A Multicenter Longitudinal MRI Study Assessing LeMan-PV Software Accuracy in the Detection of White Matter Lesions in Multiple Sclerosis Patients

Alexandra Ramona Todea, MD,[1,2] Lester Melie-Garcia, PhD,[2,3] Muhamed Barakovic, PhD,[2,3] Alessandro Cagol, MD,[2,3] Reza Rahmanzadeh, PhD,[2,3] Riccardo Galbusera, MD,[2,3] Po-Jui Lu, PhD,[2,3] Matthias Weigel, PhD,[2,3,4] Esther Ruberte, PhD,[2,3] Ernst-Wilhelm Radue, Prof, MD,[2] Sabine Schaedelin, MSc,[5] Pascal Benkert, PhD,[5] Yaldizli Oezguer, MD, PhD,[2,3] Tim Sinnecker, PhD,[2,3,6] Stefanie Müller, MD,[7] Lutz Achtnichts, MD,[8] Jochen Vehoff, MD,[7] Giulio Disanto, MD,[9] Oliver Findling, MD,[8] Andrew Chan, MD,[10] Anke Salmen, MD,[10] Caroline Pot, Prof, MD,[11] Patrice Lalive, Prof, MD,[12] Claire Bridel, Prof, MD,[12] Chiara Zecca, Prof, MD,[9,13] Tobias Derfuss, Prof, MD,[3] Luca Remonda, Prof, MD,[14] Franca Wagner, MD,[15] Maria Vargas, Prof,[16] Renaud Du Pasquier, Prof, MD,[11] Emanuele Pravata, MD,[17,13] Johannes Weber, MD,[18] Claudio Gobbi, MD,[9,13] David Leppert, Prof, MD,[3] Jens Wuerfel, PhD,[6] Tobias Kober, PhD,[19,20,21] Benedicte Marechal, PhD,[19,20,21] Ricardo Corredor-Jerez, PhD,[19,20,21] Marios Psychogios, Prof, MD,[1] Johanna Lieb, MD,[1] Ludwig Kappos, Prof, PhD, MD,[2,3] Meritxell Bach Cuadra, PhD,[22] Jens Kuhle, MD, PhD,[3] and Cristina Granziera, Prof, PhD, MD,[2,3*] for the Swiss MS Cohort Study

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.28618

Received Oct 24, 2022, Accepted for publication Jan 14, 2023.

*Address reprint requests to: C.G., Translational Imaging in Neurology (ThINk), Department of Biomedical Engineering, University of Basel, Basel, Gewerbestrasse 14, 4123 Allschwil, Switzerland. E-mail: cristina.granziera@usb.ch; cristina.granziera@unibas.ch
Alexandra Ramona Todea and Lester Melie-Garcia made equal contributions to this work and share the first authorship.

From the [1]Department of Neuroradiology, Clinic of Radiology and Nuclear Medicine, University Hospital of Basel, Basel, Switzerland; [2]Translational Imaging in Neurology (ThINk) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; [3]Department of Neurology, University Hospital Basel, Switzerland, MS Center and Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland; [4]Division of Radiological Physics, Department of Radiology, University Hospital Basel and University of Basel, Basel, Switzerland; [5]Clinical Trial Unit, Department of Clinical Research, University Hospital Basel, University of Basel, Basel, Switzerland; [6]Medical Image Analysis Center (MIAC) and qbig, Department of Biomedical Engineering, University Basel, Basel, Switzerland; [7]Department of Neurology, Cantonal Hospital St. Gallen, St. Gallen, Switzerland; [8]Department of Neurology, Cantonal Hospital Aarau, Switzerland; [9]Department of Neurology, Neurocenter of Southern Switzerland, EOC, Lugano, Switzerland; [10]Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland; [11]Service of Neurology, Department of Clinical Neurosciences, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland; [12]Department of Clinical Neurosciences, Geneva University Hospital and Faculty of Medicine, Geneva, Switzerland; [13]Faculty of Biomedical Sciences, University of Italian Switzerland, Lugano, Switzerland; [14]Department of Radiology, Cantonal Hospital Aarau, Switzerland; [15]Department of Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland; [16]Department of Radiology, Geneva University Hospital and Faculty of Medicine, Geneva, Switzerland; [17]Department of Neuroradiology, Neurocenter of Southern Switzerland, Lugano, Switzerland; [18]Department of Radiology, Cantonal Hospital St. Gallen, St. Gallen, Switzerland; [19]Advanced Clinical Imaging Technology, Siemens Healthineers International, Lausanne, Switzerland; [20]Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

Background: Detecting new and enlarged lesions in multiple sclerosis (MS) patients is needed to determine their disease activity. LeMan-PV is a software embedded in the scanner reconstruction system of one vendor, which automatically assesses new and enlarged white matter lesions (NELs) in the follow-up of MS patients; however, multicenter validation studies are lacking.
Purpose: To assess the accuracy of LeMan-PV for the longitudinal detection NEL white-matter MS lesions in a multicenter clinical setting.
Study Type: Retrospective, longitudinal.
Subjects: A total of 206 patients with a definitive MS diagnosis and at least two follow-up MRI studies from five centers participating in the Swiss Multiple Sclerosis Cohort study. Mean age at first follow-up = 45.2 years (range: 36.9–52.8 years); 70 males.
Field Strength/Sequence: Fluid attenuated inversion recovery (FLAIR) and T1-weighted magnetization prepared rapid gradient echo (T1-MPRAGE) sequences at 1.5 T and 3 T.
Assessment: The study included 313 MRI pairs of datasets. Data were analyzed with LeMan-PV and compared with a manual "reference standard" provided by a neuroradiologist. A second rater (neurologist) performed the same analysis in a subset of MRI pairs to evaluate the rating-accuracy. The Sensitivity (Se), Specificity (Sp), Accuracy (Acc), F1-score, lesion-wise False-Positive-Rate (aFPR), and other measures were used to assess LeMan-PV performance for the detection of NEL at 1.5 T and 3 T. The performance was also evaluated in the subgroup of 123 MRI pairs at 3 T.
Statistical Tests: Intraclass correlation coefficient (ICC) and Cohen's kappa (CK) were used to evaluate the agreement between readers.
Results: The interreader agreement was high for detecting new lesions (ICC = 0.97, *P*value < 10^{-20}, CK = 0.82, *P* value = 0) and good (ICC = 0.75, *P* value < 10^{-12}, CK = 0.68, *P* value = 0) for detecting enlarged lesions. Across all centers, scanner field strengths (1.5 T, 3 T), and for NEL, LeMan-PV achieved: Acc = 61%, Se = 65%, Sp = 60%, F1-score = 0.44, aFPR = 1.31. When both follow-ups were acquired at 3 T, LeMan-PV accuracy was higher (Acc = 66%, Se = 66%, Sp = 66%, F1-score = 0.28, aFPR = 3.03).
Data Conclusion: In this multicenter study using clinical data settings acquired at 1.5 T and 3 T, and variations in MRI protocols, LeMan-PV showed similar sensitivity in detecting NEL with respect to other recent 3 T multicentric studies based on neural networks. While LeMan-PV performance is not optimal, its main advantage is that it provides automated clinical decision support integrated into the radiological-routine flow.
Evidence Level: 4
Technical Efficacy: Stage 2

J. MAGN. RESON. IMAGING 2023;58:864–876.

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system characterized by focal perivenular inflammatory lesions leading to demyelinating plaques and diffuse inflammation and degeneration of the central nervous system tissue.[1]

In MS patients, MRI is used for diagnosis, prognosis, and treatment follow-up. Lesion location and also the identification of new and enlarging white matter (WM) lesions contribute to the diagnostic criteria for MS (i.e. dissemination in space and time)[2–4] and also to an evaluation of activity, which might lead to escalation of the immunomodulatory therapy.[5,6]

Identifying changes in lesion number is prone to intraobserver and interobserver variability.[7,8] Most existing tools dedicated to evaluating MS lesion evolution allow a cross-sectional analysis, and only a few support the longitudinal assessment. These tools can be roughly classified into two main groups: 1) lesion segmentation-based methods[9,10] and 2) change detection-based methods.[11–13]

LeMan-PV is software that combines lesion segmentation and change detection-based approaches to perform longitudinal lesion detection in MS patients.[14,15] This software is based on the k-nearest neighbor (k-NN) technique and is optimized to detect lesions with a volume as small as three connected voxels (typically 0.0036 mL). This software has been developed to estimate the partial volume, which improves the volume estimation of new lesions.[16] Baseline and follow-up T1-weighted magnetization prepared rapid acquisition with gradient echo (MPRAGE) and 3D fluid attenuated inversion recovery (FLAIR) images are used as inputs for the segmentation method. The output provides a segmentation mask indicating the new, enlarged, and stable WM lesions.

The advantage of this software is that it generates segmentation masks that permit visual assessment of its reliability, a fundamental characteristic of a clinical decision support tool. Moreover, these masks are directly generated after the image acquisition, rendering them immediately available in clinical routine. LeMan-PV has been clinically validated in a single-center setting,[15] but multicenter validation is lacking.

Thus, the main aim of this study was to assess the diagnostic accuracy of LeMan-PV for detecting new and enlarging WM lesions in MS patients in five centers.

[21]LTS5, École Polytechnique FÉdÉrale de Lausanne (EPFL), Lausanne, Switzerland; and [22]CIBM Center for Biomedical Imaging, Radiology Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland

Additional supporting information may be found in the online version of this article

Materials and Methods

Study Design

We performed a retrospective study on 206 subjects from the Swiss Multiple Sclerosis Cohort (SMSC) between 2012 and 2019.[17]

The ethics committee approved the study of all participating institutions, and all subjects gave written informed consent before participation in the cohort study.

Participants

Eligibility inclusion criteria were a diagnosis of MS according to the revised McDonald criteria 2017,[2] and at least two MRI visits with a 3D-FLAIR sequence and with a native 3D T1-weighted MPRAGE sequence, separated by at least 1 year. A total of 206 patients were retrospectively selected among data collected between 2012 and 2019 within the SMSC Study with the following attributes male/females: 70/136; median age at baseline 45.2 years (range 36.9–52.8 years); median disease duration at baseline: 12.2 years (range 6.5–20.7 years); relapsing–remitting MS, $n = 184$; primary progressive MS, $n = 8$, secondary progressive MS, $n = 14$.

As part of the SMSC study, all patients underwent regular, standardized neurological evaluations (performed at least annually), with the calculation of the Expanded Disability Status Scale (EDSS) score[18] conducted by certified raters (https://www.neurostatus.net/). The median EDSS at the first-time point baseline was 2.5 (minimum = 1.5, maximum: 3.5).

The clinical and demographic characteristics of the patients enrolled in the study are reported in Table 1.

TABLE 1. Demographic Details of the Patients Included in the Study

Demographic Variable	Value	
Number of patients	206	
Gender	Male	70
	Female	136
Age at the first follow-up (years)	45.2 (36.9, 52.8)	
Disease subtype at the first scan (number)	RRMS	184 (89.3%)
	PPMS	8 (3.9%)
	SPMS	14 (6.8%)
EDSS at the first scan	2.5 [1.5, 3.5]	
Disease duration at the first scan (years)	12.2 [6.5, 20.7]	
Follow-up time at the first scan (years)	2.3 (2.2)	
DMT at the first scan	All patients	

RRMS = remitting relapsing MS; PPMS = primary progressive MS; SPMS = secondary progressive MS; EDSS = expanded disability status score; DMT = disease-modifying therapy.

In the first step, 351 MRI dataset pairs from 206 patients fulfilled the criteria; 38 pairs were excluded due to missing follow-up or insufficient registration accuracy. The registration accuracy was assessed by A.T and L.M-G, and the exclusion was made by decision consensus. We did not include patients with mixed field strength visits, for example, once at 1.5 T and then at 3 T or vice versa, due to the low number of available cases in 2/5 centers. The final analysis was hence performed on 313 MRI pairs (Fig. 1). Figure S1 shows the number of patients that have from 1 to 6 scan pairs.

MRI Protocols

MRI acquisitions were performed at 1.5 T (4 centers, 190 pairs, 380 scans; MAGNETOM AvantoFit, Siemens Healthcare, Erlangen, Germany) and 3 T (2 centers, 123 pairs, 246 scans: MAGNETOM Skyra and SkyraFit, both Siemens Healthcare) using the 20-channel commercial head coils on all scanners. The number of scan pairs per site and scanner magnetic field strength is shown in Table 2. The imaging protocol included a 3D-FLAIR and a native MPRAGE (details are summarized in Tables 3 and 4).

Image Analysis

IMAGE QUALITY ASSESSMENT. We first assessed the quality of the scans performed by A.T and L.M-G. Depending on the qualitative evaluation (signal-to-noise ratio, artifacts, contrast, good registration between time points), a score of 1 (optimal/good quality), 2 (nonoptimal/medium quality but still diagnostically valid), or 3 (bad quality) was attributed to each MR image (see Supplementary Material S1 for more information). Both experts re-evaluated the mismatch in scan quality scores, and the final scores were attributed according to a final consensus.

MANUAL LESION SEGMENTATION (GROUND TRUTH). One neuroradiologist A.T (5 years of experience in neuroradiology, 9 years of experience in radiology) evaluated, identified, and marked new and enlarged MS lesions on 3D FLAIR-3D MPRAGE datasets using ITK-SNAP (www.itksnap.org).[19] In a subset of 60 MRI pairs, a second neurologist, A.C (3 years of experience in neuroimaging research), performed the same analysis. This allowed evaluation of the interrater reliability (IRR) or rating accuracy in the lesion segmentation process.

The raters were blinded to the clinical status of the patients and the results from LeMan-PV and performed the assessments independently of each other.

Lesions were classified as new or enlarged based on current clinical criteria. A lesion was considered new when it was identified in the second time point images, and its diameter was at least 3 mm in the longest axis,[20] and a lesion was defined as enlarged when in the second time point; the enlargement was at least 50% larger than the initial size in the first time point MRI.

AUTOMATIC LESION SEGMENTATION: LEMAN-PV ALGORITHM. LeMan-PV provides a lesion progression map indicating new, enlarged, and stable lesions.[15]

In summary (see flowchart Fig. 2), it takes as inputs a 3D FLAIR and a 3D T1w MPRAGE dataset obtained at two different time points. The images were previously submitted to

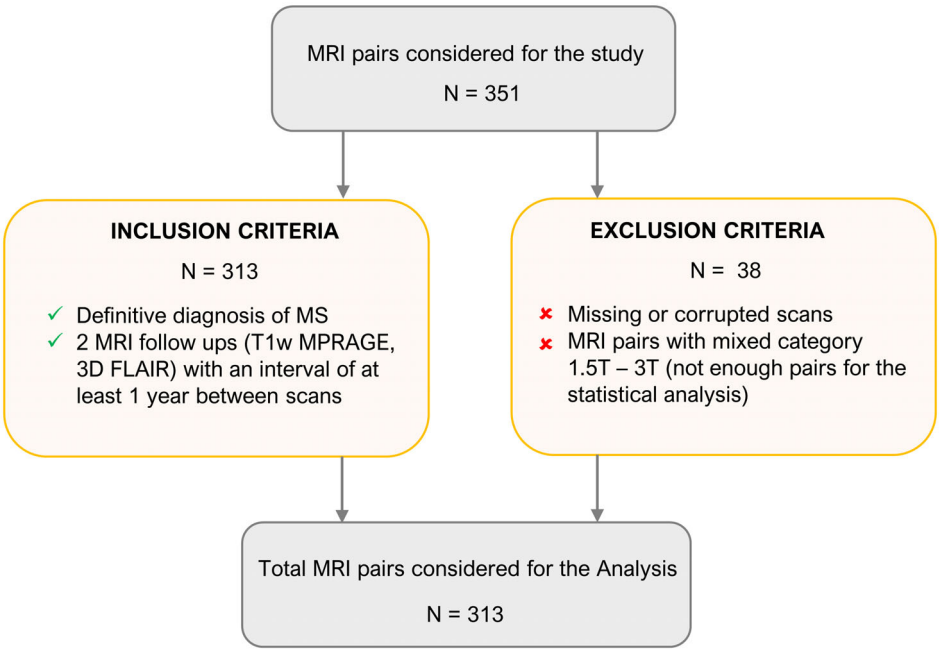

FIGURE 1: Inclusion and exclusion criteria flowchart. MS = multiple sclerosis.

TABLE 2. The Number of Scan Pairs Per Site and Scanner Magnetic Field Strength

	Scanner Magnetic Field strength		
Site	1.5 T	3 T	1.5 T and 3 T together
Center 1	51	-	51
Center 2	50	93	143
Center 3	49	-	49
Center 4	-	30	30
Center 5	40	-	40
All	190	123	313

preprocessing steps, including N4 bias field correction and intensity normalization.[16] FLAIR and MPRAGE images were rigidly registered to the second time point MPRAGE using the Elastix toolbox.[21] It first computes cross-sectional lesion segmentation maps for each time point, then creates a "joint difference image" subtracting FLAIR and MPRAGE images separately at the first and second time points and finally estimates a lesion progression map combining the cross-sectional lesion segmentation map from the second time point and the "joint difference image".[16] Voxels classified as new or enlarged lesional tissue are binarized to generate the final progression maps.

LeMan-PV defines the new lesion as identifiable on the second time point image but not on the first time point. The enlarged lesion is determined when the lesion diameter in the second time point increased by at least 50% with respect to the first time point.

EVALUATION OF THE LONGITUDINAL LEMAN-PV PERFORMANCE ON WM LESIONS. To evaluate LeMan-PV's performance, we compared new/enlarged lesion-segmentation masks with ground-truth masks generated by the first expert (A.T) using an in-house MATLAB (https://www.mathworks.com/) function.

A new/enlarged lesion was defined as a cluster of voxels interconnected over a three-dimensional 26 voxels neighborhood with a minimum volume of 0.01 mL.[22] A lesion was considered a true positive if it overlapped with a ground-truth lesion. Exclusively one longitudinal LeMan-PV lesion was deemed to overlay a ground-truth lesion in case of multiple coincidences.

As this study focused on WM lesion detection, the final lesion progression maps were masked with a white matter mask (WMM). The WMM was generated using the lesion-filled version of the follow-up MPRAGE image applying the SPM12 unified segmentation tool (with default settings) (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). The Lesion Filling toolbox developed by Valverde et al (http://atc.udg.edu/nic/slfToolbox/index.html)[23] was utilized for the lesion-filling step.

Additionally, we explored different intensity thresholds for binarizing joint difference image map (0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9) to optimize the performance of LeMan-PV in the longitudinal assessment of new and enlarged lesions compared to the ground truth. We selected the threshold that showed the best trade-off between sensitivity and 1-specificity, which is the one with the minimum Euclidean distance to the maximum sensitivity and specificity (1-specificity, sensitivity) = (0,1).

LeMan-PV performance was evaluated per center and in all centers combined: 1) for new and enlarged lesions separately, 2) considering new and enlarged lesions jointly; 3) taking into account the MRI scanner magnetic field 1.5 T and 3 T separately and combined; 4) considering different image quality scores, for new/enlarged lesions separately and together.

TABLE 3. MRI Parameters for the MPRAGE Sequence Per Site

Site code	Scanner Model	Magnetic Field (T)	TR (msec)	TE (msec)	TI (msec)	FoV (mm × mm)	Voxel size (mm × mm × mm)	FA (deg)	Slice Gap (mm)
Center 1	Aera	1.5	2200	2.67	900	256 × 256	1 × 1 × 1	8	0
Center 2	Avanto	1.5	2700	5.03	950	256 × 256	1 × 1 × 1	8	
	Skyra	3	2300	2.9	900	256 × 240	1 × 1 × 1	9	
	Skyra_fit	3	2300	3.02	900	256 × 240	1 × 1 × 1	9	
			2300	1.96	900	256 × 240	1 × 1 × 1	9	
			2300	1.98	900	256 × 240	1 × 1 × 1	9	
Center 3	Avanto	1.5	2700	5.03	950	256 × 256	1 × 1 × 1	8	
	Avanto_fit	1.5	2700	5.03	950	256 × 256	1 × 1 × 1	8	
Center 4	Skyra	3	1800	2.43	900	256 × 256	1 × 1 × 1	9	
			2300	2.98	900	256 × 240	1 × 1 × 1	9	
Center 5	Avanto	1.5	2700	3.08	950	256 × 256	1 × 1 × 1	8	
	Avanto_fit	1.5	2700	3.09	950	256 × 256	1 × 1 × 1	8	

TR = repetition time; TE = echo time; TI = inversion time; FoV = field of view; FA = flip angle.

TABLE 4. MRI Parameters for the FLAIR Sequence Per Site

Site code	Scanner Model	Magnetic Field (T)	TR (msec)	TE (msec)	TI (msec)	FoV (mm × mm)	Voxel size (mm × mm × mm)	FA (deg)	Slice Gap (mm)
Center 1	Aera	1.5	5000	333	1800	256 × 217	1 × 1 × 1	120	0
			5000	335	1800	256 × 217			
			5000	334	1800	256 × 218			
Center 2	Avanto	1.5	6000	352	2200	256 × 258			
	Avanto_fit	1.5	5000	335	1800	256 × 251			
	Skyra	3	5000	280	1800	256 × 256			
	Skyra_fit	3	5000	280	1800	256 × 256			
Center 3	Avanto	1.5	6000	257	2200	256 × 256			
			6000	352	2200	256 × 256			
	Avanto_fit	1.5	6000	353	2200	256 × 256			
Center 4	Skyra	3	5000	392	1800	256 × 256			
			5000	373	1800	256 × 256			
Center 5	Avanto	1.5	6000	354	2200	256 × 258			
	Avanto_fit	1.5	6000	355	2200	256 × 256			

TR = repetition time; TE = echo time; TI = inversion time; FoV = field of view; FA = flip Angle.

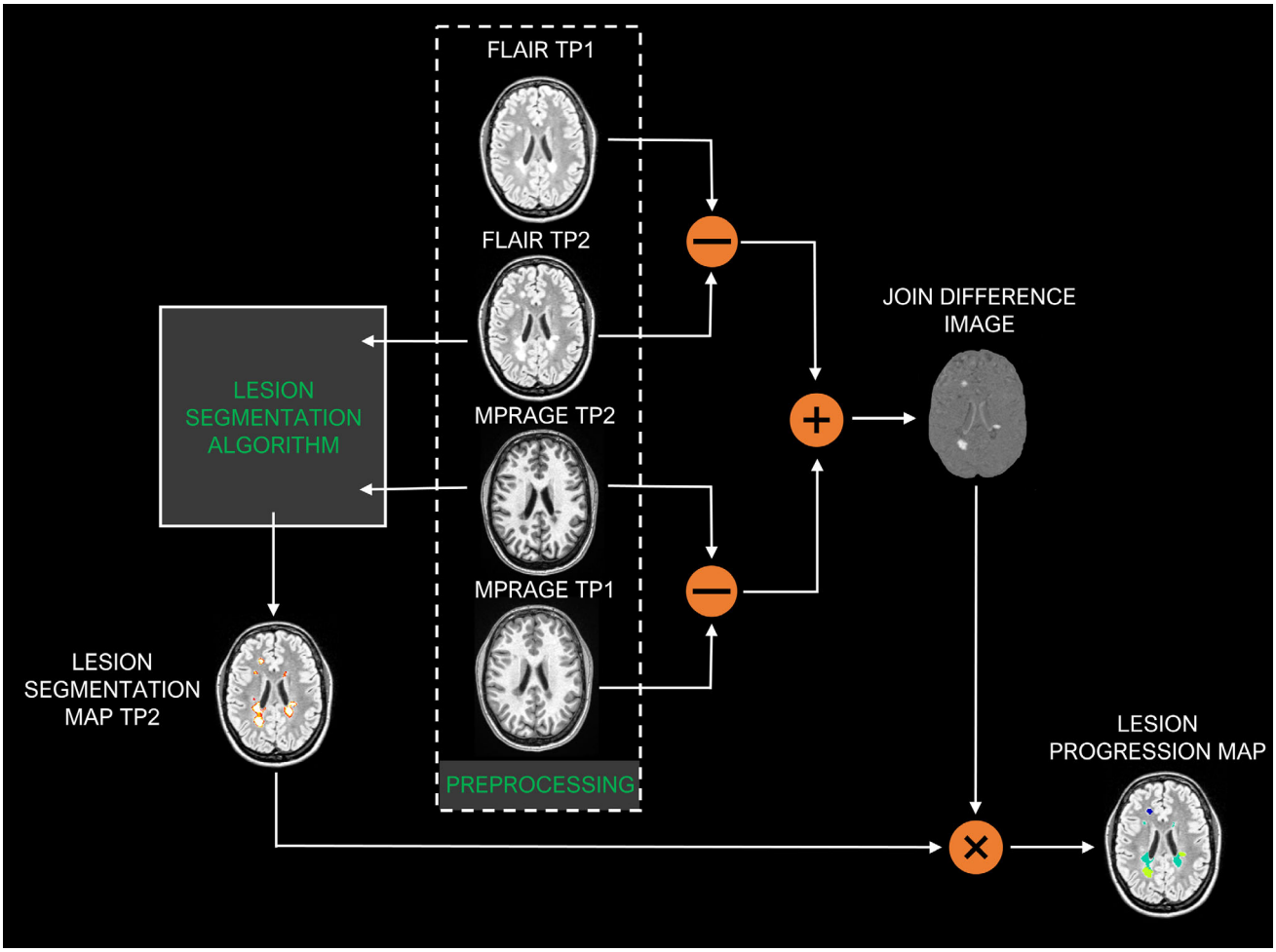

FIGURE 2: LeMan-PV methodology flowchart.

Statistical Analysis

IRR was assessed by the intraclass correlation coefficient (ICC) and Cohen's kappa coefficient (CK).[24] ICC is considered as: 1) poor ICC < 0.5; 2) moderate 0.5 < ICC < 0.75, 3) good 0.75 < ICC < 0.9, and 4) excellent ICC > 0.9.[25–27] For CK, 1) CK values ≤ 0 as indicating no agreement, 2) 0.01–0.20 as none to slight, 3) 0.21–0.40 as fair, 4) 0.41–0.60 as moderate, 5) 0.61–0.80 as substantial, and 6) 0.81–1.00 as almost perfect agreement.

Lesion-wise sensitivity, specificity, accuracy, F1 score, positive predictive value (PPV), and negative predictive value (NPV) metrics were assessed across new/enlarged WM lesions. As used in previous studies, an Adhoc lesion-wise false positive rate (aFPR) was also computed.[15,28] The Adhoc true negative rate (aTNR) allows LemanPV performance evaluation in cases with no lesions. The performance parameters are defined as follows:

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}).$$

$$\text{Specificity} = \text{TN}/(\text{TN}+\text{FP}).$$

$$\text{Accuracy} = (\text{TN}+\text{TP})/(\text{TN}+\text{TP}+\text{FN}+\text{FP}).$$

$$\text{F1 score} = \text{TP}/\left(\text{TP}+0.5^{*}(\text{FP}+\text{FN})\right).$$

$$\text{aFPR} = \text{FP}/\text{P}.$$

$$\text{aTNR} = \text{aTN}/\text{aN}.$$

$$\text{PPV} = \text{TP}/(\text{TP}+\text{FP}).$$

$$\text{NPV} = \text{TN}/(\text{TN}+\text{FN}).$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, P is the number of manually ground-truth delineated lesions; N is the number of ground-truth negative, aN is the number of ground-truth pairs images without any new/enlarged lesions, and aTN is the number of pair's images without new/enlarged lesions detected by LeMan-PV.

A true positive is defined as a LemanPV lesion overlapping with a manually delineated lesion. A false positive is a LemanPV lesion not overlying any manually delineated lesion. A true negative is defined as LemanPV correctly detecting the nonlesion WM region. The false negative is LemanPV incorrectly detecting the nonlesion WM region.

Results

Image Quality Assessment and Manual Identification of New/Enlarged Lesions

All MRI pairs ($n = 313$) were first classified into three categories according to their quality: good quality ($n = 222$) for diagnostic, medium quality not optimal but still useful for diagnostic purposes ($n = 80$), and poor quality ($n = 11$).

The first rater A.C found new/enlarged lesions in 69 MRI scans pairs out of 313. The total number of new/enlarged lesions was 190 (new lesions: 121; enlarged lesions: 69).

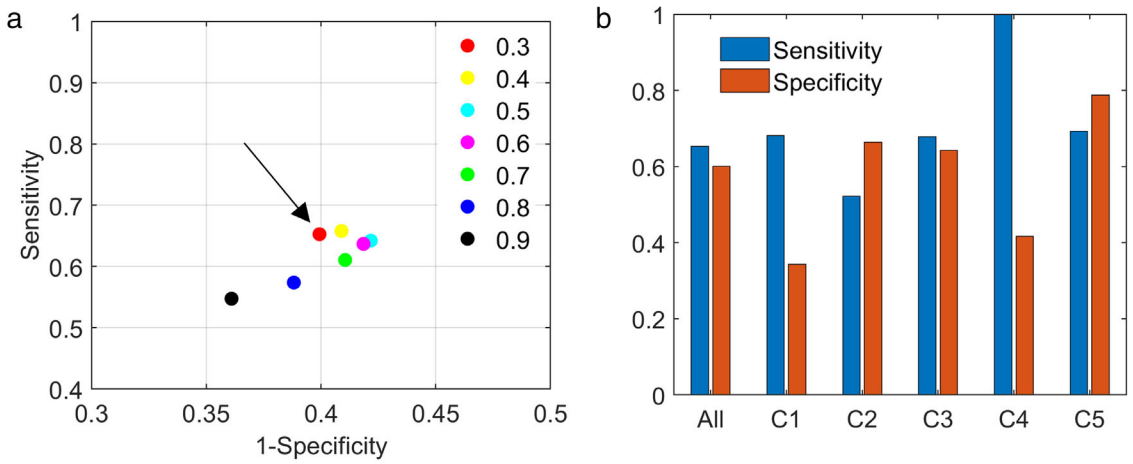

FIGURE 3: (a) Plot of sensitivity (y-axis) and 1-specificity (x-axis) to select the optimal threshold that provides the best LeMan-PV performance. The arrow indicates the optimal value = 0.3. In colors are represented the different thresholds from 0.3 to 0.9 in steps of 0.1. (b) LeMan-PV performance, sensitivity, and specificity for all centers. All = all centers together with joining new and enlarged lesions; C1 = center 1; C5 = center 5.

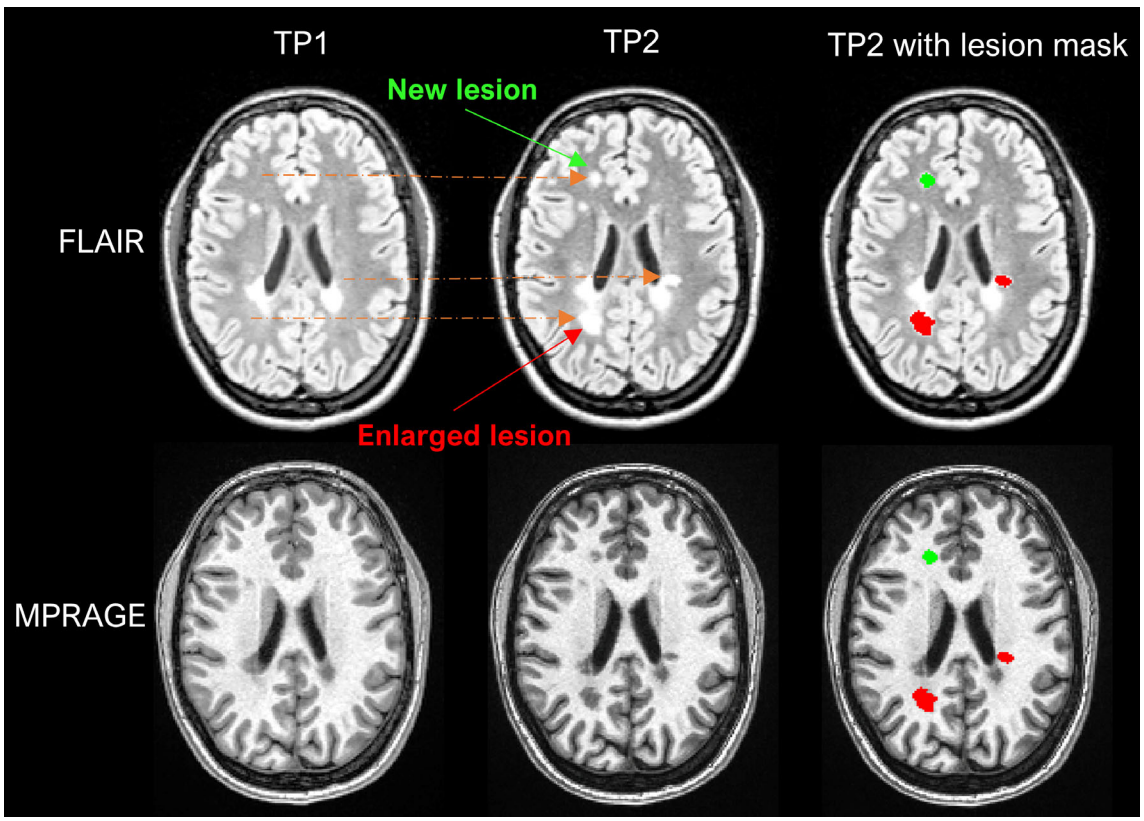

FIGURE 4: Example of LeMan-PV detecting new and enlarged lesions. The arrows in orange represented in the FLAIR image indicate the region of change from time point 1 (TP1) to time point 2 (TP2) for a new (green arrow) or enlarged (red arrow) lesion. The last column, named "TP2 with lesion mask" represents the LeMan-PV progression map overlapping the FLAIR and MPRAGE in TP2.

The two raters had a very high interrater concordance in the detection of new lesions (ICC = 0.97, confidence interval [CI] = [0.96, 0.98], $P < 10^{-20}$) and a good one for enlarged lesions (ICC = 0.75, CI = [0.61, 0.84], $P = 10^{-13}$). The Cohen's kappa coefficient was CK = 0.82, CI = [0.67, 0.96], $P = 0.00$; in this case, the observed agreement is not accidental and showed perfect agreement. For enlarged lesions, CK = 0.68, CI = [0.53, 0.84], $P = 0.00$; in this case, the observed agreement is not accidental and showed substantial agreement.

The optimal joint difference image map threshold for our dataset was 0.3 (Fig. 3a). Figure 4 shows an example of LeMan-PV detecting new and enlarged lesions with this specific threshold. Using this parameter, for all centers combined, we obtained the following performance results (Table 5):

1. new lesions were detected by LeMan-PV with Acc = 62% (Se = 52%, Sp = 66%), F1-score = 0.43 and aFPR = 0.86;
2. enlarged lesions were detected with Acc = 60%, (Se = 88%, Sp = 54%), F1-score = 0.44 and aFPR = 2.08; and
3. for new and enlarging lesions combined, LeMan-PV resulted in a detection Acc = 61% (Se = 65%, Sp = 60%), F1-score = 0.44, and aFPR = 1.31.

When analyzing the data per center, LeMan-PV performed best in the detection of new/enlarging lesions in center 3 (C3) at 1.5 T (Acc = 65%, Se = 67%, Sp = 64%, F1-score = 0.59 and aFPR = 0.62). Center 5 (C5) also showed a good performance, but the number of lesions in this center was only 13, with a higher aFPR. Figure 3b shows the sensitivity and specificity for all centers.

For new and enlarged lesions detected on scans acquired at 1.5 T (n = 190 pairs), Acc = 65% (Se = 55%, Sp = 58%), F1-score = 0.48, aFPR = 1.02; for scans acquired at 3 T (n = 123 pairs): Acc = 66% (Se = 66%, Sp = 66%), F1-score = 0.28, aFPR = 3.03(Table 6).

The LeMan-PV performance considering different image quality scores is shown in Table 7. The highest performance was achieved when only the best quality images were used (score = 1) for new/enlarged lesions separately and together. For all lesions together Acc = 68% (Se = 69%, Sp = 67%), F1-score = 0.49 and aFPR = 1.11.

TABLE 5. Leman-PV Performance Metrics for New and Enlarged Lesion Detection Overall and Per Center

Lesion Type	Site	N. Lesions	Accuracy	Sensitivity	Specificity	F1-score	aFPR	aTNR	PPV	NPV
New	All sites	121	0.62	0.52	0.66	0.43	0.86	0.68	0.37	0.78
	Center 1	45	0.51	0.55	0.47	0.52	0.58	0.49	0.49	0.54
	Center 2	34	0.65	0.47	0.7	0.34	1.26	0.73	0.27	0.85
	Center 3	29	0.59	0.45	0.67	0.45	0.55	0.63	0.45	0.67
	Center 4	4	0.59	1	0.53	0.36	3.5	0.54	0.22	1
	Center 5	9	0.79	0.55	0.85	0.5	0.66	0.85	0.45	0.89
Enlarged	All sites	69	0.60	0.88	0.54	0.44	2.08	0.58	0.3	0.95
	Center 1	24	0.44	0.92	0.21	0.51	1.66	0.24	0.35	0.85
	Center 2	12	0.63	0.67	0.63	0.22	4.41	0.67	0.13	0.96
	Center 3	27	0.72	0.92	0.61	0.70	0.70	0.69	0.57	0.94
	Center 4	2	0.34	1	0.3	0.16	10.5	0.32	0.09	1
	Center 5	4	0.75	1	0.72	0.42	2.75	0.74	0.27	1
New + Enlarged	All sites	190	0.61	0.65	0.60	0.44	1.31	0.49	0.33	0.85
	Center 1	69	0.48	0.68	0.34	0.51	0.95	0.15	0.42	0.61
	Center 2	46	0.64	0.52	0.66	0.29	2.08	0.54	0.2	0.89
	Center 3	56	0.65	0.67	0.64	0.59	0.62	0.57	0.52	0.78
	Center 4	6	0.47	1	0.42	0.25	5.83	0.30	0.15	1
	Center 5	13	0.77	0.69	0.79	0.46	1.31	0.73	0.35	0.94

TABLE 6. Leman-PV Performance for New and Enlarged Lesions for Different Scanner Magnetic Field Strengths 1.5 T and 3 T

Lesion Type	Magnetic Field at Both TP	N. Lesions	Accuracy	Sensitivity	Specificity	F1-score	aFPR	aTNR	PPV	NPV
All lesions	1.5 T	163	0.65	0.55	0.58	0.48	1.02	0.46	0.39	0.79
	3 T	27	0.66	0.66	0.66	0.28	3.03	0.55	0.18	0.94
New	1.5 T	98	0.5	0.62	0.58	0.45	0.72	0.64	0.413	0.71
	3 T	23	0.61	0.72	0.70	0.39	1.47	0.72	0.291	0.91
Enlarged	1.5 T	65	0.88	0.49	0.59	0.52	1.47	0.55	0.37	0.92
	3 T	4	1	0.61	0.62	0.14	12	0.63	0.08	1

TABLE 7. Leman-PV Performance for Different Image Quality Scores

Score	Lesion Type	N. Lesions	Accuracy	Sensitivity	Specificity	F1-score	aFPR	aTNR	PPV	NPV
Score = 1	All	132	0.68	0.69	0.67	0.49	1.11	0.6	0.38	0.88
	New	81	0.69	0.56	0.73	0.49	0.73	0.76	0.43	0.82
	Enlarged	51	0.66	0.9	0.61	0.5	1.71	0.66	0.35	0.96
Score = 2	All	55	0.46	0.58	0.41	0.36	1.69	0.24	0.26	0.74
	New	38	0.47	0.45	0.48	0.35	1.08	0.46	0.29	0.64
	Enlarged	17	0.44	0.88	0.35	0.36	3.06	0.39	0.22	0.93
Score = 3	All	3	0.52	0.33	0.55	0.14	3.33	0.33	0.09	0.86
	New	2	0.54	0.5	0.55	0.25	2.5	0.56	0.17	0.86
	Enlarged	1	0.5	0	0.55	0	5	0.5	0	0.86

The LeMan-PV PPV changes with respect to lesion volume for new and enlarged lesions were also studied. It will provide us with the percent of lesions that were correctly identified with respect to all lesions found by LeMan-PV (true and false positive lesions). LeMan-PV showed the highest PPV for new lesions with a volume lower than 300 mm^3. For the enlarged ones, lesions with lower 50 mm^3 were better identified (see Fig. S2a,b).

Discussion

The aim of this study was to assess the diagnostic accuracy of the software LeMan-PV[15] for the longitudinal detection of new and enlarging MS lesions in a multicenter controlled clinical settings. These are current radiological metrics of disease activity and are used for diagnostic, prognostic, and therapy evaluation purposes.[29]

Compared to the available automatic tools for identifying new and enlarged lesions,[28,30,31] LeMan-PV has the advantage of being integrated into the image reconstruction system, which facilitates routine use. However, multicenter validation data on the longitudinal version of LeMan-PV are lacking.

In the current study, we assessed the performance of LeMan-PV in a multicenter dataset obtained from five centers participating in the SMSC study, where protocols have been standardized, but data are acquired in a clinical setting with sequences obtained from MRI different machines.[31]

In this multicenter dataset, including data acquired at 1.5 and 3 T MRI, LeMan-PV reached a sensitivity of 65% and an aFPR of 1.31 for both new and enlarged lesions, which is lower than the 87% reported in a single-center study at 3 T.[15] This may be due to protocol variations compared to the recommended one for LeMan-PV, different hardware and software versions, as well as nonconsistent repositioning that

sometimes made the co-registration process, that is one crucial LeMan-PV methodology step, challenging.

Compared to two previous longitudinal studies on multicenter data that exploited a convolutional neural network for the detection of new and enlarged lesions, LeMan-PV showed 1) better sensitivity (65% vs. 60%) but a higher aFPR (1.31 vs. 0.48) than the Krüger et al.'s study of comparable size (217 patients)[32]; and 2) lower sensitivity than the second study by McKinley et al (65% vs. 72%) where the number of MRI pairs was lower ($n = 53$) and obtained from only two centers.[30]

The higher aFPR exhibited by LeMan-PV compared to the Krüger et al. study[32] might be due to challenges in co-registration between images acquired at different time points in a clinical scenario but also to the fact that the radiological reference standard is less susceptible to detect small and local enlargement of lesions with any amount of enlargement than the applied software. LeMan-PV detects enlargements as small as 10 voxels in MS lesions, independent of lesion size. In contrast, the radiological standard is to identify a greater enlargement (at least 50% of the main diameter). This explanation is supported by the fact that, in the current study, the aFPR for enlarged lesion detection was more than double that of new lesion detection.

As to new lesions, LeMan-PV showed a slightly better performance than the CNN-based approaches by Krüger et al and McKinley et al in a multicenter setting[30,32] (Se = 65% LeMan-PV vs. 60% in the two other longitudinal studies). The approach in the current study (k-nearest neighbor [kNN]) is different from that in these previous studies (neural networks deep-learning based). We expect that a multicenter setting favors a machine-learning algorithm more than one based on the kNN. In contrast to kNN-based classifiers, neural networks have a higher power of generalization (learning from the data, estimating a model for the data); they discover relevant features and their relative levels of importance from the training data. So one would expect that this type of algorithm would perform better if they are trained in more heterogenous data and therefore have better performance classifying new cases. Future head-to-head studies should further investigate and confirm these findings.

LeMan-PV detects changes in lesion number and size, exploiting the difference in FLAIR and MPRAGE contrast at different time points. In this study, a threshold was applied to the joint difference image provided by LeMAN-PV to detect new and enlarged lesions (Fig. 3a). Thresholding did not notably improve our results, but the number of false-positive lesions slightly decreased. The improved thresholding effects have also been observed in a previous study by Cabezas et al.[28]

When looking at the influence of the MRI field strength on the performance of LeMan-PV, in the detection of the new and enlarged lesions in all centers together, the sensitivity was better when both follow-ups were at 3 T than at 1.5 T, but this was at the penalty of a higher aFPR. The higher sensitivity but poorer aFPR values in detecting new and enlarged lesions at 3 T might be due to a better contrast of the images depicting the new and enlarging lesions. The poor aFPR values might also be due to co-registration errors or the contrast difference between FLAIR images at baseline and follow-up. At the same time, the LeMan-PV algorithm was mainly "trained" with data acquired at 3 T.[15]

The best results in detecting the new and enlarged lesions together were observed at 1.5 T in center 3. It was possible due to the higher number of lesions analyzed at this field strength in this center compared to other centers.

All MRI pairs were classified according to their quality (quality of the images and the postprocessing parameters like the co-registration, distortion correction, and prescan normalization) in the five centers. The scans that were classified as having the optimal quality for the analysis proved to increase the specificity in detecting new and enlarging lesions compared to the ones with a lower rate.

Limitations

The number of MRI scans presenting new or enlarged lesions should be increased for a better LeMan-PV performance evaluation. One crucial part of the LeMan-PV is that its performance depends on cross-sectional lesion segmentation, which is highly protocol dependent. While consistent protocols were used in the current study, deep-learning-based approaches might possibly overcome this limitation and further improve longitudinal lesion detection. In this type of study, the quality of the "ground-truth" influences the results. Therefore, using one reader and another to evaluate interrater reliability could have affected the LeMan-PV performance evaluation.

Conclusion

In this multicenter study using clinical data settings acquired at 1.5 T and 3 T, and with variations in MRI protocols, LeMan-PV showed similar sensitivity in detecting new and enlarged lesions with respect to other recent multicentric studies based on neural networks that only used 3 T scans.[30,32] While LeMan-PV performance is not optimal, its main advantage compared to previous methods is that it provides an automated clinical decision support integrated into the radiological routine flow.

Acknowledgments

The authors thank all the patients that participate in this study. The Swiss Multiple Sclerosis Cohort study received funding from the Swiss MS Society and unrestricted grant funding from Biogen, Celgene, Merck, Novartis, Roche, and Sanofi. Dr Cagol is supported by the Horizon 2020 Eurostar program (grant E!113682). Dr Granziera is supported by a Swiss National Science Foundation grant (PP00P3_176984),

the Stiftung zur Förderung der gastroenterologischen und allgemeinen klinischen Forschung, and the Horizon 2020 Eurostar program (grant E!113682). Open access funding provided by Universitat Basel.

Conflicts of Interest

Dr. Barakovic is an employee of Hays plc and a consultant for F. Hoffmann-La Roche Ltd. Dr. Cagol received compensation for activities with Actelion, Almirall, Bayer, Biogen, Celgene, Sanofi-Genzyme, Merck, Novartis, Roche, Teva, all for hospital research funds. Dr. Achtnichts served on scientific advisory boards for Celgene, Novartis Pharmaceuticals, Merck, Biogen, Sanofi Genzyme, Roche and Bayer; received funding for travel and/or speaker honoraria from Celgene, Biogen, Sanofi Genzyme, Novartis, Merck Serono, Roche, Teva and the Swiss MS Society; and research support from Biogen, Sanofi Genzyme, and Novartis. Dr. Lalive received honoraria for speaking and/or travel expense from Biogen, Merck, Novartis, Roche; consulting fees from Biogen, GeNeuro, Merck, Novartis, Roche; research support from Biogen, Merck, Novartis. None were related to this work. BMS. Dr. Müller received honoraria for travel, honoraria for lectures/consulting, and/or grants for studies from Almirall, Biogen, Celgene, Novartis, Teva, Merck Serono, Genzyme, Roche, and Bayer Schweiz. Dr. Pot received consulting fees and/or travel compensation, used exclusively for research support, for activities with Biogen, Merck, Novartis, Roche and Sanofi Genzyme. Dr. Salmen received speaker honoraria and/or travel compensation for activities with Bristol Myers Squibb, CSL Behring, Novartis, and Roche, and research support by the Baasch Medicus Foundation, the Medical Faculty of the University of Bern and the Swiss MS Society, not related to this work. Ente Ospedaliero Cantonale (employer) received compensation for Dr. Zecca's speaking activities, consulting fees, or research grants from Almirall, Biogen Idec, Bristol Meyer Squibb, Lundbeck, Merck, Novartis, Sanofi, Teva Pharma, Roche. Dr. Derfuss received speaker fees, research support, travel support, and/or served on Advisory Boards, data safety monitoring boards, or Steering Committees of Actelion, Alexion, Celgene, Polyneuron, Novartis Pharma, Merck Serono, Biogen, Teva, Bayer-Schering, GeNeuro, Mitsubishi Pharma, MedDay, Roche, and Genzyme. Dr. Yaldizli's institution (University Hospital Basel) received honoraria for lectures from Teva, Novartis and Bayer Schering exclusively used for funding of research or educational courses. Dr. Kappos' employer (University Hospital Basel) has received and dedicated to research support steering committee, advisory board, and consultancy fees (Abbvie, Actelion, Almirall, Auriga Vison AG, Bayer HealthCare, Biogen, Eisai, EMD Derono Inc, Genzyme, Genentech Inc, F. Hoffmann-La Roche, Japan Tobacco, Janssen Pharmaceuticals Inc, Merck, Minoryx Therapeutics SL, Novartis, Sanofi, Santhera, Senda Biosciences, Shionogi BV, TG Therapeutics); speaker fees (Bayer HealthCare, Biogen, Celgene, Genzyme, Janssen Pharmaceuticals Inc, Merck, Novartis, Roche, and Sanofi); support of educational activities (Allergan, Bayer HealthCare, Biogen, CSL Behring, Genzyme, Merck, Novartis, Roche, Pfizer, Sanofi, Shire, and Teva); license fees for Neurostatus products; and grants (Bayer HealthCare, Biogen, European Union, Innosuisse, Merck, Novartis, Roche Research Foundation, Swiss MS Society, and Swiss National Research Foundation). Ente Ospedaliero Cantonale (employer) received compensation for Dr. Gobbi's speaking activities, consulting fees, or research grants from Almirall, Biogen Idec, Bristol Meyer Squibb, Lundbeck, Merck, Novartis, Sanofi, Teva Pharma, Roche. The University Hospital Basel (USB), as the employer of Dr. Granziera has received the following fees which were used exclusively for research support: (i) advisory board and consultancy fees from Actelion, Novartis, Genzyme and F. Hoffmann-La Roche; (ii) speaker fees from Biogen and Genzyme-Sanofi; (iii) research support by F. Hoffmann-La Roche Ltd. Dr Granziera is also supported by a Swiss National Science Foundation grant (PP00P3_176984), the Stiftung zur Förderung der gastroenterologischen und allgemeinen klinischen Forschung, and the Horizon 2020 Eurostar program (grant E!113682). Dr. Kuhle received speaker fees, research support, travel support, and/or served on advisory boards by the Progressive MS Alliance, Swiss MS Society, Swiss National Research Foundation (320030_189140/1), University of Basel, Biogen, Celgene, Merck, Novartis, Octave Bioscience, Roche, Sanofi. Dr. Kober, Dr. Marechal and Corredor-Jerez are employed of Siemens Healthineers International, Lausanne, Switzerland. The Swiss Multiple Sclerosis Cohort study received funding from the Swiss MS Society and unrestricted grant funding from Biogen, Celgene, Merck, Novartis, Roche, and Sanofi. No other disclosures were reported. For the remaining authors none were declared.

REFERENCES

1. Lassmann H, Brück W, Lucchinetti CF. The immunopathology of multiple sclerosis: An overview. Brain Pathol 2007;17:210-218.
2. Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol 2018;17:162-173.
3. 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.
4. Lassmann H. Targets of therapy in progressive MS. Mult Scler 2017;23:1593-1599.
5. Río J, Castilló J, Rovira A, et al. Measures in the first year of therapy predict the response to interferon beta in MS. Mult Scler 2009;15:848-853.
6. Sormani MP, Bruzzi P. MRI lesions as a surrogate for relapses in multiple sclerosis: A meta-analysis of randomised trials. Lancet Neurol 2013;12:669-676.
7. Egger C, Opfer R, Wang C, et al. MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? Neuroimage (Amst) 2017;13:264-270.

8. Altay EE, Fisher E, Jones SE, Hara-Cleaver C, Lee JC, Rudick RA. Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. JAMA Neurol 2013;70: 338-344.
9. Solomon J, Sood A. 4-D lesion detection using expectation-maximization and hidden markov model. *2004 2nd IEEE Int Symp Biomed Imaging Macro to Nano*, Vol 1 United States: IEEE; 2004. p 125-128.
10. Köhler C, Wahl H, Ziemssen T, Linn J, Kitzler HH. Exploring individual multiple sclerosis lesion volume change over time: Development of an algorithm for the analyses of longitudinal quantitative MRI measures. NeuroImage Clin 2019;21:101623.
11. Lladó X, Ganiler O, Oliver A, et al. Automated detection of multiple sclerosis lesions in serial brain MRI. Neuroradiology 2012;54:787-807.
12. Moraal B, Wattjes MP, Geurts JJG, et al. Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. Radiology 2010;255: 154-163.
13. Jain S, Ribbens A, Sima DM, et al. Two time point MS lesion segmentation in brain MRI: An expectation-maximization framework. Front Neurosci 2016;10.
14. Fartaria MJ, Todea A, Kober T, et al. Partial volume-aware assessment of multiple sclerosis lesions. NeuroImage Clin 2018;18:245-253.
15. Fartaria MJ, Kober T, Granziera C, Bach Cuadra M. Longitudinal analysis of white matter and cortical lesions in multiple sclerosis. NeuroImage Clin 2019;23:101938.
16. Fartaria MJ, Bonnier G, Roche A, et al. Automated detection of white matter and cortical lesions in early stages of multiple sclerosis. J Magn Reson Imaging 2016;43:1445-1454.
17. Disanto G, Benkert P, Lorscheider J, et al. The Swiss multiple sclerosis cohort-study (SMSC): A prospective Swiss wide investigation of key phases in disease evolution and new treatment options. PLoS One 2016;11:e0152347.
18. Kurtzke JF. Rating neurologic impairment in multiple sclerosis. Neurology 1983;33:1444-1452.
19. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31:1116-1128.
20. Filippi M, Preziosa P, Banwell BL, et al. Assessment of lesions on magnetic resonance imaging in multiple sclerosis: Practical guidelines. Brain 2019;142:1858-1875.
21. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: A toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29:196-205.
22. Rovira A, Tintoré M, Álvarez-Cermeño JC, Izquierdo G, Prieto JM. Recommendations for using and interpreting magnetic resonance imaging in multiple sclerosis. Neurologia 2010;25:248-265.
23. Valverde S, Oliver A, Lladó X. A white matter lesion-filling approach to improve brain tissue volume measurements. NeuroImage Clin 2014;6: 86-92.
24. McHugh ML. Interrater reliability: The kappa statistic. Biochem Med 2012;22:276-282.
25. Liljequist D, Elfving B, Roaldsen KS. Intraclass correlation – A discussion and demonstration of basic features. PLoS One 2019;14:e0219854.
26. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15: 155-163.
27. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420-428.
28. Cabezas M, Oliver A, Roura E, et al. Automatic multiple sclerosis lesion detection in brain MRI by FLAIR thresholding. Comput Methods Prog Biomed 2014;115:147-161.
29. Kaunzner UW, Gauthier SA. MRI in the assessment and monitoring of multiple sclerosis: An update on best practice. Ther Adv Neurol Disord 2017;10:247-261.
30. McKinley R, Wepfer R, Grunder L, et al. Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. NeuroImage Clin 2020;25:102104.
31. Disanto G, Berlanga AJ, Handel AE, et al. Heterogeneity in multiple sclerosis: Scratching the surface of a complex disease. Autoimmune Dis 2010;2011:1-12.
32. Krüger J, Opfer R, Gessert N, et al. Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. NeuroImage Clin 2020;28:102445.