., McKinley, R., Wagner, F., Muschelli, J., Sweeney, E., Roura, E., Lladó, X., Santos, M.M., Santos, W.P., Silva-Filho, A.G., Tomas-Fernandez, X., Urien, H., Bloch, I., Valverde, S., Cabezas, M., Vera-Olmos, F.J., Malpica, N., Guttmann, C., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S.K., Cotton, F.c., Barillot, C., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci. Rep. 8 (1), 13650.

Commowick, O., Wiest-Daesslé, N., Prima, S., 2012. Block-matching strategies for rigid registration of multimodal medical images. In: Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 700–703.

Confavreux, C., Compston, D.A., Hommes, O.R., McDonald, W.I., Thompson, A.J., 1992. EDMUS, a European database for multiple sclerosis. J. Neurol. Neurosurg. Psychiatry 55 (8), 671–676.

Cotton, F., Kremer, S., Hannoun, S., Vukusic, S., Dousset, V., 2015. OFSEP, a nationwide cohort of people with multiple sclerosis: consensus minimal MRI protocol. J. Neuroradiol. 42 (3), 133–140.

Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27 (4), 425–441.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. 17 (1), 1–18.

Giovannoni, G., Comi, G., Cook, S., Rammohan, K., Rieckmann, P., Sørensen, P.S., Vermersch, P., Chang, P., Hamlett, A., Musch, B., Greenberg, S.J., 2010. A placebo-controlled trial of oral cladribine for relapsing multiple sclerosis. N. Engl. J. Med. 362 (5), 416–426.

Hauser, S.L., Bar-Or, A., Comi, G., Giovannoni, G., Hartung, H.P., Hemmer, B., Lublin, F., Montalban, X., Rammohan, K.W., Selmaj, K., Traboulsee, A., Wolinsky, J.S., Arnold, D.L., Klingelschmitt, G., Masterman, D., Fontoura, P., Belachew, S., Chin, P., Mairon, N., Garren, H., Kappos, L., 2017. Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. N. Engl. J. Med. 376 (3), 221–234.

Kappos, L., Bar-Or, A., Cree, B.A.C., Fox, R.J., Giovannoni, G., Gold, R., Vermersch, P., Arnold, D.L., Arnould, S., Scherz, T., Wolf, C., Wallström, E., Dahlke, F., Achiron, A., Achtnichts, L., Agan, K., Akman-Demir, G., Allen, A.B., Antel, J.P., Antiguedad, A.R., Apperson, M., Applebee, A.M., Ayuso, G.I., Baba, M., Bajenaru, O., Balasa, R., Balci, B.P., Barnett, M., Bass, A., Becker, V.U., Bejinariu, M., Bergh, F.T., Bergmann, A., Bernitsas, E., Berthele, A., Bhan, V., Bischof, F., Bjork, R.J., Blevins, G., Boehringer, M., Boerner, T., Bonek, R., Bowen, J.D., Bowling, A., Boyko, A.N., Boz, C., Bracknies, V., Braune, S., Brescia Morra, V., Brochet, B., Brola, W., Brownstone, P.K., Brozman, M., Brunet, D., Buraga, I., Burnett, M., Buttmann, M., Butzkueven, H., Cahill, J., Calkwood, J.C., Camu, W., Cascione, M., Castelnovo, G.,

Centonze, D., Cerqueira, J., Chan, A., Cimprichova, A., Cohan, S., Comi, G., Conway, J., Cooper, J.A., Corboy, J., Correale, J., Costell, B., Cottrell, D.A., Coyle, P.K., Craner, M., Cui, L., Cunha, L., Czlonkowska, A., da Silva, A.M., de Sa, J., de Seze, J., Debouverie, M., Debruyne, J., Decoo, D., Defer, G., Derfuss, T., Deri, N.H., Dihenia, B., Dioszeghy, P., Donath, V., Dubois, B., Duddy, M., Duquette, P., Edan, G., Efendi, H., Elias, S., Emrich, P.J., Estruch, B.C., Evdoshenko, E.P., Faiss, J., Fedyanin, A.S., Feneberg, W., Fermont, J., Fernandez, O.F., Ferrer, F.C., Fink, K., Ford, H., Ford, C., Francia, A., Freedman, M., Frishberg, B., Galgani, S., Garmany, G.P., Gehring, K., Gitt, J., Gobbi, C., Goldstick, L.P., Gonzalez, R.A., Grandmaison, F., Grigoriadis, N., Grigorova, O., Grimaldi, L.M.E., Gross, J., Gross-Paju, K., Gudesblatt, M., Guillaume, D., Haas, J., Hancinova, V., Hancu, A., Hardiman, O., Harmjanz, A., Heidenreich, F.R., Hengstman, G.J.D., Herbert, J., Herring, M., Hodgkinson, S., Hoffmann, O.M., Hofmann, W.E., Honeycutt, W.D., Hua, L.H., Huang, D., Huang, Y., Huang, D., Hupperts, R., Imre, P., Jacobs, A.K., Jakab, G., Jasinska, E., Kaida, K., Kalnina, J., Kaprelyan, A., Karelis, G., Karussis, D., Katz, A., Khabirov, F.A., Khatri, B., Kimura, T., Kister, I., Kizlaitiene, R., Klimova, E., Koehler, J., Komatineni, A., Kornhuber, A., Kovacs, K., Koves, A., Kozubski, W., Krastev, G., Krupp, L.B., Kurca, E., Lassek, C., Laureys, G., Lee, L., Lensch, E., Leutmezer, F., Li, H., Linker, R.A., Linnebank, M., Liskova, P., Llanera, C., Lu, J., Lutterotti, A., Lycke, J., Macdonell, R., Maciejowski, M., Maeurer, M., Magzhanov, R.V., Maida, E.-M., Malciene, L., Mao-Draayer, Y., Marfia, G.A., Markowitz, C., Mastorodimos, V., Matyas, K., Meca-Lallana, J., Merino, J.A.G., Mihetiu, I.G., Milanov, I., Miller, A.E., Millers, A., Mirabella, M., Mizuno, M., Montalban, X., Montoya, L., Mori, M., Mueller, S., Nakahara, J., Nakatsuji, Y., Newsome, S., Nicholas, R., Nielsen, A.S., Nikfekr, E., Nocentini, U., Nohara, C., Nomura, K., Odinak, M.M., Olsson, T., van Oosten, B.W., Oreja-Guevara, C., Oschmann, P., Overell, J., Pachner, A., Panczel, G., Pandolfo, M., Papeix, C., Patrucco, L., Pelletier, J., Piedrabuena, R., Pless, M., Polzer, U., Pozsegovits, K., Rastenyte, D., Rauer, S., Reifschneider, G., Rey, R., Rizvi, S.A., Robertson, D., Rodriguez, J.M., Rog, D., Roshanisefat, H., Rowe, V., Rozsa, C., Rubin, S., Rusek, S., Saccà, F., Saida, T., Salgado, A.V., Sanchez, V.E.F., Sanders, K., Satori, M., Sazonov, D.V., Scarpini, E.A., Schlegel, E., Schluep, M., Schmidt, S., Scholz, E., Schrijver, H.M., Schwab, M., Schwartz, R., Scott, J., Selmaj, K., Shafer, S., Sharrack, B., Shchukin, I.A., Shimizu, Y., Shotekov, P., Siever, A., Sigel, K.-O., Silliman, S., Simo, M., Simu, M., Sinay, V., Siquier, A.E., Siva, A., Skoda, O., Solomon, A., Stangel, M., Stefoski, D., Steingo, B., Stolyarov, I.D., Stourac, P., Strassburger-Krogias, K., Strauss, E., Stuve, O., Tarnev, I., Tavernarakis, A., Tello, C.R., Terzi, M., Ticha, V., Ticmeanu, M., Tiel-Wilck, K., Toomsoo, T., Tubridy, N., Tullman, M.J., Tumani, H., Turcani, P., Turner, B., Uccelli, A., Urtaza, F.J.O., Vachova, M., Valikovics, A., Walter, S., Van Wijmeersch, B., Vanopdenbosch, L., Weber, J.R., Weiss, S., Weissert, R., West, T., Wiendl, H., Wiertlewski, S., Wildemann, B., Willekens, B., Visser, L.H., Vorobeychik, G., Xu, X., Yamamura, T., Yang, Y.N., Yelamos, S.M., Yeung, M., Zacharias, A., Zelkowitz, M., Zettl, U., Zhang, M., Zhou, H., Zieman, U., Ziemssen, T., 2018. Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. Lancet 391 (10127), 1263–1273.

Kurtzke, J.F., 1983. Rating neurologic impairment in multiple sclerosis. Neurology 33 (11), 1444–1452.

Leray, E., Yaouanq, J., Le Page, E., Coustans, M., Laplaud, D., Oger, J., Edan, G., 2010. Evidence for a two-stage disability progression in multiple sclerosis. Brain 133 (7), 1900–1913.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió–Torrentà, L., Rovira, A., 2012. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. Inf. Sci. 186 (1), 164–185.

Manjón, J.V., Coupé, P., 2016. Volbrain: an online MRI brain volumetry system. Front. Neuroinform. 10, 30.

McKinley, R., Gundersen, T., Wagner, F., Chan, A., Wiest, R., Reyes, M., 2016. Nabla-net: a deep Dag-like convolutional architecture for biomedical image segmentation: application to white-matter lesion segmentation in multiple sclerosis. In: Proceedings of the 1st MICCAI Challenge on Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure - MICCAI-MSSEG, pp. 37–43.

Mortazavi, D., Kouzani, A.Z., Soltanian-Zadeh, H., 2012. Segmentation of multiple sclerosis lesions in MR images: a review. Neuroradiology 54 (4), 299–320.

Polman, C.H., O'Connor, P.W., Havrdova, E., Hutchinson, M., Kappos, L., Miller, D.H., Phillips, J.T., Lublin, F.D., Giovannoni, G., Wajgt, A., Toal, M., Lynn, F., Panzara, M.A., Sandrock, A.W., 2006. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. N. Engl. J. Med. 354 (9), 899–910.

Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F.D., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A.J., Waubant, E., Weinshenker, B., Wolinsky, J.S., 2011. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann. Neurol. 69 (2), 292–302.

Pugliatti, M., Rosati, G., Carton, H., Riise, T., Drulovic, J., Vécsei, L., Milanov, I., 2006. The epidemiology of multiple sclerosis in europe. Eur. J. Neurol. (13) 700–722.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S., 2008. 3D Segmentation in the clinic: a grand challenge II: MS lesion segmentation. MIDAS J. https://www.midasjournal.org/browse/publication/638.

Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S., Fujihara, K., Galetta, S.L., Hartung, H.P., Kappos, L., Lublin, F.D., Marrie, R.A., Miller, A.E., Miller, D.H., Montalban, X., Mowry, E.M., Sorensen, P.S., Tintoré, M., Traboulsee, A.L., Trojano, M., Uitdehaag, B.M.J., Vukusic, S., Waubant, E., Weinshenker, B.G., Reingold, S.C., Cohen, J.A., 2018. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol. 17 (2), 162–173.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 Bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.

Vukusic, S., Casey, R., Rollot, F., Brochet, B., Pelletier, J., Laplaud, D.-A., De Sèze, J., Cotton, F.c., Moreau, T., Stankoff, B., Fontaine, B., Guillemin, F., Debouverie, M., Clanet, M., 2020. Observatoire Français de la Sclérose en Plaques (OFSEP): a unique multimodal nationwide MS registry in France. Mult. Scler. 26 (1), 118–122.