ASSESSMENT OF AUTOMATIC DECISION-SUPPORT SYSTEMS FOR DETECTING ACTIVE T2 LESIONS IN MULTIPLE SCLEROSIS PATIENTS

Alex Rovira[a,b,c], Juan Francisco Corral[a,b], Cristina Auger[a,b,c], Sergi Valverde[d], Angela Vidal-Jordana[e,f,c], Arnau Oliver[d], Andrea de Barros[a], Yiken Karelys Ng Wong[a], Mar Tintoré[e,f,c], Deborah Pareto[a,b], Francesc Xavier Aymerich[a,b,c,g], Xavier Montalban[e,f,c], Xavier Lladó[d], Juli Alonso[a,b,c]

[a]Neuroradiology Section, Department of Radiology (IDI), Vall d'Hebron University Hospital, Barcelona, Spain

[b]Neuroradiology Research Group, Vall d'Hebron Research Institute (VHIR), Barcelona, Spain

[c]Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

[d]Department of Computer Architecture and Technology, University of Girona, Girona, Spain.

[e]Department of Neurology and Neuroimmunology, Centre d'Esclerosi Múltiple de Catalunya, (Cemcat), Vall d'Hebron University Hospital, Barcelona, Spain

[f]Clinical Neuroimmunology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain

[g]Automatic Control Department. Universitat Politècnica de Catalunya. Barcelona Tech. Barcelona, Spain

Corresponding author:

Dr Alex Rovira, Neuroradiology Section, Department of Radiology, Hospital Universitari Vall d'Hebron, Pg Vall d'Hebron 119-129, 08035 Barcelona, Spain

Email: alex.rovira.idi@gencat.cat

Telephone: +34 93 4286034

FAX: +34 93 4286059

Part of this study has been presented at the Annual Meeting of the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS) 2020 (September 11-13, 2020)

Abbreviations: MS = multiple sclerosis; DSS = decision-support system; CNN = convolutional neural networks; NEDA = no evidence of disease activity; MEDA = minimal evidence of disease activity; RONL = reference outcome of a active lesion.

ABSTRACT

Background: Active (new/enlarging) T2 lesion counts are routinely used in the clinical management of multiple sclerosis (MS). Thus, automated tools able to accurately identify active T2 lesions would be of high interest to neuroradiologists for assisting in their clinical activity.

Objective: To compare the accuracy in detecting active T2 lesions and of radiologically active patients based on different visual and automated methods.

Methods: One hundred multiple sclerosis patients underwent two magnetic resonance imaging examinations within 12 months. Four approaches were assessed for detecting active T2 lesions: 1) conventional neuroradiological reports; 2) prospective visual analyses performed by an expert; 3) automated unsupervised tool; and 4) supervised convolutional neural network. As a gold standard, a reference outcome was created by the consensus of two observers.

Results. The automated methods detected a higher number of active T2 lesions, and a higher number of active patients, but a higher number of false-positive active patients than visual methods. The convolutional neural network model was more sensitive in detecting active T2 lesions and active patients than the other automated method.

Conclusion. Automated convolutional neural network models show potential as an aid to neuroradiological assessment in clinical practice, although visual supervision of the outcomes is still required.

Keywords: Multiple sclerosis, new lesions, disease activity, convolutional neural network, automatic new lesion detection, magnetic resonance imaging.

Abbreviations: MS = multiple sclerosis; DSS = decision-support system; CNN = convolutional neural networks; NEDA = no evidence of disease activity; MEDA = minimal evidence of disease activity; RONL = reference outcome of a active lesion.

INTRODUCTION

Multiple sclerosis is characterized by the presence of demyelinating lesions scattered throughout the central nervous system (CNS). [1] Magnetic resonance imaging (MRI) has an established role in diagnosis, in assessing disease activity, and in predicting and monitoring treatment efficacy and is commonly used to determine outcome measures in clinical trials. [2,3] The presence of new/enlarging demyelinating lesions as visually detected on serial T2-weighted images is used to assess disease activity. [4,5] However, visual identification is challenging, particularly in studies with repositioning deficiencies or in the presence of diffuse and confluent chronic MS lesions, and is usually accompanied by low sensitivity and high variability across raters. [6] Moreover, visual assessment, which requires a certain degree of expertise, is a time-consuming task that slows down the reporting process and, therefore, radiologists' productivity. [7] Different research groups have developed automated software packages to help neuroradiologists identify and count new/enlarging T2 lesions. [8–13] Our group has developed a decision-support system (DSS) based on an unsupervised approach that uses intensity-derived features from subtraction images together with deformation field information obtained from the nonrigid registration between two MR scans acquired at different time points. [14] More recently, we proposed another DSS, a supervised approach based on the application of convolutional neural networks (CNNs) to basal and follow-up input modalities previously trained to detect the presence of new/enlarging T2 lesions. [15]

To introduce these automated DSSs in clinical practice, it is necessary to evaluate their performance and compare it with that of visual methods. Thus, the objective of our study was to analyze new/enlarging T2 lesion detection using the two

automated DSSs previously described and compare this performance with that of two visual methods. As a second objective, we compared these tools in terms of their identification of radiologically active patients.

MATERIAL AND METHODS

Patients

A single-center cohort of 100 MS patients (61 women) with a mean age of 39.6±10.7 years (range 18-69 years) participated in this observational, retrospective study. We recruited patients who had undergone two MRI examinations on the same device and using the same MRI protocol as part of their routine clinical assessment at our institution between December 2016 and September 2019 (mean time between MRI examinations 12 months, range 3-27 months). Fifteen percent of these patients had clinical relapses during the interval between the two MRI scans.

All procedures performed were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The protocol was approved by the hospital research and ethics committee, and informed consent was obtained from each participant.

MRI exams

MRI was performed on a 3 T scanner (MAGNETOM Trio A Tim System, Siemens Healthcare, Erlangen, Germany) equipped with a 12-channel head matrix receiver coil, with the body coil acting as a transmitter.

The standardized protocol included the following pulse sequences: 1) a 3D sagittal T1-weighted magnetization-prepared rapid gradient echo (MPRAGE) sequence (repetition time [TR]=2300 ms, echo time [TE]=2.98 ms, inversion time [TI]=900 ms, voxel size=1.0 × 1.0 × 1.0 mm^3); 2) a 2D transverse proton density and T2-weighted turbo spin-echo (TSE) sequence (TR = 2500 ms; TE = 16/91 ms; voxel size = 0.78 × 0.78 × 3.0 mm^3; and 3) a 3D sagittal fast fluid-attenuated inversion recovery (FLAIR) sequence (TR = 5000 ms, TE = 394 ms, TI = 1800 ms, flip angle = 120°, voxel size = 1.0 × 1.0 × 1.0 mm^3).

MRI analysis

Four different methods were used to determine the number of new/enlarging T2 lesions. These methods included 1) Visual analysis 1 (V1), in which the number of new/enlarging T2 (assessed on non-processed FLAIR and dual echo T2-weighted sequences) lesions was obtained as described in the radiological report performed under routine conditions (nonblinded to clinical information) by full-time academic radiologist at our institution; 2) Visual analysis 2 (V2), in which the number of new/enlarging T2 lesions (also assessed on non-processed FLAIR and dual echo T2-weighted images) was obtained after a new review of the MRI scans by an expert observer (a technician with more than 15 years of experience in assessing new T2 lesions in MS under neuroradiologist supervision) nonblinded to the number and topography of new/enlarging T2 lesions described in the radiological report; 3) Automated tool 1 (A1), in which the number of new/enlarging T2 lesions was found by an unsupervised approach that used intensity-derived features from T1-weighted and FLAIR subtraction images together with deformation field information obtained from the nonrigid

registration between the two MRI scans[14]; and 4) Automated tool 2 (A2), in which the number of active T2 lesions was found by a supervised approach based on the application of a CNN trained on the 4 available input sequences to detect the presence of new T2 lesions in the follow-up scan[15]. The CNN model was trained from MRI data acquired in an independent cohort of 35 MS patients using the same scanner, imaging protocol and preprocessing steps.

MR images for both automated approaches (A1) and (A2) were processed following the same procedure: for each subject, the brain mask was first identified on the T1-w image using ROBEX[16] and then applied to the rest of input sequences. Then, the four image sequences (T1, FLAIR, T2 and proton density) underwent bias field correction using the N4 algorithm.[17] Afterwards, all baseline images co-registered and warped to the follow-up space. To do so, each image sequence was first linearly registered to its correspondent proton density sequence using NiftyREG package[18]. Then, the baseline proton density sequence was linearly registered to the proton density follow-up while the T1 weighted, T2 and FLAIR baseline sequences were warped to the follow-up space using the combined affine transformation to avoid unnecessary interpolations. Finally, for each patient and image sequence, image intensities across baseline and the follow-up sequences were normalized using a histogram matching approach [19].

A new T2 lesion was visually defined as an area of high signal intensity that arises in an area of previously normal brain tissue signal, and an enlarged T2 lesion was defined as an area of high signal intensity that has increased in diameter by at least 100% or increased in size on at least two consecutive slices.

Radiologically active patients were defined in two ways. The first was the presence of ≥ 1 active T2 lesion based on the no evidence of disease activity (NEDA)

concept[20-22], and the second was the presence of > 2 new/enlarging T2 lesions based on the less stringent concept of minimal evidence of disease activity (MEDA), which represents a more realistic treatment goal in clinical practice [22,23]. Both groups of radiologically active patients were evaluated.

Finally, to evaluate the accuracy of the different methods, a "reference outcome of a active lesion" (RONL) was created as a "gold standard" by the consensus of two observers (the technician that performed the V2 analysis and a neuroradiologist) with more than 15 years of experience. In detail, RONL was created by comparing both automated methods with V2 results, using the non-processed source MR images as reference. We then identified all detected new T2 lesions with the automated tools and confirmed their presence on the source images with the help of the mask created on V2. Finally, we also verified if all lesions marked with V2 were identified with the automated methods.

Statistical analysis

Data analysis was performed with the Statistical Package for the Social Sciences (SPSS), version 25.0 (IBM Corp., USA). The number of new/enlarging T2 lesions detected by each method is expressed as the mean ± SD. Significant differences between lesion counts obtained by the different methods were evaluated with repeated-measures one-way ANOVA followed by a post hoc test (Bonferroni test) to evaluate pairwise differences. Significant differences between volume of true positive and false lesions were evaluated with Mann–Whitney *U*-test. For all analyses, $p < 0.05$ was considered significant. RONL information was used to define a truly radiologically active patient. Sensitivity, specificity, and accuracy were determined with MedCalc (https://www.medcalc.org/calc/diagnostic_test.php,

MedCalc Software, Ostend, Belgium). Confidence intervals for sensitivity, specificity, and accuracy are "exact" Clopper-Pearson confidence intervals. The sensitivity and accuracy with 95% confidence intervals in detecting new/enlarged lesions were calculated only for the automated methods but not for the visual methods, as active T2 lesions were only individually digitally identified with the former. However, the sensitivity, specificity, and accuracy with 95% confidence intervals in detecting the number of active patients were determined for all four methods.

RESULTS

The RONL technique identified 104 new/enlarging T2 lesions in 38 of the 100 patients included in the study that showed at least one truly new/enlarging T2 lesion, while V1 detected 59 lesions, V2 detected 73 lesions, A1 125 lesions, and A2 119 lesions. The automated tools counted approximately double the number of lesions that the visual methods identified. Table 1 and Figure 1 show the mean number of lesions per patient for the entire cohort, for the cohort of 53 patients among whom at least one of the methods identified at least one new/enlarging lesion, and for the 38 patients with at least one true new/enlarging T2 lesion.

The V2 and A1 methods showed slightly larger numbers of new/enlarged T2 lesions than the V1 and A2 methods, respectively, although these differences were not significant (Figure 1). However, the differences between both visual methods and the automated tools were significant. A2 provided a number of new/enlarging T2 lesions (total positive cases) more similar to the RONL than all other methods. Figure 2 shows examples of the performance of the different methods.

The performance of A2 was better than that of A1, with a lower number of false negative and false positive results and a higher number of true positive lesions (Table 2). Accordingly, the sensitivity and accuracy of A2 were higher than those of A1.

The combined assessment of both V1 and A2 showed 89 new lesions (Table 1) with 15 false-negative lesions but no false-positive lesions, achieving higher sensitivity for detecting new T2 lesions than any of the both automated methods tested (Table 2).

The volume of new T2 lesions was significantly lower for false-negative lesions compared to true-positive lesions for each of the four methods assessed (Table 3), but there were no differences in the topography (periventricular, juxtacortical, subcortical, infratentorial) of true-positive and false-negative lesions in each of the four methods assessed.

Thirty-eight patients were defined as active according to the presence of at least one truly new/enlarging T2 lesion in the RONL. As shown in Table 4, the visual methods identified more true-negative patients than the automated tools. Furthermore, the automated tools produced a higher number of false-positive patients and a smaller number of false-negative patients than the visual methods. A comparison of the visual methods showed that V2 correctly identified more active patients than V1. Similarly, a comparison of the automated methods showed that A2 correctly identified more active patients and fewer false-negative patients than A1. The specificity in identifying radiologically active patients was higher for the visual methods than for the automated methods. In contrast, the automated tools achieved a higher sensitivity than the visual methods, with A2 exhibiting the highest sensitivity. When comparing the automated methods, A2 presented higher

sensitivity and specificity. Accuracy was highest for V2, followed by V1 and A2, and A1 exhibited the lowest accuracy. When combining the V1 and A2 assessment, we achieved 100% specificity, a slight decrease in sensitivity than the A2 method, but the highest accuracy of all methods.

When the cutoff used to define active patients was increased to >2 truly new/enlarging T2 lesions, the number of true-positive patients according to the RONL decreased to 15 (a 60% reduction with respect to the 38 patients found with the previously used criteria). All methods subsequently reported a low number of true-positive patients. The findings were similar to those previously described, but the differences between the visual and automated methods were smaller. The sensitivity decreased for all the methods, with automated methods still providing the highest values. However, the specificity was approximately the same for the visual methods and increased for the automated methods. Finally, the accuracy was more similar among the different methods than when using the previous cutoff criterion. When combining the V1 and A2 assessment, we achieved 100% specificity, a slight decrease in sensitivity than the A2 method, but the highest accuracy of all methods (Table 5).

DISCUSSION

New/enlarging T2 lesion detection is one of the most relevant MRI biomarkers for evaluating disease activity and for monitoring and predicting treatment response in MS [3,24,25], although it is a time-consuming task requiring observer expertise [6,26,27]. Different research groups have dedicated various efforts to develop automated tools to obtain this information rapidly and accurately. In this study, we analyzed and compared two of these tools[14,15], previously developed by our

research group, and further compared them to visual assessments using a generated RONL as the gold standard. The results showed significant differences in the number of new/enlarging T2 lesions detected by each method and in the number of radiologically active patients.

As expected, false-negative lesions had significant smaller volume than true-positive lesions, indicating the limitation of both visual assessments and automated tools for detecting new small lesions.

The number of new/enlarging T2 lesions detected by each method revealed that none of the methods is perfect, as all of them were prone to different errors. The evaluation of the visual methods showed that conventional radiological reports (V1) found a lower number of active T2 lesions than the analysis performed by an expert observer under optimized conditions (V2). This is not unexpected, as the expert observer used the information included in the radiological report, and his sole objective was to focus on the identification of new/enlarging T2 lesions, without the time pressure found in reporting tasks in clinical practice.

Although there were no significant differences between the DSS methods or between either DSS and the RONL, the supervised CNN approach (A2) produced a lower number of false negative and false positive results than the unsupervised approach (A1). Since the number of true positive results yielded by A2 was more similar to RONL than that obtained by A1, and that the number of false negative results yielded by A2 was smaller than that obtained by A1, we can hypothesize that the supervised CNN approach is very well suited to evaluate new/enlarging T2 lesions. This is also supported by the quantitative analysis, showing that A2 yielded better scores than A1 in false positive, true positive and, most importantly, false negative lesion detection. A recent study[28] also highlighted the value of a CNN

approach in the evaluation of new/enlarging T2 lesions over other already available tools, such as the open-source Lesion Segmentation Toolbox.

Finally, the combined assessment of both V1 and A2, which is the most realistic approach in clinical practice, achieved a higher sensitivity compared to any of the both automated methods tested, supporting their value in assessing disease activity.

We also assessed the accuracy of the different methods in identifying active patients using two different definitions, one based on the NEDA concept[20-22] and the other on the less stringent MEDA concept[22,23]. The results showed that the neuroradiologist correctly identified radiologically active patients by eye without false positives. This high specificity was obtained at the cost of missing a substantial number of active patients. In clinical practice, this could be of concern, as we would miss disease activity in a relevant percentage of patients, which may negatively impact their proper treatment management. In contrast, the automated methods detected nearly all true positive patients but included a larger number of false positive patients than the visual methods. This means that these automated tools identified some truly radiologically inactive patients as radiologically active.

Our results show that fully automated tools are well suited to detect the presence of disease activity in MS patients. Nevertheless, their suboptimal specificity and sensitivity preclude their use in clinical practice without visual supervision of the outcomes by an expert observer. In addition, these results encourage further research with the main objective of improving their performance by decreasing the number of false negatives and false positives to reach sufficiently good accuracy to avoid the need for expert visual supervision. Increasing the MRI data set used to train the CNN model, could certainly improve its performance.

A2 exhibited a slightly better sensitivity and specificity than A1 under the NEDA concept. However, under the MEDA concept, A2 exhibited a better sensitivity and better identification of patients with active disease but at the cost of a lower specificity than A1. In light of these results and given the previously revealed finding that the number of new/enlarging T2 lesions yielded by A2 was more similar to the RONL than that yielded by A1, we can hypothesize that the supervised CNN approach is well suited not only to evaluate new/enlarging T2 lesions but also to identify active patients. These results are in agreement with a previous work that noted improved results when using a supervised classification model instead of an unsupervised rule-based approach[15].

There is increasing pressure for neuroradiologists to include quantitative information in their reports, which could certainly increase their clinical value. This leads to an increasing need for robust and available tools that require minimal or no human supervision, such as those provided by automated DSS. New/enlarging T2 lesions are usually assessed visually, although with suboptimal sensitivity and interrater and intrarater concordance[6,24]. This is of particular concern for patients with a high lesion load or when the MRI scans are poorly repositioned, leading to an underestimation of the number of new/enlarging T2 lesions, especially of small lesions. The use of a DSS method can, at least partially, solve these problems despite requiring an expert observer to review the outcomes. However, this is a much easier task than the complete visual detection of new/enlarged T2 lesions. Moreover, and probably more importantly, the number of false negatives obtained with both automated methods is lower than that obtained with the visual methods.

As a limitation of the present study, we note that we studied a relatively small cohort of 100 patients, and consequently, the number of radiologically active patients was not high (38% and 15% with the NEDA and MEDA concepts, respectively). Therefore, the specificity and sensitivity obtained in this work may be affected by the small number of active patients. Moreover, we think that generalization of the performance results between visual and automated methods is limited by the fact that we used the same MRI scanner and the same standardized MRI protocol across all the studies. The performance may therefore differ with different input data due to MR system and pulse sequence differences. Finally, it is not possible to have a true active T2 lesion "gold standard" for defining the diagnostic properties of any test. To minimize the effects of this problem, we defined an RONL by the consensus of two expert observers based on the combined visual assessment of all MR images, the radiological report, and the outcomes of the automated methods. We also must acknowledge that the RONL may contain false-negative results as the set of potential false negatives is restricted to those detected by at least one of the four methods, which might result in an overestimation of the sensitivity of the automated methods tested.

CONCLUSIONS

In summary, the results of this study show that the automated methods have greater sensitivity but slightly lower specificity in detecting new/enlarging T2 lesions and active MS patients than conventional neuroradiological reports and expert visual analysis. This indicates that the automated tools should be further developed to allow their fully unsupervised use in clinical practice. From the comparison of both DSSs, we show that the DSS based on the application of a CNN

model, even when trained with a small number of cases, is more promising than automated unsupervised approaches in detecting MR disease activity. Finally, although not specifically assessed in this study, visually supervised automated tools likely provide a higher level of confidence in the detection of radiological activity than the standard radiological report and show potential as an aid to neuroradiological visual assessment in clinical practice.

REFERENCES

1. Filippi M, Preziosa P, Banwell BL, et al. Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. Brain. 2019;142(7):1858-1875.
2. Rovira A, Wattjes, MP, Tintore, M, et al. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-clinical implementation in the diagnostic process. Nature reviews Neurology 2015, 11: 471-478
3. Wattjes MP, Rovira A, Miller, D, et al. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-establishing disease prognosis and monitoring patients. Nature Reviews Neurology 2015, 11: 597-606
4. Fahrbach K, Huelin R, Martin AL, et al. Relating relapse and T2 lesion changes to disability progression in multiple sclerosis: a systematic literature review and regression analysis. BMC Neurol 2013, 13: 180
5. Río J, Auger C, Rovira À. MR Imaging in Monitoring and Predicting Treatment Response in Multiple Sclerosis. Neuroimaging Clin N Am. 2017;27:277-287
6. Altay E, Fisher E, Jones SE, et al., Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. JAMA Neurol 2013, 70: 338-344
7. Patriarche J, Erickson B. A review of the automated detection of change in serial imaging studies of the brain. J Digit Imaging 2004, 17: 158-174
8. Moraal B, Wattjes MP, Geurts JJ, et al. Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. Radiology 2010; 255:154–63
9. Lladó X, Ganiler O, Oliver A, et al. Automated detection of multiple sclerosis lesions in serial brain MRI. Neuroradiology 2012, 54: 787-807
10. Elliott C, Arnold DL, Collins DL, et al. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. IEEE Trans Med Imaging 2013; 32:1490–503

11. Sweeney EM, Shinohara RT, Shea CD, et al. Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. AJNR Am J Neuroradiol 2013; 34:68–73
12. Battaglini M, Rossi F, Grove RA, et al. Automated identification of brain new lesions in multiple sclerosis using subtraction images. J Magn Reson Imaging 2014; 39:1543–49
13. Danelakis A, Theoharis T, Verganelakis DA, et al. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. Comput Med Imaging Graph 2018, 70: 83-100
14. Cabezas M, Corral JF, Oliver A, et al. Improved Automatic Detection of New T2 Lesions in Multiple Sclerosis Using Deformation Fields. AJNR Am J Neuroradiol 2016, 37: 1816-1823
15. Salem M, Valverde S, Cabezas M, et al. A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. Neuroimage Clin 2020, 25: 102149
16. Iglesias J, Liu C, Thompson,P et al. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging, 2011, 30: 1617-1634
17. Tustison N, Avants B, Cook P, et al. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging, 2010, 29: 1310-1320
18. Modat M, Cash D, Daga P, et al. Global image registration using a symmetric block-matching approach. J. Med. Imaging, 2014, 1: 24003
19. Nyúl L, Udupa J, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging, 2000, 19: 143-150
20. Havrdova E, Galetta S, Hutchinson M, et al., Effect of natalizumab on clinical and radiological disease activity in multiple sclerosis: a retrospective analysis of the Natalizumab Safety and Efficacy in Relapsing-Remitting Multiple Sclerosis (AFFIRM) study. Lancet Neurol 2009, 8: 254-260
21. Stangel M, Penner IK, Kallmann BA, et al., Towards the implementation of 'no evidence of disease activity' in multiple sclerosis treatment: the multiple sclerosis decision model. Ther Adv Neurol Disord 2015, 8: 3-13.

22. Gasperini C, Prosperini L, Tintoré M, et al., Unraveling treatment response in multiple sclerosis: A clinical and MRI challenge. Neurology 2019; 92: 180-192.
23. 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
24. Sormani M, Bruzzi P. MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. Lancet Neurol 2013, 12: 669-676
25. Tintore M, Vidal-Jordana A, Sastre-Garriga J. Treatment of multiple sclerosis - success from bench to bedside. Nat Rev Neurol 2019, 15: 53-58
26. Molyneux PD, Miller DH, Filippi M, et al. Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. Neuroradiology 1999, 41: 882-888
27. Tan IL, van Schijndel RA, Fazekas, F, et al., Image registration and subtraction to detect active T(2) lesions in MS: an interobserver study. J Neurol 2002, 249: 767-773
28. 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

Tables

TABLE 1. Comparison of the number of active T2 lesions and the mean number of active T2 lesions per patient for the different methods and the reference outcome of active lesions (RONL)

	V1	V2	A1	A2	V1A2	RONL
New lesions	59	73	125	119	89	104
New lesion/patient	0.59±1.15[a,b,c,d]	0.73±1.29[a,c,d]	1.25±2.32	1.19±1.95[b]	0.89±1.62[a,c]	1.04±1.85
New lesion/patient[e]	1.11±1.38[a,b,c,d]	1.38±1.51[a,c,d]	2.36±2.75	2.25±2.20[b]	1.68±1.91[a,c]	1.96±2.17
New lesion/patient[f]	1.55±1.41[a,b,c,d]	1.92±1.46[a,c]	3.00±3.00	2.82±2.28	2.34±1.88[a]	2.74±2.10

RONL: reference outcome of new/enlarging lesions; V1: standard radiological report method; V2: visual review of the MR scans by an expert nonblinded to the radiological report; A1: automated unsupervised approach; A2: automated supervised convolutional neural network-based approach; V1A2: A2 method corrected with the report information.
[a] Significantly different with respect to RONL.
[b] Significantly different with respect to V1A2.
[c] Significantly different with respect to A2.
[d] Significantly different with respect to A1.
[e] Representing the 53 patients for whom at least one new/enlarging T2 lesion was detected by one method.
[f] Representing the 38 patients with truly new/enlarging T2 lesions.

TABLE 2. Performance of the visual and automated tools in lesion detection.

	V1	V2	A1	A2	V1A2
False negatives	47[a]	31[a]	27[a]	17[a]	15[a]
True positives	56[a]	70[a]	77[a]	87[a]	89[a]
False positives	3[a]	3[a]	48[a]	32[a]	30[a]
Sensitivity (CI)	54.37 (44.26-64.22)	69.31 (59.34–78.10)	74.04 (64.52-82.14)	83.65 (75.12-90.18)	85.58 (77.33–91.70)
Accuracy (CI)	52.83 (42.89–62.60)	67.31 (57.41–76.19)	50.66 (42.44-58.85)	63.97 (55.30-72.02)	NA

V1: standard radiological report method; V2: visual review of the MR scans by an expert nonblinded to the radiological report; A1: automated unsupervised approach; A2: automated supervised convolutional neural network-based approach; V1A2: A2 method corrected with the report information; CI: confidence interval; NA: not applicable. Sensitivity and accuracy values are provided with 95% interval of confidence.

[a] Numbers represent new/enlarging T2 lesions.

Table 3. Volume of true positive and false negative lesions for the different methods.

	V1	**V2**	**A1**	**A2**
False negative	55.0 ± 60.8 #	51.2 ± 44.9 #	58.9 ± 106.5 #	40.1 ± 40.1 #
True positive	142.7 ± 189.2	122.9 ± 150.8	123.3 ± 165.5	117.8 ± 162.9

V1: standard radiological report method; V2: visual review of the MRIs by an expert nonblinded to the radiological report; A1: automated unsupervised approach; A2: automated supervised convolutional neural network-based approach.

All values are expressed in mm 3 as the mean ± standard deviation.

Significant difference between false negative and true positive ($p < 0.05$).

TABLE 4. Performance of the different methods in identifying radiologically active patients when using the presence of at least one new/enlarging T2 lesion in the RONL as the cutoff criterion (NEDA criteria).

	V1	V2	A1	A2	V1A2
True negatives	62[a]	62[a]	52[a]	54[a]	62[a]
False negatives	9[a]	4[a]	3[a]	1[a]	2[a]
True positives	29[a]	34[a]	35[a]	37[a]	36[a]
False positives	0[a]	0[a]	10[a]	8[a]	0a
Sensitivity (CI)	76.32 (59.76-88.56)	89.47 (75.20-97.06)	92.11 (78.62-98.34)	97.37 (86.19-99.93)	94.74 (85.25-99.36)
Specificity (CI)	100.00 (94.22-100.00)	100.00 (94.22-100.00)	83.87 (72.33-91.98)	87.10 (76.15-94.26)	100.00 (94.22-100.00)
Accuracy (CI)	91.00 (83.60-95.80)	96.00 (90.07-98.90)	87.00 (78.80-92.89)	91.00 (83.60-95.80)	98.00 (92.96-99.76)

NEDA, no evidence of disease activity; V1, standard radiological report method; V2, visual review of the MRIs by an expert nonblinded to the radiological report; A1, automated unsupervised approach; A2, automated supervised convolutional neural network-based approach; CI, confidence interval.

Sensitivity, specificity and accuracy values are provided with 95% interval of confidence.

[a] Numbers represent patients with at least one new/enlarging T2 lesions.

TABLE 4.- Performance of the different methods in identifying radiologically active patients when using the presence of more than two active T2 lesions as the cutoff criterion (MEDA criteria).

	V1	**V2**	**A1**	**A2**	**V1A2**
True negatives	84[a]	85[a]	82[a]	79[a]	85[a]
False negatives	8[a]	5[a]	4[a]	3[a]	4[a]
True positives	7[a]	10[a]	11[a]	12[a]	11[a]
False positives	1[a]	0[a]	3[a]	6[a]	0[a]
Sensitivity (CI)	46.67 (21.27-73.41)	66.67 (38.38-88.18)	73.33 (44.90-92.21)	80.00 (51.91-95.67)	73.33 (44.90-92.21)
Specificity (CI)	98.82 (93.62-99.97)	100.00 (95.75-100.00)	96.47 (90.03-99.27)	92.94 (85.27-97.37)	100.00 (95.75-100.00)
Accuracy (CI)	91.00 (83.60-95.80)	95.00 (88.72-98.36)	93.00 (86.11-97.14)	91.00 (83.60-95.80)	96.00 (90.07-98.90)

MEDA, minimal evidence of disease activity; V1, standard radiological report method; V2, visual review of the MRIs by an expert nonblinded to the radiological report; A1, automated unsupervised approach; A2, automated supervised convolutional neural network-based approach; CI, confidence interval.

Sensitivity, specificity and accuracy values are provided with 95% interval of confidence.

[a] numbers represent patients with more than two active T2 lesions.

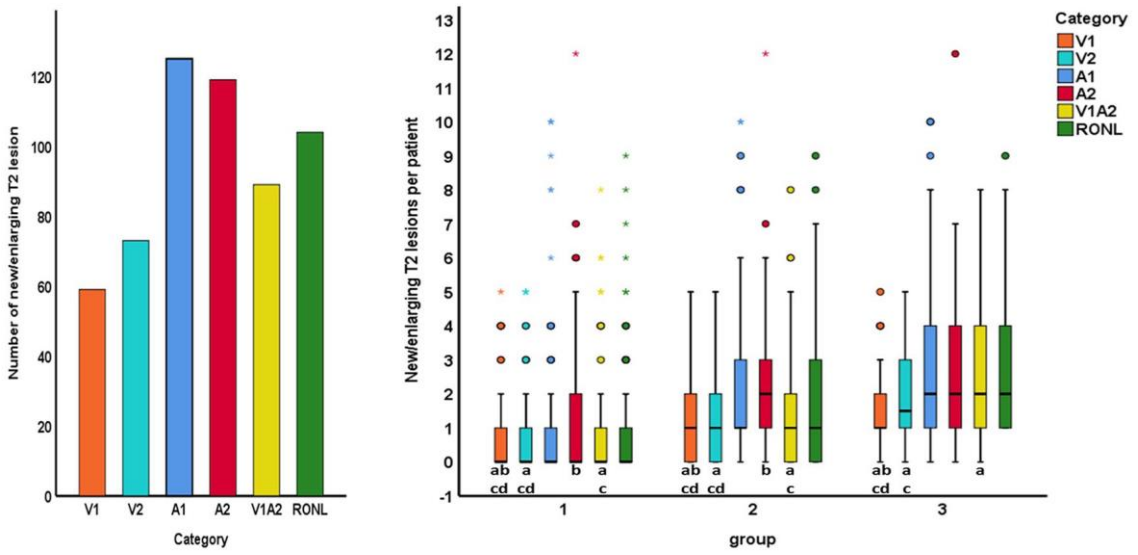

Figure 1. Boxplots comparing the mean number of new/enlarging T2 lesions per patient for the different methods and the reference outcome of new/enlarging lesions (RONL). Left: Number of new/enlarging T2 lesions for the different methods and the reference outcome of new/enlarging lesions (RONL). Right: Comparison of the new/enlarging T2 lesions per patient for the different methods considering the whole cohort (group 1), the 53 patients for whom at least one new/enlarging T2 lesion was detected by one method (group 2) and the 38 patients with truly new/enlarging T2 lesion (group 3).

V1: standard radiological report method; V2: visual review of the MR scans by an expert nonblinded to the radiological report; A1: automated unsupervised approach; A2: automated supervised convolutional neural network-based approach; V1A2: A2 method corrected with the report information; RONL: reference outcome of new/enlarging lesions.

a Significantly different with respect to RONL.

b Significantly different with respect to V1A2.

c Significantly different with respect to A2.

d Significantly different with respect to A1.

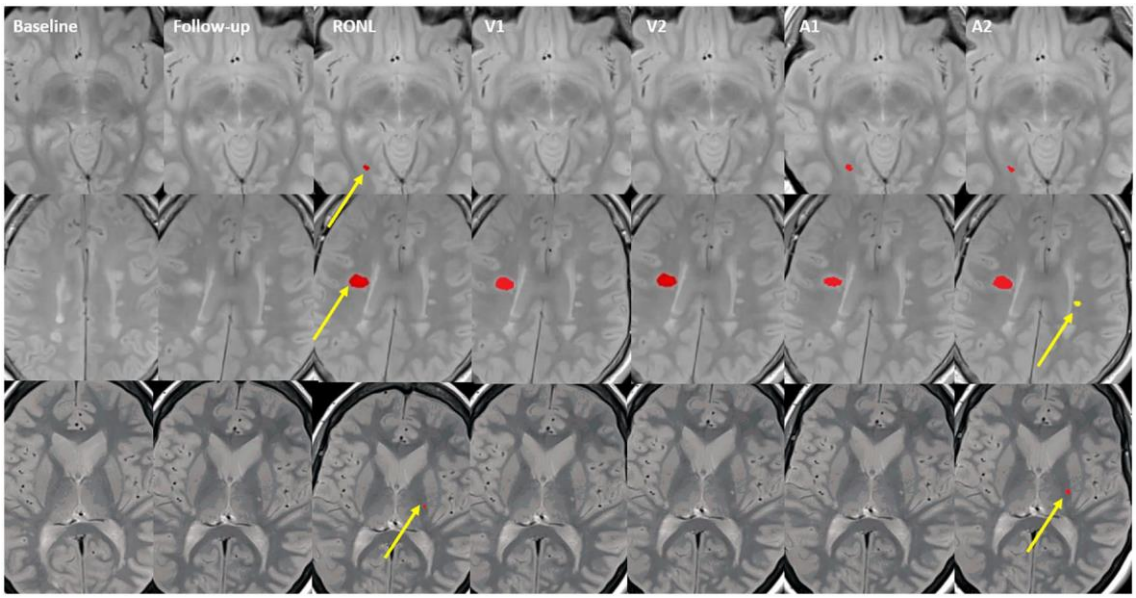

Figure2. Examples of the numbers of new/enlarging T2 lesions detected using the different methods. From left to right, the columns show baseline MRIs, follow-up MRIs, the reference outcome of new/enlarging T2 lesion (RONL) images, the results derived from the standard radiological report method (V1), the results derived from visual review by an expert nonblinded to the radiological report (V2), the results derived from the application of an automated unsupervised approach (A1), and the results derived from the application of an automated supervised convolutional neural network-based approach (A2). The detected lesions are marked with dots. The upper row shows a new lesion in the right occipital white matter missed with the V1 and V2 assessments but identified with the A1 and A2 tools (arrow in the RONL image). The middle row shows one new lesion located in the right frontal white matter that was identified by all methods (arrow in the RONL image) and a false positive lesion detected in the left periventricular white matter with A2 (arrow in the A2 image) but ignored with all other assessments. Finally, the bottom row shows a new lesion located in the posterior limb of the left internal capsule correctly identified with A2 but missed with all other assessments (arrow in the RONL and A2 images).