Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging

Daniel García-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L. Arnold, D. Louis Collins

McConnell Brain Imaging Center, Montreal Neurological Institute, McGill University, Montreal, Canada

Corresponding Author: Daniel García-Lorenzo, dgarcia@bic.mni.mcgill.ca

Abstract

Magnetic resonance (MR) imaging is often used to characterize and quantify multiple sclerosis (MS) lesions in the brain and spinal cord. The number and volume of lesions have been used to evaluate MS disease burden, to track the progression of the disease and to evaluate the effect of new pharmaceuticals in clinical trials. Accurate identification of MS lesions in MR images is extremely difficult due to variability in lesion location, size and shape in addition to anatomical variability between subjects. Since manual segmentation requires expert knowledge, is time consuming and is subject to intra- and inter-expert variability, many methods have been proposed to automatically segment lesions.

The objective of this study was to carry out a systematic review of the literature to evaluate the state of the art in automated multiple sclerosis lesion segmentation. From 1240 hits found initially with PubMed and Google scholar, our selection criteria identified 80 papers that described an automatic lesion segmentation procedure applied to MS. Only 47 of these included quantitative validation with at least one realistic image. In this paper, we describe the complexity of lesion segmentation, classify the automatic MS lesion segmentation methods found, and review the validation methods applied in each of the papers reviewed. Although many segmentation solutions have been proposed, including some with promising results using MRI data obtained on small groups of patients, no single method is widely employed due to performance issues related to the high variability of MS lesion appearance and differences in image acquisition. The challenge remains to provide segmentation techniques that work in all cases regardless of the type of MS, duration of the disease, or MRI protocol, and this within a comprehensive, standardized validation framework. MS lesion segmentation remains an open problem.

1. Introduction

Multiple sclerosis (MS) is a chronic disease that affects the central nervous system with great variability in clinical manifestation. The disease usually begins with a relapsing remitting course in which patients have "attacks" characterized by transient symptoms of focal central nervous system (CNS) dysfunction, such as numbness or weakness of a limb, incoordination, vertigo or visual dysfunction (McAlpine, 1973). These clinical attacks, or relapses, are due to focal inflammation in the CNS directed against myelin, the insulation around nerve fibers (axons). The inflammation results in focal "MS lesions", which are characterized by demyelination, axonal injury and axonal conduction block, and are the hallmark of MS. These classically-described white matter (WM) lesions are visible in conventional magnetic resonance imaging (cMRI), appearing hyperintense in T2-weighted (T2w) images and hypointense in T1-weighted (T1w) images. A useful tool for noninvasive monitoring of the disease, cMRI is used widely in clinical practice, research, and clinical trials (Fazekas et al., 1999; Simon et al., 2006).

The presence and spatial pattern of focal WM lesions on MRI (dissemination in space) and the appearance of new WM lesions (dissemination in time) are key components of current diagnostic criteria for MS (Polman et al., 2011). In clinical trials, the impact of treatment on the accrual of WM lesions is an important measure of efficacy. While more specialized MRI techniques such as magnetization transfer imaging, diffusion tensor imaging, and magnetic resonance spectroscopy (Bakshi et al., 2008; Zivadinov et al., 2008) demonstrate diffuse MS pathology outside of focal WM lesions, these techniques are complicated to implement and interpret, and are not yet widely used clinically. Also, it is not clear to what degree this diffuse damage is secondary to damage within focal lesions (in WM or grey matter (GM)), and to what extent this diffuse injury represents a primary degenerative process. Identifying and segmenting the focal WM lesions is an essential first step in characterizing the MS disease burden in MS, and in calculating and interpreting more specialized measures of damage.

Over the last decade, a number of immunomodulatory and immunosuppressive therapies have been approved for clinical use based primarily on demonstration of their ability to suppress focal inflammatory activity, i.e., lesion formation, and clinical relapses (Comi et al., 2001; Leary et al., 2003; Li and Paty, 1999; Li et al., 2001; Miller et al., 1999; Paty and Li, 1993; Simon et al., 1998; Simon et al., 2000; Zhao et al., 2000). Development of these therapies has been critically dependent on MRI because of the ability of MRI to visualize lesion formation with an order of magnitude greater sensitivity than clinical observation is able to detect relapses The number and volume of MS lesions represent the "burden of disease" and are predictive of patients' clinical course over the long term (O'Riordan et al., 1998).

Before the advent of computers in radiology, lesions were visually identified on film by experts and counted. With the integration of digital images in radiology departments, it became possible to manually segment the lesions and estimate total lesion load (TLL). However, manual segmentation of MS lesions is time consuming and suffers from large intra- and interexpert variability. Several authors (Filippi et al., 1995a; Grimaud et al., 1996; Udupa et al., 1997) have proposed semiautomatic segmentation methods whereby the computer aids the expert to reduce both segmentation time and rater variability. Ultimately, the objectives are to obtain an automatic segmentation method that is completely reproducible and that enables the efficient processing of hundreds or thousands of images acquired for research studies and in clinical trials. Although many automatic methods have been proposed in the last 15 years, no single method is widely employed, for a variety of reasons to be discussed below.

Due to the heterogeneity of lesions and variability in the magnetic resonance (MR) acquisitions, no perfect solution has yet been found. There has been only limited validation on multicenter data, which is crucial to proving the utility of the automatic methods. As we will show in this paper, much work has been done to improve the quality of the segmentation, and newer methods provide more complex approaches to deal with the variability. Validation is still limited though, probably due to the difficulty of providing a good validation framework. With the exception of the MS Lesion Segmentation Challenge at MICCAI 2008 (Styner et al., 2008), few comparisons have been made of the different methods, and only a limited number of methods are freely available to the community. Recently, a review paper was presented where some semi-automatic and automatic segmentation methods were compared (Mortazavi et al., 2011). While they give an extensive description of different methods, the two main limitations of this paper are that semi-automatic and automatic methods are mixed, and numerical results for each method are compared even though they were obtained with different sets of images. We do not compare semi-automatic and automatic methods in this review because these methods have different objectives: Automatic methods try to be as accurate as possible compared to a expert's segmentation. On the contrary, when using semi-automatic methods the expert can always modify the segmentation until satisfied. The objective of semi-automatic methods is more to reduce the processing time and the intra- and inter-variability while automatic methods are evaluated for accuracy and precision. Because of these different goals, we review only automatic methods here.

In this paper, we review the automatic segmentation of MS lesions, with a focus on three main aspects: the complexity of MS lesion segmentation; the existing literature on automatic segmentation methods of MS lesions (and their validation); and the current challenges facing those methods.

2. Segmentation of MS lesions

We first describe the appearance of MS lesions in cMRI and then focus on the aspects that render their segmentation complex, thereby resulting in high variability.

2.1. Definition of MS lesions on cMRI

The classic inflammatory demyelinating MS lesion in the WM of patients with MS, and is well visualized on cMRI. However, recent advances in neuropathological techniques have revealed that there is significant demyelination occurring in the GM of patients with MS (Peterson et al., 2001). These GM lesions are not detected on cMRI because the pathological processes responsible for these lesions do not appear to alter the MR tissue properties of T2 and T1 sufficiently to be imaged with conventional sequences. Newer sequences have been developed (Geurts et al., 2005; Nelson et al., 2007) that improve the detection of some types of GM lesion, but these sequences are not employed routinely and are out of the scope of this paper. Therefore, we focus on WM lesions (WML) only.

As yet, there exists no well-established, precise definition of a WML on MRI. In general, MS lesions are brighter than the surrounding WM in T2w, proton density-weighted (PDw), and FLAIR images, although, in the case of severely destructive lesions, the lesion center can be darker in FLAIR images. Lesions usually have an ovoid or round shape and are centered on small blood vessels. They are most likely to occur in the periventricular WM and the juxtacortical and infratentorial regions (Fazekas et al., 1999).

Lesions are often classified into three main groups (*Figure 1*) based on their intensity under different MR sequences:

- **T2w lesions:** These lesions appear hyperintense compared with normal-appearing WM in T2w, PDw, and FLAIR images. They may be iso- or hypointense in T1w images. T2w lesions are not pathologically specific and can result from inflammation, edema, demyelination, or axonal loss.

- **Gadolinium (Gd)-enhanced lesions:** These lesions show an increase in intensity on T1w images after injection with gadolinium, and are usually associated with hyperintensity in T2w, PDw, and FLAIR images. Some lesions that appear hypointense in comparison to normal-appearing WM on T1w images before gadolinium injection may only become isointense with NAWM after entry of Gd in the lesion. These lesions can often be missed if a pre-injection T1w image is not acquired for comparison. Gadolinium enhancement is associated with active inflammatory activity and breakdown of the blood-brain barrier (BBB).

- **Black holes:** This term is usually used to refer to chronic T1w hypointense lesions. These lesions usually appear hyperintense in T2w, PDw, and FLAIR images. Since transient inflammation may be associated with hypointensity in T1w images, some hypointensities in T1w images may disappear after a month or two. Thus, to qualify as a "black hole", a T1w lesion should not enhance with gadolinium and should generally have been present for at least several months. Such lesions are usually associated with relatively more severe tissue injury and axonal loss.

Inflammation and atrophy also occur outside of the focal lesions throughout the whole brain (both WM and GM). This aspect of the disease is often referred to as diffuse disease or nonlesional pathology. For this reason, WM and GM in MS patients are often referred to as "normal-appearing white matter" (NAWM) and "normal-appearing grey matter", respectively, to distinguish them from normal WM and GM in healthy subjects. In cMRI, some of the nonlesional pathology may be visible as a diffuse increase

in the intensity of WM, which is usually referred to as *dirty white matter* (Figure 2). This dirty white matter often surrounds lesions and can complicate the segmentation of the focal lesions.

In summary, there is a relative consensus in the literature regarding the imaging characteristics of lesions that enables accurate lesion detection and counting. However, the actual level of how hypo-intense or how hyper-intense a lesion should be results in subtle differences between groups working in the field when determining the lesion boundaries and thus limits agreement in metrics of total lesion volume and lesion overlap.

2.2. Description of the segmentation

Manual segmentation of WML can be separated into two processes: detection of lesions and segmentation of the lesion border.

2.2.1. Challenges in the detection of WML

The main characteristic of a WML is that its intensity is brighter than the NAWM on T2-w, PD and FLAIR, a definition that lacks precision. For example, the specific threshold of intensity increase depends on the imaging sequence parameters and can vary between different experts working in the field. Hence, detection of WML in manual segmentation methods requires a good knowledge of neuroanatomy and experience with MRI to correctly identify all WML in an image and avoid false positives. The following challenges are faced in manual detection of WML.

First, the radiologist (or other expert) must combine the information from the different MRI sequences and identify the three-dimensional (3D) structures of the brain that are shown in only two dimensions on coronal, sagittal, or transverse slices. Second, healthy tissues (e.g., the choroid plexus, blood vessels, tail of the caudate nucleus), which may resemble lesions in their intensity depending on the image modality and the amount of partial-volume, must be correctly identified as such to avoid misidentification as lesion. Similarly, lesions around the frontal or posterior horns of the lateral ventricles can be difficult to differentiate from the phenomenon of hyperintense periventricular caps found in normal subjects (Neema et al., 2009). In addition, when mixed with partial volume effects due to relatively large slice thickness, Virchow-Robin spaces can sometimes be mistaken for lesions. Third, any non-MS pathology that may occur in patients with confirmed MS, such as lesions of vascular origin or, particularly in patients treated with very strong immune modulating therapies, progressive multifocal leukoencephalopathy (PML) lesions must be identified. Fourth, the MR protocol employed to acquire the images influences the number of lesions detected for the same patient. Using thinner slices or higher magnet strength increases the number of detected lesions (Filippi et al., 1995b), while using FLAIR contrast results in more lesions detected in the supratentorial region when compared with dual-echo (T2w, PDw) images (Simon et al., 2006). Finally, juxtascortical lesions can be difficult to detect because of their resemblance to the surrounding tissues in T2w and T1w images (Figure 3).

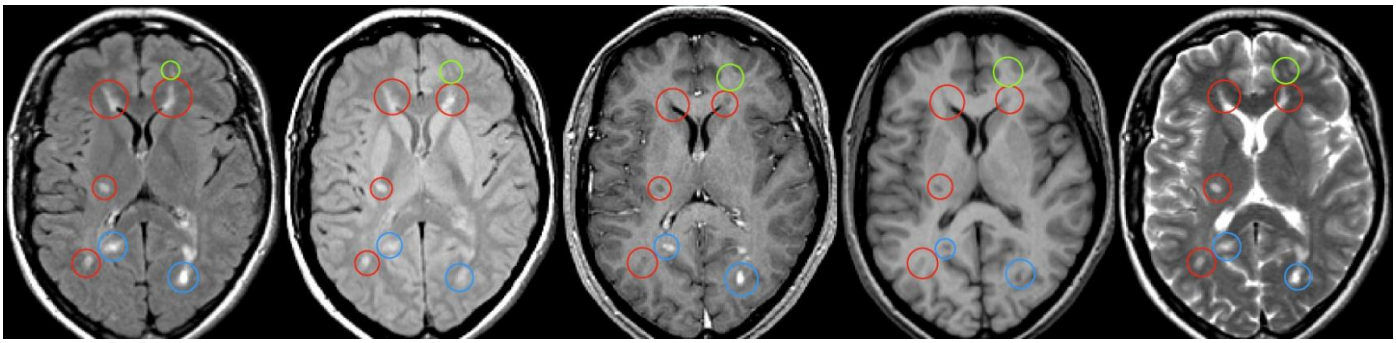

Figure 1. Example of MS lesions on MRI. From left to right: FLAIR, PDw, Gd-enhanced T1w, T1w, and T2w images. Several types of MS lesions can be observed: Enhancing lesions (blue), lesions visible only on T2w (green), black holes (red). In Gd-enhanced T1w and FLAIR images, multiple bright regions are observed that may be mislabeled as lesions.

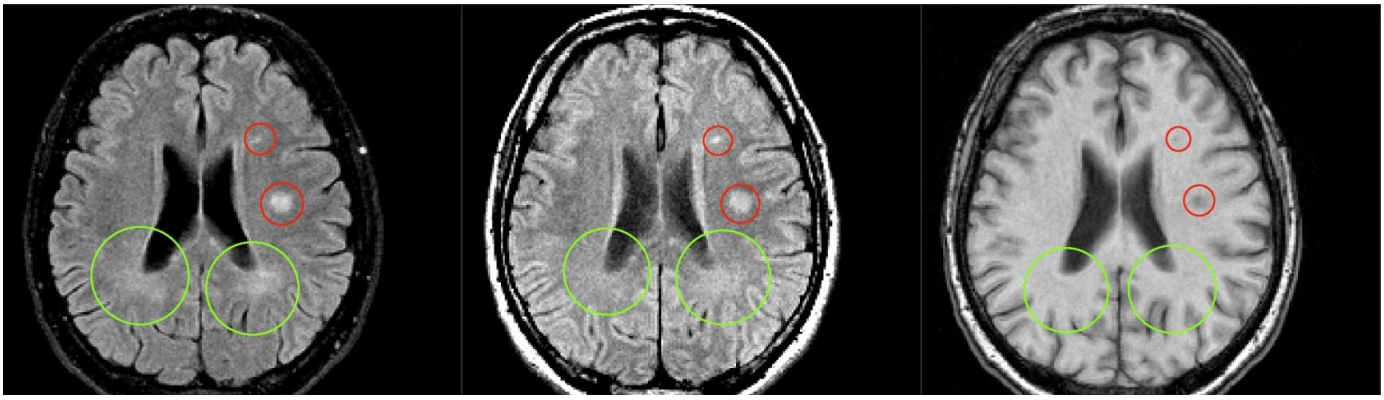

Figure 2. Difference between focal lesions (red) and diffuse regions of visible-abnormal WM (green). From left to right: FLAIR, PDw, and T1w images.

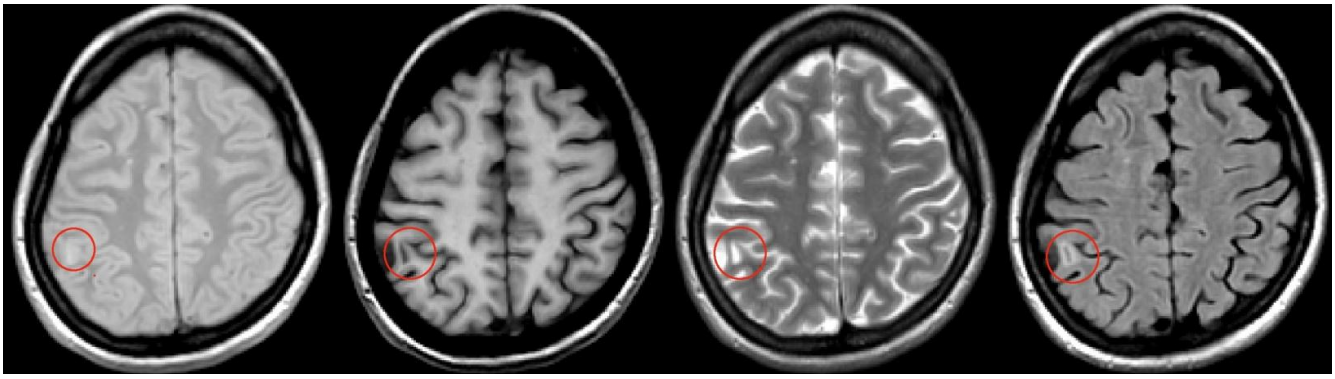

Figure 3. Example of juxtacortical lesion. From left to right: PDw, T1w, T2w, and FLAIR images. The detection of juxtacortical lesions (red) requires a good understanding of the brain anatomy and the simultaneous use of several sequences.

2.2.2. Segmentation of WML

Once a lesion is identified, its boundaries must be determined, even if its boundaries are not clear. Segmentation is the process whereby a reader decides which voxels belong to the lesion and which do not.

Segmentation is complicated for several reasons. Histologically, lesions may have sharp or fuzzy borders. When imaged with MRI at standard, clinically feasible resolutions, delineation of the lesion borders is further complicated by partial volume effects, where several tissue types may contribute to the image intensity of the border voxels. Manual segmentation of lesion borders is affected because the eye detects local contrast better than absolute intensities. Further, this contrast is much more convincing if a lesion is surrounded by NAWM as opposed to “dirty” white matter. Similarly, intensity contrast is seen more easily between two regions separated by a clear boundary than by two diffusely separated regions. Consistency in lesion segmentation throughout slices is therefore hard to achieve, and consistency

between different reading sessions is made difficult because the reader sets the overall contrast of an image, which influences the relative appearance of tissues. Many studies have investigated the wide variability inherent to manual lesion segmentation by neuroradiologists or other trained experts and have reported volume differences ranging from 10% to 68% (Grimaud et al., 1996; Styner et al., 2008; Zijdenbos et al., 2002).

Lesion borders also differ depending on the MRI protocol employed. Segmented lesions on FLAIR images have been shown to provide larger volumes than T2w images, although the correlation with the disease is similar for the two sequences (Ciccarelli et al., 2002). Thinner slices enable more accurate estimation, and may yield larger total lesion volumes (Filippi et al., 1995b). In addition, images acquired with 3T scanners show more lesions than images acquired with a 1.5T (Sicotte et al., 2003).

To reduce the variability inherent in manual lesion segmentation and improve consistency, many semiautomatic methods have been proposed to aid the reader. Many of these methods rely on the reader to detect the lesion, and then performs the segmentation (under user control) based on a local threshold (Filippi et al., 1995a), region growing (Ashton et al., 2003; Parodi et al., 2002), fuzzy connectedness (Udupa et al., 1997), or the intensity gradient (Grimaud et al., 1996). All of these methods have demonstrated reduced variability in comparison to manual segmentation for lesion quantification.

Manual detection of lesions is a complicated task, and the intra- and inter-reader agreement when reporting new and enlarging lesions is moderate, at best (Molyneux et al., 1999). Other semiautomatic methods exist to assist in both the detection and segmentation of lesions (Ashton et al., 2003; Johnston et al., 1996). Depending on the method, these may require, for example, manual segmentation of one lesion (Ashton et al., 2003) or seed points for different tissues (Johnston et al., 1996).

Although semiautomatic methods reduce intrareader variability, their use in large clinical trials is time consuming and deals incompletely with inter-rater differences. The development of automatic methods is thus necessary to avoid human interaction, reduce the corresponding rater variability, and enable efficient segmentation of many images. However, automatic methods require an exhaustive validation of their performance before they can be employed in clinical trials.

3. Methodology of the review

To capture all segmentation methods for this review, a search was performed on two main search engines, PubMed and Google Scholar, on September 29, 2011. Our criteria included both peer-reviewed journal papers and conference papers, but abstracts and theses were excluded. If both a journal and a conference paper were found for the same method, then the conference paper was discarded and the journal version was cited.

The vocabulary used to describe WML segmentation is very heterogeneous, and many related terms are employed (e.g., classification, detection or delineation). The search expressions included this heterogeneity (cf., Table 1).

Additional selection criteria were used to further focus our search. The methods described in the papers had to be:

- completely automatic (no interaction);

- employed on clinical or realistic MR images of MS patients; and

- evaluated with some quantitative measure on clinical images.

Table 1: Search expressions used to identify potential papers for review

Search Engine	Search Expression
PubMed	("Multiple sclerosis" [Title/abstract] or "MS" [Title/abstract]) AND (classif*[Title] OR segment*[Title] OR delineat*[Title] or detect*[title]) AND (MR[title/abstract] or MRI[title/abstract] or Magnetic[title/abstract])
Google Scholar	("Multiple sclerosis") and (MRI OR MR OR "Magnetic resonance") and (intitle:~classify OR intitle:~segmentation OR intitle:~delineation) AND lesions

Table 2: Selection criteria

Selection	**Rejected**	**Remaining**
Initial search: PubMed		533
Initial search Google Scholar		707
Selection by title/abstract	1108	132
- Rejected: No segmentation	12	
- Rejected: No MS	15	
- Rejected: Semiautomatic	25	
Automatic methods		80
- Rejected: Abstract only	8	
- Rejected: Conf./Journal copy	5	
- Rejected: No quantitative validation	20	
Included in the review		47
- BrainWeb only		5
- Monocenter validation		29
- Multicenter validation		13

3.1. Results of the search

Our search (PubMed + Google Scholar) resulted in 1240 hits (Table 2). During an initial parsing phase, a first selection from these papers was performed by reading the title and abstract wherein duplicated entries and papers not referring to automatic segmentation were removed, reducing our list to 132 papers. From this list, we found 10 papers that did not apply to MS, 11 papers that did not deal with segmentation, and 25 papers that described semiautomatic methods.

We were thus left with 80 papers that described an automatic method of segmenting WML in MS. Of these, 8 conference abstracts were rejected as there was no associate paper, and 20 papers had no validation or only a qualitative validation. Five conference papers were removed as a journal paper describing the method was found. In the end, 47 papers met our criteria and are the object of this review. Of these, 5 papers used only synthetic images for validation, 29 used monocenter data, and 13 used multicenter data.

4. Review of segmentation methods

4.1. Automatic segmentation pipeline

We define a segmentation *pipeline* as the entire ensemble of image processing steps necessary to accomplish the segmentation of lesions from the raw images. This pipeline can be further divided into preprocessing steps and the segmentation method itself. Preprocessing refers to all the changes made to the image prior to segmentation. For the most part, the goal of preprocessing is to minimize the effect of imaging artefacts and align the different sequences in the same space.

The majority of the papers in our review focuses on the segmentation methods, with little or no description of the segmentation pipeline. All methods that use more than one sequence (i.e., multiple modalities such as T1w, T2w, and FLAIR) assume that the images have been previously registered into a common space. In addition, some methods involve some kind of image enhancement, and the segmentation is usually restricted to the region of the brain only, after application of some kind of skull-stripping or brain extraction procedure.

The following is a brief summary of the preprocessing steps often applied prior to the segmentation procedure:

- **Registration:** MR sequences of the same patient are registered into the same space (patient space or stereotaxic space) (Collins et al., 1994; Maes et al., 1997). Registration might also be employed to align an atlas to the brain to provide some initial estimates of the brain tissues.
- **Brain extraction:** The brain is extracted from the image, and the segmentation takes place only on the remaining brain voxels (Fennema-Notestine et al., 2006; Smith, 2002).
- **Intensity inhomogeneity (IIH) correction:** The intensity of the same tissue can vary across the image due to inhomogeneity of the static or applied magnetic fields within the scanner (Sled and Pike, 1998). IIH correction methods (Vovk et al., 2007) reduce the smooth variation of intensity of the tissues to simplify subsequent segmentation, although some segmentation methods take this effect into account (Ashburner and Friston, 2005; Van Leemput et al., 1999; Wells III et al., 1996).
- **Noise reduction:** The acquisition process can induce some noise in the image (Dietrich et al., 2008). Several noise reduction methods in the literature (Bao and Zhang, 2003; Coupé et al., 2008; Gerig et al., 1992) have been applied to MR data from patients with MS. The negative

effect of noise in the image can also be reduced by incorporating spatial information in the segmentation (i.e., Markov random fields (MRF) (Ahmed et al., 2002; Zhang et al., 2001)).

- **Intensity normalization:** Some segmentation methods require the intensity of the image to be similar to the intensity of the training images and thus depend on an intensity normalization step. Intensity normalization methods modify the intensity range of the target image and map them into a predefined intensity range (Nyul et al., 2000).

The effect of image pre-processing on segmentation is important, and not often mentioned in the papers. However, the effects of each preprocessing step might affect each method differently making difficult any generalizations. A few authors (García-Lorenzo et al., 2008c; Kikinis et al., 1999; Wei et al., 2002) have evaluated their methods under different preprocessing pipelines. The comparison of preprocessing steps and their interaction is complex and is beyond the scope of the paper.

As described above, WML can be further classified into different groups using the information provided by other sequences. Most automatic methods consider all lesions in the same class of WML, although there are methods that model each lesion type independently (Subbanna et al., 2009). From this generic WML class, a second classification step can be performed to segment separately T2w lesions, black holes (Datta et al., 2006), and Gd-enhanced lesions (Datta et al., 2007; He and Narayana, 2002; Karimaghaloo et al., 2010; Khayati et al., 2008b).

4.2. Classification of methods

Several reviews of segmentation methods have classified the techniques according to the mathematical algorithms behind them (Clarke et al., 1995; Suri et al., 2002). In our review, we divide the methods into two groups according to the following classification:

- **Supervised learning methods:** These methods "learn" the definition of lesions from example images that have been previously segmented by another method, usually manual segmentation. Example images usually come from a "training database" and must be available before this class of methods can be used on other data.
- **Unsupervised methods:** These methods do not require labeled training data to perform the segmentation. Most employ clustering techniques to separate the voxels into different classes (or clusters) based on, for example, voxel intensity. Typically, these classes are then assigned to WM, GM, cerebrospinal fluid (CSF), or WML according to some a priori information.

Supervised learning methods have been widely employed and are very efficient in tasks where the training database covers all possible cases. In MS, because the heterogeneity of the disease and the potential variability of MR acquisitions are large, the training database must be chosen carefully. In addition, the training database has to be segmented; as mentioned above, manual segmentation is time consuming, and high intra- and inter-rater variability may induce errors in the training database. By contrast, unsupervised methods do not rely on a training database and segment each image independently. Unsupervised methods are strongly based on *a priori* information in order to perform the segmentation. The *a priori* information is extracted and simplified from the expert's knowledge and combined with unsupervised segmentation or classification algorithms. The complexity comes from the translation of the expert's knowledge into an algorithm that can be implemented.

In MS, clinicians are interested not only in the lesions, but also in the whole brain, as MS patients develop progressive atrophy (Figure 4) that is regionally heterogeneous. Some methods are focused only on the segmentation of WML, while others also provide classification of the normal-appearing brain tissues (NABT), thereby providing information on brain atrophy. These techniques will be identified in the review below.

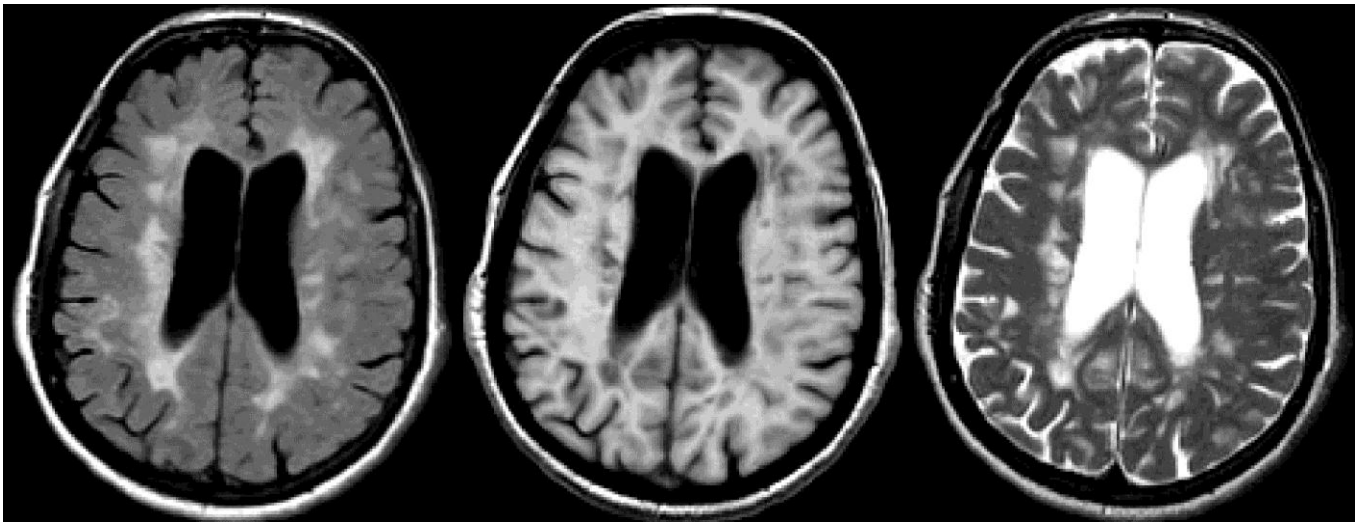

Figure 4. Patient with high lesion load and atrophy. From left to right: FLAIR, T1w, and T2w images. In patients with severe MS, the atrophy becomes more visible and the lesion load becomes very high and includes most of the white matter.

4.2.1. Supervised learning methods

The primary objective of these methods is to segment MS lesions using a database of previously segmented images. To compare any incoming image with those in the database, a set of image features must be selected to enable the discrimination of WML from other tissues. A feature is any characteristic of an image that can be employed to discriminate tissue classes (e.g., the intensity of the voxel, the local gradient of the intensity). The set of selected features is employed in one of the many available supervised learning methods to perform the segmentation (e.g., k-nearest neighbours (k-NN), artificial neural networks (ANN), Bayesian framework, or support vector machine (SVM)). For a detailed description of these methods, the reader should refer to the classic book by Duda et al. (Duda et al., 2000). A comparison of the performance of different methods can be found in Caruana and Niculescu-Mizil (Caruana and Niculescu-Mizil, 2006). The standard pipeline of supervised methods consists in four steps: preprocessing to normalize the data and reduce the effect of artefacts, feature generation, classifier training/application and post processing.

The normalization is the step whereby all the features are set in the same space in order to make the images comparable to each other. As cMRI is not quantitative, differences in brightness and contrast can be found between images, especially when they come from different scanners, and can be reduced by intensity normalization algorithms (Nyul et al., 2000). Depending on the algorithm employed, all features might also need to be normalized to have the same variance or zero mean. Some authors have proposed a segmentation method based on the normalization of images, which is in turn based on the intensity of the training datasets (Alfano et al., 2000; Erickson and Avula, 1998). For example, one normalization approach employed tissue that is constant across the images and not affected by the disease (i.e., subcutaneous fat) (Erickson and Avula, 1998). Another approach used spin-echo sequences with specific parameters to generate T1, T2, and PD relaxometry maps (Alfano et al., 2000). Once all images have the same intensity range, tissues can be classified using fixed intensity rules for all patients in the normalized multidimensional intensity feature space. Similarly, other methods employed a Parzen window density estimation for the segmentation (Datta et al., 2006) after performing histogram-based intensity normalization (Nyul et al., 2000). Yet another approach (Tomas-Fernandez and Warfield, 2011) proposed a method whereby the image is segmented by comparing its intensities on a voxel-by-voxel basis to a

group of images of healthy subjects. The WML voxels will have a large Mahalanobis distance compared to the intensities of healthy patients when using T1w, T2w and FLAIR images. This comparison requires that all images be registered to the same space using non-linear registration and normalized in terms of intensity to avoid scanner-related differences. The main advantage of this technique is that it compares the same anatomical regions in MS patients and normal controls instead of using a global (non spatial) approach.

The selection of features employed in the classifier is one of the aspects that differentiate the supervised methods. In the case of MS lesion segmentation, the classical features are the intensity of each voxel in the T1w, T2w, and PDw images, which reflects the a priori knowledge of lesion characteristics, namely, that lesions appear brighter than WM in T2w and PDw images. These sets of features were first employed in a supervised algorithm in a semiautomatic fashion using a Bayesian approach (Johnston et al., 1996), ANN (Zijdenbos et al., 1994), and k-NN (Vinitski et al., 1999), and in a decision tree classifier (Kamber et al., 1995). In these methods, the user had to select voxels from each class to train the supervised algorithm, but they could have easily been automated using a database of previously segmented datasets.

The use of intensity information enables the detection of lesions, but it is necessary to incorporate more information to improve the segmentation. Some authors proposed to combine an initial supervised intensity-based segmentation with other methods to reduce false positives (i.e., voxels incorrectly identified as lesion). For example, the automatic k-NN segmentation technique (based on image intensities) was combined with a template-driven segmentation method (Warfield et al., 2000) to include anatomical spatial information in the segmentation (Wu et al., 2006).

Other methods proposed to include spatial features directly in the segmentation method to improve its specificity. To include anatomical information in the ANN, Zijdenbos et al. (Zijdenbos et al., 2002) included the probability of the voxel belonging to a tissue class based on information derived from an atlas. Hadjiprocopis and Tofts (Hadjiprocopis and Tofts, 2003) included the Cartesian and polar coordinates of the voxel in addition to its intensity information. Younis et al. (Younis et al., 2007) added the intensity of the six neighbor voxels into the ANN to include local spatial information in the segmentation, thus avoiding a noisy segmentation. They employed the ANN twice, first to classify the brain tissue in the T1w image and, after removing the CSF, to segment the T2w image and obtain the WML. Anbeek et al. (Anbeek et al., 2008) also employed both the intensity and spatial features in a k-NN algorithm using FLAIR images. One important aspect of their method is that the intensities and spatial coordinates must be normalized such that the distance between spatial and intensity features is comparable. A recent paper (Cerasa et al., 2012) employed the Cellular Neural Networks (CNN) to segment MS lesions on 2D images. The advantage of this approach compared to the classical ANN is that neighboring cells interact with each other, providing local spatial information to the segmentation.

More recent methods rely on a large set of features, selecting the more discriminant ones and thus reducing the bias given by previous methods where ‘a priori’ features were selected by the authors. These methods (Morra et al., 2008; Wels et al., 2008) introduced more than 10,000 features in the supervised classification algorithm using an Adaboost approach (Freund and Schapire, 1995) with a probabilistic boosting tree to improve the training process. These features included Haar-like filters of various shapes and sizes, intensity values, and curvature filters. From the five modalities employed in the segmentation (T1w, T2w, FLAIR, mean diffusivity (MD), and fractional anisotropy (FA)) (Morra et al., 2008), the method selected a number of FLAIR features as those that most contributed to lesion classification.

Another method employed principal component analysis (PCA), whereby the data is transformed to a new orthogonal coordinate system in which the first dimension (or component) “covers” the greatest variance in the data. The classification can be performed more easily in this new coordinate system using only the first components of the PCA with a simple threshold. In addition to the voxel intensity, up to

225 features were used in a PCA to segment MS lesions (Kroon et al., 2008). These features included intensity of the neighbor voxels in T1w, T2w, FLAIR, MD, and FA, intensity derivatives, spatial coordinates and atlas probabilities. A threshold was then computed using training data in this new coordinate system to perform the segmentation.

The random decision forest (RDF) method (Menze et al., 2009) has also been employed to segment MS lesions (Geremia et al., 2011) using two types of features simultaneously: local and context-rich features. Local features are similar to those employed in other methods, while the context-rich features compare the intensities of the voxel of interest with distant regions either in an extended neighborhood or in the symmetric counterpart with respect to the midsagittal plane. The advantage of RDF over other methods is that it enables interpretation of the decisions made by the algorithm. In this particular case, the results showed that the FLAIR image intensity was the most discriminative feature, followed by the information of the atlas, which rejected voxels that do not belong to WM.

The Bayesian framework has also been employed in the segmentation of MS lesions. In this context, the training dataset was used to compute the prior probability model of the segmentation, which was then combined with the observed data using Bayes' theorem. An extension of this framework considers that the intensity of the voxels of each tissue follow a Gaussian distribution, and the training data is employed to determine the parameters of the Gaussian distribution for each tissue. This idea was improved upon by Harmouche et al. (Harmouche et al., 2006), who observed that different parts of the brain had different intensities and that these differences were particularly important in the posterior fossa. Their method proposed to model each region of the brain with a different Gaussian distribution to refine the segmentation. The resulting prior probability depended on the intensity and region of the voxel. They also incorporated MRF into their method to take advantage of local spatial information.

Another method (Subbanna et al., 2009) proposed to solve the MRF minimization using a simulated annealing algorithm that modeled black holes and T2w lesions as different Gaussian classes. Errors in the center of the lesions were corrected using b-spline smoothing. Scully et al. (Scully et al., 2008) proposed to employ a naïve Bayesian classifier to segment MS lesions in T1w, T2w, and FLAIR images by constructing a nonparametric model for the lesions using the training data in the form of a joint histogram. They then combined this information with the result of a k-means classification of the T1w image to obtain a rough classification of WM, GM, and CSF and merged the results to obtain the final lesion segmentation.

Other methods use the supervised classification not at the voxel level but at a lesion level. The image is partitioned into homogeneous regions and each region is classified by the method as a lesion or not. In this case, the supervised method is only used as classification algorithm because the region was detected by other means. Goldberg-Zimring et al. (Goldberg-Zimring et al., 1998) employed an adaptive thresholding technique to obtain candidate lesions, which included both lesions and other artefacts that resembled lesions. The artefacts were then discarded in a second stage by the ANN using as input the shape and intensity of the candidate lesion. Yamamoto et al. (Yamamoto et al., 2010) employed a multithreshold technique to detect candidate lesions and then a level set to improve the border of the candidate WML. Finally, false positives were rejected with an SVM that used shape and intensity features. Another approach (Akselrod-Ballin et al., 2009) proposed to first partition the image into homogeneous regions using the segmentation-by-weighted-aggregations (SWA) algorithm (Sharon et al., 2006). Each region was then classified using a decision forest classifier that used multiple features. Some of these features were given directly by the SWA (saliency, average intensity), while other features were calculated within each segmented region and included shape moments, neighborhood statistics, and anatomical probabilities.

4.2.2. Unsupervised methods

Unsupervised methods do not require a training database to perform the segmentation and are instead based on a parametric definition of lesions and the anatomy of the brain. They employ clustering methods, that is, unsupervised classification methods that regroup similar voxels into clusters or classes.

For healthy subjects, these methods have been employed to classify brain voxels into WM, GM, and CSF using tissue contrast in T1w images. For example, voxels have been classified into three classes (or clusters) according to their intensity using a fuzzy C-means (FCM) (Pham and Prince, 1999) and a finite Gaussian mixture model with the expectation-maximization (EM) algorithm (Wells III et al., 1996). The Gaussian mixture model assumes that each class of the model follows a Gaussian distribution and the parameters of each Gaussian can be easily estimated using an EM algorithm (Dempster et al., 1977). The fuzzy C-means(Bezdek, 1981) provides a fuzzy clustering of the voxels estimating the center of each class and has the advantage of removing the Gaussian assumption of the EM algorithm but the clusters should have and spherical shape to be accurate.

For MS lesion segmentation, methods take into account that lesions appear brighter than NAWM in FLAIR, T2w, and PD images and are usually located in the NAWM. The majority of methods will consider that every voxel of the brain can be classified into four classes: WM, GM, CSF and lesions. Combining the information of the T1-w image, which has a good contrast for healthy tissues, with T2-w, PD and/or FLAIR, which have a good sensibility for lesions. These methods have the advantage to segment the whole brain providing information about volume of the normal appearing healthy tissues in addition to the lesions segmentation.

Some authors extended the previous methods for healthy brains to use several sequences at the same time and added an extra class to include the WML in their classification (Forbes et al., 2010; Guttmann et al., 1999; Kawa and Pietka, 2007; Kikinis et al., 1999; Shahar and Greenspan, 2004; Shiee et al., 2009).

An EM approach (Wells III et al., 1996) employed with an extra class for WML (Guttmann et al., 1999; Kikinis et al., 1999) was later improved by being combined with a template-driven segmentation and a partial volume correction (Wei et al., 2002). In another approach, the Gaussian mixture model used multiple Gaussians (Dugas-Phocion et al., 2004) to include partial volumes; this method provided a more accurate estimation of the mean and variance of each tissue and, hence, a more accurate segmentation of WML (Souplet et al., 2008). Forbes et al. (Forbes et al., 2010) provided a model wherein the weight of each sequence differed for each class estimated using an EM algorithm. Their technique is related to the idea that FLAIR images should be more important than T1w images in the classification of WML, although T1w images provide a better distinction between the GM and WM classes. Bijar et al. (Bijar et al., 2011) proposed a monomodal EM-approach using the entropy instead of the likelihood as cost function using only FLAIR images for the segmentation.

Two approaches included longitudinal data from the same patient in their models. Shahar et al. (Shahar and Greenspan, 2004) proposed a Gaussian mixture model that incorporated the time dimension to follow the evolution of the lesions. Another option was to combine the voxels of all time points in the same joint histogram to have more samples to perfom the Gaussian mixture model estimation (Aït-Ali et al., 2005). In both cases, it was important that all images were acquired with the same parameters on the same scanner to avoid variations in the image contrast, which could lead to misclassifications.

Because the use of only one Gaussian to model the intensity of a tissue oversimplifies the problem, some authors have proposed to include *many* Gaussians in their models to overcome this limitation (Freifeld et al., 2009; Khayati et al., 2008a). In one case, a constrained Gaussian mixture model was proposed in which each Gaussian simultaneously included spatial and intensity features, thereby providing a much more complex model of the brain (Freifeld et al., 2009). In another case, an adaptive mixture model was proposed in which the final number of Gaussians of the model was computed automatically (Khayati et al., 2008a).

Other authors proposed not to model the lesions, but to consider them as outliers to the normal appearing brain tissues model. The advantages of this approach are that it avoids the need to model the intensity of the heterogeneous lesions and also provides a more robust estimation in the presence of other tissues (e.g., vessels) or artefacts. Van Leemput et al. (Van Leemput et al., 2001) proposed a weighted EM algorithm in which the voxels situated far from the model were weighted less in the estimation and considered potential lesions. The trimmed-likelihood estimator (Neykov et al., 2007) has also been used to avoid outliers in the estimation (Aït-Ali et al., 2005; Bricq et al., 2008b; Garcia-Lorenzo et al., 2011). Recently, the integrated square estimation (Scott, 2001) was employed as a cost function because it is a robust estimation method that has the advantage of not having any parameters to tune (Liu et al., 2009). Prastawa and Gerig (Prastawa and Gerig, 2008) also employed a robust estimator, called the minimum covariance determinant, using an atlas to initialize the estimation of the model. Given that the probability of a voxel being lesion is very low, the *a contrario* framework (Desolneux et al., 2003) was employed to segment the lesions as outliers to the normal appearing brain tissues model (Rousseau et al., 2008). The WML were segmented from the outliers using heuristic rules (e.g., the intensity of WML is brighter than the mean intensity of the WM on T2w images).

Most segmentation models are based only on intensity information, which is insufficient because of the presence of noise and other artefacts. The inclusion of spatial information is necessary to improve lesion segmentation. As in the case of the supervised models described above, several authors using unsupervised methods have proposed to apply MRF to include the local neighborhood in the estimation (Khayati et al., 2008a; Van Leemput et al., 2001). Bricq et al. (Bricq et al., 2008b) proposed a double Markov chain that couples information from the neighbor voxels with information from an atlas to improve the segmentation.

With the goal of including spatial information, other authors combined segmentation methods with the intensity model estimation. Some authors first performed a parcellation of the brain and then used the model to classify each local region of the brain, for example, using the mean shift (García-Lorenzo et al., 2008a; Rousseau et al., 2008) or the watershed algorithm (Prastawa and Gerig, 2008). Other authors proposed to employ boundary-based methods to improve the contour performed by a robust EM that uses the graph-cut algorithm (García-Lorenzo et al., 2009) or active contours (Freifeld et al., 2009).

Some authors remove the Gaussian assumption used in the EM method, replacing it with the fuzzy C-means. A hierarchical segmentation method was proposed whereby the fuzzy C-means is used twice in PDw images: first, to detect the lesion+CSF cluster with respect to the GM and WM and, second, to differentiate lesions from CSF (Boudraa et al., 2000). Kawa and Pietka (Kawa and Pietka, 2007) proposed to include spatial information in the clustering using the kernel fuzzy C-means combined with fuzzy connectedness theory (Udupa et al., 1997). Shiee et al. (Shiee et al., 2009) modified the fuzzy C-means algorithm to include anatomical information using both anatomical and topological atlases. The anatomical atlas provided information on where a given tissue is more likely to be present, while the topological atlas provided information about the topology of the structures (e.g., all WM must be connected in a healthy brain).

Finally, because the definition of lesions is not precise, some authors proposed to employ the fuzzy Dempster-Shafer theory of evidential reasoning (Zhu and Basir, 2003). After each sequence was classified using k-means to obtain a coarse tissue segmentation, each voxel was defined using fuzzy values (small, medium, and large), according to its intensity. Then, fuzzy rules were defined for each tissue: “If T1w intensity is medium, T2w intensity is large, and PD intensity is large, then the voxel is WML.” The advantage of this method is that the fuzzy rules resemble the verbal description of lesions. However, the definition of the fuzzy values is still complex, and it is not obvious how to maintain consistency across subjects.

5. Review of validation

Validation of a segmentation method should evaluate both the performance and limitations of the algorithm as well as clarify the context of applicability of the method (Jannin et al., 2006). In medical image processing, complete validation of a method is a necessary step before it can be applied in a clinical setting. In this review, we examine the validation methods used in the literature, including the data employed and measures performed. Finally, we describe the MS Lesion Segmentation challenge as an example of automatic segmentation method validation.

5.1. Validation data

Any validation method requires data to evaluate the performance of the algorithm. Here, we differentiate between synthetic data and real clinical data.

5.1.1. Synthetic images

Synthetic images are created by a computer without using a real scanner. The advantages of these images are that the user can fully control all the parameters in the image and the ground truth is available. Different levels of image complexity range from piecewise constant images (Kamber et al., 1995) to realistic images generated by an MRI simulator (Kwan et al., 1999).

In WML segmentation, the images most widely employed for validation are those from BrainWeb (www.bic.mni.mcgill.ca/ServicesBrainWeb), an online database of synthetic MR images (Cocosco et al., 1997; Collins et al., 1998). A healthy subject was scanned 20 times to obtain a high-SNR image (Holmes et al., 1998). From this image, a healthy anatomical phantom was created in which each voxel belongs to a specific tissue class. Simulated MR images are created using this phantom and an MRI simulator (Kwan et al., 1999). The simulator allows the choice of MR parameters (TE, TR, resolution, sequence, etc.) and certain artefact parameters (noise and intensity inhomogeneity) to generate three conventional imaging sequences: T1w, T2w, and PDw.

From the healthy phantom, three different MS phantoms (Zijdenbos et al., 2002) were created using manually segmented lesions from real patients with different levels of lesion load: mild, moderate, and severe. There are three main advantages to using BrainWeb images to evaluate WML segmentation algorithms:

- **Ground truth:** The existence of a ground truth simplifies the validation in comparison to using clinical images.
- **Different inhomogeneity and noise levels:** Synthetic images allow the evaluation of noise and inhomogeneity robustness.
- **Freely available:** All authors use the same images, simplifying the comparison of methods within the literature.

These phantoms have been employed to evaluate WML segmentation algorithms because of the simplicity of the evaluation. Yet, despite being a good method for comparing different techniques, it suffers from serious limitations:

1. **Only one phantom:** Only one brain model means no anatomical variability. Efforts have been made to address this issue (Aubert-Broche et al., 2006).
2. **Too simple:** Despite being based on real data, the final image is not completely realistic, and is much easier to segment than clinical images.
3. **Limited to T1w, T2w, and PDw:** Other useful MRI sequences such as FLAIR or Gd-enhanced T1w are not available, limiting the validation of certain methods. In addition, T2-w

and PD-w images are based on conventional spin echo, while fast spin echo is more common in actual protocols.

4. **Reduced lesion load range:** Only one type of lesion is included, and the lesion load range is very limited. The *severe lesion load* phantom has only 10 cm^3 of lesions, yet it is not uncommon to have lesion loads of more than 100 cm^3 in patients with severe disease.

In our review, only 12 methods were evaluated using BrainWeb (Aït-Ali et al., 2005; Akselrod-Ballin et al., 2009; Bricq et al., 2008a; Forbes et al., 2010; Freifeld et al., 2009; García-Lorenzo et al., 2009; García-Lorenzo et al., 2008a; Rousseau et al., 2008; Shiee et al., 2009; Younis et al., 2007; Zhu and Basir, 2003; Zijdenbos et al., 2002). Within these 12 publications, 5 had no quantitative validation with real clinical data (Aït-Ali et al., 2005; Freifeld et al., 2009; Rousseau et al., 2008; Younis et al., 2007; Zhu and Basir, 2003). In two cases, a conference paper contained only BrainWeb information, but a later paper did include clinical validation (Akselrod-Ballin et al., 2006; Bricq et al., 2008b).

Although the BrainWeb database should enable good comparison across methods, direct comparison is limited by several factors:

- *Intensity inhomogeneity and noise*: Almost every paper employs different levels of IIH and noise. Four papers used only one level of noise and IIH, and only three methods varied both noise and IIH to provide information about the robustness of the method to both (García-Lorenzo et al., 2009; Shiee et al., 2009; Zijdenbos et al., 2002). In addition, only five methods employed the three available phantoms with different lesion loads (Akselrod-Ballin et al., 2006; Forbes et al., 2010; García-Lorenzo et al., 2009; Rousseau et al., 2008; Zijdenbos et al., 2002).
- *Validation metric*: Most methods employed either kappa, or the related Dice similarity coefficient (DSC) (Dice, 1945; Zijdenbos et al., 1994), but three methods selected only certain slices to perform the segmentation (Akselrod-Ballin et al., 2006; Freifeld et al., 2009; Zhu and Basir, 2003). This selection biases the results and prevents comparison of their results with those of other published methods.
- *Supervised methods*: The lack of a training database has limited the use of BrainWeb in validating supervised methods. Three different options have been employed to validate supervised methods using BrainWeb although they suffer from different limitations:
    - Use a clinical database: The ANN were trained using clinical data and then applied to the BrainWeb images (Zijdenbos et al., 2002). This validation is only possible when the clinical images are acquired with similar MR parameters (TE/TR/IR) compared to those from BrainWeb.
    - Use part of the image for testing: One hemisphere was employed as training data, while the other hemisphere was used for testing (Akselrod-Ballin et al., 2006). This method prevents direct comparison of results with those of other methods and overestimates generalization, as the exact same image is used for both testing and training.
    - Use a noiseless image for training: The noiseless BrainWeb image was employed as the training data (Younis et al., 2007), then the same image with noise was used for testing. This validation also employs the same image to test and train, thus limiting the comparison of their results to those of other methods.

5.1.2. Clinical images

Segmenting real clinical images is a necessary step in the evaluation of a segmentation method. There are two main aspects to describe for validation using clinical images: the characteristics of the database and the ground truth.

5.1.2.1. Characteristics of the database

The validation database should be representative of the heterogeneity of the disease. It is important to have a sufficient number of patients to cover this heterogeneity. Epidemiological and clinical information is helpful to better describe the results.

In the literature reviewed, the databases used for validation varied in size from one patient with longitudinal data (Shahar and Greenspan, 2004) to 41 patients (Akselrod-Ballin et al., 2009). Ten methods were evaluated with MRI data from fewer than 10 patients, and less than half of the papers (19 methods) were evaluated using at least 20 patients.

Clinical information can help in judging the applicability of a method to a particular class of patients. For example, lesions in patients with relapsing-remitting MS have slightly different imaging characteristics than lesions in patients with secondary progressive MS, and these differences might affect segmentation quality. The evolution of the disease also includes other factors that can complicate the segmentation of lesions such as atrophy and dirty white matter that should be taken into account. Thus it is easier to segment single new punctate lesions in early relapsing remitting MS compared to the segmentation of enlarging lesions in regions of dirty white matter in secondary progressive MS. In addition, the interpretation of sensitivity and specificity is very different when the total lesion load is very low or very high. Of the 47 papers that used clinical MRI data for validation, 28 papers (60%) contained no information regarding the form (relapsing-remitting, secondary progressive, etc.), duration, or severity of the disease, or the sex or age of the patients.

Another important issue to consider is lesion load. The number and volume of lesions can vary greatly across patients, and an algorithm's performance can differ depending on whether lesion load is low or high. In addition, some metrics are sensitive to the surface-to-volume ratio of the structures segmented; therefore, it may be important to take lesion size into account to correctly understand the results. Only 14 papers included the lesion load of the patients, and the range varied greatly among the papers; the lowest lesion loads ranged from approximately 1 cm^3 (Alfano et al., 2000; García-Lorenzo et al., 2009; Harmouche et al., 2006) to 8 cm^3 (Shiee et al., 2009), and the highest, from 20.0 cm^3 (Guttmann et al., 1999) to 130 cm^3 (Alfano et al., 2000). Some authors divided the patients according to lesion load (Khayati et al., 2008a; Sajja et al., 2006), although no consensus exists on how to achieve this division.

One of the main limitations on comparing segmentation methods is evaluating the differences between the MR protocols employed. A detailed description of the MRI protocol is important to understand the differences between the approaches and make it possible to choose the method that best matches the protocol. Only 19 papers (40%) provided a description of the MRI acquisition protocol.

Finally, and importantly, clinical trials used for drug development usually include data from many centers, acquired using different scanners. Standardization of MRI acquisition sequences is generally attempted, but, even with standardization, differences between scanners will always exist. Automatic segmentation methods should be able to address this residual variability to enable their use in clinical trials. In our review, only 13 papers used images from more than one scanner. Of these, 11 papers used data from two sites, provided by the MS Lesion Segmentation Challenge described in section 5.3, and only two papers used multisite data.

5.1.2.2. The ground truth

A very important limitation on the validation of segmentation methods when using clinical data is the lack of a ground truth for the WML. Although an estimate of ground truth is not necessary to measure the reproducibility of a method, it is required to measure the accuracy of the segmentation. Although manual segmentation of lesions is often considered the gold standard validation technique for the segmentation of WML, manual methods can suffer from large variability. In practice, semiautomatic techniques are employed in the segmentation of WML, as they reduce both the variability of the segmentation and the processing time (Grimaud et al., 1996; Udupa et al., 1997).

Of the publications reviewed, 20 papers compared the results of their automatic segmentation with the segmentation of one expert using manual or semiautomatic segmentation, which was considered to be the gold standard. Although this evaluation provides an indication of the similarity of the segmentations, the significance of the results obtained with this evaluation is limited by the well-known intra- and interexpert variability that plagues manual segmentation (Grimaud et al., 1996). When results are compared against a single rater, it is not possible to distinguish the dissimilarities between manual and automatic segmentation methods caused by errors in the algorithm from those caused by variability in the manual segmentation.

For this reason, some authors have compared their methods with segmentations by more than one expert. In the MS Lesion Segmentation Challenge (Styner et al., 2008), automatic segmentations were compared independently with two expert segmentations. Then, the results of those comparisons were themselves compared to evaluate if there was a bias towards one of the experts.

As no real ground truth is available, the objective of some methods was to show that their results lie within the variability of the experts by using total lesion load (TLL) measures (Wei et al., 2002; Zijdenbos et al., 2002). Other approaches considered the manual segmentation imperfect and merged several manual segmentations to create a silver standard and reduce the variability of each expert individually (García-Lorenzo et al., 2009; Harmouche et al., 2006; Subbanna et al., 2009). This silver standard has been created both by using a consensus approach, whereby each voxel is considered lesion only if most experts considered it lesion, and by using the STAPLE algorithm (Warfield et al., 2004), whereby sensitivity and specificity are computed simultaneously with the silver standard. The objective is to understand whether the results of the automatic method vary at a magnitude similar to the variability in the results of the experts.

5.2. Validation measures

Two main aspects characterize the validation of a segmentation method: accuracy and reproducibility.

5.2.1. Accuracy

Accuracy refers to the degree of closeness of the estimated measure to the true measure. Figure 5 shows a diagram of the segmentation comparison. In binary segmentation, four scenarios are possible: true positives (TP) and true negatives (TN), wherein the segmentation is correct, and false positives (FP) and false negatives (FN), wherein there is disagreement between the two segmentations.

Several measures have been employed to evaluate the accuracy of automatic segmentation methods. In general, no single measure is capable of giving all the information desired regarding the quality of the segmentation.

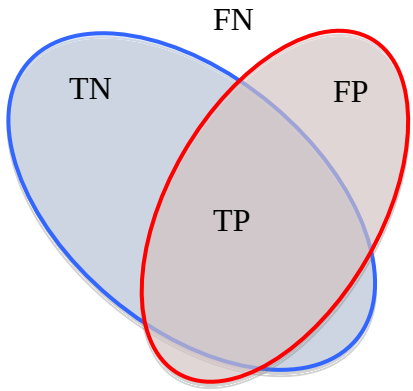

Figure 5. Diagram of the comparison between to segmentation. If we consider the blue region as our reference and the red region the segmentation, the intersection of the two regions divide the space into four regions: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

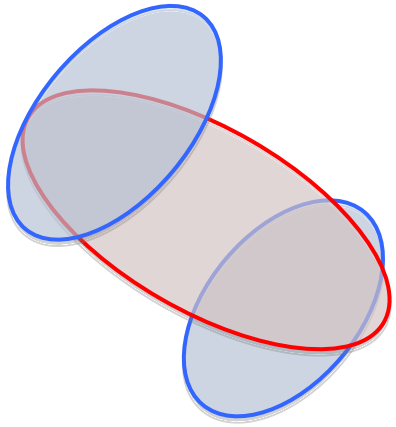

Figure 6. Example of problem in a lesion-based metric. When two manually segmented lesions (blue) overlap with only one large automatic lesion (red). Should we consider that the two lesions were detected, or should we consider that one lesion was detected but the other one was missed?

- Lesion load:

As the TLL is often used as a biomarker in clinical trials, the total volume of lesion voxels has been employed to evaluate the validity of the segmentation using different metrics, for example, the volume difference between the automatic and the expert's manual segmentations (Liu et al., 2009; Shahar and Greenspan, 2004; Styner et al., 2008; Van Leemput et al., 2001). To ensure that no bias is introduced when segmenting patients with different lesion load, authors have employed Pearson's correlation (Harmouche et al., 2006; Khayati et al., 2008a; Shiee et al., 2009; Wels et al., 2008; Wu et al., 2006; Zijdenbos et al., 2002) and Bland–Altman plots (Datta et al., 2007; Datta et al., 2006; He and Narayana, 2002).

To compare the automatic segmentation with several experts, some authors used the intraclass correlation (ICC) (Garcia-Lorenzo et al., 2011; Wu et al., 2006), while other authors used Z-scores (Wei et al., 2002). In the latter case, the Z-scores for the automatic segmentation were computed using the mean of and variance between experts, which give information about the bias of the automatic segmentation: toward either over-segmentation (positive Z-scores) or under-segmentation (negative Z-scores). Another option, considering that the error follows a Gaussian distribution, was to employ an analysis of variance to test if the TLL of the automatic method was significantly different from that computed by the experts (Zijdenbos et al., 2002).

The main limitation of using lesion volume as a measure is that there is no information regarding the overlap of the segmentation. In the extreme case, the automatic segmentation method could obtain the same TLL as the true segmentation and have no voxels in common.

- Overlap measures:

Many overlap measures have been used in the literature, but the most widely employed with respect to segmentation is the similarity index or Dice similarity coefficient (DSC) (Dice, 1945):

$$DSC = \frac{2 \times TP}{FP + FN + 2 \times TP}$$

The value of the index varies between 0 and 1 (perfect segmentation), with 0.7 normally considered a good segmentation. Derived from Cohen's kappa under the assumption that $TN >> \mathrm{TP}$ (Zijdenbos et al., 1994), this metric, unfortunately, does not give information about whether lesions are being under-segmented and is also sensitive to the TLL and size distribution of lesions.

Other authors employed sensitivity (Sens), specificity (Spec) and accuracy (Acc) (Akselrod-Ballin et al., 2009; Alfano et al., 2000; Hadjiprocopis and Tofts, 2003; Liu et al., 2009; Sajja et al., 2006; Tomas-Fernandez and Warfield, 2011; Wels et al., 2008; Wu et al., 2006):

$$Sens = \frac{TP}{TP + FN}$$

$$Spec = \frac{TN}{FP + TN}$$

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

These measures should be considered carefully, as TP (lesions) are much smaller than TN (normal appearing brain tissues). An automatic segmentation in which the TLL is 10 times greater than the ground truth can still result in Sens = 100%, Spec = 99.3%, and Acc = 99.3% when the TLL is around 1 cm^3 and brain size is 1,500 cm^3.

Other methods provide fuzzy (or probabilistic) segmentations (Anbeek et al., 2008; Datta et al., 2007; Datta et al., 2006); therefore, fuzzy metrics were adapted to measure the accuracy of the segmentation, such as the probabilistic similarity index (PSI)

$$PSI = \frac{2 \cdot \sum_x P_s(x | R(x) == 1)}{\sum_x R(x) + \sum_x P_s(x)}$$

where P_s is the probabilistic segmentation, and R is the binary reference.

- Lesion-based measures:

Number of lesions is also often used in diagnosis and as an output measure in clinical trials. Due to the problems faced in obtaining the true boundaries of WML, some authors have proposed to validate the automatic detection of lesions by counting them (Goldberg-Zimring et al., 1998; Sajja et al., 2006; Styner et al., 2008).

First, the number of lesions given by the automatic segmentation was compared with the number of lesions estimated manually (Goldberg-Zimring et al., 1998). Other authors computed the sensitivity and specificity, but used the lesion counts instead of the voxel labels (Sajja et al., 2006; Styner et al., 2008; Yamamoto et al., 2010).

The advantage of these lesion-based measures is that they are less sensitive to the variability of the manual segmentation. However, they do not give information about the accuracy of the boundary of the

lesion, or its over- or under-segmentation. In addition, the definition of true positives and false positives can be ambiguous, for example, when a lesion identified manually as a single connected lesion but overlaps with two distinct automatically segmented lesions, or vice versa. Two definitions are possible: the two automatic lesions are true positives, or only one of the lesions is considered to be a true positive and the other a false positive. (Figure 6).

5.2.2. Reproducibility

Reproducibility measures the degree of agreement between several identical experiments. Reproducibility is crucial for longitudinal trials to ensure that differences in segmentations obtained over time result from changes in the pathology and not from the variability of the automatic method.

To test reproducibility, MS patients may undergo MR imaging several times within a short period ranging from half an hour to several days in *scan-reposition-rescan* experiments. As the images are obtained within a short space of time, it is assumed that the disease has not evolved during this period. Two sources of variability are possible in the segmentation: variability of the method and interscan variability.

- **Variability of the method** refers to the differences in the segmentation when the same method is applied to the same acquisition. In manual and semiautomatic methods, this variability is mainly due to the human interaction and is termed *intra-* and *inter-rater* variability. Automatic methods that employ random processes (e.g., genetic algorithms, multiple random initializations in k-means or EM algorithms) can obtain different segmentation results for the same image. Normally, the results are similar, but there can be slight differences due to the random process during the segmentation. By contrast, deterministic segmentation methods always produce the same segmentation for the same image because there is no methodological variability. This is an important advantage of the automatic deterministic methods.

- **Interscan variability** refers to the differences in the segmentation when the same method is applied to two different scans of the same patient, where the two scans have been acquired in a sufficiently short period so that there is no real change in the disease. In practice, patients are required to exit the scanner after the first acquisition and then re-enter to undergo a second, independent acquisition. Interscan variability is affected by the acquisition protocol employed (e.g., resolution, slice thickness, etc.) and can be used to estimate the minimum change in lesion load that can be measured in longitudinal exams for a given acquisition protocol.

In our review, only four methods (Alfano et al., 2000; Erickson and Avula, 1998; Guttmann et al., 1999; Wei et al., 2002) employed scan-rescan images to evaluate the reproducibility of the segmentation. Reproducibility was measured using the coefficient of variation (CV) on the TLL

$$CV = \frac{\sigma}{\mu}$$

where $\boldsymbol{\mu}$ is the mean $\boldsymbol{\sigma}$ and is the standard deviation of the different measures of TLL. The smaller the value of CV, the more reproducible the method is. The CV on the TLL has the same limitations as mentioned for TLL as accuracy measure as there is no overlap information. For a method, CV can be low but segmentations can have low overlap. New metrics need to be defined to integrate the overlap in reproducibility metrics.

From the practical point of view, measuring reproducibility has the advantage of not requiring a ground truth; one only needs to show that the method obtains a similar result when applied to similar input , for example, from scan-rescan images. Thus, reproducibility is a necessary but not sufficient part of validation. One still needs to show that the method is accurate and sensitive to changes in input data.

Measuring accuracy requires an independent estimate of the ground truth, an often difficult task when using clinical data.

5.3. Example validation framework: MS Lesion Segmentation Challenge

At MICCAI 2008, a workshop was organized with the objective of comparing different segmentation algorithms in the form of a competition (Van Ginneken et al., 2007). One of the segmentation challenges was the segmentation of WML in MS (Styner et al., 2008), and nine methods were compared (Anbeek et al., 2008; Bricq et al., 2008b; García-Lorenzo et al., 2008b; Kroon et al., 2008; Morra et al., 2008; Prastawa and Gerig, 2008; Scully et al., 2008; Shiee et al., 2008; Souplet et al., 2008).

The validation database consisted of T1w, T2w, FLAIR, and diffusion-weighted images from two different sites. The acquisition protocol was not described, and the images were registered and upsampled to an isotropic 0.5 mm by the organizers. No information about the clinical characteristics of the patients was available. The database was divided into three sets:

- **Training set:** Participants were given 20 images along with their manual segmentations to adapt each algorithm to the MR protocol and the experts' definition of lesion. Supervised methods used this dataset to train their methods.
- **Off-site test set:** Participants were given 25 images, without the manual segmentations, to be processed in their own laboratories. The automatic segmentation results had to be sent to the organizers for evaluation.
- **On-site test set:** Participants were given 7 images during the workshop to be processed within a limited amount of time.

Four metrics were employed in the evaluation of the competing methods:

- **Relative volume difference:** Difference of total volumes divided by the reference volume;
- **Mean surface distance:** Distance between borders of the reference volume and the segmentation;
- **True positives:** Number of lesions in the segmentation that correspond to a lesion in the reference volume divided by total number of lesions in the reference volume; and
- **False negatives:** Number of lesions in the segmentation that do not correspond to a lesion in the reference volume divided by total number of lesions in the segmentation.

To compare the different metrics, the organizers normalized metric values between 0 and 100, where 100 was a perfect score and 90 was the *typical score* of an independent observer (Van Ginneken et al., 2007). Setting the experts' values to 90 allows the automatic methods to obtain better scores than the agreement between experts. The typical score was measured by comparing the manual segmentation of two experts, and scaling was done linearly using these two points (negative values were set to 0). The typical scores obtained for the two experts were:

- Relative volume difference: 68%
- Mean surface distance: 4.85 mm
- Overlap error: 75%
- True positives: 68%
- False negatives: 32%.

Despite the relatively low agreement between the two raters, it is important to underline the value of the MS Lesion Segmentation Challenge database. It represents the first publicly available clinical database for validation of WML lesions. It includes 52 sets of images from two different centers and

provides two expert manual segmentations in order to account for manual segmentation variability. Compared with many other papers, the MICCAI MS Lesion Segmentation Challenge contains a larger number of subjects. Finally, the organizers provide for an objective comparison between methods since the manual labels from the testing set are not made public. Segmentations of the testing data have to be sent to the organizers, and segmentation quality metrics are sent back to the users. This prevents users training on the testing data and facilitates comparison between methods.

It is important that future papers using this database report their results for **both** the training and testing datasets, as this is the only way to make a fair comparison with previously published papers. Reporting **only** on the testing data enables potential over learning of the training database, while results on the testing database are objective and perhaps better reflect real-world performance.

Three limitations of the MS Lesion Segmentation Challenge database must be discussed: the quality of the images, the preprocessing steps, and the manual segmentations performed by the experts. First, multiple artefacts were found in several images due to patient movement during acquisition (García-Lorenzo et al., 2008b). While these artefacts represent the reality of clinical data, they can greatly affect the quality of the automatic segmentations. A quality control step should have removed any poor quality images, a common practice in clinical studies. Another option would be to characterize the images according to their quality to assess the robustness of the automatic methods. Second, the segmentation is usually performed after preprocessing steps, such as inhomogeneity correction and denoising. The MS Lesion Segmentation Challenge proposed data that was already registered and upsampled, which reduced the effectiveness of denoising methods, as the assumption of voxel-to-voxel independence of the noise was no longer valid. Such preprocessing steps could reduce the performance of algorithms where denoising is necessary for the segmentation. Third, the manual segmentations were performed by two experts from different centers and showed great variability. For example, the two experts only agreed on 68% of the lesions, which puts a low ceiling on the maximum possible accuracy of methods, as they cannot agree with both raters simultaneously.

6. Discussion

The automatic segmentation of MS lesions has been an active research topic for more than 15 years. Although many improvements have been achieved, many challenges remain.

Definition of WML

"WML are areas that are hyperintense with respect to normal-appearing white matter on T2w or FLAIR images that are not due to normal structures." Although this statement is widely accepted, its interpretation produces high variability in the value of segmented MS lesions, mainly due to variability in deciding how much of an elevated intensity is required for a hyperintensity to be called a lesion. The definition of WML depends on the MRI sequence employed and even the contrast of the screen used for the segmentation. Variations of 40% (Zijdenbos et al., 2002) and 68% (Styner et al., 2008) were reported across different centers, demonstrating a large variability in the definition of WML. Because automatic methods rely (explicitly or not) on the definition of WML, a more precise definition of lesions would provide a better starting point for automatic methods and thus result in a better segmentation.

The diagnosis of MS (McDonald et al., 2001) is defined by a panel of experts and updated every few years (Polman et al., 2011; Polman et al., 2005); the definition of WML should be revised in a similar way to give a more accurate and precise definition of MS lesions than the qualitative visual rule above. The definition should also include recommendations for a standard MRI protocol for MS that is more specific that the MRI guidelines proposed by Simon et al. (Simon et al., 2006) to simplify the segmentation of lesions between centers.

There have been efforts in order to reduce the inter-rater variability in the detection of lesions. A reduction of these variabilities was shown after more precise guidelines were given for the different types of lesions (Barkhof et al., 1997; Filippi and Rocca, 2007). Similar efforts would also be interesting for the segmentation of lesions.

Segmentation methods

We proposed a classification of methods and described the methodological differences between them. Instead of considering each method independently, we looked for similarities between techniques, thus yielding a better overview of the methodologies and strategies employed.

When methods are compared, the conclusions remain limited because the methods are considered black boxes. For example, a supervised method (SM) can be roughly defined by its feature space (FS) and classification algorithm (CA). When two methods, SM_1 and SM_2, are compared, to understand the results, a mixed method (FS_1+CA_2 and FS_2+CA_1) should also be evaluated to determine the improvement provided by each part of each algorithm.

In our review, we decided not to compare numerical results because each paper used very different databases and metrics. To address this issue, the MICCAI MS Lesion Segmentation Challenge offered the first objective comparison of methods where 9 methods were employed in the *same* multicenter database.. The winner was (Souplet et al., 2008) and 4 out 5 best algorithms included unsupervised methods. New methods can still be evaluated using the same images and the results are publicly available (http://www.ia.unc.edu/MSseg/).

From our review, we can extract several conclusions from the methodological point of view:

- **A robust, accurate, fully-automated lesion segmentation method suitable for use in clinical trials is still not available:** While there is abundant literature on the topic of fully-automated lesion segmentation and some important advances have been made, it is clear that the problem is far from solved, and more work is needed in this area.
- **Multimodal information is necessary:** Although some lesions are obvious in only one sequence (especially in FLAIR), the lesion should be confirmed in other sequences (T2-w or PD) to avoid false positives. T1w images, although not necessary in manual segmentation, provide very good tissue contrast for automatic segmentation techniques and improve the segmentation.
- **Spatial information is necessary:** Intensity information is not sufficient for a good segmentation, as other tissues have similar intensities, and noise can reduce the performance of the algorithm. Spatial information needs to be included in the segmentation algorithms at two different spatial levels: the local neighborhood level and the anatomical level. The former refers to the information provided by neighboring voxels to reduce the impact of noise and to improve coherence of the results (i.e. MRF models, Graph Cut or kernel features). The latter refers to the particular anatomical location within the brain could include information from specific anatomical templates for MS patients, lesion probability templates or topological atlases (Shiee et al., 2009).
- **Unsupervised vs. Supervised**: Unsupervised methods have the advantage of not requiring a ground truth and makes them more easily adaptable to new acquisition protocols (as in the MS Lesion Segmentation Challenge). Still, once the supervised methods are well trained results can be comparable (Geremia et al., 2011).
- **Supervised methods**: The newer algorithms selects at the same time the optimal number of features between thousands of possible features. The challenge remains in the creation of the database were all the possible lesions are integrated and in the preprocessing of the data to reduce the variability of the images between centers and scanners.

- **Unsupervised methods**: The majority of methods are based in global clustering techniques such us Fuzzy-C means or a Gaussian mixture model. The intensity of the tissues and lesions vary across the image. New methods should take into account these local variations in order to improve the sensibility of the methods in some parts of the brain.
- **Availability of methods:** The majority of methods are not publicly available making their use by other centers hard, and making their comparison and improvement difficult. The methods that are available can be found in Table 3.

Table 3. Freely available automatic segmentation methods

Paper	Website
(Van Leemput et al., 1999)	http://mirc.uzleuven.be/MedicalImageComputing/downloads/ems.php
(Souplet et al., 2008)	http://med.inria.fr
(Garcia-Lorenzo et al., 2011)	https://www.irisa.fr/visages/benchmarks/
(Shiee et al., 2008)	http://www.nitrc.org/projects/toads-cruise/

Other remaining challenges that the literature has not yet addressed include:

- **Partial volumes:** The border of lesions is fuzzy in part because of the limited resolution of MR images that causes the partial volume effects (Choi et al., 1991). Although partial volume effects have mainly been studied for healthy tissues (Santago and Gage, 1993; Shattuck et al., 2001) or applied in MS but only for the normal appearing healthy tissues (Dugas-Phocion et al., 2004). Accounting for partial volume effects should also improve the reproducibility of WML segmentation. Of course, the validation of such techniques will become even more complicated than the evaluation of binary lesions.
- **Multicenter datasets:** Images from different scanners have different contrasts or intensities, even when same protocol is employed. Methods should be designed specifically to deal with this variation without biasing the posterior clinical studies.
- **Spinal cord imaging:** Spinal cord lesions correlate with motor disability. Automatic methods are almost always focused on the brain and, to our knowledge, little effort has been performed to segment lesions in the spinal cord.
- **Diffuse disease:** Automatic methods have concentrated on the focal lesions caused by the disease, but in some cases it is impossible to find the border between lesions and the neighboring diffuse dirty white matter. No method has attempted to address these issues with the dirty white matter yet.

Validation and results

The final goal of an automatic segmentation method is its use in a clinical setting; therefore, a complete validation of the method in real conditions is mandatory. In our initial search, we rejected 20 papers (more than 20%) because no quantitative validation had been performed of the method proposed.

Multiple sclerosis is a heterogeneous disease, and this fact needs to be taken into account in the validation of segmentation methods. Many papers included no description of the clinical information of the patients, nor any details of the MRI protocol employed in the validation. Including information about patient age, clinical type, and MRI protocol will be helpful to determine the applicability of methods and simplify their comparison.

In clinical images, no ground truth is available; therefore, simple frameworks are insufficient to provide a good validation. The "best" approximation to a ground truth is manual segmentation, but, as mentioned previously, there is a great variability among experts. As a result, the objective of the validation cannot be to obtain the "exact" same segmentation as an expert (e.g., DSC = 1), as we know the manual segmentation includes errors. The validation framework should also consider that the manual segmentation is imperfect. In this sense, the use of several manual segmentations can improve the validation by the creation of an improved silver standard. STAPLE (Commowick and Warfield, 2009; Warfield et al., 2004) provides a framework to compare the accuracy of multiple experts and create a silver standard. While this has been a major contribution, there still is work needed to define the number of experts required and how to integrate the automatic methods and manual experts in the algorithm. As we mentioned, new metrics are also required to measure the reproducibility beyond that of the total lesion load.

The majority of the papers published on WML segmentation only use one metric to evaluate the accuracy of their method, in particular the DSC. This has the advantage of simplifying the results to one number, but reduces the information we can extract from the evaluation. For example, the DSC does not give information about over- or under-segmentation, nor does it provide any notion about the consistency across disease severity, as does the volume correlation. In addition, lesions in some anatomical regions are more complicated to segment than others, an issue not addressed by these global measures. Multiple complementary metrics are needed, and reporting the location of errors would provide a better understanding of the performance of the proposed methods.

Although accuracy is the most widely performed type of validation, clinical studies are very concerned with the precision/reproducibility of measures over time, and this is rarely evaluated in the literature. The validation of reproducibility is partially dealt with using a scan-reposition-rescan protocol. Although this has the advantage of not requiring a ground truth, it still does not deal with real-world variability due to hardware drift over time. The three papers that evaluated reproducibility reported it using only volumetric measures. New metrics should be developed to include the overlap into the reproducibility measures.

In order to provide a good validation of methods, we propose four validation steps:

1) BrainWeb:

This step should be considered optional, but every method that employs BrainWeb should use the same validation framework: use the whole brain (not selected slices) and the three different phantoms, and vary the levels of noise and inhomogeneity using the predefined values on the website.

Using BrainWeb should provide a proof of concept for new methods and evaluate the robustness to noise and IIH. BrainWeb is a good database to compare methods–although clinical validation is more important; hence, validation with BrainWeb is a necessary, but not sufficient, condition for validation.

2) Small group with several manual segmentations (10–20 patients):

The main objective of using a small group is to measure the accuracy of the segmentation method. Patients should have a wide range of TLL to measure the different stages of the disease. Several manual segmentations should be used for comparison to provide an accurate silver standard. Several metrics should be used to provide a complete validation and should include ICC, Hausdorff distance and lesion-based metrics such as rates of false positives and false negatives.

3) Small group with scan-reposition-rescan acquisitions (5–10 patients):

The objective is to measure the reproducibility of the method in the presence of data where no change is expected. No ground truth is required, but metrics other than the CV should be employed to evaluate the consistency of lesion label overlap. The reproducibility is very important complementary information that will give an idea of how many patients are required in large clinical trials.

4) Large multicenter validation (50–100+ patients):

The objective is to show the robustness of the method within a large database containing data from multiple scanners. Semiautomatic segmentation (or automatic segmentation with manual correction) should provide a rough estimate of the ground truth. Robustness should be evaluated both in terms of volume and number of lesions.

The creation of these databases cannot be the effort of only one center if we want to create an agreement in the validation. An international initiative should be proposed to study the Multiple Sclerosis in a similar way to the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008). The main objective of this initiative should not be, of course, the validation of automatic methods, but the standardization of scanners and acquisition and the creation of a public database will provide the necessary environment to provide a better and complete framework of validation.

Conclusions

We presented a comprehensive review of the automatic methods for MS lesion segmentation. We described the complexity of the lesion segmentation problem and identified solutions that have been proposed in the literature. In addition to describing the techniques, we focused on the strengths and weaknesses of the validation methods used to characterize the published methods.

Although many papers have been published on the subject, automatic segmentation of lesions in MS remains an open problem. Many methods provide limited solutions where they deal with only one type of MR protocol or identify only one type of MS lesion, and they rarely address the complexity of the disease. Nevertheless, many advances have been made over the years, and many methods have demonstrated promising results with MRI data from small groups of patients. The challenge remains to provide segmentation techniques that can work in all cases regardless of the type of MS, duration of the disease, or MRI protocol, and this within a comprehensive, standardized validation framework.

Reference

Ahmed, M.N., Yamany, S.M., Mohamed, N., Farag, A.A., Moriarty, T., 2002. A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data. Medical Imaging, IEEE Transactions on 21, 193--199.

Aït-Ali, L.S., Prima, S., Hellier, P., Carsin, B., Edan, G., Barillot, C., 2005. STREM: A Robust Multidimensional Parametric Method to Segment MS Lesions in MRI, In: Duncan, J.S., Gerig, G. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005. Springer Berlin / Heidelberg, pp. 409-416.

Akselrod-Ballin, A., Galun, M., Basri, R., Brandt, A., Gomori, M.J., Filippi, M., Valsasina, P., 2006. An Integrated Segmentation and Classification Approach Applied to Multiple Sclerosis Analysis, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1122-1129.

Akselrod-Ballin, A., Galun, M., Gomori, J.M., Filippi, M., Valsasina, P., Basri, R., Brandt, A., 2009. Automatic Segmentation and Classification of Multiple Sclerosis in Multichannel MRI. Biomedical Engineering, IEEE Transactions on 56, 2461-2469.

Alfano, B., Brunetti, A., Larobina, M., Quarantelli, M., Tedeschi, E., Ciarmiello, A., Covelli, E.M., Salvatore, M., 2000. Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. Journal of Magnetic Resonance Imaging 12, 799-807.

Anbeek, P., Vincken, K.L., Viergever, M.A., 2008. Automated MS-Lesion Segmentation by K-Nearest Neighbor Classification. MS Lesion Segmentation (MICCAI 2008 Workshop).

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839-851.

Ashton, E.A., Takahashi, C., Berg, M.J., Goodman, A., Totterman, S., Ekholm, S., 2003. Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI. Journal of Magnetic Resonance Imaging 17, 300-308.

Aubert-Broche, B., Evans, A.C., Collins, L., 2006. A new improved version of the realistic digital brain phantom. Neuroimage 32, 138-145.

Bakshi, R., Thompson, A.J., Rocca, M.A., Pelletier, D., Dousset, V., Barkhof, F., Inglese, M., Guttmann, C.R., Horsfield, M.A., Filippi, M., 2008. MRI in multiple sclerosis: current status and future prospects. Lancet Neurol 7, 615-625.

Bao, P., Zhang, L., 2003. Noise reduction for magnetic resonance images via adaptive multiscale products thresholding. IEEE Trans Med Imaging 22, 1089-1099.

Barkhof, F., Filippi, M., van Waesberghe, J.H., Molyneux, P., Rovaris, M., Lycklama a Nijeholt, G., Tubridy, N., Miller, D.H., Yousry, T.A., Radue, E.W., Ader, H.J., 1997. Improving interobserver variation in reporting gadolinium-enhanced MRI lesions in multiple sclerosis. Neurology 49, 1682-1688.

Bezdek, J., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers.

Bijar, A., Khanloo, M.M., Peñalver Benavent, A., Khayati, R., 2011. Segmentation of MS lesions using entropy-based EM algorithm and Markov random fields. Journal of Biomedical Science and Engineering 4, 552-561.

Boudraa, A.-O., Dehak, S.M.R.d., Zhu, Y.-M., Pachai, C., Bao, Y.-G., Grimaud, J.r.m., 2000. Automated segmentation of multiple sclerosis lesions in multispectral MR imaging using fuzzy clustering. Computers in Biology and Medicine 30, 23-40.

Bricq, S., Collet, C., Armspach, J.P., 2008a. Lesions detection on 3D brain MRI using trimmmed likelihood estimator and probabilistic atlas, Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, pp. 93-96.

Bricq, S., Collet, C., Armspach., J., 2008b. MS Lesion Segmentation based on Hidden Markov Chains. MS Lesion Segmentation (MICCAI 2008 Workshop).

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms, Proceedings of the 23rd international conference on Machine learning. ACM, Pittsburgh, Pennsylvania, pp. 161-168.

Cerasa, A., Bilotta, E., Augimeri, A., Cherubini, A., Pantano, P., Zito, G., Lanza, P., Valentino, P., Gioia, M.C., Quattrone, A., 2012. A Cellular Neural Network methodology for the automated segmentation of multiple sclerosis lesions. J Neurosci Methods 203, 193-199.

Choi, H.S., Haynor, D.R., Kim, Y., 1991. Partial volume tissue classification of multichannel magnetic resonance images-a mixel model. Medical Imaging, IEEE Transactions on 10, 395-407.

Ciccarelli, O., Brex, P.A., Thompson, A.J., Miller, D.H., 2002. Disability and lesion load in MS: a reassessment with MS functional composite score and 3D fast FLAIR. J Neurol 249, 18-24.

Clarke, L.P., Velthuizen, R.P., Camacho, M.A., Heine, J.J., Vaidyanathan, M., Hall, L.O., Thatcher, R.W., Silbiger, M.L., 1995. MRI segmentation: Methods and applications. Magnetic Resonance Imaging 13, 343-368.

Cocosco, C.A., Kwan, K., Evans, R., 1997. BrainWeb: online interface to a 3-D MRI simulated brain database. Neuroimage 5, part 2/4, S425.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J Comput Assist Tomogr 18, 192-205.

Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., Evans, A.C., 1998. Design and construction of a realistic digital brain phantom. Medical Imaging, IEEE Transactions on 17, 463-468.

Comi, G., Filippi, M., Wolinsky, J.S., 2001. European/Canadian multicenter, double-blind, randomized, placebo-controlled study of the effects of glatiramer acetate on magnetic resonance imaging--measured disease activity and burden in patients with relapsing multiple sclerosis. European/Canadian Glatiramer Acetate Study Group. Ann Neurol 49, 290-297.

Commowick, O., Warfield, S.K., 2009. A continuous STAPLE for scalar, vector, and tensor images: an application to DTI analysis. IEEE Trans Med Imaging 28, 838-846.

Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An Optimized Blockwise Nonlocal Means Denoising Filter for 3-D Magnetic Resonance Images. Medical Imaging, IEEE Transactions on 27, 425--441.

Datta, S., Sajja, B., He, R., Gupta, R., Wolinsky, J., Narayana, P., 2007. Segmentation of gadolinium-enhanced lesions on MRI in multiple sclerosis. Journal of Magnetic Resonance Imaging 25, 932-937.

Datta, S., Sajja, B.R., He, R., Wolinsky, J.S., Gupta, R.K., Narayana, P.A., 2006. Segmentation and quantification of black holes in multiple sclerosis. Neuroimage 29, 467-474.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1-38.

Desolneux, A., Moisan, L., More, J.-M., 2003. A grouping principle and four applications. Pattern Analysis and Machine Intelligence, IEEE Transactions on 25, 508--513.

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26, 297-302.

Dietrich, O., Raya, J.G., Reeder, S.B., Ingrisch, M., Reiser, M.F., Schoenberg, S.O., 2008. Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. Magnetic Resonance Imaging 26, 754--762.

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification, 2nd ed. Wiley.

Dugas-Phocion, G., González Ballester, M.Á., Malandain, G., Lebrun, C., Ayache, N., 2004. Improved EM-Based Tissue Segmentation and Partial Volume Effect Quantification in Multi-Sequence Brain MRI, International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Saint-Malo, France, pp. 26--33.

Erickson, B., Avula, R., 1998. An algorithm for automatic segmentation and classification of magnetic resonance brain images. Journal of Digital Imaging 11, 74-82.

Fazekas, F., Barkhof, F., Filippi, M., Grossman, R.I., Li, D.K., McDonald, W.I., McFarland, H.F., Paty, D.W., Simon, J.H., Wolinsky, J.S., Miller, D.H., 1999. The contribution of magnetic resonance imaging to the diagnosis of multiple sclerosis. Neurology 53, 448-456.

Fennema-Notestine, C., Ozyurt, I.B., Clark, C.P., Morris, S., Bischoff-Grethe, A., Bondi, M.W., Jernigan, T.L., Fischl, B., Segonne, F., Shattuck, D.W., Leahy, R.M., Rex, D.E., Toga, A.W., Zou, K.H., Brown, G.G., 2006. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Human Brain Mapping 27, 99--113.

Filippi, M., Horsfield, M.A., Bressi, S., Martinelli, V., Baratti, C., Reganati, P., Campi, A., Miller, D.H., Comi, G., 1995a. Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis. A comparison of techniques. Brain 118 (Pt 6), 1593-1600.

Filippi, M., Horsfield, M.A., Campi, A., Mammi, S., Pereira, C., Comi, G., 1995b. Resolution-dependent estimates of lesion volumes in magnetic resonance imaging studies of the brain in multiple sclerosis. Ann Neurol 38, 749-754.

Filippi, M., Rocca, M.A., 2007. Conventional MRI in multiple sclerosis. J Neuroimaging 17 Suppl 1, 3S-9S.

Forbes, F., Doyle, S., García-Lorenzo, D., Barillot, C., Dojat, M., 2010. Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation, Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on Rotterdam, Netherlands pp. 69 - 72

Freifeld, O., Greenspan, H., Goldberger, J., 2009. Multiple sclerosis lesion detection using constrained GMM and curve evolution. Int J Biomed Imaging 2009, 715124.

Freund, Y., Schapire, R., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting, In: Vitányi, P. (Ed.), Computational Learning Theory. Springer Berlin / Heidelberg, pp. 23-37.

García-Lorenzo, D., Lecoeur, J., Arnold, D.L., Collins, D.L., Barillot, C., 2009. Multiple Sclerosis Lesion Segmentation Using an Automatic Multimodal Graph Cuts, Medical Image Computing and Computer-Assisted Intervention, pp. 584-591.

Garcia-Lorenzo, D., Prima, S., Arnold, D.L., Collins, D.L., Barillot, C., 2011. Trimmed-likelihood estimation for focal lesions and tissue segmentation in multisequence MRI for multiple sclerosis. IEEE Trans Med Imaging 30, 1455-1467.

García-Lorenzo, D., Prima, S., Collins, D.L., Arnold, D., Morrissey, S.P., Barillot, C., 2008a. Combining Robust Expectation Maximization and Mean Shift algorithms for Multiple Sclerosis Brain Segmentation, MICCAI workshop on ¨Medical Image Analysis on Multiple Sclerosis (MIAMS)", pp. 82-91.

García-Lorenzo, D., Prima, S., Morrissey, S.P., Barillot, C., 2008b. A robust Expectation-Maximization algorithm for Multiple Sclerosis lesion segmentation, MS Lesion Segmentation (MICCAI 2008 Workshop), New York,USA.

García-Lorenzo, D., Prima, S., Parkes, L., Ferré, J.C., Morrissey, S.P., Barillot, C., 2008c. The impact of processing workflow in performance of automatic white matter lesion segmentation in Multiple Sclerosis, MICCAI Workshop in Medical Image Analysis for Multiple Sclerosis (MIAMS), New York,USA, pp. 104-112.

Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57, 378-390.

Gerig, G., Kubler, O., Kikinis, R., Jolesz, F.A., 1992. Nonlinear anisotropic filtering of MRI data. IEEE Trans Med Imaging 11, 221-232.

Geurts, J.J.G., Bö, L., Pouwels, P.J.W., Castelijns, J.A., Polman, C.H., Barkhof, F., 2005. Cortical Lesions in Multiple Sclerosis: Combined Postmortem MR Imaging and Histopathology. American Journal of Neuroradiology 26, 572-577.

Goldberg-Zimring, D., Achiron, A., Miron, S., Faibel, M., Azhari, H., 1998. Automated detection and characterization of multiple sclerosis lesions in brain MR images. Magn Reson Imaging 16, 311-318.

Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G.J., Plummer, D.L., Tofts, P.S., McDonald, W.I., Miller, D.H., 1996. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magnetic Resonance Imaging 14, 495-505.

Guttmann, C.R., Kikinis, R., Anderson, M.C., Jakab, M., Warfield, S.K., Killiany, R.J., Weiner, H.L., Jolesz, F.A., 1999. Quantitative follow-up of patients with multiple sclerosis using MRI: reproducibility. Journal of Magnetic Resonance Imaging 9, 509-518.

Hadjiprocopis, A., Tofts, P., 2003. An automatic lesion segmentation method for fast spin echo magnetic resonance images using an ensemble of neural networks, Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on, pp. 709-718.

Harmouche, R., Collins, L., Arnold, D., Francis, S., Arbel, T., 2006. Bayesian MS Lesion Classification Modeling Regional and Local Spatial Information, Pattern Recognition, International Conference on. IEEE Computer Society, Los Alamitos, CA, USA, pp. 984-987.

He, R., Narayana, P.A., 2002. Automatic delineation of Gd enhancements on magnetic resonance images in multiple sclerosis. Med Phys 29, 1536-1546.

Holmes, C.J., Hoge, R., Collins, L., Woods, R., Toga, A.W., Evans, A.C., 1998. Enhancement of MR Images Using Registration for Signal Averaging. Journal of Computer Assisted Tomography 22, 324-333.

Jack, C.R., Jr., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., J, L.W., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 27, 685-691.

Jannin, P., Krupinski, E., Warfield, S., 2006. Validation in medical image processing. IEEE Trans Med Imaging 25, 1405-1409.

Johnston, B., Atkins, M.S., Mackiewich, B., Anderson, M., 1996. Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. Medical Imaging, IEEE Transactions on 15, 154-169.

Kamber, M., Shinghal, R., Collins, D.L., Francis, G.S., Evans, A.C., 1995. Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images. Medical Imaging, IEEE Transactions on 14, 442-453.

Karimaghaloo, Z., Shah, M., Francis, S., Arnold, D., Collins, D., Arbel, T., 2010. Detection of Gad-Enhancing Lesions in Multiple Sclerosis Using Conditional Random Fields, In: Jiang, T., Navab, N., Pluim, J., Viergever, M. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010. Springer Berlin / Heidelberg, pp. 41-48.

Kawa, J., Pietka, E., 2007. Kernelized Fuzzy c-means Method in Fast Segmentation of Demyelination Plaques in Multiple Sclerosis, 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, Lyon, France, pp. 5616-5619.

Khayati, R., Vafadust, M., Towhidkhah, F., Nabavi, M., 2008a. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and markov random field model. Computers in Biology and Medicine 38, 379-390.

Khayati, R., Vafadust, M., Towhidkhah, F., Nabavi, S.M., 2008b. A novel method for automatic determination of different stages of multiple sclerosis lesions in brain MR FLAIR images. Computerized Medical Imaging and Graphics 32, 124-133.

Kikinis, R., Guttmann, C.R.G., Metcalf, D., III, W.M.W., Ettinger, G.J., Weiner, H.L., Jolesz, F.A., 1999. Quantitative follow-up of patients with multiple sclerosis using MRI: Technical aspects. Journal of Magnetic Resonance Imaging 9, 519-530.

Kroon, D., Oort, E.V., Slump, K., 2008. Multiple Sclerosis Detection in Multispectral Magnetic Resonance Images with Principal Components Analysis. MS Lesion Segmentation (MICCAI 2008 Workshop).

Kwan, R.K., Evans, A.C., Pike, G.B., 1999. MRI simulation-based evaluation of image-processing and classification methods. IEEE Trans Med Imaging 18, 1085-1097.

Leary, S.M., Miller, D.H., Stevenson, V.L., Brex, P.A., Chard, D.T., Thompson, A.J., 2003. Interferon beta-1a in primary progressive MS: an exploratory, randomized, controlled trial. Neurology 60, 44-51.

Li, D.K., Paty, D.W., 1999. Magnetic resonance imaging results of the PRISMS trial: a randomized, double-blind, placebo-controlled study of interferon-beta1a in relapsing-remitting multiple sclerosis. Prevention of Relapses and Disability by Interferon-beta1a Subcutaneously in Multiple Sclerosis. Ann Neurol 46, 197-206.

Li, D.K., Zhao, G.J., Paty, D.W., 2001. Randomized controlled trial of interferon-beta-1a in secondary progressive MS: MRI results. Neurology 56, 1505-1513.

Liu, J., Smith, C.D., Chebrolu, H., 2009. Automatic Multiple Sclerosis detection based on integrated square estimation, 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, Miami, FL, pp. 31-38.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 16, 187-198.

McAlpine, D., 1973. Multiple sclerosis: a review. Br Med J 2, 292-295.

McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., McFarland, H.F., Paty, D.W., Polman, C.H., Reingold, S.C., Sandberg-Wollheim, M., Sibley, W., Thompson, A., van den Noort, S., Weinshenker, B.Y., Wolinsky, J.S., 2001. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 50, 121-127.

Menze, B., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10, 213.

Miller, D.H., Molyneux, P.D., Barker, G.J., MacManus, D.G., Moseley, I.F., Wagner, K., 1999. Effect of interferon-beta1b on magnetic resonance imaging outcomes in secondary progressive multiple sclerosis: results of a European

multicenter, randomized, double-blind, placebo-controlled trial. European Study Group on Interferon-beta1b in secondary progressive multiple sclerosis. Ann Neurol 46, 850-859.

Molyneux, P.D., Miller, D.H., Filippi, M., Yousry, T.A., Radü, E.W., Adèr, H.J., Barkhof, F., 1999. Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. Neuroradiology 41, 882-888.

Morra, J., Tu, Z., Toga, A., Thompson, P., 2008. Automatic Segmentation of MS Lesions Using a Contextual Model for the MICCAI Grand Challenge. MS Lesion Segmentation (MICCAI 2008 Workshop).

Mortazavi, D., Kouzani, A.Z., Soltanian-Zadeh, H., 2011. Segmentation of multiple sclerosis lesions in MR images: a review. Neuroradiology.

Neema, M., Guss, Z.D., Stankiewicz, J.M., Arora, A., Healy, B.C., Bakshi, R., 2009. Normal findings on brain fluid-attenuated inversion recovery MR images at 3T. AJNR Am J Neuroradiol 30, 911-916.

Nelson, F., Poonawalla, A.H., Hou, P., Huang, F., Wolinsky, J.S., Narayana, P.A., 2007. Improved identification of intracortical lesions in multiple sclerosis with phase-sensitive inversion recovery in combination with fast double inversion recovery MR imaging. AJNR Am J Neuroradiol 28, 1645-1649.

Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P., 2007. Robust fitting of mixtures using the trimmed likelihood estimator. Computational Statistics & Data Analysis 52, 299-308.

Nyul, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. Medical Imaging, IEEE Transactions on 19, 143--150.

O'Riordan, J.I., Thompson, A.J., Kingsley, D.P., MacManus, D.G., Kendall, B.E., Rudge, P., McDonald, W.I., Miller, D.H., 1998. The prognostic value of brain MRI in clinically isolated syndromes of the CNS. A 10-year follow-up. Brain 121 (Pt 3), 495-503.

Parodi, R.C., Sardanelli, F., Renzetti, P., Rosso, E., Losacco, C., Ferrari, A., Levrero, F., Pilot, A., Inglese, M., Mancardi, G.L., 2002. Growing Region Segmentation Software (GRES) for quantitative magnetic resonance imaging of multiple sclerosis: intra- and inter-observer agreement variability: a comparison with manual contouring method. Eur Radiol 12, 866-871.

Paty, D.W., Li, D.K., 1993. Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial. UBC MS/MRI Study Group and the IFNB Multiple Sclerosis Study Group. Neurology 43, 662-667.

Peterson, J.W., Bö, L., Mörk, S., Chang, A., Trapp, B.D., 2001. Transected neurites, apoptotic neurons, and reduced inflammation in cortical multiple sclerosis lesions. Annals of Neurology 50, 389-400.

Pham, D.L., Prince, J.L., 1999. Adaptive fuzzy segmentation of magnetic resonance images. Medical Imaging, IEEE Transactions on 18, 737--752.

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

Polman, C.H., Reingold, S.C., Edan, G., Filippi, M., Hartung, H.P., Kappos, L., Lublin, F.D., Metz, L.M., McFarland, H.F., O'Connor, P.W., Sandberg-Wollheim, M., Thompson, A.J., Weinshenker, B.G., Wolinsky, J.S., 2005. Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". Ann Neurol 58, 840-846.

Prastawa, M., Gerig, G., 2008. Automatic MS Lesion Segmentation by Outlier Detection and Information Theoretic Region Partitioning. MS Lesion Segmentation (MICCAI 2008 Workshop).

Rousseau, F., Blanc, F., de Seze, J., Rumbach, L., Armspach, J.-P., 2008. An a contrario approach for outliers segmentation: Application to Multiple Sclerosis in MRI, Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, pp. 9-12.

Sajja, B.R., Datta, S., He, R., Mehta, M., Gupta, R.K., Wolinsky, J.S., Narayana, P.A., 2006. Unified approach for multiple sclerosis lesion segmentation on brain MRI. Ann Biomed Eng 34, 142-151.

Santago, P., Gage, H.D., 1993. Quantification of MR brain images by mixture density and partial volume modeling. Medical Imaging, IEEE Transactions on 12, 566-574.

Scott, D.W., 2001. Parametric Statistical Modeling by Minimum Integrated Square Error. Technometrics 43, 274-285.

Scully, M., Magnotta, V., Gasparovic, C., Pelligrimo, P., Feis, D., Bockholt, H., 2008. 3D Segmentation In The Clinic: A Grand Challenge II at MICCAI 2008 - MS Lesion Segmentation. MS Lesion Segmentation (MICCAI 2008 Workshop).

Shahar, A., Greenspan, H., 2004. Probabilistic Spatial-Temporal Segmentation of Multiple Sclerosis Lesions, In: Sonka, M., Kakadiaris, I.A., Kybic, J. (Eds.), Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis. Springer Berlin / Heidelberg, pp. 269-280.

Sharon, E., Galun, M., Sharon, D., Basri, R., Brandt, A., 2006. Hierarchy and adaptivity in segmenting visual scenes. Nature 442, 810-813.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic Resonance Image Tissue Classification Using a Partial Volume Model. Neuroimage 13, 856-876.

Shiee, N., Bazin, P., Pham, D.L., 2008. Multiple Sclerosis Lesion Segmentation Using Statistical and Topological Atlases. MS Lesion Segmentation (MICCAI 2008 Workshop).

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2009. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage In Press, Corrected Proof, -.

Sicotte, N.L., Voskuhl, R.R., Bouvier, S., Klutch, R., Cohen, M.S., Mazziotta, J.C., 2003. Comparison of multiple sclerosis lesions at 1.5 and 3.0 Tesla. Invest Radiol 38, 423-427.

Simon, J.H., Jacobs, L.D., Campion, M., Wende, K., Simonian, N., Cookfair, D.L., Rudick, R.A., Herndon, R.M., Richert, J.R., Salazar, A.M., Alam, J.J., Fischer, J.S., Goodkin, D.E., Granger, C.V., Lajaunie, M., Martens-Davidson, A.L., Meyer, M., Sheeder, J., Choi, K., Scherzinger, A.L., Bartoszak, D.M., Bourdette, D.N., Braiman, J., Brownscheidle, C.M., Whitham, R.H., et al., 1998. Magnetic resonance studies of intramuscular interferon beta-1a for relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group. Ann Neurol 43, 79-87.

Simon, J.H., Li, D., Traboulsee, A., Coyle, P.K., Arnold, D.L., Barkhof, F., Frank, J.A., Grossman, R., Paty, D.W., Radue, E.W., Wolinsky, J.S., 2006. Standardized MR imaging protocol for multiple sclerosis: Consortium of MS Centers consensus guidelines. AJNR Am J Neuroradiol 27, 455-461.

Simon, J.H., Lull, J., Jacobs, L.D., Rudick, R.A., Cookfair, D.L., Herndon, R.M., Richert, J.R., Salazar, A.M., Sheeder, J., Miller, D., McCabe, K., Serra, A., Campion, M.K., Fischer, J.S., Goodkin, D.E., Simonian, N., Lajaunie, M., Wende, K., Martens-Davidson, A., Kinkel, R.P., Munschauer, F.E., 3rd, 2000. A longitudinal study of T1 hypointense lesions in relapsing MS: MSCRG trial of interferon beta-1a. Multiple Sclerosis Collaborative Research Group. Neurology 55, 185-192.

Sled, J.G., Pike, G.B., 1998. Standing-wave and RF penetration artifacts caused by elliptic geometry: an electrodynamic analysis of MRI. Medical Imaging, IEEE Transactions on 17, 653--662.

Smith, S.M., 2002. Fast robust automated brain extraction. Human Brain Mapping 17, 143--155.

Souplet, J., Lebrun, C., Ayache, N., Malandain, G., 2008. An Automatic Segmentation of T2-FLAIR Multiple Sclerosis Lesions. MS Lesion Segmentation (MICCAI 2008 Workshop).

Styner, M., Lee, J., Chin, B., Chin, M.S., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S., 2008. 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. MS Lesion Segmentation (MICCAI 2008 Workshop).

Subbanna, N.K., Shah, M., Francis, S.J., Narayanan, S., Collins, D.L., Arnold, D.L., Arbel, T., 2009. MS Lesion Segmentation using Markov Random Fields, MICCAI Workshop in Medical Image Analysis for Multiple Sclerosis (MIAMS), London, UK.

Suri, J.S., Singh, S., Reden, L., 2002. Computer Vision and Pattern Recognition Techniques for 2-D and 3-D MR Cerebral Cortical Segmentation (Part I): A State-of-the-Art Review. Pattern Analysis & Applications 5, 46--76.

Tomas-Fernandez, X., Warfield, S.K., 2011. A new classifier feature space for an improved Multiple Sclerosis lesion segmentation, Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, pp. 1492-1495.

Udupa, J.K., Wei, L., Samarasekera, S., Miki, Y., van Buchem, M.A., Grossman, R.I., 1997. Multiple sclerosis lesion quantification using fuzzy-connectedness principles. Medical Imaging, IEEE Transactions on 16, 598-609.

Van Ginneken, B., Heimann, T., Styner, M., 2007. 3D Segmentation in the Clinic: A Grand Challenge, 3D Segmentation in the Clinic: A Grand Challenge. Miccai Workshop, pp. 7-15.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans Med Imaging 20, 677-688.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based bias field correction of MR images of the brain. Medical Imaging, IEEE Transactions on 18, 885--896.

Vinitski, S., Gonzalez, C.F., Knobler, R., Andrews, D., Iwanaga, T., Curtis, M., 1999. Fast tissue segmentation based on a 4D feature map in characterization of intracranial lesions. Journal of magnetic resonance imaging : JMRI 9, 768-776.

Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. Medical Imaging, IEEE Transactions on 26, 405--421.

Warfield, S.K., Kaus, M., Jolesz, F.A., Kikinis, R., 2000. Adaptive, template moderated, spatially varying statistical classification. Medical Image Analisys 4, 43-55.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. Medical Imaging, IEEE Transactions on 23, 903-921.

Wei, X., Warfield, S.K., Zou, K.H., Wu, Y., Li, X., Guimond, A., Mugler, J.P., Benson, R.R., Wolfson, L., Weiner, H.L., Guttmann, C.R.G., 2002. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. Journal of Magnetic Resonance Imaging 15, 203-209.

Wells III, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of MRI data. Medical Imaging, IEEE Transactions on 15, 429--442.

Wels, M., Huber, M., Hornegger, J., 2008. Fully automated segmentation of multiple sclerosis lesions in multispectral MRI. Pattern Recognition and Image Analysis 18, 347-350.

Wu, Y., Warfield, S.K., Tan, I.L., Wells, W.M., Meier, D.S., van Schijndel, R.A., Barkhof, F., Guttmann, C.R.G., 2006. Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. Neuroimage 32, 1205-1215.

Yamamoto, D., Arimura, H., Kakeda, S., Magome, T., Yamashita, Y., Toyofuku, F., Ohki, M., Higashida, Y., Korogi, Y., 2010. Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society 34, 404-413.

Younis, A., Soliman, A., Kabuka, M., John, N., 2007. MS Lesions Detection in MRI Using Grouping Artificial Immune Networks, 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering. IEEE, Boston, MA, USA, pp. 1139-1146.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. Medical Imaging, IEEE Transactions on 20, 45-57.

Zhao, G.J., Koopmans, R.A., Li, D.K., Bedell, L., Paty, D.W., 2000. Effect of interferon beta-1b in MS: assessment of annual accumulation of PD/T2 activity on MRI. UBC MS/MRI Analysis Group and the MS Study Group. Neurology 54, 200-206.

Zhu, H., Basir, O., 2003. Automated brain tissue segmentation and MS lesion detection using fuzzy and evidential reasoning, Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference on, pp. 1070-1073Vol.1073.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. Medical Imaging, IEEE Transactions on 13, 716-724.

Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans Med Imaging 21, 1280-1291.

Zivadinov, R., Stosic, M., Cox, J.L., Ramasamy, D.P., Dwyer, M.G., 2008. The place of conventional MRI and newly emerging MRI techniques in monitoring different aspects of treatment outcome. J Neurol 255 Suppl 1, 61-74.