RECENT ADVANCES IN THE LONGITUDINAL SEGMENTATION OF MULTIPLE SCLEROSIS LESIONS ON MAGNETIC RESONANCE IMAGING: A REVIEW

Marcos Diaz-Hurtado[1], Eloy Martínez-Heras[2], Elisabeth Solana[2], Jordi Casas-Roma[1], Sara Llufriu[2], Baris Kanber[3,4,5], Ferran Prados[1,3,4,5]

1 - e-Health Center, Universitat Oberta de Catalunya, Barcelona, Spain

2 - Center of Neuroimmunology, Laboratory of Advanced Imaging in Neuroimmunological Diseases, Hospital Clinic Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) and Universitat de Barcelona, Barcelona, Spain

3 - Centre for Medical Image Computing, University College London, London, United Kingdom

4 - National Institute for Health Research Biomedical Research Centre at UCL and UCLH, London, United Kingdom

5 - Queen Square MS Centre, Department of Neuroinflammation, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, United Kingdom

Author ORCIDs

Marcos Diaz-Hurtado - 0000-0003-1528-5873
Eloy Martínez-Heras - 0000-0001-9937-3162
Elisabeth Solana - 0000-0001-7973-2439
Jordi Casas-Roma - 0000-0002-0617-3303
Sara Llufriu - 0000-0003-4273-9121
Baris Kanber - 0000-0003-2443-8800
Ferran Prados - 0000-0002-7872-0142

Email of corresponding author: mdiazhu@uoc.edu
Full postal address:
ADaS Lab - e-Health Center
Universitat Oberta de Catalunya
Rambla del Poblenou, 156
08018 Barcelona
(SPAIN)

KEYWORDS: Multiple sclerosis, MRI, longitudinal, lesion segmentation, review

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CODE AVAILABILITY

This publication does not have any code related to its development and no code has been published.

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analysed in this study just reviewing previous literature research.

ETHICS APPROVAL

This study has been approved by the Ethics Committee of the University Oberta de Catalunya (UOC) stating that this research does non include human subjects participation or any processing of personal data and the research fulfils current legislation on data protection.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by M. D.-H., J. C.-R. and F.P. The first draft of the manuscript was written by M.D.-H., E. M.-H. and F.P. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

STATEMENTS AND DECLARATION OF CONFLICTS OF INTEREST

M.D.-H., E.M.-H., J.C.-R., B.K. and F.P. have nothing to disclose. E.S received travel reimbursement from Sanofi. S.L. received compensation for consulting services and speaker honoraria from Biogen Idec, Novartis, TEVA, Genzyme, Sanofi and Merck.

CONSENT TO PARTICIPATE

This research does non include human subjects participation or any processing of personal data.

CONSENT FOR PUBLICATION

All authors consent for publication of this paper.

ABSTRACT

INTRODUCTION: Multiple sclerosis (MS) is a chronic autoimmune disease characterized by demyelinating lesions that are often visible on magnetic resonance imaging (MRI). Segmentation of these lesions can provide imaging biomarkers of disease burden that can help monitor disease progression and the imaging response to treatment. Manual delineation of MRI lesions is tedious and prone to subjective bias, while automated lesion segmentation methods offer objectivity and speed, the latter being particularly important when analysing large datasets. Lesion segmentation can be broadly categorised into two groups: cross-sectional methods, which use imaging data acquired at a single time-point to characterise MRI lesions; and longitudinal methods, which use imaging data from the same subject acquired at two or more different time-points to characterise lesions over time. The main objective of longitudinal segmentation approaches is to more accurately detect the presence of new MS lesions, and the growth or remission of existing lesions, which may be effective biomarkers of disease progression and treatment response.

PURPOSE: This paper reviews articles on longitudinal MS lesion segmentation methods published over the past ten years. These are divided into traditional machine learning methods, and deep learning techniques.

METHODS: PubMed articles using longitudinal information and comparing fully-automatic two time point segmentations in any step of the process were selected.

RESULTS: 19 articles were reviewed.

CONCLUSION: There is an increasing number of deep learning techniques for longitudinal MS lesion segmentation that are promising to help better understand disease progression.

1 INTRODUCTION

Multiple sclerosis (MS) is a chronic autoimmune disease of unknown aetiology, and is characterized by demyelination of the central nervous system (CNS), often visible by magnetic resonance imaging (MRI). MS can be diagnosed after a single clinical episode if lesions are visible on MRI [1–3]. Lesions may appear anywhere in the CNS parenchyma, for example the brain, the spinal cord or the optic nerve. The latest McDonald criteria emphasizes the importance of dissemination of the lesions in both time and space, shortening MS diagnosis time [2]. Dissemination in space (DIS) mean snew T2 lesions, which can be within the brain, optic nerve or the spinal cord. Dissemination in time (DIT) is proven when asymptomatic gadolinium-enhancing lesions appear together with non-enhancing lesions, or new lesions appear in follow-up studies. Lesion load quantification on serial MRI provides a sensitive and objective measure of disease activity, and is a surrogate marker in treatment trials [4]. Clinical and anatomical reality supports the usefulness of using serial MRI studies to delineate and characterise lesions; however, very few approaches exploit the consistency of this longitudinal data to define new [5], enlarging [6] and shrinking lesions [7]. Increasing the number of neuroradiologists in order to cope with increasing numbers of MRIs and the time-consuming task of looking for new, enlarging or shrinking lesions puts hospital radiology departments under considerable pressure. Hence, the importance of translating automated methods for detecting new or enlarging lesions into clinical practice, thus enabling neuroradiologists to simply verify lesions, and amend existing delineations if necessary, rather than having to visually inspect or, in the case of clinical trials, manually delineate all lesions. It is expected that these automated methods will increase the sensitivity of measures used to detect changes (better confidence in findings), reduce inter-rater variability and decrease assessment time.

MRI lesion segmentation enables us to extract imaging biomarkers of disease burden, and can assist in MS diagnosis and monitoring disease progression, as well as imaging treatment response. MRI lesion segmentation consists of a set of techniques focused on differentiating focal MS lesions from normal appearing tissues. Until recently, MRI lesion segmentation was primarily performed manually [8, 9], or with the help of semi-automated tools such as JIM software (Xinapse Systems, Northans, UK; http://www.xinapse.com). However, new advances in machine learning (ML) algorithms, and the improved quality of MRI acquisition have enabled automatic segmentation techniques to achieve remarkable performance across a wide range of medical applications, including more accurate MRI automatic MS lesion segmentation.

Over the past decade, several attempts have been made to improve MS lesion segmentation using fully automatic algorithms. One example is cross-sectional methods, which take a single MRI time-point. The majority of studies in the literature use this method on the brain [5, 10–16], a few use it on the spinal cord [11, 17], and a semi-automatic method is used on the optic nerve region [18]. Cross-sectional methods have been widely used to analyse longitudinal data, including most teams in the ISBI 2015 longitudinal lesion segmentation challenge [12]. However, the main disadvantage of using this method to analyse longitudinal data is the lack of consistency in timing between consecutive scans. Furthermore, longitudinal approaches specifically use the biological temporal consistency from two or more images of the same subject acquired at different time-points to detect changes in size or appearance of new lesions more accurately (see Figure 1 for an example of new and enlarging MS lesions). Recently, Longitudinal MS lesion segmentation methods have become a very active field of research, and several longitudinal segmentation challenges have been organized such as the ISBI 2015 [12], and more recently, the MICCAI21 MS new lesions segmentation challenge (MSSEG-2- https://portal.fli-iam.irisa.fr/msseg-2/). Two of the main challenges of using longitudinal MRI segmentation methods are their effectiveness in the registration procedure in highly pathological brains, and the differing image acquisition protocols in diverse MRI machines. Evaluating longitudinal segmentation methods can be addressed in two ways: (1) according to the number of new or enlarging lesions detected; or (2) by assessing the whole volumetric lesion segmentation. From a clinical perspective, the detection of two or more new MS lesions, separated in time, is used as diagnostic criteria [2]. On the other hand, the precise segmentation and the volume change derived over time is of interest for clinical trials, as outcomes can be used to assess treatment response [19].

This article reviews published research on fully automatic longitudinal MS lesion segmentation on MRI over the past decade. This study expands on the previous longitudinal lesion segmentation review by Llado et al., [20] in 2012. Our review includes only fully automated longitudinal approaches, and these have been classified according to their core architectures into either deep learning or traditional machine learning techniques.

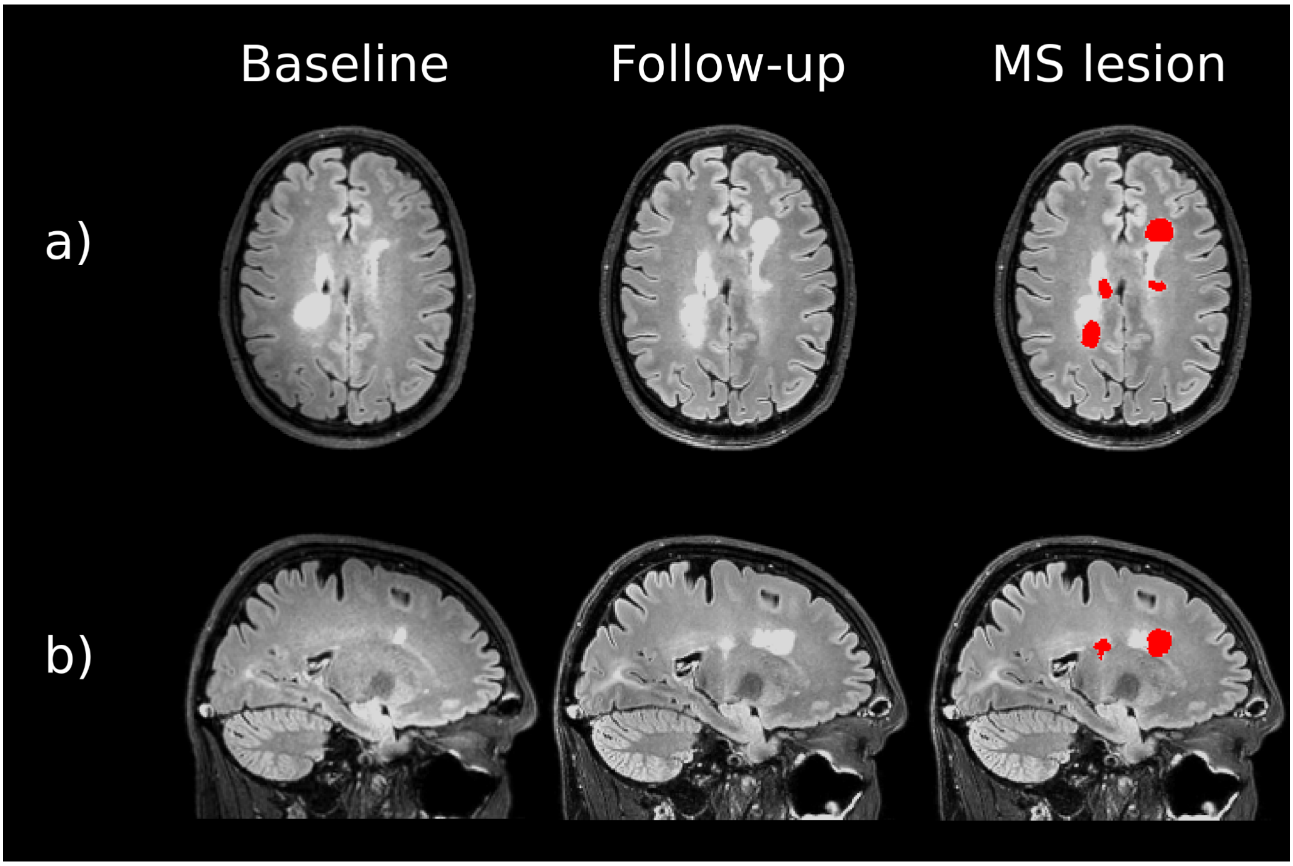

Figure 1 - Example of new and enlarging MS lesions in a MS patient (ISBI 2015 dataset [12]). First row (a) axial view of the baseline and follow-up FLAIR MR images; in the last column, new or enlarging lesions superimposed in red. Second row (b) shows the sagittal view.

2 MATERIAL AND METHODS

Searches on PubMed were made to identify relevant peer-reviewed papers. Search terms combined “image segmentation”, and “multiple sclerosis”, with “longitudinal” or "serial" with and without “MR” or “magnetic resonance” keywords. Bibliographic references cited on the papers found were also reviewed if they included these keywords. All articles using longitudinal data to automatically compare two time-points for new, enlarging or shrinking MS lesion segmentation in any step of the process were selected. The only limit applied was to only selected articles published after Llado et al., 2012 review [20], as this exhaustive review reflected the current state-of-the-art at that time. Papers focusing on other areas were excluded. The flowchart (Figure 2) summarizes the steps taken to refine the search and obtain the papers included in this review. Non-peer-reviewed works (i.e. ArXiV) or conference proceedings were excluded, unless they appeared in the bibliographic references of the PubMed search results.

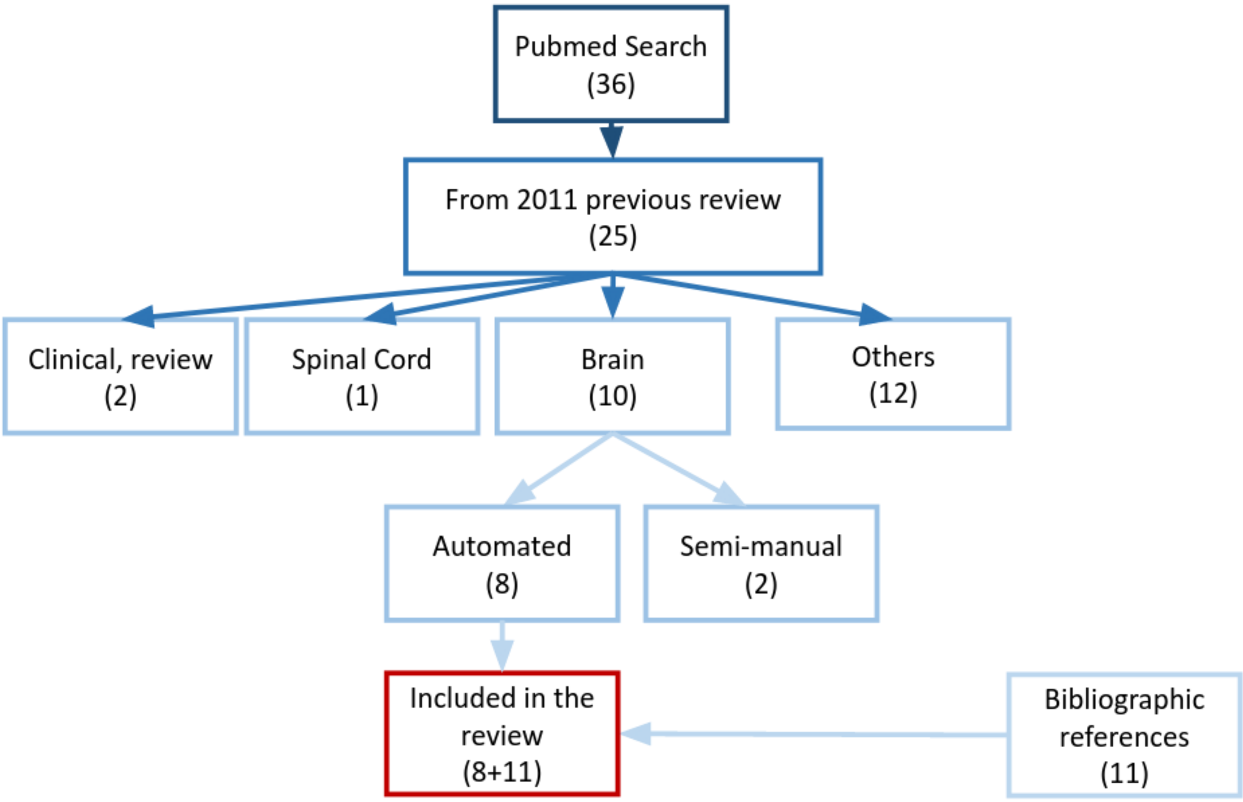

Figure 2 - Schematic representation of the search strategy used to select the papers included in this review

3 RESULTS

Since the review by Llado et al. in 2012 [20], only 19 papers met our research criteria for fully automatic longitudinal MS lesion segmentation. The results section divides findings into four subsections: the common pre-processing steps used in these methods; the validation metrics used to analyse performance; the datasets included in each study; and finally, a chronological summary of the techniques used, categorised into traditional machine learning techniques and deep learning based methods.

3.1 PREPROCESSING STEPS FOR LONGITUDINAL LESION SEGMENTATION

It is not usually possible to make comparisons between MRI studies directly. All studies require pre-processing steps to homogenize the distinct MRI patterns so different image acquisitions can be accurately compared across time. Indeed, the pre-processing steps are a critical procedure that play a key role in identifying new or enlarging lesions. In this section, we number the most common pre-processing steps used in the reviewed articles on improving longitudinal lesion segmentation. Whether these pre-processing steps are used or not may vary, and will depend on the method and dataset selected.

1. Bias field correction

Bias inhomogeneity corrupts MRI images, especially in older MRI scanners, and also presents similar regions with different intensity values [21]. The most commonly used pre-processing techniques for counteracting the effect of the intensity inhomogeneity are N3 or N4 (Nonparametric Nonuniform Intensity Normalization) techniques [22]. The nonparametric nonuniform intensity normalization algorithm corrects the bias field without any pre-segmentation of the image, and is robust to the presence of pathologies [23]. Uncorrected MRI bias field signals could be problematic in longitudinal lesion segmentation procedures based on intensity differences across timepoints.

2. Noise reduction

The primary source of noise that needs to be filtered in MRI images is thermal noise from the patient's body. A widely used approach is the Non-Local Means (NLM) technique, which smooths small intensity variations [23]. Noise reduction often affects the accuracy of brain tissue segmentation methods, and plays an important role in the processing pipeline by distinguishing new or enlarged lesions.

3. Image registration

This is a mandatory step in any longitudinal lesion segmentation method. To compare MRI images across time, the longitudinal acquisitions need to be aligned in the same coordinate system. A rigid linear transformation is employed to register all the MRI modalities acquired from the same subject at different scan times. Specifically, it has been proven that registration needs to be in a halfway position between baseline and follow-up images in order to avoid any potential bias caused by the transformation being applied to a single image [24–26]. Registration can also be made to a standard space such as the MNI (Montreal Neurological Institute) [27] or to a within-subject template [25].

4. Skull stripping

This refers to the process of removing extra-meningeal tissue from the brain MRI image to subtract all non-brain signals from the image. There are four methods: morphology-based; intensity-based; deformable surface with templates; atlas-based; and hybrid. One of the most frequently used methods is the Brain Extraction Tool (FSL-BET), which uses a deformable surface model that evolves until it locates the brain’s boundary [28]. However, this approach can introduce bias and false positives [29; 30]. Recently, newer and more robust techniques such as the HD-BET algorithm [31] have been developed.

5. Longitudinal intensity normalization

Longitudinal intensity normalization by applying linear intensity correction functions are mandatory to compensate for global intensity changes [32]. This step is crucial for homogenising serial MRI image signal intensity and to avoid the appearance of false positives caused by the distinct acquisition parameters, or changes in the scanner hardware settings.

6. Priors

This widely used step applies anatomical prior constraints from predefined atlases in order to define areas where the targeted lesions could not be presented [14] (i.e. white matter (WM) lesions in the cortical grey matter (CGM)). These priors are commonly propagated using registration methods and can include cerebrospinal fluid (CSF), or CGM regions in cases where MS lesion segmentation solely centres on detecting WM lesions.

3.2 VALIDATION METRICS FOR LONGITUDINAL LESION SEGMENTATION METHODS

A number of measures are used to calculate goodness of fit. Some are used at segmentation level (the overall voxel mask is considered, which means lesion segmentation), and others at regional level (one lesion overlap if there is at least one voxel overlapping, which means lesion detection) [16]. These two different approaches to presenting results make it difficult to compare methods as there is no agreement on which method and measure best reflects each algorithm's performance, and each study uses them to their own convenience. Also it is important to note that we can report the metrics over the new lesion only, or over the whole lesion load at followup.

The most frequently used measures evaluate the relationships between true positives (TP), false negatives (FN) and true negatives (TN) at voxel and lesional level. These can be summarized as:

- Sensitivity (also known as recall, or true positive rate, or fraction when it is a percentage), is the proportion of those with a positive test result (correctly identified) out of those having the condition: $\frac{TP}{TP+FN}$, the best value being 100%, as there are no false positives.
- Specificity (also known as true negative rate or fraction, or 1-false positive rate), is the proportion of those with a negative test result (correctly identified) out of those not having the condition: $\frac{TN}{TN+FP}$, the best value being 100%, as tests identify all non-lesional voxels as negative.
- Accuracy, degree of the result of a measurement conforms to the correct value given by the formula $\frac{TP+TN}{TP+TN+FP+FN}$, the best value being 1, as there are no false positives nor false negatives] .

- Positive Predictive Value (PPV), also known as precision, is the proportion of predicted positives actually being positive: $\frac{TP}{TP+FP}$
- Sorensen-Dice Similarity Coefficient (DSC or DICE score) is a statistic of similarity between two samples: $DSC = \frac{(2\,x\,TP\,)}{(2\,x\,TP\,+\,FP\,+\,FN)}$. Its values range between 0 (no overlap) and 1 (perfect agreement).
- Area Under ROC Curve (AUC) and Receiver Operating Characteristic (ROC): ROC is a probability curve obtained by plotting True Positive Rate (TPR) against False Positive Rate (FPR), and its area under the curve represents the measurement separating and distinguishing between two classes, where a higher AUC value means better performance.
- Kappa (K) index: statistic to measure inter-rater reliability. Given pe (expected ratio if raters gave a random prediction) and po (observed prediction of inter-rater agreement). The formula is $K = \frac{(po-pe)}{(1-pe)}$.
- False discovery rate: the ratio of false positive discoveries over the sum of true positives and false positives, then $FDR = \frac{FP}{FP+TP}$.

Table 1 shows a meta comparison of published methods for detecting new or enlarging lesions at longitudinal level over the past decade. As each method has been validated with different input data and different gold standards that include whole lesion segmentation at follow-up, it is not possible to make a direct comparison. Moreover, not all the methods used are publicly available. However, the overall results point in the same direction: at segmentation level mean DSC is 0.59 with 95% CI [0.53-0.64]: and at lesional level, mean DSC increases to 0.69 with 95% CI [0.59-0.80].

In 2015 and 2021, two longitudinal MS segmentation challenges were organised: ISBI 2015 (https://smart-stats-tools.org/lesion-challenge); and MSSEG2 (https://portal.fli-iam.irisa.fr/msseg-2/). Most of the proposed measures were similar, and the organisers of these challenges made a website and/or the code available so the metrics explored could be computed automatically. We therefore encourage researchers working in this field to use these measures in the coming years to better understand the outcomes of each new method.

3.3 DATASETS

The design of test datasets directly affects results. However, in recent years, little effort has been made to homogenise or release a common longitudinal, rich dataset, in comparison to work carried out on cross-sectional datasets [33]. Key aspects of conceptualising a dataset are number of subjects, MS phenotype, length of follow-ups, scan strength, and which MRI modalities to include. Furthermore, during the first half of the last decade, inhouse datasets were widely used as publicly available datasets were scarce. However, the 2015 ISBI challenge, and the release of subsequent datasets and associated labels meant that the methods published began to include both inhouse and ISBI 2015 challenge datasets. In 2016, Lesjak et al. [34] released a dataset with twenty subjects imaged twice, which covered different MS phenotypes and variable length between scans. In 2021, a new public dataset was released following the MSSEG-2 new lesion segmentation challenge.

In the studies reviewed, the ISBI dataset is the most widely used public MR image set. The dataset comprises twenty-one studies from five subjects, and includes T1, T2, proton density (PD) and fluid-attenuated inversion recovery (FLAIR) with the ground-truth segmentations, and is known as the training dataset. There is also a test dataset without segmentations. All the images were skull-stripped using the Brain Extraction Tool (BET) [28], rigidly registered to the 1mm^3 MNI-ICBM152 template [28, 35, 36] using FMRIB's Linear Image Registration tool (FLIRT) [37] and N3 intensity normalization [38]. Two experts manually detected and delineated all the MS lesions in the longitudinal data.

Similarly, Lesjak et al.'s dataset [34] also has conventional MRI images (T1w, T2w, FLAIR and post gadolinium enhancement image), which are co-registered to the FLAIR image and the N4 bias correction was performed. The raw, pre-processed images are available here: https://github.com/muschellij2/open_ms_data.

3.4 METHODS INCLUDED

In 2012, **Lladó** et al. [20] provided the last comprehensive review on using longitudinal methods in MRI segmentation in MS in 2012. This exhaustive review examined 34 longitudinal image segmentation papers published up to that date. The authors proposed classifying techniques into two categories: a first lesion detection, and change detection methods. Lesion detection methods could be further classified as supervised or unsupervised; and change detection methods could be further classified as intensity and deformation, the latter including vector displacement field and deformation field morphometry. The largest group from the reviewed models used subtraction of baseline and follow-up images. Associated techniques, ordered by frequency, were expectation-maximization algorithms, K-nearest neighbours, Bayesian classifiers and artificial neural networks.

Since 2012, nineteen new studies have been published (see Figure 3). The emergence of deep learning methods has had an enormous impact over the last decade, hence the various methods have been classified into either traditional machine learning techniques or deep learning methods.

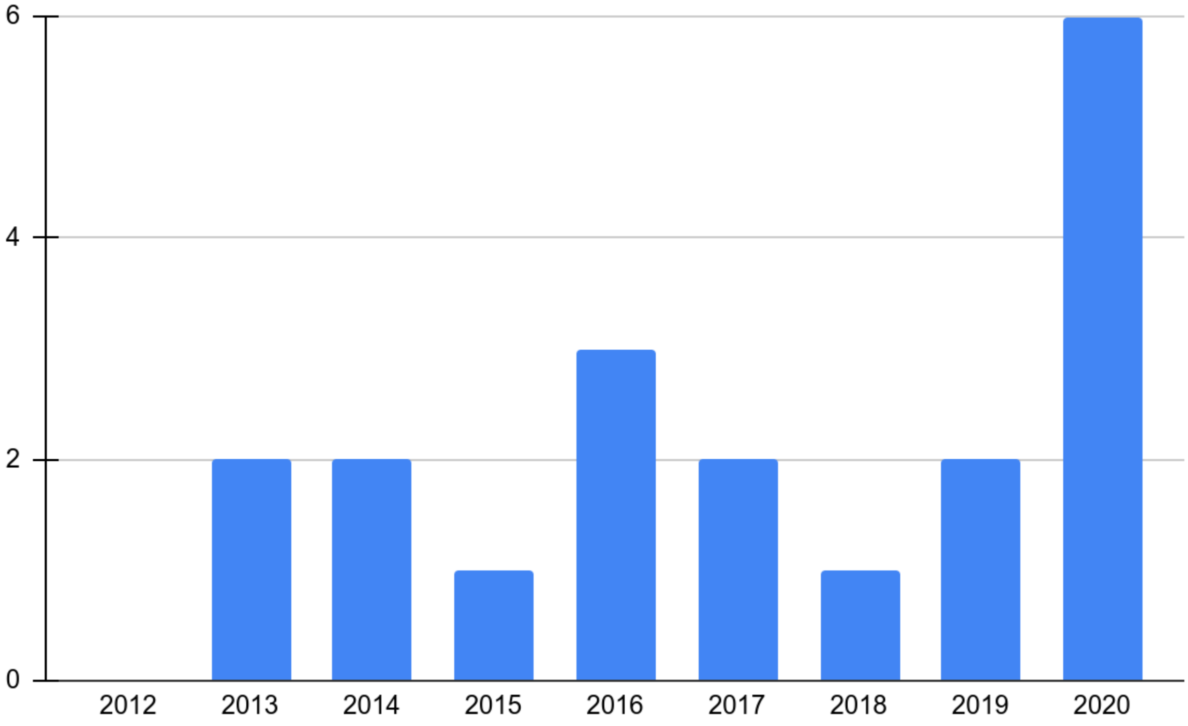

Figure 3 - Bar plot showing the number of papers included in this review (y-axis) for each publication year (x-axis).

3.4.1-Traditional machine learning techniques

This group includes methods based on the latest advances in image analysis techniques.

In 2013, **Sweeny** et al. [5] developed SuBLIME, an unsupervised fully automated and computationally fast method to identify lesion incidence between two voxels from different time-points. The validation was performed using an in-house dataset from 10 patients with longitudinal MRI studies. The gold standard was a manual segmentation by neuroradiologists between consecutive studies. This was an intensity-based approach that used logistic regression coefficients to identify possible changing voxels. After normalizing and registering the images, the T2 hyperintense candidate voxel selection was made. FLAIR, PD, T2 and T1-weighted voxels were subtracted using logistic regression coefficients obtained from the modelled probability of a voxel being part of a lesion. Model performance had 95% sensitivity and 99% specificity with AUC on the ROC curve was 99%. The main weakness of the proposed model was sensitivity to registration errors. In 2013, **Elliot** et al. [39] also proposed a two-stage automated probabilistic framework to identify new MS lesions, providing a multimodal MRI baseline and follow-up images. A generative Bayesian model was used to infer classes at each voxel, and those identified as possible new lesions were assigned a confidence value using a random forest classifier. There were four approaches to developing the Bayesian model, and that gave the probability of each voxel a random variable, which defined class label. Bayes' model was intended to define the probability of a voxel being a lesion and the probability of a new lesion appearing.

The following year, **Battaglini** et al. [40] published a paper using intensity thresholding over the subtracted image from the baseline and follow-up PD MR images. After pre-processing input images of brain extraction, spatial alignment and intensity normalization, the subtraction image was generated. The authors applied 30% intensity threshold over the voxels within the WM mask, then filtered the clusters using shape, extent and intensity constraints. Model segmentation was tested on healthy volunteer datasets and two MS patient datasets. In 2014, **Ganiler** et al. [14] proposed a fully unsupervised automated subtraction pipeline based on a thresholding intensity strategy for detecting new MS lesions. After normalization and pre-processing, images were subtracted and differences selected using an atlas-based WM tissue mask. Two types of thresholding were used: intensity values larger than mean-plus-five standard deviation; and volume lesions occupying three or more voxels.

In 2015, **Roy** et al. [41] proposed a supervised 4D image segmentation from serial MPRAGE, and FLAIR images from a relapsing remitting MS patients' dataset, each with three time points one year apart. Taking a new approach, and instead of using differences between time points, the authors used features from all the time-points simultaneously, building patches that contained information about the temporal trajectories of every lesion. Using norm minimization and other algebraic techniques, the authors developed several atlases of brain characteristics for lesion segmentation.

In 2016, the work of **Lesjak** et al. [34] validated three previously published strategies: confidence level thresholding (CLT) by Ganiler et al. [14]; change vector angular thresholding (CVAHT) by Simoes and Slump [42]; and manual thresholding based on logistic regression by Sweeney et al. [5]. Although all three are intensity subtraction-based studies, they differ in the way they compute dissimilarity maps between baseline and follow-up images. In order to avoid false positives, Ganiler et al. used a threshold of mean-plus-five standard deviations, Simoes and Slump used the generalized likelihood ratio, and Sweeney et al. used logistic regression. Ground truth data, dissimilarity map segmentation and postprocessing varied among models, which made precise comparison between methods extremely difficult. Surprisingly, the authors were unable to reproduce results comparable to the original authors, arguing that this may be due to data overfitting, or a different rater accuracy defining ground truth.

In 2016 **Cabezas** et al. [43] developed a deformation field-based approach to longitudinal segmentation using non-rigid registration of baseline and follow-up images to obtain a deformation field between images. Only candidates with WM lesions were selected, and an intensity threshold used. The deformation vectors around a candidate's lesion were used to detect lesions by employing three metrics: divergence, jacobian and concentricity.

Finally, also in 2016, **Jain** et al. [15] also presented a framework for MSmetrix-long, an iterative WM lesion segmentation based on a joint expectation-maximization (EM) framework, taking 3D T1 and FLAIR of two time points as input. Step one of the segmentation process is a cross-sectional segmentation of the two time-points into WM, GM, CSF and lesions. Step two differentiates between an image subtracting baseline and follow-up FLAIR after bias correction, co-registration and intensity normalization. Step three is joint EM, and uses the output of steps one and two. Step four is a pruning process to eliminate non-lesion candidates.

In 2017, **Carass** et al. [12] published the ISBI 2015 conference longitudinal lesion segmentation challenge outcomes, comprising the eleven approaches presented in the challenge. Since the dataset was made publicly available, it has been the most widely used longitudinal MS dataset for testing longitudinal and cross-sectional lesion segmentation methods. Moreover, the challenge site remains open to submissions to enable method comparison.

In 2018, **Salem** et al. [16] published their first paper on longitudinal segmentation using non-rigid registration between baseline and follow-up. At each voxel of the deformation field, Jacobian, Divergence and NormDiv operators were computed. A logistic regression model was computed at voxel level, and several model configurations explored using four image densities T1, T2, PD and FLAIR, both from baseline and follow-up, in order to decide if a voxel belongs to a lesion or not. **Schmidt** et al. [44], also used this approach to apply a lesion growth algorithm, also published by the same author, to segmentate the baseline and follow-up images individually, first using T1-weighted images to map CSF, GM, WM, and then combining this information with FLAIR to obtain a lesion probability map. After registering both images to the within-subject common space, each voxel was classified one by one as lesion containing or not, both in baseline and follow-up giving six patterns of lesion evolution.

Research by **Cerri** et al. work [13] expands on a cross-sectional generative Bayesian model with prior segmentation for longitudinal tissue and lesion segmentation. This method can be used longitudinally by computing an unbiased within-subject template to estimate the initial model, which is then propagated to the different time points. The intensity probability segmentation approach uses a Gaussian intensity model for each structure.

3.4.2-Deep learning based methods

In 2016, developments in machine learning techniques and their dissemination drove research in this field towards advanced deep learning models.

In 2017 **Birenbaum** et al. [45] published the first study on applying convolutional neural networks (CNN) to longitudinal data for MS lesion segmentation**.** This segmentation method involves three phases: pre-processing, candidate extraction, and CNN prediction. Candidate voxels were selected from hyperintense FLAIR lesions. A probabilistic WM template was used to locate lesions in WM only, or near the union of GM and WM, helping to reduce the CNN computation load. The input was an image voxel, and the output a lesion probability value. Four types of CNN architecture were used: SCSTP (Single Contrast Image Single Time Point), MCSTP (Multiple Contrast Images, Single Time Point), SCMTP (single contrast image, multiple time points), and MCMTP (Multiple Contrast images Multiple Time Points). All models used Single View V-Net of convolution and max pool layers. For longitudinal models (SCMTP and MCMTP), a longitudinal net (L-Net) was made from a V-Net, with input for each time point and output then concatenated. In turn, this output was convoluted and its output fed to a fully connected layer. The multi-view longitudinal network processes the input axial, coronal, and sagittal views separately using a different L-Net for each concatenated view, and finally connecting them to two fully connected layers with final binary output. Several techniques were used to avoid overfitting: weight sharing using identical weights in all V-Nets, dropout layer, and data augmentation. The authors performed cross-validation (four patients for training and one as a test) to evaluate the model. Best Dice score was obtained using MCMTP. Using all contrast images and two consecutive time points improves segmentation accuracy with close to human rater performance. CNNs that make use of longitudinal information can produce better segmentation than standard CNNs.

In 2019, **Fartaria** et al. [46] also presented two models to differentiate WM and GM, and LeMan-PV, a Bayesian partial volume estimation algorithm: LeMan, a supervised KNN classifier using features from images, and an atlas for prior probability maps.

In 2020, the number of publications increased, and convolutional neural networks (CNN) became, as still is, the dominant technique. **Salem** et al. [47] published a model inspired by the VoxelMorph method developed by Balakrishnan et al. [48]. Using voxel-type patches of paired images at consecutive time points from T1, T2, PD and FLAIR, their deformation field was obtained during a registration process. Two fully convolutional neural networks (FCNN) U-Nets are used, and although the authors defined various models, their main model trains all registration and deformation blocks end to end simultaneously. The first segmentation U-Net learns the nonlinear deformation between baseline and follow-up time point. The second U-Net takes as input baseline and follow-up voxels together with its deformation field and outputs the new T2 lesion segmentation mask. In 2020, **Denner** et al. [49] also published a model inspired by the Voxelmorph method, where concatenating multimodal (FLAIR, T1, PD, T2) images are fed to an encoder that carries out a segmentation task using a CNN. The non-rigid registration task learns deformation fields between the already coregistered images at different time points in order to find what has changed (mainly lesions). Segmentation is performed on the three orthogonal slices crossing the voxel of interest, also called 2.5D approach. A fully convolutional and densely connected neural network (known as Tiramisu) was used for each slide image segmentation [50]. The authors of this study developed a multitask learning framework that joined segmentation and non-rigid registration tasks sharing the same encoder, but using separate decoders. A combination of techniques enabled the use of different methods: longitudinal network, multitask longitudinal network, and longitudinal network with pretraining. In another study, **Krüger** et al. [51] developed a two-path 3D CNN in a U-Net- like architecture. This included a fully convolutional network with residual blocks and deep supervision, and Leaky ReLU, instance normalization, Adam optimizer and data augmentation were used to increase performance. In the same year, 2020, **McKinley** et al. [52] introduced DeepSCAN MS, which was a hybrid of U-net [53] and Densenet [54], a CNN based architecture with pooling and upscaling replaced by dilated convolutions for automatic lesion load change detection. Here, a new loss function called label-flip was used, where the probability of containing a certain tissue class and the probability of corresponding to the ground-truth annotation was calculated at each voxel. Given the baseline image, baseline and follow-up segmentations and the label-flip probability, each voxel was classified as a new lesion with segmentation confidence. In 2020 **Gessert** et al. [55] proposed the ResNet-based multiencoder-decoder 3D CNN using convolutional recurrent units (convGRUs) to address longitudinal segmentation of new and enlarged lesions. The authors transformed the single path CNN U-Net [55] into a two-path architecture, where baseline and follow-up volumes are first processed individually, and then jointly, combining convolution down, ResBlock and fusion blocks. The authors then introduced an attention-guided interaction block to control the flow of information between the two paths.

Finally, results from the recent MSSEG-2 challenge are still pending. This extremely well-organised challenge was attended by 20 international teams, and is yet further proof of the interest this topic is currently drawing.

Table 3 summarises the methods published on longitudinal MS lesion segmentation over the past decade.

4 DISCUSSION

A reliable method for assessing the presence of new MS lesions and evidence of no new disease activity is key to evaluating the efficacy of disease-modifying therapies [56]. MRI is a biomarker of MS progression and useful for both diagnosis and monitoring disease activity, as well as treatment response. Radiological MS progression is defined as the appearance of new or enlarging lesions in T2 weighted imaging and new enhancing lesions on T1 weighted imaging with Gadolinium-based contrast [2]. Conventional MRI provides reliable markers of acute inflammatory activity, but has low specificity and sensitivity for those tissue changes that characterize the chronic phase of MS. Sometimes lesions do not show intensity changes, but can affect surrounding tissues. Noise and residual MRI artifacts must also be taken into account, and this can be evaluated by comparing baseline and follow-up scans [43, 49]. Manual delineation is the highest sensitivity technique, and provides higher reliability in the detection of enlarging lesions and new lesions close to areas with large lesion accumulation (such as periventricular regions) when performed by a single rater. However, manual expert segmentation is time-consuming and is subject to inter-observer variability [15]. Furthermore, automatic lesion segmentation has three advantages over manual segmentation in that it offers more consistent segmentations, especially in longitudinal studies; it displays more reproducible results between datasets; and it improves processing speed. Numerous cross-sectional automatic image segmentation methods have been published, but few focus on longitudinal approaches, which appear to obtain better results [12]. The main advantage of lesion segmentation methods is the precise quantification of the volume of the brain lesion, which is extremely useful for lesion filling, and therefore improving brain atrophy quantification [57, 58].

Over the past decade, comparing methods has been an impossible task because this requires using the same dataset, but the data used in each paper is highly heterogeneous. Two issues hinder making comparisons: firstly, using inhouse vs publicly available datasets; and secondly, the different MRI modalities needed for each method. In addition to this, further issues have appeared, some of which are related to the disease, or others to the ground truth design. The main issue associated with the disease is that the methods are focused on detecting new and enlarging lesions. However, disease activity varies across phenotypes; for example, the number of new or enlarging lesions will differ in CIS, SPMS or PPMS subjects, hence results may vary. It is therefore important to choose a large dataset with subjects from all the phenotypes and different levels of disease activity. Another issue is labelling, or creating ground truth. Some methods have used a single rater as ground truth; however, labelling new and enlarging lesions has proven to be extremely difficult, with considerable variability between experts [12]. Using a small number of experts has failed to reach a good agreement; hence, including a large number of experts and understanding inter-method variability (e.g. automatic methods vs manual labelling) will be helpful in showing a better performance for any method.

Using a common dataset across the papers would help estimate the proposed methods more accurately. Therefore, in coming years it will be useful for all new methods to use the publicly available datasets and their labels, such as those released by the ISBI 2015 and MSSEG-2 organisers. The inclusion of these datasets, along with other inhouse datasets, will benefit the entire MS research community. However, ISBI 2015 and MSSEG-2 are small datasets and have been labelled by only a small number of experts. Therefore, it would be advantageous in the near future to organise a national or international consortium to gather and release a bigger, multi-centre, labelled (ideally by a large number of experts) dataset that could boost the development of this key research field.

This review has incorporated papers published on longitudinal MS lesion segmentation between 2012 to January 2022. The techniques that stand out, from older to newer, are as follows: image subtraction and thresholding, pure mathematical image differences, Bayesian generative models and deep neural network-based methods.

Chronologically, the main features of the first articles published used baseline and follow-up image segmentation subtraction, with thresholds on image intensity values being this a trade-off between sensitivity and specificity [5]. Bayesian-inspired models are a very powerful method as their effectiveness rests on the anatomical probability distribution of finding a lesion in a tissue, which is supported by the pathophysiology of the disease. Mathematical techniques based on the matrix comparison of the intensity values of the images have given way to more powerful techniques such as deformation fields, which have also been used with CNN successfully [47]. Deep learning U-Nets and other CNN-based architectures comprise the majority of more recent articles on cross-sectional and longitudinal approaches to MRI image segmentation, and can be classified into patch-wise, semantic and cascaded CNNs[59]. However, CNNs still have challenges that limit their potential. One is data class imbalance when the number of voxels with lesions is much smaller than those without, which is a common MS segmentation problem that causes overfitting. A special loss function based on the Tversky index is used to mitigate the issue of MS data imbalance[60], and transfer learning is used for small datasets.

As stated above, direct comparison between the techniques reviewed is difficult because they use different datasets. We only found one article comparing different longitudinal models using a common dataset [34], and that most datasets are private with implementation codes that are not usually published. Nevertheless, some authors offer resources in public code repositories such as Github [43, 49], and some procedures are available in software suites such as FreeSurfer (http://surfer.nmr.mgh.harvard.edu/)) [13]. Table 2 summarises the code and data availability for each method. Several initiatives have been launched to keep track of the accuracy of different methods and enable direct comparison between them. For example, the organizers of the ISBI 2015 challenge have made their dataset publicly available and its website continues to accept submissions. However, detailed descriptions of the models submitted to the website are not always provided.

Our literature search results show that there is an increasing interest in finding automated methods to enable the detection of new or enlarging lesions, and ways to compare the effectiveness of current longitudinal models (Figure 3). Most studies in this review used FLAIR acquisitions (clinical settings being the most common) and when additional MRI modalities are used (e.g. T1, T2, PD) the models become more accurate. This requires more computational resources and is more difficult to translate to clinical practice, as some image modalities may not be routinely available at all healthcare centres [61]. The methods reviewed could also be classified under the categories "supervised" and "unsupervised". The main drawback of unsupervised approaches is that they assume perfect registration and intensity normalization [15], while supervised approaches often require large training datasets. As demonstrated, longitudinal image segmentation is a very active research field that has shifted from image analysis to machine learning and AI techniques, which are mainly CNN based architectures. Future challenges cover both domain shifting (i.e. enabling a change in input MRI modality) and domain adaptation (i.e. performing to the same level, regardless of the MR scanner used) and the introduction of attention mechanisms such as transformer architectures, as proposed by Gessert for MS brain MR segmentation [55].

Recently, the MSSEG-2 challenge has only focused on new lesions, but for an effective translation to clinical practice, and to draw up precise patient prognosis, the methods employed will not only need to detect new MS lesions, but also lesion changes between time-points, such as lesion growth or shrinkage. New and enlarging lesions have been widely explored in recent years; however, research on shrinking lesions is still lacking, and is an area that needs to be carefully examined in the future. Accomplishing automated segmentation of new and changing MS lesions would be beneficial for disease and care management programs. Hence, implementing and adopting automated aid systems to assist radiologists and neurologists is key to leveraging reading time and ensuring similar diagnoses by different radiologists. Adopting quantitative image-based biomarkers in the clinic addresses these issues [62], and is a process with three main steps. Firstly, a credibility study is carried out, which involves a technical validation and a limited clinical inspection by an expert clinician who confirms usability and findings. Secondly, the tool needs to undergo an accuracy study to confirm its beneficial impact on clinical routine. A large multi-centre dataset with clinical MRI quality is used to do this. Thirdly, the methods developed need to meet safety, health and environmental protection requirements in order to get approval from regulatory bodies prior to final integration into care centre workflows.

5 CONCLUSIONS

In MS, longitudinal MR image segmentation is key to assessing disease progression and response to treatment. This review describes approaches published on longitudinal MRI lesion segmentation in MS patients over the past ten years. Over this period, we have seen an increasing interest in automatically detecting new and enlarging or shrinking MS lesions, and a transition from image-processing based techniques towards mainly machine learning-based methods. To boost this research field, gathering and making available a larger, longitudinal, multicentre MS dataset with the associated labelling should be mandatory. Finally, these new methods need to be validated through clinical trials to increase automatisation and then transferred to a clinical setting by integrating them into automatic quantitative reports. In coming years, this important endeavour will have a significant positive impact on patient care and the overall healthcare system.

CONFLICTS OF INTEREST

M.D.-H., E.M.-H., J.C.-R., B.K. and F.P. have nothing to disclose. E.S received travel reimbursement from Sanofi. S.L. received compensation for consulting services and speaker honoraria from Biogen Idec, Novartis, TEVA, Genzyme, Sanofi and Merck.

REFERENCES

1. Igra MS, Paling D, Wattjes MP, et al (2017) Multiple sclerosis update: use of MRI for early diagnosis, disease monitoring and assessment of treatment related complications. Br J Radiol 90:20160721

2. Thompson AJ, Banwell BL, Barkhof F, et al (2018) Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol 17:162–173

3. Rovira À, on behalf of the MAGNIMS study group, Wattjes MP, et al (2015) MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process. Nature Reviews Neurology 11:471–482

4. Molyneux PD (1998) Precision and reliability for measurement of change in MRI lesion volume in multiple sclerosis: A comparison of two computer assisted techniques. J Neurol Neurosurg Psychiatry 65:42–47

5. Sweeney EM, Shinohara RT, Shea CD, et al (2013) Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. AJNR Am J Neuroradiol 34:68–73

6. Calvi A, Haider L, Prados F, et al (2020) In vivo imaging of chronic active lesions in multiple sclerosis. Mult Scler 1352458520958589

7. Sethi V, Nair G, Absinta M, et al (2017) Slowly eroding lesions in multiple sclerosis. Mult Scler 23:464–472

8. Miller DH (1994) Magnetic resonance in monitoring the treatment of multiple sclerosis. Ann Neurol 36 Suppl:S91–4

9. Weeda MM, Brouwer I, de Vos ML, et al (2019) Comparing lesion segmentation methods in multiple sclerosis: Input from one manually delineated subject is sufficient for accurate lesion segmentation. NeuroImage: Clinical 24.: https://doi.org/10.1016/j.nicl.2019.102074

10. Valverde S, Cabezas M, Roura E, et al (2017) Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. Neuroimage 155:159–168

11. Gros C, de Leener B, Badji A, et al (2019) Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. Neuroimage 184:901–915

12. Carass A, Roy S, Jog A, et al (2017) Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. Neuroimage 148:77–102

13. Cerri S, Puonti O, Meier DS, et al (2021) A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. Neuroimage 225:117471

14. Ganiler O, Oliver A, Diez Y, et al (2014) A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. Neuroradiology 56:363–374

15. Jain S, Ribbens A, Sima DM, et al (2016) Two time point MS lesion segmentation in brain MRI: An expectation-maximization framework. Front Neurosci 10:1–11

16. Salem M, Cabezas M, Valverde S, et al (2018) A supervised framework with intensity subtraction and deformation field

features for the detection of new T2-w lesions in multiple sclerosis. NeuroImage: Clinical 17:607–615

17. Horsfield MA, Sala S, Neema M, et al (2010) Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: application in multiple sclerosis. Neuroimage 50:446–455

18. Hickman SJ (2007) Optic nerve imaging in multiple sclerosis. J Neuroimaging 17 Suppl 1:42S–45S

19. Doyle A, Elliott C, Karimaghaloo Z, et al (2018) Lesion Detection, Segmentation and Prediction in Multiple Sclerosis Clinical Trials. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries 15–28

20. Lladó X, Ganiler O, Oliver A, et al (2012) Automated detection of multiple sclerosis lesions in serial brain MRI. Neuroradiology 54:787–807

21. Juntu J, Sijbers J, Dyck D, Gielen J (2008) Bias field correction for MRI images. In: Advances in Soft Computing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 543–551

22. Tustison N, Gee J (2010) N4ITK: Nick's N3 ITK implementation for MRI bias field correction. The Insight Journal. https://doi.org/10.54294/jculxw

23. Freire PGL, Ferrari RJ (2016) Automatic iterative segmentation of multiple sclerosis lesions using Student's t mixture models and probabilistic anatomical atlases in FLAIR images. Comput Biol Med 73:10–23

24. Leung KK, Ridgway GR, Ourselin S, et al (2012) Consistent multi-time-point brain atrophy estimation from the boundary shift integral. Neuroimage 59:3995–4005

25. Reuter M, Schmansky NJ, Rosas HD, Fischl B (2012) Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61:1402–1418

26. Alexa M (2002) Linear combination of transformations. ACM Trans Graph 21:380–387

27. Brett M, Johnsrude IS, Owen AM (2002) The problem of functional localization in the human brain. Nat Rev Neurosci 3:243–249

28. Smith SM (2002) Fast robust automated brain extraction. Hum Brain Mapp 17:143–155

29. Eskildsen SF, Coupé P, Fonov V, et al (2012) BEaST: brain extraction based on nonlocal segmentation technique. Neuroimage 59:2362–2373

30. Iglesias JE, Liu C-Y, Thompson PM, Tu Z (2011) Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans Med Imaging 30:1617–1634

31. Isensee F, Schell M, Pflueger I, et al (2019) Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp 40:4952–4964

32. Bosc M, Heitz F, Armspach JP, et al (2003) Automatic change detection in multimodal serial MRI: Application to multiple sclerosis lesion evolution. Neuroimage 20:643–656

33. Commowick O, Istace A, Kain M, et al (2018) Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Sci Rep 8:13650

34. Lesjak Ž, Pernuš F, Likar B, Špiclin Ž (2016) Validation of White-Matter Lesion Change Detection Methods on a Novel Publicly Available MRI Image Database. Neuroinformatics 14:403–420

35. Mori S, Oishi K, Jiang H, et al (2008) Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. NeuroImage 40:570–582

36. Oishi K, Zilles K, Amunts K, et al (2008) Human brain white matter atlas: identification and assignment of common anatomical structures in superficial white matter. Neuroimage 43:447–457

37. Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Medical Image Analysis 5:143–156

38. Sled JG, Bruce Pike G (1998) Understanding intensity non-uniformity in MRI. Medical Image Computing and Computer-Assisted Intervention — MICCAI'98 614–622

39. Elliott C, Arnold DL, Collins DL, Arbel T (2013) Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. IEEE Trans Med Imaging 32:1490–1503

40. Battaglini M, Rossi F, Grove RA, et al (2014) Automated identification of brain new lesions in multiple sclerosis using subtraction images. J Magn Reson Imaging 39:1543–1549

41. Roy S, Carass A, Prince JL, Pham DL (2015) Longitudinal Patch-Based Segmentation of Multiple Sclerosis White Matter Lesions. Machine Learning for Medical Imaging 9352:194–202

42. Simões R, Slump C (2011) Change detection and classification in brain MR images using change vector analysis. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 7803–7807

43. Cabezas M, Corral JF, Oliver A, et al (2016) Improved automatic detection of new t2 lesions in multiple sclerosis using deformation fields. AJNR American journal of neuroradiology 37:1816–1823

44. Schmidt P, Pongratz V, Küster P, et al (2019) Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. NeuroImage: Clinical 23:101849

45. Birenbaum A, Greenspan H (2017) Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. Eng Appl Artif Intell 65:111–118

46. Fartaria MJ, Kober T, Granziera C, Bach Cuadra M (2019) Longitudinal analysis of white matter and cortical lesions in multiple sclerosis. NeuroImage: Clinical 23:101938

47. Salem M, Valverde S, Cabezas M, et al (2020) A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. NeuroImage: Clinical 25.: https://doi.org/10.1016/j.nicl.2019.102149

48. Balakrishnan G, Zhao A, Sabuncu MR, et al (2019) VoxelMorph: A Learning Framework for Deformable Medical Image

Registration. IEEE Trans Med Imaging 38:1788–1800

49. Denner S, Khakzar A, Sajid M, et al (2020) Spatio-temporal Learning from Longitudinal Data for Multiple Sclerosis Lesion Segmentation. In: BrainLes Workshop in MICCAI2020

50. Zhang H, Valcarcel AM, Bakshi R, et al (2019) Multiple Sclerosis Lesion Segmentation with Tiramisu and 2.5D Stacked Slices. Med Image Comput Comput Assist Interv 11766:338–346

51. Krüger J, Opfer R, Gessert N, et al (2020) Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. NeuroImage: Clinical 28:102445

52. McKinley R, Wepfer R, Grunder L, et al (2020) Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence. NeuroImage: Clinical 25:102104

53. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp 234–241

54. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2016) Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

55. Gessert N, Krüger J, Opfer R, et al (2020) Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs. Comput Med Imaging Graph 84:101772

56. Giovannoni G, Turner B, Gnanapavan S, et al (2015) Is it time to target no evident disease activity (NEDA) in multiple sclerosis? Mult Scler Relat Disord 4:329–333

57. Prados F, Cardoso MJ, Kanber B, et al (2016) A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. Neuroimage 139:376–384

58. Battaglini M, Jenkinson M, De Stefano N (2012) Evaluating and reducing the impact of white matter lesions on brain volume measurements. Human brain mapping 33:2062–2071

59. Akkus Z, Galimzianova A, Hoogi A, et al (2017) Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. Journal of Digital Imaging 30:449–459

60. Hashemi SR (2017) Asymmetric Loss Functions and Deep Densely Connected Networks for Highly Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection. Physiol Behav 176:139–148

61. Goodkin O, Prados F, Vos SB, et al (2021) FLAIR-only joint volumetric analysis of brain lesions and atrophy in clinically isolated syndrome (CIS) suggestive of multiple sclerosis. Neuroimage Clin 29:102542

62. Pemberton HG, Goodkin O, Prados F, et al (2021) Automated quantitative MRI volumetry reports support diagnostic interpretation in dementia: a multi-rater, clinical accuracy study. Eur Radiol 31:5312–5323

Table 1 - Summary of the published results for each multiple sclerosis longitudinal lesion segmentation method included in this review at segmentation (voxel by voxel) and lesional (whole mask) level. Only the most common measures and the best results have been included. Results were obtained using different datasets.

	YEAR	**REFERENCE**	**Segmentation level**			**Lesional level**			**Others**
			DSC	**Sensitivity/TPR**	**Specificity/TNR**	**DSC**	**Sensitivity/TPR**	**Specificity/ TNR**	
1	2013	[5]	--	0.95	0.99	--	--	-	AUC=0.99
2	2013	[39]	--	--		0.84	0.80	-	FDR=0.08
3	2014	[40]	--	--	--	--	0.91	--	Kappa=0.82 [95%CI: 0.77-0.87] Spearsman R=0.92
4a/12M	2014	[14]	0.51	0.91	--	0.64	--	--	FDR=0.5
4b/48M	2014	[14]	0.58	0.80	--	0.63	--	--	FDR=0.48
5	2015	[41]	0.50	0.46	0.42	--	--	--	VD=0.03
6	2016	[34]	0.58	--	0.91	--	--	--	--
7	2016	[43]	0.52	0.91	--	0.68	--	--	Average surface distance = 7.89
8	2016	[15]	0.63	0.57	0.75	--	--	--	Pearson correlation coefficient=0.96 Absolute volume difference=1.48 ml
9	2017	[12]	0.64	-	-	-	-	-	-
10	2017	[45]	0.73	--	--	--	--	--	-
11	2018	[16]	0.56	0.74	0.89	0.77	--	--	-
12	2019	[44]	0.72	0.74	--	--	--	--	FDR=0.08
13	2019	[46]	--	0.87	--	--	--	--	-
14	2020	[47]	0.55	0.83	0.91	0.83	-	--	-

15	2020	[52]	--	--	--	0.45	0.60	0.59	-
16	2020	[49]	--	--	--	0.70	0.68	0.80	ISBI Score=92.12 PPV=0.77 Volume difference=0.22
17	2020	[51]	0.45	0.60	0.59	--	--	--	-
18	2020	[13]	0.57	0.42	-	-	-	-	-
19	2020	[55]	0.64	--	--	-	0.77	0.75	-
SUMMARY	**Mean**		***0.58***	***0.73***	***0.78***	***0.69***	***0.75***	***0.71***	
	95% CI		***[0.54-0.63]***	***[0.620.85]***	***[0.59-0.97]***	***[0.59-0.80]***	***[0.61-0.90]***	***[0.44-0.99]***	

Table 2 - Brief dataset description, data and code availability for each longitudinal multiple sclerosis lesion segmentation method included in this review.

	YEAR	REFERENCE	DATASET	DATASET AVAILABILITY	CODE AVAILABILITY
1	2013	[5]	10 MS patients, 11 scans each separated a mean of about 2 months between them	Not available	Only web interface: https://smart-stats-tools.org/sublime-interface
2	2013	[39]	a) Dataset A: 95 RRMS patients, 2-7 scans separated 3-12 months apart b) Dataset B: 160 MS patients, 2 scans separated 2-7 months	Not available	Not available
3	2014	[40]	a) 10 healthy volunteers, 2 scans separated 24 months b) 19 MS patients, 2 scans separated 9 months c) 103 MS patients, 2 scans separated 18 months	Not available	Not available
4	2014	[14]	a) 10 MS patients, 2 scans separated 12 months b) 10 MS patients, 2 scans separated 48 months	Not available	Not available
5	2015	[41]	10 MS patients, 3 scans separated 12 months	Not available	Not available
6	2016	[34]	20 MS patients, 2 scans separated by different follow-up lengths	https://johnmuschelli.com/open_ms_data/	Not available
7	2016	[43]	36 CIS patients, 2 scans separated 12 months	Not available	https://github.com/NIC-VICOROB/braintools
8	2016	[15]	a) 12 RRMS patients, 2 scans separated 12 months b) 10 MS patients, 2 scans with repositioning (5-10 minutes) using 3 different scanners	Not available	Not available
9	2017	[12]	ISBI Challenge dataset	https://smart-stats-tools.org/lesion-challenge	Not available
10	2017	[45]	ISBI Challenge dataset	https://smart-stats-tools.org/lesion-challenge	Not available

11	2018	[16]	60 MS patients, 2 scans separated 12 months	Not available	https://github.com/NIC-VICOROB/LR-T2-w-Lesions
12	2019	[44]	55 MS patients from different hospitals in the National cohort study of the German Competence Network Multiple Sclerosis	Not available	Lesion Segmentation Tool (LST): https://www.applied-statistics.de/lst.html
13	2019	[46]	32 RRMS patients, 2 scans separated 12 months	Not available	Not available
14	2020	[47]	60 MS patients, 2 scans separated 12 months	Not available	Not available
15	2020	[52]	a) 26 MS patients, 4-5 scans separated a mean of about 4 months (Bernese MS bank dataset) b) 8 MS patients, 4 scans separated without specifying length of the follow-ups (Zurich dataset) c) 53 MS patients, 2 scans separated without specifying length of the follow-ups (Munich dataset)	Not available	Not available
16	2020	[49]	a) 70 MS patients, 2 scans separated without specifying length of the follow-ups b) ISBI Challenge dataset	Not available	https://github.com/StefanDenn3r/Spatio-temporal-MS-Lesion-Segmentation
17	2020	[51]	a) 1574 MS patients, 2 scans separated a mean of about 12 months (Rou2 and PhIng datasets) b) 89 MS patients, 2 scans separated a mean of about 27 months (Zurich dataset) c) 32 MS patients, 4 scans separated a mean of about 12 months (Dresden dataset)	Not available	Not available
18	2020	[13]	a) 2 MS patients, 6 scans in 3 differents scanners within 3 weeks b) 86 RRMS patients, 3-6 scans separated 6-12 months c) 135 non-MS patients 2-6 scans separated 6-12 months	‘a’ and ‘b’ are not available ‘c’ can be found at: http://adni.loni.usc.edu	Available through Freesurfer software package

19	2020	[55]	a) 89 MS patients, 2 scans separated a mean of about 27 months (Zurich dataset) b) 33 MS patients, 3 scans separated 12 months	Not available	Not available

Table 3 - Longitudinal multiple sclerosis lesion segmentation methods summary in chronological order. Publication year, reference, descriptive summary of the method, combination of used MRI modalities and segmented lesion type for each method are provided.

	YEAR	REFERENCE	APPROACH	MRI MODALITIES	LESION TYPE
0	2012	[20]	Latest review on longitudinal multiple sclerosis lesion segmentation techniques, it divides them into intensity-based and deformation-based methods.	Depending on paper: T1, T2, PD, FLAIR	Mainly WM, also GM and Gadolinium enhancing lesions
1	2013	[5]	SuBLIME (subtraction-based logistic inference for modeling and estimation): lesion incidence in a voxel as a probability map calculated by a logistic regression model	T1, T2, PD, FLAIR	WM
2	2013	[39]	Bayesian classifier and Random Forest Classifier	T1, T2, PD, FLAIR, some T1 with Gadolinium	WM and GM
3	2014	[40]	Subtraction and threshold using PD images	PD	WM
4	2014	[14]	Fully automated subtraction pipeline based on threshold strategy	T1, T2, PD	WM
5	2015	[41]	Supervised 4D image segmentation using features from all the time points simultaneously.	T1, FLAIR	WM
6	2016	[34]	Validation of approaches proposed by Ganiler et al., Simoes and Slump and Seeney et al.	T1, T2, FLAIR	WM
7	2016	[43]	Mask obtained by image subtraction and thresholding. Then, around each lesion a deformation field is calculated. The lesion mask is refined applying divergence, jacobian and concentricity metrics.	T2, PD, FLAIR	WM
8	2016	[15]	Used MSmetrix-long an expectation maximization (EM) framework. First cross-sectional, after subtraction, next joint EM, next pruning non-lesions candidates.	T1, FLAIR	WM and GM
9	2017	[12]	ISBI Longitudinal MS lesion segmentation challenge. Several methods that segment longitudinal lesions in a cross-sectional fashion.	T1, T2, PD, FLAIR	WM

10	2017	[45]	First CNN applied to longitudinal data, different architecture, Longitudinal net (L-Net) made from V-Net	FLAIR	WM
11	2018	[16]	At every jacobian voxel, the Divergence and NormDiv operators were computed. Baseline and follow-up voxels were subtracted, and the deformation field calculated. A logistic regression model at voxel level was computed using these features.	T1, T2, PD, FLAIR	WM
12	2019	[44]	Baseline and follow-up images were individually segmented using a lesion growth algorithm and after a probabilistic map was obtained. After registration of both images is done, six patterns of lesion evolution at every voxel are determined.	T1, FLAIR	WM and GM
13	2019	[46]	LeMAN (knn plus atlas approach to differentiate WM from GM) LeMAn-PV (Bayesian partial volume estimation with spatial GM constraints)	T1, FLAIR, DIR	WM and GM
14	2020	[47]	U-Net to learn deformation fields and registering. The U-Net has as input a deformation field, baseline and follow-up images and it outputs final segmentation	T1, T2, PD, FLAIR	WM
15	2020	[52]	2.5D approach (3 orthogonal slices over voxel) CNN adapted to 2D inputs using Tiramisu architecture (fully convolutional and densely connected neural network) a) Early multimodal fusion as input to network b) Adding structural changes to a first model with deformable registration	T1, T2, PD, FLAIR	WM
16	2020	[49]	The main model is a two-path 3D CNN in a U-Net- like architecture.	FLAIR	WM
17	2020	[51]	Fully convolutional encoder-decoder U-Net using residual blocks and deep supervision, and leaky ReLU as activation function. Instance normalization, Adam optimizer and data augmentation were used.	FLAIR	WM
18	2020	[13]	Bayesian generative model	T1, FLAIR	WM
19	2020	[55]	Multiple sclerosis lesion activity segmentation with attention-guided by two-path CNNs	FLAIR	WM