



UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA IN NEUROSCIENZE

CICLO XXXIV

COORDINATORE Prof. Felicità Pedata

Facing the current challenges in multiple sclerosis lesions: from automated segmentation to assessing the role of inflammation in neurodegeneration

Settore Scientifico Disciplinare MED/26


Dottorando

Dott. Gentile Giordano



Tutore

Prof. De Stefano Nicola



Coordinatore

Prof. Felicità Pedata

Anni 2018/2021

Index

1. Introduction.....	4
2. MS and MRI	5
2.1 Lesions	6
2.2 Atrophy.....	9
3. Aim of the thesis	14
3.1 MS Challenge 1: Automated lesion segmentation	14
3.2 MS Challenge 2: The inter-role between inflammation and neurodegeneration	15
4. Study 1: BIANCA-MS: an optimised tool for automated multiple sclerosis lesion segmentation	16
5. Study 2: The concurrent Spatio-temporal relationship between inflammation and neurodegeneration in early Multiple Sclerosis: A Post-hoc Analysis of the REFLEXION Study	51
6. Study 3: The Spatio-temporal Relationship Between White Matter Lesions and Brain Atrophy in Clinically Isolated Syndrome and Early Multiple Sclerosis: A Post-hoc Analysis of the REFLEXION Study.....	72
7. Summary and future perspectives	92
8. References.....	98

Abbreviations

AI: Artificial Intelligence

BBB: Blood Brain Barrier

BET: Brain Extraction Tool

BIANCA: Brain Intensity AbNormality Classification Algorithm

CDMS: Clinically Definite Multiple Sclerosis

CIS: Clinically Isolated Syndrome

CNN: Convolutional Neural Network

CNS: Central Nervous System

CSF: Cerebrospinal Fluid

DIS: Dissemination In Space

DIT: Dissemination in Time

DT: Delayed Treatment

ET: Early Treatment

FLIRT: FMRIB's Linear Image Registration Tool

FNIRT: FMRIB's Non-Linear Image Registration Tool

FSL: FMRIB Software Library

Gd: Gadolinium

GM: Gray Matter

HC: Healthy Control

LCM: Lesion Change Map

LGA: Lesion Growth Algorithm

LMM: Linear Mixed Model

LPA: Lesion Prediction Algorithm

LTP: Location of Training Points

MRI: Magnetic Resonance Imaging

MS: Multiple Sclerosis

NBV: Normalized Brain Volume

NAWM: Normal-Appearing White Matter

nFNC: number of False Negative Clusters

nFPC: number of False Positive Clusters

NTP: Number of Training Points

PBVC: Percentage Brain Volume Change

PD: Proton Density
PPMS: Primary Progressive Multiple Sclerosis
PS: Patch Size
PVE: Partial Volume Effect
PVVC: Percentage Ventricular Volume Change
QNL: Quantitative NeuroImaging Laboratory
REFLEXION: REbif FLEXible dosing in early MS extensION
RRMS: Relapsing Remitting Multiple Sclerosis
RQ: Research Question
SPM: Statistical Parametric Mapping
SPMS: Secondary Progressive Multiple Sclerosis
SW: Spatial Weighting
TE: Echo Time
TLVC: Total Lesion Volume Change
TR: Repetition Time
WM: White Matter
WMH: White Matter Hyper-intensities

1. Introduction

Multiple sclerosis (MS) is a widespread inflammatory, demyelinating and neurodegenerative disease of the central nervous system (CNS) (Filippi, Preziosa, & Rocca, 2018). The incidence of MS is not homogeneous, but changes in a considerable way over the world. Where the prevalence is higher (North Europe, North America and Australia), about 30 on 100,000 people are affected (GBD 2016 Neurology Collaborators, 2019). The causes of MS are still unknown, but it probably includes genetic as well as environmental factors.

Different subtypes of disease exist: about the 85% of patients belong to that one named Relapsing Remitting (RR). These patients alternate attacks of focal neurological deficits (relapses) followed by clinically silent period (remitting phase). In RR patients the accumulation of clinical disability can be considered as the sum of residual disability not resolving after the attacks (Weinshenker, et al., 1989). After about 10-15 years the accumulation of disability usually change its course, becoming continue and progressive. This phase of disease is named Secondary Progressive MS (SPMS). Finally, about 15% of patients show a progressive accumulation of clinical disability from the disease onset. For this reason, the name of this last MS subtype is Primary Progressive MS (PPMS) (Lublin & Reingold, 1996). There is not a privileged area of CNS involved at the onset of the disease, but different areas can be contemporary affected from the beginning of the MS. Thus, clinical symptoms greatly vary across patients, ranging from optics neuritis and muscle weakness to cognitive decline and fatigue (Noseworthy, Lucchinetti, Rodriguez, & Weinshenker, 2000).

The following paragraph will briefly describe Magnetic Resonance Imaging (MRI) basic concepts, with emphasis on the sequences that are routinely used in MS. Afterwards, a detailed overview of the MS pathological hallmarks that are regularly analysed on MRI in both clinical

and research setting will be provided. Within this respect, brain lesions and atrophy biology, their appearance on MRI, how they are quantified, and the related challenges that are currently faced will be explained. Finally, the aim of the thesis will be provided.

2. MS and MRI

MRI is an imaging tool that currently offers the most sensitive non-invasive way of imaging the brain, spinal cord, or other areas of the body. Briefly, MRI works on the principle of nuclear magnetic resonance, a phenomenon where nuclei of atoms get excited in the magnetic field by electromagnetic waves and emit signals. Such signals reflect the characteristics of the imaged tissues and mostly derive from the hydrogen protons. Thus, the intensities of acquired images are influenced by the protons density and by the local environment of water molecules.

By properly choosing the acquisition parameters, different kind of images can be obtained. In MS clinical practice the sequences that are routinely acquired are the T1-weighted, the proton density (PD) and the T2-weighted images (including fluid-attenuated inversion recovery [FLAIR]). T1-weighted sequence provide good contrast between gray (GM) and white matter (WM) resulting in images that most closely approximate the appearances of tissues macroscopically, although this is a gross simplification. Thus, this sequence is best for detecting anatomical abnormalities like tissue loss (i.e. atrophy). Further, contrast-enhanced T1-weighted imaging is a sensitive method for detecting active MS lesions. On such sequence, gadolinium (Gd) enhancement appears as an hyperintense area which reflects blood brain barrier (BBB) breakdown and histologically correlates with the inflammatory phase of lesion development.

T2-weighted and PD images provide a good depiction of disease because pathological processes, such as demyelination or inflammation, are often related to increase in water content; thus, the affected areas appear bright on these sequences.

Given this context, MRI has the potential to depict the MS pathological hallmark: the concurrent presence of focal areas of inflammation (i.e. lesions) and of diffuse damage and neurodegeneration (i.e atrophy) (Lassmann, Brück, & Lucchinetti, 2007).

2.1 Lesions

Biological features. Lesions are characterized by the infiltration of inflammatory cells and the breakdown of the BBB (Lassmann, van Horssen, & Mahad, 2012). The mechanisms of BBB breakdown are incompletely understood but seem to involve direct effects of pro-inflammatory cytokines and chemokines produced by resident cells and endothelial cells, as well as indirect cytokine-dependent and chemokine-dependent leukocyte mediated injury (Minagar & Alexander, 2003; Ortiz, et al., 2014). The dysregulation of the BBB increases the infiltration of activated leukocytes, including macrophages, T cells and B cells, into the CNS, which leads to further inflammation and demyelination, followed by oligodendrocyte loss, reactive gliosis and neuro-axonal degeneration (Filippi, et al., 2018; Reich, Lucchinetti, & Calabresi, 2018). In active lesions, biopsies and autopsies showed a profound pathologic heterogeneity with four major patterns of immunopathology, suggesting that the targets of injury and mechanisms of demyelination in MS are heterogeneous and evolve over the course of months (Lucchinetti, et al., 2000). In this respect, WM inflammation plaques could be further classified as chronic active lesions, slowly expanding lesions and inactive lesions. The first ones are more frequent in MS patients with a longer disease duration and SPMS and are characterized by macrophages and microglia at the edge of lesions, with an inactive centre (Filippi, et al., 2018). Inactive lesions are sharply circumscribed, hypocellular and show reduced axonal density and lower density of lymphocytes than active plaques (Lassmann, van Horssen, & Mahad, 2012; Prineas, et al., 2001). Finally, mixed active/inactive lesions are also detected in MS.

Appearance on MRI. On PD, T2-weighted and FLAIR images MS lesions are easily detected in the WM as ovoid or round hyperintense elements (Filippi, et al., 2012; Filippi, et al., 2019) with the major axis usually perpendicular to the corpus callosum (Dawson's fingers). Dimensions can range from few millimetres to more than 1 centimetre. MS lesions are disseminated throughout the CNS but have a predilection for optic nerves, subpial spinal cord, brainstem, cerebellum, and juxtacortical and periventricular WM regions (Filippi, et al., 2019). A subset (from 10 to 30%) of these T2-hyperintense MS lesions may appear hypointense on corresponding T1-weighted images (Filippi, et al., 2012). These hypointense lesions are commonly referred to as black holes and indicate areas with pathologically confirmed severe tissue destruction (Filippi, et al., 2012).

Clinical relevance. The identification of lesions in the CNS on both cross-sectional and longitudinal MRI scans is a MS crucial diagnostic step for the demonstration of dissemination in space (DIS) and time (DIT) (Thompson, et al., 2018). Cross-sectional MRI scans provide an overview of the lesion damage accumulated this far in a patient and thus are used for assessing DIS. Longitudinal MRI scans are currently used for demonstrating DIT and are very useful for evaluating lesions activity over time in terms of newly formed, enlarging, shrinking or disappearing lesions. Thus, they provide an overview of disease evolution. Cross-sectional and longitudinal MRI studies demonstrated how lesions number and volume are associated with short and long-term changes in physical disability as well as clinical progression (O'Riordan, et al., 1998; Tintore, et al., 2015; Fisniku, et al., 2008; Brownlee, et al., 2019). For this reason, lesion load (i.e. number and volume), the number of Gd enhancing and newly formed lesions are often used as MRI outcomes in MS clinical trials (van Munster & Uitdehaag, 2017). Given this background, lesion identification and quantification on MRI is an important step in MS diagnosis, in monitoring disease progression and evaluating treatment efficacy.

Quantification approaches. To date, manual approach is considered the golden standard procedure for detecting MS lesions. However, this procedure is time consuming and prone to intra and inter-rater variability, which in turn could result in a large difference in the extracted lesions values (García-Lorenzo, Francis, Narayanan, Arnold, & Collins, 2013), thus limiting its use in large studies. Several authors (Filippi, et al., 1995; Grimaud, et al., 1996; Udupa, et al., 1997) have proposed semiautomated segmentation methods, whereby the computer aids the expert to reduce both segmentation time and rater variability. However, their use in large clinical trials is time consuming and deals incompletely with inter-rater differences. Given this background, automated tools represent a fascinating solution. However, automated lesions quantification is a complex task for several reasons, and no satisfactory solution has yet been reached.

Current challenges. Several automated tools with different operating mechanisms and architecture have been proposed in the last years. However, diverse issues limit their use in clinical practice. First, automated accurate identification of MS lesions on MRI brain image is extremely difficult due to variability in lesion location, size and shape in addition to anatomical variability between subjects (García-Lorenzo, Francis, Narayanan, Arnold, & Collins, 2013). Second, most of these tools are developed to be protocol specific (Mårtensson, et al., 2020) or are poorly validated (Griffanti, et al., 2016). Consequently, such tools failed to be generalizable when applied to MRI with different acquisition conditions and often did not guarantee the same performances on new or “unseen” MRI datasets. Finally, the validation of these tools suffers from two limitations. The first one concerns with the limited and scarce use of high-resolution FLAIR images, now considered the preferable sequences where to perform segmentation due to their high sensitivity in lesion detection (Filippi, et al., 2019). Second, these tools are usually validated using manual segmentation as ground truth. As this approach is prone to error and is

highly subjective and difficult to reproduce, the validation process must be critically considered. Indeed, recent evidence suggested how the intra and inter-rater variability could affect the automated segmentation performances (Shwartzman, Gazit, Shelef, & Riklin-Raviv, 2019). Given this context, the automated identification of MS lesions is still an open challenge, and no standardized tool has been widely employed.

2.2 Atrophy

Biological features. Different pathological substrates may be responsible for brain atrophy: loss of myelin, glial cells, neurons and axons due to demyelination and neurodegeneration. Establishing to which extent these components contribute to tissue loss is complex as they could depend on many factors, such as disease stage, brain region affected, type of pharmacological treatment, presence of comorbidities and other factors unrelated to the disease (Giorgio, Battaglini, Smith, & De Stefano, 2008). Brain atrophy affects both WM and GM. Atrophy of non-lesional or normal-appearing WM (NAWM) is likely secondary to demyelination, axonal damage and loss, the latter partially caused by Wallerian degeneration (Filippi, et al., 2012) and slow, progressive axonal loss throughout the brain, due to diffuse inflammation and oxidative stress (Kornek, et al., 2000). GM atrophy is common in neocortical areas, but is also found in other GM areas, such as the thalamus, hippocampus, and cerebellum. Evidence suggest how WM damage partially contribute to GM atrophy (Battaglini, et al., 2009; Riccitelli, et al., 2011). Thus, several mechanism drive GM atrophy. A study relating GM atrophy patterns measured using post-mortem MRI to histopathology showed that atrophy is explained predominantly by (neuro)axonal loss and neuronal shrinkage and is largely independent of demyelination (Popescu, et al., 2015).

Appearance on MRI. Cerebral atrophy is simply the compensatory enlargement of the cerebrospinal fluid (CSF) spaces derived from the reducing brain parenchymal volume. Thus, on the majority of MS brain T1-weighted images, brain atrophy can be detected as ventricular system enlargement, widening of cortical sulci and gyri and cortex displacement from inner skull.

Clinical Relevance. Over the last 2 decades, several cross-sectional and longitudinal MRI studies have been performed to elucidate the clinical relevance of brain atrophy in MS. Within this respect, cross-sectional MRI scans are used to quantify brain atrophy accumulated this far (i.e. the atrophy state), whereas longitudinal MRI data allow to assess the brain volume changes over time (i.e. the atrophy rate). Evidence highlights how periventricular atrophy is detected in patients with clinically isolated syndrome (CIS) who evolve to MS compared to those who does not (Dalton, et al., 2002). Greater brain atrophy develops in patients with worsening disability than in those who are clinically stable (Ingle, Stevenson, Miller, & Thompson, 2003). Whole brain atrophy correlates with cognitive dysfunction (Rao, Leo, Haughton, St Aubin-Faubert, & Bernardin, 1989) and mood disturbances (Bakshi, et al., 2000). Finally, quantification of brain volume on early scans provides prognostic measures of clinical status not only for medium and long-term follow-up (Fisher, et al., 2002) but also for short-term (over 6 months) decline (Gauthier, et al., 2007). In the field of regional atrophy, GM tissue loss provides more clinically relevant information than does WM (Rocca, et al., 2017) and better explain physical and cognitive impairment (Roosendaal, et al., 2011). Further, subcortical deep GM volume may be present in the early stage of the disease, and it is strongly correlated with the disease course (Eshaghi, et al., 2018). Given this context, cross-sectional and longitudinal assessment of brain atrophy on MRI are valid measures of disease burden and progression and thus are of great relevance for better understanding the pathophysiology of MS. Further, as such

measures reflect neurodegenerative processes, brain volumetric changes are regularly quantified for assessing neuroprotection in clinical trials (Zivadinov & Bakshi, 2004; De Stefano, et al., 2014; Barkhof, Calabresi, Miller, & Reingold, 2009).

Quantification approaches. As for lesions, manual outlining is considered the most accurate method for both whole and regional brain segmentation. However, its use on large studies is limited for several reasons (see lesions quantification approaches paragraph). Semi-automated procedure certainly reduce the time needed for the segmentation, but do not offer a definitive solution and still have the disadvantages of user intervention, albeit to a lesser extent. Given this context, several approaches for the automated quantification of brain volumes have been developed. Currently, FSL (FMRIB Software Library, <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki>) SIENAX and SIENA (Smith, et al., 2002) are ones of the most widely used tools for assessing cross-sectional brain volumes and longitudinal brain volumes change over time, respectively. Further, the SIENA method has been extended (SIENAr) (Bartsch, et al., 2007) to allow the voxel-wise statistical analysis of brain atrophy across subjects, which results in a regional analysis of differences in brain volume occurring over time between two groups of subjects (Battaglini, et al., 2009).

Along with FSL, SPM (Statistical Parametric Mapping, <https://www.fil.ion.ucl.ac.uk/spm/>) and FreeSurfer (<https://surfer.nmr.mgh.harvard.edu/>) have provided automated pipelines that are widely used by the neuroimaging communities. However, all these tools are still considered far from manual segmentation and can produce considerable errors.

Current challenges. When dealing with brain volume assessment, different challenges are routinely faced in both clinical and research settings. First, automated reliable quantification of brain tissues volume is not an easy task as image noises, artifacts, poor contrast between tissues

and bias field inhomogeneity certainly affect the segmentation procedure. Further, the complex anatomy of the brain and the finite resolution of MRI images lead to a phenomenon called partial volume effect (PVE). If ignored, PVE can bias brain measurements in the range of 20%-60% (González Ballester, Zisserman, & Brady, 2002). Finally, producing an accurate separation of GM and WM at their interface is quite difficult and thus, their measures are relatively unstable (Battaglini, Jenkinson, De Stefano, & ADNI, 2018). Given this context, the automated segmentation of brain tissues represents the first challenge in MS brain atrophy assessment. In this respect, our laboratory (Quantitative Neuroimaging Laboratory, QNL) has recently developed both cross-sectional (SIENAX 2.0) (Luchetti, Gentile, Battaglini, Giorgio, & De Stefano, 2019) and longitudinal tools (SIENA-XL) (Battaglini, Jenkinson, De Stefano, & ADNI, 2018) for the accurate and robust assessment of whole and regional (GM/WM) brain volumes which demonstrated significant decrease in the measurements errors on multicentre-datasets.

The second challenge is represented by the presence of factors now recognized as potential modifiers of pathological brain atrophy estimates. Particular attention should be given to pseudoatrophy, the paradoxical acceleration of brain atrophy following the initiation of anti-inflammatory therapies. Such phenomenon, which is thought to reflect fluid shift related to resolution of inflammatory oedema, certainly complicates the interpretation and the clinical impact of brain volume measurements in MS. Further, only recently the contribution of normal aging to brain volume loss has begun to be explored. In this respect, our laboratory has recently provided “pathologic cut-offs” values for whole-brain atrophy to discriminate patients with MS from healthy controls (HCs) (De Stefano, et al., 2016). Given this context, establish to which extent brain atrophy measures are related to ‘true’ MS pathological mechanism needs to be elucidated and thus caution is needed when moving atrophy measures into clinical practice (Rocca, et al., 2017).

The third challenge that greatly limits the use of brain volume measurements in clinical practice is related to the absence of a common normative database against which compare the raw volumes obtained on MS patients (Rocca, et al., 2017). In this respect, our laboratory has recently provided both cross-sectional and longitudinal brain atrophy normative values for assessing the deviation from the expected brain volume loss in patients with neurological disorders (De Stefano, et al., 2018; Battaglini, et al., 2019).

Finally, the fourth challenge concerns with the complex inter-role between WM lesions (inflammation) and brain atrophy (neurodegeneration) in MS. Although these two pathological mechanisms are present early in the disease course of MS, the dynamics of accumulation of WM lesions and brain atrophy is not completely understood (Fisher et al., 2002). Genetic data, observations from most experimental models and MRI studies appeared to favour a pathogenesis model in which inflammation precedes neurodegeneration (Dalton, et al., 2002; Chard, et al., 2003; Paolillo, et al., 2004; Milo, Korczyn, Manouchehri, & Stüve, 2020). Within this respect, several studies tried to understand the mechanisms that relates lesions to brain volume loss in MS. For example, axonal loss in lesions could cause atrophy by two mechanisms: tissue loss within the lesion per se, and Wallerian degeneration in related fibre pathways (Miller, Barkhof, Frank, Parker, & Thompson, 2002). Another plausible mechanism is the existence of destructive lesions especially in the periventricular regions resulting in an increase in ventricular volume, thus leading to increased brain volume loss (Dalton, et al., 2002). Other studies suggested how neurodegeneration is related to pathogenic mechanisms primed by the preceding inflammation and later perpetuating with disease progression (Andravizou, et al., 2019).

Conversely, evidence from clinical trials suggest how anti-inflammatory therapies exerts only a moderate effect on brain volume loss and no available treatment does completely halt neurodegeneration (Bross, Hackett, & Bernitsas, 2020; Dendrou, Fugger, & Friese, 2015; De

Stefano, et al., 2014). Further, serial MRI studies show that subtle focal changes in the WM can be seen weeks before a classical new lesion is formed (Filippi, Rocca, Martino, Horsfield, & Comi, 1998; Narayana, Doyle, Lai, & Wolinsky, 1998). Other MRI evidence suggested that lesion damage contribute only partially to brain atrophy (Battaglini, et al., 2009; Cappellani, et al., 2014; Roostendaal, et al., 2011). These findings suggest how brain atrophy mechanisms could be indirectly related to or are independent from traditional measures of overt lesions (Bermel & Bakshi, 2006).

In conclusion, although evidence suggest the presence of inflammation in every stage of the disease, establishing the role of MS lesions in driving brain atrophy is not straightforward and needs further clarification (Lassmann, 2007) as a more complex interplay across these two pathological processes might occur.

3. Aim of the thesis

The aim of this thesis is to separately face the two MS challenges strictly related to WM lesions. In this respect, we will deal with both the more “practical” issue of automated lesions segmentation and the more “conceptual” topic related to the role of MS lesions in brain atrophy. In the following, a short description of the content of each chapter composing the main body of this thesis will be provided.

3.1 MS Challenge 1: Automated lesion segmentation

As lesions quantification is a key biomarker in MS diagnosis, monitoring disease course and treatment response, the need of a widely validated tool that can be considered totally comparable to manual segmentation is of utmost relevance in both clinical and research settings. Thus, in the fourth chapter of this thesis, we addressed the automated MS lesions segmentation technical challenge by describing a novel segmentation tool specifically tailored on MR brain images of MS patients. Within this respect, we developed a pipeline able to reduce

the impact of the inter-rater variability on lesions segmentation and whose settings are generalizable across different acquisition protocols.

3.2 MS Challenge 2: The inter-role between inflammation and neurodegeneration

Understanding the spatio-temporal relation between inflammation and neurodegeneration in the early phase of MS is of key relevance not only for developing more targeted and effective therapeutic strategies, but also for broaden our comprehension of the underlying disease mechanisms, how these relate to disease progression and whether these can be either modified by treatment or disease worsening. Further, exploring the spatio-temporal evolution of these two pathological processes will make data interpretation in both clinical and research settings clearer and more straightforward. To investigate the role of WM lesions in driving brain atrophy over time and how the evolution of inflammation and neurodegeneration is related, two complementary studies were performed. In study:

1. We focused on the assumption that inflammation and neurodegeneration are two separate pathological mechanisms. Thus, in the fifth chapter of this thesis, we investigated whether WM lesions and brain atrophy developed simultaneously over time and tested whether these two processes were spatially interconnected within the same follow-up period.
2. We focused on the assumption that inflammation precedes neurodegeneration. Thus, in the sixth chapter of the thesis, we investigated whether WM lesions were spatio-temporally related to subsequent atrophy.

4. Study 1: BIANCA-MS: an optimised tool for automated multiple sclerosis lesion segmentation

This study was performed in collaboration with Professor Mark Jenkinson and the post-doctoral researcher Ludovica Griffanti of Oxford FSL (FMRIB software library) laboratory.

Introduction

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system (CNS), characterized by focal areas of inflammation resulting in lesions on Magnetic Resonance Imaging (MRI). The identification of these lesions is the fundamental diagnostic step for the demonstration of dissemination in space and time (Thompson, et al., 2018) and plays an important role in predicting disease course, as lesion number and volume are associated with short and long-term changes in physical disability (Brownlee, et al., 2019; Tintoré, et al., 2006). To date, manual segmentation is considered the best approach for segmenting lesions, but this procedure is time consuming and is affected by intra/inter-rater variability, thus limiting its use in large studies (García-Lorenzo, Francis, Narayanan, Arnold, & Collins, 2013). Against this background, automated tools have been increasingly developed and employed over the last years (Danelakis, Theoharis, & Verganelakis, 2018; Zeng, Gu, Liu, & Zhao, 2020; Shanmuganathan, Almutairi, Aborokbah, Ganesan, & Ramachandran, 2020; Kaur, Kaur, & Singh, 2021). Among automated methods, artificial intelligence (AI) approaches showed higher performance; however, their use in clinical practice is limited for several reasons. Firstly, the majority of AI tools are developed, trained and validated on a limited amount of data, mostly using a single scanning protocol (Mårtensson, et al., 2020). This often leads to algorithm “default” settings that fail to be generalizable when applied to MRI dataset with different acquisition protocols. To guarantee similar performances of AI software on new MRI, thus, a long and complex optimization pipeline needs to be performed (Griffanti, et al., 2016; Popescu, et al., 2012). Another, MS specific, limit consists in the lack of an extensive validation of these tools on large datasets of high-resolution 3D FLAIR images, now considered the most sensitive sequence for the detection of lesions in MS (Filippi, et al., 2019; Paniagua Bravo, et al., 2014). Finally, the inter-rater differences in lesions contouring affects both manual and automated lesion segmentation and it has been shown that it reduces the performances of AI

tools even when the same test data, algorithm settings and architecture are employed (Shwartzman, Gazit, Shelef, & Riklin-Raviv, 2019).

In 2016, the FSL group developed a machine learning tool (BIANCA) to segment WM hyperintensities (WMH) of presumed vascular origin (Griffanti, et al., 2016). Recently, two studies employed BIANCA in MS lesion segmentation on FLAIR images (Weeda, et al., 2019; Duong, et al., 2019). However, in the former, very few algorithm settings (n=18) were tested on a limited amount of MRI data (n=14) and in the latter detailed information about the number of MS subjects and the setting employed were not provided. Therefore, the lack of data or an insufficient optimisation process may have limited the algorithm's performance.

In this work BIANCA-MS, a new pipeline to automatically detect MS WM lesions that is based on the original version of BIANCA, is introduced with the aims to:

1. provide the best parameters setting to be implemented in BIANCA for MS lesion segmentation, irrespective of scanning protocols, magnetic field strength, set of modalities and image resolutions.
2. refine the lesions mask and reduce the impact of inter-rater variability on automated tools segmentation by introducing a post-processing cleaning step.

Once designed, the following analyses were performed to validate BIANCA-MS. We (i) compared the performance of our approach with other currently available, and widely used, tools for automatic lesion segmentation in MS; on a multicentre dataset, we (ii) tested BIANCA-MS cross-centre generalisation by comparing algorithm performances using mixed and site-specific training and test sets; finally, we (iii) evaluated BIANCA-MS performance when all the datasets were pooled together.

Material and Methods

MRI data

In this study, MRI data from 470 Relapsing-Remitting MS subjects were included, belonging to 3 datasets that were different with respect to scanner manufacturers, magnetic field strengths and MRI acquisition protocols (see table 1). Dataset 1 consisted of 200 scans acquired at the University of Siena on a 1.5T Philips Gyroscan MRI (Philips Medical Systems, Best, Netherlands); Dataset 2 consisted of 120 scans acquired at the Meyer Hospital on a 3T Philips Achieva dStream (Philips Medical Systems, Best, Netherlands); Dataset 3 consisted of 150 scans from a multicentre retrospective study, whose scans were collected at different imaging centres using similar acquisition parameters. No selection criteria were used for lesion characteristics (i.e. large confluent lesions or focal distinct T2-hyperintense areas).

	Dataset 1			Dataset 2		Dataset 3		
Image	Flair	T1-weighted	PD	Flair	T1-weighted	PD	T1-weighted	T2-weighted
Voxel size (mm³)	0.97X0.97X3	0.97X0.97X3	0.97X0.97X3	1X0.97X0.97	1X1X1	0.97X0.97X3	0.97X0.97X3	0.97X0.97X3
TR (ms)	9000	35	30	11000	25	2200-3000	550-700	1800-2800
TE1 (ms)	150	10	90	125	4.6	15-50	10-20	30-50
TE2 (ms)	-	-	-	-	-	80-120	-	60-100
Inversion Recovery Delay (ms)	2725	-	-	2800	-	-	-	-
Image Resolution	2D	2D	2D	3D	3D	2D	2D	2D

Table 1. Acquisition protocols for each Dataset. TR=Repetition Time. TE= Echo Time

Gold Standard lesion segmentation

WM lesions were outlined by expert tracers (M.I. and M.L.) using a semi-automated segmentation technique based on local thresholding (www.xinapse.com/jim-7-software/). Recently published guidelines were followed to enhance the proper recognition of MS lesions (Filippi, et al., 2019). Briefly, WM lesions were defined as areas of focal hyperintensity on a T2-weighted (T2, T2-FLAIR or similar) or a proton density (PD)-weighted sequence. The raters had access to all the available sequences, thus segmentation was achieved through consensus of information among all modalities. These lesion masks were used as the “gold standard” for measuring the performance of the automatically obtained segmentations.

Training, validation and test set creation

Each dataset was divided into training, validation and test sets. The training and validation sets were built as follows: 100 subjects from the first and third datasets, and 80 subjects from the second dataset were randomly selected, following the alphabetical order of their pseudo-anonymised codes. Following the proposed suggestion to use patients with higher lesion loads to train the algorithm (Griffanti, et al., 2016), these randomly selected subjects were further split into 2 groups: the half with the higher lesion loads were inserted into the training set (Dataset 1 and Dataset 3: 50; Dataset 2: 40), the others in the validation set (Dataset 1 and Dataset 3: 50; Dataset 2: 40). The remaining subjects from each dataset, with variable lesion loads, formed the three test sets (Dataset 1: 100; Dataset 2: 40; Dataset 3: 50) (see table 2). Details about the choice of the optimal number of subjects to be included in the training set are described in the supplementary materials.

	Dataset 1		Dataset 2		Dataset 3	
	N° of subjects	Lesion Volume	N° of subjects	Lesion Volume	N° of subjects	Lesion Volume
Training	50	16.74±14.62	40	11.82±11.65	50	23.16±13.13
Validation	50	1.33 ±0.93	40	1.32±1.03	50	5.18±3.2
Test	100	6.01±7.42	40	5.51±7.39	50	13.51±12.7

Table 2. Subdivision of subjects for each dataset. Lesion volume is reported in cm^3

MRI Analysis

Data preprocessing

Image quality was assessed for aliasing, ghosting and other type of artefacts. After this step, 3 subjects from Dataset 3 were excluded from this study. For each dataset all modalities were co-registered to the reference sequence where lesions have been segmented (FLAIR 2D and 3D respectively for Dataset 1 and 2 and PD for Dataset 3) using FMRIB's Linear Image Registration tool (FLIRT) (Jenkinson & Smith, 2001; Jenkinson, Bannister, Brady, & Smith, 2002). Brain masks were obtained on T1-weighted images using the Brain Extraction Tool (BET) (Smith, 2002) from FSL and then registered on the main modality to obtain FLAIR/PD brain tissues.

BIANCA-MS

Here we present the two steps performed to obtain BIANCA-MS, a modified version of BIANCA that implements one fixed optimal setting, which we identified after a large optimization procedure, and a post-processing cleaning step for reducing the inter/intra-variability in gold-standard lesion creation.

BIANCA optimisation

BIANCA is a flexible, multimodal supervised method based on the k-nearest neighbour algorithm. Briefly, the algorithm learns the definition of a lesion from a set of manually

segmented masks, by using the voxel intensities and the spatial distribution of the intensities as features. BIANCA can be optimised using different options like:

- The possibility of weighting the spatial coordinates (i.e. spatial weighting option) which can increase the accuracy of segmentation, as in some brain regions lesions are more likely to occur.
- The inclusion of intensity information about a small neighbourhood of each voxel (patch size option) that can make the segmentation more robust to misregistration and provides local context.
- The selection of the number and the position of training points which is important to establish to what extent information around lesion edges is crucial for segmentation (i.e. Location of non-lesion training points and Number of Training points options).

There are also post-processing steps to perform on BIANCA outputs: threshold selection and masking. The former highly influences the results: more restrictive thresholds reduce false positives but increase false negatives (Anbeek, Vincken, van Osch, Bisschops, & van der Grond, 2004). The latter consists of applying an exclusion mask to BIANCA outputs (Griffanti, et al., 2016) to reduce false positives.

To select the optimized BIANCA setting, the training and validation sets of each dataset were used. There were 108 different sets of options used for training BIANCA (table 3). The parameters that were varied were the spatial weighting (SW), number of lesion and non-lesion training points (NTP), patch size (PS) and the location of non-lesion training points (LTP). All the available MRI sequences for each dataset were used, exclusion masks were applied and a threshold of 0.9 was used to obtain the binarized lesion mask. Finally, each of the 108 trained BIANCA models was applied to the validation set of the 3 datasets (see table 3).

Features	Values Tested	Number of options
Spatial weighting (SW)	1-5-10	3
Patch size (PS)	None-3-6-9	4
Location of non WMH training points (LTP)	Surround-Any-No Border	3
N° of training points (WMH VS non WMH) (NTP)	500-1000; 2000-2000; 2000- 10000	3
Total combinations		108

Table 3. The list of the different values tested in this study during the phase of algorithm optimization.

Post-processing cleaning step

Supervised algorithms trained on lesion masks outlined by different users provide slightly different results when applied to new images (Bordin, et al., 2021). To reduce the impact of the inter-rater variability, we developed a post-processing cleaning step to be applied to lesions mask, either automatically or manually outlined. This procedure starts by obtaining pure WM, GM and CSF masks by removing lesion clusters from the three different tissue classes provided by FAST (Zhang, Brady, & Smith, 2001). Two stages then follow: a “Recovery” phase to avoid the loss of lesion voxels erroneously not included in lesion mask and a “Refining” phase that aims to create a cleaned lesion mask by considering local intensity contrast and B0 and radiofrequency (RF) inhomogeneity which reflected smooth spatial intensity changes through the whole image (figure 1).

To perform the “Recovery” step, lesion masks ($mask_0$) were 3D dilated by a single voxel 4 times ($mask_{d1} = dilate3d(dilate3d(dilate3d(dilate3d(mask_0))))$) and only those voxels not adjacent to lesions were retained to form the dilated strip masks ($dilated_{strip} = mask_{d1} - dilate3d(mask_0)$). The WM surrounding the lesions was defined by applying a one voxel 3D dilation ($mask_{d2} = dilate3d(mask_0)$), then taking this new dilated strip and masking this by the pure WM masks obtained from FAST:

$$mask_{swm} = (mask_{d2} - mask_0) * mask_{purewm}$$

From this the Mahalanobis distance (MD) was defined as:

$$MD = \frac{|mask_{swm} - \mu_{dilated_strip}|}{\sigma_{dilated_strip}}$$

where $\mu_{dilated_strip}$ and $\sigma_{dilated_strip}$ are the mean and standard deviation, respectively, of the intensities of the voxels in the dilated strips. Only those voxels of the surrounding WM with a MD bigger than a mild threshold (2.0) from the dilated strip were included to form the “recovery lesion” mask ($mask_{rec}$); see “Recovery” in figure 1.

The “Refining” phase consists of two consecutive steps. In the first refinement step, the intensities of all the voxels from the “recovery lesion” mask were compared to the intensity of the strip of the newly defined surrounding WM (after dilating 5 times the “recovery lesion” mask)

$$mask_{swm2} = (dilated3d(dilate3d(dilate3d(dilate3d(dilate3d(mask_{rec})))))) - mask_{rec}) * mask_{purewm}$$

and retained when the MD was bigger than a sequence-dependent threshold (4.0 for PD, 7.0 for FLAIR images). This step creates a new “roughly refined” mask ($mask_{ref1}$); see Refine-step 1 in figure 1. From the $mask_{ref1}$ we removed all the lesions clusters with less than 4 voxels because lesions below that size are not well defined (Filippi, et al., 2019) and thus could impair segmentation reproducibility. Further, these clusters are very likely to be false positive findings of the algorithm. In the second refinement phase, a strip of WM was created by dilating 3 times the “roughly refined” mask

$$mask_{swm3} = (dilate2d(dilate2d(dilate2d(mask_{ref1}))) - mask_{ref1}) * mask_{purewm}$$

All the voxels with an MD distance from the surrounding WM bigger than 4.0 were retained to create the final output. This last step (MD distance calculation and thresholding) was performed separately at each slice of each lesion since differences of intensity due to the RF

inhomogeneities lead to systematic differences in the selection criteria of the retained voxels (Refine- step 2 in figure 1). To select all the thresholds, we tested several values for each sequence (e.g. FLAIR or PD images) from an internal dataset, that was not used in the analysis, and chose the ones providing the best results upon visual inspection.

To test the impact of this post-processing cleaning step, 5 high resolution FLAIR images (different from the one used for selecting the MD thresholds) were segmented by two experienced tracers. Afterwards, the lesion masks obtained were processed with the procedure previously described and the differences across raters prior and after cleaning was assessed. Once validated, we implemented this post-processing cleaning step into BIANCA thus obtaining, together with the “optimised” setting, the BIANCA-MS pipeline. We then compared BIANCA and BIANCA-MS performances. Briefly, we separately trained BIANCA and BIANCA-MS on the training set of each dataset and then ran the trained algorithms on the corresponding test set. As the cleaning procedure was meant to be an integrative part of the new BIANCA-MS, manual masks were not cleaned when compared to BIANCA outputs. Instead, to allow a fair comparison and to reduce the inter-rater variability in lesions contouring, BIANCA-MS outputs were compared to the manually outlined masks being processed with the cleaning procedure.

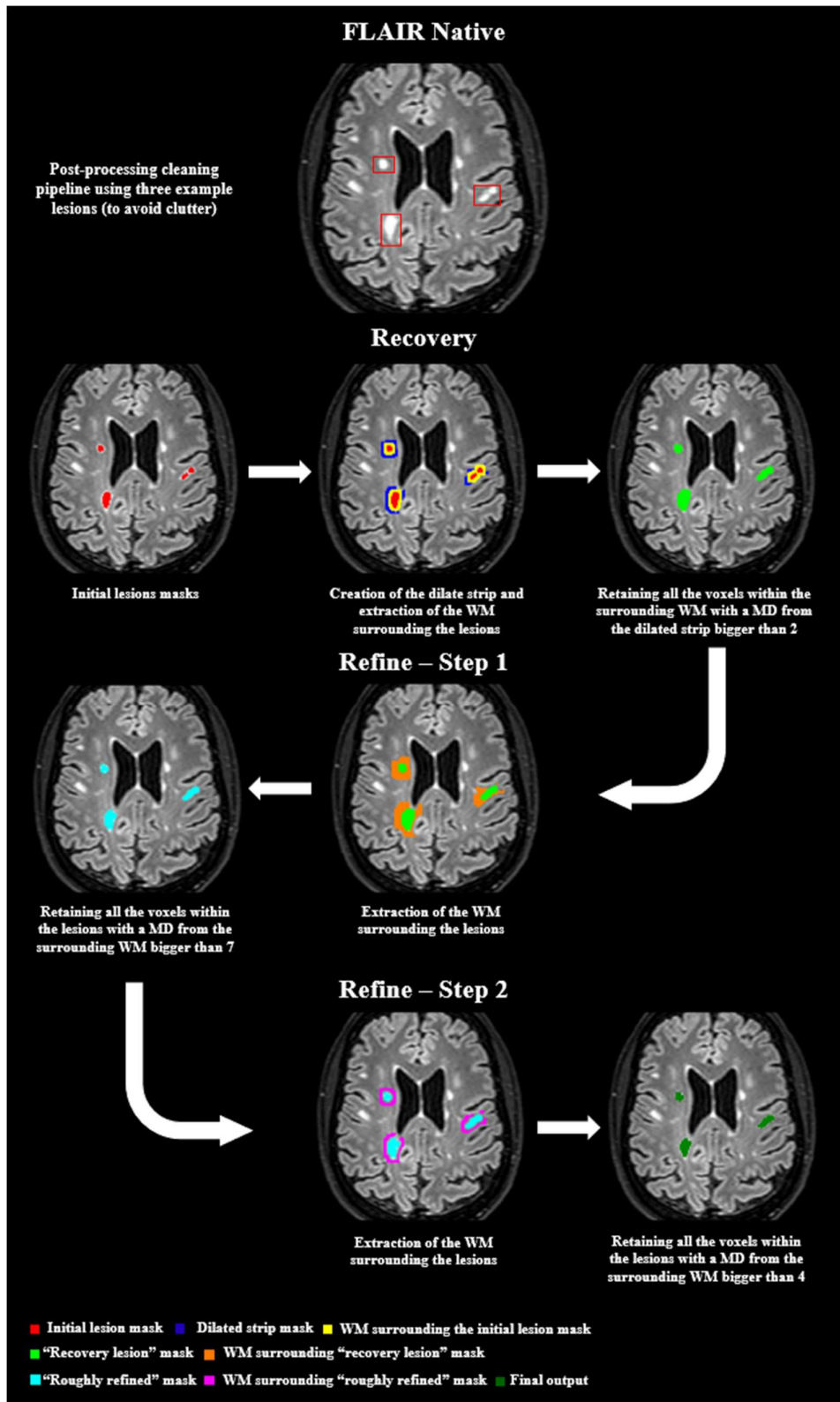


Figure 1. Illustration of the cleaning step pipeline using three example lesions, to avoid clutter (example lesions shown with red boxes in the top image). The original lesion mask underwent an initial "Recovery" phase, followed by a two steps "Refinement" process. For each phase, the corresponding surrounding WM is depicted. Note MD = Mahalanobis Distance, WM = White Matter

BIANCA-MS Validation

To validate BIANCA-MS, all the analyses were performed on the test set of each dataset. During algorithm validation, the inter-rater bias was handled into two different ways. First, BIANCA-MS was trained with the manual lesion masks being not cleaned as the source of variability across raters could ensure a training procedure as heterogeneous as possible. Second, on the test set of each dataset, the cleaning procedure was applied to the manual lesions masks to reduce the inter-rater bias impact on algorithm performance.

Comparison of BIANCA-MS with existing approaches

Firstly, we investigated whether the performance of BIANCA-MS was similar to those obtained with other existing approaches. Three different automated lesion segmentation algorithms were deployed:

- Lesion growth algorithm (LGA) (Schmidt, et al., 2012) from SPM12 (<https://www.applied-statistics.de/lst.html>): the algorithm first segments the T1-weighted images into the three main tissue classes (CSF, GM and WM). This information is then combined with the co-registered FLAIR intensities in order to calculate lesion belief maps. By thresholding these maps with a pre-chosen initial threshold (κ) an initial binary lesion map is obtained which is subsequently grown along voxels that appear hyperintense in the FLAIR image. The result is a lesion probability map. In this work, the κ -value was set to 0.3 as previously suggested (Schmidt et al 2012).
- Lesion prediction algorithm (LPA) (Schmidt, 2017) from SPM12 (<https://www.applied-statistics.de/lst.html>): LPA requires a FLAIR image only (although also T1-weighted images can be provided to the algorithm) and does not require the initial thresholding κ -value. This algorithm consists of a binary classifier in the form of a logistic regression model trained on the data of 53 MS patients with severe lesion patterns. As covariates for this model a similar lesion belief map as for the lesion growth algorithm (Schmidt, et al.,

2012) was used as well as a spatial covariate that considers voxel specific changes in lesion probability. Parameters of this model fit are used to segment lesions in new images by providing an estimate for the lesion probability for each voxel. No pre-processing is applied, because the algorithm performs the necessary bias field correction and affine registration of T1 to FLAIR images as part of the pipeline.

- nicMS (Valverde, et al., 2017; Valverde, et al., 2019): the algorithm is a deep learning method based on cascaded convolutional neural networks (CNN) that, in contrast to most supervised machine learning or deep learning methods, can be used when limited amounts of manual input data are available. As for BIANCA-MS, nicMS was trained on the training set. No pre-processing was required.

As the cleaning procedure was developed as a BIANCA-MS component, SPM tools and nicMS performance were evaluated in comparison with the manual lesion masks being not cleaned. For each of these tools we performed threshold adjustment by varying the probability threshold from 0 to 1 with step size of 0.1 and choosing the value that gave the highest degree of similarity with the manual masks in the test sets. Analyses were performed on Datasets 1 and 2, as these tools were developed to work using FLAIR sequences.

BIANCA-MS behaviour across datasets

Second, we tested how BIANCA-MS behaves for each dataset individually. Thus, BIANCA-MS was separately trained on each dataset, and for each training case it was run only on the corresponding test set. This provided for each dataset one set of performance measures, which we then compared to test the relative performance achieved across the different datasets.

BIANCA-MS segmentation using mixed training sets

Third, using images acquired in different centres, but with similar acquisition protocols, (Dataset 3) we investigated the influence of using mixed training and test sets on BIANCA-

MS performance. Therefore, the selection of subjects in training, validation and test sets for Dataset 3 was made in 2 distinct ways:

- non-stratified: sampling is mixed across all centres (as used for the previous analysis).
- stratified by centre: the subjects selected for the training and validation subsets belonged to different centres from those used for the test set (a leave-one-centre out approach). This is used to test cross-centre generalisation.

We separately trained BIANCA-MS on the training set of both the stratified and non-stratified dataset. For each training case, we ran BIANCA-MS on the corresponding test set and compared the performances achieved on the stratified and non-stratified datasets.

Validation on pooled MRI dataset

Finally, we created a fourth dataset (i.e. a global dataset) by merging the three datasets. All the images from the three training sets were used for training BIANCA-MS, whereas the MRI scans from all the different test sets were used for evaluating the algorithm. To ensure we used a consistent training and validation procedure, it was necessary to provide to BIANCA-MS a fixed set of modalities. Thus, we decided to include in the analyses FLAIR and T1-weighted images as they are present in both Dataset 1 and Dataset 2. For Dataset 3, we created artificial “PseudoFLAIR” images (figure 2) by using the following formula (Battaglini, De Stefano, & Jenkinson, 2012):

$$PseudoFlair: \frac{2(PD * T2)}{PD + T2} * T1$$

After training the algorithm, we then evaluated the performance of BIANCA-MS on the combined test set.

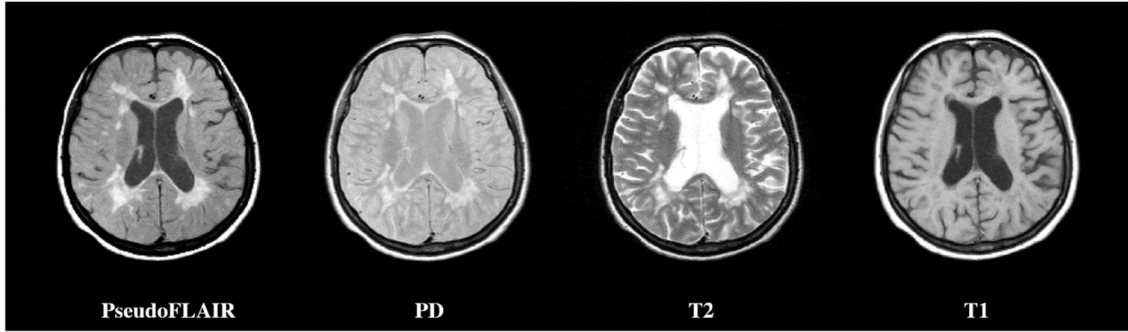


Figure 2. Example of PseudoFLAIR as obtained from PD, T2 and T1 on one subject for Dataset 3.

Performance Evaluation

To test the sensitivity, specificity and accuracy of the algorithm, three different metrics were evaluated: number of false positive clusters (nFPC), defined as the number of clusters incorrectly labelled as a WM lesion; number of false negative clusters (nFNC), defined as the number of clusters incorrectly labelled as non-WM lesion; DICE spatial similarity index (SI) defined as

$$SI = \frac{2 * Tp}{2 * Tp + Fp + Fn}$$

where Tp, Fp and Fn are the true positive, false positive and false negative WM lesion voxels respectively. All these metrics were assessed in comparison to manual segmentation.

Statistical analyses

BIANCA-MS

Optimization step: For each dataset, the 108 different settings were ranked accordingly to the SI, nFPC and nFNC achieved on the validation set. As in the original work of BIANCA (Griffanti, et al., 2016), the SI index was considered the main metric for determining the rankings. In cases where the SI index was equal, a higher priority for the ranking was then given to nFNC compared to nFPC, as we are interested in achieving high sensitivity for lesion detection. Finally, the setting that had a high ranking across all the datasets was retained and considered as the common optimal setting.

Post-processing cleaning step: On the five high resolution FLAIR images that had been segmented by two raters, two kinds of analyses were performed.

- *Qualitative analysis:* to ensure that the proposed approach did not influence the quality of lesion segmentation, a third rater (RC, a neurologist who is an expert in MRI analysis) blindly assessed the lesion masks manually outlined by the 2 raters and those obtained after the cleaning procedure is applied. A total of 10 pairs of masks (manual without cleaning versus manual with cleaning) were assessed, with the third rater assigning to each pair a “winner” (i.e. the mask that is best at outlining the real lesions).
- *Quantitative analysis:* the SI indices measured between the lesion segmentations from each rater, on the same subject, both with and without the cleaning step were compared using a paired t-test.

To evaluate the influence of the cleaning procedure on BIANCA outputs on the three test sets, we compared the performances achieved with and without the refine step using a Wilcoxon signed rank test.

BIANCA-MS Validation

Comparison with existing approaches: SI, nFPC and nFNC values obtained by BIANCA-MS and the 3 software tools that we tested were separately compared using the Kruskal-Wallis test followed by post hoc tests with Bonferroni correction. The volumetric correlation between each tool outputs and the manually segmented masks was assessed using Pearson coefficients.

BIANCA-MS behaviour across datasets: SI, nFPC and nFNC values obtained by BIANCA-MS on the 3 datasets were compared using the Kruskal-Wallis test followed by post hoc tests with Bonferroni correction.

BIANCA-MS segmentation using mixed training sets: the SI, nFPC and nFNC values obtained by BIANCA-MS on the third dataset, with and without data stratification per centre for the training and test sets, were compared using the Wilcoxon test.

Validation on the pooled MRI dataset: SI, nFPC and nFNC values obtained on the pooled dataset were compared to the ones achieved by BIANCA-MS on each separate dataset using the Wilcoxon test.

All the analyses were performed using MATLAB. Statistical significance was considered when p values were < 0.05 .

Results

BIANCA Optimization

Table 3 shows the values for each BIANCA parameter that we decided to vary in order to optimize the algorithm. The best five ranked option combinations are listed in table 4. The setting with SW value of 5, different number of training points for WMH (2000) and non WMH (10000) classes, local average intensity within a kernel of size of 3 voxels and the absence of any location preferences in non-lesion training points is the first one to be shared between the 3 datasets. This setting was indicated as the “optimal setting”, which demonstrated that was the least dependent on the acquisition protocols and therefore should be able to be used with less variation across datasets and hence be employed for all datasets without further adaptation required.

	Dataset 1					Dataset 2					Dataset 3				
Setting	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Ranking															
SW	5	1	1	10	1	1	1	1	5	1	5	5	10	10	5
PS	3	6	9	3	3	9	6	None	3	None	3	None	3	3	6
LTP	Any	Any	Any	Any	Surround	Any	Any	Any	Any	Surround	Any	No Border	No Border	Any	Any
NTP	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000	2000- 10000
SI	0.48±0.17	0.48±0.17	0.47±0.17	0.47±0.17	0.46±0.17	0.44±0.18	0.42±0.18	0.42±0.17	0.41±0.17	0.41±0.18	0.51±0.16	0.5±0.15	0.5±0.15	0.5±0.16	0.5±0.16
nFPC	23±15	23±38	25±43	19±10	15±18	53±37	48±31	69±36	26±10	28±15	31±20	49±21	26±13	27±13	25±13
nFNC	2±4	3±4	2±4	3±4	5±5	2±4	2±5	2±4	3±7	5±7	6±5	5±5	7±5	7±6	8±6

Table 4. Rankings of the best five BIANCA settings found based on performance in the validation set of each dataset. The common optimal setting is highlighted in bold. SW = spatial weighting, PS = Patch size, LTP = Location of WMH training points, NTP = Number of WMH/non-WMH training points. SI = Dice similarity Index, nFPC = number of false positive cluster, nFNC = number of false negative clusters

Post-processing cleaning step

Qualitative analysis: according to the third rater, no relevant differences were detected between the manually segmented and the cleaned lesion masks, suggesting the two segmentation outputs are indistinguishable (5/10 were better with cleaning).

Quantitative analysis: the implementation of the post-processing cleaning step greatly reduced the inter-operator variability and strongly increased the spatial overlap between the two raters (SI without cleaning: 0.75 ± 0.02 ; SI with cleaning: 0.87 ± 0.02 . $p < 0.01$) (Figure 3). Further, the implementation of the cleaning step greatly improved BIANCA performances (Figure 4, table 5). For all the three datasets, significantly higher SI was observed when the post processing cleaning step is introduced, compared to when BIANCA outputs are not refined ($p < 0.01$). Moreover, the use of the cleaning step was found to be an effective method for increasing algorithm precision (i.e. reducing nFPC, $p < 0.01$). Finally, the implementation of the cleaning procedure slightly increased the algorithm sensitivity (i.e. reduced the nFNC, $p < 0.01$).

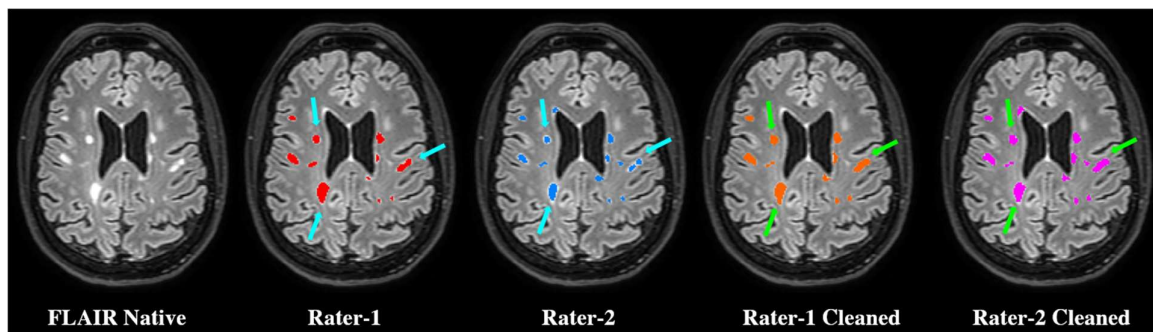
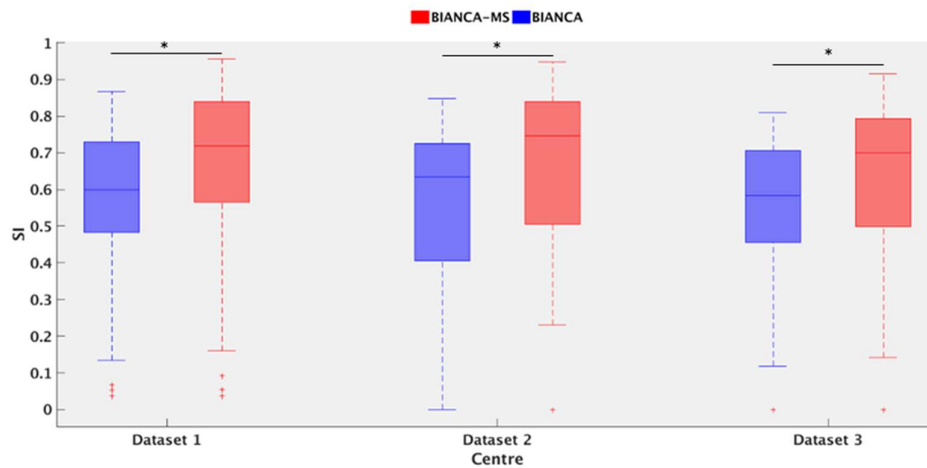


Figure 3. Example of lesion segmentation outputs obtained by two raters without (Red and Blue, in the second and third columns) and with the post processing cleaning step (Orange and Magenta, in the fourth and fifth columns). Light Blue arrows indicate the regions where the raters showed differences in lesion contouring. These differences are reduced (green arrows) when a cleaning step is introduced.

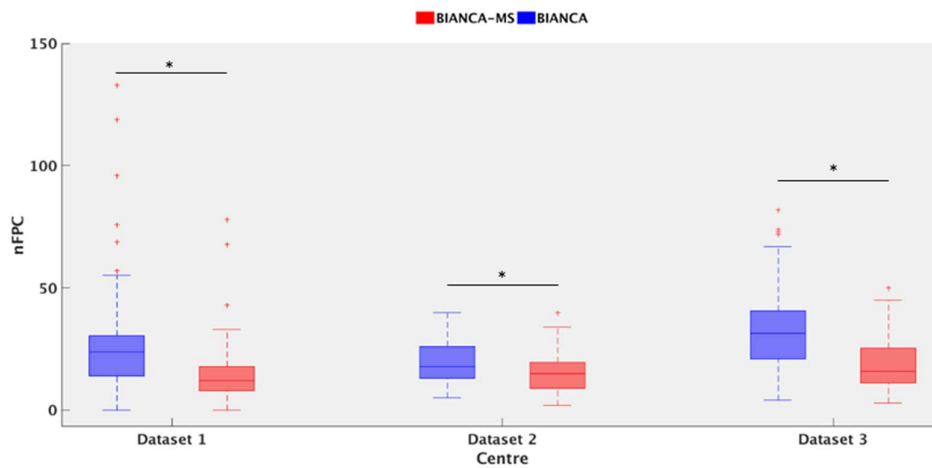
		SI	nFPC	nFNC
Dataset 1	BIANCA	0.6±0.18	24±21	4±8
	BIANCA-MS	0.72±0.2*	12±12*	3±6*
Dataset 2	BIANCA	0.63±0.22	18±9	4±10
	BIANCA-MS	0.75±0.23*	15±8*	3±8*
Dataset 3	BIANCA	0.58±0.18	31±19	10±11
	BIANCA-MS	0.7±0.21*	16±11*	8±11*

Table 5. BIANCA versus BIANCA-MS comparison. Note that BIANCA performances were assessed with respect to lesions masks not cleaned, whereas BIANCA-MS outputs were compared with the manually outlined masks being cleaned. * Indicates results where the implementation of the refine procedure significantly ($p < 0.05$) altered tool performance

A



B



C

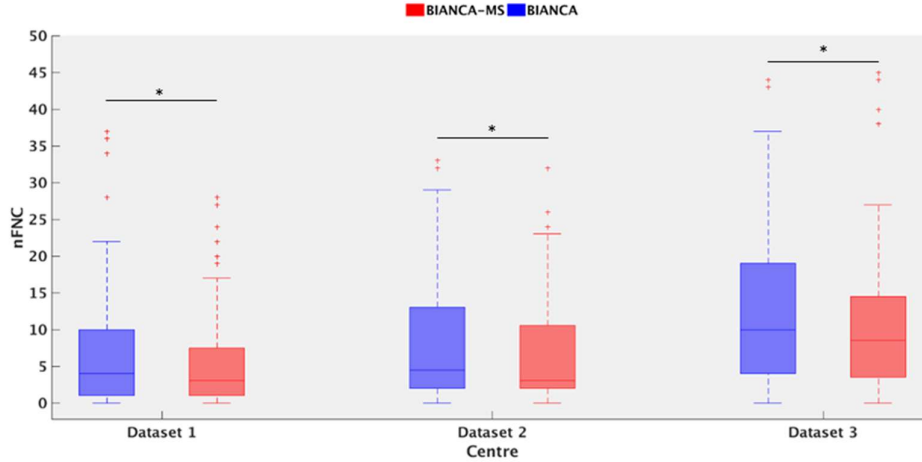


Figure 4. Boxplots of BIANCA (blue) and BIANCA-MS (red) performance measures showing the SI (A), nFPC (B) and nFNC (C) obtained using the test sets of each dataset. Note that BIANCA performances were assessed with respect to lesions masks not cleaned, whereas BIANCA-MS outputs were compared with the manually outlined masks being cleaned * indicates results where the implementation of the refine procedure significantly ($p < 0.05$) altered tool performance

Comparison with existing approaches

To allow a fair comparison, only the datasets where FLAIR images were provided have been analysed. For this reason, Dataset 3 was not included in this experiment.

Dataset 1: The optimal lesion probability threshold was 0.4 for nicMS, 0.5 for LPA and 0 (no threshold) for LGA. Figures 5 and 6 show examples where all the evaluated tools provided optimal and suboptimal lesion segmentation for a single subject. On the test set, the tools showed significant different SI values (BIANCA-MS: 0.72 ± 0.2 ; LGA: 0.33 ± 0.18 ; LPA: 0.56 ± 0.18 ; nicMS: 0.67 ± 0.23 ; $p < 0.01$). A post-hoc test revealed how BIANCA-MS showed the highest SI ($p < 0.01$). Similarly, overall differences across tools were achieved for nFNC (BIANCA-MS: 3 ± 6 ; LGA: 16 ± 16 ; LPA: 9 ± 11 ; nicMS: 5 ± 7 ; $p < 0.01$). A post-hoc test revealed how BIANCA-MS showed the lowest nFNC ($p < 0.01$). Significantly different nFPC were achieved across the different tools: (BIANCA-MS: 12 ± 12 ; LGA: 1 ± 3 ; LPA: 7 ± 10 ; nicMS: 8 ± 7 ; $p < 0.01$). A post-hoc test revealed how LGA showed the lowest nFPC ($p < 0.01$). Finally, BIANCA-MS and nicMS showed the highest volumetric correlation (i.e. Pearson coefficients) with the manually outlined masks (BIANCA-MS: 0.97; nicMS: 0.97; LPA = 0.92; LGA = 0.86).

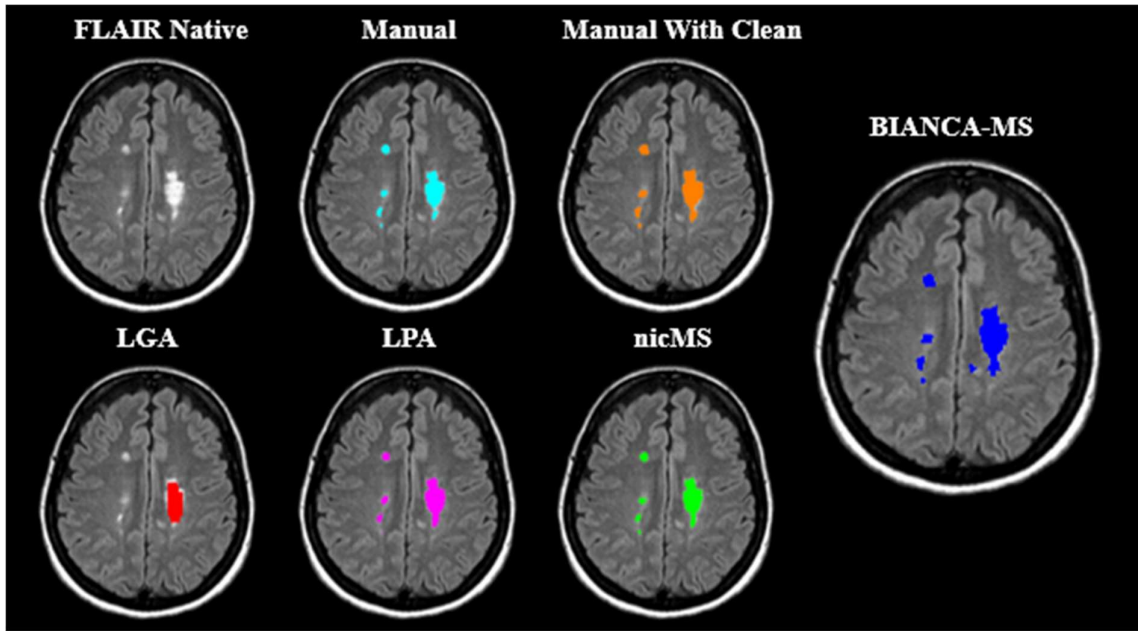


Figure 5. Example of optimal lesion segmentation using the different lesion segmentation tools on the same subject from the test set of dataset 1: manual without (light blue) and with clean (orange); BIANCA-MS (blue, $SI : 0.9$, $nFPC : 8$, $nFNC : 3$), LST-LGA (red, $SI : 0.55$, $nFPC : 2$, $nFNC : 19$), LST-LPA (magenta, $SI : 0.76$, $nFPC : 9$, $nFNC : 10$) and nicMS (green, $SI : 0.83$, $nFPC : 12$, $nFNC : 5$).

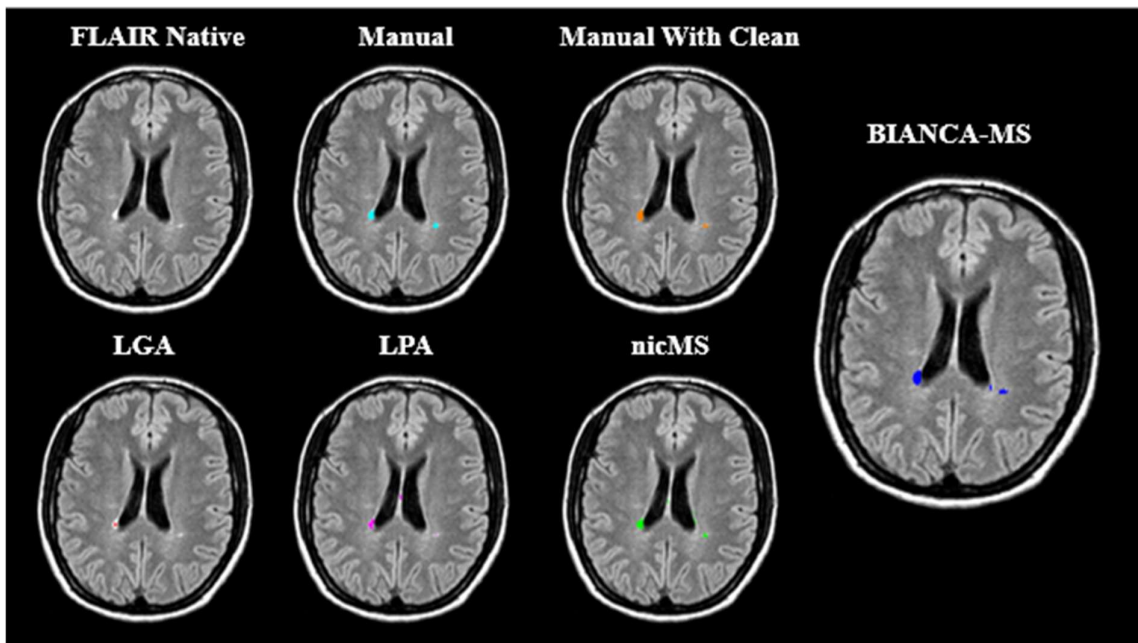


Figure 6. Example of suboptimal lesion segmentation using the different lesion segmentation tools on the same subject from the test set of dataset 1: manual without (light blue) and with clean (orange); BIANCA-MS (blue, $SI : 0.59$, $nFPC : 14$, $nFNC : 7$), LST-LGA (red, $SI : 0.02$, $nFPC : 0$, $nFNC : 16$), LST-LPA (magenta, $SI : 0.33$, $nFPC : 6$, $nFNC : 9$) and nicMS (green, $SI : 0.5$, $nFPC : 15$, $nFNC : 1$).

Dataset 2: The optimal lesion probability threshold was 0.3 for nicMS, 0.4 for LPA and 0.2 for LGA. Figures 7 and 8 show examples where all the evaluated tools provided optimal and suboptimal lesion segmentation for a single subject. On the test set, the tools showed significant different SI values (BIANCA-MS: 0.75 ± 0.23 ; LGA: 0.59 ± 0.24 ; LPA: 0.59 ± 0.21 ; nicMS: 0.71 ± 0.3 ; $p < 0.01$). A post-hoc test revealed how BIANCA-MS showed the highest SI ($p < 0.01$). Significant differences were achieved for nFNC (BIANCA-MS: 3 ± 8 ; LGA: 14 ± 19 ; LPA: 8 ± 11 ; nicMS: 6 ± 4 ; $p < 0.01$). A post-hoc test revealed how BIANCA-MS and nicMS showed the lowest nFNC ($p < 0.01$). Significantly different nFPC were achieved across the different tools: (BIANCA-MS: 15 ± 8 ; LGA: 6 ± 5 ; LPA: 18 ± 21 ; nicMS: 5 ± 8 ; $p < 0.01$). A post-hoc test revealed how LGA and nicMS showed the lowest nFPC ($p < 0.01$). Finally, nicMS and LGA showed the highest volumetric correlation with the manually outlined masks (BIANCA-MS: 0.98; nicMS: 0.99; LPA = 0.97; LGA = 0.99).

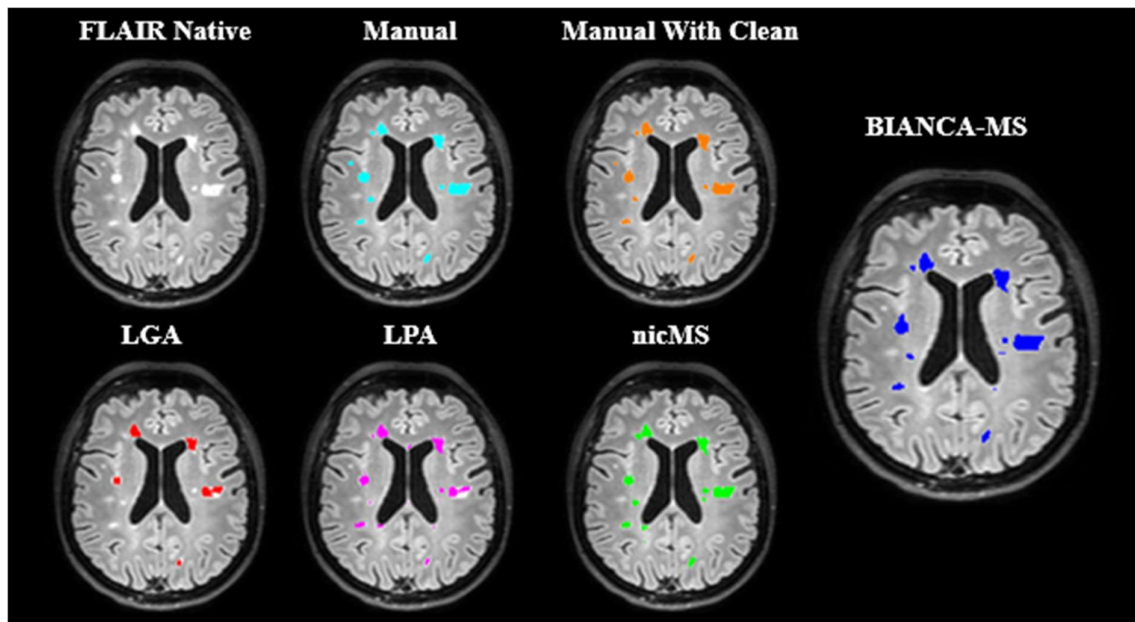


Figure 7. Example of optimal lesion segmentation using the different lesion segmentation tools on the same subject from the test set of dataset 2: manual without (light blue) and with clean (orange); BIANCA-MS (blue, SI : 0.94, nFPC : 6, nFNC : 23), LST-LGA (red, SI : 0.69, nFPC : 2, nFNC : 31), LST-LPA (magenta, SI : 0.73, nFPC : 18, nFNC : 11) and nicMS (green, SI : 0.78, nFPC : 11, nFNC : 5).

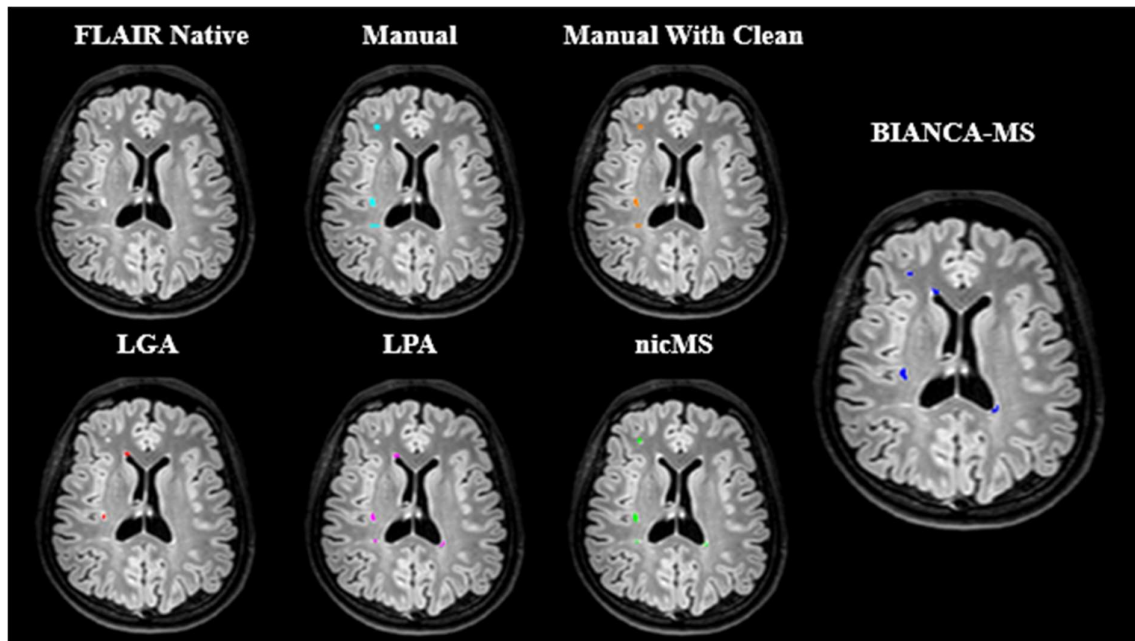


Figure 8. Example of suboptimal lesion segmentation using the different lesion segmentation tools on the same subject from the test set of dataset 2: manual without (light blue) and with clean (orange); BIANCA-MS (blue, SI : 0.62, nFPC : 12, nFNC : 4), LST-LGA (red, SI : 0.18, nFPC : 7, nFNC : 44), LST-LPA (magenta, SI : 0.36, nPC : 31, nFNC : 22) and nicMS (green, SI : 0.55, nFPC : 7, nFNC : 9).

BIANCA-MS behaviour across datasets

When BIANCA-MS was run on the three test sets, no differences were found for SI (Dataset 1: 0.72 ± 0.2 ; Dataset 2: 0.75 ± 0.23 ; Dataset 3: 0.7 ± 0.21). Significant different nFNC were achieved across datasets (Dataset 1: 3 ± 6 ; Dataset 2: 3 ± 8 ; Dataset 3: 8 ± 11 ; $p < 0.01$). A post-hoc test revealed how the highest nFNC were achieved on Dataset 3 ($p < 0.01$). BIANCA-MS showed different nFPC across datasets (Dataset 1: 12 ± 12 ; Dataset 2: 15 ± 8 ; Dataset 3: 16 ± 11 ; $p < 0.01$). A post-hoc test revealed increased nFPC in Dataset 3 compared to Dataset 1 ($p < 0.01$).

BIANCA-MS segmentation using mixed training sets

No statistically significant differences were found for SI, nFPC and nFNC when BIANCA-MS was trained on the stratified versus the non-stratified datasets (Table 6).

	SI	nFPC	nFNC
Non-stratified			
Dataset 3	0.7±0.21	16±11	8±11
Stratified Dataset 3	0.68±0.18	17±12	12±13

Table 6. Comparison of BIANCA-MS performances measures using the stratified and non-stratified sets extracted from Dataset 3 (to test for cross-centre generalisation). Note that none of the statistical tests were significant here

Validation on pooled MRI dataset

No statistically significant differences were found in BIANCA-MS SI and nFNC measures when the tool was trained and validated using pooled sets of images acquired with different scanning protocols (SI: 0.70±0.21, nFNC: 5±11) and when each dataset is separately analysed (median values across datasets, SI: 0.72±0.21; nFNC: 4±9). Slightly higher nFPC were achieved on the pooled dataset (global Dataset: 20±14; median value across datasets, nFNC: 14±11; $p < 0.01$).

Discussion

In this work we presented BIANCA-MS, a novel automated procedure for the segmentation of WM brain lesions in MS. This tool has been validated on MR images acquired using different scanners, imaging protocols and resolutions, demonstrating that it is robust, flexible and accurate. With the pipeline proposed here, we wanted to overcome some of the main issues limiting AI tools generalization and their implementation in medical imaging: the identification of a unique set of parameters able to deal with a variety of scanning protocols and the variability in results induced by inter-rater variability.

The first contribution of this study is the unique algorithm setting identified after a large optimization procedure, whose parameters proved to be relatively independent from scanning protocol and reflected typical MS lesion features (SW = 5, PS = 3, LTP = any, NTP WMH/non-WMH= 2000/10000). The selection of a value of 5 for the SW option was most likely needed due to the specific regional distribution of MS lesions (Filippi, et al., 2019). The absence of any regional preferences in non WMH-voxels (LTP option) could be due to the variability of MS lesion borders, which could be either sharp or ill-defined (Lucchinetti, et al., 2000). Since the inflammation process could affect different brain regions, information from the anatomic context around lesions (PS option) proved to improve algorithm accuracy (Lao, et al., 2008). The imbalance in the number of training voxels between WMH and non-WMH was related to the wide heterogeneity of non-WMH voxels. Thus, a higher number of non-WMH voxels was needed for better depicting non-lesional tissues. Importantly, the set of options proved to be relatively independent of the acquisition protocol, the set of images provided, image resolution and main modality of reference, based on the fact that it ranked highly for all datasets. Therefore, the use of this harmonized setting could avoid the complex and time-consuming optimization procedures needed for adapting algorithm parameters to each dataset. In the original work of BIANCA, different optimal settings were found (Griffanti, et al., 2016). This

is likely related to the different spatial location, shape and contrast between WMH of presumed vascular origin and MS lesions.

Another crucial and innovative step of the BIANCA-MS pipeline is the cleaning procedure. The implementation of this approach was motivated by the evidence of the inter-rater bias impact on segmentation (Shwartzman, Gazit, Shelef, & Riklin-Raviv, 2019). This is of great relevance if we consider that automated tools are usually validated in comparison to manual segmentation. Such procedure is highly subjective and difficult to reproduce, thus resulting in insufficiently reliable gold standard. Even in standardized platforms, that are increasingly being used for the validation of automated lesion segmentation tools (i.e. MICCAI and ISBI), differences across raters are detected. Given this context, obtaining a reliable gold standard makes the automated tools validation process itself a challenge. In this work, the application of the cleaning step significantly increased the concordance across raters, thus reducing the source of variability when validating BIANCA-MS. Moreover, the absence of any relevant difference between the lesion masks with and without the cleaning procedure suggested that our pipeline reduced the inter-rater variability without influencing the overall quality of the initial segmentation. Finally, our cleaning approach greatly improved BIANCA accuracy, precision and slightly affected its sensitivity. Several strategies have been proposed to refine lesion segmentation, including the use of lesion location information (Datta & Narayana, 2013), continuity across slices (Abdullah, Younis, Pattany, & Saraf-Lavi, 2011), ratio maps across modalities (Sajja, et al., 2006) and classification of FP as outlier clusters far from lesion and not lesion tissues (Lao, et al., 2008). Other approaches simultaneously employed information coming from different sources (Ganna, Rombaut, Goutte, & Zhu, 2002; Khastavanehm & Haron, 2014; Abdullah, Younis, Pattany, & Saraf-Lavi, 2011; Roura, et al., 2015; Battaglini, et al., 2014). A point of strength of our pipeline relies on the objective analysis of intensity distribution of the pure tissues surrounding the lesions in both 3D and 2D, which solves the

problem of local inhomogeneities strongly influencing lesion segmentation. Although future studies are needed to further test the reliability of the approach, the large variability in MRI data analysed here and the results obtained demonstrate the apparent robustness of our approach.

BIANCA-MS clearly outperformed all the existing tools that we evaluated in terms of SI in both high and low-resolution images (Dataset 2 and 1 respectively). LGA achieved the lowest SI value whereas nicMS and LPA ranked second and third respectively. Further, BIANCA-MS achieved on both datasets the lowest nFNC thus proving the high sensitivity of the approach. LGA showed the highest nFNC, providing overall the worst performance across all the tools. BIANCA-MS showed lower precision in lesion detection (i.e. higher nFPC) in comparison to nicMS and LGA. However, the higher SI achieved by our approach suggested that the higher nFPC did not seem to alter the segmentation accuracy. On both high and low-resolution images, BIANCA-MS and nicMS showed constantly high volumetric correlation with the manually outlined masks, whereas LGA and LPA showed relatively lower degree of concordance on 2D acquired images. Taken together, our findings suggested that BIANCA-MS provided overall the best performance across all the evaluated tools. Noteworthy, these results could be influenced by the way we selected tools optimal thresholds: we used SI as main metrics, giving more importance in achieving lower nFNC (i.e. higher sensitivity) than lower nFPC (i.e. higher precision). Thus, different selection criteria might provide different results. Importantly, the nicMS and LST tools require FLAIR images as a predefined reference modality; thus, although artificial pseudoFLAIR images were provided, Dataset 3 was excluded to allow a fair comparison across tools. BIANCA-MS is very flexible in terms of reference modality and can perform segmentation using any set of images provided. Further, several studies have reported variations in algorithm performances across datasets (Griffanti, et al., 2016; Roura, et al., 2015; Guo, et al., 2019), even with consistent scanner field strength and after protocol harmonization (Shinohara, et al., 2017). The consistency of the performance measures obtained across datasets

demonstrated the robustness and flexibility of BIANCA-MS when different acquisition protocols, image modalities and resolutions are provided to the algorithm.

It is also worth noting that BIANCA-MS performed slightly better (although not significantly) on FLAIR-3D images. Algorithms are validated using mostly images with 2D/slice-wise acquisitions, with only a small number of images with 3D acquisitions employed, often referring to online databases. In this respect we did not limit our analyses to images with 2D acquisitions, but we validated BIANCA-MS, to the best of our knowledge, on the biggest private dataset of MS subjects where images with 3D acquisitions have been manually segmented. The results achieved are of the utmost relevance if we consider that high resolution FLAIR-3D sequences are now the preferable acquisition due to their high sensitivity in lesion detection (Filippi, et al., 2019; Paniagua Bravo, et al., 2014), and thus in the future such high-resolution images will be more and more commonly acquired.

Another key observation in our study is that training BIANCA-MS on stratified or non-stratified data did not influence the performance of the algorithm. In clinical trials the accrual of WM lesions is one of the most commonly used MRI outcomes (van Munster & Uitdehaag, 2017) and images are acquired from different centres. Thus, an automatic segmentation tool that provides accurate and robust segmentation of WM lesions in multicentre data is greatly needed. The results of this study highlight how BIANCA-MS is insensitive to data stratification per centre, making it easier to apply in clinical trials where training with data from the same centre is difficult or impossible, whereas mixed training sets are straightforward. In a recent work, a similar cross validation approach was performed (Gabr, et al., 2020). Our results are in line with those obtained by Gabr and colleagues, however we did not focus only on SI, but we also included nFPC and nFNC as further analysis metrics.

Finally, the introduction of PseudoFLAIR images allowed the creation of a global multicentre dataset. When trained on this unified dataset, the performance of BIANCA-MS was comparable

to the ones achieved when separately trained on each centre. Lesion segmentation across heterogeneous acquisition protocols is a challenging task, with studies reporting from poor (Heinen, et al., 2019) to moderate reproducibility across centres (de Sitter, et al., 2017). Recently, an AI tool demonstrated high consistency across a wide range of imaging parameters (Duong, et al., 2019). However, the great amount of training data (n=295) needed to achieve such performance could in part limit its use in a real-world setting. Pooling MRI data represents a very practical solution for increasing both the external validity and transposability of research findings to clinical settings (Heinen, et al., 2019). Further, training AI tools on more heterogeneous data increased the performance in out of distribution (OOD) MRI data (Mårtensson, et al., 2020). We hope large multicentre datasets will be more and more employed in the future. The wide training procedure is likely to further the performance of BIANCA-MS for lesion segmentation in OOD MRI data, avoiding the need to be retrained.

This study is not without limitation. Firstly, BIANCA-MS is not completely automatic, as it needs to be re-trained whenever applied to data acquired with different acquisition protocols. However, we showed that a promising solution to this would be training BIANCA-MS on big multicentre datasets. Secondly, we focused on segmenting WM lesions, employing an exclusion mask of cerebellum and gray matter (GM). Usually, these lesions are not easily detectable on conventional MR images (García-Lorenzo, Francis, Narayanan, Arnold, & Collins, 2013). This is certainly matter for further improvement of this promising tool. Future efforts could address the implementation of sequences able to improve GM lesion detection (Nelson, et al., 2007; Geurts, et al., 2005). To be implemented in clinical settings, future studies are necessary to validate BIANCA-MS on both healthy subjects and longitudinal data.

To conclude, in this work we have presented BIANCA-MS, a novel automated procedure developed to overcome some of the issues limiting the generalizability of results achieved by AI tools. Our method clearly outperformed other available tools and proved to be robust,

accurate and flexible across different scanning protocols. Further, the insensitivity of BIANCA-MS to data stratification per centre makes it suitable when a mixed training set is provided, as in clinical trial settings. Finally, pooling MRI data acquired with different scanning protocols did not influence BIANCA-MS performance. This introduces the possibility of obtaining a BIANCA-MS version that is pre-trained on some large datasets and can perform lesion segmentation on OOD MRI data without needing to be retrained. These encouraging results suggested that BIANCA-MS is a promising tool for the segmentation of WM brain lesions in MS.

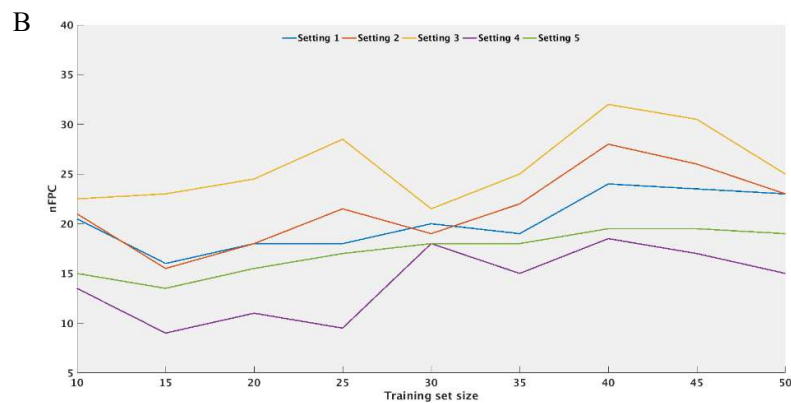
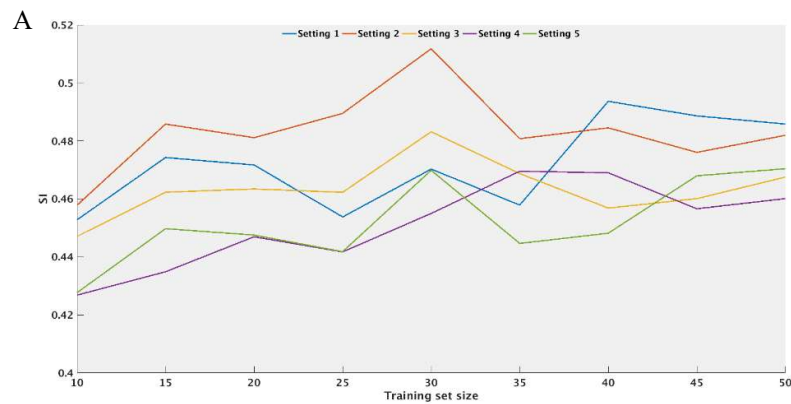
Supplementary Materials

Choice of optimal training size

For determining the optimal number of subjects to include in the training set, training data from Dataset 1 were partitioned into 9 subsets: from 10 (20%) to 50 (100%) subjects with step size of 5. Using a restricted set of algorithm settings ($n=5$), BIANCA was trained separately using these various training subsets and the corresponding performance values were evaluated using the validation set (50 subjects). Finally, a Spearman regression was used for assessing the relation between training size and the BIANCA performance measures (SI, nFPC and nFNC).

	SI	nFPC	nFNC
r	0.8	0.67	-0.86

Table 7. Spearman correlation coefficients between training size and BIANCA performance measures



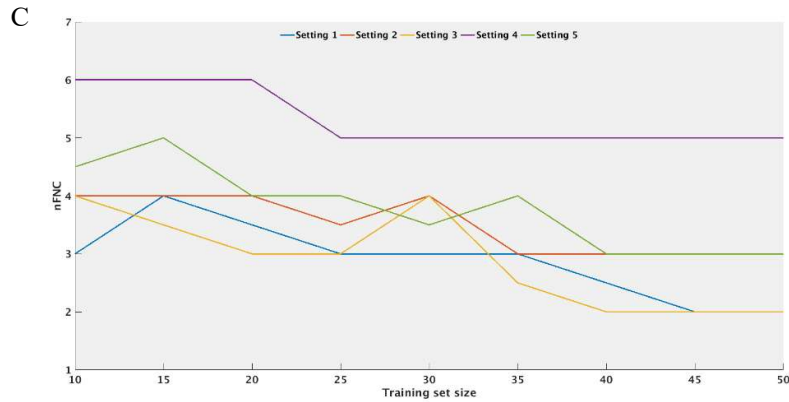


Figure 9. Regression analyses between BIANCA performance measures (A: SI; B: nFPC; C: nFNC) achieved using the validation set and varying the size of the training set. Each line represents a different BIANCA setting.

Results showed a dependence on the training size (Table 7, Figure 9), with performances improving when using bigger training sets. In particular, with training size varying from 10 to 50 subjects, the SI index increased by 5.22% ($r=0.8$, $p< 0.05$; median at 10 vs median at 50: 0.44 vs 0.47), the nFPC increased by 12.19% ($r=0.67$, $p=0.05$; median at 10 vs median at 50: 20.5 vs 23) and the nFNC decreased by 25% ($r=-0.86$, $p< 0.01$; median at 10 vs median at 50: 4 vs 3).

These results are consistent with the finding of another work (Narayana, et al., 2020), where using bigger training sets (≥ 50) led to better segmentation performances. Given this context, we decided to include in the training set a number of subjects (Dataset 1 and 3: 50; Dataset 2: 40) that is likely to be able to be used in a real-world setting allowing, at the same time, meaningful lesion segmentation.

5. Study 2: The concurrent Spatio-temporal relationship between inflammation and neurodegeneration in early Multiple Sclerosis: A Post-hoc Analysis of the REFLEXION Study

This study was performed in collaboration with professor Hugo Vrenken and PhD candidate Rozemarijn Mattiesing* of the Amsterdam VU University Medical Center.

*Equally contributed to this work and share co-first authorship

Introduction

Multiple sclerosis (MS) is a chronic demyelinating disease of the central nervous system (CNS) characterized by the concomitant presence of focal area of inflammation in the white (WM) and gray (GM) matter (lesions) and diffuse damage and neurodegeneration in the entire brain (atrophy) (Lassmann, Brück, & Lucchinetti, 2007). Although these two processes are present early in the disease course of MS, the dynamics of accumulation of WM lesions and brain atrophy is not completely understood.

At current, the few longitudinal MRI studies investigating the relation between WM lesions and brain atrophy have focused on the assumption that inflammation precedes neurodegeneration (Chard, et al., 2003; Paolillo, et al., 2004; Dalton, et al., 2002). However, these two biological processes could represent somewhat unrelated aspects of the disease (Tauhid, Neema, Healy, Weiner, & Bakshi, 2014) working in parallel, with one prevailing over the other at different stages of disease or in different brain regions (Bodini, et al., 2016; Bodini, et al., 2009). Indeed, weak-to-modest associations between the development of regional brain atrophy and lesion changes in number (newly detected lesions) and volume (progressive tissue damage in pre-existing lesions) suggested that lesions damage contribute only partially to brain atrophy (Battaglini, et al., 2009; Cappellani, et al., 2014; Roosendaal, et al., 2011). At present, only few studies investigated how WM lesion changes (activity) and brain volume changes (atrophy) are linked within the same time interval (Richert, et al., 2006; Sailer, et al., 2001). Further, studies within the early phase of MS are even scantly (Varosanec, et al., 2015; Dalton, et al., 2004) and often limited to a single follow-up when looking at the longitudinal spatial correlation (i.e. voxel-wise analyses) between WM damage and brain atrophy (Raz, et al., 2010; Rocca, et al., 2016). To develop more targeted therapeutic strategies which can effectively intervene in the early stage of disease, it is crucial to better understand the underlying disease mechanisms, how these relate to disease progression and whether these can be either modified by treatment or

disease worsening. Thus, it is of great relevance to investigate whether inflammation and neurodegeneration are two independent processes which might develop simultaneously within the early phase of the disease. Further, exploring the concurrent temporal evolution of these two pathological processes will make data interpretation in both clinical and research settings clearer and more straightforward.

Longitudinal MRI studies are mandatory to explore the dynamic associations between WM lesions and brain atrophy. The randomized, double-blind, placebo-controlled, multi-center clinical trial REFLEX and its extension REFLEXION (REbif FLEXible dosing in early MS extensION) provided such opportunity. In this trial, patients presenting with a first clinical demyelinating event (FDCE) were followed over a period of five years with yearly MRI scans. Primary analyses on REFLEX/ION study showed how treatment with interferon β -1a was associated to overall MRI reduced activity (Comi, et al., 2012) (Comi, et al., 2017). However, the link between atrophy and WM lesions was not investigated.

In this study we investigated whether, in the early phase of MS, WM lesions and brain atrophy were spatially interconnected within the same follow-up period and tested whether these two processes developed simultaneously over time. Further, the REFLEX/ION study design provided us the opportunity to investigate how treatment influenced the concurrent relation between inflammation and neurodegeneration. Therefore, we first tested whether WM lesions and brain atrophy were differentially related prior and after treatment onset. Second, we examined on how these two processes were associated within the first year of treatment. Third, we explored how brain atrophy and WM lesions were linked during a stable treatment period. Finally, we assessed whether the relation between inflammation and neurodegeneration differed between patients who converted to MS and those who did not.

Methods

Population

REFLEXION was a preplanned extension of the randomized, double-blind, placebo-controlled, multi-center clinical trial REFLEX. Procedures and the design of the study have been described in detail elsewhere (Comi, et al., 2012) (Comi, et al., 2017). Briefly, patients experiencing a first clinical demyelinating event at high risk of converting to MS and with at least two clinically silent lesions of 3 mm or more on T2-weighted brain MRI scan, at least one of which was ovoid, periventricular, or infratentorial were included. Patients were either randomized to one of the two early treatment (ET) arms, where treatment was initiated with subcutaneous interferon beta-1a (sc IFN β -1a) once a week or three times a week, or to the delayed treatment (DT) arm where, during the first two years (i.e., the REFLEX phase), patients did not receive treatment (placebo group) but at the start of REFLEXION received sc IFN β -1a three times per week (figure 1). Clinically definite MS (CDMS) was defined by a relapse accompanied by an abnormal magnetic resonance imaging (MRI) scan or a sustained increase in EDSS score of ≥ 1.5 points. For each patient, the yearly interval-specific CDMS status was provided. If patients converted to CDMS they switched to open-label treatment with sc IFN β -1a three times per week.

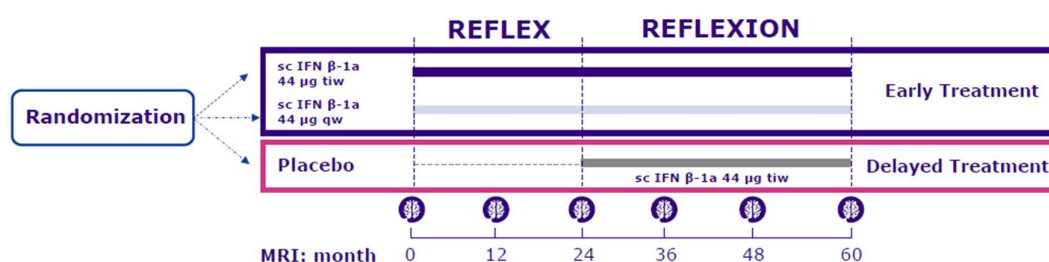


Figure 1. Schematic representation of the REFLEX and its extension REFLEXION study design.

MRI data

For the current post-hoc analyses, multicenter yearly MRI scans over a period of 5 years were provided by PAREXEL International Corporation. These consisted of 1x1x3 mm proton-density- (PD), T2-, T1-, and Gadolinium-enhanced T1-weighted images. The Image Analysis

Center, VU University Medical Center, Amsterdam, The Netherlands, provided manual delineations of the PD/T2-weighted lesions for each yearly visit and manually edited brain extraction masks originally obtained by using FSL BET (Smith, 2002) on T1-weighted images, part of FSL (Smith, et al., 2004). Scans were included if the input data was of sufficient quality and the different processing steps and output of the image analyses described below passed the quality control (specific criteria are described in more detail in the Results section).

MRI analysis

Longitudinal atrophy quantification

The T1-weighted images were corrected for slice-to-slice variations (interleaved acquisition) and subsequently lesion filled with the linearly registered PD/T2-weighted manual delineations. Yearly percentage brain volume change (PBVC) and percentage ventricular volume change (PVVC) were estimated with SIENA (Smith, De Stefano, Jenkinson, & Matthews, 2001) and its extension VIENA (Vrenken, et al., 2014), both part of FSL (Smith, et al., 2004). The normalised and lesion filled T1-weighted image and manually edited brain mask were used as input. PBVC and PVVC were used as longitudinal measures of whole-brain atrophy and central atrophy, respectively.

Longitudinal lesion change quantification

Yearly lesion changes were automatically segmented by an in-house developed method that is based on the use of subtraction images (SI) (Moraal, et al., 2010) (Moraal, et al., 2010). Details of this method are described in a separate work investigating the effect of sc IFN β -1a treatment on WM lesions accrual and the relation between WM lesions distribution and conversion to CDMS in the REFLEX study (Battaglini, et al., In Preparation). Briefly, before creating the SIs, the slice-to-slice variation in signal intensity on the PD-weighted images was corrected. The PD images of both visits of a yearly interval (e.g., baseline-month12) were registered to a common halfway space using a similar procedure to that used in FSL-SIENA software. To

obtain the SIs, the PD-weighted image of the first visit of each interval was subtracted from the second visit. To give a robust analysis, the SIs were further normalized to account for the differences between the sites and MRI scanners from which the images were obtained. Voxels with a normalized intensity difference exceeding 1.5 standard deviations were labeled as changing. Based on the baseline and follow-up lesion masks and the voxel-wise lesion changes, each individual lesion was labelled as new, enlarging, shrinking or disappearing. The yearly total lesion volume change (TLVC) was calculated by subtracting the sum of the negative lesion volume change (shrinking + disappearing) from the positive lesion volume change (new + enlarging) for each interval.

Voxel-wise Input images

The following steps were performed to produce the input images that were submitted to the voxel-wise analyses. First, we created a study-specific template. FSL-SIENAX (Smith, et al., 2002) was used to obtain normalized brain volume (NBV) for all baseline T1-weighted images. Afterwards, 100 patients were selected based on the percentile NBV distribution (from 1st to the 100th percentile). For each of these 100 subjects, the T1-weighted images were intensity normalized (divided by the 99th intensity percentile and multiplied by 10000), non-linearly registered to the MNI standard space (voxel size = 2x2x2 mm³) using FSL-FNIRT (Andersson, Jenkinson, & Smith, 2010) and averaged to create the study-specific template.

Second, to ensure that all the images of a subject underwent the same preprocessing and to avoid interpolation bias, a subject-specific template was created. For each subject, the T1-weighted images were intensity-normalized using the N4 algorithm (Tustison, et al., 2010), linearly registered to the baseline T1-weighted scan using FLIRT (Jenkinson & Smith, 2001; Jenkinson, Bannister, Brady, & Smith, 2002) and averaged to create the subject-specific template. The subject-specific template was then non-linearly registered on the study-specific

template space and the warp-fields generated from this registration were used for the subsequent registration of SIENA outputs on the study-specific template.

To study local atrophy, cerebral edge displacement maps were created. For each subject, the yearly brain edge flow images provided by SIENA were spatially dilated, non-linearly registered to the study-specific template using the warp fields generated through the procedure described above, masked with a standard space brain edge image, smoothed with an isotropic Gaussian kernel with a sigma of 5 mm and remasked (Bartsch, et al., 2007) (De Stefano, et al., 2003)

Study Design

Population subgroups were formed to address at both whole brain and voxel-wise level the following research questions (RQ):

1. Investigate the relation between lesion volume changes and brain atrophy within the same year. All available datapoints were analysed.
2. Test whether treatment could influence this relation (all available datapoints were analysed). Within this respect, we investigated the relation between lesion volume changes and brain atrophy during:
 - a. An untreated period: year 1 and 2 DT patients of the REFLEX period (i.e. placebo) were analysed, while excluding the interval-specific converters. To test whether this relation differed during a treatment period, all the REFLEXION period DT patients (year 4 and 5) were analysed.
 - b. The first year of treatment: year 1 ET and year 3 DT patients were analysed, excluding the DT patients who converted during the first two years of the study (because, as per the study protocol, they received treatment upon conversion, which in their case was earlier than year 3).

- c. A stable treatment period. To prevent the confounding effects by resolving edema and pseudo-atrophy at the start of treatment, the datapoints where the patients have received at least one year of treatment were analysed (ET: year 2, 3, 4 and 5; DT: year 4 and 5).
3. Assess whether the relation between lesion volume changes and brain atrophy differed between converters and non-converters. All available datapoints were analysed.

Statistical analyses

Whole brain: Statistical analyses were performed in Rstudio. Linear mixed models (LMM) were used to deal with the yearly repeated measurements we have for the atrophy and lesion volume change measures. We incorporated a random intercept with a three-level structure where observations are clustered within the patients and the patients are clustered within the different study sites. All LMMs were corrected for age and sex. An alpha of 0.05 was used as the cut-off for significance. Significant interactions were further explored by post-hoc tests. To address RQ 1, we applied a LMM with PBVC/PVVC as the dependent variable and TLVC as the independent variable, while also inserting treatment and the interval-specific CDMS status as additional fixed factors. For RQ 2 and 3, we used similar LMMs but now we also incorporated an interaction between treatment and TLVC and interval-specific CDMS status and TLVC, respectively. To address RQ 2a, we incorporated an interaction between TLVC and period, and corrected for the interval-specific CDMS status in a separate LMM. The same LMM was used for RQ 2b and 2c.

Voxel-wise: For each yearly interval, regional statistical inference was carried out using permutation testing (Nichols & Holmes, 2002) (5000 permutations) as implemented in the FSL “randomise” program (Winkler, Ridgway, Webster, Smith, & Nichols, 2014). Design matrices within the GLM framework were used, with age, sex and site as covariates. Threshold-Free

Cluster Enhancement randomise option was used. To address RQ 1, TLVC was used as regressors with treatment and the interval-specific CDMS status as additional covariates. For RQ 2 and 3, we used similar voxel-wise statistics but now incorporating an interaction between treatment and TLVC and between interval-specific CDMS status and TLVC, respectively. Similar voxel-wise statistics were used for RQ 2a, with the CDMS status being inserted as covariate for the REFLEXION treatment period. The same model used for RQ 2b and 2c. Only results with at least 15 significant voxels ($p < 0.05$) were reported. The anatomical location where brain atrophy significantly correlated with lesion volume changes was determined by using predefined standard space masks (<http://www.fmrib.ox.ac.uk/fsl/>) as provided by the MNI structural atlas. The number (V) and location of significant voxels were reported.

Results

Population

A total of 400 patients enrolled in the REFLEXION study provided MRI data for the extension period and the input data of 392 were included in the current analyses (see table 1). Concerning the input data: 4 subjects were excluded because of incomplete trial data, 2 subjects because of an inconsistent acquisition protocol, 5 visits because of incorrect/incomplete image(s), 8 visits because of movement, 6 visits because of missing data, and 2 visits were excluded from the lesion change analyses because of corrupted PD-weighted images. The quality check of the output from the lesion change quantification and longitudinal atrophy measurements resulted in 158 excluded visits (23 lesion change quantification, 133 longitudinal atrophy, 2 shared rejections). Reasons for exclusion were: low quality of the images, artefacts, registration problems and pipeline failure.

	Converters to CDMS	Non-converters to CDMS	ET	DT	Overall
Subjects (N)	162	230	262	130	392
Gender (F/M)	95/67	147/83	162/100	80/50	242/150
Age (mean±SD)	30.32±7.99	32.23±8.52	31.68±8.43	30.97±8.19	31.44±8.35

Table 1. Demographics of the included patients. CDMS = clinically definite multiple sclerosis (across the whole study period), ET = early treatment, DT = delayed treatment, SD = standard deviation

RQ 1: Relation between atrophy and concurrent lesion volume changes

All the datapoints available from the entire dataset of 392 subjects were analyzed (see table 1 for demographic).

Whole brain: A Significant positive relation between PBVC and TLVC ($B = 0.046$, $SE = 0.013$, $p < 0.001$) was found. A Significant negative relation between PVVC and TLVC ($B = -0.466$, $SE = 0.118$, $p < 0.001$) was found.

Voxel-wise: Results are summarized in table 2. In year 1, faster atrophy was associated with lower TLVC. From year 2 to year 5, faster periventricular atrophy was related to higher TLVC.

Time Interval	Number of significant Voxels	Location of significant voxels
Year 1	4868	PV/FL/PL/TL
Year 2	220	PV
Year 3	2629	PV
Year 4	87	PV
Year 5	121	PV

Table 2. Schematic representation of the Voxel-wise significant results for the concurrent relation between atrophy and TLVC. PV

=Periventricular, FL = Frontal Lobe, PL = Parietal Lobe, TL = Temporal Lobe

RQ 2: Influence of treatment on the relation between atrophy and concurrent lesion volume changes

All the datapoints available from the entire dataset of 392 subjects were analyzed (see table 1 for demographic).

Whole brain: ET and DT patients showed significant different relation between PVVC and TLVC ($B = -0.812$, $SE = 0.24$, $p < 0.001$). A post-hoc test revealed how the relation was significant only for the ET group ($B = -0.781$, $SE = 0.149$, $p < 0.001$).

Voxel-wise: Results are summarized in table 3. In year 1, 4 and 5, ET and DT patients showed significant different relation between atrophy and TLVC. In particular, faster periventricular, infratentorial and frontal lobe atrophy was associated with higher TLVC in ET patients. In year 3, faster periventricular atrophy was related to higher TLVC in both ET and DT patients, but the relation was stronger for DT.

Time Interval	Number of significant Voxels	Location of significant voxels
Year 1	77	INF
Year 3	446	PV
Year 4	172	PV
Year 5	643	PV/FL

Table 3. Schematic representation of the Voxel-wise significant results for the influence of treatment on the concurrent relation between atrophy and TLVC. PV =Periventricular, FL = Frontal Lobe, INF = Infratentorial

RQ 2a: REFLEX placebo versus REFLEXION DT patients

Datapoints from 97 placebo subjects were analyzed (mean age \pm SD: 31.65 \pm 8.25, Number of Male/Female: 36/61). 101 DT subjects were analyzed (mean age \pm SD: 31.06 \pm 8.17, Number of Male/Female: 41/60). 57 patients did not convert to CDMS.

Whole brain: REFLEX period DT patients showed significant positive relation between PBVC and TLVC ($B = 0.072$, $SE = 0.029$, $p = 0.013$). An opposite trend was found between PVVC and TLVC ($B = -0.917$, $SE = 0.306$, $p = 0.003$). No significant relation was reached for the REFLEXION DT patients. Overall, similar relation between WM lesion changes and brain atrophy was observed through REFLEX and REFLEXION period DT patients.

Voxel-wise: Results are summarized in table 4. In year 1, faster atrophy was associated with lower TLVC in placebo patients (figure 2). An opposite trend was found in year 2, where faster periventricular atrophy was related to higher TLVC (figure 2). During a treated period (year 4 and 5), faster periventricular atrophy was related to lower TLVC in DT patients.

Time Interval	Number of significant Voxels	Location of significant voxels
Year 1 (untreated)	2034	PV/PL/TL/INF
Year 2 (untreated)	1464	PV
Year 4 (treated)	223	PV
Year 5 (treated)	1000	PV/FL/TL

Table 4. Schematic representation of the Voxel-wise significant results for the concurrent relation between atrophy and TLVC in REFLEX and REFLEXION period DT patients. PV = Periventricular, FL = Frontal Lobe, PL = Parietal Lobe, TL = Temporal Lobe, INF = Infratentorial

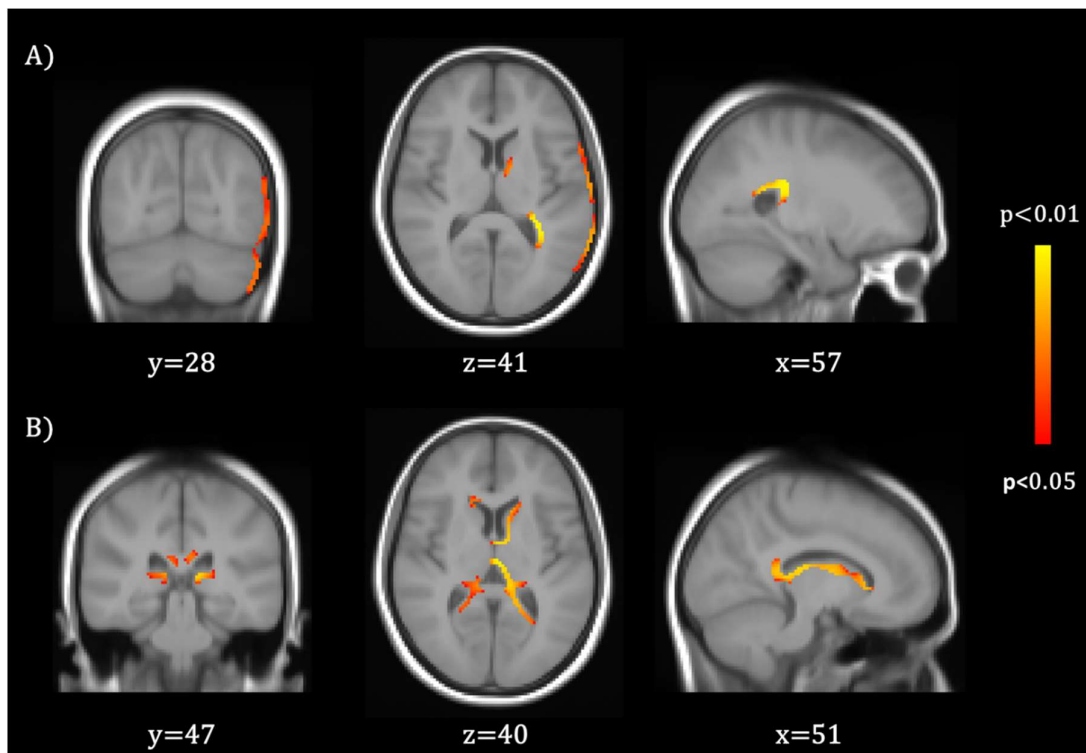


Figure 2. Voxel-wise analyses within the first two years of the untreated placebo period: yellow-orange show voxels of significant regions where lower TLVC was related to faster atrophy (A, top row) and where higher TLVC was related to faster atrophy (B, bottom row)

RQ 2b: First year of treatment

231 ET subjects were analyzed (mean age \pm SD: 31.9 ± 8.38 , Number of Male/Female: 82/149).

208 patients did not convert to CDMS. 65 DT subjects were analyzed (mean age \pm SD: 31.72 ± 7.62 , Number of Male/Female: 24/41). 59 patients did not convert to CDMS.

Whole brain: ET and DT patients showed significant different relation between PBVC and TLVC in the first year of treatment ($B = 0.222$, $SE = 0.07$, $p = 0.002$). A post-hoc test revealed that the direction of the relation was different between ET ($B = 0.081$, $SE = 0.027$, $p = 0.003$) and DT ($B = -0.141$, $SE = 0.065$, $p = 0.032$) patients. Similar results were found for the relation between PVVC and TLVC, with ET and DT showing a significant difference in the first year of treatment ($B = -4.489$, $SE = 0.732$, $p < 0.001$). A post-hoc test revealed that the direction of the relation was different between ET ($B = -1.08$, $SE = 0.284$, $p < 0.001$) and DT ($B = 3.41$, $SE = 0.677$, $p < 0.001$) patients.

Voxel-wise: Within the first year of treatment of the ET patients (year 1), faster periventricular and frontal lobe atrophy was associated with lower TLVC ($V = 2192$, figure 3). No significant relation was found for DT patients in year 3.

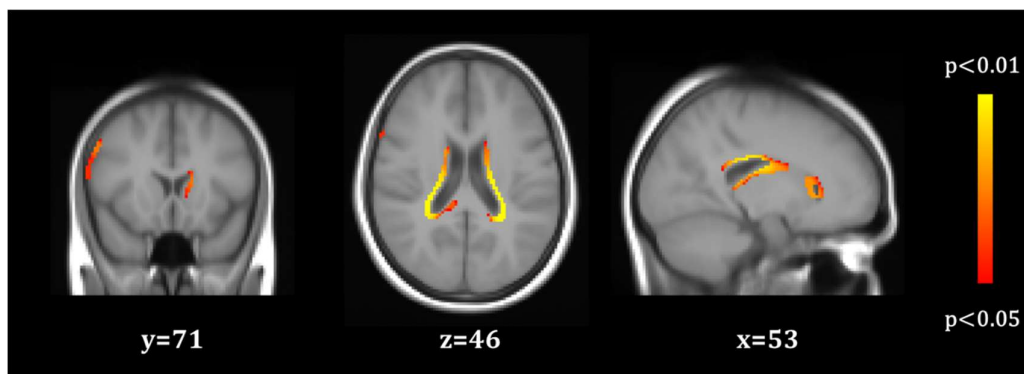


Figure 3. Voxel-wise analyses within the first year of treatment of ET patients (Year 1): yellow-orange shows voxels of significant regions where lower TLVC was related to faster atrophy.

RQ 2c: Stable treatment period

Datapoints from 256 ET subjects were analyzed (mean age \pm SD: 31.66 ± 8.46 , Number of Male/Female: 96/160). 162 patients did not convert to CDMS. Datapoints of the RQ 2a 101 DT patients were used.

Whole brain: The relation between PBVC/PVVC and TLVC did not differ between ET and DT patients.

Voxel-wise: During stable treatment period, faster occipital lobe atrophy was associated with higher TLVC (Year 4 $V = 283$) in ET patients. For DT patients, results are summarized in table 4 (years 4 and 5).

RQ 3 - Influence of conversion to CMDS on the relation between atrophy and concurrent lesion volume changes

All the datapoints available from the entire dataset of 392 subjects was analyzed (see table 1 for demographic).

Whole brain: Converters and non-converters patients showed significant different relation between PBVC and TLVC ($B = -0.113$, $SE = 0.027$, $p < 0.001$). A post-hoc test revealed that the relation was significant only for patients who did not convert to CDMS ($B = 0.084$, $SE = 0.016$, $p < 0.001$). Similar results were achieved when looking at the relation between PVVC and TLVC ($B = 1.087$, $SE = 0.245$, $p < 0.001$). A post-hot test revealed that this relation was significant only for non-converters ($B = -0.837$, $SE = 0.145$, $p < 0.001$).

Voxel-wise: Results are summarized in table 5. In years 3 and 4, converters and non-converters showed significant different relation between atrophy and TLVC. In particular, faster periventricular and occipital lobe atrophy was associated with higher TLVC in patients who did convert to CDMS.

Time Interval	Number of significant Voxels	Location of significant voxels
Year 3	3779	PV
Year 4	153	OL

Table 5. Schematic representation of the Voxel-wise significant results for the influence of conversion on the concurrent relation between atrophy and LTVC. PV =Periventricular, OL = Occipital Lobe

Discussion

In this work we found that inflammation and neurodegeneration developed simultaneously in the early phase of MS, thus suggesting that these two processes partially resulted from different and independent pathological mechanisms. Interestingly, the spatio-temporal concordance between these two processes seems to take place mostly in the periventricular region. Further, WM lesion changes and brain atrophy seemed to be differentially related across an untreated and treated period.

To better elucidate the complex spatio-temporal dynamics between inflammation and neurodegeneration, it is important to consider two key points in the REFLEX/ION study design. First, patients were recruited after their first attack. This implies that most of the subjects had active inflammation when entering in the study. Thus, several competitive mechanisms should be considered: the anti-inflammatory effect of the treatment, the neurodegeneration (both pseudo-atrophy and “true” atrophy) and the slow focal damage accrual. Second, DT subjects who converted to CDMS during the first two years received treatment earlier than year 3. Thus, the patients who truly received delay treatment from year 3 were the less severe cases showing lower brain activity (i.e. non-converters). Taken together, these factors certainly influenced the relation between inflammation and neurodegeneration.

Contrary to the studies which assumed that inflammation precedes neurodegeneration (Chard, et al., 2003; Paolillo, et al., 2004; Dalton, et al., 2002), our first relevant finding is that WM lesions and brain atrophy developed simultaneously over time, suggesting an uncoupling between these two processes. Few other MRI studies investigated the relationship between lesion changes and brain atrophy within the same follow-up period in the early phase of MS (Varosanec, et al., 2015; Dalton, et al., 2004). Such studies found how WM lesions accrual and brain volume changes occurred simultaneously. Our results largely confirmed this: higher WM lesion volume changes was related to faster widespread atrophy within the same follow-up

period. Whether there is a pathological explanation for this association remains to be elucidated. Neuropathology observations revealed how profound axonal loss in the normal appearing white matter (NAWM) seemed to develop independently from axonal injury in demyelinated lesions (DeLuca, Williams, Evangelou, Ebers, & Esiri, 2006). Further, in long-term MS patients, ongoing myelin destruction, associated with axonal and neuronal degeneration, was detected in the absence of parenchymal inflammatory infiltration (Peterson, Bö, Mörk, Chang, & Trapp, 2001; Bø, Vedeler, Nyland, Trapp, & Mørk, 2003). These observations suggested that neurodegeneration in MS could occur independently from inflammation (Trapp & Nave, 2008). Whole brain analyses results were not entirely replicated at voxel-wise level. This is not unexpected given that whole brain analyses used all datapoints together, whereas voxel-wise analyses were performed within each time interval. This implies that the results achieved within a specific time-interval could in some extent be “diluted” by other years datapoints and thus could not be detected at whole brain level. Further, if we consider the REFLEX/ION study design, the initial shifts in fluid and changes in the volume of inflammation could have obscured the real relation between WM lesion volume changes and brain atrophy. Given this context, first years datapoints could have driven the overall detection of any relationships between inflammation and neurodegeneration. To support this theory, it is important to mention how the voxel-wise analyses in the first year of the study achieved the same results to those obtained at whole brain level (i.e. faster atrophy being associated to lower lesion volume change).

Inflammation and neurodegeneration were related mostly in the periventricular area. This finding is in line with other studies that showed how these two processes are greater in the periventricular areas and in proximity to CSF spaces (Brown, et al., 2017; Liu, et al., 2015). Such greater periventricular activity is related to the presence of locally secreted proinflammatory cytokines derived from CSF compartments harbor B cells that reside within the CSF space (Magliozzi, et al., 2018; Genovese, et al., 2019).

At a voxel-wise level we found a different relation between WM lesion changes and brain atrophy across an untreated and treated period. In particular, REFLEX period DT patients (i.e. placebo) showed faster whole and central atrophy being associated with lower TLVC. These findings are not surprising if we consider the two crucial points of REFLEX/ION study design previously discussed: the inclusion of only the non-converters patients (i.e. the ones showing less brain activity) in this analysis since they are the ones receiving truly delay treatment and the presence of baseline active inflammation. Indeed, previous studies have shown how MS patients presenting gadolinium-enhancing lesions have an accelerated decrease in brain volume in comparison to those without sign of active inflammation (Radue, et al., 2015; Vidal-Jordana, et al., 2013; Sastre-Garriga, et al., 2015). These observations were largely confirmed by the results achieved in this work. If we exclude the initial effect of active inflammation and shifts in fluids, at voxel-wise level we did find how inflammation and neurodegeneration are differentially related prior and after treatment. In the second year of the untreated period faster periventricular atrophy was associated with concurrent increase of TLVC. During a treated period, faster (mostly periventricular) atrophy was related to concurrent lower TLVC. These results demonstrated how treatment could influence the relation between inflammation and neurodegeneration. A possible explanation is that anti-inflammatory medication does not seem to stop the chronic accrual of pathology, which is not surprising. While acute inflammation could be largely suppressed, chronic inflammation and neurodegeneration could progress and may cause further neuronal and axonal death. Another possible theory could be that treatment initially exerts its effect on inflammation while effects on neurodegeneration requires more time to be detected.

Another key observation in our study is that, within the first year of treatment, faster whole and central brain atrophy was associated to lower TLVC. Paradoxically, anti-inflammatory drugs have often been associated with an acceleration of brain volume loss following the initiation of

therapy. This phenomenon, referred to as “pseudo-atrophy”, is generally assumed to be related to resolution of inflammation and fluid shifts (De Stefano, et al., 2014; Zivadinov, et al., 2008). Although its dynamics are still largely unknown, pseudo-atrophy certainly complicates the interpretation of brain atrophy measurements in both clinical and research settings (De Stefano, et al., 2021). Thus, it is crucial to investigate to what extent the pseudo-atrophy may be related to the resolution of inflammation as opposed to neurodegeneration. Further, it could be of great relevance to localize the brain tissues or regions where this phenomenon occurs. Congruent to pseudo-atrophy effect, in this work lower TLVC was associated to faster periventricular and frontal lobe atrophy within the first year of treatment of the ET patients. Conversely, such relation was not detected in the DT patients. This result was confirmed by the voxel-wise statistics looking at treatment effect on the relation between inflammation and neurodegeneration (RQ 2). Normally, one would expect that both ET and DT would have shown the same response to treatment onset. The lack of pseudo-atrophy effect on the DT group could be addressed to several reasons. First, it is well assumed how pseudo-atrophy is found only in patients who showed active inflammation (Vidal-Jordana, et al., 2013; Sastre-Garriga, et al., 2015; Radue, et al., 2015). Within this respect, DT patients showed relatively stable WM lesions activity (i.e. TLVC value close to 0, data not shown) during their first year of treatment. Second, it should also be considered the low sample size of the DT group (ET: 231; DT: 65). Finally, and accordingly to REFLEX/ION study design, most of the DT patients receiving treatment from year 3 were mostly the less severe cases (non-converters: 59; converters: 6). Taken together, these factors might have hampered the detection of the pseudo-atrophy effect within DT patients first year of treatment.

To prevent the confounding effect of resolving oedema and pseudo-atrophy during the first year of treatment, we restricted our analyses to datapoints where patients had received at least one year of therapy. Significantly different relationships were found at voxel-wise level: in ET

patients, faster occipital lobe atrophy was associated to higher TLVC, whereas in DT faster (mostly periventricular) atrophy was associated to lower TLVC. These results were confirmed by the voxel-wise statistics looking at treatment effect on the relation between WM lesion changes and brain atrophy (RQ 2). Although not straightforward, one might hypothesize a sort of prolonged pseudo-atrophy effect on the DT patients. Indeed, the course of pseudo-atrophy is not completely understood and thus, the assumption that pseudo-atrophy occurs only during the first year of therapy is not necessarily valid (De Stefano & Arnold, 2015).

Finally, we investigated whether conversion to CDMS could influence the relation between inflammation and neurodegeneration. It is well established how increased inflammatory activity and brain atrophy are related to higher risk of conversion to CDMS (Kalincik, et al., 2012; Tintoré, et al., 2006). Our results showed that in patients who did convert to MS, WM lesions and brain atrophy developed simultaneously, with higher TLVC being related to concurrent periventricular and occipital lobe faster atrophy.

This study is not without limitation. First, the REFLEX/ION study design made it difficult to precisely assess to what extent inflammation and neurodegeneration are the results of two independent pathologic mechanisms. Second, we did focus our analyses only on the relation between WM lesion changes and global/central brain atrophy. Several studies have highlighted the presence of GM damage in the early phase of MS (Raz, et al., 2010; Dalton, et al., 2004; Henry, et al., 2008). In this work the low quality of the T1-weighted images and the poor contrast across tissues made it difficult to look at the relation between WM lesions and GM damage. Future studies will address this issue by implementing new generation of imaging processing methods (i.e. SIENA-XL, Jacobian integration methods) (Battaglini, Jenkinson, De Stefano, & ADNI, 2018; Nakamura, et al., 2013) able to provide robust and accurate GM volumes estimates. Further, it would be very interesting to test whether WM lesions accrual in specific brain tracts is related to damage in “anatomically contiguous/connected” cortical lobes.

Third, the pathological explanation of the uncoupling between these inflammation and neurodegeneration remains to be elucidated. Indeed, although our results suggested that these two processes develop independently, genetic data and observation from most experimental models appear to favour a pathogenesis model in which inflammation precedes neurodegeneration (Milo, Korczyn, Manouchehri, & Stüve, 2020). Thus, the question whether inflammation and neurodegeneration are causally related or could develop independently is still a topic of discussion and our results did not provide a definite solution. Future studies should focus not only on the correlated and not causally linked changes but should also investigate the causal relation between WM lesions and brain atrophy. Within this respect, the causal relation between inflammation and neurodegeneration has been investigated in a separate study (6th chapter of this thesis) for the present dataset.

To conclude, we found that inflammation and neurodegeneration occur simultaneously in the early phase of MS, thus suggesting how WM lesions contribute only partially to the loss of overall brain tissue or vice versa. Interestingly, the periventricular regions are always affected by atrophy, while the parietal and temporal lobe seems to be involved at different temporal intervals and in relation with the treatment and the activities of the patients.

6. Study 3: The Spatio-temporal Relationship Between White Matter Lesions and Brain Atrophy in Clinically Isolated Syndrome and Early Multiple Sclerosis: A Post-hoc Analysis of the REFLEXION Study

This study was performed in collaboration with professor Hugo Vrenken and PhD candidate Rozemarijn Mattiesing* of the Amsterdam VU University Medical Center.

*Equally contributed to this work and share co-first authorship

Data of this work submitted to peer review scientific publication

Introduction

Multiple sclerosis (MS) is a chronic demyelinating inflammatory disease of the central nervous system with a neurodegenerative component. Already present from the earliest stages of the disease, the two most prominent pathological processes that lead to tissue damage in the brain are the formation of focal white matter (WM) lesions and atrophy (Simon, 2014). The development of these two pathological processes is presumed to be interrelated (Fisher, et al., 2002) but the underlying mechanisms remain to be elucidated.

In order to effectively intervene and target the underlying pathologies in the early phase of the disease, it is crucial to broaden our understanding of the underlying mechanisms, their interactions, and whether these can be modified by early treatment. For this reason, it is especially important to uncover the association between inflammation and neurodegeneration in patients with clinically isolated syndrome (CIS) and early MS.

There are relatively few longitudinal studies of the relationship between atrophy and WM damage in patients with CIS and early MS (Chard, et al., 2003; Dalton, et al., 2004; Dalton, et al., 2002; Paolillo, et al., 2004; Varosanec, et al., 2015), and these studies have focused on the assumption that inflammation precedes neurodegeneration. For example, Chard et al. (2003) found that, in long-term follow-up of patients with CIS, subsequent atrophy was more strongly related to the accumulation of focal T2 lesions in the early phase (0–5 years) rather than later phases in the study (5–10 years and 10–14 years). Dalton et al. (2002) found that baseline lesion measures are related to the development of subsequent ventricular enlargement over one year in those with CIS. According to a review, inflammatory damage and ongoing WM changes, such as gadolinium enhancing lesions, seem to be predictive of later atrophy in relapsing remitting MS (Zivadinov & Zorzon, 2002).

Currently it is unknown whether neurodegeneration in MS is secondary to the inflammatory processes leading to WM lesions or whether neurodegeneration is a primary disease process,

that (also) leads to secondary WM damage. Alternatively, both options might occur simultaneously. A recent review (Milo, Korczyn, Manouchehri, & Stüve, 2020) concluded that evidence from animal models and genetic studies favors a pathogenesis in which inflammation precedes neurodegeneration.

To investigate how WM lesions and atrophy in MS develop over time and how their evolution is related, long-term follow-up with regular magnetic resonance imaging (MRI) is required. The randomized, double-blind, placebo-controlled, multi-center REFLEXION clinical trial provides such an opportunity. In this trial, patients presenting with CIS were followed over a period of 5 years with yearly MRI scans. In the primary analyses of the study by Comi et al. (2017), overall MRI activity was reduced in patients receiving early treatment compared to patients receiving delayed treatment with subcutaneous interferon β -1a (sc IFN β -1a). However, that study did not look at the relationship between atrophy and WM lesion measures.

In the current study, we therefore conducted advanced post-hoc image analyses on the REFLEXION dataset. Our main goal was to investigate whether WM lesions are spatio-temporally related to subsequent atrophy in patients with CIS and early MS. In turn, we also studied if this possible association differed between patients receiving either early or delayed treatment with sc IFN β -1a, or between patients who converted to clinically definite MS during the study and those who did not.

Methods

Study

REFLEXION was a preplanned extension of the randomized, double-blind, placebo-controlled, multicenter REFLEX clinical trial (REbif FLEXible Dosing in Early Multiple Sclerosis). Procedures and the design of the study have been described in detail elsewhere (Comi, et al., 2012) (Comi, et al., 2017). Briefly, patients with CIS at high risk of converting to MS were included and either randomized to one of two early treatment arms, where treatment with sc IFN β -1a 44 μ g was initiated once a week or three times a week, or to the delayed treatment arm in which, during the first 2 years (i.e., the REFLEX phase), patients did not receive treatment (placebo group) but at the start of REFLEXION received sc IFN β -1a 44 μ g three times a week (Figure 1). If patients converted to clinically definite MS (CDMS), they received open-label treatment with sc IFN β -1a 44 μ g three times a week. CDMS was defined by a relapse accompanied by an abnormal MRI scan or a sustained increase in Expanded Disability Status Scale score of ≥ 1.5 points, i.e. as defined by the 2005 McDonald criteria (Polman, et al., 2005).

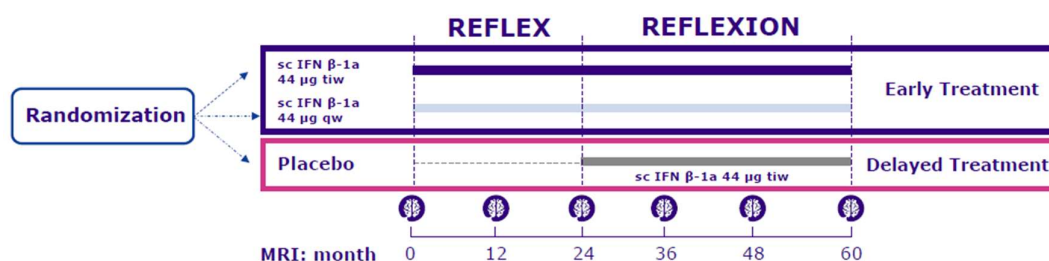


Figure 1. Schematic representation of the REFLEX and its extension REFLEXION study design.

MRI data

For the current post-hoc analyses, multicenter yearly MRI scans over the full REFLEX/REFLEXION study period of 5 years were evaluated. These consisted of 1x1x3 mm proton-density- (PD), T2-, T1-, and gadolinium-enhanced T1-weighted images. The Image Analysis Center of Amsterdam UMC (Location VUmc, Amsterdam, The Netherlands)

provided manual delineations of the PD-/T2-weighted lesions for each yearly visit and manually edited T1-weighted brain extraction masks originally obtained by using the FSL brain extraction tool (Smith, 2002), part of FMRIB's software library (Smith, et al., 2004). Scans were included if the input data were of sufficient quality and the different processing steps and output of the image analyses described below passed quality control measures (specific criteria are described in more detail in the Results section).

Longitudinal atrophy measurement

The T1-weighted images were corrected for slice-to-slice intensity variations (due to interleaved acquisitions) and subsequently lesion filled with the linearly registered PD-/T2-weighted manual delineations. Yearly percentage brain volume change (PBVC) and percentage ventricular volume change (PVVC) were estimated with SIENA (Smith, et al., 2002; Smith, De Stefano, Jenkinson, & Matthews, 2001) and its extension VIENA (Vrenken, et al., 2014), both part of FMRIB's software library (Smith, et al., 2004). The normalized and lesion filled T1-weighted image and the manually edited brain mask were used as input. Yearly PBVC and PVVC were used as longitudinal measures of whole brain atrophy and central atrophy, respectively.

Longitudinal lesion change quantification

Yearly lesion volume changes were automatically segmented by an in-house developed method that is based on the use of subtraction images (SI) (Moraal, et al., 2010; Moraal, et al., 2010). Details of this method are described in a separate work investigating the effect of sc IFN β -1a treatment on WM lesions accrual and the relation between WM lesions distribution and conversion to CDMS in the REFLEX study (Battaglini, et al., In Preparation). Briefly, before creating the SIs, the slice-to-slice variation in signal intensity on the PD-weighted images was corrected. The PD images of both visits of a yearly interval (e.g., baseline-month 12) were registered to a common halfway space using a similar procedure to that used in SIENA

software, based on the T2-weighted images. To obtain the SIs, the PD-weighted image of the first visit of each interval was subtracted from the second visit. To give a robust analysis, the SIs were further normalized to account for the differences between the sites and MRI scanners from which the images were obtained. Voxels with a normalized intensity difference exceeding 1.5 standard deviations were labelled as changing. Based on the baseline and follow-up lesion masks and the voxel-wise lesion changes, each individual lesion was labelled as new, enlarging, shrinking, or disappearing. The yearly total lesion volume change (TLVC) was thereafter calculated by subtracting the sum of the negative lesion volume change (shrinking + disappearing) from the positive lesion volume change (new + enlarging) for each interval. To allow lesion probability map analyses of the anatomical distribution of the four lesion types, lesion labels were stored in a lesion change map (LCM) for each of the four lesion types, for each patient and for each interval.

Voxel-wise input images

The following steps were performed in order to produce the input images for voxel-wise statistical analyses. First, we created a study-specific template. FSL-SIENAX (Smith, et al., 2002) was used to obtain normalized brain volume for all baseline T1-weighted images. Afterwards, 100 patients were selected based on the percentile distribution of normalized brain volume (from 1st to the 100th percentile). For each of these 100 patients, the T1-weighted images were intensity normalized (divided by the 99th intensity percentile and multiplied by 10,000), non-linearly registered to the MNI standard space (resolution = 2x2x2 mm) using FSL-FNIRT (Andersson, Jenkinson, & Smith, 2010) and averaged to create the study-specific template. Second, to ensure that all the images of each patient underwent the same pre-processing and to avoid interpolation bias, a patient-specific template was created. Accordingly, the T1-weighted images were intensity-normalized using the N4 algorithm (Tustison, et al., 2010), linearly registered to the baseline T1-weighted scan using FLIRT (Jenkinson, Bannister,

Brady, & Smith, 2002), and averaged to create the patient-specific template. The patient-specific template was then non-linearly registered on the study-specific template space using FNIRT and the warp-fields generated from this registration were used for the subsequent registration of SIENA and LCM outputs on the study-specific template.

To study local atrophy, brain edge shift maps were created. For each patient, the yearly brain edge shift images provided by SIENA were spatially dilated, non-linearly registered to the study-specific template using FNIRT, masked with a standard space brain edge image, smoothed with an isotropic Gaussian kernel with a sigma of 5 mm, and remasked (Bartsch, et al., 2007; De Stefano, et al., 2003)

To study local lesion activity, the yearly LCMs of growing, new, shrinking, and disappearing lesions were non-linearly registered to the study-specific template using the warp-fields generated through the procedure described above.

Statistical analyses

Whole brain: Statistical analyses were performed in Rstudio. For the whole brain analyses, linear mixed models were used to deal with the repeated measurements we have for the yearly atrophy and lesion change measures. We incorporated a random intercept with a three level structure where observations are clustered within the patients and the patients are clustered within the different study sites. All linear mixed models were corrected for age and sex. Regarding conversion to CDMS, depending on the research question, we categorized patients into converters and non-converters either considering the full 5-year study period or using each patient's time-dependent CDMS status for the yearly interval under consideration. An alpha of 0.05 was used as the cut-off for significance. Significant interactions were further explored by post-hoc tests.

First, we performed separate analyses in order to assess if atrophy and lesion volume changes differed between treatment groups and between converters and non-converters. Linear mixed

models were used with treatment and interval-specific CDMS status as fixed factors. To assess the treatment effect over time, we used a similar model but incorporated an interaction between treatment and the yearly intervals. For atrophy, PBVC and PVVC were alternately used as the dependent variable; for lesion volume changes, TLVC was the dependent variable. Second, to test whether lesion volume changes were related to atrophy in the next year, we incorporated a time-lag in our linear mixed models to link TLVC in year 1 to PBVC or PVVC in year 2, and TLVC in year 2 to PBVC or PVVC in year 3, etc... We then applied a linear mixed model with TLVC as the independent variable and PBVC or PVVC in the next year as the dependent variable, correcting for treatment and CDMS status across the whole study period. Additionally, to test whether the relationship between lesion volume changes and atrophy in the next year differed between treatment groups or between converters and non-converters, we used similar models but also incorporated an interaction between treatment and TLVC, and CDMS status across the whole study period and TLVC, respectively.

In order to prevent confounding effects by resolving edema and pseudo-atrophy at the start of treatment, all linear mixed models were performed on the data points where the patients have received at least one year of treatment. This means that for the early treatment group, TLVC in years 2, 3, and 4 and PBVC or PVVC in the next year were included; and for the delayed treatment group, only TLVC in year 4 and PBVC or PVVC in year 5. This will be called the stable treatment period.

To test the relationship between lesion volume changes and atrophy in the next year in an untreated period, we included the data points of TLVC in year 1 and PBVC or PVVC in year 2 of the delayed treatment (then placebo) group in the REFLEX period, while excluding the delayed treatment patients who converted during the first 2 years of the study and controlling for CDMS status across the whole study period. To see if the relationship differed between the REFLEX placebo period and the REFLEXION treatment period (TLVC in year 4 and PBVC

or PVVC in year 5) excluding the first year of treatment of the delayed treatment patients (TLVC in year 3 and PBVC or PVVC in year 4), we incorporated an interaction between TLVC and period and corrected for CDMS status across the study period in a separate linear mixed model.

Voxel-wise: For each yearly interval, regional statistical inference was carried out using permutation testing (Nichols & Holmes, 2002) (5000 permutations) as implemented in the Randomise program of FMRIB's software library (Winkler, Ridgway, Webster, Smith, & Nichols, 2014). The Threshold-Free Cluster Enhancement randomise option was used. When looking at the difference between early and delayed treatment patients and converters/non-converters in terms of atrophy and LCMs within each interval, design matrices within the GLM framework were used with treatment and interval-specific CDMS status as variables of interest, and age, sex, and study site as covariates. TLVC was used as regressor when looking at the relationship with brain edge shifts in the next year. Only results with at least 15 significant voxels were reported. The anatomical location was determined by using pre-defined standard space masks (see <http://www.fmrib.ox.ac.uk/fsl/fslwiki/>), as provided by the MNI structural atlas and the JHU WM tractography atlas. The number (V) and the location of significant voxels ($p < 0.05$) were reported. The voxel-wise analyses were matched to the whole-brain analyses but could only be performed for each yearly interval.

Results

A total of 400 patients enrolled in the REFLEXION study provided MRI data for the extension period, and the input data of 392 patients were included in the current analyses (Table 1). Concerning the input data, 4 patients were excluded because of incomplete trial data, 2 patients because of an inconsistent acquisition protocol, and 2 patients because no consecutive visits were available that were needed to calculate yearly atrophy and lesion change measures. Regarding visit data, 5 visits were excluded because of incorrect/incomplete image(s), 8 visits because of movement, 6 visits because of missing data, and 2 visits were excluded from the lesion change analyses because of corrupted PD-weighted images. The quality check of the output from the lesion change quantification and longitudinal atrophy measurements resulted in 158 excluded visits (23 lesion change quantification, 133 longitudinal atrophy, and 2 shared rejections). Reasons for exclusion were: low quality of the images, artefacts, registration problems, and pipeline failure.

	Converters to CDMS	Non-converters to CDMS	ET	DT	Overall
Subjects (N)	162	230	262	130	392
Gender (F/M)	95/67	147/83	162/100	80/50	242/150
Age (mean±SD)	30.32±7.99	32.23±8.52	31.68±8.43	30.97±8.19	31.44±8.35

Table 1. Demographics of the included patients. CDMS = clinically definite multiple sclerosis (across the whole study period), ET = early treatment, DT = delayed treatment, SD = standard deviation.

Atrophy

Table 2 and Figure 2, panel A and C, provide the yearly PBVC and PVVC for the different treatment groups. When looking separately within each yearly interval, global atrophy was higher in the first year for early versus delayed treatment (PBVC: B=-0.198, SE=0.069,

$p=0.004$), indicative of resolving edema and pseudo-atrophy. In the second and fourth year, global atrophy rate was slower in the early vs delayed treatment group (PBVC, year 2: $B=0.159$, $SE=0.069$, $p=0.021$; PBVC, year 4: $B=0.176$, $SE=0.076$, $p=0.021$). For central atrophy, similar results were found in year 1 ($B=2.538$, $SE=0.616$, $p<0.001$) and year 2 ($B=-1.560$, $SE=0.614$, $p=0.011$). Taken across the whole 5-year study period, atrophy rate (PBVC and PVVC) did not differ significantly between patients in the early and delayed treatment group. On a voxel-wise level, similar results were observed in the first year; there was faster periventricular atrophy and atrophy in the temporal lobe for the early vs delayed treatment group ($V=4165$). In the second year, the early treatment group showed slower atrophy in the frontal lobe compared with the delayed treatment group ($V=807$).

Table 2 and Figure 2, panel B and D, provide the yearly PBVC and PVVC for the interval-specific converters/non-converters to CDMS. Across the whole 5-year study period, compared to non-converters, patients who converted to CDMS showed faster global atrophy ($B=-0.112$, $SE=0.035$, $p=0.001$) but not central atrophy. In voxel-wise analyses, patients who converted during an interval showed faster atrophy compared with non-converters, mostly periventricular, in years 1 ($V=1720$), 2 ($V=93$), and 4 ($V=919$).

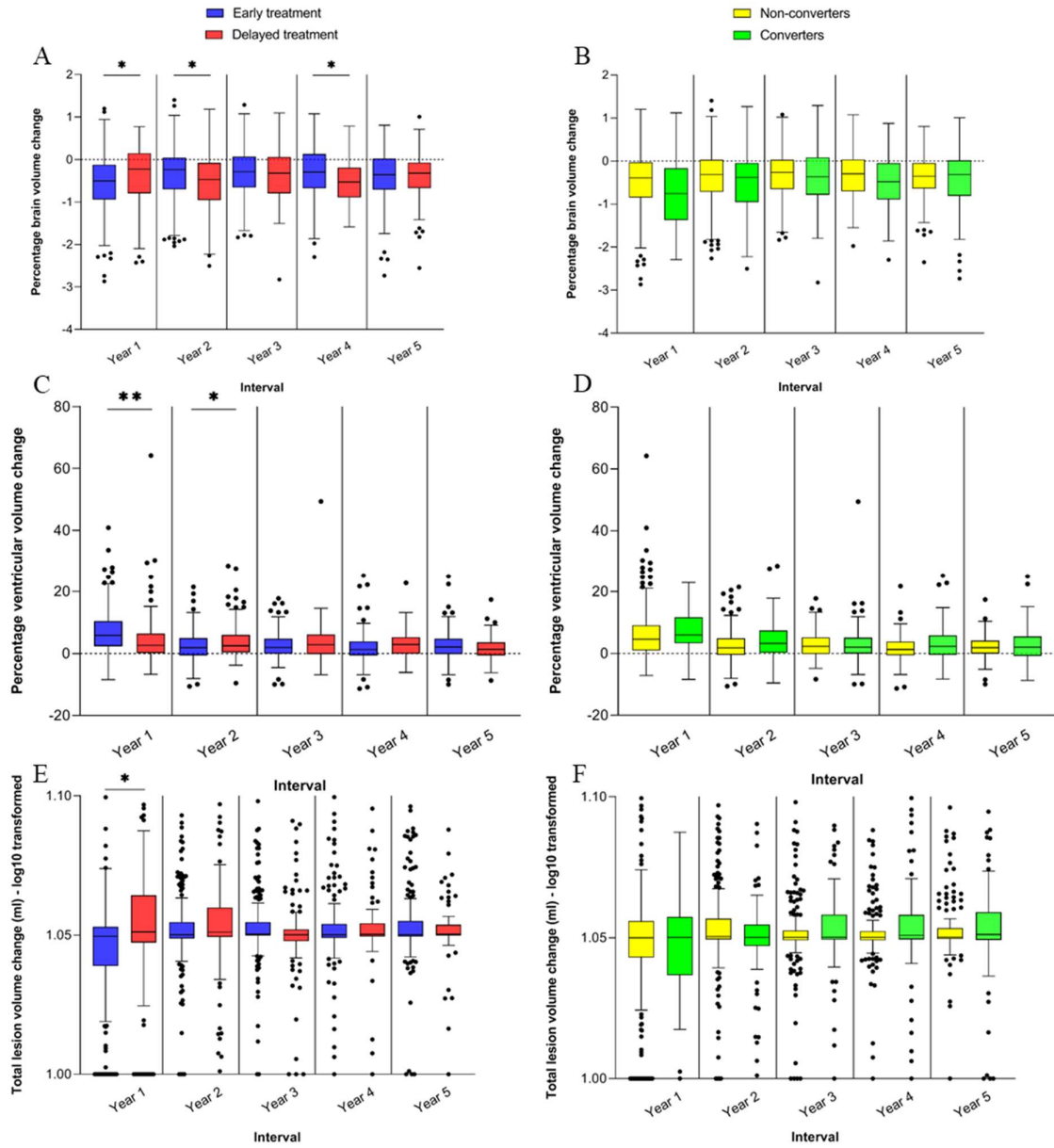


Figure 2. Boxplots depicting the percentage brain volume change, percentage ventricular volume change, and total lesion volume change across all years for the early and delayed treatment groups and interval-specific converters and non-converters. Outliers beyond the y-axis range are shown on the x-axis (panel E and F). * $p < 0.05$. ** $p < 0.001$

Measure	Group	Year 1	Year 2	Year 3	Year 4	Year 5
PBVC (%/y)	ET	-0.541±0.722	-0.353±0.603	-0.322±0.547	-0.311±0.584	-0.412±0.597
	DT	-0.362±0.675	-0.536±0.750	-0.371±0.591	-0.508±0.533	-0.416±0.573
	Nonc	-0.443±0.685	-0.375±0.620	-0.324±0.522	-0.316±0.525	-0.382±0.518
	Conv	-0.752±0.828	-0.521±0.754	-0.366±0.633	-0.480±0.640	-0.464±0.687
PVVC (%/y)	ET	6.868±7.095	2.373±4.582	2.518±4.195	1.968±4.725	2.602±4.354
	DT	4.413±8.479	4.022±5.965	3.604±6.394	2.901±4.431	1.850±3.933
	Nonc	5.870±7.741	2.412±4.620	2.843±4.186	1.713±3.935	2.098±3.435
	Conv	7.283±7.045	4.319±6.149	2.941±6.414	3.254±5.540	2.816±5.270
TLVC (mL/y)	ET	-0.339±1.568	0.207±0.983	0.214±0.837	0.150±1.002	0.185±0.874
	DT	-0.006±1.945	0.218±0.838	0.065±0.960	0.222±1.078	0.041±0.697
	Nonc	-0.254±1.731	0.217±0.927	0.097±0.664	0.077±0.643	0.073±0.542
	Conv	-0.057±1.556	0.196±0.961	0.296±1.205	0.345±1.469	0.250±1.135

Table 2. Longitudinal atrophy and lesion volume changes measures across treatment groups and interval-specific converters to clinically definite multiple sclerosis and non-converters, in each year of the study. Values are mean ± standard deviation. ET = early treatment, DT = delayed treatment, Conv = converters to clinically definite multiple sclerosis (CDMS), Nonc = non-converters to CDMS, PBVC = percentage brain volume change, PVVC = percentage ventricular volume change, TLVC = total lesion volume change.

White matter lesion volume changes

In Table 2 and Figure 2, panel E, the yearly TLVC is provided for the early and delayed treatment groups. When looking separately within each interval, the early treatment group showed a lower TLVC compared with the delayed treatment group only in year 1 (B=-0.318, SE=0.125, p=0.011). Overall, TLVC did not differ between the early and delayed treatment groups across the whole 5-year study period. In voxel-wise analyses, early treatment patients showed lower activity of growing lesions compared with the delayed treatment group in year 1 in the forceps minor and anterior thalamic radiation (V=49), and higher activity of shrinking lesions in year 5 in the posterior corona radiata (V=98).

Yearly TLVC is provided for converters/non-converters using the interval-specific CDMS status in Table 2 and Figure 2, panel F. Converters showed a higher TLVC compared with non-converters across the whole study period ($B=0.217$, $SE=0.062$, $p<0.001$). In year 5, converters showed higher activity of growing lesions compared with non-converters in the superior and posterior corona radiata, and the forceps major ($V=358$).

Relationship between WM lesion volume changes and subsequent atrophy

During stable treatment, after at least one year of treatment (so excluding the first year for the early treatment group and the third year for the delayed treatment group), we found a negative relationship between TLVC and PBVC in the next year ($B=-0.113$, $SE=0.022$, $p<0.001$), consistent with a higher lesion volume change being related to faster atrophy in the next year. This relationship did not differ between the treatment groups and also not between patients who did and did not convert to CDMS across the whole study period. We found a similar relationship for central atrophy: TLVC was positively related to PVVC in the next year ($B=1.156$, $SE=0.164$, $p<0.001$). This relationship did not differ between converters and non-converters, but treatment seemed to modulate the effect (TLVC*treatment: $B=0.972$, $SE=0.421$, $p=0.021$). A post-hoc test revealed that TLVC was only significantly related to PVVC in the next year in the early treatment group ($B=1.348$, $SE=0.181$, $p<0.001$). Voxel-wise analyses showed that, in early treatment patients, higher TLVC in years 2 ($V=3926$) and 3 ($V=1369$) were related to faster periventricular atrophy in the next year, as shown in Figure 3. In year 4, TLVC was related to faster periventricular atrophy in the next year in the early ($V=322$) and delayed treatment ($V=472$) groups. A separate analysis that included an interaction term indicated that this relationship was stronger in the delayed treatment group ($V=113$), as shown in Figure 4. The (temporal-spatial) relationship between TLVC and atrophy in the next year did not differ between CDMS converters and non-converters.

In the untreated period of patients who received delayed treatment, there was no significant relationship between TLVC and PBVC or PVVC in the next year. This relationship also did not differ significantly between the REFLEX and REFLEXION period. In voxel-wise analyses, the relationship

was also not present in the REFLEX period; but in the REFLEXION period, higher TLVC in year 4 was related to faster periventricular atrophy in the next year (V=472).

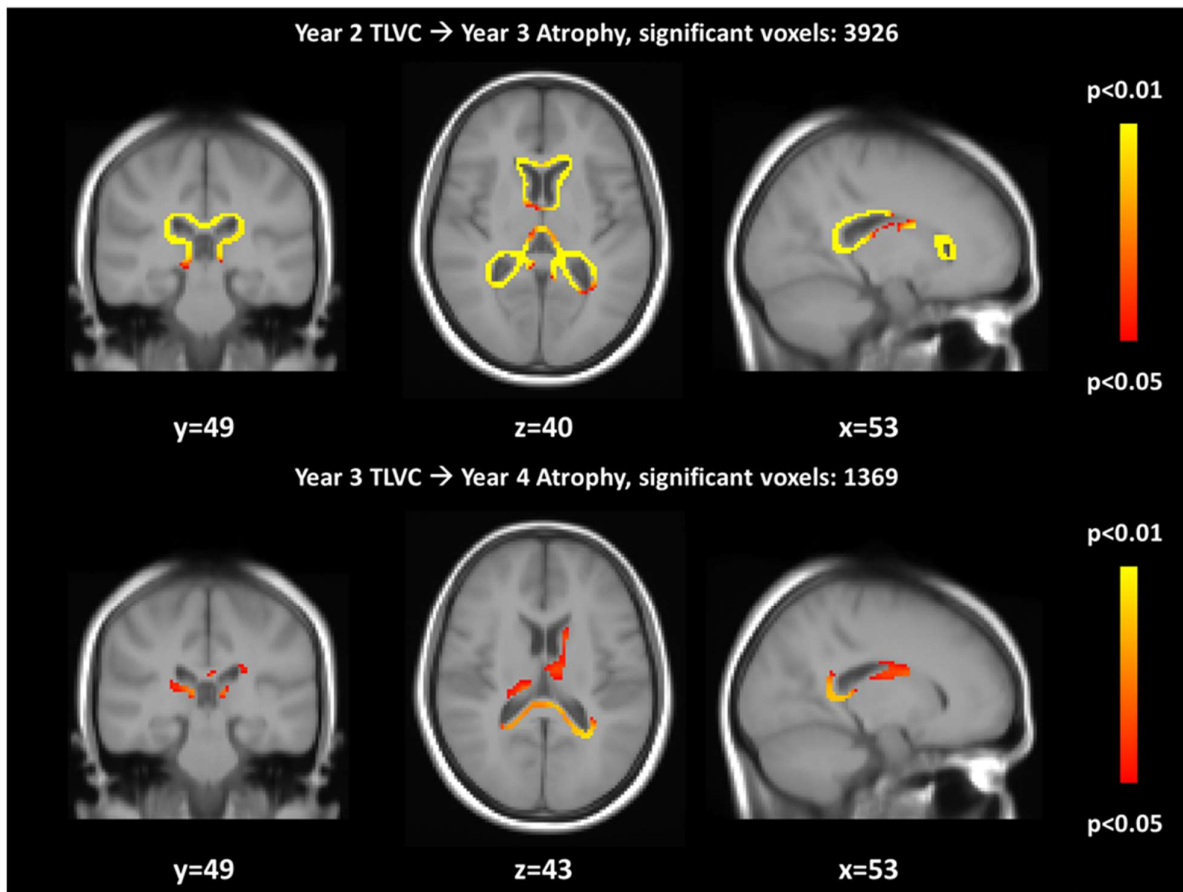


Figure 3. Voxelwise analyses in the early treatment group: significant regions where higher TLVC in year 2 (top row) and year 3 (bottom row) was related to faster periventricular atrophy in the next year.

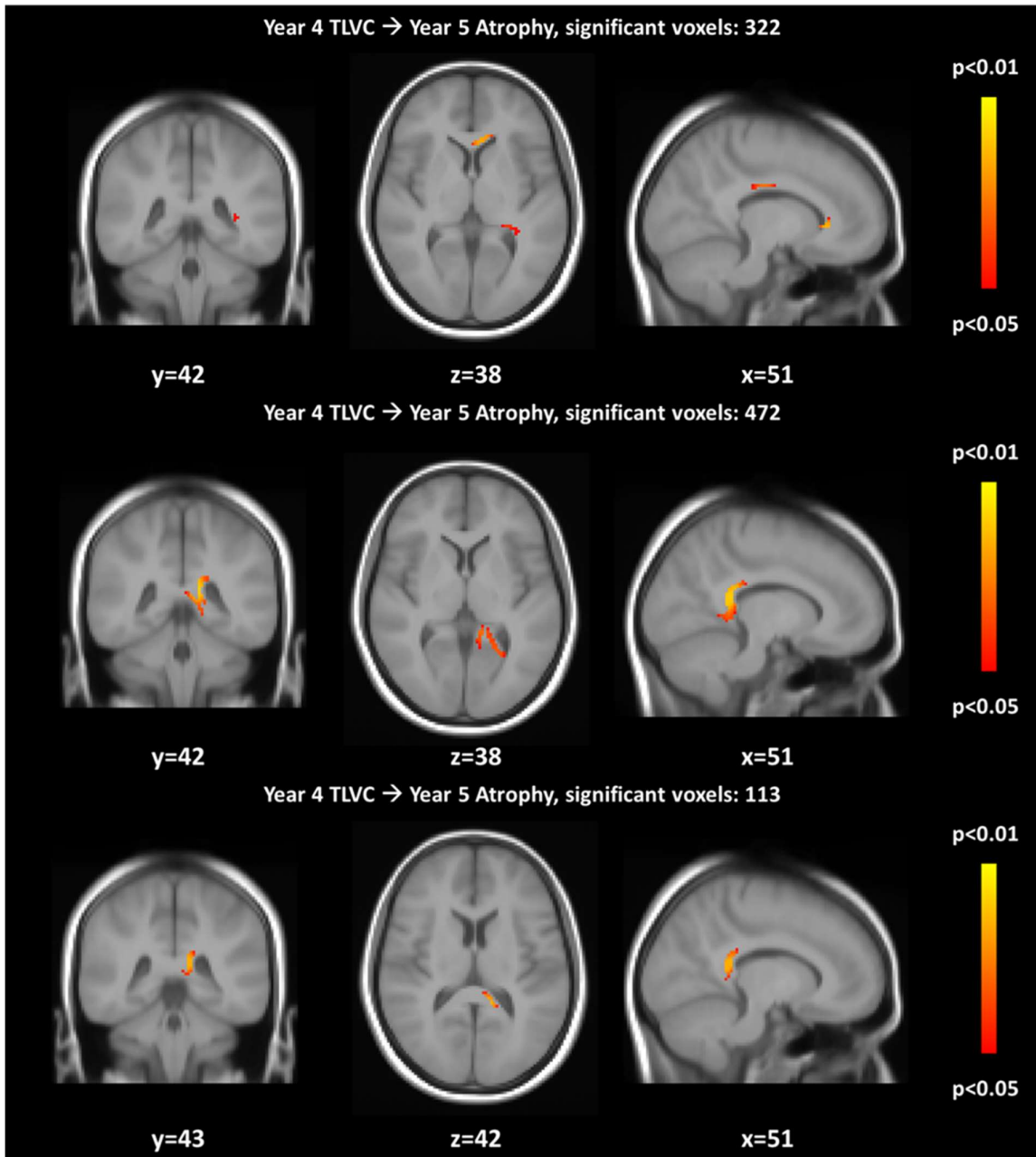


Figure 4. Voxelwise results showing the relation between TLVC in year 4 and subsequent atrophy in year 5 in the early treatment group (top row) and the delayed treatment group (middle row), as well as the regions in which this relation between TLVC and subsequent atrophy was significantly stronger in the delayed treatment group compared to the early treatment group (bottom row).

Discussion

This study found that in patients with CIS and early MS, during stable treatment (at least one year) with sc IFN β -1a, higher WM lesion volume changes were related to faster periventricular atrophy in the next year. In the short untreated period for patients with delayed treatment, the relationship between lesion volume changes and atrophy in the next year was not significant.

In year 1, patients who received early treatment showed faster brain volume loss (in line with expected pseudo-atrophy) and lesion volume decrease (in line with expected resolving edema), while in years 2 and 4 they showed slower atrophy compared with patients who received delayed treatment. In voxel-wise analyses, similar results were found in the periventricular areas and temporal lobe in year 1, and in the frontal lobe in year 2.

Previous studies found that lesion measures were associated with subsequent atrophy (Chard, et al., 2003; Dalton, et al., 2004; Dalton, et al., 2002; Paolillo, et al., 2004; Varosanec, et al., 2015). Our results largely confirmed this: higher WM lesion volume changes were related to faster periventricular atrophy in the next year in patients with CIS and early MS. The association was not the same for all investigated subsets of the trial data. We first looked at the stable treatment period, in which we only included the data points where patients had received at least 1 year of therapy. This selection was made to prevent the confounding effects of resolving edema and pseudo-atrophy during the first year of treatment with sc IFN β -1a. Anti-inflammatory medication is known to induce an initial reduction in brain volume during the first 6 months to 1 year, which is not associated with a loss of cell structures but rather fluid shifts (De Stefano, et al., 2014; Zivadinov, et al., 2008). The relationship between WM change and subsequent atrophy was found in this stable treatment period. However, we did not observe this relationship in untreated patients, i.e. when we focused on the placebo period of the delayed treatment group in the first 2 years of the REFLEX study.

Our findings do not imply that treatment triggers a relationship between WM changes and subsequent atrophy. Actually, it might be expected that this relationship would be more apparent in the delayed treatment group, while these patients were receiving placebo, since the inflammation is not (yet)

suppressed. However, this (exploratory) analysis only included one data point (TLVC at year 1 and PBVC or PVVC at year 2), had a relatively small sample size, and this group also has the potential for bias, because of the presumably non-random removal of patients who converted to CDMS and who were then treated with open label sc IFN β -1a; all of these factors may have prevented the detection of a relationship in the placebo period. Since patients were recruited just after their first attack, during the first two years (the REFLEX period), placebo recipients may have exhibited shifts in fluid and changes in the volume of inflammatory cells that might obscure any relationship between true lesion accrual and true atrophy during this period. Moreover, the relationship did not differ between the REFLEX and REFLEXION period for the delayed treatment patients, meaning that there was no difference between the untreated and treated period for such patients. Therefore, treatment does not cause the association between WM changes and subsequent atrophy to appear.

For central atrophy, the relationship between WM changes and subsequent atrophy seemed to be modulated by treatment with sc IFN β -1a, since a post-hoc test revealed that it was only present in the early treatment group. However, this additional interaction analysis is limited by the fact that the delayed treatment patients only had one data point included (because they started treatment in year 3, unless they converted before) and the early treatment group was overrepresented with three data points for each patient. Therefore, this result should be interpreted with caution.

The very few studies that have investigated the relation between WM lesions and brain atrophy in CIS and early MS have focused on the hypothesis that inflammation precedes neurodegeneration. To the best of our knowledge, the causal relation between WM lesion changes and brain atrophy has not been investigated in a similar longitudinal manner before. However, evidence suggest that lesion accrual and brain atrophy could be largely independent processes that both occur at increased rates in patients or during periods with more severe disease. Future studies should in more detail investigate both hypothesized causal relations between lesions and atrophy, as well as the possibility of correlated but not causally linked changes. Such studies should employ imaging techniques that are able to zoom in on the local microstructure and tissue properties, such as diffusion tensor imaging. Furthermore,

the relation between concurrent changes in lesions and atrophy has been investigated in a separate study (5th chapter of this thesis) for the present dataset.

Treatment with sc IFN β -1a only had a significant effect on atrophy and lesion volume changes during certain periods of the study. The results in the first year were indicative of resolving edema and pseudo-atrophy, which is to be expected based on what is known about the effects of anti-inflammatory treatment during the first year (De Stefano, et al., 2014; Zivadinov, et al., 2008). Interestingly, we did not see this effect in the delayed treatment group in the first year of treatment (i.e. year 3 of the study). This might be because this group is smaller, but we did find a significant difference in the fourth year of the study, which could be speculated to reflect a delayed pseudo-atrophy effect in such patients. It would have been easier to explain if this effect occurred in the third year of the study, since the delayed treatment patients started treatment at that time (unless they converted to CDMS beforehand). In the fourth year, these patients had already received at least 1 year of treatment and it seems counterintuitive that the atrophy rate had not become more similar between the early and delayed treatment groups by then. This potentially has an important clinical consequence, by highlighting a different response to early and delayed treatment in terms of brain atrophy and lesion accrual.

Patients who converted to CDMS showed faster global atrophy and lesion volume change across the whole study period compared to non-converters, and since these patients have a worse disease progression this seems to be expected. However, because they received treatment with sc IFN β -1a three times a week upon conversion to CDMS, this result is somewhat difficult to interpret.

This brings us to one of the limitations of the study. Since the study design is fairly complicated because treatment (dosing) and conversion status are intertwined, it was not possible to correct for all potential confounders. We tried to account for this in the linear mixed models, and for this reason we analyzed specific subsets of the trial data. For example, in order to look at a pure untreated placebo group we excluded the converters in the first and second year of the study, but this also introduced a potential selection bias of including only the cases with a more benign disease progression.

Ideally, we would have looked at the relationship between WM lesions and gray matter (GM) atrophy. However, due to the quality of the 2D scans, with limited contrast in the images, which makes it difficult to detect the WM-GM border, this was not possible. For this reason we focused on global and central atrophy measures, and we cannot make statements about the relationship between WM and GM pathology in MS. Future studies using 3D imaging with adequate WM-GM contrast could address these issues, including the relationship between specific cortical lobes and WM tracts in the brain.

A strength of this trial was regular, tightly controlled MRI scans over a fairly long period, and a large sample that enabled us to investigate the relationship between atrophy and WM changes in patients with CIS early in the disease process. Ideally these patients would have been followed-up for an even longer duration. Therefore, future trials should aim to increase the follow-up period to better elucidate the relationship between the two pathological processes, and make use of the most recent advances in imaging such as 3D-FLAIR and 3D-T1 images to be able to look in more detail at GM and WM; for example, by measuring cortical thickness. These studies should also focus on WM damage potentially being secondary to neurodegeneration.

In conclusion, we found that higher lesion volume changes were related to subsequent faster periventricular atrophy in patients with CIS and early MS. The question remains whether these processes are causally related or whether they are merely two pathological processes that occur simultaneously in such patients. This needs to be investigated further in future studies.

7. Summary and future perspectives

The studies presented in this thesis faced two MS lesions related challenges. We first dealt with the technical difficulties of developing an accurate tool for automated WM lesion segmentation. Afterwards, we investigated the complex inter-role across inflammation and neurodegeneration. For these two distinct MS challenges, a summary followed by possible future directions of research will be provided in the following paragraphs.

MS Challenge 1: Automated lesion segmentation

Summary. As mentioned in the Introduction, accurate lesion segmentation is a complex task for several reasons. In certain respects, one could argue that some of the tools developed this far have reached a higher degree of accuracy and sensitivity. However, these tools usually showed poor consistency, robustness and generalizability. Such limitation could be due to the lack of an extensive validation approach, with most of the tools showing to be “tailored” on specific acquisition protocols. Further, there is very scarce employment of high-resolution FLAIR images, now considered the preferable sequence where to detect MS lesions. Finally, the use of the manual segmentation as gold standard to measure tools performances should be critically considered. Indeed, the lack of reproducibility and the intra/inter-rater variability of the manual approach could affect the automated segmentation performances, thus impairing the validation procedure.

Against this background, in this thesis a novel approach, named BIANCA-MS, was introduced. This procedure comprises two innovative key elements. The first one is a harmonized setting tested under different acquisition protocols which avoid the long optimization procedure needed to tune algorithm settings to each dataset. The second key element is a post-processing cleaning-step. This was designed to be applied to both the manual segmentation, to reduce the impact of the inter-rater variability during algorithm validation, and to BIANCA-MS outputs, to further refine lesions segmentation.

Our experiments highlighted how BIANCA-MS achieved, on both low and high-resolution FLAIR images, significant higher degree of similarity to the gold standard compared to other widely used tools. Further, the consistency and reproducibility of performances achieved across different datasets (i.e. different scanning acquisition protocols) proved BIANCA-MS robustness and flexibility. On a multicentre dataset, BIANCA-MS demonstrated to be insensitive to data stratification per centre, making it easier to apply in clinical trials context. Finally, when all the scanning protocols are mixed into one pooled dataset, BIANCA-MS performances were comparable to the ones separately achieved on each centre. This introduces the possibility of obtaining a BIANCA-MS version that is pre-validated on large datasets and can perform robust and accurate lesion segmentation on “unseen” or new MRI data without needing to be revalidated.

Taken together, these encouraging results suggested how BIANCA-MS is a promising tool able to overcome some of the technical issues that still makes MS automated lesion an open challenge.

Future perspectives. As lesion identification on MRI is a crucial diagnostic step in MS, it is important to develop a tool as accurate as possible. In this respect, it would be greatly important to test BIANCA-MS behaviour on healthy controls. Such test will provide crucial information about BIANCA-MS reliability and specificity, opening the way to its possible implementation in clinical practice.

Further, it would be of utmost relevance to develop a BIANCA-MS longitudinal pipeline able to provide robust lesions volumetric assessment over-time and accurately classify lesions accordingly to their degree of activity (i.e. new, shrinking, disappearing, enlarging or stable). Finally, future efforts could address the implementation of sequences to allow GM lesion detection.

MS Challenge 2: The inter-role between inflammation and neurodegeneration

To date, whether inflammation and neurodegeneration are two independent or causally related processes is still a topic of debate. Uncover the association between these two pathological processes

in the early phase of MS is of utmost relevance to effectively intervene and target the underlying pathologies. Noteworthy, most of the studies have investigated the inter-role between inflammation and neurodegeneration at whole brain level, with regional analyses being poorly used and often limited to a single follow-up.

Given this context, by using both whole brain and voxel-wise analyses, two complementary studies on the REFLEX/ION clinical trial were performed to assess the spatio-temporal relation between WM lesion changes and brain atrophy. We first investigated whether inflammation and neurodegeneration are two independent processes which develop simultaneously over time; second, we explored whether and to which extent WM lesion changes are related to subsequent atrophy thus implying a causal relationship between these two pathological processes. Further, in the REFLEX/ION trial patients received either early (from baseline) and delay (from year 3) treatment and underwent yearly MRI for a 5-year period. This design allowed us to further investigate whether treatment and disease worsening could influence the relation between inflammation and neurodegeneration.

Concurrent relation between WM lesion volume changes and brain atrophy

Summary. In this study we found that inflammation and neurodegeneration developed simultaneously in the early phase of MS, thus suggesting that these two processes partially resulted from different and independent pathological mechanisms. We then restricted our analyses within the first year of treatment to better explore the pseudoatrophy effect. As mentioned in the Introduction, this phenomenon certainly complicates the interpretation of brain atrophy measurements in both clinical and research settings. Thus, it is crucial to investigate to what extent the pseudoatrophy may be related to the resolution of inflammation as opposed to neurodegeneration. In our work, lower WM lesion volume changes was related to faster periventricular and frontal lobe atrophy. Interestingly, this phenomenon was detected only in patients who showed signs of active inflammation (i.e. early treated patients).

Another key finding in this work was that WM lesion changes and brain atrophy seemed to be differentially related prior and after treatment onset. However, this was detected only at voxel-wise level and not with whole brain analyses. This finding could suggest specific regional mechanisms, which presence is “diluted” when the whole brain analyses are performed. In this work, faster periventricular atrophy was associated with higher WM lesion volume changes during an untreated period, whereas faster periventricular atrophy was related to lower inflammatory activity during a treated period. These results suggested that while treatment largely suppressed acute inflammation, it did not stop the chronic accrual pathology and neurodegeneration. However, it may be possible that treatment effects on neurodegeneration requires more time to be detected.

We have also investigated whether the relation between inflammation and neurodegeneration could be influenced by disease worsening. In this respect, higher brain activity in terms of WM lesions and atrophy were detected in patients who converted to MS.

Interestingly, the spatio-temporal concordance between inflammation and neurodegeneration seemed to take place mostly in the periventricular region, while the parietal and temporal lobe seems to be involved at different temporal intervals and in relation with the treatment and the activities of the patients. This greater periventricular activity could be related to the presence of locally secreted proinflammatory cytokines derived from CSF compartments harbor B cells that reside within the CSF space. Such inflammatory environment could lead to the presence of destructive lesions resulting in an increase in ventricular volume. Cortical atrophy could be explained by retrograde degeneration of axons injured in WM lesions (i.e. Wallerian degeneration). However, this phenomenon alone could not entirely account for cortex volume loss as several mechanisms, like neuronal shrinkage and demyelination, could contribute to cortical neurodegeneration.

Relation between WM lesion volume changes and subsequent brain atrophy

Summary. The main finding of this study was that higher WM lesion volume changes were related to faster periventricular atrophy in the next year. Further, no significant different association between

inflammation and neurodegeneration was detected across an untreated and treated period, thus meaning that treatment does not influence the causal relationship between these two processes.

Treatment only had a significant effect on atrophy and lesion volume changes during certain periods of the study. In the first year of treatment, our results were indicative of resolving oedema and pseudoatrophy. Congruent with the study on the concurrent relation between inflammation and neurodegeneration, this effect was not detected in the first year of treatment of the delay treated patients (year 3 of the study). Interestingly, in these patients we did find faster brain atrophy in year 4, which could be speculated to reflect a delayed pseudoatrophy effect. This potentially has an important clinical consequence, by highlighting a different response to early and delayed treatment in terms of brain atrophy and lesion accrual.

General Conclusions. Our experiments provided complementary results. Indeed, if on one hand higher WM lesion volume changes were related to subsequent faster atrophy, it is also true that inflammation and neurodegeneration developed simultaneously in the early phase of MS, thus suggesting that these two processes partially resulted from different and independent pathological mechanisms. These findings highlighted how the relation between inflammation and neurodegeneration is not restricted to a single direction but is more probably the sum of different models that are not mutually exclusive and could coexist at the same time.

Future perspectives. In elucidating the relation between inflammation and neurodegeneration, we did focus our analyses only on the relation between WM lesions and global brain and ventricular atrophy. Several studies have highlighted the presence of GM damage in the early phase of MS. Thus, future studies will address this issue by implementing new generation of imaging processing methods able to provide robust and accurate GM volumes estimates (i.e. SIENA-XL). Further, it would be very interesting to test whether WM lesions accrual in specific brain tracts is related to damage in “anatomically contiguous/connected” cortical lobes. Future efforts should aim to increase the follow-

up period to better elucidate the relation between the two pathological processes. Finally, the question whether inflammation and neurodegeneration are causally related or could develop independently is still a topic of discussion and our results did not provide a definite solution. In this respect, future studies should also focus on the pathological model where WM damage could be secondary to neurodegeneration.

8. References

- Abdullah, A. B., Younis, A. A., Pattany, P. M., & Saraf-Lavi, E. (2011). Textural Based SVM for MS Lesion Segmentation in FLAIR MRIs. *Open Journal of Medical Imaging, 1*(2), 26-42. doi:10.4236/ojmi.2011.12005
- Anbeek, P., Vincken, K. L., van Osch, M. J., Bisschops, R. H., & van der Grond, J. (2004). Probabilistic segmentation of white matter lesions in MR imaging. *NeuroImage, 21*(3), 1037-1044. doi:https://doi.org/10.1016/j.neuroimage.2003.10.012
- Andersson, J. L., Jenkinson, M., & Smith, S. (2010). Non-linear registration, aka spatial normalisation. *FMRIB technical report TR07JA2*.
- Andravizou, A., Dardiotis, E., Artemiadis, A., Sokratous, M., Siokas, V., Tsouris, Z., . . . Hadjigeorgiou, G. M. (2019). Brain atrophy in multiple sclerosis: mechanisms, clinical relevance and treatment options. *Auto- immunity highlights, 10*(1), 7. doi:https://doi.org/10.1186/s13317-019-0117-5
- Bakshi, R., Czarnecki, D., Shaikh, Z. A., Priore, R. L., Janardhan, V., Kaliszky, Z., & Kinkel, P. R. (2000). Brain MRI lesions and atrophy are related to depression in multiple sclerosis. *Neuroreport, 11*(6), 1153–1158. doi:https://doi.org/10.1097/00001756-200004270-00003
- Barkhof, F., Calabresi, P. A., Miller, D. H., & Reingold, S. C. (2009). Imaging outcomes for neuroprotection and repair in multiple sclerosis trials. *Nature reviews. Neurology, 5*(5), 256-266. doi:https://doi.org/10.1038/nrneurol.2009.41
- Bartsch, A. J., Homola, G., Biller, A., Smith, S. M., Weijers, H. G., Wiesbeck, G. A., . . . Bendszus, M. (2007). Manifestations of early brain recovery associated with abstinence from alcoholism. *Brain : a journal of neurology, 130*(Pt 1), 36-47. doi:https://doi.org/10.1093/brain/awl303

- Battaglini, M., De Stefano, N., & Jenkinson, M. (2012). A fully automated, hierarchical classification methods for detecting white matter lesions in Multiple Sclerosis. *20th ISMRM Congress*.
- Battaglini, M., Gentile, G., Luchetti, L., Giorgio, A., Vrenken, H., Barkhof, F., . . . Preziosa, P. (2019). Lifespan normative data on rates of brain volume changes. *Neurobiology of aging*, *81*, 30-37. doi:<https://doi.org/10.1016/j.neurobiolaging.2019.05.010>
- Battaglini, M., Giorgio, A., Stromillo, M. L., Bartolozzi, M. L., Guidi, L., Federico, A., & De Stefano, N. (2009). Voxel-wise assessment of progression of regional brain atrophy in relapsing-remitting multiple sclerosis. *Journal of the neurological sciences*, *282*(1-2), 55-60. doi:<https://doi.org/10.1016/j.jns.2009.02.322>
- Battaglini, M., Jenkinson, M., De Stefano, N., & ADNI. (2018). SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI. *Human brain mapping*, *39*(3), 1063-1077. doi:<https://doi.org/10.1002/hbm.23828>
- Battaglini, M., Rossi, F., Grove, R. A., Stromillo, M. L., Whitcher, B., Matthews, P. M., & De Stefano, N. (2014). Automated identification of brain new lesions in multiple sclerosis using subtraction images. *Journal of magnetic resonance imaging : JMRI*, *39*(6), 1543-1549. doi:<https://doi.org/10.1002/jmri.24293>
- Battaglini, M., Vrenken, H., Tappa Brocci, R., Gentile, G., Luchetti, L., Versteeg, A., . . . De Stefano, N. (In Preparation). Evolution from FCDE to MS in the REFLEX trial: Regional susceptibility in the conversion to MS at disease onset and their amenability to subcutaneous interferon beta-1a.
- Bermel, R. A., & Bakshi, R. (2006). The measurement and clinical relevance of brain atrophy in multiple sclerosis. *Neurology*, *5*(2), 158-170. doi:[https://doi.org/10.1016/S1474-4422\(06\)70349-0](https://doi.org/10.1016/S1474-4422(06)70349-0)

- Bø, L., Vedeler, C. A., Nyland, H., Trapp, B. D., & Mørk, S. J. (2003). Intracortical multiple sclerosis lesions are not associated with increased lymphocyte infiltration. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 9(4), 323-331.
doi:<https://doi.org/10.1191/1352458503ms917oa>
- Bodini, B., Chard, D., Altmann, D. R., Tozer, D., Miller, D. H., Thompson, A. J., . . . Ciccarelli, O. (2016). White and gray matter damage in primary progressive MS: The chicken or the egg? *Neurology*, 86(2), 170-176. doi:<https://doi.org/10.1212/WNL.0000000000002237>
- Bodini, B., Khaleeli, Z., Cercignani, M., Miller, D. H., Thompson, A. J., & Ciccarelli, O. (2009). Exploring the relationship between white matter and gray matter damage in early primary progressive multiple sclerosis: an in vivo study with TBSS and VBM. *Human brain mapping*, 30(9), 2852-2861. doi:<https://doi.org/10.1002/hbm.20713>
- Bordin, V., Bertani, I., Mattioli, I., Sundaresan, V., McCarthy, P., Suri, S., . . . Mackay, C. E. (2021). Integrating large-scale neuroimaging research datasets: Harmonisation of white matter hyperintensity measurements across Whitehall and UK Biobank datasets. *NeuroImage*, 237, 118189. doi:<https://doi.org/10.1016/j.neuroimage.2021.118189>
- Bross, M., Hackett, M., & Bernitsas, E. (2020). Approved and Emerging Disease Modifying Therapies on Neurodegeneration in Multiple Sclerosis. *International journal of molecular sciences*, 21(12), 4312. doi:<https://doi.org/10.3390/ijms21124312>
- Brown, J. W., Pardini, M., Brownlee, W. J., Fernando, K., Samson, R. S., Prados Carrasco, F., . . . Chard, D. T. (2017). An abnormal periventricular magnetization transfer ratio gradient occurs early in multiple sclerosis. *Brain : a journal of neurology*, 140(2), 387-398.
doi:<https://doi.org/10.1093/brain/aww296>
- Brownlee, W. J., Altmann, D. R., Prados, F., Miskiel, K. A., Eshaghi, A., Gandini Wheeler-Kingshott, C., . . . Ciccarelli, O. (2019). Early imaging predictors of long-term outcomes in

relapse-onset multiple sclerosis. *Brain : a journal of neurology*, 142(8), 2276-2287.

doi:<https://doi.org/10.1093/brain/awz156>

Cappellani, R., Bergsland, N., Weinstock-Guttman, B., Kennedy, C., Carl, E., Ramasamy, D. P., . . .

Zivadinov, R. (2014). Diffusion tensor MRI alterations of subcortical deep gray matter in clinically isolated syndrome. *Journal of the neurological sciences*, 338(1-2), 128-134.

doi:<https://doi.org/10.1016/j.jns.2013.12.031>

Chard, D. T., Brex, P. A., Ciccarelli, O., Griffin, C. M., Parker, G. J., Dalton, C., . . . Miller, D. H.

(2003). The longitudinal relation between brain lesion load and atrophy in multiple sclerosis: a 14 year follow up study. *Journal of neurology, neurosurgery, and psychiatry*,

74(11), 1551-1554. doi:<https://doi.org/10.1136/jnnp.74.11.1551>

Comi, G., De Stefano, N., Freedman, M. S., Barkhof, F., Polman, C. H., Uitdehaag, B. M., . . .

Kappos, L. (2012). Comparison of two dosing frequencies of subcutaneous interferon beta-1a in patients with a first clinical demyelinating event suggestive of multiple sclerosis (REFLEX): a phase 3 randomised controlled trial. *The Lancet. Neurology*, 11(1), 33-41.

doi:[https://doi.org/10.1016/S1474-4422\(11\)70262-9](https://doi.org/10.1016/S1474-4422(11)70262-9)

Comi, G., De Stefano, N., Freedman, M. S., Barkhof, F., Uitdehaag, B. M., de Vos, M., . . . Kappos,

L. (2017). Subcutaneous interferon β -1a in the treatment of clinically isolated syndromes: 3-year and 5-year results of the phase III dosing frequency-blind multicentre REFLEXION study. *Journal of neurology, neurosurgery, and psychiatry*, 88(4), 285-294.

doi:<https://doi.org/10.1136/jnnp-2016-314843>

Dalton, C. M., Brex, P. A., Jenkins, R., Fox, N. C., Miskiel, K. A., Crum, W. R., . . . Miller, D. H.

(2002). Progressive ventricular enlargement in patients with clinically isolated syndromes is associated with the early development of multiple sclerosis. *Journal of neurology, neurosurgery, and psychiatry*,

73(2), 141-147. doi:<https://doi.org/10.1136/jnnp.73.2.141>

- Dalton, C. M., Chard, D. T., Davies, G. R., Miszkiel, K. A., Altmann, D. R., Fernando, K., . . . Miller, D. H. (2004). Early development of multiple sclerosis is associated with progressive gray matter atrophy in patients presenting with clinically isolated syndromes. *Brain : a journal of neurology*, *127*(Pt 5), 1101-1107. doi:<https://doi.org/10.1093/brain/awh126>
- Danelakis, A., Theoharis, T., & Verganelakis, D. A. (2018). Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, *70*, 83-100. doi:<https://doi.org/10.1016/j.compmedimag.2018.10.002>
- Datta, S., & Narayana, P. A. (2013). A comprehensive approach to the segmentation of multichannel three-dimensional MR brain images in multiple sclerosis. *NeuroImage. Clinical*, *2*, 184-196. doi:<https://doi.org/10.1016/j.nicl.2012.12.007>
- de Sitter, A., Steenwijk, M. D., Ruet, A., Versteeg, A., Liu, Y., van Schijndel, R. A., . . . Sastre-Garriga, J. (2017). Performance of five research-domain automated WM lesion segmentation methods in a multi-center MS study. *NeuroImage*, *163*, 106-114. doi:<https://doi.org/10.1016/j.neuroimage.2017.09.011>
- De Stefano, N., & Arnold, D. L. (2015). Towards a better understanding of pseudoatrophy in the brain of multiple sclerosis patients. *Multiple sclerosis (Houndmills, Basingstoke, England)*, *21*(6), 675-676. doi:<https://doi.org/10.1177/1352458514564494>
- De Stefano, N., Airas, L., Grigoriadis, N., Mattle, H. P., O'Riordan, J., Oreja-Guevara, C., . . . Kieseier, B. C. (2014). Clinical relevance of brain volume measures in multiple sclerosis. *CNS drugs*, *28*(2), 147-156. doi:<https://doi.org/10.1007/s40263-014-0140-z>
- De Stefano, N., Gentile, G., Luchetti, L., Giorgio, A., Zhang, J., Vrenken, H., . . . Gallo, A. (2018). Large-scale normative volumes of brain structures as assessed by SIENAX. *34th ECTRIMS Congress*.

- De Stefano, N., Giorgio, A., Gentile, G., Stromillo, M. L., Cortese, R., Gasperini, C., . . . Battaglini, M. (2021). Dynamics of pseudo-atrophy in RRMS reveals predominant gray matter compartmentalization. *Annals of clinical and translational neurology*, 8(3), 623-630. doi:<https://doi.org/10.1002/acn3.51302>
- De Stefano, N., Jenkinson, M., Guidi, L., Bartolozzi, M. L., Federico, A., & Smith, S. M. (2003). Voxel-Level Cross-Subject Statistical Analysis of Brain Atrophy in early Relapsing Remitting MS patients. *Int Soc Magn Reson Med (Book Of Abstracts)*, 2625, 2.
- De Stefano, N., Stromillo, M. L., Giorgio, A., Bartolozzi, M. L., Battaglini, M., Baldini, M., . . . Sormani, M. P. (2016). Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. *Journal of neurology, neurosurgery, and psychiatry*, 87(1), 93-99. doi:<https://doi.org/10.1136/jnnp-2014-309903>
- DeLuca, G. C., Williams, K., Evangelou, N., Ebers, G. C., & Esiri, M. M. (2006). The contribution of demyelination to axonal loss in multiple sclerosis. *Brain : a journal of neurology*, 129(Pt 6), 1507-1516. doi:<https://doi.org/10.1093/brain/awl074>
- Dendrou, C. A., Fugger, L., & Friese, M. A. (2015). Immunopathology of multiple sclerosis. *Nature reviews. Immunology*, 15(9), 545–558. doi:<https://doi.org/10.1038/nri3871>
- Duong, M. T., Rudie, J. D., Wang, J., Xie, L., Mohan, S., Gee, J. C., & Rauschecker, A. M. (2019). Convolutional Neural Network for Automated FLAIR Lesion Segmentation on Clinical Brain MR Imaging. *AJNR. American journal of neuroradiology*, 40(8), 1282-1290. doi:<https://doi.org/10.3174/ajnr.A6138>
- Eshaghi, A., Prados, F., Brownlee, W. J., Altmann, D. R., Tur, C., Cardoso, M. J., . . . Rovira, A. (2018). Deep gray matter volume loss drives disability worsening in multiple sclerosis. *Annals of neurology*, 83(2), 210-222. doi:<https://doi.org/10.1002/ana.25145>

- Filippi, M., Bar-Or, A., Piehl, F., Preziosa, P., Solari, A., Vukusic, S., & Rocca, M. A. (2018). Multiple sclerosis. *Nature reviews. Disease primers*, *4*(1), 43.
doi:<https://doi.org/10.1038/s41572-018-0041-4>
- Filippi, M., Horsfield, M. A., Bressi, S., Martinelli, V., Baratti, C., Reganati, P., . . . Comi, G. (1995). Intra- and inter-observer agreement of brain MRI lesion volume measurements in multiple sclerosis. A comparison of techniques. *Brain : a journal of neurology*, *118*(Pt 6), 1593-1600. doi:<https://doi.org/10.1093/brain/118.6.1593>
- Filippi, M., Preziosa, P., & Rocca, M. A. (2018). MRI in multiple sclerosis: what is changing? *Current opinion in neurology*, *31*(4), 386-395.
doi:<https://doi.org/10.1097/WCO.0000000000000572>
- Filippi, M., Preziosa, P., Banwell, B. L., Barkhof, F., Ciccarelli, O., De Stefano, N., . . . Rocca, M. A. (2019). Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain : a journal of neurology*, *142*(7), 1858-1875.
doi:<https://doi.org/10.1093/brain/awz144>
- Filippi, M., Rocca, M. A., Barkhof, F., Brück, W., Chen, J. T., Comi, G., . . . Lassmann, H. (2012). Association between pathological and MRI findings in multiple sclerosis. *The Lancet. Neurology*, *11*(4), 349-360. doi:[https://doi.org/10.1016/S1474-4422\(12\)70003-0](https://doi.org/10.1016/S1474-4422(12)70003-0)
- Filippi, M., Rocca, M. A., Martino, G., Horsfield, M. A., & Comi, G. (1998). Magnetization transfer changes in the normal appearing white matter precede the appearance of enhancing lesions in patients with multiple sclerosis. *Annals of neurology*, *43*(6), 809-814.
doi:<https://doi.org/10.1002/ana.410430616>
- Fisher, E., Rudick, R. A., Simon, J. H., Cutter, G., Baier, M., Lee, J. C., . . . Simonian, N. A. (2002). Eight-year follow-up study of brain atrophy in patients with MS. *Neurology*, *59*(9), 1412-1420. doi:<https://doi.org/10.1212/01.wnl.0000036271.49066.06>

Fisniku, L. K., Brex, P. A., Altmann, D. R., Miszkief, K. A., Benton, C. E., Lanyon, R., . . . Miller, D. H. (2008). Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain : a journal of neurology*, *131*(Pt 3), 808-817.

doi:<https://doi.org/10.1093/brain/awm329>

Gabr, R. E., Coronado, I., Robinson, M., Sujit, S. J., Datta, S., Sun, X., . . . Narayana, P. A. (2020). Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. *Multiple sclerosis (Houndmills, Basingstoke, England)*,

26(10), 1217-1226. doi:<https://doi.org/10.1177/1352458519856843>

Ganna, M., Rombaut, M., Goutte, R., & Zhu, Y. (2002). Improvement of brain lesions detection using information fusion approach. *6th International Conference on Signal Processing Proceedings*.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., & Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical image analysis*, *17*(1), 1-18.

doi:<https://doi.org/10.1016/j.media.2012.09.004>

Gauthier, S. A., Mandel, M., Guttmann, C. R., Glanz, B. I., Khoury, S. J., Betensky, R. A., & Weiner, H. L. (2007). Predicting short-term disability in multiple sclerosis. *Neurology*,

68(24), 2059–2065. doi:<https://doi.org/10.1212/01.wnl.0000264890.97479.b1>

GBD 2016 Neurology Collaborators. (2019). Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016.

The Lancet. Neurology, *18*(5), 459-480. doi:[https://doi.org/10.1016/S1474-4422\(18\)30499-](https://doi.org/10.1016/S1474-4422(18)30499-X)

X

Genovese, A. V., Hagemeier, J., Bergsland, N., Jakimovski, D., Dwyer, M. G., Ramasamy, D. P., . . . Zivadinov, R. (2019). Atrophied Brain T2 Lesion Volume at MRI Is Associated with

- Disability Progression and Conversion to Secondary Progressive Multiple Sclerosis.
Radiology, 293(2), 424-433. doi:<https://doi.org/10.1148/radiol.2019190306>
- Geurts, J. J., Pouwels, P. J., Uitdehaag, B. M., Polman, C. H., Barkhof, F., & Castelijns, J. A. (2005). Intracortical lesions in multiple sclerosis: improved detection with 3D double inversion-recovery MR imaging. *Radiology*, 236(1), 254-260.
doi:<https://doi.org/10.1148/radiol.2361040450>
- Giorgio, A., Battaglini, M., Smith, S. M., & De Stefano, N. (2008). Brain atrophy assessment in multiple sclerosis: importance and limitations. *Neuroimaging clinics of North America*, 18(4), 675-xi. doi:<https://doi.org/10.1016/j.nic.2008.06.007>
- González Ballester, M. A., Zisserman, A. P., & Brady, M. (2002). Estimation of the partial volume effect in MRI. *Medical image analysis*, 6(4), 389-405. doi:[https://doi.org/10.1016/s1361-8415\(02\)00061-0](https://doi.org/10.1016/s1361-8415(02)00061-0)
- Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., . . . Jenkinson, M. (2016). BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141, 191-205.
doi:<https://doi.org/10.1016/j.neuroimage.2016.07.018>
- Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G. J., . . . Miller, D. H. (1996). Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. *Magnetic resonance imaging*, 14(5), 495-505.
doi:[https://doi.org/10.1016/0730-725x\(96\)00018-5](https://doi.org/10.1016/0730-725x(96)00018-5)
- Guo, C. N., Shi, L., Wang, Z., Zhao, M., Wang, D., Zhu, W., . . . Sun, L. (2019). Intra-Scanner and Inter-Scanner Reproducibility of Automatic White Matter Hyperintensities Quantification. *Frontiers in neuroscience*, 13, 679. doi:<https://doi.org/10.3389/fnins.2019.00679>

Heinen, R., Steenwijk, M. D., Barkhof, F., Biesbroek, J. M., van der Flier, W. M., Kuijf, H. J., . . .

group, T.-V. s. (2019). Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset. *Scientific reports*, *9*(1), 16742.

doi:<https://doi.org/10.1038/s41598-019-52966-0>

Henry, R. G., Shieh, M., Okuda, D. T., Evangelista, A., Gorno-Tempini, M. L., & Pelletier, D.

(2008). Regional gray matter atrophy in clinically isolated syndromes at presentation.

Journal of neurology, neurosurgery, and psychiatry, *79*(11), 1236-1244.

doi:<https://doi.org/10.1136/jnnp.2007.134825>

Ingle, G. T., Stevenson, V. L., Miller, D. H., & Thompson, A. J. (2003). Primary progressive

multiple sclerosis: a 5-year clinical and MR study. *Brain : a journal of neurology*, *126*(Pt

11), 2528-2536. doi:<https://doi.org/10.1093/brain/awg261>

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of

brain images. *Medical image analysis*, *5*(2), 143–156 . doi:[https://doi.org/10.1016/s1361-](https://doi.org/10.1016/s1361-8415(01)00036-6)

[8415\(01\)00036-6](https://doi.org/10.1016/s1361-8415(01)00036-6)

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust

and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2),

825-841. doi:[https://doi.org/10.1016/s1053-8119\(02\)91132-8](https://doi.org/10.1016/s1053-8119(02)91132-8)

Kalincik, T., Vaneckova, M., Tyblova, M., Krasensky, J., Seidl, Z., Havrdova, E., & Horakova, D.

(2012). Volumetric MRI markers and predictors of disease activity in early multiple

sclerosis: a longitudinal cohort study. *PloS one*, *7*(11), e50101.

doi:<https://doi.org/10.1371/journal.pone.0050101>

Kaur, A., Kaur, L., & Singh, A. (2021). State-of-the-Art Segmentation Techniques and Future

Directions for Multiple Sclerosis Brain Lesions. *Arch Computat Methods Eng*, *28*, 951–977.

doi:<https://doi.org/10.1007/s11831-020-09403-7>

- Khastavanehm, H., & Haron, H. (2014). False Positives Reduction on Segmented Multiple Sclerosis Lesions Using Fuzzy Inference System by Incorporating Atlas Prior Anatomical Knowledge: A Conceptual Model. *Computational Collective Intelligence. Technologies and Applications. Springer International Publishing*. doi:10.1007/978-3-319-11289-3_2
- Kornek, B., Storch, M. K., Weissert, R., Wallstroem, E., Stefferl, A., Olsson, T., . . . Lassmann, H. (2000). Multiple sclerosis and chronic autoimmune encephalomyelitis: a comparative quantitative study of axonal injury in active, inactive, and remyelinated lesions. *The American journal of pathology*, *157*(1), 267-276. doi:https://doi.org/10.1016/S0002-9440(10)64537-3
- Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., . . . Davatzikos, C. (2008). Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Academic radiology*, *15*(3), 300-313. doi:https://doi.org/10.1016/j.acra.2007.10.012
- Lassmann, H. (2007). Multiple sclerosis: is there neurodegeneration independent from inflammation? *Journal of the neurological sciences*, *259*(1), 3-6. doi:https://doi.org/10.1016/j.jns.2006.08.016
- Lassmann, H., Brück, W., & Lucchinetti, C. F. (2007). The immunopathology of multiple sclerosis: an overview. *Brain pathology (Zurich, Switzerland)*, *17*(2), 210-218. doi:https://doi.org/10.1111/j.1750-3639.2007.00064.x
- Lassmann, H., van Horssen, J., & Mahad, D. (2012). Progressive multiple sclerosis: pathology and pathogenesis. *Nature reviews. Neurology*, *8*(11), 647-656. doi:https://doi.org/10.1038/nrneurol.2012.168

- Liu, Z., Pardini, M., Yaldizli, Ö., Sethi, V., Muhlert, N., Wheeler-Kingshott, C. A., . . . Chard, D. T. (2015). Magnetization transfer ratio measures in normal-appearing white matter show periventricular gradient abnormalities in multiple sclerosis. *Brain : a journal of neurology*, *138*(Pt 5), 1239-1246. doi:<https://doi.org/10.1093/brain/awv065>
- Lublin, F. D., & Reingold, S. C. (1996). Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. *Neurology*, *46*(4), 907-911. doi:<https://doi.org/10.1212/wnl.46.4.907>
- Lucchinetti, C., Brück, W., Parisi, J., Scheithauer, B., Rodriguez, M., & Lassmann, H. (2000). Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. *Annals of neurology*, *47*(6), 707-717. doi:[https://doi.org/10.1002/1531-8249\(200006\)47:6<707::aid-ana3>3.0.co;2-q](https://doi.org/10.1002/1531-8249(200006)47:6<707::aid-ana3>3.0.co;2-q)
- Luchetti, L., Gentile, G., Battaglini, M., Giorgio, A., & De Stefano, N. (2019). SIENAX2.0, an update of SIENAX tool for cross sectional brain. *X AIRMM (Italian Association of Magnetic Resonance Imaging in Medicine) CONGRESS*.
- Magliozzi, R., Howell, O. W., Nicholas, R., Cruciani, C., Castellaro, M., Romualdi, C., . . . Reynolds, R. (2018). Inflammatory intrathecal profiles and cortical damage in multiple sclerosis. *Annals of neurology*, *83*(4), 739-755. doi:<https://doi.org/10.1002/ana.25197>
- Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., . . . Vellas, B. (2020). The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. *Medical image analysis*, *66*, 101714. doi:<https://doi.org/10.1016/j.media.2020.101714>
- Miller, D. H., Barkhof, F., Frank, J. A., Parker, G. J., & Thompson, A. J. (2002). Measurement of atrophy in multiple sclerosis: pathological basis, methodological aspects and clinical

relevance. *Brain : a journal of neurology*, 125(Pt 8), 1676-1695.

doi:<https://doi.org/10.1093/brain/awf177>

Milo, R., Korczyn, A. D., Manouchehri, N., & Stüve, O. (2020). The temporal and causal relationship between inflammation and neurodegeneration in multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 26(8), 876-886.

doi:<https://doi.org/10.1177/1352458519886943>

Minagar, A., & Alexander, J. S. (2003). Blood-brain barrier disruption in multiple sclerosis.

Multiple sclerosis (Houndmills, Basingstoke, England), 9(6), 540-549.

doi:<https://doi.org/10.1191/1352458503ms965oa>

Moraal, B., van den Elskamp, I. J., Knol, D. L., Uitdehaag, B. M., Geurts, J. J., Vrenken, H., . . .

Barkhof, F. (2010). Long-interval T2-weighted subtraction magnetic resonance imaging: a powerful new outcome measure in multiple sclerosis trials. *Annals of neurology*, 67(5), 667-675. doi:<https://doi.org/10.1002/ana.21958>

Moraal, B., Wattjes, M. P., Geurts, J. J., Knol, D. L., van Schijndel, R. A., Pouwels, P. J., . . .

Barkhof, F. (2010). Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. *Radiology*, 255(1), 154-163. doi:<https://doi.org/10.1148/radiol.09090814>

Nakamura, K., Guizard, N., Fonov, V. S., Narayanan, S., Collins, D. L., & Arnold, D. L. (2013).

Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. *NeuroImage. Clinical*, 4, 10-17.

doi:<https://doi.org/10.1016/j.nicl.2013.10.015>

Narayana, P. A., Coronado, I., Sujit, S. J., Wolinsky, J. S., Lublin, F. D., & Gabr, R. E. (2020).

Deep-Learning-Based Neural Tissue Segmentation of MRI in Multiple Sclerosis: Effect of Training Set Size. *Journal of magnetic resonance imaging : JMRI*, 51(5), 1487-1496.

doi:<https://doi.org/10.1002/jmri.26959>

Narayana, P. A., Doyle, T. J., Lai, D., & Wolinsky, J. S. (1998). Serial proton magnetic resonance spectroscopic imaging, contrast-enhanced magnetic resonance imaging, and quantitative lesion volumetry in multiple sclerosis. *Annals of neurology*, 43(1), 56-71.

doi:<https://doi.org/10.1002/ana.410430112>

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. American journal of neuroradiology*, 28(9), 1645-1649.

doi:<https://doi.org/10.3174/ajnr.A0645>

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1), 1-25.

doi:<https://doi.org/10.1002/hbm.1058>

Noseworthy, J. H., Lucchinetti, C., Rodriguez, M., & Weinshenker, B. G. (2000). Multiple sclerosis. *The New England journal of medicine*, 343(13), 938-952.

doi:<https://doi.org/10.1056/NEJM200009283431307>

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

doi:<https://doi.org/10.1093/brain/121.3.495>

Ortiz, G. G., Pacheco-Moisés, F. P., Macías-Islas, M. Á., Flores-Alvarado, L. J., Mireles-Ramírez, M. A., González-Renovato, E. D., . . . Alatorre-Jiménez, M. A. (2014). Role of the blood-brain barrier in multiple sclerosis. *Archives of medical research*, 45(8), 687-697.

doi:<https://doi.org/10.1016/j.arcmed.2014.11.013>

- Paniagua Bravo, Á., Sánchez Hernández, J. J., Ibáñez Sanz, L., Alba de Cáceres, I., Crespo San José, J. L., & García-Castaño Gandariaga, B. (2014). A comparative MRI study for white matter hyperintensities detection: 2D-FLAIR, FSE PD 2D, 3D-FLAIR and FLAIR MIP. *The British journal of radiology*, *87*(1035), 20130360. doi:<https://doi.org/10.1259/bjr.20130360>
- Paolillo, A., Piattella, M. C., Pantano, P., Di Legge, S., Caramia, F., Russo, P., . . . Pozzilli, C. (2004). The relationship between inflammation and atrophy in clinically isolated syndromes suggestive of multiple sclerosis: a monthly MRI study after triple-dose gadolinium-DTPA. *Journal of neurology*, *251*(4), 432-439. doi:<https://doi.org/10.1007/s00415-004-0349-8>
- 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*(3), 389-400. doi:<https://doi.org/10.1002/ana.1123>
- Polman, C. H., Reingold, S. C., Edan, G., Filippi, M., Hartung, H. P., Kappos, L., . . . Wolinsky, J. S. (2005). Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Annals of neurology*, *58*(6), 840-846. doi:<https://doi.org/10.1002/ana.20703>
- Popescu, V., Battaglini, M., Hoogstrate, W. S., Verfaillie, S. C., Sluimer, I. C., van Schijndel, R. A., . . . Group, M. S. (2012). Optimizing parameter choice for FSL-Brain Extraction Tool (BET) on 3D T1 images in multiple sclerosis. *NeuroImage*, *61*(4), 1484–1494. doi:<https://doi.org/10.1016/j.neuroimage.2012.03.074>
- Popescu, V., Klaver, R., Voorn, P., Galis-de Graaf, Y., Knol, D. L., Twisk, J. W., . . . Geurts, J. J. (2015). What drives MRI-measured cortical atrophy in multiple sclerosis? *Multiple sclerosis (Houndmills, Basingstoke, England)*, *21*(10), 1280-1290. doi:<https://doi.org/10.1177/1352458514562440>

- Prineas, J. W., Kwon, E. E., Cho, E. S., Sharer, L. R., Barnett, M. H., Oleszak, E. L., . . . Morgan, B. P. (2001). Immunopathology of secondary-progressive multiple sclerosis. *Annals of neurology*, *50*(5), 646-657. doi:<https://doi.org/10.1002/ana.1255>
- Radue, E. W., Barkhof, F., Kappos, L., Sprenger, T., Häring, D. A., de Vera, A., . . . Cohen, J. A. (2015). Correlation between brain volume loss and clinical and MRI outcomes in multiple sclerosis. *Neurology*, *84*(8), 784-793. doi:<https://doi.org/10.1212/WNL.0000000000001281>
- Rao, S. M., Leo, G. J., Haughton, V. M., St Aubin-Faubert, P., & Bernardin, L. (1989). Correlation of magnetic resonance imaging with neuropsychological testing in multiple sclerosis. *Neurology*, *39*(2 Pt 1), 161-166. doi:<https://doi.org/10.1212/wnl.39.2.161>
- Raz, E., Cercignani, M., Sbardella, E., Totaro, P., Pozzilli, C., Bozzali, M., & Pantano, P. (2010). Clinically isolated syndrome suggestive of multiple sclerosis: voxelwise regional investigation of white and gray matter. *Radiology*, *254*(1), 227-234. doi:<https://doi.org/10.1148/radiol.2541090817>
- Raz, E., Cercignani, M., Sbardella, E., Totaro, P., Pozzilli, C., Bozzali, M., & Pantano, P. (2010). Gray- and white-matter changes 1 year after first clinical episode of multiple sclerosis: MR imaging. *Radiology*, *257*(2), 448-454. doi:<https://doi.org/10.1148/radiol.10100626>
- Reich, D. S., Lucchinetti, C. F., & Calabresi, P. A. (2018). Multiple Sclerosis. *The New England journal of medicine*, *378*(2), 169-180. doi:<https://doi.org/10.1056/NEJMra1401483>
- Riccitelli, G., Rocca, M. A., Forn, C., Colombo, B., Comi, G., & Filippi, M. (2011). Voxelwise assessment of the regional distribution of damage in the brains of patients with multiple sclerosis and fatigue. *AJNR. American journal of neuroradiolog*, *32*(5), 874-879. doi:<https://doi.org/10.3174/ajnr.A2412>
- Richert, N. D., Howard, T., Frank, J. A., Stone, R., Ostuni, J., Ohayon, J., . . . McFarland, H. F. (2006). Relationship between inflammatory lesions and cerebral atrophy in multiple

- sclerosis. *Neurology*, 66(4), 551-556.
doi:<https://doi.org/10.1212/01.wnl.0000197982.78063.06>
- Rocca, M. A., Battaglini, M., Benedict, R. H., De Stefano, N., Geurts, J. J., Henry, R. G., . . . Filippi, M. (2017). Brain MRI atrophy quantification in MS: From methods to clinical application. *Neurology*, 88(4), 403-413.
doi:<https://doi.org/10.1212/WNL.0000000000003542>
- Rocca, M. A., Preziosa, P., Mesaros, S., Pagani, E., Dackovic, J., Stosic-Opincal, T., . . . Filippi, M. (2016). Clinically Isolated Syndrome Suggestive of Multiple Sclerosis: Dynamic Patterns of Gray and White Matter Changes-A 2-year MR Imaging Study. *Radiology*, 278(3), 841-853.
doi:<https://doi.org/10.1148/radiol.2015150532>
- Roosendaal, S. D., Bendfeldt, K., Vrenken, H., Polman, C. H., Borgwardt, S., Radue, . . . Geurts, J. J. (2011). Gray matter volume in a large cohort of MS patients: relation to MRI parameters and disability. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 17(9), 1098-1106.
doi:<https://doi.org/10.1177/1352458511404916>
- Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J. C., . . . Lladó, X. (2015). A toolbox for multiple sclerosis lesion segmentation. *Neuroradiology*, 57(10), 1031-1043.
doi:<https://doi.org/10.1007/s00234-015-1552-2>
- Sailer, M., Losseff, N. A., Wang, L., Gawne-Cain, M. L., Thompson, A. J., & Miller, D. H. (2001). T1 lesion load and cerebral atrophy as a marker for clinical progression in patients with multiple sclerosis. A prospective 18 months follow-up study. *European journal of neurology*, 8(1), 37-42. doi:<https://doi.org/10.1046/j.1468-1331.2001.00147.x>
- 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. *Annals of biomedical engineering*, 34(1), 142-151. doi:<https://doi.org/10.1007/s10439-005-9009-0>

- Sastre-Garriga, J., Tur, C., Pareto, D., Vidal-Jordana, A., Auger, C., Río, J., . . . Montalban, X. (2015). Brain atrophy in natalizumab-treated patients: A 3-year follow-up. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 21(6), 749-756.
doi:<https://doi.org/10.1177/1352458514556300>
- Schmidt, P. (2017). Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. *Dissertation, LMU München: Faculty of Mathematics, Computer Science and Statistics*.
- Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förchler, A., Berthele, A., . . . Mühlau, M. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage*, 59(4), 3774-3783.
doi:<https://doi.org/10.1016/j.neuroimage.2011.11.032>
- Shanmuganathan, M., Almutairi, S., Aborokbah, S. M., Ganesan, G., & Ramachandran, V. (2020). Review of advanced computational approaches on multiple sclerosis segmentation and classification. *IET Signal Processing*, 14(6), 333. doi:<https://doi.org/10.1049/iet-spr.2019.0543>
- Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J., . . . Sicotte, N. L. (2017). Volumetric Analysis from a Harmonized Multisite Brain MRI Study of a Single Subject with Multiple Sclerosis. *AJNR. American journal of neuroradiology*, 38(8), 1501-1509. doi:<https://doi.org/10.3174/ajnr.A5254>
- Shwartzman, O., Gazit, H., Shelef, I., & Riklin-Raviv, T. (2019). The Worrisome Impact of an Inter-rater Bias on Neural Network Training. *In arXiv, [eess.IV]*.
- Simon, H. (2014). MRI outcomes in the diagnosis and disease course of multiple sclerosis. *Handbook of clinical neurology*, 122, 405-425. doi:<https://doi.org/10.1016/B978-0-444-52001-2.00017-0>

- Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping, 17*(3), 143-155. doi:<https://doi.org/10.1002/hbm.10062>
- Smith, S. M., De Stefano, N., Jenkinson, M., & Matthews, P. M. (2001). Normalized accurate measurement of longitudinal brain change. *Journal of computer assisted tomography, 25*(3), 466-475. doi:<https://doi.org/10.1097/00004728-200105000-00022>
- Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., . . . Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage, 23*(Suppl 1), S208-219. doi:<https://doi.org/10.1016/j.neuroimage.2004.07.051>
- Smith, S. M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P. M., Federico, A., & De Stefano, N. (2002). Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage, 17*(1), 479-489. doi:<https://doi.org/10.1006/nimg.2002.1040>
- Tauhid, S., Neema, M., Healy, B. C., Weiner, H. L., & Bakshi, R. (2014). MRI phenotypes based on cerebral lesions and atrophy in patients with multiple sclerosis. *Journal of the neurological sciences, 346*(1-2), 250-254. doi:<https://doi.org/10.1016/j.jns.2014.08.047>
- Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., . . . Miller, D. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet. Neurology, 17*(2), 162-173. doi:[https://doi.org/10.1016/S1474-4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2)
- Tintoré, M., Rovira, A., Río, J., Nos, C., Grivé, E., Téllez, N., . . . Montalban, X. (2006). Baseline MRI predicts future attacks and disability in clinically isolated syndromes. *Neurology, 67*(6), 968-972. doi:<https://doi.org/10.1212/01.wnl.0000237354.10144.ec>
- Tintore, M., Rovira, À., Río, J., Otero-Romero, S., Arrambide, G., Tur, C., . . . Montalban, X. (2015). Defining high, medium and low impact prognostic factors for developing multiple

- sclerosis. *Brain : a journal of neurology*, 138(Pt 7), 1863-1874.
doi:<https://doi.org/10.1093/brain/awv105>
- Trapp, B. D., & Nave, K. A. (2008). Multiple sclerosis: an immune or neurodegenerative disorder? *Annual review of neuroscience*, 31, 247-269.
doi:<https://doi.org/10.1146/annurev.neuro.30.051606.094313>
- Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6), 1310-1320. doi:<https://doi.org/10.1109/TMI.2010.2046908>
- 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. *IEEE transactions on medical imaging*, 16(5), 598-609. doi:<https://doi.org/10.1109/42.640750>
- Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., . . . Lladó, X. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155, 159-168.
doi:<https://doi.org/10.1016/j.neuroimage.2017.04.034>
- Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., . . . Lladó, X. (2019). One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage. Clinical*, 21, 101638.
doi:<https://doi.org/10.1016/j.nicl.2018.101638>
- van Munster, C. E., & Uitdehaag, B. M. (2017). Outcome Measures in Clinical Trials for Multiple Sclerosis. *CNS drugs*, 31(3), 217-236. doi:<https://doi.org/10.1007/s40263-017-0412-5>
- Varosanec, M., Uher, T., Horakova, D., Hagemeyer, J., Bergsland, N., Tyblova, M., . . . Zivadinov, R. (2015). Longitudinal Mixed-Effect Model Analysis of the Association between Global and Tissue-Specific Brain Atrophy and Lesion Accumulation in Patients with Clinically

- Isolated Syndrome. *AJNR. American journal of neuroradiology*, 36(8), 1457-1464.
doi:<https://doi.org/10.3174/ajnr.A4330>
- Vidal-Jordana, A., Sastre-Garriga, J., Pérez-Miralles, F., Tur, C., Tintoré, M., Horga, A., . . .
Montalban, X. (2013). Early brain pseudoatrophy while on natalizumab therapy is due to
white matter volume changes. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 19(9),
1175-1181. doi:<https://doi.org/10.1177/1352458512473190>
- Vrenken, H., Vos, E. K., van der Flier, W. M., Sluimer, I. C., Cover, K. S., Knol, D. L., & Barkhof,
F. (2014). Validation of the automated method VIENA: an accurate, precise, and robust
measure of ventricular enlargement. *Human brain mapping*, 35(4), 1101-1110.
doi:<https://doi.org/10.1002/hbm.22237>
- Weeda, M. M., Brouwer, I., de Vos, M. L., de Vries, M. S., Barkhof, F., Pouwels, P., & Vrenken,
H. (2019). Comparing lesion segmentation methods in multiple sclerosis: Input from one
manually delineated subject is sufficient for accurate lesion segmentation. *NeuroImage.
Clinical*, 24, 102074. doi:<https://doi.org/10.1016/j.nicl.2019.102074>
- Weinshenker, B. G., Bass, B., Rice, G. P., Noseworthy, J., Carriere, W., Baskerville, J., & Ebers, G.
C. (1989). The natural history of multiple sclerosis: a geographically based study. 2.
Predictive value of the early clinical course. *Brain : a journal of neurology*, 112(6), 1419-
1428. doi:<https://doi.org/10.1093/brain/112.6.1419>
- Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014).
Permutation inference for the general linear model. *NeuroImage*, 92(100), 381-397.
doi:<https://doi.org/10.1016/j.neuroimage.2014.01.060>
- Zeng, C., Gu, L., Liu, Z., & Zhao, S. (2020). Review of Deep Learning Approaches for the
Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Frontiers in Neuroinformatics*,
610697, 14. doi:[10.3389/fninf.2020.610967](https://doi.org/10.3389/fninf.2020.610967)

- Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 45-47. doi:<https://doi.org/10.1109/42.906424>
- Zivadinov, R., & Bakshi, R. (2004). Central nervous system atrophy and clinical status in multiple sclerosis. *Journal of neuroimaging : official journal of the American Society of Neuroimaging*, 14((3 Suppl)), 27S-35S. doi:<https://doi.org/10.1177/1051228404266266>
- Zivadinov, R., & Zorzon, M. (2002). Is gadolinium enhancement predictive of the development of brain atrophy in multiple sclerosis? A review of the literature. *Journal of neuroimaging : official journal of the American Society of Neuroimaging*, 12(4), 302-309. doi:<https://doi.org/10.1111/j.1552-6569.2002.tb00137.x>
- Zivadinov, R., Reder, A. T., Filippi, M., Minagar, A., Stüve, O., Lassmann, H., . . . Khan, O. (2008). Mechanisms of action of disease-modifying agents and brain volume changes in multiple sclerosis. *Neurology*, 71(2), 136-144. doi:<https://doi.org/10.1212/01.wnl.0000316810.01120.05>