The relationship between cognition and white matter abnormalities in multiple sclerosis as detected by magnetic resonance imaging

Daisy Mollison

Doctor of Philosophy
The University of Edinburgh
2018

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text or where work forming part of a jointly authored publication has been included. This work has not been submitted for any other degree or professional qualification.

The work presented in Chapter 3 is adapted from a previous publication in PLoS One (2017) titled *'The clinico-radiological paradox of cognitive function and MRI burden of white matter lesions in people with multiple sclerosis: A systematic review and meta-analysis'* authored by myself, Robin Sellar, Mark Bastin, Denis Mollison, Siddharthan Chandran, Joanna Wardlaw and Peter Connick. All authors contributed to study conception and design. I was responsible for data collection, analysis and drafting of the report.

Daisy Mollison
March 2019

Abstract

Background

Multiple sclerosis (MS) is a highly variable disease of the central nervous system with inflammatory and neurodegenerative components, associated with both physical and cognitive disability. Abnormalities are visible on routine magnetic resonance imaging (MRI) of the brain, with ‘white matter hyperintensities’ (WMHs) representing sites of previous inflammation. Techniques for measuring WMHs have not been standardised, although manual outlining is conventionally taken to be the reference standard, despite its subjective element.

WMHs have been found to only partly explain the degree of cognitive impairment, forming part of the ‘clinico-radiological paradox’. Research interest has largely moved to advanced imaging techniques, one such technique being diffusion tensor imaging (DTI). Through sensitivity to water molecule movement, DTI reflects the integrity of white matter tracts and thus its measures may be relevant to both the inflammatory and degenerative disease components.

Aims

The work described in this thesis aims to improve our understanding of the true relationship between measures of white matter damage and cognitive impairment in people with MS, to determine the optimum measurement technique(s) for quantifying WMHs, including developing and testing a novel visual rating scale, and to assess whether information provided by DTI can strengthen the association of imaging and clinical findings.

Methods

A systematic review of the literature and meta-analysis relating WMHs to cognition was conducted, focussing on image analysis technique. Three separate methods for quantifying WMHs were then investigated. The reproducibility of manual outlining was assessed using scans available from 43 people with secondary progressive MS (SPMS). An automated software method was optimised for the same cohort, based on the results of the manual outlining. A novel semi-quantitative visual rating scale was developed, with validation using the same scans within a larger, more varied cohort. All available information regarding the participants studied was then used to construct a linear regression model predicting cognitive outcomes and determining the utility of the various imaging markers derived from conventional imaging techniques. A non-linear relationship for WMHs was also considered. White matter DTI metrics in the same smaller cohort of 43 people were then investigated, primarily considering tissue outwith WMHs, as well as that within major tracts and the novel diffusion marker 'peak width of skeletonised mean diffusivity'. The additional explanatory power of DTI metrics within the linear models developed previously was then determined.

Results

High variability was found in the literature regarding imaging marker measurement and reporting of technique reproducibility. Manual outlining was found to be associated with considerable measurement error, dependent on observer and cohort factors. It was possible to optimise the automated software for a particular cohort, either for volumetric or spatial outputs. Visual rating of MS imaging features was found to be feasible and measures of WMH burden were closely related to fully quantitative measures. The overall association of WMHs to cognitive function was similar to that found in the published literature, with no additional association following addition of DTI metrics. A trend towards a greater effect of WMH volume at higher levels was found, consistent with a non-linear relationship between imaging metrics and cognitive phenotype.

Conclusions

Substantial heterogeneity in the reporting of the reproducibility of WMH measurement supports a move towards benchmarking against reference datasets. Poor reliability of the current reference standard, manual segmentation, should be recognised as a key limitation for the field. Rich information can be captured quickly using visual rating of imaging features. The close correlation of visual ratings of WMHs with quantitative measures may represent a practical alternative in the appropriate circumstances. Combining visual rating features provided additional explanatory power, supporting a multidimensional substrate for the cognitive phenotype. Finally, both automated and visual rating analyses support a non-linear relationship between disease burden and cognitive performance in MS.

Lay Summary

Multiple sclerosis (MS) is a highly variable, progressive, disabling disease affecting the brain and spinal cord. Many people with MS experience problems with cognitive function, such as poor memory and slowed processing of new information. Characteristic signs related to inflammation show up on magnetic resonance imaging (MRI) scans of the brain and these tend to worsen with time. Previous research has shown that the changes on scans are partly associated with cognitive performance but are not enough to accurately predict who will have these problems. The aim of the work described here is to improve our understanding of how the changes visible on brain scans are associated with cognition.

A review was carried out of all the research into the relationship between the most common changes seen on brain scans, 'white matter hyperintensities' or WMHs, and cognitive function in people with MS. This found that the techniques for measuring both these were very variable, leading to a recommendation that researchers work together to establish common standards. In addition the relationship between imaging and cognitive features may depend on how advanced the disease is.

Three different approaches to measuring the severity of the WMHs were then investigated. Drawing round all the abnormal areas by hand is usually considered to be the best method. However differences in the results were found when different people performed this and even when the same person repeated the process. An alternative, fully computerised method, sensitive to subtle differences around the edge of the abnormalities, was tested and adjusted so as to closely match the results from hand-drawing. A third method, involving assigning scores to scans using a set of sample images was also developed and the scores for the total amount of abnormality were closely related to the volumes measured by the outlining technique.

Combining information about the WMHs and other characteristics of the people being tested, it was possible to show that cognitive function could be partly predicted using the computerised measurements or the scores assigned. Overall the results were similar to those already published, although there was a suggestion that the brain might be able to compensate for WMHs up to a certain level of damage.

An advanced imaging technique called diffusion tensor imaging (DTI) was also used in the same group of people, looking for microscopic damage not visible on routine MRI scans. DTI picks up information on water movement in the brain, giving abnormal results when there has been any disruption to the nerve fibres. Although it was possible to demonstrate subtle damage using this technique, the overall ability to predict cognitive function was not improved.

From this work, the advantages and disadvantages of different WMH measurement techniques are clarified, and a novel method of visual scoring is suggested. In addition this work suggested that the brain may be able to compensate for lower levels of disease without substantial impact on cognitive performance.

Acknowledgements (I)

This work was supported by a Rowling Scholars Clinical Academic Fellowship between 2015 and 2018.

Essential computer programming and processing, necessary for calculation and comparison of imaging volumes, as well as expert quality control of all diffusion data was provided by Dr Mark Bastin, Centre for Clinical Brain Sciences, University of Edinburgh.

Development work necessary for the use of the automated tissue segmentation software in the MS-SMART cohort was performed by Dr David Dickie, Institute of Cardiovascular and Medical Sciences, University of Glasgow.

Acknowledgements (II)

I am immensely grateful to all the participants in the three studies considered here for their time and commitment to MS research, without which no part of this work would have been possible.

Many individuals have been instrumental in helping me along the way, not least my supervisors Peter Connick, Siddharthan Chandran and Joanna Wardlaw, who have all provided inspiration, support and practical assistance. Mark Bastin deserves a special thank you for huge volumes of technical support and good sense. Robin Sellar has provided invaluable wisdom and encouragement, often when much needed.

I am grateful to all my colleagues in the Anne Rowling Regenerative Neurology Clinic for their excellent work in recruitment and support for clinical research, particularly Dawn Lyle and Denise Cranley for their work on MS-SMART, Chris Batchelor for her work on FutureMS and Shuna Colville for all her advice and support. I am very grateful to Mara Sittampalam for recruitment and testing of the 'Cognition in MS' cohort. The radiographers at the Brain Research Imaging Centre and Clinical Research Imaging Centre have ensured the ongoing success of the research and found time to answer my many questions. Many other colleagues have also provided suggestions and insight, including members of the Chandran lab group and the CCBS image analysis group.

Joanna Wardlaw, Robin Sellar, Grant Mair, Zoe Morris, Mark Rodrigues and Lorna Gibson gave up their time to participate in the pilot(s) for the visual rating work; Grant Mair gave up even more time for manual outlining and general distraction. Denis Mollison provided statistical advice on all sorts of questions thrown at him and Francesca Chappell introduced me to 'bubble' plots. Charlie Mollison assisted with emergency laptop dehydration and rebuilding. I'm also very grateful to the owners of 139 Clachtoll for its use as a writing retreat.

Finally, data analysis was carried out almost exclusively using *R* and the thesis was typeset using LaTeX; I'm very grateful to all the relevant package developers and many dedicated contributors to online support forums.

Contents

List of Figures

List of Tables

List of Abbreviations

All abbreviations are written out in full at their first use within each chapter. For convenience, a list of all abbreviations used is provided below.

AD	Axial diffusivity
AIC	Akaike information criteria
ANTs	Advanced Normalisation Tools
BDI	Beck depression index
BIC	Bayesian information criteria
BICAMS	Brief international cognitive assessment for multiple sclerosis
BRB	Brief repeatable battery
CI	Confidence interval
CIS	Clinically-isolated syndrome
CSF	Cerebrospinal fluid
CNS	Central nervous system
df	Degrees of freedom
DTI	Diffusion tensor imaging
EDSS	Expanded disability status scale
EPVS	Enlarged perivascular spaces
FA	Fractional anisotropy
FLAIR	Fluid attenuated inversion recovery
FMRI	Functional magnetic resonance imaging
FMRIB	(Oxford Centre for) Functional Magnetic Resonance Imaging of the Brain
FSE	Fast spin echo
FSL	Functional Magnetic Resonance Imaging of the Brain software library
FSPGR	Fast spoiled gradient echo
GM	Grey matter
GWAS	Genome wide association studies
ICC	Intra-class correlation
ICV	Intracranial volume
IQR	Interquartile range
JC	Juxtacortical/cortical

MACFIMS	Minimal assessment of cognitive function in multiple sclerosis
MD	Mean diffusivity
ml	Millilitre
MPRAGE	Magnetisation-prepared rapid gradient echo
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MS	Multiple sclerosis
MSFC	Multiple Sclerosis Functional Composite
MS-SMART	Multiple Sclerosis Secondary Progressive Multi-Arm Randomisation Trial
MT	Magnetisation transfer
NAWM	Normal-appearing white matter
NPV	Negative predictive value
PACS	Picture archiving and communication system
PASAT	Paced Auditory Serial Addition Test
PD	Proton density
PNT	Probabilistic neighbourhood tractography
PPMS	Primary progressive multiple sclerosis
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSMD	Peak width of skeletonised mean diffusivity
RD	Radial diffusivity
RRMS	Relapsing-remitting multiple sclerosis
SD	Standard deviation
SDMT	Symbol Digit Modality Test
SE	Spin echo
SPMS	Secondary progressive multiple sclerosis
STROBE	Strengthening the Reporting of Observational studies in Epidemiology
SVD	Small vessel disease
T1w	T1-weighted
T2w	T2-weighted
TBSS	Tract-based spatial statistics
TE	Echo time
TI	Inversion time
TR	Repetition time
TSE	Turbo spin echo
WM	White matter
WMH	White matter hyperintensity

A note on terminology

The terms ‘lesion’, ‘plaque’ and ‘hyperintensity’ have all been used in imaging research to describe abnormally bright regions seen on T2-weighted imaging sequences in the context of multiple sclerosis (MS). In this thesis, the generic term ‘white matter hyperintensity’ (WMH) is preferred, widely used elsewhere in imaging research without assumptions about underlying pathology. Reflecting the widespread use of ‘lesion’ in clinical imaging and MS-specific research, this term is used in the review of the literature presented in Chapter 3 and the visual rating work presented in Chapter 5.

Chapter 1

Introduction

1.1 Multiple sclerosis

1.1.1 Historical perspective

The pathological and clinical features of multiple sclerosis were bound together as a distinct disease ('la sclerose en plaques') in 1868 by the French neurologist Jean-Martin Charcot [1], launching 150 years of research into its aetiology and treatment. Charcot was the first to associate clinical observations with the work of the Scottish pathologist Robert Carswell and his French counterpart Jean Cruveilhier earlier in the nineteenth century [2], but it is possible that the disease has existed, unnamed, for considerably longer [3].

Despite significant advances in knowledge since its recognition, multiple sclerosis (MS) today remains an incurable, disabling disease and is currently estimated to affect around 2.3 million people worldwide [4]. In an evolving literature,whether the incidence is currently increasing [5] or decreasing [6] is unclear, confounded by changes in diagnostic criteria, technology and the advent of disease-modifying treatments, but its prevalence in the UK, particularly Scotland, is among the highest in the world [6].

1.1.2 Clinical features

MS is a highly heterogeneous condition, manifestations varying between people and over time. Clinically, the disease is divided into different phenotypes [7] by its pattern of attacks and disability accumulation, but these broad categories conceal high inter-individual variability in terms of neurological deficits, rates of relapse and progression of disability. The disease is characterised by distributed damage to the central nervous system (CNS), with initial symptoms most commonly relating to sensory, motor or visual disturbances. Problems with cognition, balance, fatigue, pain, bowel, bladder and sexual function can also occur, although may be under-recognised.

An initial attack suggestive of the disease is termed 'clinically-isolated syndrome' (CIS), with further evidence of disseminated damage needed to firmly establish a diagnosis of MS. The majority of people with MS are initially diagnosed with a relapsing-remitting form (RRMS) in which there is full or partial recovery between attacks. Recovery eventually becomes incomplete, leading to a secondary progressive (SPMS) phase of the disease, in which relapses are less frequent but accumulation of disability continues. The median time to progressive disease is around 21 years [8]. In a minority of people, the disease follows a progressive course from disease onset, known as primary progressive MS (PPMS).

The diagnosis of MS rests on evidence of disease dissemination in time and space, following exclusion of alternative diagnoses that may mimic its findings. Formal diagnostic criteria are regularly updated to reflect available evidence and advances in diagnostic technology. Since 2001 the diagnostic criteria have included magnetic resonance imaging (MRI) features, which can partially substitute for clinical findings [9].

While establishing the diagnosis has become relatively straightforward, predicting future disease activity and long term outcomes remains difficult, restricting the ability of patients and clinicians to make informed decisions. Since the advent of disease-modifying therapies, with their associated risks and costs, this inability to predict untreated disease outcomes carries still greater significance.

1.1.3 Epidemiology

The variable nature of the disease extends to its geographical profile. Different populations register very different levels of disease incidence and prevalence, with the highest rates found in European and North American populations [4]. The causes of this variation are not clear and may involve a complex interplay of genetic and environmental risk factors. Differing vitamin D levels have been suggested to explain an apparent variation with latitude, but no modification of risk factors has proven beneficial. The estimated incidence in Scotland is 15.3 per 100,000 population, with a prevalence of 255.2 per 100,000, both higher than the UK average [6].

MS is primarily a disease of working-age adults, with a female predominance. In this regard the UK follows the usual pattern, with the peak incidence at 40 in women and 45 in men and a female-to-male ratio of approximately 2.4. Peak disease prevalence in the UK is at 56 for women and 59 for men and at all ages is higher in women than men, in common with many diseases with an autoimmune component. Mortality rates are more than twice that of the general population at all ages and overall life expectancy for people with MS is lower, at 71.6 for women and 65.4 for men, than it is for the general population, where it is 81.8 and 78.3 respectively [6].

1.1.4 Genetics

The complex interactions between genetic and environmental risk factors in triggering the development of multiple sclerosis remain incompletely understood. There are clearly inherited factors, with monozygotic twins of affected people having an approximately 30% risk of developing the disease, and other relatives carrying an increased risk, related to genetic proximity [10]. A complex polygenic risk profile underlies this, with no one gene identified as either sufficient or necessary to cause the disease.

With advances in genomic screening technology, genome wide association studies (GWAS) have allowed over two hundred genetic susceptibility loci to be identified [11]. These mostly, but not exclusively, relate to immune system regulation, particularly polymorphisms in the human leucocyte antigen region [12]. The high number of susceptibility genes, all conferring modest increases in risk, make robust studies of their individual effect on phenotype difficult. However there has been some evidence that certain alleles may influence age of disease onset and disease activity as measured by MRI [12, 13].

1.1.5 Pathology

MS is an inflammatory demyelinating disease leading to chronic neurodegeneration. The processes by which damage to the CNS occur are complex and the balance between them alters through the disease course. In its most common early form, clinical relapses correspond to acute inflammatory attacks, with neurodegenerative processes predominating in later stages and associated with progressive disability. Although originally considered a disease limited to the white matter and its myelinated axons, involvement of the cortical, deep and spinal grey matter is now also recognised, where demyelination may occur with relatively little immune infiltrate [14].

The initial trigger and causal pathway is unknown, but acute inflammatory attacks are associated with breakdown of the blood-brain barrier, autoreactive lymphocytes entering the CNS and launching an inflammatory cascade, leading to activation and accumulation of local immune cells. The sclerotic plaques defining the disease are the combined result of the inflammatory attacks and subsequent repair processes, involving acute inflammation, demyelination, remyelination, astrocytosis, axonal and neuronal loss [10, 15]. Specific targets to the immune response have not been determined and may differ between individuals, but the end result is loss of the protective myelin sheaths surrounding nerve axons, depletion of the oligodendrocytes producing the myelin, and acute axonal loss. Recruitment of oligodendrocyte precursor cells and capacity for remyelination show individual variation [16] and exhaustion of these processes may coincide with the transition to progressive stages of the disease.

In later stages of the disease, axonal and neuronal loss are the predominant features [14]. The blood-brain barrier appears to remain intact, but ongoing inflammation continues, confined within the CNS, in the form of diffuse microglial activation and meningeal B-cell aggregates. The exact relationship between inflammation and neurodegeneration is unclear [17].

Although valuable insights into MS have come from pathology, these studies will always be limited by the highly variable, usually chronic nature of the disease in the specimens studied. Obtaining biopsy specimens carries a substantial risk to the patient and is usually only undertaken when considerable doubt exists over the diagnosis, suggesting that these samples may represent atypical examples of the disease. Autopsy tissue may be more easily available, but is likely to represent the end stage of a long and complex disease, with limited information regarding the intervening processes.

1.1.6 Laboratory models

MS is only known to occur in humans and its many complexities cannot be fully replicated in the laboratory. However animal models of disease, ex vivo slice cultures, and human stem cell-derived glia and neurons can all mimic components of the inflammatory or degenerative processes, offering opportunities for exploring disease mechanisms and testing potential neuroprotective treatments which would not otherwise be available.

Animal research in particular allows in vivo studies with greater potential for experimental manipulation and availability of tissue samples. Various systems, predominantly in rodents, have been developed to mimic aspects of MS [18, 19], including viral, autoimmune, genetic and toxin-based models. The mutant ‘shiverer’ mice show hypo- and dysmyelination in the CNS [20], allowing study of the associated axonopathy; experimental autoimmune encephalomyelitis, with susceptible mice exposed to CNS proteins/peptides, is used to study autoimmune demyelination [21]; toxin-induced damage, for example with cuprizone, can be used to cause apoptosis in metabolically active mature oligodendrocytes and subsequent demyelination, allowing more controlled study of the remyelination process [22]. Advantages to use of animal models are clear, but their limitations in attempting to model a complex disease with an as yet unknown trigger will always restrict their translational power.

A different approach to gaining access to cell mechanisms involved in CNS inflammation, axonal injury and repair is the use of in vitro cell cultures [19]. Advances in biotechnology have made it possible to direct the differentiation of human stem cell-derived glia and neurons and use these to explore disease mechanisms at cellular and molecular levels. Ex vivo slice cultures from animal models have been used to promote rapid screening of potential drug therapies but testing in human stem cell-derived cells is becoming possible and increases possibilities for translation to trials in humans.

1.1.7 Treatment

A number of drugs are currently licensed for use in the UK as disease-modifying therapies in RRMS, all acting to reduce the neuroinflammatory disease component. First line treatments have a variety of mechanisms of action, including inhibition of lymphocyte proliferation, reduced migration of inflammatory cells across the blood-brain barrier and increasing the presence of anti-inflammatory cytokines [23]. In some cases the mechanism of action is not fully understood. These treatments have been shown to reduce the rate of relapses, disability progression and accumulation of new inflammatory lesions on imaging, while for the most part side effects are well-tolerated.

More recently three monoclonal antibody treatments have become available, acting either to deplete lymphocytes or block their CNS infiltration. These have shown greater efficacy in reducing relapse rates and disability progression, but are associated with more serious side effects, including an increased risk of other autoimmune-mediated conditions, as well as the life-threatening condition progressive multifocal leucoencephalopathy [10, 23].

While their short term efficacy in preventing relapses in RRMS is well-established, what remains unproven is that any treatment can delay or prevent conversion to SPMS, or that a reduction in inflammatory activity prevents longer term neurodegeneration. This absence of evidence may simply reflect the relatively recent advent of disease-modifying therapies and the practical time frames for running fully-blinded randomised control trials, but effective prevention or treatments for the neurodegenerative component of the disease remains a major unmet need.

1.2 Cognitive impairment in MS

1.2.1 Prevalence and impact

Cognitive impairment in multiple sclerosis is common. It is estimated to affect up to 70% of people with the disease [24], although this will depend both on the particular population and the tests used. While rarely a presenting symptom, cognitive impairment can be present in the earliest disease stages [25]. The development of cognitive impairment can be in conjunction with physical disability, or distinct from it, but the prevalence and severity appear to increase with time since diagnosis [26]. Its onset is unlikely to be as apparent to a patient or healthcare professional as a physical relapse or other forms of disability, necessitating good screening tools. As with other aspects of this disease, the risk factors for development of cognitive impairment have not been resolved.

Cognitive impairment is associated with lower measures of quality of life [27]. It reduces physical independence, competence in daily activities, medication adherence and rehabilitation potential. It also predicts both under- and un-employment [28, 29]. As a disease of predominantly working age adults, this further increases the economic impact of MS. Early recognition of cognitive features of the disease may allow greater opportunities for suitable lifestyle adaptations, as well as more relevant measures to assess treatment outcomes.

Despite its frequency and clinical significance, the pathological substrate that causes cognitive impairment in MS is not fully understood. This is discussed in greater detail below (Section 1.4). However, given the context above, it is clear that a better understanding of the relationship between pathology and phenotype would be valuable to support targeted therapeutic intervention at the relevant biological level. From a cognitive neuroscience perspective, there may also be value in providing novel insights on the relationship between brain tissues and function. The extent to which white matter pathology in MS can explain impairments of cognition is therefore the central theme of this thesis.

1.2.2 The structure of human cognition

Human cognition is a multidimensional construct with distinct cognitive abilities identified which can vary independently to an extent. Various models have therefore been proposed to describe how different cognitive domains or specific cognitive abilities are related. A common finding is that measurements on a wide variety of cognitive tests all correlate and a general factor of cognitive ability ('g') is used to explain this shared variance [30]. The Cattell-Horn-Carroll model of human intelligence [31] is one widely-used model, developed through factor analysis of psychometric data and proposing a three-tiered hierarchy of cognitive skills with an overall general factor. Reduced processing speed is the most common deficit identified in MS and deteriorates with time. It has been suggested this is the core cognitive deficit, corresponding to 'g' and mediating other deficits via disconnection of critical cortical regions [32].

This fundamental complexity of human cognitive structure presents a substantial challenge to evaluation of the relationship between pathology and phenotype, raising the question of how best to approach measurement. Options include at one extreme lengthy multiple domain neuropsychological evaluation to detect all possible deficits. At the other extreme is single domain evaluation, typically targeting processing speed as the most responsive feature. Intermediate positions are also possible and the optimum approach to measurement is discussed further below.

1.2.3 Measurement approaches and patterns of deficits

Patterns of cognitive deficits vary between individuals with MS [24], with those most frequently detected affecting information processing speed, executive function, attention and long-term memory. Different patterns of impairment may in part relate to the random nature of the sites of inflammatory damage. However many of the more commonly affected functions, such as information processing speed, appear unlikely to localise to a single brain region or small group of regions, with preservation of widespread tissue integrity more relevant.

Recognising that comprehensive testing by an expert neuropsychologist may not always be feasible, a number of set test 'batteries' have been developed, covering a range of cognitive domains, targeted towards functions found to be disproportionately affected in people with MS. Most cognitive test outcomes are known to vary with sex and age and results must be interpreted in relation to population norms. Commonly used batteries in the research setting are the Brief Repeatable Battery (BRB) [33] and the Minimal Assessment of Cognitive

Function in MS (MACFIMS) [34], taking approximately 45 and 90 minutes respectively to administer. Used correctly, standardised tests should facilitate the use of cognitive outcomes in research, including longitudinal and multicentre trials. However Fischer et al [35] found that the criteria used to interpret test results varied widely, affecting estimates of cognitive impairment prevalence at different disease stages.

The Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) initiative has sought to standardise a shorter test battery more easily fitted within a typical clinical consultation [36]. Taking tests from both the BRB and MACFIMS, this has now been validated across several countries [37]. Three short tests are recommended, with priority given to the Symbol Digit Modality Test (SDMT), a 90 second test of information processing speed, in which numbers are matched to arbitrary symbols, found to be highly sensitive to cognitive impairment in MS [36]. It is to be hoped that a shorter more practical test procedure will enable more routine testing of asymptomatic people, leading to an improved understanding of the true prevalence and development of cognitive dysfunction.

1.2.4 Modifiers of cognitive performance and cognitive reserve

The capacity for any individual to maintain performance on cognitive testing with a given level of disease appears highly variable. A number of disease and non-disease factors have been proposed to explain this, some amenable to intervention, others not.

Theories of brain and cognitive reserve suggest that lifetime maximal brain volume, estimated by measuring intracranial volume, protects against cognitive decline, as do premorbid IQ and participation in enriching cognitive leisure activities [38, 39]. Cognitive scores are also known to vary with age, sex and education level [40]. Depression is common in MS and associated with poorer performance on cognitive testing, including processing speed [24] and many anti-spasticity drugs can also have cognitive side effects [41].

The relative importance of these modifying factors may vary between individuals and over time, but failure to consider their potential impact may obscure or confound assessment of any relationship between cognitive function and disease markers.

1.2.5 Impact of cognitive assessment approach on detection of underlying pathology-phenotype relationship

At a fundamental level, the choice of cognitive assessment method may be critical to detection and characterisation of the underlying pathology-phenotype relationship. For cognitive abilities that are highly neuroanatomically localised, the relationship to pathology in relevant regions may be strong, but this relationship may also go undetected if pathology is evaluated across the whole brain with the inclusion of additional neuroanatomical regions that are irrelevant to function. It may be critical to identify the 'right' region in which to measure pathology. The corollary of this neuroanatomical targeting is that characterisation of pathology in discrete, potentially small, brain regions inflates the importance of strong psychometric performance for the brain imaging metrics used to quantify pathology. The feasibility of evaluating the pathology-phenotype relationship for neuroanatomically localised cognitive abilities is therefore explored throughout this thesis by evaluating whether the psychometric performance of quantitative brain imaging of white matter in MS is adequate.

An alternative approach is to focus on cognitive abilities that are widely distributed from a functional neuroanatomical perspective, such as information processing speed, a common and profound deficit in MS [24]. Tests of information processing speed are known to be associated with widespread brain activation [42]. Although this reduces sensitivity in those patients who have isolated deficits of neuroanatomically localised cognitive abilities, it also minimises the effects of psychometric limitations of brain imaging quantification. Investigation of the relationship between pathology and phenotype for distributed cognitive functions therefore represents the optimum approach to explore fundamental relationships between pathology in brain compartments (for example white matter) and phenotype, although brain imaging metrics must of course be optimised for their psychometric performance. The focus on neuroanatomically distributed cognitive abilities is therefore central in this thesis to evaluate the relationship between white matter pathology and cognitive impairment.

1.3 Brain imaging in MS

The advent of MRI has brought major changes to the diagnosis and monitoring of MS. Many of the known pathological features are linked to imaging findings, allowing less invasive research methods. The use of brain imaging techniques to produce biomarkers of pathology in MS can be described within multiple different frameworks. The following introduction to brain imaging metrics in MS classifies them by the underlying pathology for which they are generally considered a biomarker, while acknowledging that few of them are considered specific for a particular feature. This approach is chosen to facilitate attempts to address the question of pathology-phenotype relationships. The necessity of using non-specific (imperfect) biomarkers remains a challenging issue for the field, balancing valid but low biological value descriptions of metric-phenotype relationships against less valid but high biological value pathology-phenotype claims.

1.3.1 Focal neuroinflammation

1.3.1.1 T2-weighted white matter hyperintensities

Acute inflammation is the most readily demonstrable of the pathological features of MS. Characteristic white matter hyperintensities (WMHs) on T2-weighted (T2w) MRI sequences within the brain and/or spinal cord provide supporting evidence used in making the diagnosis and are widely accepted as a marker of the historical burden of focal inflammation. As MRI techniques and experience have improved, the imaging criteria for diagnosis have changed to reflect this. The latest criteria [43] highlight abnormalities specifically involving cortical/juxtacortical brain tissues and optic nerves as suggestive of the diagnosis - locations which were not previously visible on imaging.

Two major barriers to use of these WMHs as a biomarker, for example in disease and treatment monitoring, is both attributing them to MS rather than other comorbidities, as well as their pathological heterogeneity [44] within MS itself. A developing MS plaque comprising any combination of acute inflammation, demyelination, remyelination or astrocytosis will appear bright on T2w imaging. In clinical practice, combinations of features are used as imperfect predictors of the degree of chronicity. Although non-specific, the number and volume of enlarged perivascular spaces (EPVS) visible on T2w imaging have been linked to the presence of acute inflammation [45], although this requires confirmation. In investigating chronic damage, scant literature exists regarding the frequency of lesion progression to cavitation, in which severely damaged tissue is replaced by a fluid-filled cavity, and this is thought to be an uncommon feature in MS [46]. However the concept of ‘T1 black holes’ [47, 48] (see Section 1.3.1.3) is clearly related as a means to identifying more severely damaged tissue. Furthermore, a variety of other conditions are also associated with WMHs, including small vessel disease, other vascular and inflammatory conditions and normal ageing [49].

While particular patterns and numbers of focal WMHs have a well-established utility in clinical diagnosis, their use in disease/treatment monitoring is less clearly evidence-based, which may relate to this pathological non-specificity as well as the changing role of diffuse pathology.

A further barrier to the use of measures of WMH volume is the existence of ‘dirty’ or diffusely-abnormal white matter, a category of tissue falling between those of focal abnormalities with distinct borders and a likely inflammatory origin, and the normal-appearing white matter (NAWM) [50,51]. This intermediate category of white matter appearances may be extensive, particularly in chronic disease and presents clear challenges to volumetric approaches to image analysis.

1.3.1.2 Contrast enhancement

Current disease activity can be estimated by gadolinium contrast-enhanced imaging, with gadolinium taken up into brain tissues where the blood-brain barrier has been breached. The rate of underlying disease activity, as measured by contrast-enhanced imaging, is roughly an order of magnitude higher than clinical evidence of relapses [52–54].

This increased sensitivity of imaging to disease activity and burden, as well as its perceived objectivity, is exploited by trials where surrogate imaging outcomes allow for increased sensitivity to the effect of interventions.

1.3.1.3 T1-weighted white matter hypointensities

The relative brightness of white matter on T1-weighted (T1w) imaging relates to fat signal from myelin. Focal abnormalities appearing hypointense on T1w sequences are often considered a more specific disease marker than T2w WMHs, reflecting myelin loss and indicating more severe damage and/or a failure of repair following acute inflammation [47, 48]. Nevertheless, both T1- and T2-visible lesions remain non-specific, with the added confounding effect of acute white matter oedema appearing hypointense on T1w imaging.

1.3.1.4 Cortical lesions

Cortical lesions are only partly visible on current routine imaging, likely related to lower levels of inflammatory infiltrate, lower myelin density and partial volume effects from proximity to cerebrospinal fluid (CSF) [17]. Advanced imaging techniques, such as double inversion recovery and phase-sensitive inversion recovery sequences, have been developed to optimise evaluation of cortical pathology, but do not necessarily increase sensitivity or reliability [55, 56]. Nevertheless it has been suggested that they may be particularly relevant in determining cognitive status [57].

1.3.1.5 Quantification methods

Counts of new or enlarging T2w WMHs by adequately trained human observers have been the benchmark outcome measure for phase II trials of disease-modifying drugs. This provides an ordinal level measurement of WMH burden, but is labour intensive, user-dependent and limited by the presence of WMH coalescence as well as a failure to capture information on the size or distribution of abnormalities. The reference standard for quantitative analysis is manual outlining of all lesions, but this is user-dependent and time-intensive. For research purposes, the field has therefore largely moved to use various software analysis packages for a semi- or fully-automated quantitative assessment of T2w- and/or T1w-visible abnormality [58]. These have the advantage of providing fully quantitative, ratio level measurements of the imaging burden of disease. With the increasing use of such technology-based quantification techniques, it is not clear what their psychometric performance characteristics are, a key question that is addressed in this thesis.

A small volume of research has applied visual rating scales to the imaging features of MS [59–64]. These studies used relatively small cohorts and were predominantly conducted in the early years of clinical MRI use, predating the move towards technology-based quantification. No visual rating scale for MS imaging features has entered common use. However semi-quantitative visual rating scales are frequently used for research in other conditions, including those such as small vessel disease (SVD) [65] with a similar range of imaging features. Visual rating scales may therefore have value in providing an intermediate approach between ordinal assessment of new or enlarging WMHs and technology-based volumetric quantification. The development of a novel visual rating scale is described in this thesis and its performance evaluated.

1.3.2 Neurodegeneration and repair

1.3.2.1 Tissue volume

Brain atrophy is accelerated from the earliest stages of MS and progresses with the disease [66, 67], marking the known tissue loss. Atrophy is frequently used as an outcome in clinical trials of potentially neuroprotective drugs [66], using software-based volumetric measurements, either relying on registration of longitudinally-acquired imaging or segmentation of individual scans [68]. However, as with evidence of focal damage, atrophy is a pathologically non-specific marker, representing a global sum of rates which may differ by tissue type and anatomical location. The changes involved are small, require longitudinal data, preferably at distant time points, and may vary due to drugs and hydration status [66, 69].

1.3.2.2 Quantitative markers of tissue microstructure

Diffusion tensor imaging (DTI) is an advanced MRI technique that is frequently used in MS research, providing information on tissue microstructure through measuring the random motion of water molecules at the voxel level [70]. Healthy white matter is a tightly packed, highly coherent structure containing myelinated axons, with water movement predominantly constrained to follow the paths of white matter tracts. Any damage to these tracts is associated with altered DTI metrics. The DTI parameters fractional anisotropy (FA) and mean diffusivity (MD) measure the directional coherence and magnitude of water diffusivity respectively and are sensitive probes of tissue microstructure. Typical patterns of change seen in MS are reduced FA and increased MD in affected white matter.

As with measuring focal lesions, there are many ways to analyse DTI data and the chosen methodology will depend on the research question. Similar to the use of total WMH volume, DTI-derived measures of total disease burden can be extracted for use by calculating mean voxel metrics for white matter or other tissue compartments of interest. Several papers have examined the relationship of mean DTI metrics within the normal-appearing white matter to cognitive function [71–74]. All found significant associations, although these were not consistently stronger than the relationship between cognitive performance and WMH volume.

Tract-based spatial statistics (TBSS) [75] has become widely used for combining and comparing data from multiple subjects. This technique involves thinning and alignment of multiple subject DTI data to a common white matter skeleton, allowing voxelwise comparisons of diffusion metrics with outcomes of interest. Many groups have used TBSS to locate voxels with significant associations with cognitive outcomes [71–74, 76–91].

Baykara et al [92] have recently proposed an alternative summary marker from DTI data in cerebral SVD - the ‘peak width of skeletonised mean diffusivity’ (PSMD), based on TBSS and histogram analysis of the aligned voxels. PSMD summarises the spread of values of white matter MD in a single measure, derived from an automated pipeline. This objective metric may provide a more biologically relevant marker of total disease burden as it allows for the known individual variation in mean MD values, instead summarising their spread within a scan. In their study of its use in cohorts with SVD, Baykara et al showed that it outperformed conventional imaging markers in explaining variance in processing speed. They also showed its stability across multiple sets of healthy control data from studies on SVD and Alzheimer’s disease, but true normative data is not yet available, nor has any data yet been published on the use of PSMD in studies of MS.

Separately from TBSS, major tracts can also be tracked and segmented from DTI data by various methods, most commonly tractography. Probabilistic neighbourhood tractography (PNT) is one such technique [93], optimising starting voxels or ‘seed’ points for fibre-tracking based on a comparison with standard reference tracts. Tractography has the advantage of minimising the confounding effect of crossing fibres on directionality markers and focussing on tissues with more readily attributable functional significance.

1.3.2.3 Quantitative markers of tissue composition

More specific markers of neuronal health and tissue integrity have been proposed, derived from different advanced imaging techniques. Magnetic resonance spectroscopy (MRS) can be used to non-invasively assess changes in neuroaxonal metabolites [94, 95], such as N-acetyl-aspartate, within lesional and non-lesional white matter. While potentially a useful biomarker of neurodegeneration, MRS is currently limited by its spatial resolution and the high processing demands. Magnetisation transfer (MT) imaging is based on signal sensitivity to the presence of tissue macromolecules, such as myelin [96, 97]. While potentially providing useful biomarkers of myelination status for trials of neuroprotective agents, metrics derived from MT imaging remain non-specific. As yet, no advanced imaging technique has been recommended for clinical use.

1.3.2.4 Imaging markers of connectivity

Neuronal plasticity as a response to pathological changes is well-recognised [98], from local synaptic reorganisation to recruitment of distant cortical sites and parallel pathways. Obtaining direct evidence of synaptic changes is beyond the resolution of current in vivo imaging techniques, but larger scale alterations in structural and functional connectivity can be inferred using information derived from tractography as well as functional MRI (fMRI). FMRI measures dynamic changes in regional blood flow, demonstrating recruitment and synchronicity of activity in cortical regions either in response to a particular task or in the resting state. It has been used to show altered connectivity even in the absence of clinical deficits [99] and increased regional activation correlating with measures of damage to NAWM [100,101], suggesting that adaptive changes may in part be responsible for limiting the extent of clinical impairment.

1.4 The cognitive clinicoradiological paradox

The mismatch between radiological and clinical findings in MS is well-recognised [52]. Although measurable imaging changes are visible on MRI from the earliest disease stages and progress with the disease, these can so far not be used to accurately predict accumulated disability. This disconnection has important implications for the use of imaging in monitoring disease progression and as a surrogate outcome of treatment success in clinical trials, limiting the efficient collection of relevant information.

In the case of motor function, the modest correlation between imaging and clinical outcomes can be attributed to the frequent presence of spinal pathology and the complex hierarchy of relevant brain regions. Spinal pathology allows a focal interruption in the system of motor control that will inevitably reduce correlation to the total burden of pathology. In contrast, tests of distributed cognitive functions, such as processing speed, can be designed with minimal dependence on physical function, and the reasons for these not reflecting more closely the brain imaging burden of disease are less clear. Moreover, MS is primarily a disease of working age adults; the clinical features usually starting once brain development is complete and before the onset of age-related cognitive decline. Any impairment of cognitive function can thus be reasonably hypothesised to reflect the total burden of brain disease. This prediction has not been borne out by the evidence. The entire body of published research on the relationship between white matter brain imaging features on routine MRI and cognitive performance in MS is summarised and synthesised in Chapter 3 of this thesis.

Various authors have previously sought to summarise the extensive literature regarding the relationship between cognitive function and WMH burden in MS [102] without fully investigating the potential reasons for the modest association.

Two primary considerations are important that might account for the cognitive clinicoradiological paradox.

1. Attenuation of the observed correlation due to measurement error affecting metrics for cognition, brain imaging or both. Despite a recommendation in the original publication by Spearman in 1904 [103] to consider this possibilty, it has been largely ignored by the field. Defining and optimising the psychometric performance characteristics of brain imaging metrics is therefore a key objective of this thesis.

2. Adjustment for known modifiers of cognitive performance (as discussed in Section 1.2.4) has been attempted by some research groups but is by no means universal. Adjustment for relevant covariates is therefore performed in this thesis to support a more accurate evaluation of the underlying pathology-phenotype relationship.

These considerations are critical to establish the true extent of any residual mismatch between imaging and cognitive measures and should perhaps be an obligatory step before attempting to define the further pathological features that are responsible for any 'missing explanatory power'. Nevertheless, the MS imaging research community has largely moved to using advanced imaging techniques aiming to quantify more subtle features and the volume of published work on this subject increases yearly. However without addressing the relative importance of measurement technique and error, cognitive modifiers and confounders, the explanatory power of advanced imaging techniques may similarly be limited.

1.5 Overview of thesis

1.5.1 Hypothesis

The overarching hypothesis tested in this work is that optimised measurement of white matter MRI characteristics will lead to a more accurate determination of the relationship between the overall imaging burden of disease and cognitive performance.

Specific tested hypotheses are:

- The reference standard for WMH segmentation is imperfect, with error substantial enough to obscure relevant relationships.

- Visual rating can be used to accurately capture data on imaging features in MS relevant to cognitive status.

- Routine MR imaging features in MS will contribute significantly to accurate prediction of cognitive phenotype.
- The addition of DTI-derived measures of microstructural tissue abnormality in the NAWM to predictive models of cognitive function based on routine imaging and non-disease factors will increase the overall predictive power.

1.5.2 Aims and thesis structure

The thesis aims and the related work are outlined below.

1. To review the published literature on the relationship between imaging measures of white matter pathology and cognitive performance in people with MS, confirming the modest correlations found previously and exploring methodological issues that may affect this.

   A systematic review of the literature and meta-analysis is described in Chapter 3.

2. To develop tools for reproducible quantification of WMH burden on structural brain MRI. Differing approaches to evaluation of white matter imaging features were identified in the literature, and three alternative methodologies are examined in Chapters 4 to 6. The reference standard, manual segmentation, is evaluated in Chapter 4, with an investigation of intra- and inter-observer reproducibility. The development and evaluation of a novel visual rating scale is described in Chapter 5. Optimisation and evaluation of an automated segmentation software tool is described in Chapter 6.

3. To evaluate the relationship between reproducible tools for quantifying WMH burden on structural brain MRI and cognitive performance in people with MS.

   Using the optimised volumetric and semi-quantitative measures of WMH burden evaluated previously, their relationship to cognitive performance is evaluated in Chapter 7, taking into account other relevant imaging and non-imaging features using linear modelling.

4. To evaluate the potential additional value of DTI techniques in accounting for the relationship between brain imaging metrics of pathology and cognitive performance in people with MS.

   In Chapter 8, the ability of DTI to demonstrate measurable changes in the NAWM is first assessed and different DTI-derived metrics are evaluated.

Using the best predictive model developed in the previous chapter, the additional value of DTI measures is tested.

Chapter 2

Description of cohorts studied

Hypotheses in the work described in Chapters 4 to 8 were tested using imaging, clinical and demographic data available locally from people with multiple sclerosis (MS) participating in ongoing or recently completed research. Three cohorts were chosen for study, encompassing a range of clinical phenotypes.

The 'MS-SMART' cohort, a group of 93 participants with secondary progressive MS (SPMS), was used for the work described in Chapters 4, 6 and 8; as well as the validation work in Chapter 5 and part of the linear modelling work in Chapter 7. The 'Cognition in MS' (n = 60) cohort, a mixed phenotype group recruited for a previous PhD thesis, was used in the initial development work on the visual rating scale described in Chapter 5. 'FutureMS', a prospective cohort of 67 people newly diagnosed with relapsing-remitting MS (RRMS), was used in the validation stage of the visual rating scale work described in Chapter 5 and the related linear modelling in Chapter 7.

2.1 MS-SMART

2.1.1 Study aims, protocol & recruitment

Multiple Sclerosis Secondary Progressive Multi-Arm Randomisation Trial (MS-SMART) was an ongoing multicentre, multi-arm, randomised, double blind, placebo-controlled trial. Participants were randomised to receive either placebo or one of three potentially neuroprotective drugs (fluoxetine, riluzole and amiloride) for 96 weeks. The primary outcome was magnetic resonance imaging (MRI)-derived percentage brain volume change.

Participants were recruited into the trial after referral from their neurologist, or self-referral following media campaigns. The main eligibility criteria were a confirmed diagnosis of SPMS, age between 25 and 65, an Expanded Disability Status Scale (EDSS) score of 4.0 to 6.5, a Beck Depression Index of < 20 and neither having taken disease-modifying therapies within the 6 months prior to recruitment, nor having had oral or intravenous steroids within 3 months.

All participants underwent MRI at three time points. The baseline scans, performed prior to treatment randomisation, in participants at the University of Edinburgh site were used for work presented in this thesis and the standard protocol for these is described below. Participants were also invited to participate in an 'advanced' imaging protocol (see below for further details).

The chief investigator was Dr Jeremy Chataway (University College, London) and the research was funded through the Efficacy and Mechanism Evaluation Programme (a partnership between the Medical Research Council and the National Institute for Health Research) and the MS Society. The trial was registered with the European Medicines Agency with EudraCT number 2012-005394-31 and with the International Standard Randomised Controlled Trial Number Registry, number 28440672.

2.1.2 Image acquisition

Baseline imaging for all of the Edinburgh centre participants was carried out between February 2015 and May 2016 at 3T (Magnetom Verio, Siemens AG, Healthcare Division GmbH, Erlangen, Germany) at the Clinical Research Imaging Centre, University of Edinburgh, using a standard 12-channel head coil. Imaging acquired included a volumetric T1-weighted (T1w) (1mm isotropic voxels) sequence, as well as proton density (PD), T2-weighted (T2w), T1w and fluid attenuated inversion recovery (FLAIR) (all 3mm slices) sequences, acquired as contiguous axial slices, parallel to a line joining the inferior points of the corpus callosum. See Table 2.1 for sequence details.

A subset of University of Edinburgh participants was also enrolled in the Advanced MRI substudy, undergoing additional magnetic transfer imaging, proton magnetic resonance spectroscopy and diffusion tensor imaging (DTI) at their baseline and 96-week scans. The baseline DTI sequences were used for the work described in Chapter 8 of this thesis.

The diffusion imaging protocol consisted of 6 T2w sequences (b = 0 s mm^{-2}) and sets of diffusion-weighted (b = 1000 s mm^{-2}) whole brain single-shot spin-echo echo-planar imaging volumes acquired with diffusion encoding gradients applied in 56 non-collinear directions. The acquisition parameters were: field-of-view 240 x 240mm; imaging matrix 96 x 96; 60 contiguous 2.5mm thick axial slices, giving

2.5mm^3 isotropic voxels. Repetition and echo times were 11500ms and 73.6ms respectively.

Sequence name	Field-of-view (mm)	Slices	Voxel size (mm)	TR/TE/TI (ms)	Flip angle (°)
PD/T2w dual echo TSE	250×250	60	$1 \times 1 \times 3$	3050/31/- & /82/-	150
T2w FLAIR BLADE TSE	250×250	60	$1 \times 1 \times 3$	9500/124/2400	150
T1w SE	250×250	60	$1 \times 1 \times 3$	600/8.4/-	70/180
T1w MPRAGE	250×250	160	$1 \times 1 \times 1$	2400/2.97/1000	8

Table 2.1: Sequence details for standard baseline imaging protocol for MS-SMART participants. TR: repetition time; TE: echo time; TI: inversion time; TSE: turbo spin echo; SE: spin echo; MPRAGE: magnetisation-prepared rapid gradient echo.

2.1.3 Image post-processing

Post-processing of MS-SMART imaging was performed locally, using fully-automated processes unless specified otherwise, supervised by MB.

2.1.3.1 Tissue segmentation

The 3D T1w and 2D FLAIR sequences were co-registered to the T2w sequence using affine transformations (12 degrees of freedom), using tools freely available in the Functional Magnetic Resonance Imaging of the Brain (FMRIB) software library (FSL, https://fsl.fmrib.ox.ac.uk) [104]. Using Advanced Normalisation Tools (ANTs, http://www.picsl.upenn.edu/ANTs) [105], both the FLAIR and T1w sequences were corrected for bias field inhomogeneities. Again using ANTs, voxels in the T1w sequence were segmented into different tissue classes by assigned voxel probabilities of belonging to cerebrospinal fluid (CSF), cortical grey matter, white matter, subcortical grey matter, cerebellum and brainstem based on the MICCAI 2012 Multi-Atlas Challenge Data atlas (https://my.vanderbilt.edu/masi/ [106]). Volumes for each tissue compartment were then generated using weighted sums of all voxels multiplied by the relevant probability. A threshold probability of 0.05 was used for all compartments to exclude noise.

2.1.3.2 White matter hyperintensity masks

The white matter hyperintensity (WMH) masks used for the work described in Chapters 6 to 8 were automatically generated using a method combining statistical transformation and atlas-based segmentation, developed by DD. This was based on work completed for a previous PhD thesis [107], with code ran in Matrix Laboratory (MatLab) on a Linux workstation.

Initial tissue segmentations were as above, using the MICCAI 2012 Multi-Atlas Challenge Data atlas. A standard deviation map of the FLAIR volume was created for each individual scan and this was used to update the initial probabilities of belonging to the white matter tissue class to produce a WMH probability for each voxel. A probabilistic mask of WMH voxels was created by identifying any voxel with a standard deviation greater than the user specified threshold, based on standard deviations above the mean FLAIR intensity. It was then possible to create hard (binary) masks of WMH by selecting an arbitrary probability threshold and this was used for the optimisation work described in Chapter 6, comparing the masks with binary manual segmentations. An example of a binary WMH mask overlaid on the FLAIR sequence from one of the MS-SMART participants is shown in Figure 2.1.

Following this optimisation work, two different FLAIR thresholds were used for the work described in Chapters 7 and 8. First, the FLAIR threshold associated with the highest correlation between absolute WMH volumes derived from manual and automated segmentation in the Advanced MRI subgroup was determined. The resulting mask was then used to generate an absolute WMH volume for all participants, by summing probabilities voxelwise. Second, the FLAIR threshold that maximised spatial overlap with the manual masks, as determined by the Dice index [108], was used. This also retained the probabilistic output and was used as a template for overlaying the water diffusion imaging parameters.

2.1.3.3 Diffusion post-processing

Diffusion-weighted images were corrected for eddy current-induced distortions and subject motion with the ‘eddy_correct’ tool (FSL). After brain tissue extraction using the Brain Extraction Tool, diffusion tensors and scalar diffusion parameters (fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD) and radial diffusivity (RD)) were calculated using DTIFit (FSL).

For each tissue compartment, mean water diffusion metrics were derived using a weighted mean; each voxel metric was multiplied by its probability of belonging to

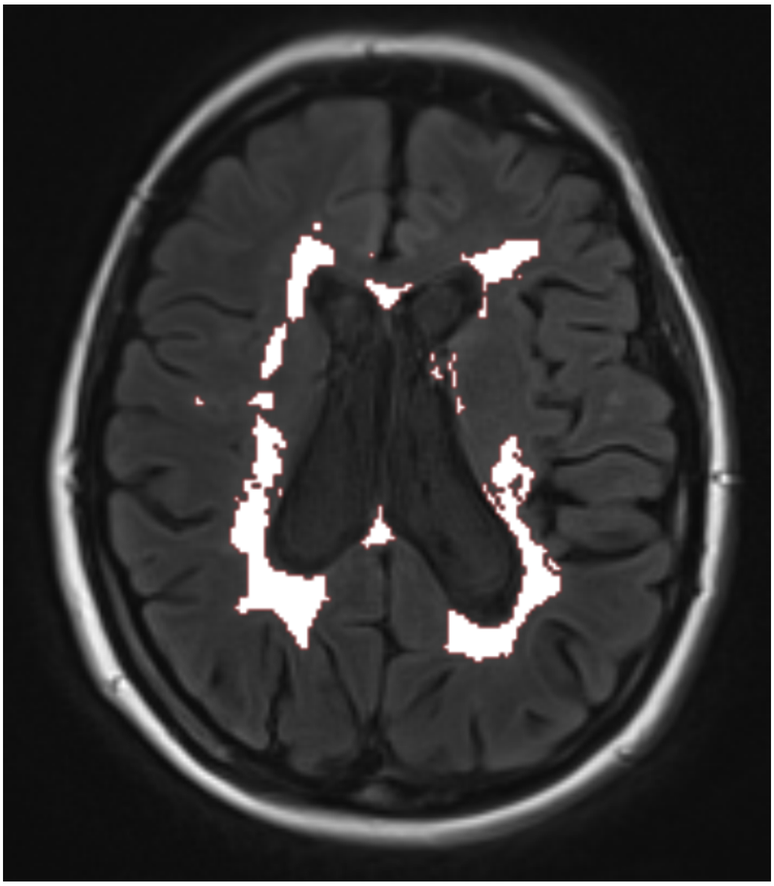


Figure 2.1: Sample image showing automated WMH segmentation overlaid on FLAIR sequence from one of the MS-SMART participants.

the compartment of interest and the sum divided by the sum of probabilities. The probabilistic WMH mask was generated using the FLAIR threshold (see Section 2.1.3.2) that showed the closest spatial agreement with manual segmentation, as described above.

2.1.3.4 Peak Width of Skeletonised Mean Diffusivity

The novel imaging marker 'peak width of skeletonised mean diffusivity' (PSMD), developed for use in small vessel disease, was derived by the method described by Baykara et al [92] using the shell script provided at http://www.psmd-marker.com/. DTI data were first skeletonised using the Tract-Based Spatial Statistics procedure [75], part of FSL, aligning all participants' FA data into common space using nonlinear registration (FNIRT) and the standard space FMRIB 1mm FA template. Individual subject FA data was then projected onto the skeleton derived from the standard space template, thresholded at an FA of 0.2. Subject MD data were then projected onto this, using the FA-derived parameters. The final MD skeleton was masked at an FA value of 0.3 to avoid CSF contamination. The PSMD parameter was then calculated as the difference between the 95th and 5th percentiles of the voxel-based MD values within the white matter skeleton.

2.1.3.5 Probabilistic Neighbourhood Tractography

Twelve tracts of interest were identified from the diffusion MRI data using probabilistic neighbourhood tractography (PNT) as implemented in the TractoR package for fibre tracking analysis (http://www.tractor-mri.org.uk). This technique optimises the choice of seed point for tractography by estimating the best matching tract to a reference tract derived from a white matter atlas, using a series of candidate seed points placed in a $7 \times 7 \times 7$ voxel neighbourhood. Tracts assessed were the genu and splenium of the corpus callosum, and bilaterally the cingulum (divided into dorsal and ventral portions), corticospinal tracts, arcuate fasciculi and inferior longitudinal fasciculi. All generated tracts were visually assessed by an experienced observer (MB) and those that were deemed not to be anatomically acceptable representations of the fasciculi of interest were discarded from further analysis.

2.1.4 Clinical and cognitive assessments

Participants underwent a structured baseline assessment by trained assessors prior to randomisation and within one month of the initial MRI. This included assessment of the Expanded Disability Status Scale (EDSS), the Multiple Sclerosis Functional Composite (MSFC) and the Symbol Digit Modality Test (SDMT). The EDSS is a widely used method for quantifying neurological disability in MS, covering eight functional systems (pyramidal, cerebellar, brainstem, sensory, bowel and bladder, visual, cerebral/mental and 'other', http://www.neurostatus.net). The MSFC [109] is a short, three part standardised test designed for use as an outcome in clinical trials in MS, assessing upper and lower limb function as well as cognitive function. The SDMT is a ninety second task to assess processing speed, consisting of matching numbers to arbitrary symbols using a given code [36]. The SDMT was assessed by three raters

(DM, DL, DC), all trained locally and observing each other during the training period.

2.1.5 Cohort characteristics

At the University of Edinburgh, 111 people with MS were screened for trial eligibility. Fourteen did not meet eligibility criteria, and four withdrew before baseline assessment; 93 participants were successfully enrolled. Of these, 43 people consented to participate in the Advanced MRI substudy and the remainder were enrolled in the standard protocol study. Further details of the study participants are given in Table 2.2.

	Standard	Advanced	Overall
No. of participants	50	43	93
Female : Male	39:11	30:13	69:24
Mean age (years) ± SD	54.9 ± 6.6	55.5 ± 8.3	55.2 ± 7.4
Age range (years)	$41.4 - 65.9$	$34.4 - 65.6$	$34.4 - 65.9$
Mean disease duration (years) ± SD	21.0 ± 10.8	23.1 ± 10.2	22.0 ± 10.6

Table 2.2: Characteristics of participants enrolled in MS-SMART at the University of Edinburgh, also presented separately for those having the standard imaging protocol, and those participating in the Advanced MRI substudy. SD: standard deviation.

2.2 Cognition in MS

2.2.1 Study aims, protocol & recruitment

A cross-sectional cohort of 108 people with MS were recruited to the 'Cognition in MS' study as part of a PhD research project at the University of Edinburgh Centre for Clinical Brain Sciences. The aim of that study was to explore the prevalence of cognitive impairment in people with MS. Participants were recruited from secondary care in the Lothian area of Scotland, referred by local neurologists and specialist nurses between August 2010 and August 2012. Potential participants were screened against the following eligibility criteria: a diagnosis of MS according to the revised (2010) McDonald criteria [110], age 18 to 65 years inclusive and absence of psychiatric or physical comorbidity (including major affective disorder, significant dementia or other significant comorbidities). Subjects showing any ophthalmological condition not related to MS that might interfere with testing were also excluded. For this purpose subjects having a Snellen acuity worse than 20/70 were excluded.

Of 972 patients screened, 108 patients fulfilled the eligibility criteria and agreed to participate. It was projected to accomplish a sample with the ratio of RRMS, SPMS and primary progressive MS (PPMS) similar to their natural proportions in the population. A subset of sixty participants agreed to undergo MRI. Data from this subgroup were used for the work described in Chapter 5 of this thesis, with details of the participants given in Table 2.3.

Female : Male	32:28
Disease course (RRMS : SPMS : PPMS)	27:18:15
Mean age (years) ± SD	46.4 ± 8.2
Age range	28 - 61
Mean disease duration (years) ± SD	9.7 ± 6.2

Table 2.3: Characteristics of participants enrolled in the Cognition in MS study. PPMS: primary progressive MS; RRMS: relapsing-remitting MS; SD: standard deviation; SPMS: secondary progressive MS.

2.2.2 Imaging protocol

All MRI data were acquired in the Brain Research Imaging Centre, University of Edinburgh, using a GE Signa Horizon HDx 1.5T clinical scanner (General Electric, Milwaukee, WI) equipped with a self-shielding gradient set (33 mT/m maximum gradient strength) and manufacturer supplied eight-channel phased-array head coil, between May 2011 and July 2012. Details of the basic sequence parameters are shown in Table 2.4. A subset of participants had an additional T2 CUBE volume sequence.

2.3 FutureMS

2.3.1 Study aims, protocol & recruitment

FutureMS was an ongoing multicentre observational cohort study, using baseline clinical, laboratory and genomic data to predict neuroinflammatory disease activity over a 12 month period. The main inclusion criteria were: age over

Sequence name	Acquisition method	Field-of-view (mm)	Matrix	Slices	Voxel (mm)	TR/TE/TI (ms)
T2w	FSE	256 × 256	256 × 256	80	1 × 1 × 2	11320/102
T2*w	Gradient echo	256 × 256	256 × 192	80	1 × 1 × 2	940/15
FLAIR	FSE	256 × 256	256 × 192	40	1 × 1 × 4	9000/140/2200
T1w	3D IR-Prep FSPGR	256 × 256	192 × 192	160	1 × 1 × 1.3	10/4/500

Table 2.4: Sequence details for standard imaging protocol for Cognition in MS participants. FSE: fast spin echo; TR: repetition time; TE: echo time; TI: inversion time; FSPGR: fast spoiled gradient echo.

18 years, having been diagnosed with relapsing-onset MS within the preceding 6 months and not having been started on any disease-modifying therapy. Participants underwent brain imaging at baseline and after 12 months, with detailed clinical assessment including the MSFC performed by trained assessors at the same timepoints.

At the time of the visual rating work reported in Chapter 5, sixty-seven participants had been recruited and scanned at the University of Edinburgh. Details of these participants are given in Table 2.5.

Female : Male	49:18
Mean age (years) ± SD	39.3 ± 9.6
Age range	21.5 - 58.6

Table 2.5: Characteristics of participants enrolled in the FutureMS study. SD: standard deviation

2.3.2 Imaging protocol

All scans were performed at 3T (Magnetom Verio, Siemens AG, Healthcare Division GmbH, Erlangen, Germany) at the Clinical Research Imaging Centre, University of Edinburgh, using a standard 12-channel head coil.

These 67 participants included those imaged during the scan protocol development phase. All protocols included a volumetric T1w sequence (1mm isotropic voxels), with FLAIR and T2w sequences either as 3D or axial 2D acqusitions. After the initial 23 participants, all scans included both 2D and 3D FLAIR sequences; after the initial 25 participants, all scans included a 2D T2w sequence, replacing a 3D T2w sequence. 3D sequences were acquired sagittally and 2D sequences were acquired axially, parallel to a line joining the inferior points of the corpus callosum. The axial T2w sequence had a slice gap of 30%. Details of the final scan protocol are given in Table 2.6.

Sequence name	Field-of-view (mm)	Slices	Voxel size (mm)	TR/TE/TI (ms)	Flip angle (°)
T1w MPRAGE	256 × 256	176	1 × 1 × 1	5300/3.37/1100	7
T2w	220 × 220	33	0.7 × 0.7 × 4	6000/96/-	150
T2 FLAIR BLADE	250 x 250	60	1 × 1 × 3	9500/124/2400	150
T2 SPACE FLAIR	256 x 256	176	1 × 1 × 1	5000/715/1800	-

Table 2.6: Sequence details for standard imaging protocol for FutureMS participants. Abbreviations are as for Table 2.1.

Chapter 3

Systematic review of literature: relationship between cognitive performance and total white matter lesion burden

3.1 Introduction

Moderate correlations have been reported between the imaging quantification of brain white matter hyperintensities (WMHs) and cognitive performance in people with multiple sclerosis (MS). This forms part of the 'clinicoradiological paradox'. A number of factors may account for this, including aspects of MS pathology that are neither measured nor closely correlated with WMHs, insensitivity of MRI techniques resulting in 'subvisible' pathology, and methodological limitations of the current approaches to quantifying WMH burden. A systematic review and meta-analysis of the published literature describing the relationship between cognitive function and the total burden of white matter pathology detected by standard structural brain MRI was therefore performed. The specific aims were to summarise the cognitive clinicoradiological paradox, confirming the modest correlations previously described [102], and to define the potential methodological factors that could have influenced the assessment of this relationship. The design of the systematic review, meta-analysis and structured report were based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [111].

3.2 Methods

3.2.1 Protocol, information sources and search strategy

The study protocol was documented in advance (see Appendix A). Medline, Embase, and Web of Science databases were searched for English language papers on 1st July 2015, with no date restrictions. The search terms were: ‘magnetic resonance imaging’, ‘multiple sclerosis’, ‘cognitive’, ‘cognition’, related terms, including relevant medical subject headings (‘MeSH’) and abbreviations of these. (See Appendix B for details of search strategy.) Review articles were excluded, with relevant reviews published in the last 10 years being screened for references. Archives of the journals Neurology, Multiple Sclerosis (Multiple Sclerosis Journal from 2011) and the American Journal of Neuroradiology were ‘hand-searched’ for relevant articles published in the previous ten years. These journals were identified as relevant examples of the literature, being widely read by clinicians and academics with an interest in MS.

3.2.2 Study selection and eligibility criteria

Initial screening of abstracts was performed by a single investigator (DM). Full articles were then retrieved and eligibility assessment performed in a standardised manner, with a final decision over study inclusion taken in consensus with a second investigator (PC).

Eligibility criteria were: English language and peer-reviewed publications reporting data from adults with clinically-definite MS as primary research with a primary aim of relating cognition to routine MRI (T1-weighted (T1w), T2-weighted (T2w), fluid-attenuated inversion recovery (FLAIR) or proton density (PD)) metrics of total brain white matter lesion burden.

Imaging outcomes given as total lesion volumes or areas, lesion counts or scores, were all accepted as valid measures of whole brain lesion burden. Similarly, any measure of cognitive function with face validity, taken as any credible test of brain function, was accepted.

Studies were excluded if reporting exploratory or secondary analysis, or if lesion burden was only related to longitudinal change in cognitive function. Where studies examined both cross-sectional and longitudinal outcomes, or cross-sectional outcomes at more than one time point, the baseline cross-sectional analyses were used. When overlap of reported cohorts was identified and clarification from the original investigators was not possible, a conservative approach was adopted with inclusion of only the earliest dated relevant article.

Studies within the systematic review were suitable for meta-analysis if they reported an overall effect size for the relationship of imaging metrics to a single measure of cognition defined by either a single cognitive test, or a summary result from a cognitive battery.

3.2.3 Data collection

Data was extracted by a single investigator (DM) using a standardised form, which captured:

- Study structure, including design, hypotheses, recruitment pattern and time between cognitive testing and imaging;
- Characteristics of the participants, including age, sex and disease phenotype;
- Cognitive testing methods including blinding and identity of the tester, tests and scoring system used;
- Image acquisition methods;
- Image analysis methods including training and blinding of investigators, software tools used, whether measures of intra- and inter-rater reliability were provided;
- Statistical analysis methods, including controlling for potential confounding factors

A study quality assessment tool (see Appendix C) was also developed, based on STROBE (Strengthening the Reporting of Observational studies in Epidemiology) guidelines [112] to evaluate the risk of bias in individual studies.

The authors for one paper [113] were contacted for further information and numerical data were provided.

3.2.4 Summary measures and synthesis of results

Summary measures were recorded if relating MRI metrics to an overall measure of cognitive function or to a single cognitive test. Where summary measures were provided both unadjusted and adjusted for potentially confounding clinical covariates, adjusted results were used. Correlation coefficients or the difference in lesion burden between groups defined by cognitive status were accepted as summary measures, with preference given to correlations if both were available [114].

All reported summary measures were converted into effect sizes and inverted as necessary so that negative values always indicated an association of lower cognitive scores with higher lesion burdens. Standardised mean differences were calculated from studies reporting group comparisons, prior to conversion to equivalent correlations [114]. An approximation to the standard deviation was estimated as necessary based on available measures of dispersion (e.g. interquartile range or range) [115]. In studies with two impaired groups defined by specific cognitive deficits, these groups were combined before calculation of a standardised difference from a non-impaired group. The Fisher's z transformation [114] was used prior to calculation of an aggregate summary effect, with conversion back to correlations for reporting of overall meta-analysis findings and confidence intervals.

An aggregate summary effect was calculated using maximum likelihood estimation [116] taking into account the size of the various studies; this method allows incorporation of those studies reporting non-significant results without providing their estimate. Separate analyses were carried out for studies measuring hyperintense lesion burden on T2w, FLAIR and/or PD sequences, and for the subgroup of studies evaluating T1w hypointense lesion volume.

Heterogeneity was assessed using Cochran's Q and the I^2 statistic [115]. Tests for heterogeneity test the null hypothesis that all the included studies are evaluating the same effect, with the I^2 statistic quantifying the effect that any inconsistency between studies has on the overall estimate.

3.2.5 Risk of bias across studies

The eligibility criteria required a stated primary aim to evaluate the relationship between cognitive status and brain imaging metrics. This was pre-specified in order to minimise the influence of reporting bias from post hoc analyses. Within the included studies, all analyses that were described without results being provided were recorded. A funnel plot was evaluated visually for asymmetry and tested formally using Egger's regression test.

3.2.5.1 Study quality

An alternative aggregate effect size was calculated using quality scores as an additional scaling factor to study size. This was pre-specified, with the hypothesis that the aggregate effect size would differ with the methodological quality of the study. Study quality was also investigated as a predictor of effect size, using general linear modelling, with all component quality scores or the overall summary score as predictors.

3.2.5.2 Sensitivity analyses

Following discovery of considerable heterogeneity in the image analysis methodology, sensitivity analyses were carried out to investigate the effect of scanner magnet strength and lesion quantification method. Similarly, to investigate heterogeneity in cognitive assessment, a further sensitivity analysis into the effect of using adjusted or unadjusted cognitive scores was carried out.

To explore the possibility of 'true heterogeneity' between study effect sizes, a sensitivity meta-analysis was carried out using a random effects model.

3.2.6 Subgroup analyses - information processing speed tests

Subgroup analyses of studies using the Paced Auditory Serial Addition Test (PASAT) and the Symbol Digit Modalities Test (SDMT), two common tests of information processing speed, were pre-specified to investigate whether focusing on distributed cognitive function would improve correlations with overall lesion burden and replicate previous findings [102].

3.2.7 Additional analyses

3.2.7.1 Disease phenotype

Between-study heterogeneity was further investigated by considering the effect of disease phenotype on effect size. The studies from which an overall effect size could be estimated were classified as having cohorts with relapsing-remitting, progressive, 'benign' or mixed disease courses and separate effect sizes were calculated for each group.

3.2.7.2 Effect of lesion volume

The effect of lesion volume on effect size was investigated where enough information was provided to estimate both a study-specific effect size and a mean cohort lesion volume with standard error. Equivalent lesion volumes were estimated from lesion areas using slice thickness. The effect sizes, on the z-scale were then entered in a linear model, using lesion volume as the predictor, with studies weighted by size.

3.3 Results

3.3.1 Study selection

A total of 3882 studies were identified from the initial literature search, 1975 of which were duplicates (see Figure 3.1). Year-on-year increases were seen in the publication rate identified through the initial search (see Figure 3.2). No additional studies were included following hand searching of journal archives, taken to indicate good coverage by the initial search strategy. After review of abstracts, 139 manuscripts were retrieved. Ninety were subsequently excluded, most frequently ($35/90 = 39\%$) because the primary study aim was not relevant. A total of fifty papers met all inclusion criteria [61, 63, 72–74, 83, 113, 117–159], spanning the period 1987 - 2015.

Thirty studies provided usable summary measures relating hyperintense T2w/FLAIR/PD lesion burden to cognitive function. Two studies reported a 'non-significant' result and one study was excluded from meta-analysis as the reported summary measure was internally inconsistent with other reported results and significance levels. The remaining seventeen studies did not provide results suitable for use in meta-analysis, reporting only individual results for each cognitive subtest ($n = 12$) or multiple regression modelling with simultaneous assessment of several brain imaging metrics ($n = 5$). Thirteen studies reported equivalent summary measures relating cognition to T1w hypointense lesion burden. One study examined the relationship of lesion burden to longitudinal change in cognition, as well as providing baseline cross-sectional data.

3.3.2 Participant characteristics

The total number of subjects from all included studies was 2891. Individual study size ranged from 17 to 327 participants (mean 58, median 45; see Figure 3.3). Forty-four studies specified the sex ratio, all but one having a female majority. The range of mean participant age (provided in 47/50 studies) was 31 to 55 years. No study used age of disease onset in its eligibility criteria. Twenty-six studies included participants with a mixture of disease courses; thirteen studies

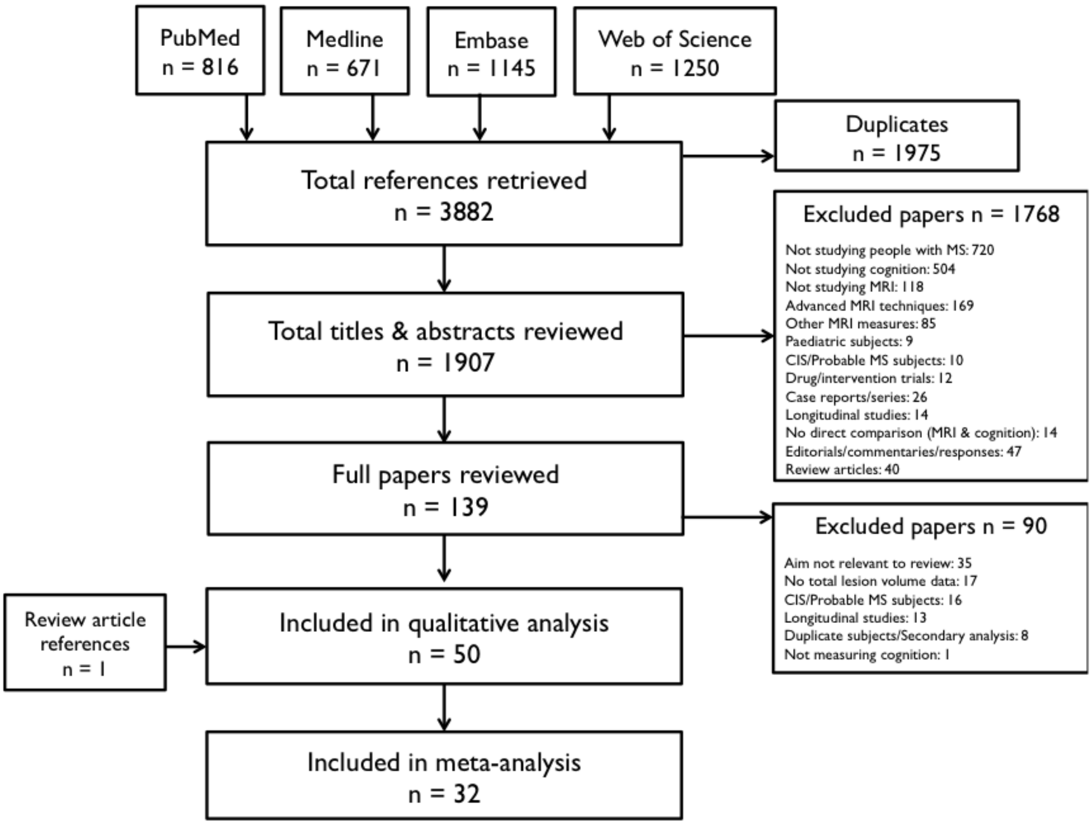

PubMed
n = 816
Medline
n = 671
Embase
n = 1145
Web of Science
n = 1250
Duplicates
n = 1975
Total references retrieved
n = 3882
Excluded papers n = 1768
Not studying people with MS: 720
Not studying cognition: 504
Not studying MRI: 118
Advanced MRI techniques: 169
Other MRI measures: 85
Paediatric subjects: 9
CIS/Probable MS subjects: 10
Drug/intervention trials: 12
Case reports/series: 26
Longitudinal studies: 14
No direct comparison (MRI & cognition): 14
Editorials/commentaries/responses: 47
Review articles: 40
Total titles & abstracts reviewed
n = 1907
Full papers reviewed
n = 139
Excluded papers n = 90
Aim not relevant to review: 35
No total lesion volume data: 17
CIS/Probable MS subjects: 16
Longitudinal studies: 13
Duplicate subjects/Secondary analysis: 8
Not measuring cognition: 1
Review article references
n = 1
Included in qualitative analysis
n = 50
Included in meta-analysis
n = 32

Figure 3.1: Flowchart showing articles retrieved and considered at each stage of the review process.

recruited exclusively relapsing-remitting disease, six studies progressive disease, two ‘benign’, and three did not specify the participants’ disease course.

3.3.3 Image acquisition

The majority (29/50 studies) used 1.5T scanners. Ten studies used scanners with below 1.5T magnets for some or all participants’ imaging, seven used 3T scanners, one used both 1.5 and 3T scanners and three did not specify the scanner field strength. Details of the imaging protocol were given in all but seven studies.

3.3.4 Image analysis

The sequence(s) used to measure lesion volume was specified in 43 studies. Twenty-six specified the number of people involved in the lesion analysis; this was a single observer in 14 studies. The anatomical boundaries of evaluation were explicitly defined in two studies and a sample image was provided by five

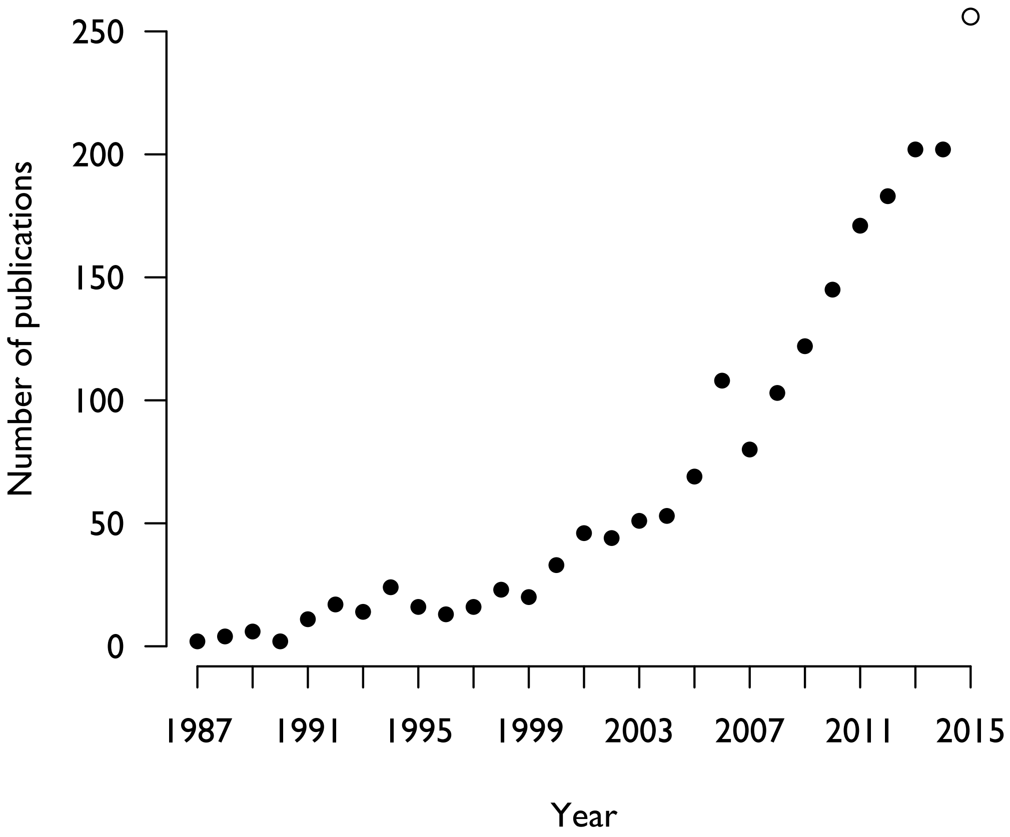


Figure 3.2: Number of results retrieved from database search by year of publication. The point for the year 2015 is an extrapolated value from the 6-month figure.

studies. Only five per cent of studies calculating a lesion volume or area (2/42) normalised to intracranial volume.

A wide variety of approaches were used for the quantification of lesion burden. These included lesion counts (two studies) or weighted lesion scores (six studies), manual lesion outlining either on hard copies (two studies) or within viewing software (six studies), and the use of semi-automated software methods (thirty-one studies). Of the six studies using lesion scores, five different scoring systems were used. One study used both manual and semi-automated measurements (for different sequences), one used manual lesion outlining and an absolute lesion count, and in one study the methodology was unclear.

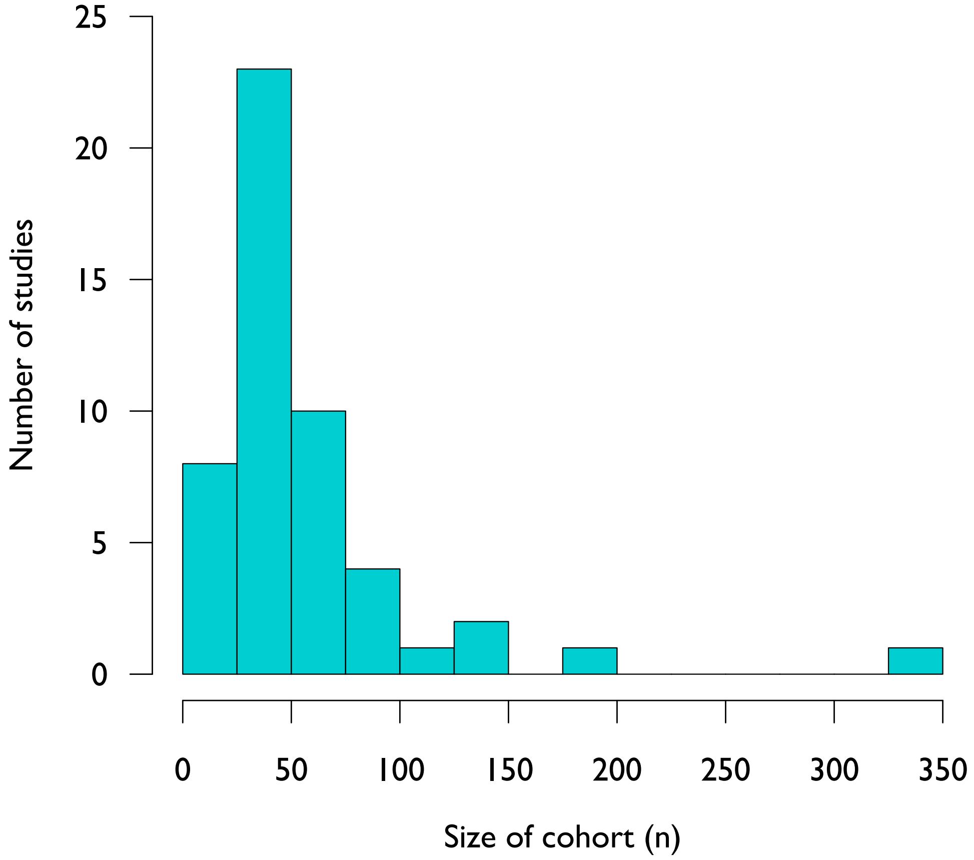


Figure 3.3: Histogram of study sizes

In the thirty-two studies using semi-automated measurement tools, the software used was specified or references provided in 25 studies (78%), covering 14 different software packages. In 18 of these studies the named software was publicly available (11 different softwares). The remaining studies did not specify their software. A manual editing stage for software-generated lesion masks was specified in five studies (16%) and the person performing this was described in two studies. In the ten studies using fully manual lesion outlining, the person performing this was described in six.

Only two studies provided an indication of inter-observer agreement and one study intra-observer reproducibility. Seven studies gave previous measures of reproducibility or results on training data sets.

3.3.5 Cognitive testing

The cognitive assessor and their training were unclear in 38 studies. Of defined batteries, the most commonly used was Rao's Brief Repeatable Battery (12/50), followed by the Minimal Assessment of Cognitive Function in MS (5/50), used with modifications or additional tests in eight (67%) and two (40%) studies respectively. Unique collections of tests were found in 27 studies. The SDMT or PASAT were used either exclusively or as part of a wider battery in 30 studies.

Substantial variability was seen in how raw cognitive scores were processed prior to their use in the evaluation of a possible relationship with imaging metrics. Methods included use of unadjusted scores, standardisation and the deployment of group classifiers. Standardisation was performed using either historic (published or unpublished) or contemporary (matched or unmatched for participant characteristics) control data.

Group classifiers were either based on internal (patient) or external (normative) reference cohorts. The specific thresholds used to define impairment on individual tests were also variable, including 1, 1.5, and 2 standard deviations from the reference mean, and those based on centiles. Moreover, the number of failed tests used to define overall cognitive impairment was also variable (see Appendix D).

Consideration of the effect of potential confounders also varied between studies, both in the recording of relevant data and whether it was adjusted for in the analysis. Some studies adjusted for age ($n = 18$), sex ($n = 12$), education level ($n = 13$) and/or affective disorders ($n = 15$). Drug treatments and premorbid IQ were both adjusted for in three studies. Cognitive leisure activities were neither measured nor adjusted for in any study.

3.3.6 Statistical analysis

Summary measures were provided through univariate correlations ($n = 37$) and/or group comparisons based on cognitive status ($n = 24$). Four studies divided participants into groups dependent on radiological features. Fourteen studies constructed statistical models predicting cognitive performance based on imaging and other laboratory, demographic, or clinical markers.

3.3.7 Reporting quality and risk of bias within studies

A range of study-specific quality scores was seen (mean 42%, SD 11%; Figure 3.4). Among individual elements of the composite quality score, complete reporting was provided most frequently for eligibility criteria and outcome measures (Table 3.1). In contrast, no study provided complete reporting of potential confounding factors, measurement methodology, or a justification of study size.

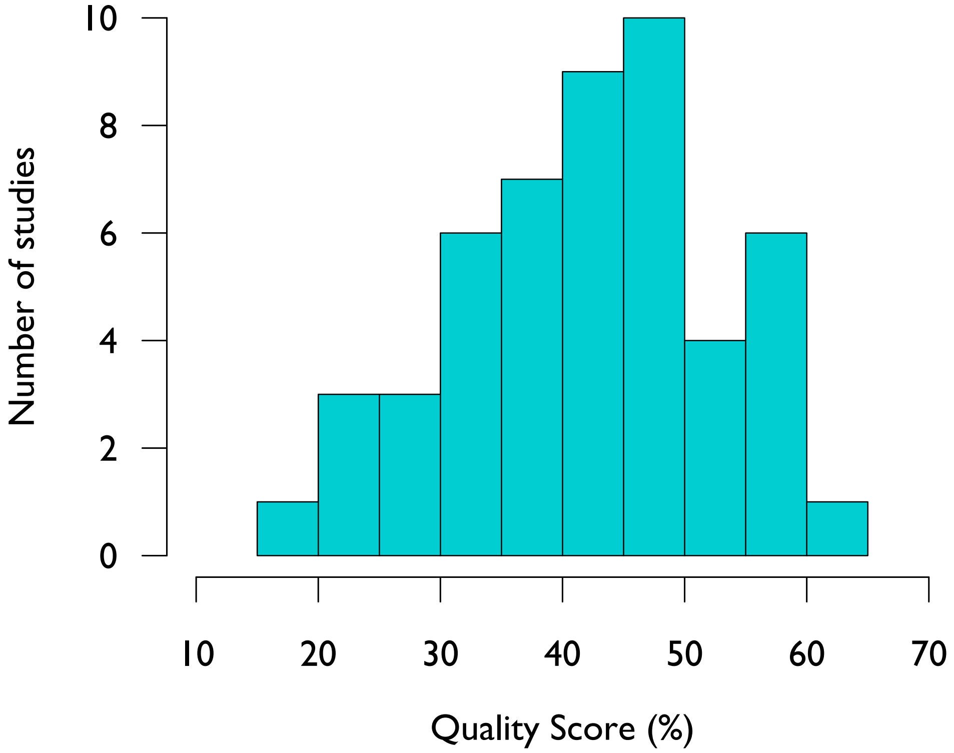


Figure 3.4: Histogram of overall quality scores, expressed as a percentage of the maximum possible score.

3.3.8 Results of individual studies

Studies directly reporting correlation coefficients relating cognitive performance to T2w hyperintense lesion burden gave correlations ranging from −0.6 to

Information reported	Studies gaining each mark (%)		
	0	0.5	1
Eligibility criteria	18	22	60
Individual outcome variables results	8	40	52
Overall outcome results with precision	14	40	46
Quantitative variable handling	10	48	42
Recruitment pattern	60	-	40
Participant characteristics	14	48	38
Statistical methodology	18	50	32
Blinding of assessors	32	50	18
Participant dropout	86	-	14
Objective clearly stated	42	50	8
Cognitive testing & imaging delay	36	58	6
Study design specified	84	10	6
Clearly defined outcomes	8	88	4
Potential confounding factors	36	64	0
Measurement methodology	44	56	0
Study size rationale	100	-	0

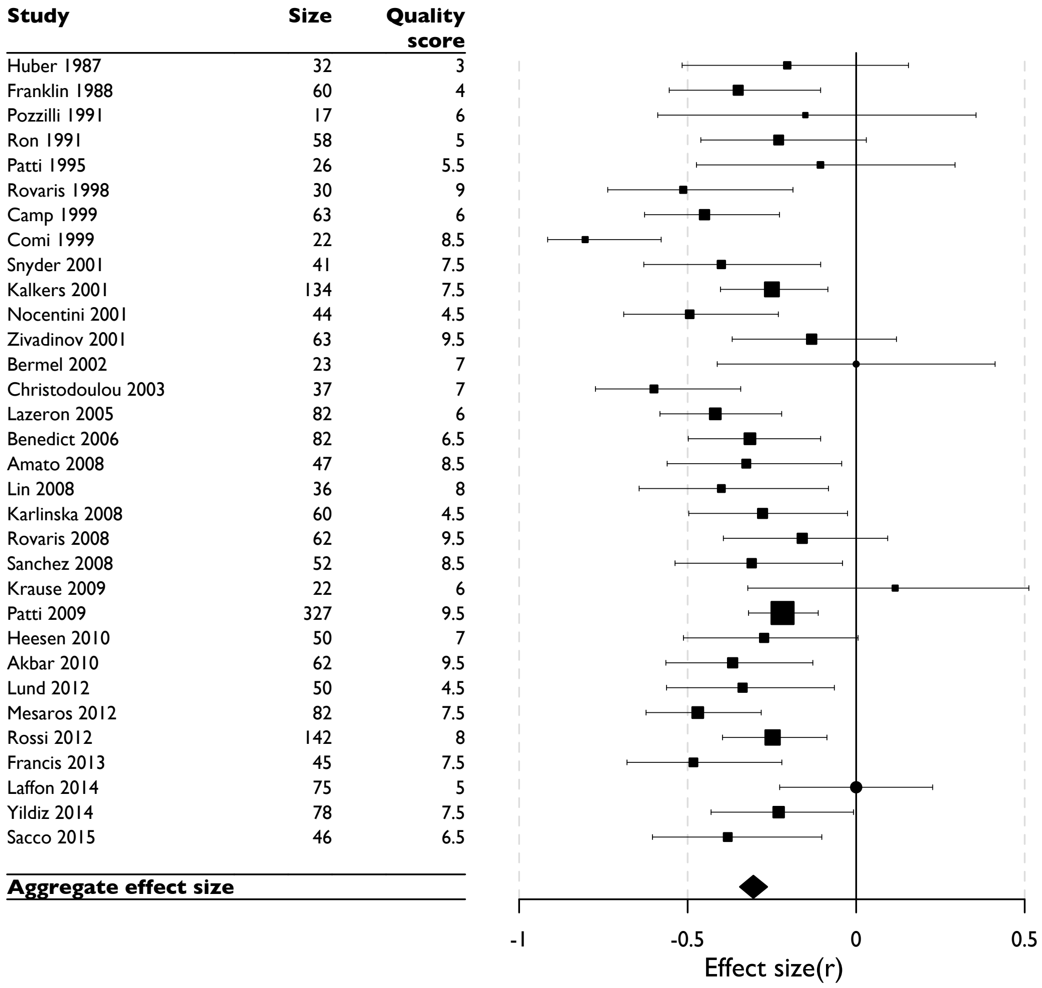
Table 3.1: Table showing percentage of studies gaining 0/0.5/1 for each component of the quality assessment tool.

−0.23. Standardised mean differences ranged from −2.70 to +0.23, equivalent to correlations of −0.80 to +0.11.

3.3.9 Synthesis of results

3.3.9.1 T2w hyperintense lesion burden

The aggregate effect size relating cognitive performance to T2w hyperintense lesion burden was r = −0.30 (95% confidence interval (CI): −0.34 to −0.26; Figure 3.5, n = 32). There was evidence of possible heterogeneity (Q = 43.62, df = 29, p = 0.04; I^2 = 33.5%).

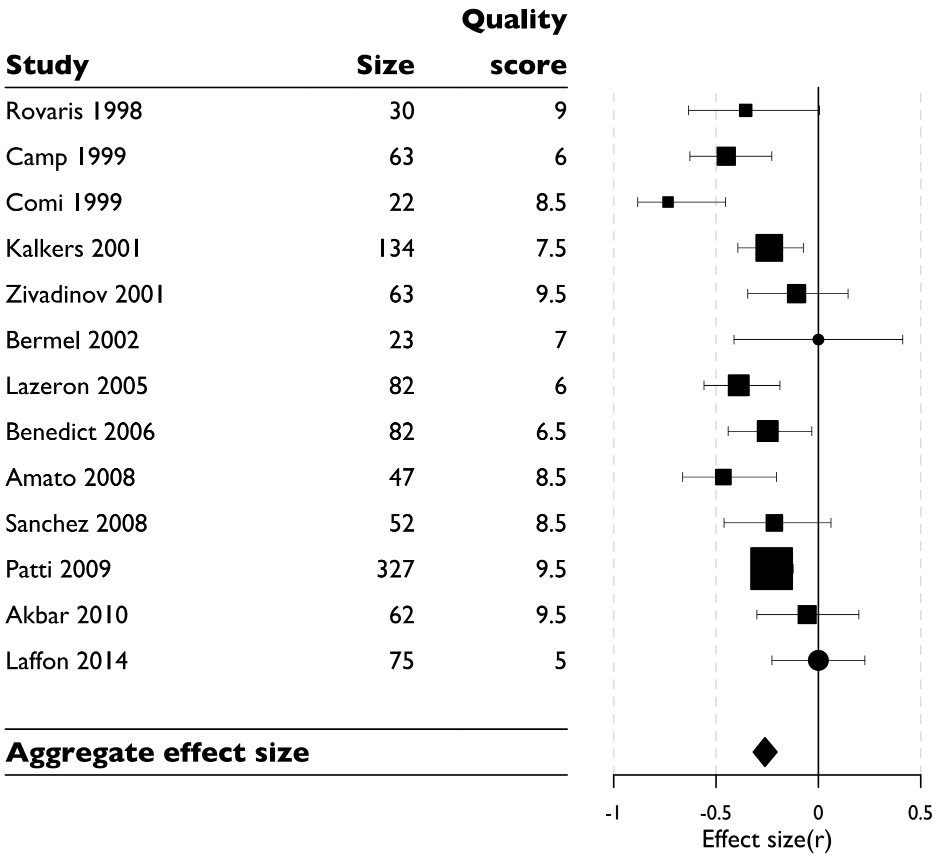


Study	Size	Quality score
Huber 1987	32	3
Franklin 1988	60	4
Pozzilli 1991	17	6
Ron 1991	58	5
Patti 1995	26	5.5
Rovaris 1998	30	9
Camp 1999	63	6
Comi 1999	22	8.5
Snyder 2001	41	7.5
Kalkers 2001	134	7.5
Nocentini 2001	44	4.5
Zivadinov 2001	63	9.5
Bermel 2002	23	7
Christodoulou 2003	37	7
Lazeron 2005	82	6
Benedict 2006	82	6.5
Amato 2008	47	8.5
Lin 2008	36	8
Karlinska 2008	60	4.5
Rovaris 2008	62	9.5
Sanchez 2008	52	8.5
Krause 2009	22	6
Patti 2009	327	9.5
Heesen 2010	50	7
Akbar 2010	62	9.5
Lund 2012	50	4.5
Mesaros 2012	82	7.5
Rossi 2012	142	8
Francis 2013	45	7.5
Laffon 2014	75	5
Yildiz 2014	78	7.5
Sacco 2015	46	6.5
Aggregate effect size		

Figure 3.5: Forest plot of the individual studies using T2w/FLAIR/PD sequences, showing their effect sizes as correlation coefficients. Box sizes are inversely proportional to study variance. Aggregate effect size: r = -0.30; 95% confidence interval: $-0.34, -0.26$.

3.3.9.2 T1w hypointense lesion burden

The aggregate effect size relating cognitive performance to T1w hypointense lesion burden was r = -0.26 (95% CI: $-0.32, -0.20$; Figure 3.6, n = 13). There was evidence of heterogeneity (Q = 20.4, df = 10, p = 0.025, $I^2 = 51.0\%$).



Study	Size	Quality score
Rovaris 1998	30	9
Camp 1999	63	6
Comi 1999	22	8.5
Kalkers 2001	134	7.5
Zivadinov 2001	63	9.5
Bermel 2002	23	7
Lazeron 2005	82	6
Benedict 2006	82	6.5
Amato 2008	47	8.5
Sanchez 2008	52	8.5
Patti 2009	327	9.5
Akbar 2010	62	9.5
Laffon 2014	75	5
Aggregate effect size		

Figure 3.6: Forest plot of effect sizes from individual studies relating T1w hypointense lesion burden to overall cognitive performance, with 95% confidence interval (total n = 1062). Box sizes are inversely proportional to study variance. The overall effect size was r = -0.26 (95% CI: $-0.32, -0.20$).

3.3.10 Risk of bias across studies

Funnel plot inspection (Figure 3.7) and Egger's test of asymmetry (p = 0.05) gave equivocal results. Possible underlying sources of heterogeneity were therefore explored [160].

In order to explore the possibility of 'true heterogeneity' between study effect sizes measured using T2w/FLAIR/PD lesion burden, we performed a sensitivity meta-analysis using a random effects model, giving an overall effect size similar to that of our primary analysis (r = -0.33; 95% CI $-0.38, -0.27$, n = 30). This method did not allow inclusion of the two studies reporting a non-significant result.

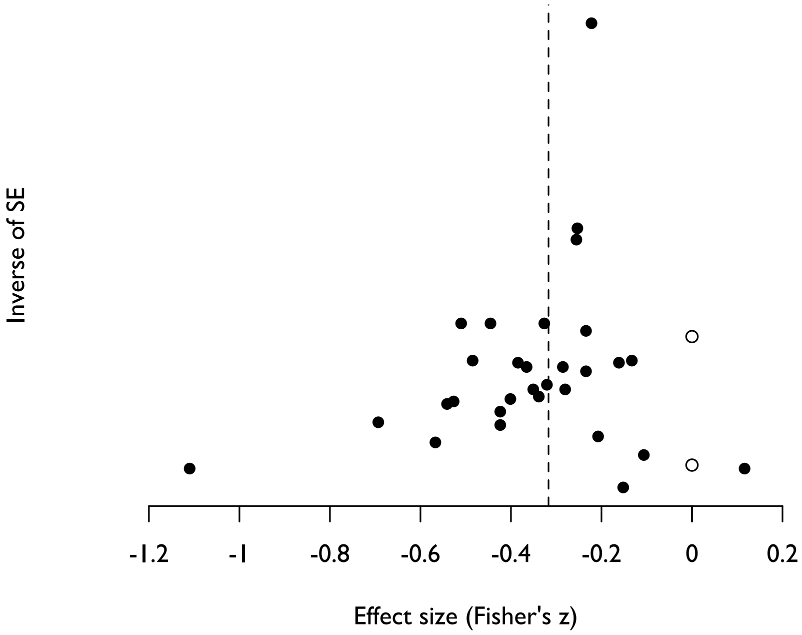


Figure 3.7: Funnel plot of effect sizes, on Fisher's z scale, against the inverse of their standard error (SE, itself inversely related to study size) with asymmetry towards increased reporting of stronger correlations for smaller study sizes. The vertical dashed line indicates the summary effect on the same scale ($z = -0.32$). The unfilled circles correspond to the two studies reporting non-significant results.

In the case of the T1w hypointense lesion burden, an alternative random effects meta-analysis gave a summary effect size of $r = -0.30$ (95% CI: $-0.39, -0.20$). A funnel plot (not shown) showed asymmetry, confirmed by Egger's regression test ($p = 0.032$).

Reporting biases could not be adequately evaluated as study protocols were not published prospectively. Despite methodological heterogeneity apparent from our quality scoring, a significant correlation was not seen between overall quality score and effect size, where reported ($r = -0.18$, $p = 0.34$, $n = 32$). General linear modelling using the individual component scores as predictors of the study effect size identified no statistically significant results ($p = 0.07$ to 0.97).

An exploratory meta-analysis using quality scores as an additional weighting factor returned an effect size similar to that of our primary analysis ($r = -0.30$; 95% CI: $-0.36, -0.24$).

Further sensitivity analyses, comparing scanner field strength and type of lesion quantification method did not demonstrate a measurable subgroup difference in heterogeneity from the small number of studies using high (3T) or low (below 1T) field scanners (see Table 3.2), or from those using lesion counts or scores (see Table 3.3).

Field strength	Studies	Participants	r	95% CI	Q	df	p	I^2
3T	2	91	-0.43	(-0.59, -0.25)	0.33	1	0.56	0
1/1.5T	21	1478	-0.29	$(-0.34, -0.24)$	34.43	20	0.02	41.9%
Below 1T	4	188	-0.32	$(-0.45, -0.18)$	3.56	3	0.32	0
Overall	30	1952	-0.31	$(-0.35, -0.26)$	43.62	29	0.04	33.5%

Table 3.2: Results of sensitivity analysis, comparing study effect sizes (r) by scanner field strength. CI: confidence interval.

Method	Studies	Participants	r	95% CI	Q	df	p	I^2
Manual outlining	6	209	-0.30	$(-0.42, -0.16)$	6.87	6	0.14	41.7%
Semi-automated	21	1320	-0.33	$(-0.38, -0.28)$	30.55	20	0.03	41.1%
Scores/counts	4	217	-0.24	$(-0.38, -0.10)$	0.9	3	0.92	0
Overall	30	1952	-0.31	$(-0.35, -0.26)$	43.62	29	0.04	33.5%

Table 3.3: Results of sensitivity analysis, comparing study effect sizes (r) by lesion burden quantification method. CI: confidence interval.

A further post hoc sensitivity analysis was also performed using the same methodology as the main analysis, incorporating all potentially analysable data from the 139 studies considered at the full paper review stage. This returned an aggregate effect size of $r = -0.31$ (95% CI: $-0.34, -0.28$; $n = 65$ studies, total participant number $= 3430$).

3.3.11 Subgroup analyses - alternative cognitive endpoints

Exploratory meta-analyses were performed on two widely used measures of information processing speed, the SDMT and PASAT. The a priori hypothesis was that total lesion burden would have a stronger correlation with these tests of distributed cognition function compared to the mixture of distributed and localised functions in our primary analysis. The summary effect size for SDMT was r = −0.37 (95% CI: −0.43, −0.31; n = 13 studies) and for PASAT was r = −0.28 (95% CI: −0.34, −0.22; n = 15 studies). See Figures 3.8 and 3.9.

A post hoc sensitivity analysis considering the effect of using raw or adjusted cognitive scores was also performed. Twenty-one of the 32 studies included in our primary endpoint meta-analysis were identified to have adjusted their cognitive scores, with an aggregate effect size (between T2w hyperintense lesion volume and cognitive performance) of r = $-0.31(-0.36, -0.26)$. Eleven studies were identified not to have adjusted their cognitive scores, with an aggregate effect size of r = $-0.29(-0.37, -0.21)$.

3.3.12 Additional analyses

3.3.12.1 Effect of disease phenotype

Twenty-nine of the 32 studies in the primary meta-analysis provided information on the disease course of their participants. Summary effect sizes and tests of heterogeneity for each group are displayed in Table 3.4. The mixed phenotype group was the only group to show evidence of heterogeneity, although the sample sizes may have been insufficient to exclude this in the 'benign' and progressive groups.

Disease course	Studies	Participants	r	95% CI	Q	df	p	I^2
Benign	2	109	−0.23	$(-0.41, -0.04)$	0.79	1	0.37	0
Relapsing-remitting	11	946	−0.24	$(-0.30, -0.18)$	3.98	9	0.91	0
Mixed phenotypes	11	626	−0.37	$(-0.44, -0.30)$	23.3	9	0.006	61.4%
Progressive	5	238	−0.41	$(-0.51, -0.29)$	3.72	4	0.445	0

Table 3.4: Summary effect sizes, with studies grouped by participant phenotype. *N.B.* The relapsing-remitting and mixed phenotype groups both included one study reporting only a non-significant result. This could be used in the calculation of a summary effect size but not the heterogeneity measures, hence the drop in the degrees of freedom (df). CI: confidence interval.

Study	Size	Quality score
Franklin 1988	60	4
Hohol 1997	44	7
Bermel 2002	23	7
Christodoulou 2003	37	7
Deloire 2005	58	8.5
Lazeron 2005	82	6
Houtchens 2007	31	10
Patti 2009	327	9.5
Heesen 2010	50	7
Mike 2011	26	7.5
Bomboi 2011	19	9.5
Mesaros 2012	82	7.5
Sacco 2015	46	6.5
Aggregate effect size		

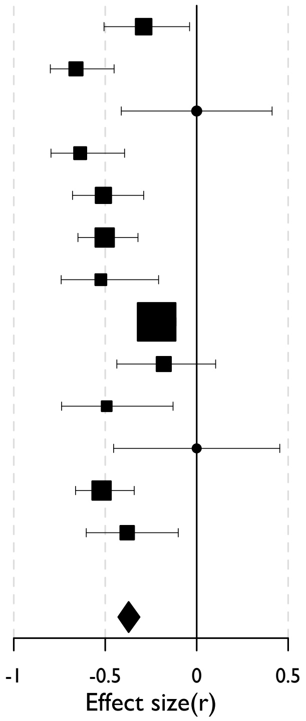


Figure 3.8: Forest plot of effect sizes from individual studies relating T2w hyperintense lesion burden to SDMT performance, with 95% confidence interval (total n = 885). Box sizes are inversely proportional to study variance. The overall effect size was r = -0.37 (95% CI: $-0.43, -0.31$). There was evidence of heterogeneity (Q = 30.7, df = 10, p = 0.001, I^2 = 67.4%). An alternative random effects meta-analysis gave a summary effect size of r = -0.45 (95% CI: $-0.55, -0.33$). To investigate the heterogeneity, a funnel plot was drawn. Egger's regression test confirmed evidence of funnel plot asymmetry (p = 0.0001).

3.3.12.2 Effect of total lesion volume on the reported strength of association

Twenty-one of the 30 studies providing data from which to calculate an effect size also gave relevant summary statistics for lesion volume. Five studies employed lesion counts or scores and four studies did not provide summary statistics for lesion burden.

Individual study effect sizes are plotted against lesion volume in Figure 3.10. In a linear model predicting effect size, cohort mean lesion volume did not

Study	Size	Quality score
Hohol 1997	42	7
Snyder 2001	41	7.5
Kalkers 2001	134	7.5
Christodoulou 2003	37	7
Deloire 2005	58	8.5
Lazeron 2005	82	6
Houtchens 2007	31	10
Lin 2008	36	8
Patti 2009	327	9.5
Heesen 2010	50	7
Mike 2011	26	7.5
Mesaros 2012	82	7.5
Sbardella 2013	36	7
Laffon 2014	75	5
Sacco 2015	46	6.5
Aggregate effect size		

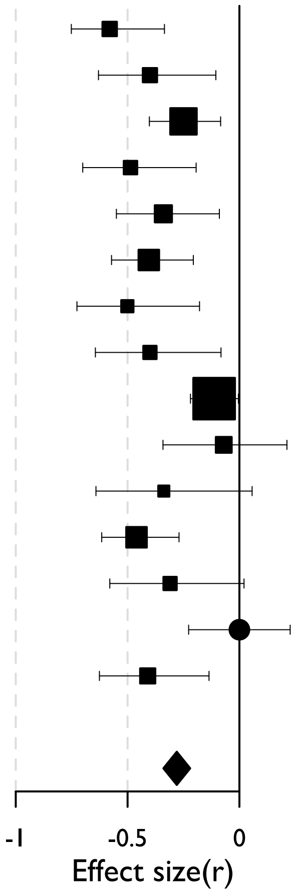


Figure 3.9: Forest plot of effect sizes from individual studies relating T2w hyperintense lesion burden to PASAT performance, with 95% confidence interval (total n = 1103). Box sizes are inversely proportional to study variance. The summary effect size was $r = -0.28$ (95% CI: $-0.34, -0.22$). There was evidence of heterogeneity ($Q = 29.2$, $df = 13$, $p = 0.006$, $I^2 = 55.5\%$). An alternative random effects meta-analysis gave a summary effect size of $r = -0.35$ (95% CI: $-0.44, -0.26$). To investigate the heterogeneity, a funnel plot was drawn. Egger's regression test confirmed evidence of funnel plot asymmetry ($p < 0.0001$).

reach significance as a predictor ($p = 0.066$) but showed a trend towards larger magnitude effect sizes with increasing mean cohort lesion volumes.

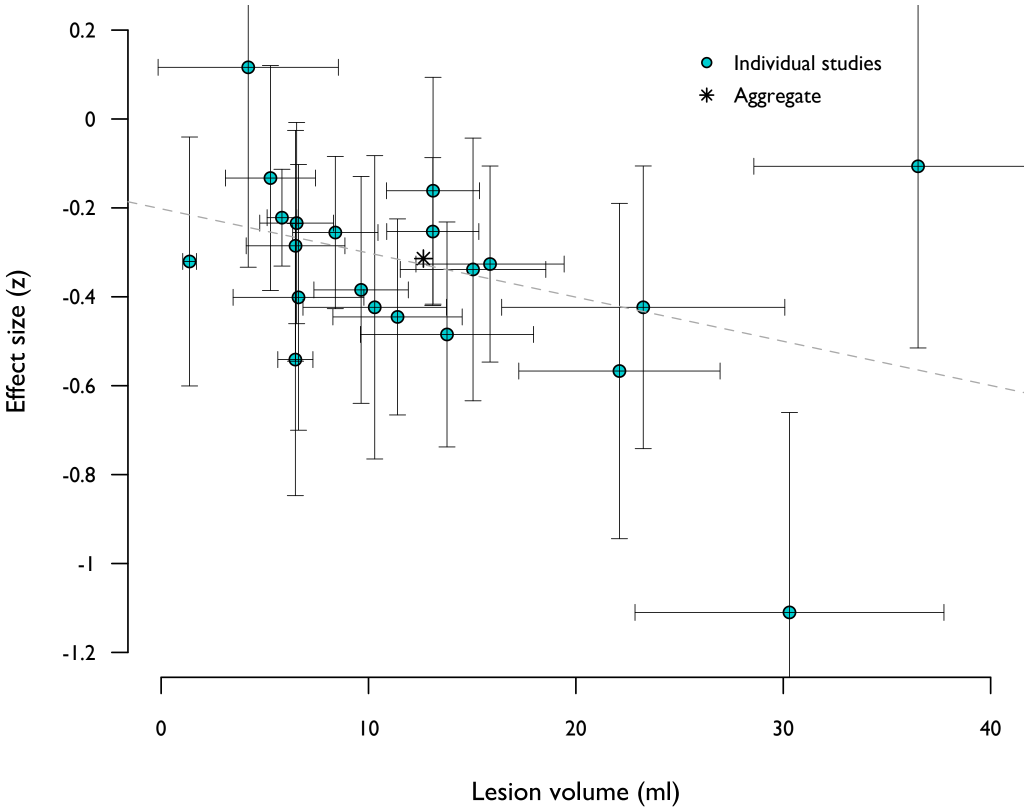


Figure 3.10: Plot of individual study effect sizes with 95% confidence intervals against estimated cohort mean lesion volume with 95% confidence interval for estimate, based on standard error. Regression line from linear model showing relationship between lesion volume and effect size (p = 0.066).

3.4 Discussion

Synthesis of published findings confirms a modest correlation (r = -0.30) between MRI measures of total brain white matter lesions and cognitive function in people with MS. Although variability was observed between studies in the magnitude of the reported relationship, no large (> 100 participants) single study demonstrated a strong correlation. Technical and methodological factors were therefore examined to determine their potential impact on the reported correlation. These were broadly divisible into three dimensions: variability in cognitive assessment and scoring; variability in cohorts studied; and variability in image acquisition and analysis.

Substantial variability was seen with respect to both the techniques used to evaluate cognitive function and the adjustment for other variables that might influence cognition (e.g. education, premorbid IQ and drugs). This may however represent a largely historic issue [35], as a global movement is emerging to harmonise evaluation and scoring through the Brief International Cognitive Assessment for MS (BICAMS) initiative [36]. In contrast, the optimum method to generate quantifiable measures of lesion burden from brain imaging data lacks emergent consensus. Recent attempts to harmonise MRI acquisition protocols [161, 162] have been made, however no similar initiative exists for image analysis techniques.

Considerable heterogeneity was observed in the clinical cohorts studied. The importance of this lies in the possibility that the fundamental relationship between white matter lesions and cognition may differ between cohorts depending on their characteristics. There was some suggestion from the secondary analyses in support of this. Larger effect sizes were found in cohorts of participants with greater lesion burdens and later stage/progressive disease courses. If confirmed, the existence of a dynamic association (dependent on cohort characteristics) raises questions about the fundamental relationship between white matter lesions and cognition. Possible reasons for this variation include interactions with other aspects of MS pathology that emerge independently from disease duration and progressive lesion burden, progression of lesions with disease course (e.g. greater pathological homogeneity in chronic lesions), or that the ability to compensate (functionally) for pathology declines in a non-linear form dependent upon time and/or total lesion burden.

Semi-automated approaches were the most frequently used for image analysis (62%) and merit particular consideration. While effective manual editing is clearly dependent on adequate training of the operator, the automated (software) component is more challenging to benchmark. Authors should routinely report the software used. Separately, the field risks delaying progress and reducing the potential for collaboration due to the many differing software packages used. Of the 24 studies naming software, ten different publicly available (commercial or open source) packages were used, and a further three packages that were developed 'in house'. As yet no comparative study has been performed on a common dataset to evaluate agreement between these varied approaches. A new consensus initiative to support an image analysis framework in MS would enable benchmarking while also supporting ongoing innovation.

Despite the finding of substantial methodological variability between studies, formal testing for heterogeneity in the primary meta-analysis returned an equivocal result. This indicates that methodological variability between studies cannot provide a sufficient explanation for the cognitive clinicoradiological paradox. Nevertheless, measurement errors within all published studies may have attenuated observed correlations in the face of a higher 'true' correlation [103]. Greater recognition and transparency around measurement error for both cognitive and lesion burden quantification would therefore be beneficial to the field.

The findings may have been limited by an overly inclusive approach to both the evaluation of cognition and white matter lesion burden. With respect to the former, a higher aggregate correlation was observed between white matter lesion burden and cognition measured by the SDMT, a measure of information processing speed, understood to reflect widely distributed brain connectivity, than was seen for cognition as defined in the primary analysis. Furthermore, a substantial body of potentially relevant data was excluded from this review as the primary aim of the study was unclear or reported findings were secondary/exploratory analyses. Notably, relatively few studies used >1.5T field strength scanners, in part reflecting the recent shift away from exploring the relationship between phenotype and T2w hyperintense lesion burden, focusing instead on the possible relevance of other MRI metrics. Finally, despite best efforts to apply a systematic approach, all reviews are conducted by researchers who bring unconscious bias [163] and the lack of replication of the literature search and data extraction by a second investigator is a limitation.

In conclusion, a modest correlation ($r = -0.30$) exists between MRI measures of total brain white matter lesion burden and cognitive function in people with MS. This review has highlighted the substantial variability existing in the literature addressing this question, particularly with respect to cognitive methodologies, cohort characteristics and imaging methodology. This variability was insufficient to fully account for the cognitive clinicoradiological paradox and resolving this will therefore likely require simultaneous evaluation of multiple components of the complex pathology using optimum measurement techniques for both cognitive and imaging feature quantification [164]. Nevertheless, measurement errors from the existing techniques to quantify lesion burden act to attenuate the strength of the observed relationship, obscuring any current attempt to quantify the true strength of that relationship. Optimised measurement of lesion burden is therefore essential. Against that background, the move to harmonise cognitive assessment in MS is valuable, but no similar move to optimise and harmonise quantification of lesion burden has emerged in the MRI community. This frames the central issue that is addressed further throughout this thesis.

It appears that the strength of association may also vary dependent upon the population being studied, in particular varying with respect to the total burden of white matter pathology and the emergence of progressive disease. This raises questions about the potential mechanism(s) of a dynamic relationship between white matter pathology and cognitive function. These questions are also explored in subsequent chapters.

Chapter 4

Assessing the reliability of the reference standard for white matter hyperintensity quantification

4.1 Introduction

Research studies in people with multiple sclerosis (MS) have to date frequently used imaging-based outcomes, often involving quantification of white matter hyperintensity (WMH) volume. This can then be used to investigate imaging correlates of disability or monitor treatment effects. Spatial templates of WMHs can be used to interpret advanced imaging markers, such as those derived from magnetic resonance spectroscopy (MRS), magnetisation transfer (MT) imaging or diffusion tensor imaging (DTI). Any lack of accuracy in the measurement of WMHs will attenuate the results derived.

As highlighted in the previous chapter, a wide variety of approaches are used to quantify white matter disease burden in MS and their comparability is not always clear. In order to establish the validity of any method, its relationship to a set of reference standard measurements should be demonstrated. Ideally this should be within a population with a similar profile to that of its intended use.

A true evaluation of disease burden would require pathological correlation. Such studies are valuable [165] but for practical and ethical reasons are not possible in large numbers. They can also only ever represent a limited sample of the varied disease courses seen in people with MS, most often those with early atypical inflammatory disease undergoing biopsies or from autopsy material showing endstage disease. In the absence of pathological confirmation, the reference standard for imaging research in practice is usually taken to be manual segmentation by a user experienced in interpreting imaging changes. This is frequently a neuroradiologist or another user supervised by a radiologist. However very little data on the reliability of this reference standard has been published and the systematic review reported in the previous chapter found poor reporting of any reproducibility measures.

In this chapter, the accepted practice used elsewhere is followed, establishing the reference standard in a relevant population for later use. This is undertaken using the Advanced MRI substudy cohort of MS-SMART (see Chapter 2, Section 2.1). Inherent to this process, but often omitted or not reported in the literature, is an assessment of its reliability, both in providing a unidimensional quantification of the WMH burden and a spatial template ('mask' or 'map') of these changes. Factors affecting the reliability and stability of the reference standard are considered. As an experienced neuroradiologist is generally accepted as an optimal observer for providing the reference quantification, the reliability seen between two neuroradiologists is investigated.

4.2 Methods

4.2.1 Participants and Imaging

This work was performed using the routine structural imaging sequences performed at the baseline assessment of all participants (n = 43) recruited in Edinburgh to the Advanced MRI substudy of MS-SMART. See Chapter 2, Section 2.1 for further details of the cohort, image acquisition and post-processing.

4.2.2 Segmentation protocol

A single observer (DM, neuroradiologist with 4 years' experience), blinded to all clinical and demographic information, outlined all WMHs using freehand drawing tools available in the Mango image analysis software (http://ric.uthscsa.edu/mango/). This segmentation was performed on the registered fluid attenuated inversion recovery (FLAIR) sequences with T1-weighted (T1w) and T2-weighted (T2w) sequences available for reference. Segmentation was completed for the entire cohort over a period of seven months, partly covering their recruitment stage, in sessions lasting up to 3 hours.

The segmentation process was performed primarily in the axial plane, with adjustments as necessary, using reformatted coronal and sagittal projections. An inferior boundary for WMH segmentation was set at the foramen magnum. Where possible, WMH boundaries were chosen to enclose areas of abnormal signal on both T2w and FLAIR sequences. Viewing windows were optimised on an individual subject basis to make WMHs as clearly distinct from surrounding tissue as possible. No minimum size for WMH segmentation was set. Enlarged perivascular spaces were not marked, unless they appeared inseparable from focal or diffuse WMHs.

Completed masks were saved as binary files in the NIfTI (Neuroimaging Informatics Technology Initiative) format. Absolute volumes for each WMH mask were calculated using tools available in the FSLstats software package (FMRIB software library (FSL) [104]).

4.2.2.1 Intra-observer reliability

Following completion of the initial WMH masks, the manual segmentation process was repeated for the whole cohort on two further occasions, with random reordering of the scans for each set of masks. The time interval between segmentations of the same scan was at least six weeks.

4.2.2.2 Inter-observer reliability

A subset of 12 scans were pseudo-randomly selected to ensure even coverage of all quartiles by WMH volume, based on the initial mask of the first observer. The manual segmentations were repeated by a second observer (neuroradiologist [GM] with 4 years' experience), using the same protocol, blind to all clinical and demographic information and the results of the initial segmentation. These were used for investigation of unidimensional reliability and spatial agreement, using comparisons to the third (final) manual segmentation by the first observer.

4.2.3 Statistical analysis

4.2.3.1 Unidimensional reliability

Reliability of absolute WMH volumes were compared for the initial, repeat and second observer segmentations using intraclass correlations (ICCs; Class 2, a random effects, two-way model based on single observations), reflecting absolute agreement, and assessed visually using Bland-Altman plots [166] of volume ratios.

Spearman correlation coefficients were calculated for each combination of two mask sets.

The effect of WMH volume on reliability measures was explored graphically, using the Bland-Altman ratio plots, and by calculating ICCs separately for the 21 scans with the highest and lowest mean WMH volumes.

The 'psych' package in R software was used for calculation of ICCs.

4.2.3.2 Spatial agreement

Spatial agreement between masks was assessed using the Dice similarity coefficient [108] generated using script written in Matlab (provided by MB). This measures the voxel overlap between the two masks and is defined as twice the ratio of the number of overlapping voxels to the sum of the voxels in each segmentation.

Visual evaluation of discrepancies was carried out in FSL viewing software following completion of successive mask sets.

4.3 Results

4.3.1 Intra-observer reliability

4.3.1.1 Participant characteristics

The baseline imaging from all 43 participants in the Advanced MRI substudy of MS-SMART was used. There were 30 female and thirteen male participants with a median age of 55.5 years (interquartile range (IQR): 49.9, 62.0). All participants had a diagnosis of secondary progressive MS (SPMS) with a median total disease duration of 23.4 years (IQR: 15.6, 27.3).

4.3.1.2 Summary statistics

Manual WMH segmentation by the initial observer was completed over a period of 7, 2 and 4 months respectively for mask sets 1 to 3, with a delay of at least six weeks between sets 1 & 2 and sets 2 & 3. The mean time (± standard deviation) taken for each mask was 44 ± 29, 50 ± 26 and 38 ± 21 minutes for mask sets 1 to 3 respectively.

Overall cohort WMH volumes were highest for the second set of masks, with mask set 3 intermediate between the first two. All three volume sets were positively skewed. Cohort WMH volumes for each of the three sets of masks are represented in boxplots shown in Figure 4.2 and summarised in Table 4.1.

An example of a manual WMH segmentation overlaid on the FLAIR sequence is shown in Figure 4.1.

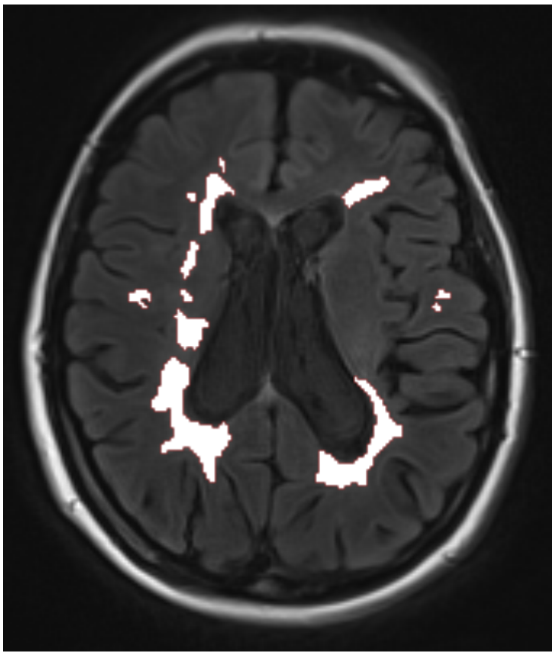


Figure 4.1: Sample image showing manual WMH segmentation overlaid on FLAIR sequence from one of the MS-SMART participants.

Mask	Mean ± SD	Median (IQR)
1	19.7 ± 20.8	$12.9\,(6.1, 22.6)$
2	26.7 ± 23.0	$21.2\,(12.5, 34.7)$
3	22.9 ± 22.0	$16.9\,(7.6, 30.1)$

Table 4.1: Summary statistics for mask sets 1 to 3, given as volumes (ml). IQR: interquartile range; SD: standard deviation.

4.3.1.3 Unidimensional reproducibility

Sets of WMH volumes for the different segmentations showed (Spearman) correlations of r = 0.92 to 0.94, and ICCs of 0.91 to 0.96 for each possible two-way

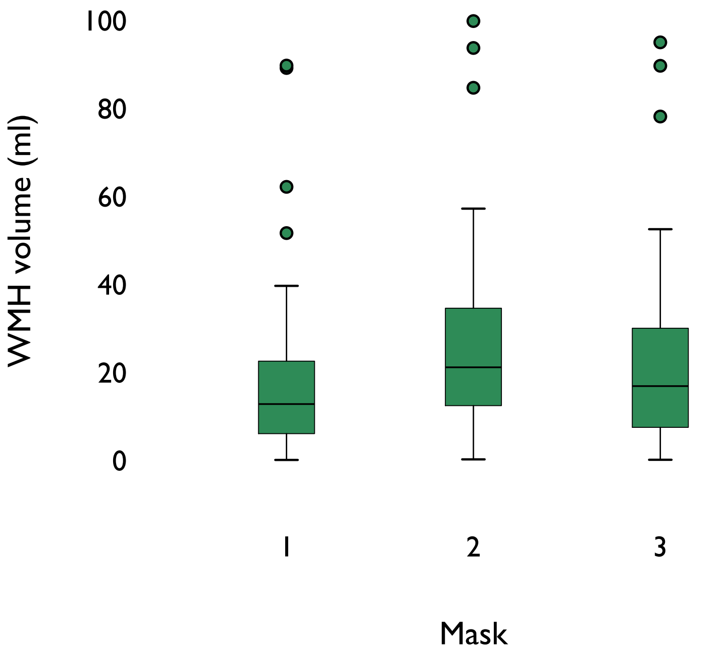


Figure 4.2: Boxplots of cohort WMH volumes for mask sets 1 to 3.

.

comparison. The overall ICC for all three mask sets was 0.94. Full results are given in Table 4.2.

Table 4.2 also shows the comparison of ICCs for the 21 scans with the lowest and highest mean WMH volume. Although the confidence intervals are overlapping, higher ICCs are associated with higher WMH volumes. The Bland-Altman plots of the volume ratios between masks from different sets are shown in Figure 4.3 and suggest a similar finding. Greater relative discrepancies between segmentations are seen at lower WMH volumes, with a trend towards convergence on more similar values at higher volumes.

The improvement in agreement with successive mask sets is also apparent from the Bland-Altman plots. The largest range of relative discrepancies and widest confidence interval for the mean ratio is seen for the comparison between the first two mask sets, and the range and confidence interval are smallest for the final comparison (mask sets 2 & 3). The final comparison mean ratio below 1 (= 0.81) reflects the overall larger WMH volumes generated in mask set 2.

Masks	r	ICC (95% CI)	Mean Dice
1 v. 2	0.92	0.91 (0.42,0.97)	0.68
1 v. 3	0.94	0.96 (0.88,0.98)	0.74
2 v. 3	0.93	0.96 (0.85,0.98)	0.73
Overall	-	0.94 (0.83,0.98)	0.71
Lowest 21	-	0.67 (0.31,0.86)	0.65
Highest 21	-	0.92 (0.74,0.97)	0.78

Table 4.2: Spearman correlations (r), intra-class correlations (ICCs, Class 2) with 95% confidence interval (CI), and mean Dice indices for each two-way mask comparison and overall for the three mask sets where possible. The overall comparison is also divided into upper and lower groups by mean lesion volume, with these examined separately.

4.3.1.4 Spatial agreement

The overall grand mean of the Dice indices for spatial overlap was 0.71, covering all mask comparisons. Mean values for each two-way comparison are given in Table 4.2. As with the unidimensional measures of reliability, the highest mean values for the Dice index are seen for comparisons with the final mask set (3), suggesting convergence on a stable segmentation.

Similar trends in spatial agreement within the cohort were seen to unidimensional measures, with increasing Dice indices at higher WMH volumes. This trend is shown graphically in the scatterplots of Figure 4.4. This is a recognised limitation of the Dice index, which measures only agreement in the regions considered of interest, where larger regions are clearly more likely to overlap, disregarding agreement on tissue outwith these regions.

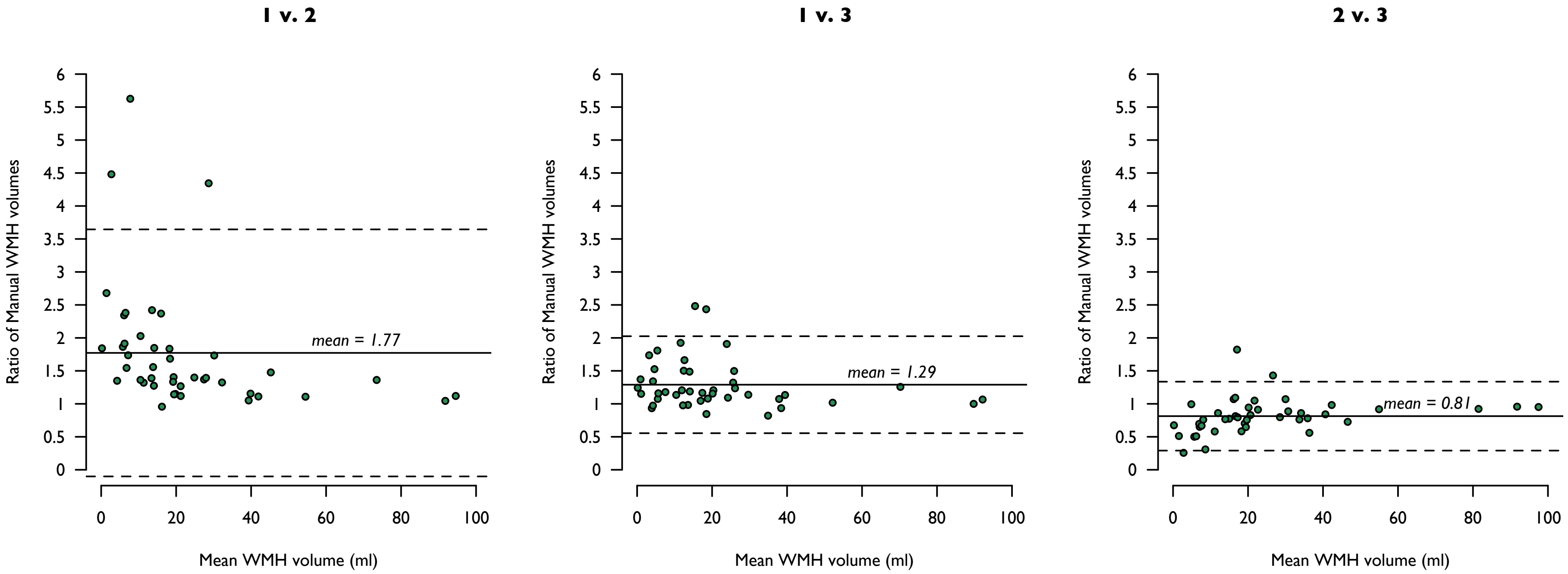


Figure 4.3: Bland-Altman plots, showing ratio of WMH volumes for each scan (n = 43) compared between different mask sets (1 to 3). In each case the ratio is that of the later to the earlier mask. The solid line shows the mean ratio for the mask comparison and the dashed lines are 95% confidence intervals for the mean.

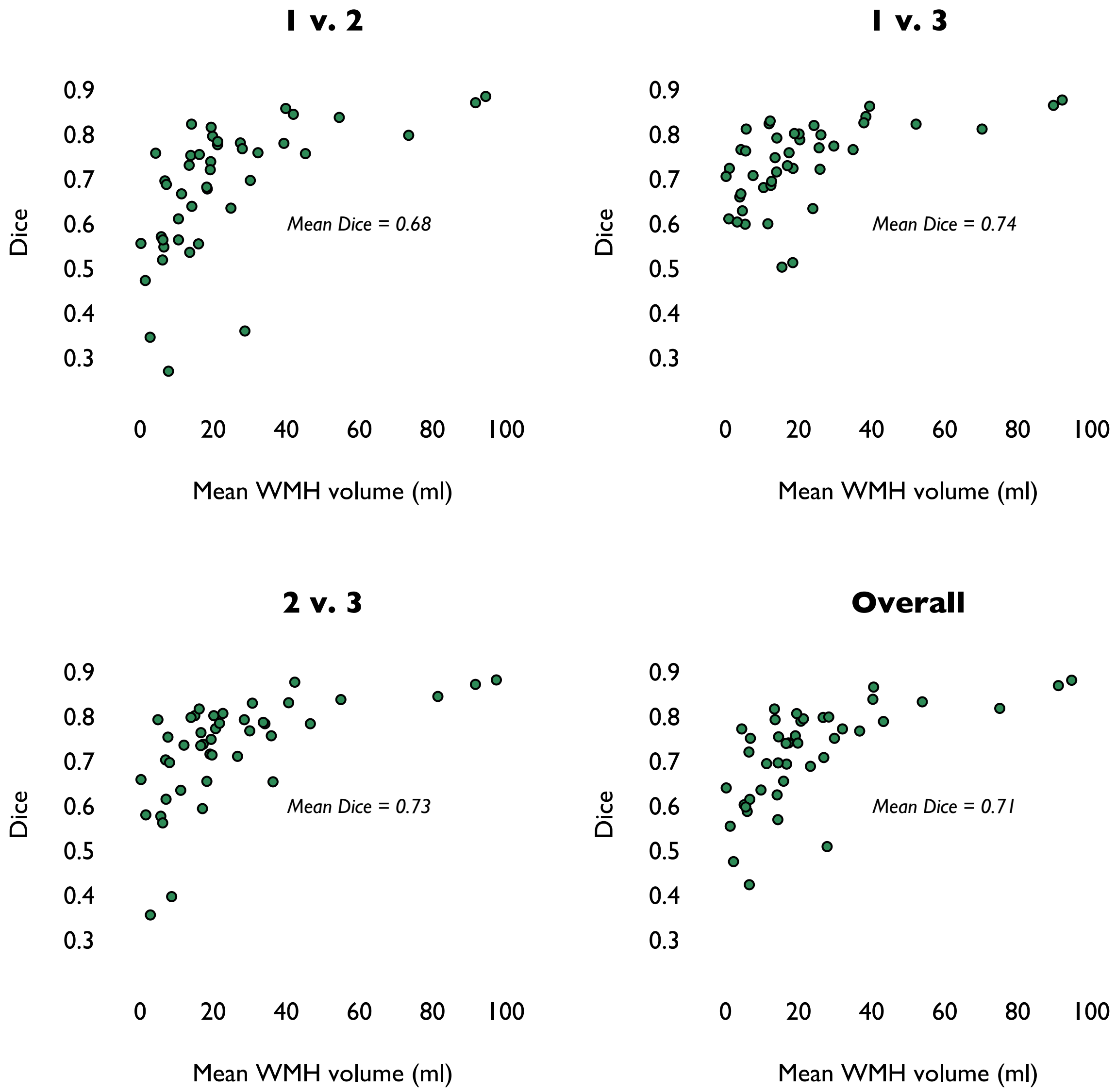


Figure 4.4: Scatterplots of Dice index of intra-observer spatial agreement against mean WMH volume for each two-way comparison between mask sets 1 to 3. The lower right plot shows the mean of the Dice index for all comparisons against the mean WMH volume for each scan.

4.3.1.5 Sources of discrepancy

The review of WMH masks with the largest discrepancies revealed that the main source was large regions of 'dirty' white matter. This was white matter with signal characteristics intermediate between that of the overlapping segmented regions and those voxels designated 'normal-appearing' white matter in both segmentations. For many subjects, this intermediate signal tissue represented a major portion of the white matter. Using the Mango software to adjust viewing windows in some cases highlighted apparent edges to the abnormal white matter, but this was not universal.

In several scans, high signal was seen to extend along the corticospinal tracts, on one or both sides, suggestive of Wallerian-like degeneration triggered by focal damage within the tract. This was a further contributor to the volume of white matter with signal intermediate between that of focal inflammatory WMHs and apparently unaffected white matter.

The FLAIR sequence had a slice thickness of 3mm and 'partial volume' effects were apparent, leading to uncertainty in delineating WMHs. This was particularly noticeable at the ventricular surface and cortical boundary, both frequent sites for MS-related WMHs. The lateral ventricles were the relatively larger contributor to this effect, although segmentation using primarily the FLAIR sequence had been chosen in order to minimise this.

While failure to recognise small focal WMHs did occur, this will have only had a significant effect on agreement metrics for the scans with the very lowest WMH volumes.

4.3.2 Inter-observer reliability

4.3.2.1 Participant characteristics

The 12 scans used were the baseline imaging from a subset of the forty-three participants in the Advanced MRI substudy of MS-SMART described earlier. There were nine female and three male participants with a median age of 57.6 years (IQR: 45.7, 61.0) and a median disease duration of 21.9 years (IQR: 14.9, 29.5).

As quantified in the third mask set by the initial observer, the median WMH volume of these participants was 19.5ml (IQR: 12.4, 32.8).

4.3.2.2 Summary statistics

Manual segmentation by the second observer was completed over a period of two weeks. The median WMH volume from the manual segmentations of the second observer were all higher than those for the initial observer, with a median volume of 44.8ml (IQR: 26.4, 61.7). A boxplot comparing the two observers is shown in Figure 4.5.

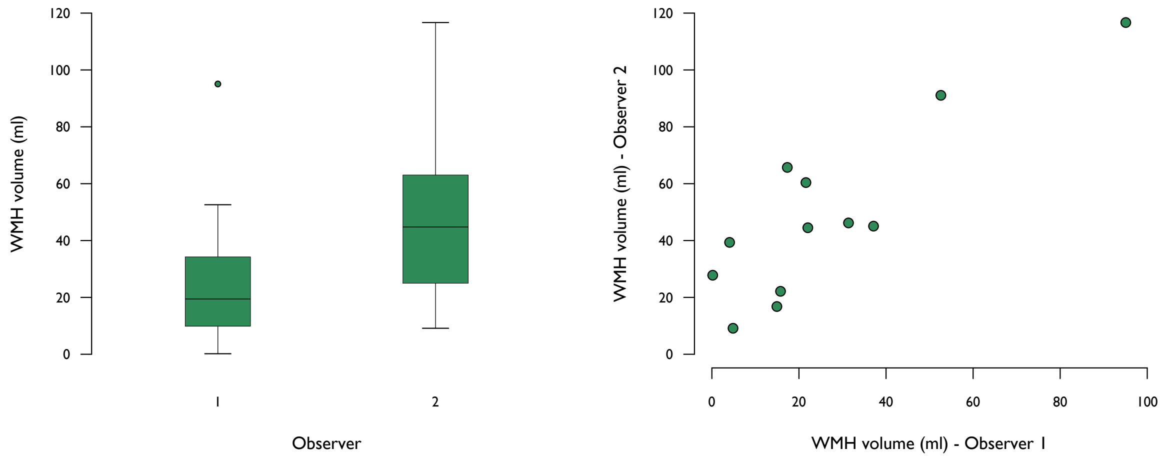


Figure 4.5: Left - Boxplot of cohort WMH volumes for segmentations of twelve scans by two observers. Right - Scatterplot of mask volumes by the two different observers. The Spearman correlation was r = 0.78.

4.3.2.3 Unidimensional reproducibility

The (Spearman) correlation between WMH volumes was r = 0.78 and the intra-class correlation was 0.66 (95% CI: -0.09, 0.91). However, the high correlation between WMH volumes concealed a large discrepancy between the two raters for two cases, in both of which one rater had identified low or very low ($<$ 1ml) volumes of visible disease. A Bland-Altman ratio plot is shown in Figure 4.6, with the most extreme outlier removed (mean WMH volume = 14ml, ratio between observer volumes = 147).

Similar to the findings for intra-observer agreement (see Figure 4.3), the potential for large discrepancies between segmentations appeared greatest at lower WMH volumes.

4.3.2.4 Spatial agreement

The mean of the Dice indices for spatial overlap between different observers' segmentations was 0.54. A plot of Dice indices against mean WMH volume is shown in Figure 4.7. As with the volume ratios, improved agreement, as measured by the Dice index, was seen at larger WMH volumes.

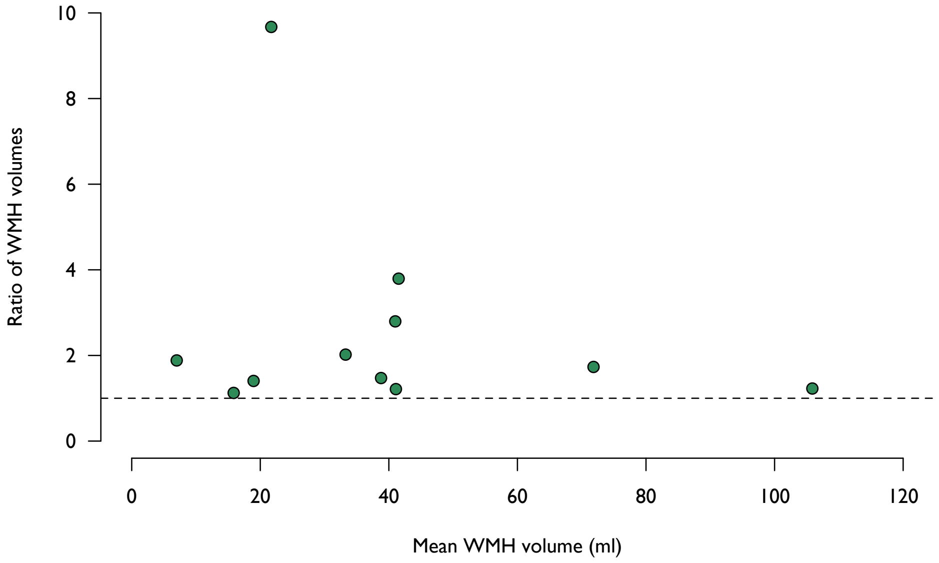


Figure 4.6: Bland-Altman plot, showing ratio of WMH volumes for each scan comparing that of the second observer to that of the first ($n = 11$). One extreme outlier with a volume ratio of 147 has been omitted. The dashed line indicates a ratio of 1 and all ratios were above this.

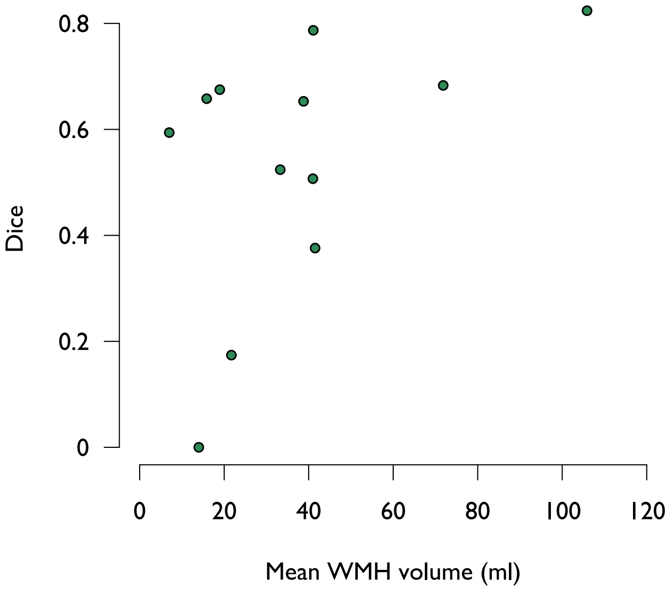


Figure 4.7: Scatterplot of Dice index of inter-observer spatial agreement against mean WMH volume (n = 12). The overall mean Dice index was 0.54

4.3.2.5 Sources of discrepancy

As for intra-observer agreement, segmentation discrepancies again occurred at sites where both partial volume effects and WMHs were frequent, particularly around the lateral ventricles and involving the corpus callosum.

A second source of discrepancy related to a bias in individual observer ‘thresholds’ for considering white matter abnormal. For the majority of scans this resulted in the second observer outlining similar regions to the first observer but with wider boundaries, contributing to the overall marked difference in summary WMH volumes. In two cases there was marked disagreement on whether there was widespread diffuse involvement of the white matter. In both these cases the initial segmentations only covered a small volume of more clearly demarcated focal WMHs.

4.4 Discussion

Before the validity of other methods for disease quantification can be tested, the reliability of the reference data they will be compared with should itself be established. With only rare availability of pathological samples, expert opinion on imaging appearances, in the form of manual segmentation, has become accepted as the reference standard. However the reliability of this reference standard is often ignored.

The results presented here demonstrate that the reference standard is imperfect, and its reliability is not constant, depending on both observer and cohort factors. There is an effect of observer experience and unconscious bias, with the potential for substantial error. Small differences in the subjective threshold used can make large differences to the overall output, particularly in decisions on how much intermediate signal white matter to include within segmentations of abnormal tissue.

An overall shift towards including more of the intermediate signal white matter in the tissue segmentation was apparent in the second mask set. Following completion of the initial masks, work had begun on optimising a software method for WMH segmentation using the same cohort, which tended to include more diffuse white matter changes. It is assumed that awareness of this influenced subsequent manual segmentation. This highlights the significant effect of observer biases, even when blinding is apparently complete. Practice effects would also likely have affected the work of the second observer, although time and resources did not provide the opportunity to confirm this.

The comparison between two observers with similar experience following the same protocol further highlights the difficulty of drawing sharp boundaries on images of diffuse disease processes. For the majority of scans, small local differences in marking edges added up to large total differences in WMH volumes. For a small number of cases, inter-observer disagreement on whether to mark diffuse regions of mildly raised signal led to very marked discrepancies in WMH volume. Even with these discrepancies in determining the absolute WMH volume, the reasonable inter-observer correlation (r = 0.78) indicates that agreement in distinguishing between different levels of disease was less affected.

The reliability of a method applies only to the particular population in which it has been tested, although this can clearly be used to make assumptions about its performance in similar situations. How far all populations of people with MS can be considered similar is debatable and even in this clinically relatively homogenous sample of people the reliability of the segmentation was clearly dependent on disease burden.

That the Dice index increases with WMH volume is a recognised limitation. Although widely used in the imaging community, the Dice index was originally developed for an entirely separate and not obviously relevant purpose, measuring ecological associations between species [108]. Numerous alternatives to the Dice have been developed (see Chapter 6).

Acknowledging the disadvantages of reliance on any one metric of reproducibility, the principle that reproducibility should be tested and reported remains highly important for standardisation of imaging research practice. Gains from research studies will be maximised only when the optimal methods for that population are used and their limitations understood.

It may be that the optimal method for WMH segmentation varies dependent on the research question. Subtle and diffuse abnormality in the white matter may be highly relevant for understanding the role of global white matter integrity in clinical outcomes. However its inclusion in disease measures when investigating the effect of treatments targeting acute inflammation pathways may be less relevant. When investigating differences in advanced imaging markers, such as those from DTI and MRS, between lesional and normal-appearing white matter, how these tissues are delineated will clearly alter what is found.

The reference standard quantification method is imperfect and this should be taken into account in presenting work reliant on it, including both validation of alternative imaging tools to quantify WMH burden and research exploring the relationship to cognitive performance. Errors associated with measurement tools will attenuate correlations derived using their outcomes [103]. Accepting this, the evidence here is interpreted to show experience resulted in a more stable definition of normal and abnormal tissue, and as such the final set of masks was chosen for later use in comparison with other methods.

Chapter 5

Development of a visual rating scale for MS imaging features

5.1 Introduction

The integration of several different and complementary sequences is universal in clinical and research imaging protocols and is particularly useful in diseases with complex and variable appearances. However, the measured outcomes in multiple sclerosis (MS) are often reduced to simple binary or scalar measures, such as stable or progressive disease, lesion or tissue volume, and information is lost.

Advanced imaging techniques partly address this issue, providing quantitative markers related to tissue damage, or focussing on anatomical structures of interest, such as the involvement of cortical grey matter. These can offer useful and objective measurement tools, but require standardisation and validation and may not be widely or routinely available. Even simple volumetric measures from routine imaging sequences, such as white matter hyperintensity (WMH) volume are not yet widely available or standardised (see Chapters 3 and 4). In other conditions, visual rating systems are commonly used to assess disease status, offering robust markers which can more easily be translated between scanners and centres and also to clinical practice, without the need for additional software. With the evolution of MS treatments, there is an unmet need for scalable and practical tools for quantification of MS imaging features that complement existing radiology reporting systems.

Imaging appearances reflect a complex interaction of disease and host, as well as potentially treatments; visible pathology may not accumulate in a straightforward or predictable manner. The pathological non-specificity of WMHs is well-recognised [44], encompassing acute inflammatory lesions and partly or fully remyelinated lesions, as well as permanent tissue damage, so limiting the utility of any single measurement. The balance between inflammation, repair and neurodegeneration is not necessarily the same in all people or at all disease stages.

In this chapter a novel semi-quantitative visual rating scale is developed for application to routine structural brain imaging in people with MS, the aim being to maximise efficient capture of information regarding different aspects of visible pathology, the degree of damage and structures involved. A series of imaging features of potential relevance to cognitive function are considered and three stages in the development of a rating system to assess these are described.

At each stage of the rating system development the frequency of feature presence is recorded, both to aid interpretation of reliability measures and to better characterise the range of disease appearances. The homogeneity of rated items is assessed for evidence of different dimensions within the data and potential item redundancy. Critical to its use as a research and imaging tool, measures of the reliability of individual items and summary scores are presented. Finally, the relationship of visual rating assessments of lesion volume to volumetric measures is considered.

5.2 Methods

5.2.1 Initial development and pilot study

5.2.1.1 Development process - design and item selection

Brain imaging features of potential relevance to disease severity and cognitive impairment were considered in consensus discussion by three consultant neuroradiologists (DM, RS, JW) with academic interests in white matter disease imaging. Existing disease-specific scales, identified in the systematic review reported in Chapter 3, were considered for relevance, responsiveness and practicality. Scales already in use for particular imaging features of interest were also considered and illustrative images were assessed. Outcomes of this process are reported in Section 5.3.1.

5.2.1.2 Initial pilot study

The first ten consecutive scans were selected from the 'Cognition in MS' study (see Chapter 2, Section 2.2 for further details). Three neuroradiologists completed a structured rating proforma, reviewing all scans using PACS (picture archiving and communication system) viewing software (Carestream manufacturer). One radiologist (DM) repeated all the ratings following a four week interval, with their initial set of ratings being taken as the reference set. All raters were blinded to the other assessments and all clinical information.

5.2.1.3 Statistical analysis

Item endorsement rates, indicating the frequency of feature presence, were calculated as the proportion of non-zero ratings assigned for each item.

Overall scale homogeneity was evaluated using item-(partial-)total correlations, split-half reliabilities and Cronbach's α, using data from the reference rater. Reliability for individual items was assessed using intra-class correlations (ICCs; Class 2), equivalent to a weighted kappa [167], comparing the three independent raters, with separate examination of the single rater repeat data. Six possible dimensions within the scale - white matter lesions, the presence of juxtacortical and cortical lesions, lesion cavitation, atrophy and enlarged perivascular spaces (EPVS) - were used to create subscores by summing all individual item scores within them. Intra- and inter-rater reliability for these subscores was assessed using ICCs.

A volumetric measure of lesion volume was also available, having been previously generated for the original study using a semi-automated software [168]. Spearman correlations were used for an exploratory comparison between this data and the mean white matter lesion dimension subscore averaged across the three raters.

Homogeneity statistics were calculated using the R software 'psychometric' and 'multicon' packages; ICCs were calculated using the R 'psych' package.

5.2.2 Further development and re-evaluation

5.2.2.1 Item refinement

The results of the initial pilot study were reviewed by the same three consultant neuroradiologists, considering item endorsement rates and reliability, rater agreement and overall practicality for data collection and analysis.

5.2.2.2 Further pilot

Twelve scans were pseudo-randomly selected from the ‘Cognition in MS’ study, to ensure an equal spread across quartiles of lesion volume (as determined by the available software measurement) and reasonable ratios of sex and clinical phenotype. Seven neuroradiologists (five consultants (DM, RS, JW, GM, ZM) and two senior trainees (MR, LG), post-fellowship examination) were recruited, with individual training prior to reviewing and rating all scans using the Carestream viewing software. One radiologist (DM) repeated all ratings, following an interval of four weeks, with their initial set of ratings being taken as the reference set. All raters were blind to other assessments and all clinical information.

5.2.2.3 Statistical analysis

Analysis of results was carried out as for the initial pilot. The initial ratings of the one radiologist with repeat data were designated the reference standard where necessary for comparison. The hierarchical arrangement of the items in the revised scale allowed the creation of an additional summary subscore for global (‘Fazekas-style’) white matter ratings [169] and the replacement of separate juxtacortical and cortical lesion subscores, with a combined score. Both cavitation and juxtacortical/cortical lesions were considered both as binary (present/absent) features for each region and also by their total number. Systematic between-rater biases were explored using the dimension subscores and agreement on definitions of cavitated lesions and juxtacortical/cortical lesions were examined using correlations of counts of the total numbers identified with those of the reference rater.

As in the initial pilot study, a volumetric measure of lesion volume was available and the relationship to this of the global ratings and the summed regional lesion score was assessed graphically and through correlations.

5.2.3 External validation study

5.2.3.1 Power calculation

Following review of the results from the pilot studies, testing of the scale in a larger study and its relationship to cognitive status was planned. With the assumption that a correlation of $r = -0.35$ with cognitive performance would be of interest (see Chapter 3, for a review of relevant literature), a power calculation was performed, using the G*Power software (version 3.1), indicating 61 subjects would be needed for a 0.8 probability of finding a significant result at a significance level of $p < 0.05$.

5.2.3.2 Study design and participants

Use of the visual rating scale was assessed in two separate cohorts of people with MS, predicted to have different imaging features. Sample sizes for each were chosen to meet or exceed the number suggested by the power calculation described above.

Sixty-seven baseline scans from participants with early stage relapsing-remitting MS (RRMS) in the FutureMS study (see Chapter 2, Section 2.3) were available at the time of this work, representing people with early stage disease. Baseline imaging for the University of Edinburgh MS-SMART participants with secondary progressive MS (SPMS, $n = 93$, see Chapter 2, Section 2.1) was complete, representing participants with more advanced disease.

The scans from these two cohorts were reviewed separately, using Mango image viewing software (http://ric.uthscsa.edu/mango/), by a single rater (DM) blind to all clinical and demographic information, other than study participation. All ratings were repeated following an interval of at least four weeks.

5.2.3.3 Statistical analysis

Separate analyses were carried out for the two different cohorts. Assessment of item endorsement rates, scale homogeneity and subscore reliability was performed as for the previous pilot studies.

Manual lesion segmentation (see Chapter 4) was also available for a subset of the MS-SMART cohort. Its relationship to the global white matter ratings and summed regional lesion scores was assessed graphically and through correlations.

5.3 Results of phase I: Initial development and piloting

5.3.1 Item selection

In the process of consensus development of an initial visual rating system, initial discussion and review of the literature identified white matter lesions and atrophy as the aspects of MS-related imaging changes of most interest to the academic and clinical imaging communities [9, 170]. No suitable pre-existing MS-specific imaging rating scale was identified. However analogous scales existed for other conditions, such as the 'Fazekas scale' [169] for age-related white matter changes, which is well-established and has been extensively tested.

Spatial localisation of white matter lesions, including laterality, was agreed to be valuable data to collect, of potential relevance to cognitive impairment. Structural brain subdivisions were agreed, with consideration to familiarity and practicality. The potential for disagreement over lobar divisions was discussed and brief guidance considered useful to ameliorate this.

Visible markers of the degree of damage, such as lesion cavitation, were identified to be an aspect of imaging appearances where visual assessment could add information to software-generated quantitative measures.

Current research interest in cortical and juxtacortical lesions was identified, confirmed by their use in the most recently published diagnostic guidelines [9]. It was recognised that routine imaging sequences may not always demonstrate these features, but felt that where visible they should be recorded.

Enlarged perivascular spaces, with their relationship to atrophy and potential relevance to cognition [171] and inflammation [45], were identified as a feature of interest. The existence of a validated rating scale [172] was recognised, albeit developed in a vascular disease context, with no adaptations felt necessary.

Where possible, having all items rated on the same scale, 0 to 3, with uniform directionality, was considered advantageous. Sample images being available at the time of rating was also identified to be beneficial, where possible without making the rating form unwieldy.

Incorporating the features described above, a data collection form was drafted, sample images reviewed and agreed. See Appendix E for a copy of the structured data collection form used in the initial pilot study.

5.3.2 Participant characteristics

The ten scans rated were from six female and four male subjects, with mean age 44.5 ± 5.3 years. The disease phenotype was RRMS in two subjects, SPMS in two and primary progressive MS (PPMS) in six. A semi-automated lesion segmentation software tool had previously been used in this cohort, and from this the median lesion volume was 8.5ml (interquartile range (IQR): 6.5, 17.9).

5.3.3 Item endorsement rates

The rating scale included 68 individual items, eleven of which were not endorsed (i.e. given a non-zero rating) in any rater-scan trial. These were one region for WMHs, nine cavitation regions and one cortical lesion region. A histogram of endorsement rates for all items is shown in Figure 5.1, showing that the majority of items were endorsed in fewer than 40% of cases. Full summary statistics for all items, including individual item endorsement rates, are shown in Table 5.1.

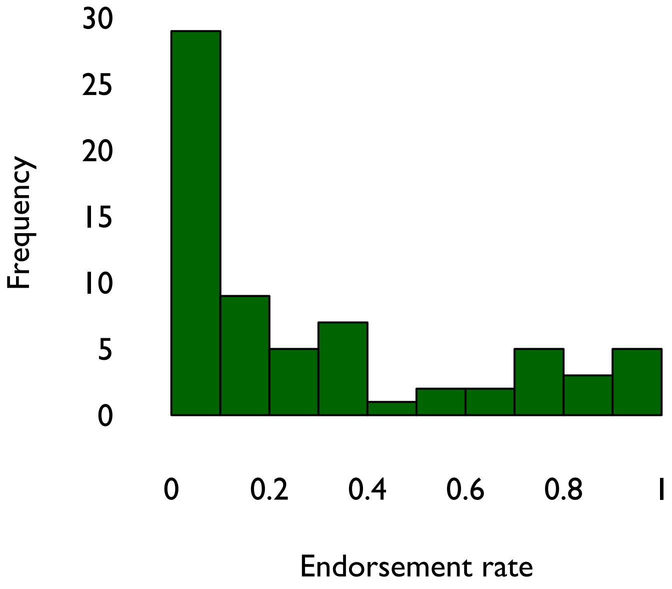


Figure 5.1: Histogram of item endorsement rates in initial pilot study, where endorsement indicates any non-zero rating. The total number of items rated was 68.

	Item	Endorsement rate	Mean	SD	Item-total correlation	Inter-rater ICC	Intra-rater ICC
Regional WMH	Frontal (R)	0.90	1.40	0.81	0.71	0.52	0.68
	Frontal (L)	0.97	1.43	0.73	0.64	0.40	0.64
	Parietal (R)	0.63	0.90	0.92	0.53	0.29	0.84
	Parietal (L)	0.77	1.00	0.79	0.47	0.27	0.72
	Temporal (R)	0.30	0.37	0.67	0.55	0.57	0.74
	Temporal (L)	0.40	0.47	0.68	0.51	0.52	0.31
	Occipital (R)	0.53	0.70	0.84	0.37	0.00	0.45
	Occipital (L)	0.50	0.73	0.91	0.13	0.27	0.79
	Insular (R)	0.07	0.10	0.40	-	0.40	0.00
	Insular (L)	0.00	0.00	0.00	-	-	0.00
	Periventricular (R)	1.00	1.63	0.89	0.45	0.47	0.21
	Periventricular (L)	0.97	1.63	0.93	0.45	0.51	0.35
	Corpus callosum	0.87	1.33	0.80	0.38	0.16	0.54
	Basal ganglia (R)	0.30	0.33	0.55	0.16	0.23	0.80
	Basal ganglia (L)	0.03	0.07	0.37	-	0.00	0.00
	Brainstem	0.37	0.50	0.78	0.46	0.30	0.05
	Cerebellar peduncles (R)	0.13	0.20	0.61	-	0.21	0.00
	Cerebellar peduncles (L)	0.33	0.40	0.67	0.33	0.65	0.70
	Cerebellar hemispheres (R)	0.10	0.13	0.43	-	0.13	0.00
	Cerebellar hemispheres (L)	0.13	0.17	0.46	-0.16	0.39	0.64
Cavitation	Frontal (R)	0.27	0.40	0.77	-0.03	0.16	0.64
	Frontal (L)	0.20	0.30	0.70	-	0.15	0.00
	Parietal (R)	0.10	0.10	0.31	-0.03	-0.08	0.64
	Parietal (L)	0.10	0.13	0.43	-	0.00	0.00
	Temporal (R)	0.03	0.03	0.18	-	0.00	-
	Temporal (L)	0.00	0.00	0.00	-	-	-
	Occipital (R)	0.07	0.07	0.25	-	0.00	-
	Occipital (L)	0.03	0.03	0.18	-	0.00	0.00
	Insular (R)	0.00	0.00	0.00	-	-	-
	Insular (L)	0.00	0.00	0.00	-	-	-
	Periventricular (R)	0.17	0.20	0.48	0.10	0.30	-0.17
	Periventricular (L)	0.10	0.10	0.31	0.10	-0.13	0.00
	Corpus callosum	0.03	0.03	0.18	-	0.00	-
	Basal ganglia (R)	0.00	0.00	0.00	-	-	-
	Basal ganglia (L)	0.00	0.00	0.00	-	-	-
	Brainstem	0.07	0.07	0.25	-	0.00	-
	Cerebellar peduncles (R)	0.00	0.00	0.00	-	-	-
	Cerebellar peduncles (L)	0.00	0.00	0.00	-	-	-
	Cerebellar hemispheres (R)	0.00	0.00	0.00	-	-	-
	Cerebellar hemispheres (L)	0.00	0.00	0.00	-	-	-
Juxta-cortical lesions	Frontal (R)	0.73	1.37	1.10	0.63	0.69	0.77
	Frontal (L)	0.80	1.17	0.95	0.78	0.51	0.87
	Parietal (R)	0.27	0.40	0.81	0.80	0.61	0.85
	Parietal (L)	0.37	0.53	0.86	0.87	0.62	0.89
	Temporal (R)	0.33	0.40	0.67	0.30	0.29	0.57
	Temporal (L)	0.20	0.27	0.64	0.45	0.29	0.84
	Occipital (R)	0.13	0.17	0.46	-	0.24	-
	Occipital (L)	0.20	0.23	0.50	0.82	0.52	0.64
	Insular (R)	0.27	0.30	0.53	0.34	0.12	0.37
	Insular (L)	0.07	0.10	0.40	0.82	0.40	1.00
Cortical lesions	Frontal (R)	0.33	0.43	0.68	-0.14	0.38	-0.17
	Frontal (L)	0.17	0.20	0.48	-	0.33	-
	Parietal (R)	0.10	0.13	0.43	-	-0.06	-
	Parietal (L)	0.13	0.17	0.46	-0.16	0.05	0.00
	Temporal (R)	0.03	0.03	0.18	-	0.00	-
	Temporal (L)	0.03	0.03	0.18	-	0.00	-
	Occipital (R)	0.10	0.10	0.31	-	0.00	-
	Occipital (L)	0.00	0.00	0.00	-	-	-
	Insular (R)	0.03	0.03	0.18	-	0.00	-
	Insular (L)	0.03	0.03	0.18	-	0.00	-
Atrophy	Deep	0.70	1.03	0.96	0.86	0.76	0.83
	Superficial	0.73	1.13	0.94	0.89	0.67	0.79
	Corpus callosum	0.60	0.87	0.82	0.70	0.62	0.77
	Posterior fossa	0.40	0.57	0.77	0.50	0.08	0.00
Enlarged perivascular spaces	Basal ganglia (R)	0.97	1.20	0.48	-	0.18	0.00
	Basal ganglia (L)	0.97	1.23	0.50	-	0.11	0.00
	Centrum semiovale (R)	0.80	1.57	1.17	-0.21	0.29	0.70
	Centrum semiovale (L)	0.83	1.57	1.14	-0.16	0.34	0.65

Table 5.1: Descriptive statistics for each individual item in visual rating scale (initial pilot study). Endorsement rate: proportion of non-zero ratings; SD: Standard deviation.

5.3.4 Scale homogeneity

Cronbach's α was $0.88\,(0.74, 0.96)$, with a split-half reliability of 0.92 ± 0.29. This indicates a high degree of homogeneity in the items rated, but is also related to the large number of individual items assessed. Certain items could be considered redundant if their sole value was in contributing to an overall score.

It was not possible to calculate item-(partial-)total correlations for 32 items in this sample due to either no non-zero ratings ($n = 30$), or no variance in the ratings ($n = 2$). Where available, these correlations ranged from -0.21 to 0.89, with a mean of 0.40. This does not provide any evidence of items with variation in an opposing direction to that of the full scale.

Eleven items had item-total correlations of < 0.2. These were three regions for white matter lesions, 4 cavitation regions, two cortical lesion regions and two EPVS regions. This could be partly explained by infrequent endorsement, as five of these items had endorsement rates < 0.2.

Item-total correlations for all individual items, where available, are shown in Table 5.1.

5.3.5 Reproducibility

5.3.5.1 Individual items

Intra-class correlations for all individual items are included in Table 5.1. The intra-rater ICCs for individual items ranged from -0.17 to 1, with mean 0.76 and median 0.43. In comparison, the inter-rater ICCs ranged from -0.13 to 0.76, with mean 0.26 and median 0.27. All cases of negative inter-rater ICCs were related to items with very low endorsement rates (< 0.2). Similarly the two items with negative intra-rater ICCs both had endorsement rates of 0.15 for the reference rater.

5.3.5.2 Dimension subscores

Intra-class correlations for dimension subscores, created by summing all ratings in each of six classes, are shown in Table 5.2. These dimension subscores focus on the rater reliability in identifying and scoring certain imaging features, removing any effect from disagreement over anatomical boundaries and lessening the effect of low endorsement rates for individual regional items.

Although likely still influenced by endorsement rates, Table 5.2 does suggest that inter-rater and intra-rater reliability varies for assessment of different imaging features, with reliability being higher for assessment of white matter lesions, juxtacortical lesions and atrophy, compared with cavitation, cortical lesions and EPVS.

The statistical significance of the association between subscores for different raters, as indicated by the p-value, confirms that the negative ICC for intra-rater cortical lesion rating is a non-significant result, again related to low endorsement rates. There was an overall endorsement rate of 0.02 for cortical lesions ratings by the reference rater.

	Inter-rater ICC (p)	Intra-rater ICC (p)
White matter lesions	0.53 (< 0.01)	0.83 (< 0.01)
Cavitation	0.12 (0.18)	0.56 (0.03)
Juxta-cortical lesions	0.69 (< 0.01)	0.94 (< 0.01)
Cortical lesions	0.17 (0.17)	-0.20 (0.72)
Atrophy	0.69 (< 0.01)	0.91 (< 0.01)
EPVS	0.33 (0.01)	0.46 (0.02)

Table 5.2: Showing reliability, assessed with intra-class correlations (ICCs), of summary scores for each of six subtotals within the scale.

5.3.6 Validation with pre-existing semi-automated lesion volume

From previous work using a semi-automated lesion segmentation software, white matter lesion volumes for the scans were available. The Spearman correlation between the mean of the white matter lesion dimension subscore for the three raters and the volumetric measurement was r = 0.69; a scatterplot of results is shown in Figure 5.2. Excluding the single highest value, this correlation was r = 0.57. The relationship appeared plausibly linear when considering the full range, but with most subjects grouped at the lower end of lesion volumes/scores and not clearly separated out by different scores.

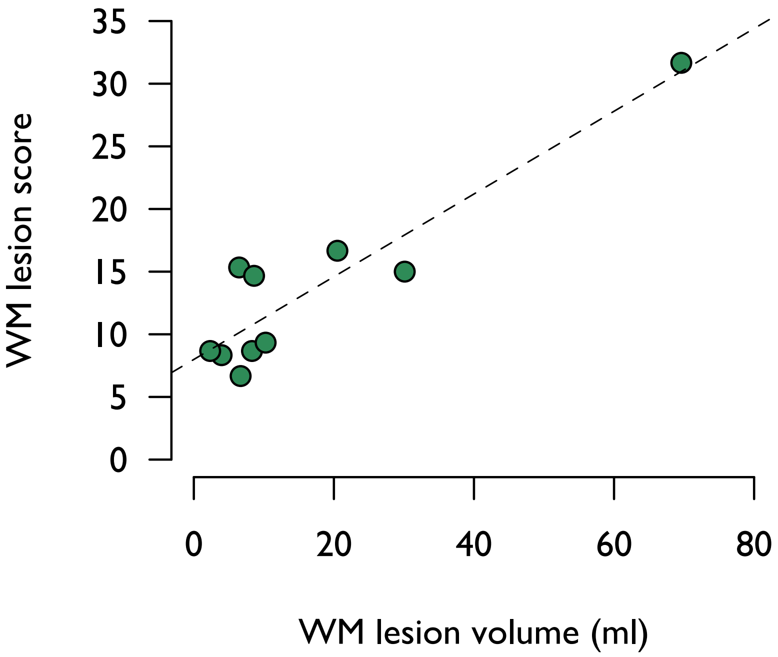


Figure 5.2: Scatterplot showing the mean of the white matter (WM) lesion dimension subscore in the initial pilot study, against the semi-automated lesion volume, annotated with a line of best fit. The Spearman correlation was $r = 0.69$.

5.4 Results of phase II: Further development and re-evaluation

5.4.1 Item refinement

Following review of the results of the initial pilot, modifications to the rating scale were agreed in consensus. Consideration was given to improving the clarity of item definitions, potential item redundancy and overall practicality. Improved descriptions of some items was felt to be of benefit, with more images available for guidance, particularly for assessing lesion cavitation.

The sample of scans selected for the initial pilot study was not thought to be an optimal representation of the range of imaging appearances seen in MS. A high proportion of the scans were from people with primary progressive MS, which may be associated with lower brain lesion loads [173] and this may in part have led to the frequently low item endorsement rates. Accepting this, endorsement rates were not felt to be a rigid guide to item inclusion, as this was a small sample and uncommon features may remain relevant. Non-endorsed items were not felt to add significantly to the time taken for rating completion.

In relation to the overall intention of the rating scale for use in routine MRI sequences, without advanced cortical imaging, it was agreed that reliable cortical lesion identification was not a priority. Not differentiating cortical and juxtacortical lesions initially would be more practical, while retaining the option to more accurately anatomically localise any cortical/juxtacortical lesion if identified.

The lack of a global white matter rating was considered a limitation and an overall deep white matter rating, modified from the 'Fazekas' scale [169] was introduced. Overall a more hierarchical structure was thought to be optimal for both rating and analysis and the form was redrafted to facilitate this.

See Appendix F for a copy of the structured data collection form used in the second pilot study.

5.4.2 Participant characteristics

The twelve scans rated were from six female and six male subjects, with mean age 47.7 ± 8.0 years. The disease phenotype was relapsing-remitting MS in five subjects, secondary progressive MS in four and primary progressive MS in three. One subject overlapped with those studied in the initial pilot. Using the available semi-automated lesion volumes, as before, the median lesion volume for the scans used in the second pilot study was 20.3ml (IQR: 10.9, 34.8). This was higher than in the first pilot, with a greater spread of values.

5.4.3 Item endorsement rates

There were 60 items available for rating in the second pilot study. Of these, two items (cavitation in the basal ganglia bilaterally) had zero endorsements in any rater-scan trial. A histogram of endorsement rates for all items is shown in Figure 5.3, suggesting a trimodal distribution, with items being endorsed in nearly all, nearly none, or around 40% of rater-scan trials. A greater dynamic range in terms of endorsement rates was achieved than in the initial pilot.

Summary statistics, including individual item endorsement rates are shown in Table 5.3.

	Item	Endorsement rate	Item-total correlation	Mean	SD	Inter-rater ICC	Intra-rater ICC
Global WM	Deep WM (R)	0.98	0.70	1.35	0.61	0.66	0.69
	Deep WM (L)	0.98	0.87	1.43	0.68	0.77	1.00
	Periventricular WM (R)	1.00	0.75	2.10	0.74	0.58	0.83
	Periventricular WM (L)	1.00	0.58	2.13	0.74	0.61	0.92
Regional WM	Frontal (R)	0.96	0.61	1.30	0.58	0.59	0.77
	Frontal (L)	0.98	0.64	1.32	0.60	0.65	0.59
	Parietal (R)	0.79	0.51	1.07	0.77	0.48	0.81
	Parietal (L)	0.77	0.72	1.07	0.80	0.55	0.52
	Temporal (R)	0.65	0.75	0.88	0.81	0.47	0.83
	Temporal (L)	0.58	0.54	0.79	0.82	0.40	0.70
	Occipital (R)	0.46	0.60	0.68	0.85	0.48	0.74
	Occipital (L)	0.48	0.62	0.71	0.87	0.31	0.34
	Insular (R)	0.48	0.61	0.58	0.73	0.28	0.29
	Insular (L)	0.33	0.66	0.48	0.78	0.59	0.80
	Corpus callosum	0.86	0.65	1.23	0.73	0.27	0.77
	Basal ganglia (R)	0.35	-0.23	0.38	0.56	0.49	0.52
	Basal ganglia (L)	0.27	0.00	0.35	0.61	0.45	0.09
	Brainstem	0.56	0.15	0.70	0.72	0.63	0.72
	Cerebellar peduncles (R)	0.43	0.25	0.44	0.52	0.25	0.75
	Cerebellar peduncles (L)	0.57	0.18	0.60	0.54	0.51	0.69
	Cerebellar hemispheres (R)	0.44	0.13	0.44	0.50	0.54	1.00
	Cerebellar hemispheres (L)	0.46	0.63	0.48	0.53	0.60	0.42
Cavitation	Periventricular WM (R)	0.37	0.48	0.68	1.09	0.45	0.33
	Periventricular WM (L)	0.42	0.75	0.95	1.40	0.51	0.52
	Frontal (R)	0.24	0.27	0.32	0.75	0.38	0.77
	Frontal (L)	0.30	0.86	0.50	0.90	0.55	0.81
	Parietal (R)	0.14	0.38	0.19	0.50	0.09	0.65
	Parietal (L)	0.12	-0.36	0.15	0.45	-0.04	0.00
	Temporal (R)	0.05	-	0.05	0.21	0.00	-
	Temporal (L)	0.04	-	0.04	0.19	0.10	-
	Occipital (R)	0.07	0.32	0.11	0.41	0.14	0.00
	Occipital (L)	0.06	-	0.07	0.30	-0.04	-
	Insular (R)	0.04	-	0.04	0.19	0.00	-
	Insular (L)	0.07	-	0.07	0.26	0.31	0.00
	Corpus callosum	0.06	-	0.08	0.35	-0.04	-
	Basal ganglia (R)	0.00	-	0.00	0.00	-	-
	Basal ganglia (L)	0.00	-	0.00	0.00	-	-
	Brainstem	0.05	-	0.10	0.51	0.00	-
	Cerebellar peduncles (R)	0.06	-	0.06	0.24	0.10	-
	Cerebellar peduncles (L)	0.06	-	0.07	0.30	0.03	-
	Cerebellar hemispheres (R)	0.04	-	0.04	0.19	-0.02	-
	Cerebellar hemispheres (L)	0.04	-	0.06	0.36	-0.02	-
(Juxta-) cortical lesions	Frontal (R)	0.85	0.42	3.10	3.69	0.54	0.33
	Frontal (L)	0.80	0.54	3.64	4.50	0.40	0.46
	Parietal (R)	0.51	0.64	1.01	1.44	0.32	0.84
	Parietal (L)	0.54	0.58	1.23	1.56	0.55	1.00
	Temporal (R)	0.50	0.79	0.77	1.01	0.31	0.66
	Temporal (L)	0.39	0.81	0.70	1.08	0.38	0.68
	Occipital (R)	0.31	0.59	0.45	0.77	0.35	0.45
	Occipital (L)	0.36	0.70	0.49	0.74	0.29	0.42
	Insular (R)	0.35	0.34	0.44	0.68	0.15	0.18
	Insular (L)	0.26	0.20	0.36	0.67	0.40	-0.22
Atrophy	Deep	0.94	0.55	1.67	0.83	0.59	0.88
	Superficial	0.89	-0.02	1.67	0.81	0.32	0.21
	Corpus callosum	0.86	0.52	1.44	0.88	0.54	0.77
	Posterior fossa	0.39	0.67	0.54	0.75	0.09	0.47
Enlarged perivascular spaces	Basal ganglia (R)	0.98	-0.39	1.15	0.48	0.07	-0.14
	Basal ganglia (L)	0.96	-0.35	1.18	0.56	0.18	0.30
	Centrum semiovale (R)	0.86	0.44	1.27	0.77	0.31	0.27
	Centrum semiovale (L)	0.86	0.13	1.32	0.87	0.31	0.75

Table 5.3: Descriptive statistics for each individual item in visual rating scale (second pilot study). Global and regional white matter (WM) ratings and atrophy were scored $0-3$. Cavitation and juxtacortical/cortical scores were counts. Enlarged perivascular spaces were rated $0-4$. Endorsement rate: proportion of non-zero ratings; SD: standard deviation. Dashed lines indicate undefined metric values due to no non-zero ratings.

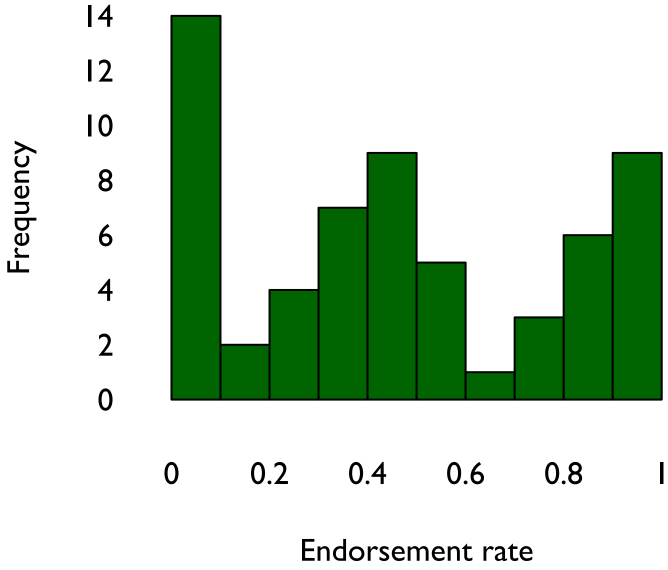


Figure 5.3: Histogram of item endorsement rates in second pilot study. The total number of items rated was 60.

5.4.4 Scale homogeneity

Cronbach's α was $0.92\,(0.84, 0.97)$, with a split-half reliability of 0.94 ± 0.24. As in the initial pilot study, this indicates a high degree of item homogeneity, but also reflects the large number of items rated.

It was not possible to calculate item-(partial-)total correlations for 13 items, all lesion cavitation regions, due to no non-zero ratings by the reference rater for these regions. Where available, item-total correlations ranged from -0.39 to 0.87, with a mean of 0.44. Six items had correlations with the full scale of between -0.2 and 0.2: four regional WMH items (1 for basal ganglia and three in the posterior fossa), 1 atrophy item and one EPVS item. Four items had correlations $r < -0.2$, raising the possibility of variation in an opposing direction to that of the full scale. These were one regional WMH item (basal ganglia), 1 cavitation item and two EPVS items. Item-total correlations for all individual items, where available, are shown in Table 5.3.

5.4.5 Reproducibility

5.4.5.1 Individual items

Intra-class correlations for all individual items are included in Table 5.3. The intra-rater ICCs for individual items ranged from -0.22 to 1, with mean 0.54 and median 0.65. Inter-rater ICCs ranged from -0.03 to 0.77, with mean 0.34 and median 0.37.

Very low inter-rater ICCs were found for all cavitation items, with the exception of the largest regions - periventricular white matter and the frontal lobes. There were very low overall endorsement rates for these items, which likely explains the five cavitation items with negative inter-rater ICCs (all with low absolute values). Intra-rater ICCs were undefined for many of these items due to no non-zero ratings by the reference rater.

Two items in other categories had negative intra-rater ICCs. In one case (a juxtacortical lesion item) this was related to a low endorsement rate by the reference rater, but in the other this was not (EPVS in the basal ganglia). All four EPVS items had high endorsement rates, but were mostly associated with low intra-rater and inter-rater ICCs.

Using the ICC as a measure of reliability combines assessment of two concepts - whether raters have the same understanding of a feature being present, e.g. lesion cavitation, and how they interpret the categories to assign different scores to the imaging appearances. Similarity of understanding of different features is assessed in Section 5.4.5.2 through the dimension subscores, as is rater bias towards using higher or lower scores.

Agreement on individual item scores was examined graphically using 'bubble' plots, providing a visual indication of how many raters agreed with the reference rating and each other. Bubble plots for the global white matter ('Fazekas-style') ratings for each scan are shown in Figure 5.4 and for the remaining scored items in Appendix G.

The ratings for deep white matter shown in the top row of Figure 5.4 demonstrate closer agreement, with perfect or near perfect agreement in most cases. There was greater variation in the scores assigned to the periventricular white matter, although the majority of raters were in agreement with the reference rater for all but 3/24 (= 13%) of ratings.

In the case of deep white matter ratings, Figure 5.4 highlights the narrow range of scores used. Scans rated are plotted in order of increasing lesion volume, but the first eight (= 67%) were all assigned scores of 1 bilaterally by the majority of raters. Although available, a score of zero was rarely (for deep white matter) or never (for periventricular white matter) used, thus further narrowing the range. The range of scores used will affect agreement, and an indication of this is also given by the standard deviation for each item, provided in Table 5.3.

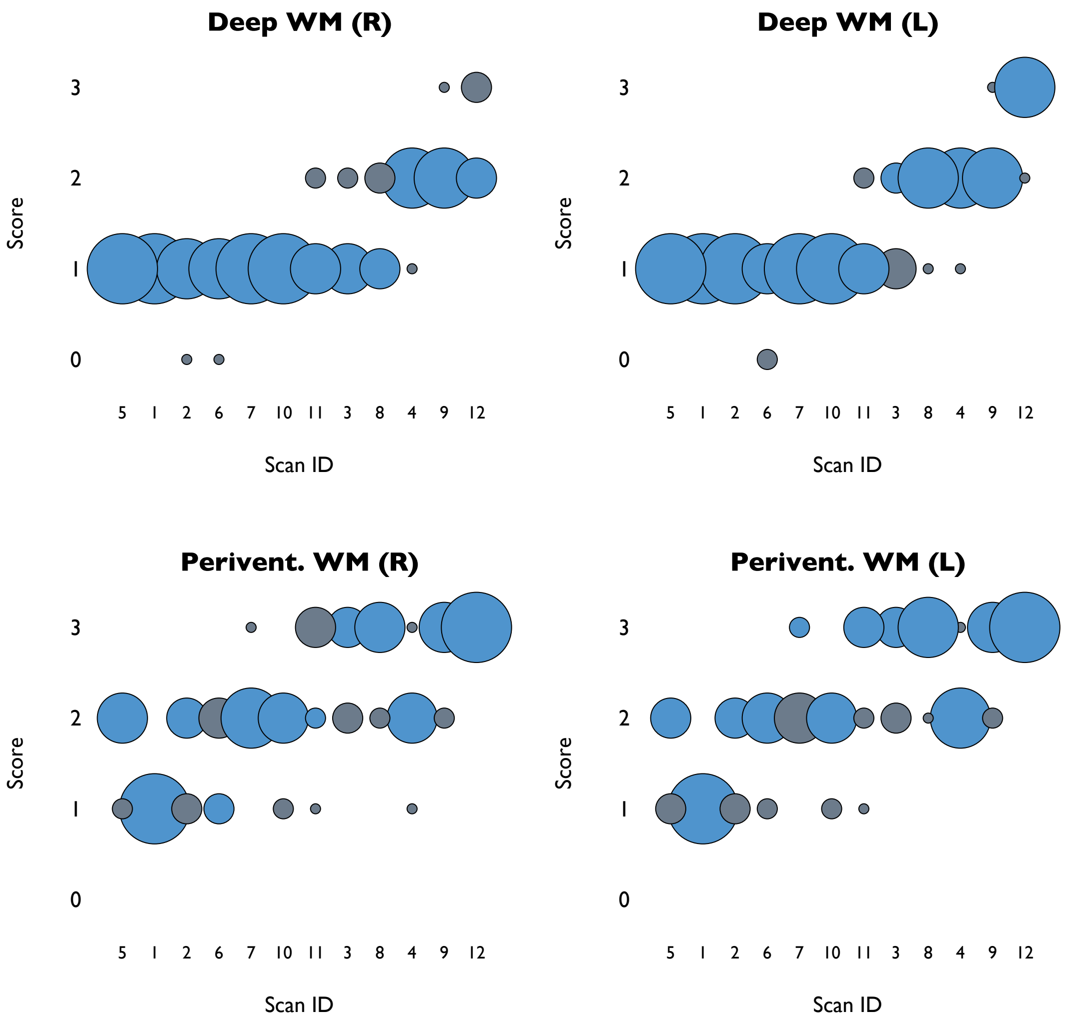


Figure 5.4: 'Bubble' plots of deep and periventricular white matter (WM) scores for each scan. The radius of each point is proportional to the number of raters assigning that score. Blue indicates agreement with the reference standard. The scans are plotted in order of increasing lesion volume.

5.4.5.2 Dimension subscores and imaging features of interest

The ICCs for dimension subscores, similar to those of the initial pilot study, are shown in Table 5.4. One rater had not provided counts of cavitated and juxtacortical lesions and was excluded from analysis related to these two subscores. Unlike the initial pilot study, all p-values are low enough to reject the null hypothesis of no association between rater scores.

	Inter-rater ICC (p)	Intra-rater ICC (p)
Global summary WM lesions	0.80 (< 0.001)	0.98 (< 0.001)
Regional WM lesions	0.64 (< 0.001)	0.91 (< 0.001)
Cavitation (regions)	0.29 (< 0.001)	0.69 (< 0.001)
Cavitation (counts)	0.17 (0.002)	0.62 (0.004)
(Juxta-)cortical lesions (regions)	0.53 (< 0.001)	0.87 (< 0.001)
(Juxta-)cortical lesions (counts)	0.51 (< 0.001)	0.93 (< 0.001)
Atrophy	0.55 (< 0.001)	0.72 (0.002)
EPVS	0.35 (< 0.001)	0.66 (0.002)

Table 5.4: Showing reliability, assessed with intra-class correlations (ICCs), of summary scores for each of eight subtotals within the scale. Cavitation and (juxta-)cortical lesion subscores were calculated both for the number of regions identified as affected as well as the total lesion count.

As in the initial pilot study, the dimension subscore ICCs suggest that reliability varies with the imaging feature of interest, again being lower for identifying cavitated lesions and rating EPVS. Poor reliability may relate to rater differences in defining the feature of interest as well as rater biases in using higher or lower scores.

Univariate correlations of the numbers of cavitated lesions identified by raters compared with the reference standard were examined to investigate similarity of underlying rater definitions. One rater had not provided counts of cavitated lesions and was excluded from this analysis. Of the remaining 5 raters, four showed total counts for each scan which strongly correlated with the reference standard (r $= 0.68, 0.89, 0.89, 0.94$) and one which only weakly correlated (r $=$ 0.26, rater D), suggesting this rater may have used different imaging appearances to define cavitation. Correlations for counts of fully cavitated lesions were slightly lower (r $=$ 0.25 to 0.78).

Similar correlations were examined for the total number of juxtacortical/cortical lesions identified. One rater had not provided counts of juxtacortical/cortical lesions and was excluded from more detailed analysis. Of the remaining 5 raters, all showed total lesion counts which correlated strongly with the reference standard (r = 0.88 to 0.95). Only two raters identified purely cortical lesions in any scan, so more detailed analysis of this was not performed.

Systematic biases between different raters were explored using barplots of the mean dimension subscore for each rater, shown in Figure 5.5. In the case of identifying cavitated lesions it is apparent that there is one outlying rater (rater D), who tended to identify much larger numbers. Excluding this rater in calculating the ICC for the cavitation (count) subscores, resulted in an increase to 0.64 ($p < 0.001$). Although less marked, the same rater also showed a tendency to identify more (juxta-)cortical lesions and excluding them in calculating the (juxta-)cortical (count) subscore ICC resulted in an increase to 0.61 ($p < 0.001$).

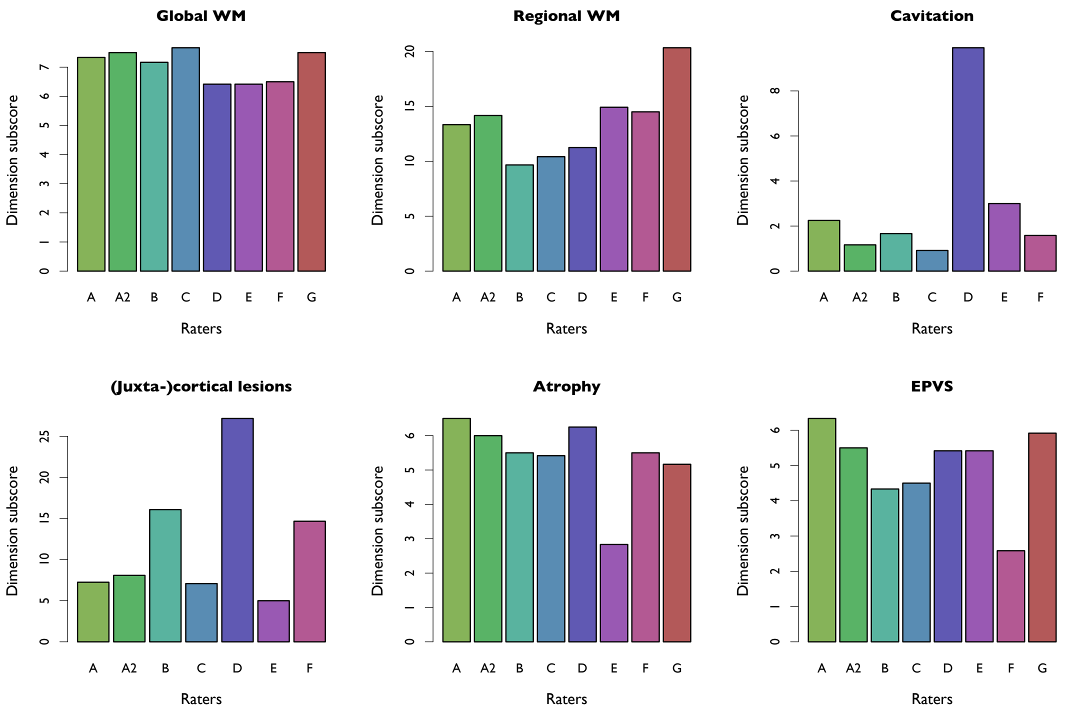


Figure 5.5: Barplots showing mean dimension subscores for each rater A - G (A2 is repeat rating by rater A). Cavitation and cortical/juxtacortical lesions are treated as binary features for each region (i.e. present or absent), so the mean represents the mean number of regions considered affected. Rater G did not provide counts of cavitated lesions or juxtacortical lesions for all scans.

5.4.6 Validation with pre-existing semi-automated lesion volume

Two visual markers of overall lesion burden were calculated: (1) the sum of bilateral deep and periventricular global white matter ratings and (2) the sum of all lesion scores for individually rated brain regions. Figure 5.6 shows the mean value from the seven raters for these two markers plotted against the available volumetric measure of lesion burden. The visual markers both correlated strongly with the volumetric measure and the relationship appeared plausibly linear across the whole range.

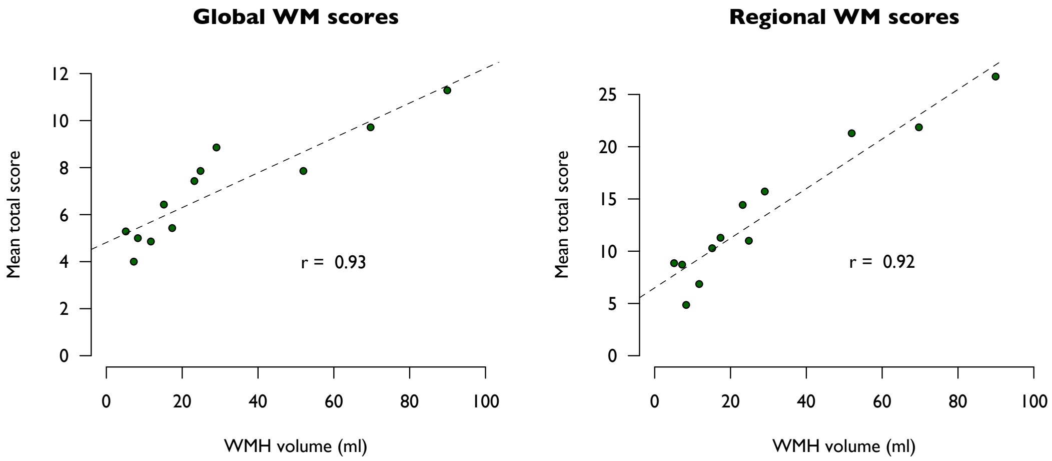


Figure 5.6: Scatterplots showing relationship of visual rating scores for white matter (WM) to semi-automated lesion (white matter hyperintensity - WMH) volume, annotated with lines of best fit and Spearman correlation coefficients. All scores plotted represent the mean value of seven raters for the sum of right and left hemisphere scores. Left: Summed deep and periventricular WM lesion scores; Right: Summed lesion scores for individual regions

5.5 Independent validation study

5.5.1 Participant characteristics

5.5.1.1 MS-SMART

One subject in the MS-SMART cohort was excluded, as a T2-weighted (T2w) imaging sequence was not available. Scans from the remaining 92 participants were included. These scans were from 68 female and 24 male participants, with a mean age of 55 ± 7.5 years. The mean disease duration, taken as the time since initial symptom, was 20 ± 10 years, median 20.5 years.

5.5.1.2 FutureMS

Sixty-seven participants had been recruited locally by the time of this work and their imaging was included. This covered a period of development of scan protocols. After the initial 23 participants, all scans included a 2D fluid attenuated inversion recovery (FLAIR) sequence, and after the initial 25 participants, all scans included a 2D T2w sequence, instead of a 3D T2w sequence. 3D T1-weighted and FLAIR sequences were available for all scans.

This cohort comprised 49 female and 18 male participants, with a mean age of 39 ± 9.6 years. The mean disease duration, taken as time since initial symptom, was 5.1 ± 5.5 years, median 2.7 years.

5.5.2 Scale homogeneity

Cronbach's α was $0.92\,(0.90, 0.94)$ with a split-half reliability of 0.94 ± 0.11. As before this reflects both the high degree of item homogeneity and the large number of items rated.

It was not possible to calculate item-(partial-)total correlations for 9 lesion cavitation items, due to no non-zero ratings in the relevant regions. Where available, item-total correlations ranged from 0.07 to 0.80, with a mean of 0.41 ± 0.21. A histogram of item-total correlations is shown in Figure 5.7. Thirteen items had correlations < 0.2, including 7 further cavitation items and all four EPVS items. Although this is likely influenced by the low endorsement rates for the cavitation items, this suggests that scores for these items may not vary with the majority of the other ratings.

5.5.3 Summary statistics for individual items

Summary statistics for each individual item, including the mean, standard deviation and endorsement rate, are given in Table 5.5. These are given both overall and separately for the two cohorts rated. Endorsement rates varied from 0 to 1, with an overall mean of 0.33. The mean endorsement rate was 0.36 for the cohort with SPMS (MS-SMART) compared with 0.28 for the cohort with RRMS (FutureMS).

		FutureMS				MS-SMART				Overall			
		End. rate	Mean	SD	ICC	End. rate	Mean	SD	ICC	End. rate	Mean	SD	ICC
Global WM	Deep WM (R)	0.75	0.91	0.65	0.73	0.90	1.40	0.76	0.80	0.84	1.19	0.75	0.80
	Deep WM (L)	0.82	1.03	0.65	0.72	0.95	1.48	0.72	0.73	0.89	1.29	0.72	0.76
	Perivent. WM (R)	0.84	1.36	0.88	0.69	0.91	2.03	0.94	0.85	0.88	1.75	0.97	0.82
	Perivent. WM (L)	0.85	1.49	0.88	0.77	0.93	2.05	0.89	0.85	0.90	1.82	0.93	0.84
Regional WM	Frontal (R)	0.72	0.85	0.66	0.78	0.91	1.37	0.74	0.78	0.83	1.15	0.75	0.81
	Frontal (L)	0.72	0.82	0.60	0.73	0.93	1.46	0.72	0.75	0.84	1.19	0.74	0.79
	Parietal (R)	0.60	0.81	0.82	0.73	0.70	1.25	1.02	0.53	0.65	1.06	0.97	0.63
	Parietal (L)	0.70	1.01	0.84	0.64	0.76	1.39	0.97	0.52	0.74	1.23	0.94	0.59
	Temporal (R)	0.36	0.43	0.63	0.80	0.39	0.46	0.64	0.58	0.38	0.45	0.63	0.67
	Temporal (L)	0.40	0.54	0.77	0.76	0.36	0.45	0.69	0.51	0.38	0.48	0.72	0.62
	Occipital (R)	0.12	0.16	0.48	0.85	0.15	0.22	0.57	0.52	0.14	0.19	0.53	0.61
	Occipital (L)	0.13	0.19	0.53	0.74	0.14	0.27	0.73	0.51	0.14	0.24	0.65	0.57
	Insular (R)	0.12	0.18	0.55	0.53	0.22	0.37	0.78	0.41	0.18	0.29	0.70	0.45
	Insular (L)	0.10	0.19	0.61	0.46	0.24	0.39	0.78	0.49	0.18	0.31	0.72	0.50
	Corpus callosum	0.75	0.94	0.69	0.60	0.39	0.42	0.58	0.57	0.54	0.64	0.68	0.62
	Basal ganglia (R)	0.37	0.40	0.55	0.29	0.46	0.54	0.65	0.60	0.42	0.48	0.61	0.40
	Basal ganglia (L)	0.30	0.30	0.46	0.55	0.51	0.63	0.69	0.61	0.42	0.49	0.63	0.61
	Brainstem	0.36	0.45	0.68	0.83	0.50	0.64	0.75	0.79	0.44	0.56	0.73	0.81
	Cereb. ped. (R)	0.18	0.22	0.52	0.80	0.24	0.32	0.63	0.80	0.21	0.28	0.58	0.80
	Cereb. ped. (L)	0.24	0.30	0.60	0.70	0.36	0.41	0.61	0.67	0.31	0.36	0.61	0.68
	Cereb. hemi. (R)	0.22	0.22	0.42	0.69	0.34	0.37	0.55	0.75	0.29	0.31	0.50	0.74
	Cereb. hemi.(L)	0.15	0.16	0.41	0.48	0.34	0.37	0.55	0.69	0.26	0.28	0.50	0.64
Cavitation	Perivent. WM (R)	0.10	0.27	0.95	0.59	0.49	1.60	2.82	0.85	0.33	1.04	2.32	0.81
	Perivent. WM (L)	0.12	0.28	0.98	0.68	0.49	1.66	2.73	0.77	0.33	1.08	2.27	0.79
	Frontal (R)	0.07	0.07	0.26	0.36	0.10	0.18	0.69	0.53	0.09	0.14	0.56	0.48
	Frontal (L)	0.01	0.01	0.12	-0.02	0.14	0.17	0.48	0.43	0.09	0.11	0.38	0.39
	Parietal (R)	0.04	0.04	0.21	0.31	0.04	0.04	0.21	0.79	0.04	0.04	0.21	0.61
	Parietal (L)	0.04	0.06	0.30	0.38	0.07	0.09	0.38	0.35	0.06	0.08	0.35	0.36
	Temporal (R)	0.00	0.00	0.00	-	0.01	0.01	0.10	-0.01	0.01	0.01	0.08	-0.01
	Temporal (L)	0.00	0.00	0.00	-	0.02	0.02	0.15	0.66	0.01	0.01	0.11	0.66
	Occipital (R)	0.00	0.00	0.00	-	0.00	0.00	0.00	-	0.00	0.00	0.00	-
	Occipital (L)	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Insular (R)	0.00	0.00	0.00	-	0.01	0.01	0.10	1.00	0.01	0.01	0.08	1.00
	Insular (L)	0.00	0.00	0.00	-	0.01	0.01	0.10	1.00	0.01	0.01	0.08	1.00
	Corpus callosum	0.00	0.00	0.00	-	0.02	0.02	0.15	0.66	0.01	0.01	0.11	0.67
	Basal ganglia (R)	0.00	0.00	0.00	-	0.00	0.00	0.00	-	0.00	0.00	0.00	-
	Basal ganglia (L)	0.00	0.00	0.00	-	0.00	0.00	0.00	-	0.00	0.00	0.00	-
	Brainstem	0.00	0.00	0.00	-	0.00	0.00	0.00	-	0.00	0.00	0.00	-
	Cerebel. ped. (R)	0.00	0.00	0.00	-	0.00	0.00	0.00	-	0.00	0.00	0.00	-
	Cerebel. ped. (L)	0.00	0.00	0.00	-	0.00	0.00	0.00	-	0.00	0.00	0.00	-
	Cerebel. hemi. (R)	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Cerebel. hemi. (L)	0.00	0.00	0.00	-	0.00	0.00	0.00	-	0.00	0.00	0.00	-
(Juxta-) cortical lesions	Frontal (R)	0.13	0.19	0.58	0.20	0.43	0.83	1.23	0.64	0.31	0.56	1.05	0.53
	Frontal (L)	0.18	0.21	0.48	0.60	0.45	0.75	1.23	0.55	0.33	0.52	1.02	0.59
	Parietal (R)	0.07	0.12	0.48	0.57	0.08	0.09	0.32	0.45	0.08	0.10	0.39	0.49
	Parietal (L)	0.13	0.16	0.45	0.67	0.18	0.21	0.48	0.65	0.16	0.19	0.47	0.66
	Temporal (R)	0.07	0.10	0.43	0.45	0.07	0.07	0.25	0.33	0.07	0.08	0.34	0.40
	Temporal (L)	0.13	0.21	0.66	0.67	0.17	0.20	0.47	0.35	0.16	0.20	0.56	0.46
	Occipital (R)	0.01	0.01	0.12	-0.02	0.04	0.05	0.27	0.58	0.03	0.04	0.22	0.48
	Occipital (L)	0.00	0.00	0.00	0.00	0.07	0.07	0.25	0.39	0.04	0.04	0.19	0.37
	Insular (R)	0.12	0.16	0.51	0.13	0.13	0.17	0.51	0.55	0.13	0.17	0.51	0.44
	Insular (L)	0.09	0.13	0.46	0.31	0.09	0.10	0.33	0.46	0.09	0.11	0.39	0.41
Atrophy	Deep	0.48	0.72	0.88	0.80	0.82	1.49	1.02	0.84	0.67	1.16	1.04	0.85
	Superficial	0.82	1.15	0.70	0.70	0.96	1.83	0.76	0.61	0.90	1.54	0.81	0.70
	Corpus callosum	0.40	0.58	0.80	0.66	0.73	1.25	1.00	0.78	0.59	0.97	0.98	0.75
	Posterior fossa	0.18	0.19	0.43	0.30	0.29	0.37	0.62	0.53	0.25	0.30	0.56	0.46
Enlarged peri- vascular spaces	Basal ganglia (R)	0.91	1.00	0.43	0.39	1.00	1.20	0.43	0.43	0.96	1.11	0.44	0.46
	Basal ganglia (L)	0.94	1.04	0.41	0.31	0.97	1.09	0.41	0.42	0.96	1.07	0.41	0.37
	Cent. semiovale (R)	0.99	1.76	0.85	0.55	0.99	2.08	0.74	0.40	0.99	1.94	0.81	0.49
	Cent. semiovale (L)	0.99	1.70	0.84	0.46	0.99	1.93	0.74	0.55	0.99	1.84	0.79	0.52

Table 5.5: Descriptive statistics for each individual item in visual rating scale (independent validation study). Global and regional white matter (WM) ratings and atrophy were scored $0-3$. Cavitation and juxtacortical/cortical scores were counts. Enlarged perivascular spaces were rated $0-4$. End. rate: proportion of non-zero ratings; SD: standard deviation; ICC: (intra-rater) intra-class correlation.

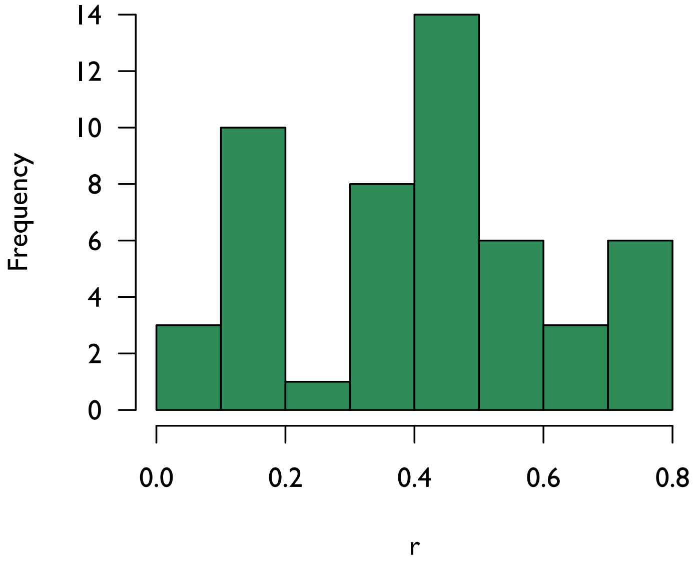


Figure 5.7: Histogram of item-(partial)-total correlations for the independent validation study.

5.5.4 Reliability

5.5.4.1 Individual items

Intra-rater ICCs for individual items are shown in Table 5.5. These ICCs had mean 0.58, median 0.61 and range −0.01 to 1. Table 5.5 shows that the lowest values and the greatest variability for item ICCs were found for the cavitation items, which also had the lowest endorsement rates.

5.5.4.2 Dimension subscores and features of interest

The ICCs for the dimension subscores are shown in Table 5.6. All p-values associated with the ICCs are < 0.00001, indicating that the null hypothesis of no association between repeat scores can be rejected with a high degree of probability. The intra-rater reliability appears notably lower for the EPVS ratings, with a similar value to that seen in the previous pilot study.

Overall agreement on whether lesion cavitation was present or absent on each scan is summarised in Table 5.7, showing that the initial rating identified at least

	FutureMS	MS-SMART	Overall
Global WMH	0.90	0.90	0.92
Regional WMH	0.94	0.79	0.86
Cavitation	0.74	0.85	0.86
(Juxta-)cortical lesions	0.83	0.71	0.75
Atrophy	0.78	0.87	0.86
EPVS	0.52	0.55	0.56

Table 5.6: Intra-rater ICCs for dimension subscores. MS-SMART: secondary progressive MS cohort (n = 92); FutureMS: Relapsing-remitting MS cohort (n = 67). The data for the cavitation and (juxta-)cortical lesion dimensions relates to the total counts.

one cavitated lesion on 45% of scans (72/159), with disagreement from the repeat rating in 9% (14/159) of cases.

		Repeat rating	
		No	Yes
Initial	No	78	9
	Yes	5	67

Table 5.7: Summary table showing whether cavitation was identified as being present on the initial and repeat scan ratings.

Overall agreement on whether there was involvement of juxtacortical or cortical tissue is summarised in Table 5.8, showing that the initial rating identified at least one (juxta-)cortical lesion on 59% of scans (= 94/159), with disagreement from the repeat rating in 20% (= 32/159) of cases.

		Repeat rating	
		No	Yes
Initial	No	44	21
	Yes	11	83

Table 5.8: Summary table showing whether (juxta-)cortical lesions were identified as being present on the initial and repeat scan ratings.

5.5.5 Cohort descriptive statistics

Sample images, showing scans from the MS-SMART cohort assigned low, intermediate and high global lesion scores, are shown in Figure 5.8. Summary statistics for each item are given broken down by cohort in Table 5.5 and for dimension subscores in Table 5.9, demonstrating significantly different distributions between cohorts for each subscore. Histograms comparing the two cohorts for the two summary WM subscores, equivalent to the first two dimension subscores, are shown in Figures 5.9 and 5.10. The histograms in Figure 5.11 show the sums of all items scored for the two cohorts. While there is a clear tendency for lower scores for scans from participants with earlier stage disease, there is still considerable overlap and disease stage cannot be distinguished from these scores alone.

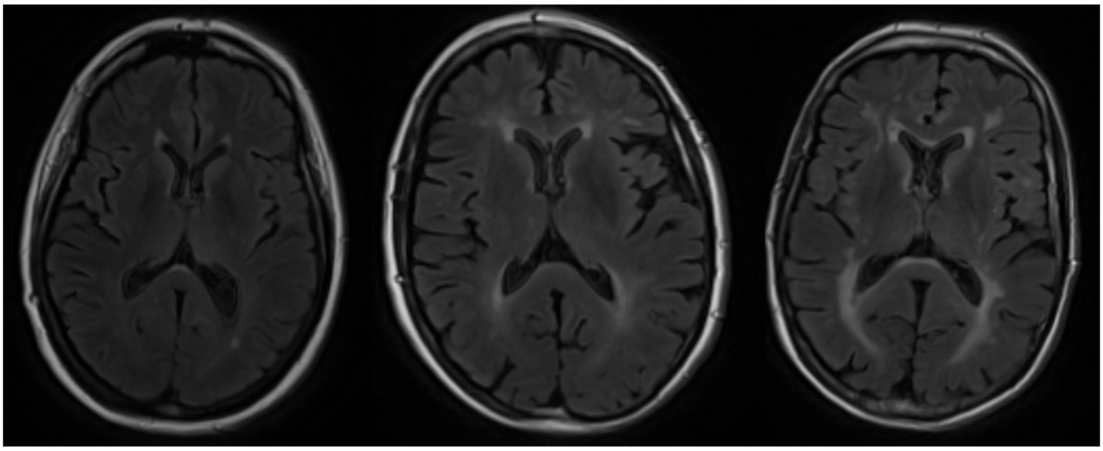


Figure 5.8: Sample images from MS-SMART cohort, demonstrating scans with (from left to right) low, intermediate and high global white matter scores.

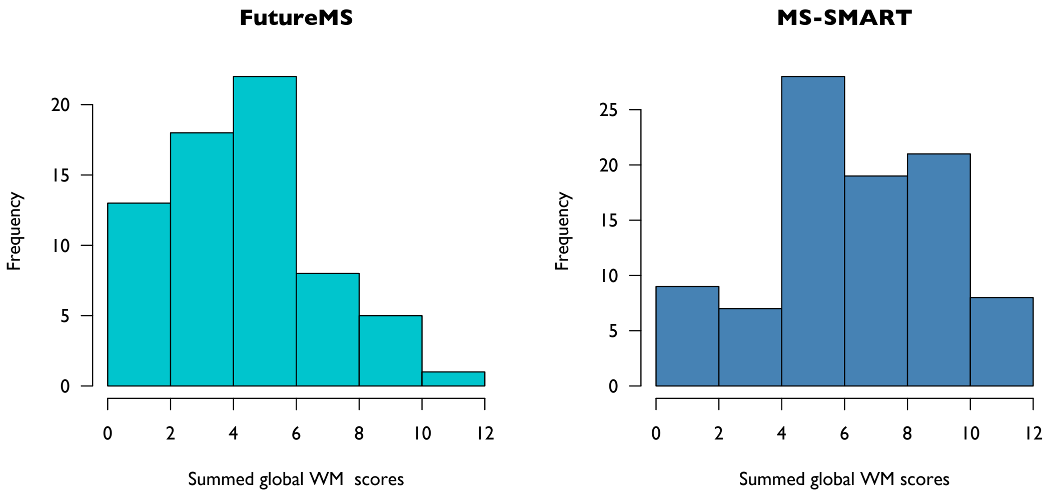


Figure 5.9: Histograms of summed global white matter scores assigned to scans belonging to each cohort.

	FutureMS		MS-SMART		Mann-Whitney test	
	Mean	SD	Mean	SD	U	p
Global WMH	4.8	2.5	7.0	2.9	4448	< 0.001
Regional WMH	8.2	6.7	11.3	7.0	3978	0.002
Cavitation (count)	0.7	1.9	3.8	5.9	4397	< 0.001
(Juxta-)cortical lesions (count)	1.3	2.8	2.5	3.3	3944	0.002
Atrophy	2.6	2.2	4.9	2.7	4592	< 0.001
EPVS	5.5	1.9	6.3	1.7	3870	0.005

Table 5.9: Summary of dimension subscores for the two cohorts studied. The data for the cavitation and (juxta-)cortical lesion dimensions relates to the total counts. SD: Standard deviation; EPVS: Enlarged perivascular spaces; U: Test-statistic from Mann-Whitney U test, with associated p-value.

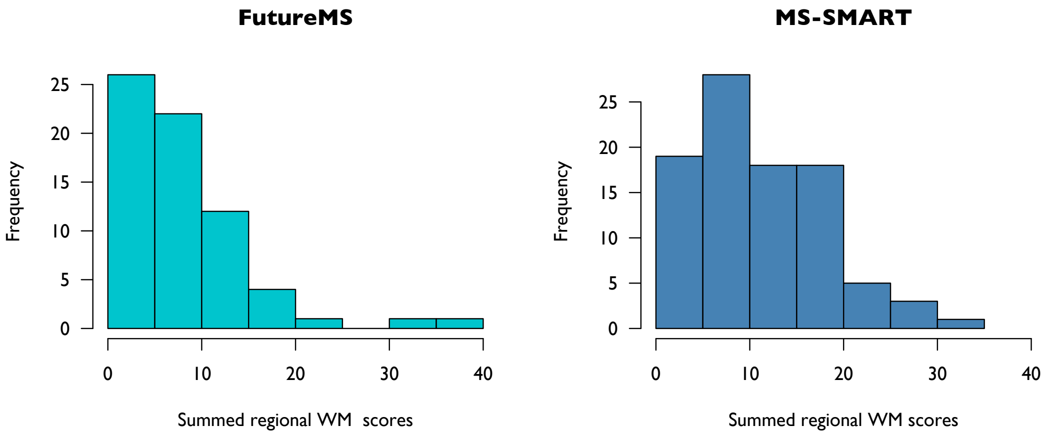


Figure 5.10: Histograms of summed regional white matter scores assigned to scans from two cohorts.

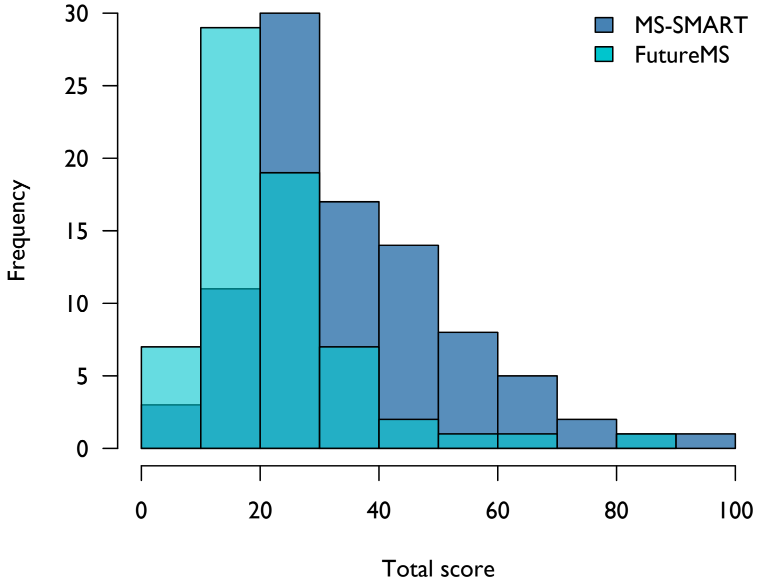


Figure 5.11: Overlapping histograms of total scores for scans from the two cohorts of interest.

5.5.5.1 Enlarged perivascular spaces

Enlarged perivascular spaces are associated with ageing in healthy populations, although not necessarily in people with MS [45]. A scatterplot of their association with age in the participants studied here is shown in Figure 5.12. The (Spearman) correlation was r = 0.40, and the relationship seemed similar in both cohorts when examined separately (r = 0.40 and r = 0.26 for FutureMS and MS-SMART participants respectively). However there was no significant relationship with cerebral atrophy (r = −0.02), the summed global (r = 0.06) or regional (r = 0.12) WM lesion scores.

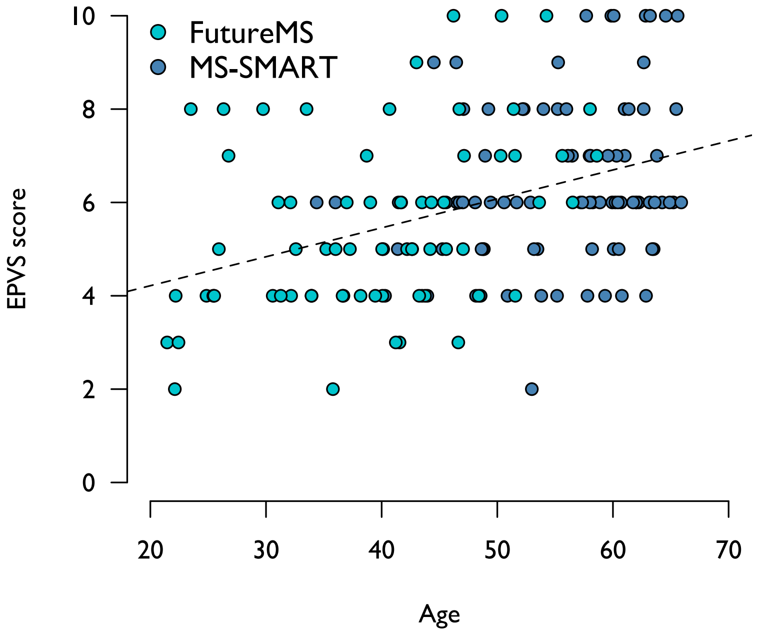


Figure 5.12: Scatterplot of total EPVS score against age for participants from both cohorts.

5.5.6 Validation with reference standard

Scatterplots of the two summary white matter scores (the summed global WM score and the summed regional WM score) are shown plotted against manual WM lesion volume in Figure 5.13 for the scans from the 43 participants in the MS-SMART Advanced MRI substudy. For the regional WM scores, the relationship appeared plausibly linear, but for the global 'Fazekas-style' scores there appeared to be a ceiling effect at high lesion volumes.

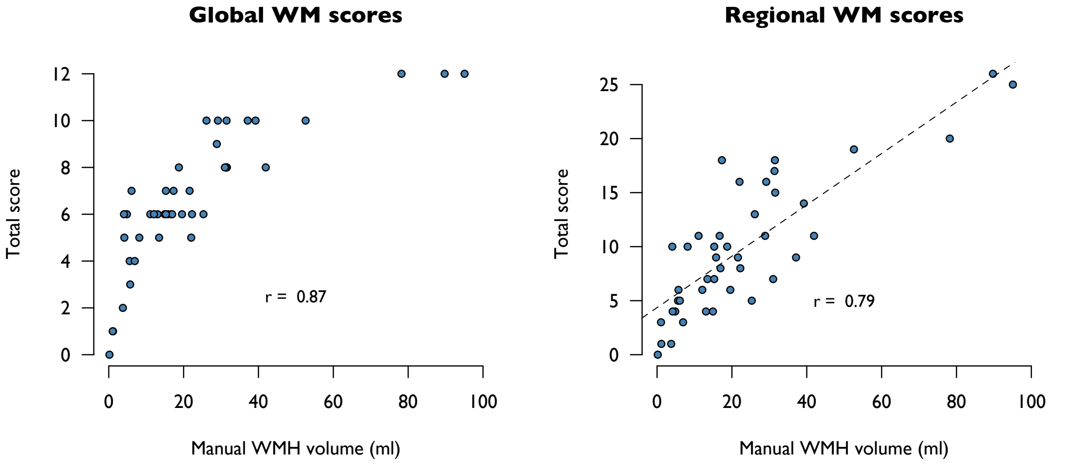


Figure 5.13: Scatterplots of summed global and regional WM scores against manual white matter hyperintensity (WMH) volume (n = 43), annotated with Spearman correlation and line of best fit for regional WM plot.

5.6 Discussion

Different forms of assessment will be appropriate to different situations. Automated methods of image analysis used in research are not currently applicable or practical on an individual patient basis and bear little resemblance to assessments used in clinical practice. While offering sensitive and reproducible measures, computer software may not be able to access and utilise potentially relevant features, apparent to human assessors. Judgments regarding the importance of particular imaging features and patterns may be better made by trained observers. However qualitative assessments are less easily subjected to statistical analysis and may be less sensitive to small changes, so information may be lost to research studies.

The failure of unidimensional quantitative imaging markers to fully explain disability in people with MS is well-established and pursuing greater measurement accuracy for any single outcome is unlikely to solve this. While the MS imaging community has largely moved on to investigation of advanced imaging techniques, there may nevertheless be additional information already available within routine sequences that is not currently captured. The aim of developing a structured imaging assessment was to recapture information regarding a broader range of imaging features for use in research, combining aspects of the descriptive reports provided in clinical radiology and the quantitative measurement tools more frequently used in research. The investigation reported in this chapter demonstrates that relevant information on white matter lesion burden can be recorded reliably using visual rating.

The visual rating system described here is entirely novel to MS and was developed using standard image viewing software, with no additional time needed for image processing. The overall time taken varied with the level of disease burden, but the global summary white matter ratings, adapted from the 'Fazekas' scale, showed a close relationship to quantitative markers with very little associated time burden. The practical advantages of a rating system allowing straightforward translation between clinical practice and research as well as between centres is clear. However at present there is no similar system used in MS imaging research.

The relationship of semi-quantitative markers of lesion burden to the reference standard volumetric measurement showed a high correlation, similar to that found in the context of vascular disease [174], and was clearly able to distinguish between different levels of disease. While sensitivity to small increments in lesion volume may not be achievable from visual assessments, the utility of these is unproven, and simple stratification based on visual ratings may prove equally useful. Van Straaten et al [174] compared different vascular disease rating scales and suggested that those with wider ranges could be more useful to differentiate between clinical groups but Fazekas scores were more useful to define groups based on imaging appearances. The flexibility of the rating system developed here for MS imaging could allow both these uses, dependent on context and time availability, with a shorter version based only on the adapted Fazekas scores likely to be more suitable when aiming to maximise subject numbers.

The limited range of responses available for most items was still found to allow reasonable dispersion of measurements, even within quite different cohorts. It was also possible to detect clear differences between cohorts at different disease stages, although the study was not specifically designed to address this.

Understanding the reproducibility of any measurement tool is critical for its use. Straightforward assessment of rater agreement and reliability was possible and appeared reasonable for the majority of features of interest. These measures will be specific to the cohort it is used in, but here it has been tested in two cohorts with differing imaging appearances and proved robust. The issue of whether observers need to be experts, such as neuroradiologists, remains open. There was however evidence at all stages of the development and testing process that some features could be identified more consistently than others. Global and regional lesion scores, atrophy scores and identification of (juxta-)cortical lesions showed consistency both for inter- and intra-rater agreement. This is perhaps partly attributable to familiarity, reflecting similarities with clinical practice. Conversely, identification of lesion cavitation and scoring of EPVS proved more variable. There was some evidence that individual rater biases may have influenced this, with stronger intra-rater reliability, and further training and feedback could be used to address this.

A trade off with practicality may have limited the results of this study. Standardisation between raters could be improved with the development of training datasets and time for feedback, group training sessions and access to more guidance pictures during data collection. These would be valuable areas for future testing. The number of raters was relatively low, although large within the context of a reproducibility study. Opportunistic use was made of available scans, which did not all follow the same protocol, although they were chosen to provide a range of disease burden. Reflecting this pragmatic cohort, no sequences were acquired which were highly sensitive to cortical lesions (e.g. phase-sensitive inversion recovery and double inversion recovery). Despite these limitations, the system appears robust, demonstrating reasonable reliability and a clear relationship to reference measurements. The rating system presented here builds on the extensive experience of using the same or similar markers in other fields.

Although significant differences were identified between cohorts, as expected, this was not the purpose of the work. There was no rater blinding as to the study cohort participants belonged to and the imaging, although performed on the same scanner, did not follow identical protocols. Further work would be needed to determine if the rating system described could be used to distinguish between disease phenotypes.

In pursuit of the thesis aim of extracting the most relevant information on white matter health from imaging and relating this to cognitive function, the visual ratings described here are used for this purpose in Chapter 7.

Chapter 6

Optimisation of an automated method for white matter hyperintensity quantification

6.1 Introduction

The practical advantages of using semi- or fully-automated software to generate image analysis outputs are readily apparent. Software-based segmentation techniques can be incorporated into image processing pipelines, supporting the generation of standard outcomes at different times and potentially across different sites. Manual segmentation as an alternative is a subjective process, both time-consuming and user-dependent, as shown in Chapter 4. This is particularly the case when the extent of abnormalities is large, with unclear margins. There is extensive use of software-based techniques for white matter hyperintensity (WMH) segmentation [58, 175] in multiple sclerosis (MS) imaging research, although many of these are not publicly available and no one technique has become standard.

Dichotomised maps of 'normal' and 'abnormal' white matter can be produced using signal intensity thresholds or contour-following methods that identify steep gradients of changing white matter signal. However WMHs do not necessarily have absolute boundaries resolvable on magnetic resonance imaging (MRI) and locating the best approximation is likely to become an increasing problem in later stage MS, when white matter surrounding focal lesions may become progressively and diffusely abnormal. Pathological-radiological correlation studies are rare, with small subject numbers, and therefore optimisation of any imaging-based method of segmentation must be based on a more practical approach. The software tested here allows recognition of the uncertainty at the border between normal and abnormal tissue by assigning probabilities to each voxel of belonging to each classification.

This chapter presents an investigation into optimising a novel method in MS imaging for generating both unidimensional WMH volumes and three-dimensional masks. The optimal software for segmenting WMHs in MS would produce outputs with the closest match to a reference standard, but the measures used to determine this optimal fit may depend on the use for which the output is required. For that reason an in house hyperintensity segmentation software was used to allow complete flexibility during the optimisation process. First, the unidimensional output (total quantity of abnormal tissue) is considered, either to best distinguish between different levels of disease or to best match the reference data in absolute terms. Secondly the spatial agreement is evaluated, using metrics that examine specifically the degree of overlap of segmented abnormality, as well as those also accounting for agreement on excluded, 'normal', tissue. The imaging data used for this work were taken from participants in a chronic progressive disease stage, in which diffuse and widespread white matter involvement was considered likely.

6.2 Methods

6.2.1 Participants and Imaging

The scans used for this work formed part of the baseline assessment of the 43 people enrolled at the University of Edinburgh centre for the MS-SMART Advanced MRI substudy. See Chapter 2, Section 2.1 for further details of the cohort and their imaging.

6.2.2 Reference standard

All software outputs were compared to manual segmentations performed on fluid attenuated inversion recovery (FLAIR) axial sequences by a single neuroradiologist, blinded to all clinical and demographic information. (See Chapter 4 for further detail of assessment of this reference standard.) To ensure comparability of segmentation volumes, the reference segmentations were multiplied by the atlas-derived (cerebral) white matter probability mask before any comparison. This removed any segmented hyperintensities that involved posterior fossa structures or extended into subcortical grey matter, which the software was not designed to assess.

6.2.3 Automated segmentation

The white matter segmentation method is described in Chapter 2, Section 2.1.3.1. Probabilistic maps of WMHs were derived through statistical conversion of the FLAIR sequence, multiplied by the atlas-based white matter probabilistic mask. The threshold below which all FLAIR intensities were immediately assigned zero probability was altered between the range of 0.7 to 1.7 standard deviations above the mean intensity, based on prior experience. Cumulative distribution transformation of the standard deviation (SD) maps produced an initial map of probabilities where each voxel value represented the probability of that voxel being classified as a WMH based on its intensity. A second, probability-based threshold could be set, above which all voxels were considered WMHs, creating a binary segmentation. Probability thresholds (P_t) from 0.05 to 0.9 were tested, i.e., all voxels with probability greater than or equal to P_t were given a binary classification as 1 (WMH) and those less than P_t were given a binary classification as 0 (not WMH). Probabilities below 0.05 were excluded at all thresholds to remove noise.

Masks were examined in FSL viewing software [104] to identify sources of discrepancy between automated and manual segmentations, but no manual adjustments were made to the automatically generated masks.

6.2.4 Statistical analysis

6.2.4.1 Unidimensional agreement

Agreement on the total WMH volume between manual and automated segmentations were assessed using Spearman correlations and percentage differences between output volumes. The threshold combinations which optimised these parameters were identified. Agreement was also examined visually using Bland-Altman plots of the volume ratio between the two segmentations.

6.2.4.2 Spatial agreement

All voxels within the white matter mask were classified as to whether they were identified as WMH on the manual segmentation only, the automated segmentation only, both or neither. Sums of these voxel categories were used to generate spatial agreement metrics, including Dice indices for each comparison. As described previously (Chapter 4, Section 4.2.3.2), this measures the voxel overlap between the two masks and is defined as twice the ratio of the number of overlapping voxels to the sum of the voxels in each segmented mask. Concerns that the Dice index increases with the size of the regions of interest, not taking into account the size of regions excluded by both segmentations, were addressed by additionally calculating sensitivities, specificities, negative and positive predictive values based on all voxels, as well as Youden's Index (defined as Sensitivity + Specificity −1).

6.3 Results

6.3.1 Participant characteristics

The scans used were the baseline imaging from participants in the Advanced MRI substudy of MS-SMART. Automated tissue segmentation failed for one scan, which had been acquired with atypical imaging parameters, and data from this participant was excluded. The remaining 42 scans were successfully segmented. This represented data from 29 female and thirteen male participants, with a mean age of 55.5 ± 8.4 years. All participants had a diagnosis of secondary progressive MS (SPMS) with median disease duration of 22.1 years (interquartile range (IQR): 15.5, 27.0). The reference standard WMH volumes (as described in Chapter 4) had median 17.1ml (IQR: 7.3, 30.6).

6.3.2 Unidimensional reproducibility

6.3.2.1 Reliability

Correlations (Spearman) with the manual lesion volume were high (> 0.9) for 80% of the threshold combinations examined, only falling below this when both low FLAIR SD and probability thresholds were used. This demonstrated that the software method reliably distinguished between participants with different levels of disease. The maximum correlation found was r = 0.96 (thresholds: FLAIR SD = 1.0, probability = 0.7). A scatterplot of volumes derived at this threshold combination against the reference manual segmentation is shown in Figure 6.1. As expected from the correlation, this showed a tight correspondence between the variables. However the software-derived volume tended to be lower than the manual volume, with the line of best fit having a gradient of 0.73.

6.3.2.2 Absolute agreement

The mean absolute percentage difference in WMH volumes between the manual and automated segmentations was minimised at 26.2% with thresholds FLAIR

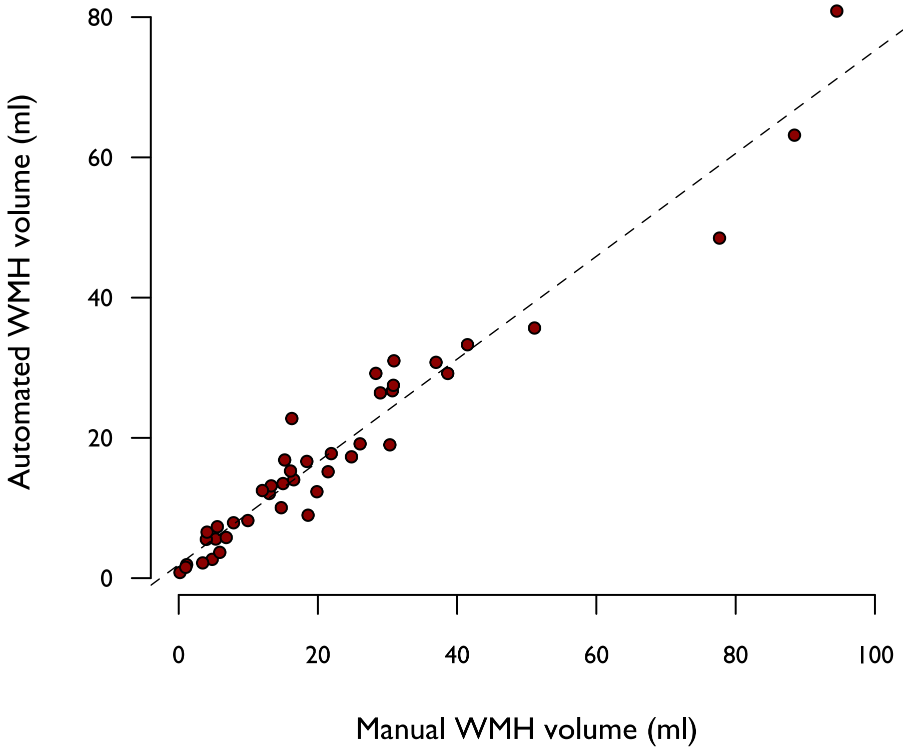


Figure 6.1: Scatterplot showing automated and manual WMH volumes (n = 42) at the threshold combination (FLAIR SD = 1.0, probability = 0.7) chosen to maximise their correlation (Spearman r = 0.96), with line of best fit.

SD = 1.1 and probability = 0.6. The effect of varying these thresholds on the cohort mean percentage difference is shown in the 'contour' plot in Figure 6.2 and a Bland-Altman plot of the ratios between volumes produced using the two methods for the optimal threshold combination is shown in Figure 6.3. Although the majority of points lie within the 95% confidence intervals, the Bland-Altman plot shows a trend towards improving agreement at higher WMH volumes. This would be expected when a larger proportion of the white matter is involved. There also appears to be a tendency for the automated segmentation to produce smaller WMH volumes at higher levels, compared with the manual segmentation. This may relate to the assignment of probabilities below the tested threshold by the software in regions with unclear WMH margins, compared with the binary manual segmentation. Alternatively it may reflect the method through which individual thresholds are chosen. Specifically, individual thresholds for lesion identification were based on SD multiples from the individual mean of whole brain FLAIR signal. For those individuals with a greater of pathology, the threshold may be substantially higher.

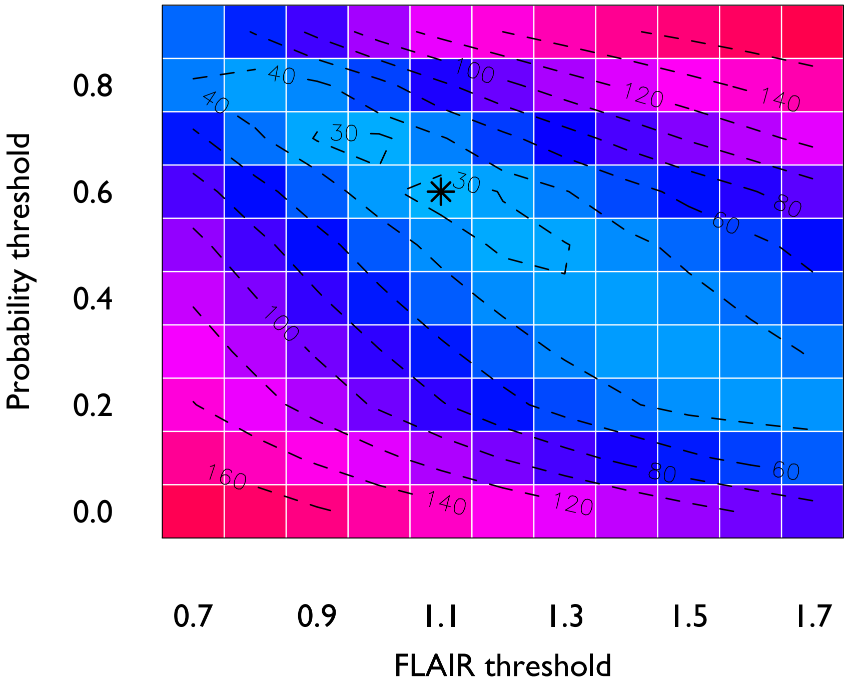


Figure 6.2: Contour plot showing the mean absolute percentage difference for all tested combinations of FLAIR SD and probability thresholds. Values along contours indicate approximate levels of mean absolute percentage differences along them. An asterisk marks the threshold combination at which this is minimised.

By considering two measures of unidimensional reproducibility, either seeking to maximise the correlation or minimise the (percentage) volume difference between masks, two different, although close, threshold combinations were found to be optimal.

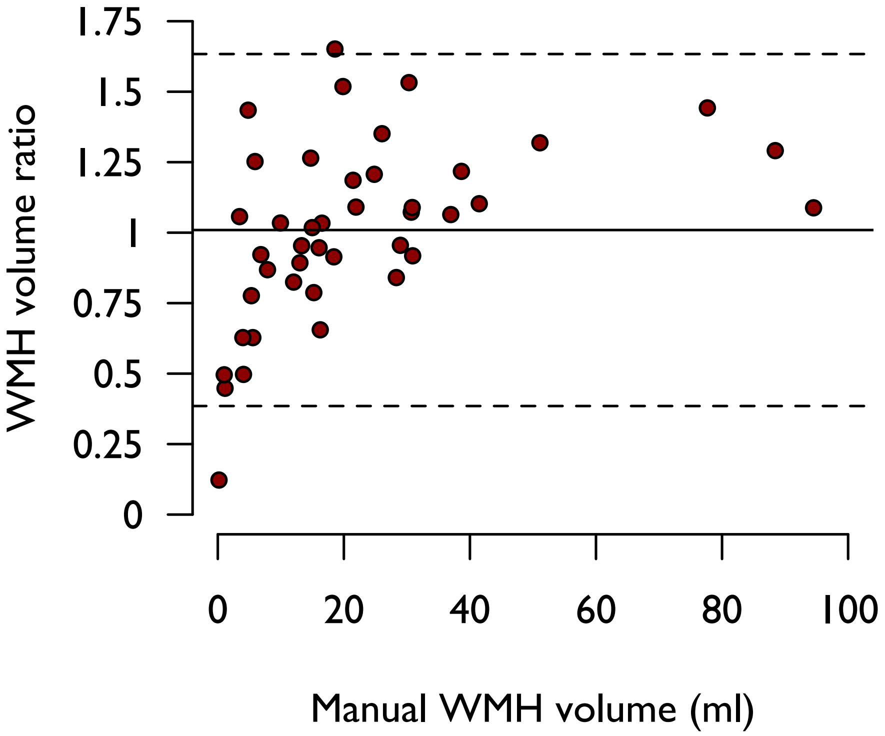


Figure 6.3: Bland-Altman plot showing the ratio of the reference standard (manual) WMH volumes to that of the automated segmentation volumes (n = 42) at the threshold combination minimising the mean absolute percentage difference (thresholds: FLAIR SD = 1.1, probability = 0.6). The solid horizontal line indicates the mean volume ratio (= 1.01) with dashed lines indicating the 95% confidence limits.

6.3.3 Spatial agreement

6.3.3.1 Dice index

The mean Dice index for spatial overlap was assessed at all tested threshold combinations and the effect of this is shown in the contour plot in Figure 6.4. The 0.6 boundary encloses a range of threshold combinations which led to a mean cohort Dice of 0.6 or greater. This was maximal, at a value of 0.62, for thresholds FLAIR SD = 1.3, probability = 0.4.

A scatterplot of the Dice index against manual WMH volume at the optimal threshold combination is shown in Figure 6.5. At this optimal threshold, the

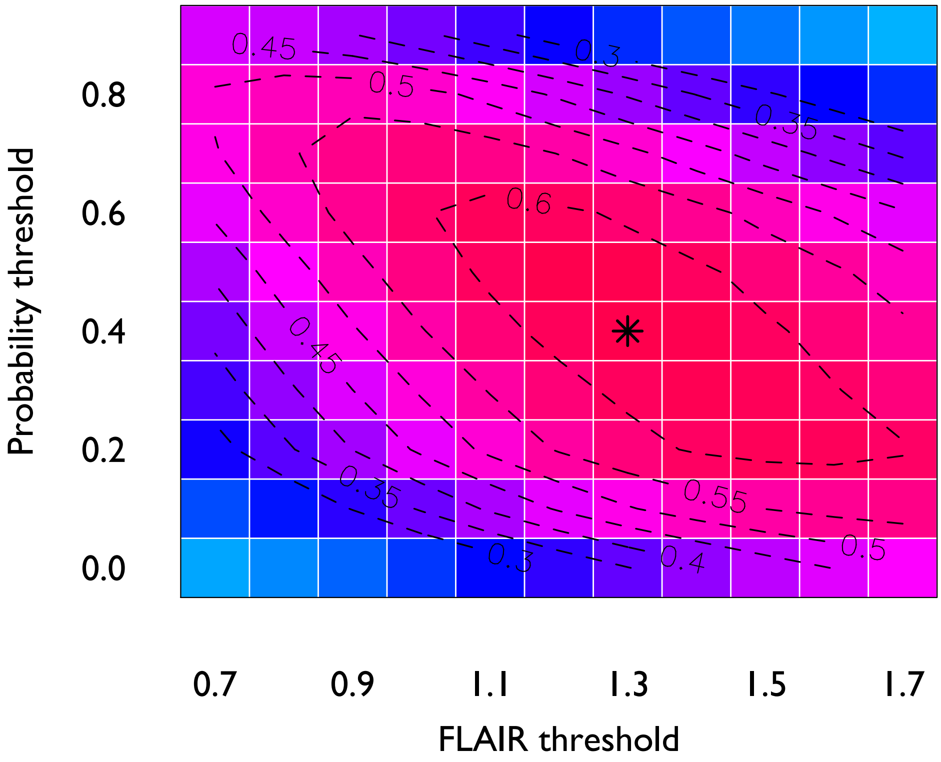


Figure 6.4: Contour plot showing the cohort mean Dice index for all tested combinations of FLAIR and probability thresholds. Values along contours indicate approximate levels of the mean Dice along them. An asterisk marks the threshold combination at which the mean Dice is maximal.

individual scan Dice indices still showed wide variation; they increased with increasing WMH volume, a recognised feature of the Dice index, which does not adjust for agreement by chance alone.

6.3.3.2 Youden's index

The classification of each white matter voxel was determined by its status on manual and automated segmentations; summing these allowed the calculation of sensitivities, specificities, positive (PPV) and negative predictive values (NPV)

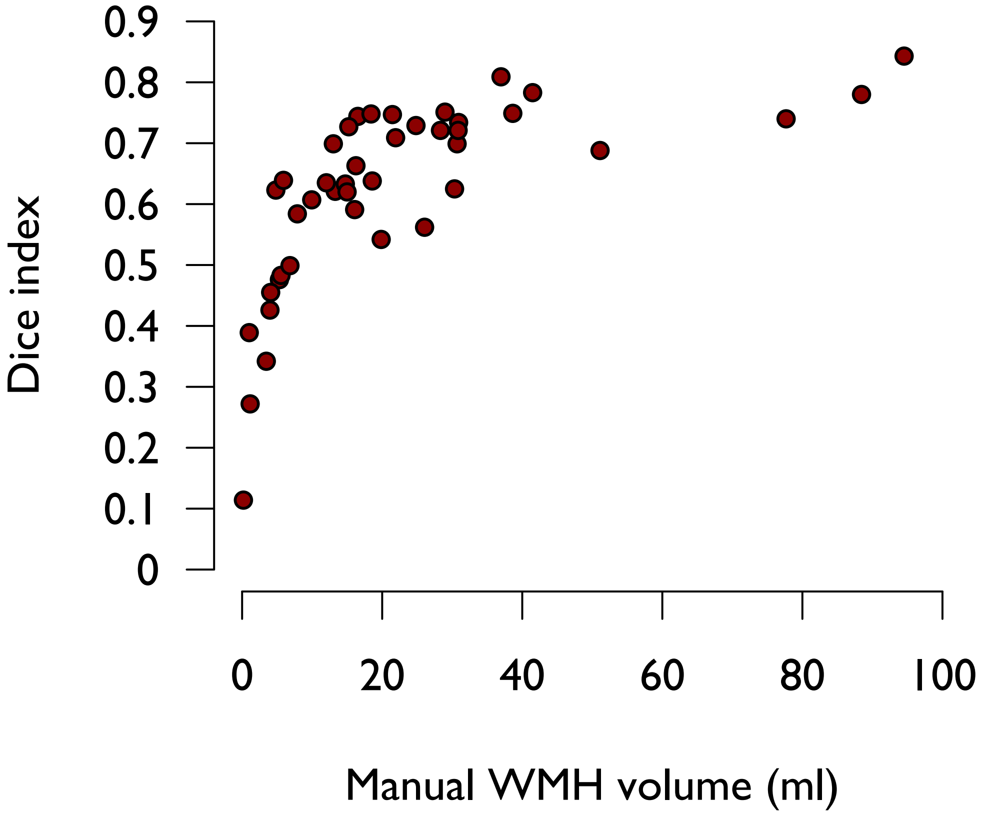


Figure 6.5: Scatterplot of individual scan Dice indices (n = 42) against manual WMH volume for the optimal threshold combination (FLAIR = 1.3, probability = 0.4). The mean cohort Dice was 0.62.

for each scan. Contour plots showing the effect of varying thresholds on these measures are shown in Figure 6.6.

Youden's Index combines information from the sensitivity and specificity, thus taking into account the number of white matter voxels not assigned to the WMH mask by either the automated or manual segmentation (ignored by the Dice index). This was maximised at 0.87 for a threshold combination of FLAIR SD = 1.3, probability = 0.0 (the lowest probability threshold, only excluding probabilities designated noise). This gave a sensitivity of 0.95 and a specificity of 0.92, a 'perfect' NPV (1.00) but a very low PPV (0.23), highlighting the fact

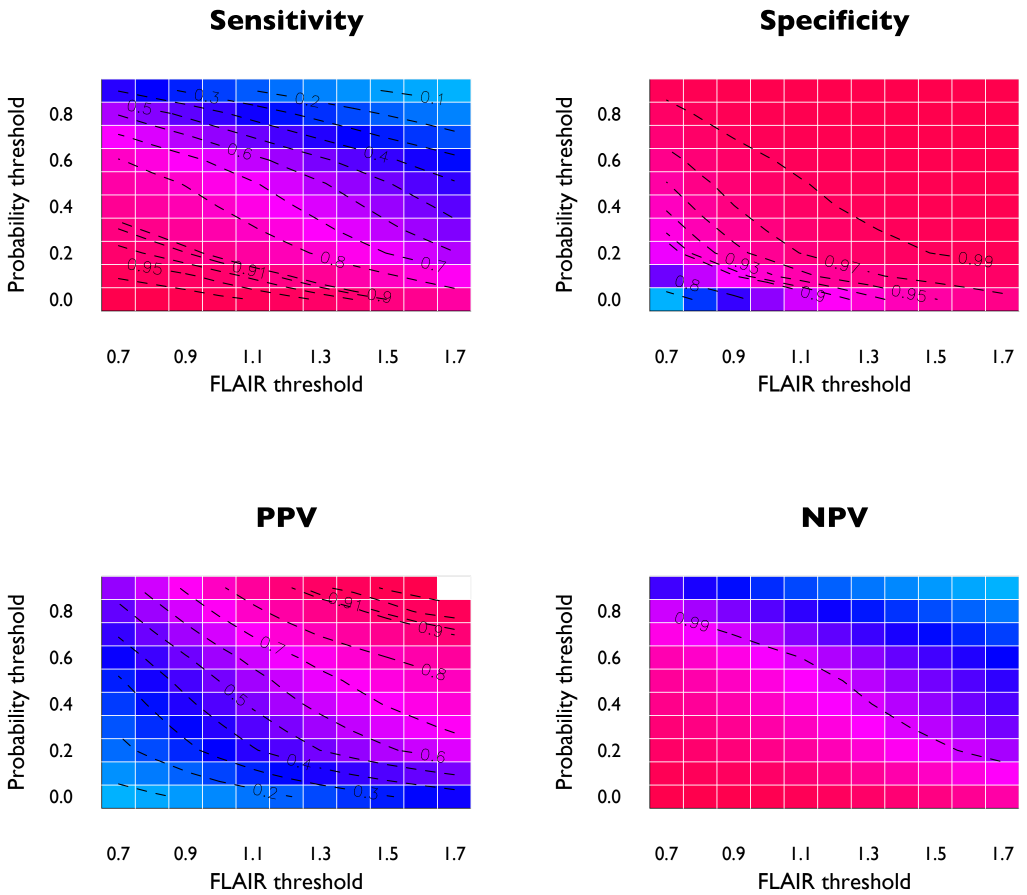


Figure 6.6: Contour plots for additional measures of spatial agreement for all tested FLAIR SD and probability thresholds. The blank square for threshold combination 1.7/0.9 in the PPV plot is due to one scan being assigned a WMH volume of 0ml by the software, leading to an undefined PPV. PPV: positive predictive value; NPV: negative predictive value.

that a high sensitivity and specificity may not necessarily produce a close match to the reference segmentation.

6.3.4 Sources of discrepancy

A major source of discrepancy contributing to large volume differences was in assigning boundaries to diffusely abnormal white matter, where signal varied gradually towards an unclear edge. Thresholds could be altered to maximise agreement on these scans but this was not necessarily optimal for scans showing much lower volumes of disease, with sharper edges to WMHs. Additionally both manual and automated segmentations were based primarily on the axial FLAIR sequence which had a 3mm slice thickness and partial volume effects between slices led to further blurring of WMH boundaries.

The corpus callosum was a recurrent source of spatial discrepancy, caused in part by partial volume effects from its proximity to CSF and exacerbated by frequent atrophy. For manual segmentation it would normally be best reviewed in the sagittal plane but the 3mm slice thickness had made this difficult.

An example image is shown in Figure 6.7 highlighting some of the causes of discrepancy.

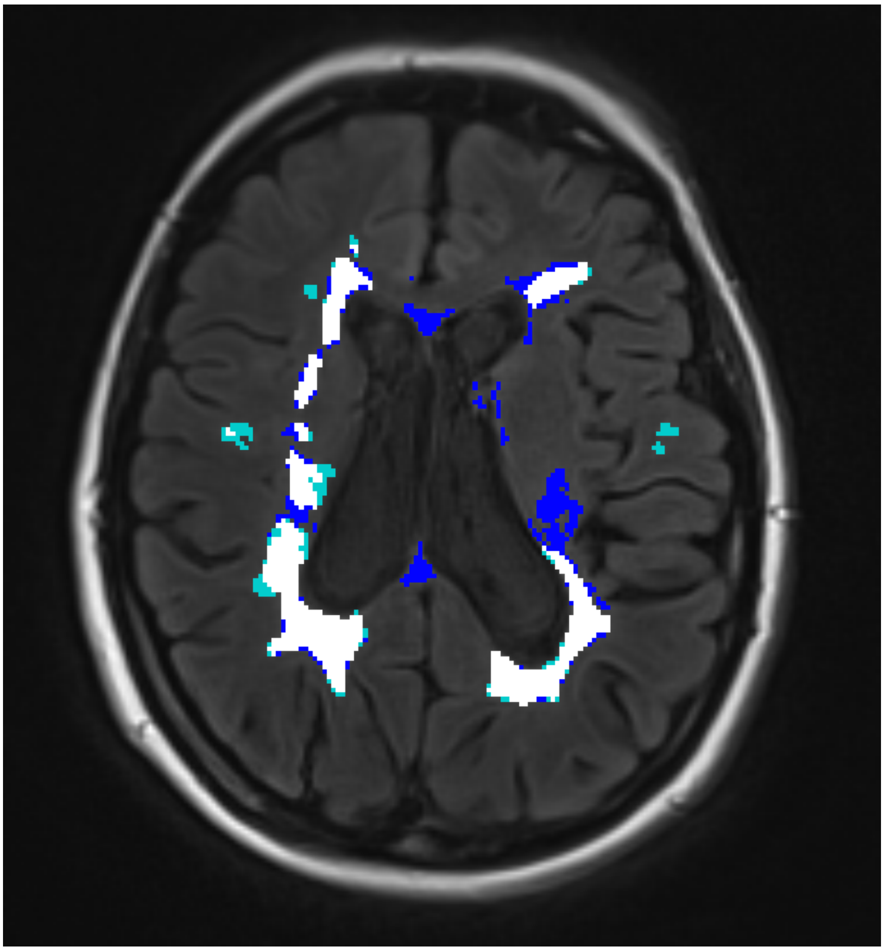


Figure 6.7: Sample image from MS-SMART participant showing manual WMH segmentation in turquoise, automated WMH segmentation in blue and overlapping regions in white. There is agreement on the majority of the tissue, although the automated segmentation boundary tends to be extend further. Only the automated segmentation has identified abnormal tissue in the corpus callosum. Conversely, only the manual segmentation has identified smaller focal WMHs distant from the ventricular surface.

6.4 Discussion

Multiple sclerosis is a diffuse disease of the central nervous system and hard lines drawn around visible regions of diseased brain are unlikely to represent true anatomical or pathological boundaries. This creates a persistent problem for quantification of MS pathology by brain imaging and cannot currently be addressed by defining a 'ground truth' (reference standard) with external validity. One approach to handling this problem is to incorporate a metric of uncertainty into the automated segmentation process, allowing the user to determine group classifications using a threshold of their choice. The novel method described here respects the gradient of abnormality, reflecting this in the voxelwise probabilistic output, which distinguishes it from most other currently used segmentation softwares.

The 'correct' statistical method used to optimise agreement remains an open question and even simple measures are often not reported (see Chapter 3 for a review of relevant literature). Taha & Hanbury [176] summarise the many different metrics used in the imaging literature. Whether the background ('true negative') rate is included, or not, as in the commonly used Dice index, affects assessment of good and bad segmentations. Several different desirable factors have been considered in this chapter, but others are possible with varying degrees of practicality. The mean Dice index found here (0.62), as well as the sensitivity and specificity, are comparable to that found in other validation studies [58], which mostly tested their software in smaller cohorts. The work described here demonstrates that this software can be optimised for particular cohorts and using the statistical measure of choice.

The probabilistic output can be used in different forms without repeating the segmentation process. For the purposes of linear modelling (see Chapter 7), the threshold producing an absolute WMH volume with the highest correlation (a measure of linear association) to manual segmentation, was selected. However for overlaying the diffusion maps (see Chapter 8), the probabilistic output was retained, with each voxel diffusion parameter multiplied by its probability of belonging to the tissue compartment of interest.

In most people with advanced multiple sclerosis, the majority of their brain WMH burden will be in supratentorial white matter, which was the case for this cohort. The results presented show that the automated software could accurately identify these. However lesions in the posterior fossa were not assessed, as appears to be the case with most segmentation softwares, although this information is not readily available in the public domain. While this cohort had low posterior fossa lesion volumes, as assessed with manual segmentation, this will not necessarily be the case in others. Lesions in the corpus callosum, a characteristic feature of MS, could not be assessed reliably, likely related to the 3mm slice thickness and associated partial volume effects with adjacent cerebrospinal fluid (CSF). This made confident manual segmentation of corpus callosum lesions difficult, and a conservative approach was adopted. Notable mismatches were observed here with the software classifying large proportions of the corpus callosum as abnormal in most subjects. This issue could be addressed by manual adjustment after software segmentation; not doing so here was chosen in order to separately assess the manual and automated methods.

Further validation of this software in independent datasets with different disease appearances is needed for it to be used more widely. Its performance here was tested using two-dimensional FLAIR sequences, but as there is a shift to greater use of volumetric three-dimensional acquisitions this will need re-evaluation. A limitation of this work is the lack of direct comparison with existing approaches to WMH segmentation, which in the majority of cases was due to the algorithms not being publicly available at the time. In the future it would be useful to compare the software tested here with that available in the Lesion Segmentation Toolbox (http://www.applied-statistics.de/lst.html) for use in the Statistical Parametric Mapping (SPM) software (http://www.fil.ion.ucl.ac.uk/spm/) as well as the BIANCA software [177], part of FSL, which has recently become available.

The method developed in this chapter was implemented in a fully-automated fashion to produce both binary and probabilistic maps of WMHs; the specific advantages are twofold. Thresholds both for the relative FLAIR signal intensity and the probability level can be adjusted for particular scans, subjects or larger studies, in order to optimise outputs for particular criteria of interest, generating binary masks and WMH volumes as required. Additionally, when an absolute boundary to WMHs is not required, the underlying probabilities assigned can be retained in the output, generating WMH volumes by summing these voxelwise and used in an analogous fashion with co-registered masks derived from advanced imaging techniques. Both WMH volumes and probabilistic masks generated by this method are used in the following two chapters in order to assess the relationship of brain imaging measures of disease burden to cognitive performance.

Chapter 7

Determining the relationship of white matter hyperintensity burden to cognitive performance

7.1 Introduction

This chapter describes work using the tools developed in Chapters 5 (a visual rating scale for the imaging features of multiple sclerosis (MS)) and 6 (automated quantification of white matter hyperintensities (WMHs)) to evaluate the relationship between the burden of WMHs and cognitive performance in people with MS.

Linear regression models were developed to evaluate a first hypothesis of a linear relationship between the WMH burden and cognitive performance in people with MS. A hierarchical approach was used for model construction, first addressing the relationship between cognitive performance and non-disease related covariates of relevance (see Chapter 1). The potential contribution of disease-related imaging metrics was then tested through addition to this 'core' model.

The second hypothesis tested was of a non-linear relationship between the burden of WMHs and cognitive performance in people with MS. This possibility was identified through evidence from previously published studies (see Chapter 3) that the pathology-phenotype relationship may be stronger at higher levels of disease burden. The question of a potentially dynamic relationship between pathology and phenotype was therefore addressed by exploring non-linear relationships and interactions between variables. If confirmed, this finding would raise the possibility that redundancy and neuroplasticity are able to compensate for pathology at low levels, with diminishing capacity as the disease progresses.

7.2 Methods

7.2.1 Construction of a linear model based on hyperintense white matter hyperintensity volume

7.2.1.1 Cohort and available data

Data from the baseline assessments of participants in MS-SMART, a cohort of subjects with secondary-progressive MS (SPMS), were used for construction of a predictive model of cognitive performance. (See Chapter 2, Section 2.1 for further details of the cohort.) Scores for the Symbol Digit Modality test (SDMT) were available from their baseline assessment and were taken as a measure of cognitive ability for use as the dependent variable in all statistical models.

7.2.1.2 Participant characteristics

Information was available on all participants with regards to sex, age and disease duration. Where possible, data from other known modifiers of cognitive performance were also considered for inclusion in the predictive model.

Due to inconsistencies in recording of education status, either as full-time equivalent years in education or as the highest level of educational attainment, a decision was made to dichotomise education data. Participants were classified by the presence or absence of educational exposure beyond compulsory schooling (i.e. entry to higher education), with a cut-off of twelve full-time equivalent years of education used where this was unclear.

Certain licensed drugs are known to be potential modifiers of cognitive performance. For each participant, all prescribed drugs were recorded at baseline, and these were classified into three groups according to whether they were likely to be associated with better or worse cognitive performance, or if there was no reason to expect any effect on cognition. On this basis, participants were split into three groups.

Participants' scores on the Beck Depression Index (BDI) were recorded during screening for study entry (up to one month prior to baseline assessment) as part of the study eligibility criteria.

7.2.1.3 Imaging data

The following volumetric data were available from the standard study baseline imaging protocol and subsequent automated tissue segmentation: intracranial volume (ICV), total brain volume and white matter (WM) volume.

After the investigations described in Chapter 4, the automated software was used to generate WMH volumes for all MS-SMART participants, using the threshold that maximised correlation with the (third) manual segmentation in the Advanced MRI substudy cohort. This was run on the entire MS-SMART cohort for consistency.

7.2.1.4 Univariate relationships

Scatterplots of individual predictors against SDMT scores were examined for outliers and evidence of non-linear relationships. Spearman correlations between all individual numerical predictors and SDMT scores were calculated. For binary predictors, t-tests were used to test for significant differences between groups.

7.2.1.5 Model construction

Generalised linear modelling was used, with an assumption of normally distributed errors. A strong prior literature exists regarding the influence of age, sex and educational status on SDMT performance [40]. The side effects of the drugs classified as having a potentially detrimental effect on SDMT are also well-established. A hierarchical approach to model construction was adopted, with an initial model based on age, sex, drugs and education status only. Education status and being prescribed drugs with a potentially detrimental effect on cognition were both modelled as binary predictors.

Initial model:

$$\text{SDMT} = \beta_0 + \beta_1\text{age} + \beta_2\text{sex} + \beta_3\text{education} + \beta_4\text{cognitive inhibiting drugs}$$

Intracranial volume reflects peak adult brain volume and represents a fundamental non-disease metric, with a known relationship to cognition and cognitive decline [38,39]. In a second model using both imaging and non-imaging metrics, ICV was therefore included as the sole imaging metric, to produce the optimal predictive model without markers of disease burden.

Second phase model:

$$\begin{aligned}\text{SDMT} = {} & \beta_0 + \beta_1\text{age} + \beta_2\text{sex} + \beta_3\text{education} + \beta_4\text{cognitive inhibiting drugs} \\ & + \beta_5\text{intracranial volume}\end{aligned}$$

The initial disease marker considered was WMH volume, modelling the impact of focal inflammatory disease. This was added to the prior model containing non-disease-related variables to form the third phase linear model.

Third phase model:

$$\begin{aligned}\text{SDMT} = {} & \beta_0 + \beta_1\text{age} + \beta_2\text{sex} + \beta_3\text{education} + \beta_4\text{cognitive inhibiting drugs} \\ & + \beta_5\text{intracranial volume} \\ & + \beta_6\text{WMH volume}\end{aligned}$$

In the final (fourth) linear model, WM volume was added in an attempt to model the impact of neuroaxonal loss. White matter volume was chosen in preference to total brain volume as possibly the more relevant to a distributed function such as processing speed and the effect of WMH burden.

Fourth phase model:

$$\begin{aligned}\text{SDMT} = {} & \beta_0 + \beta_1\text{age} + \beta_2\text{sex} + \beta_3\text{education} + \beta_4\text{cognitive inhibiting drugs} \\ & + \beta_5\text{intracranial volume} \\ & + \beta_6\text{WMH volume} \\ & + \beta_7\text{WM volume}\end{aligned}$$

7.2.1.6 Assumption checking

Individual predictor scatterplots were examined for outliers. Correlations between all individual predictors were calculated. For all models constructed, histograms of the residuals, Q-Q plots and plots of residuals against predicted values were assessed.

7.2.1.7 Investigation of non-linear relationships

The possibility of a non-linear relationship between WMH volume and SDMT being a better fit for the data was initially explored graphically using a loess fit [178], a locally-weighted smoothing function.

Two breakpoints were suggested by the smoothed plot, splitting the cohort into thirds. A piecewise linear regression model was therefore constructed with optimal breakpoints (one or two) sought iteratively. These piecewise linear regression models were constructed to evaluate both the univariate relationship between SDMT and WMH volume and this relationship within the context of other relevant covariates.

7.2.1.8 Sensitivity analyses

Disease duration was omitted as an independent variable in construction of the primary model, with a view to assessing the accumulated disease burden through imaging variables alone. In order to assess whether this could be an important predictor in itself, or provide greater explanatory power than imaging-derived predictors, sensitivity analyses were performed with the addition of disease duration to both the full model and that using only non-imaging variables, with assessment of model fit.

Related to disease duration, the possibility that age could act as a surrogate variable for pathological changes occurring over time was also considered. The age range of the cohort was not one at which substantial effects of cognitive ageing would be expected and this could have the unwanted effect of attenuating power to detect an effect of accumulated pathology due to its covariance with age. Correlations between age and imaging markers of brain pathology (WM volume and WMH volume) were considered. The effect on the goodness of fit parameters of models containing imaging markers was considered when age was omitted, with or without the addition of WM volume as an alternative variable.

The effect of adding total brain volume rather than WM volume as an additional imaging marker, following ICV and WMH volume was also considered.

7.2.1.9 Model comparisons and fit

Parameters of goodness of fit were compared for all models, including values for adjusted R^2, Akaike information criteria (AIC) and Bayesian information criteria (BIC), with the Wald test used to quantify the statistical significance of additional model predictors.

7.2.2 Construction of a linear model based on visual rating lesion scores

7.2.2.1 Cohort and available data

Data from the visual assessments of imaging from the baseline assessments of participants in FutureMS, a cohort of subjects with early relapsing-remitting MS (RRMS), and MS-SMART were used for construction of a predictive model of cognitive performance. (See Chapter 2 for further details of the cohorts and Chapter 5 for details of the visual assessments of their imaging.) Scores for the SDMT were available from the baseline assessment for each study and were taken as a measure of cognitive ability. There was no prior hypothesis that any imaging feature under consideration would have a different relevance for groups with early and later stage disease and the two cohorts were modelled together as one group.

7.2.2.2 Participant characteristics

Information was available on all participants with regards to age, sex and disease duration. Where possible, data from other known modifiers of cognitive performance were considered for inclusion in the predictive model. Participants were again classified according to whether they were prescribed drugs that could have a detrimental or beneficial effect on cognition. Data on education status and BDI were not available for the FutureMS cohort.

7.2.2.3 Imaging data

The complete initial set of rater scores from the independent validation study reported in Chapter 5 were available. Each of the dimension subscores constructed, representing summed scores for particular features of interest, were considered as predictors. Binary predictors, based on presence or absence of juxtacortical/cortical (JC) lesions and cavitated lesions, were also considered. Relevant to the focus on diffuse cerebral white matter disease, the atrophy score used was the mean of the deep and superficial cerebral atrophy scores, i.e. not incorporating the posterior fossa and corpus callosum ratings.

7.2.2.4 Univariate relationships

Plots of individual predictors against SDMT scores were examined for outliers and evidence of non-linear relationships. Spearman correlations between all individual predictors and SDMT scores were calculated. For binary predictors, t-tests were used to test for significant differences between groups.

7.2.2.5 Model construction

Generalised linear modelling was used, with an assumption of normally distributed errors. A hierarchical approach to model construction was adopted as previously, with an initial model based on age, sex and prescribed drugs.

Two summary white matter lesion/hyperintensity (WMH) scores (the global summary 'Fazekas-style' score and the summed regional white matter score) were available. The global summary score was chosen as a predictor for construction of a second model, as it was considered to be more representative of the total burden of cerebral white matter disease, given the summed regional score's over-representation of smaller regions and inclusion of grey matter and posterior fossa structures. Finally, further dimension subscores and binary predictors based on these were considered in turn as additional predictors.

7.2.2.6 Assumption checking

Individual predictor scatterplots were examined for outliers, with consideration of appropriate handling where necessary. Correlations between all individual predictors were calculated. For all models constructed, histograms of the residuals, Q-Q plots and plots of residuals against predicted values were assessed.

7.2.2.7 Linear modelling for separate cohorts

The univariate relationships between the independent variables in the main model and SDMT performance were considered separately for the two smaller cohorts, in order to explore whether predictor variables had different effects at different disease stages. The overall model was constructed separately for the two cohorts to examine the effect of the individual predictors within a multivariate context.

7.2.2.8 Sensitivity analyses

As for the modelling based on volumetric imaging markers, disease duration was not initially included with the aim of assessing the accumulated disease burden using imaging markers. The effect of its inclusion on model fit was tested both as an additional predictor in the full model and within a smaller model based only on non-imaging predictors. The potential for age and being prescribed drugs with a detrimental effect on cognitive performance to be acting as surrogate markers of disease stage was also considered. This was addressed both in the separate modelling for the two cohorts described above and investigating the effect of removing age from these smaller models.

7.2.2.9 Model comparisons and fit

Parameters of goodness of fit were compared for all models, including values for adjusted R^2, AIC and BIC, with the Wald test used to quantify the statistical significance of additional model predictors.

7.3 Results (I): Construction of a linear model based on automated white matter hyperintensity volume

7.3.1 Data completeness and participant characteristics

Scores for the SDMT were available from their baseline assessment for 91 of the 93 participants in MS-SMART. For the two instances of missing data, one participant was unable to complete either the Paced Auditory Serial Addition Test or SDMT due to cognitive limitations and for the other participant confusion during test completion invalidated the result. A histogram of SDMT scores showed no evidence of substantial deviation from a normal distribution.

Data on premorbid IQ and cognitive leisure activities were not available in this cohort. Data on educational status were unavailable for 17 participants. Four prescribed drugs (diazepam, clonazepam, baclofen and amitriptyline) were identified as potentially having a detrimental effect on cognitive performance and 36 participants were taking at least one of these. Two participants were prescribed modafinil, with a potentially beneficial effect on cognitive performance.

In two participants, volumetric imaging predictors were unavailable, in one case due to a T2-weighted sequence not being available and in one case due to failure of software segmentation.

Descriptive statistics for this cohort are presented in Table 7.1, including participant characteristics and volumetric imaging markers.

	n	Summary figures
Age (years)	93	57.2 (49.0, 61.0)
Sex (F/M)	93	69/24
Disease duration (years)	93	20.4 (14.0, 28.7)
BDI	93	6 (4, 11)
Drugs (Beneficial/Detrimental/Neither)	93	2/37/54
Education $\leq 12/> 12$ years	76	31/45
SDMT (mean ± SD)	91	43.2 ± 11.7
ICV (ml)	91	1308 (1243, 1395)
Brain volume (ml)	91	1136 (1063, 1209)
WM volume (ml)	91	425.4 (403.3, 453.0)
WMH volume (ml)	91	32.0 (23.4, 45.5)

Table 7.1: Summary statistics for MS-SMART cohort, used in predictive modelling. All continuous/numerical variables are given as median (interquartile range) other than for SDMT. The second column ('n') indicates the number of subjects for whom that data was available. BDI: Beck Depression Index; ICV: intracranial volume; WM: white matter; WMH: white matter hyperintensity.

7.3.2 Participant characteristics as predictors of cognitive performance

7.3.2.1 Univariate relationships

The relationships of the non-imaging characteristics presented in Table 7.1 to SDMT performance were examined. Only two participants were prescribed medication with a potentially beneficial effect on cognitive performance (modafinil), and their SDMT scores were the lowest and second highest in the cohort. This raised concerns over the timing of taking medication on the day of testing and it was not possible to ascertain this information.

Plots of individual non-imaging predictors against SDMT score are presented in Figure 7.1 for the MS-SMART cohort (n = 91). Numerical predictors gave Spearman correlations with SDMT of $r = 0.01$ for age, $r = -0.04$ for disease duration and $r = -0.10$ for depression score. There was no evidence of a non-linear relationship from the scatterplots and none of the correlations were significantly different from zero at the 5% level. For the binary predictors (sex ($p = 0.63$), educational status ($p = 0.19$) and being prescribed drugs with a potentially detrimental effect ($p = 0.08$)), t-tests did not show any significant differences between groups. Group differences both for those taking potentially detrimental drugs and those having a higher level of education were nevertheless in the expected direction.

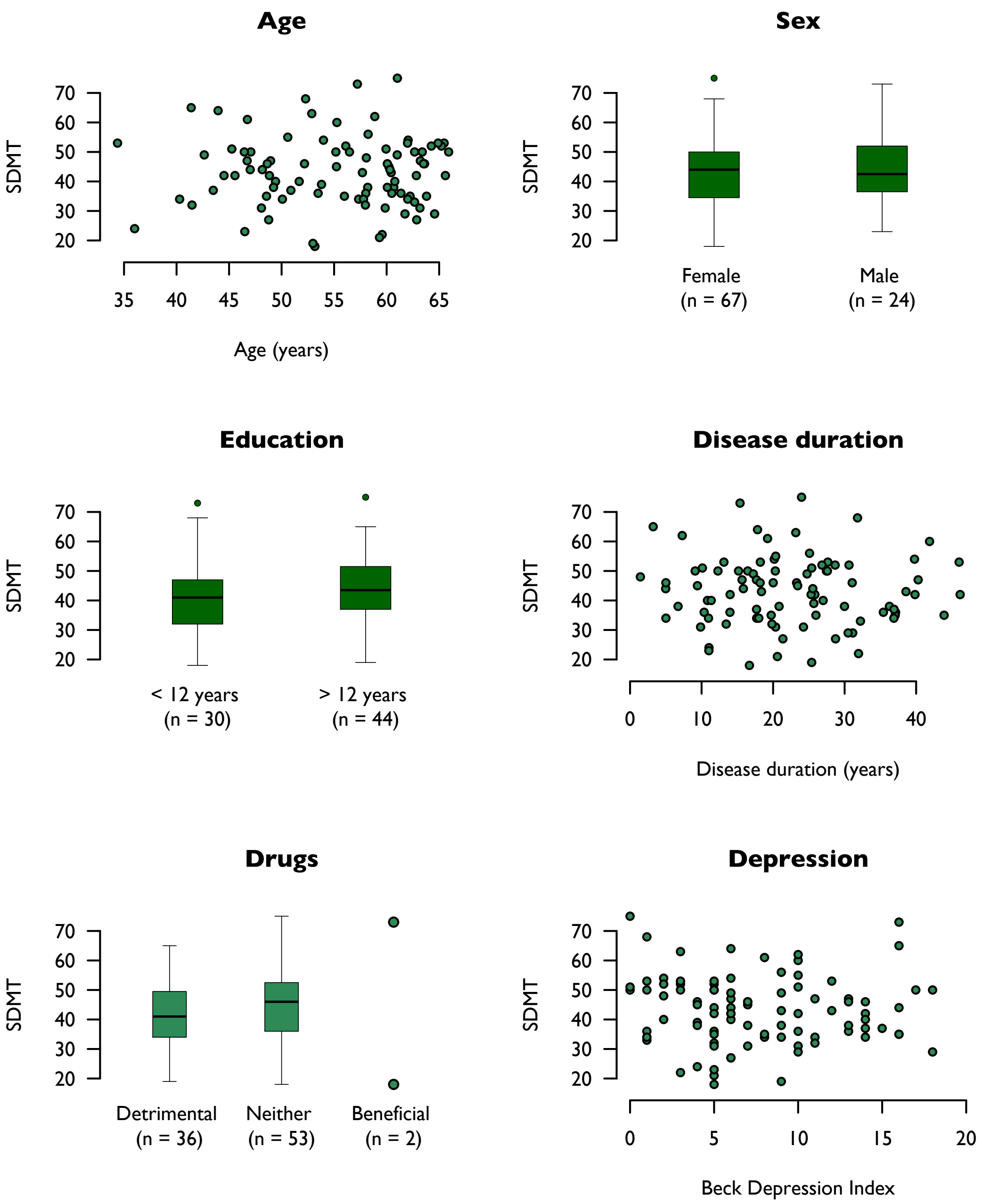


Figure 7.1: Plots of individual non-imaging predictors against SDMT (n = 91, for all except Education status (n = 74).) The bottom left plot shows participant SDMT scores grouped by whether they were prescribed at least one drug with either a potentially detrimental or beneficial effect on cognition or neither, no participant being prescribed both.

7.3.2.2 Construction of multiple linear regression model

Due to the uncertainties over the use and timing of medication with a potentially beneficial effect on cognition, the two participants prescribed modafinil were excluded from all predictive models using medication information.

Although pseudo-dementia due to depression is recognised, people with scores on the BDI thought high enough for this to have an influence on cognitive performance were not eligible for MS-SMART. No obvious relationship was found in the range of the MS-SMART cohort so BDI scores were also not included in the model.

Disease duration showed a significant correlation with age (r = 0.45) and was not used as an additional independent predictor, in preference of assessing accumulated disease burden through imaging markers.

7.3.2.3 Multiple linear regression model summary

A description of the model containing age, sex, education status and use of drugs with potentially detrimental effects as independent variables is presented in the first column of Table 7.2. In comparison with a model based on the intercept alone (the mean SDMT score), there was no definite evidence (p = 0.17) that the model based on four non-imaging predictors was a better fit to the data. With respect to individual predictors, there was some support for a possible influence of educational status (p = 0.12) and detrimental drugs (p = 0.08) on SDMT scores.

7.3.2.4 Interactions between variables

In the model described in column 1 of Table 7.2, each independent variable was added separately, without interaction terms, allowing more straightforward interpretation of coefficients. An alternative model, using the same non-imaging predictors but allowing interactions between them, did not improve model fit given the number of parameters (AIC 561.1, BIC 599.8) although there was a suggestion that education may interact with age (p = 0.08) and sex (p = 0.11).

	Dependent variable:			
	SDMT			
	1	2	3	4
Constant	50.51	3.85	14.69	21.51
	$p < 0.0001^{***}$	$p = 0.85$	$p = 0.40$	$p = 0.25$
Age	−0.14	−0.22	−0.13	−0.16
	$p = 0.43$	$p = 0.21$	$p = 0.38$	$p = 0.29$
Sex	−0.83	−7.71	−8.89	−7.74
	$p = 0.78$	$p = 0.04^{**}$	$p = 0.01^{***}$	$p = 0.03^{**}$
Education (over 12 years)	4.32	4.44	4.40	4.64
	$p = 0.12$	$p = 0.10^{*}$	$p = 0.06^{*}$	$p = 0.05^{**}$
Detrimental drugs	−4.87	−6.00	−3.36	−3.28
	$p = 0.08^{*}$	$p = 0.03^{**}$	$p = 0.15$	$p = 0.16$
ICV (ml)		0.04	0.04	−0.002
		$p = 0.01^{***}$	$p = 0.01^{***}$	$p = 0.96$
WMHV (ml)			−0.29	−0.29
			$p = 0.0001^{***}$	$p = 0.0001^{***}$
WMV (ml)				0.10
				$p = 0.35$
AIC	557.1	534.6	516.5	518.9
BIC	570.8	550.3	534.5	539.1
Observations	72	70	70	70
R^2	0.09	0.21	0.41	0.42
Adjusted R^2	0.04	0.15	0.35	0.35
Residual Std. Error	11.05	10.42	9.10	9.11
	(df = 67)	(df = 64)	(df = 63)	(df = 62)
F Statistic	1.68	3.49^{***}	7.29^{***}	6.36^{***}
	(df = 4; 67)	(df = 5; 64)	(df = 6; 63)	(df = 7; 62)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$			

Table 7.2: Summary of linear models with SDMT as dependent variable and participant characteristics and imaging markers derived from routine sequences as independent variables. Numbers shown in the main table are model coefficients followed by associated p values. Predictor abbreviations are as in Table 7.1.

7.3.3 Imaging markers as predictors

7.3.3.1 Univariate relationships

Imaging-derived variables summarised in Table 7.1 were considered initially as univariate predictors of SDMT performance. These gave Spearman correlations with SDMT of r = 0.28 for ICV, r = 0.33 for brain volume, r = 0.29 for WM volume and r = −0.48 for WMH volume. All correlations were significantly different from zero at the 1% level. For comparison with published literature (see Chapter 3), the Pearson correlation between SDMT and WMH volume was r = −0.45, with 95% confidence interval $(-0.61, -0.27)$. When adjusted for ICV, the correlations with SDMT were r = 0.33 for brain volume, r = 0.15 for WM volume and r = −0.49 for WMH volume.

Plots of individual imaging predictors against SDMT score are presented in Figure 7.2 for the MS-SMART cohort (n = 89). The scatterplots were open to multiple interpretations, but there was no strong evidence of a non-linear relationship for any of the variables, with the possible exception of WMH volume.

7.3.3.2 Construction of multiple linear regression model

Intracranial volume, brain volume and WM volume were all highly correlated (r = 0.93 to 0.96). Brain volume and WMH volume were significantly negatively correlated (r = −0.26, p = 0.01).

7.3.3.3 Multiple linear regression model summary

The model with the addition of just ICV was a significantly better fit to the data (p = 0.005) than the model based on non-imaging participant characteristics alone. The addition of WMH volume as a predictor produced a further significant improvement in model fit ($p < 0.00001$). The further addition of WM volume did not improve model fit (p = 0.35), an unsurprising result given the high covariance of WM volume and ICV (r = 0.95). Descriptions of these models are presented in columns 2 to 4 of Table 7.2.

7.3.3.4 Interactions between variables

A limited model using only the imaging predictors did not improve model fit when allowed interaction terms, given the number of parameters (increased AIC and BIC), although there was a suggestion that WM volume interacted with WMH volume (p = 0.09).

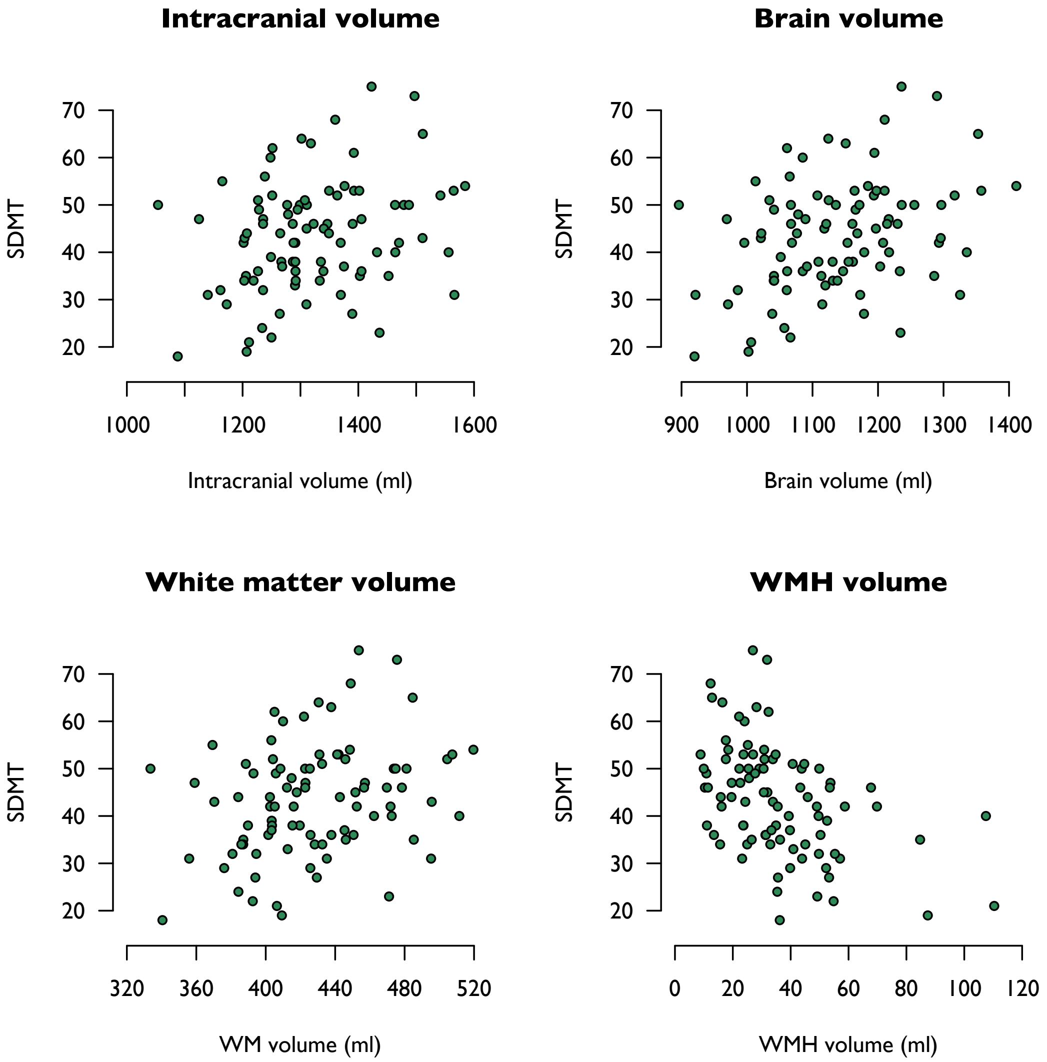


Figure 7.2: Plots of individual imaging predictors against SDMT (n = 89).

7.3.4 Model assumption checking

For the model described in column 3 of Table 7.2, a scatterplot of residuals against fitted values and a Q-Q plot are shown in Figure 7.3, with some suggestion that the fitted residuals deviate from normality at the tails of the distribution. This has implications for the use of a model predicting phenotype outside the mid-range and also raises the possibility of an underlying non-linear relationship.

There was a high correlation between ICV and WM volume (r = 0.95), consistent with the lack of improvement in model fit with the introduction of WM volume as a predictor.

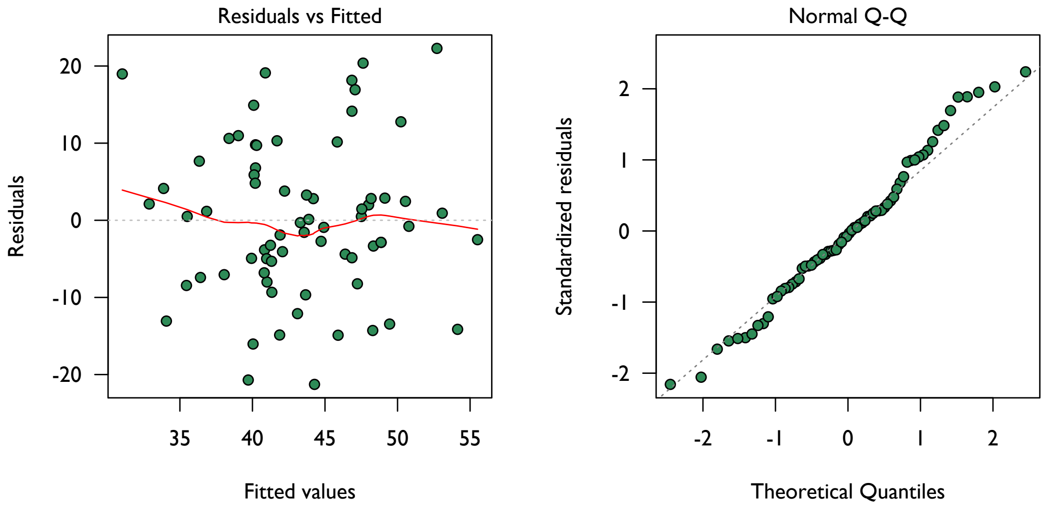


Figure 7.3: Left - plot of residuals against fitted values for linear model including participant characteristics, ICV, WM volume and WMH volume (column 3 of Table 7.2). Right - Q-Q plot of residuals for the same model.

7.3.5 Potential non-linearity in the relationship between WMH volume and SDMT

A scatterplot of SDMT against WMH volume is shown in Figure 7.4 with the added loess fit, suggesting three phases to the relationship. The fitted results of a piecewise regression produced by splitting the WMH data into thirds, also shown superimposed, were able to closely match the loess fit. Compared with a simple univariate model allowing only one slope, the model obtained by fitting the data piecewise in thirds showed a non-significant improvement in fit ($p = 0.11$; AIC 675.9, BIC 683.4 for simple model, compared with AIC 675.3, BIC 687.8 for non-linear model). Within the multivariate model, equivalent to that summarised in column 3 of Table 7.2, allowing three slope parameters for WMH volume also resulted in a non-significant improvement in model fit ($p = 0.18$; AIC 515.9, BIC 538.4).

7.3.6 Sensitivity analyses

Disease duration did not improve model fit when considered as an additional variable in models containing either only non-imaging or both non-imaging and imaging variables.

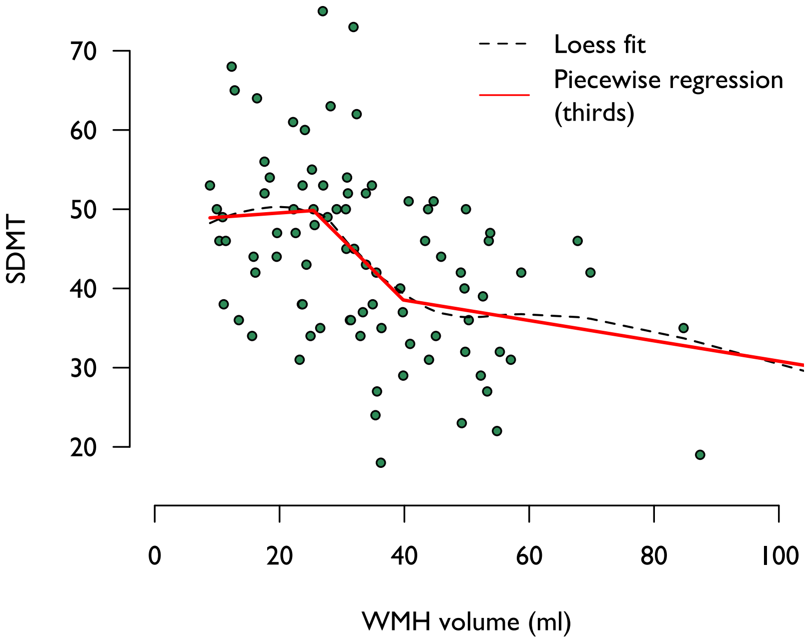


Figure 7.4: Scatterplot of SDMT against WMH volume. Superimposed lines show the piecewise regression fit produced by dividing the cohort into thirds by WMH volume and the loess fit.

A trend towards significance was seen in the (Spearman) correlation between age and WM volume (r = 0.20, p = 0.06), but not with WMH volume (r = 0.09, p = 0.38). However omitting age from the models described in columns 3 and 4 of Table 7.2 did not result in significant changes in model fit (p = 0.38 and p = 0.28 respectively) and the result of adding WM volume as a predictor to the smaller model remained a non-significant improvement in model fit (p = 0.47).

Adding total brain volume rather than WM volume as an additional imaging marker was associated with a non-significant improvement in model fit compared to the model described in column 3 of Table 7.2 (p = 0.11, AIC 515.7, BIC 537.7). WMH volume remained the only significant imaging-derived predictor, although brain volume showed a trend towards significance (p = 0.12).

7.4 Results (II): Construction of a linear model based on visual rating lesion scores

7.4.1 Data completeness and participant characteristics

A score for the SDMT was available for 91 of 93 MS-SMART participants (as described in section 7.3.1) and all FutureMS participants. A histogram of SDMT scores showed no strong evidence of deviation from a normal distribution.

Descriptive statistics for the combined FutureMS (n = 67) and MS-SMART (n = 93) cohorts are shown in Table 7.3, including participant characteristics and visual rating markers. Seven participants in FutureMS were prescribed at least one of the previously-identified drugs with a potentially detrimental effect on cognition. The participants from MS-SMART were the same as those described in the previous section, with the addition of one participant for whom automated scan segmentation had failed. Education status and depression scores were not available for the FutureMS cohort, so were not considered as predictors. Data on pre-morbid IQ and cognitive leisure activities were not available for either cohort.

7.4.2 Participant characteristics as predictors

7.4.2.1 Univariate relationships

The relationships of the non-imaging characteristics presented in Table 7.3 to SDMT performance were examined. Plots of individual non-imaging predictors against SDMT score are presented in Figure 7.5 for the combined cohort. As described in section 7.3.2.1, two participants in MS-SMART were prescribed medication with a potentially beneficial effect on cognition, with widely different SDMT scores. Again, to avoid misinterpretation of this predictor, these two participants were excluded from all predictive models using medication information.

Numerical predictors gave Spearman correlations with SDMT of $r = -0.53$ for age and $r = -0.51$ for disease duration (both with associated $p < 0.0001$). There was no evidence of a non-linear relationship from the scatterplots. There was no significant difference in SDMT performance between sexes ($p = 0.74$). Participants prescribed drugs with a potentially detrimental effect on cognition performed less well than participants not prescribed any drugs with potential effects on cognition ($p < 0.0001$).

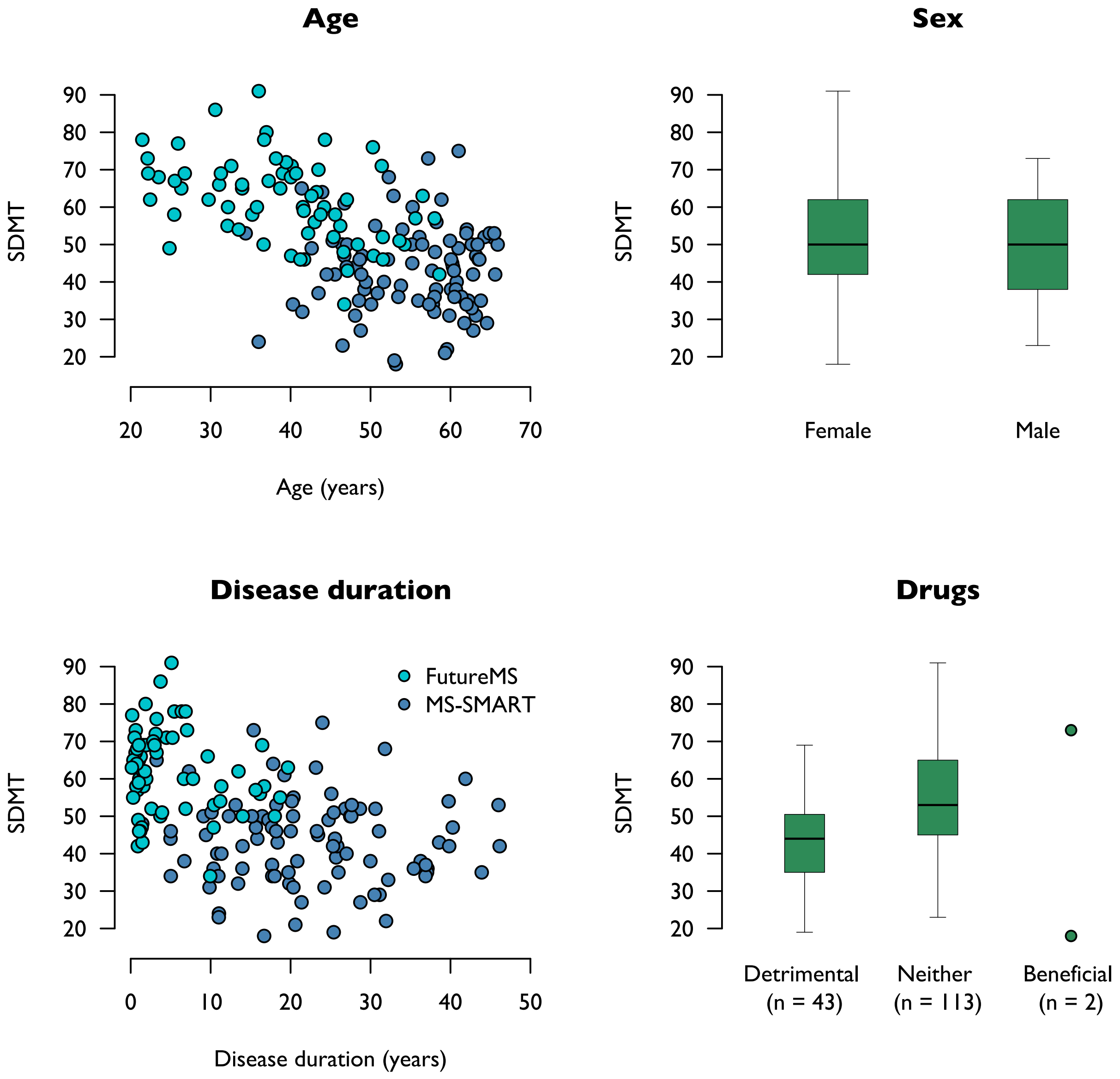


Figure 7.5: Plots of participant characteristics against SDMT for the combined MS-SMART and FutureMS cohorts (n = 158, for all except disease duration where n = 157).

	Combined cohort		FutureMS		MS-SMART	
	n	Summary	n	Summary	n	Summary
Age (years)	160	49.1 (41.4, 58.2)	67	40.1 (32.4, 45.9)	93	57.2 (49.0, 61.0)
Sex (F/M)	160	118/42	67	49/18	93	69/24
Disease duration (years)	159	13.5 (3.5, 46.2)	66	2.7 (1.0, 7.0)	93	20.4 (14.0, 28.7)
Drugs (B/D/N)	160	2/44/114	67	0/7/60	93	2/37/54
SDMT	158	51.1 ± 14.7	67	61.7 ± 11.2	91	43.2 ± 11.7
Global WMHs	159	6 (4, 8)	67	5 (3,6)	92	7 (5,10)
Regional WMHs	159	9 (5, 14)	67	7 (3,11)	92	10 (6,16)
Atrophy	159	1.5 (0.5, 2)	67	1 (0.5,1.5)	92	1.5 (1,2)
Cavitation (Y/N)	159	71/86	67	16/51	92	56/36
Cavitation (count)	159	0 (0, 3)	67	0 (0,0)	92	1.5 (0,6)
JC lesions (Y/N)	159	93/64	67	30/37	92	64/28
JC lesions (count)	159	1 (0, 2)	67	0 (0,2)	92	1 (0,3)
EPVS	159	6(5, 7)	67	5 (4,7)	92	6 (5,7)

Table 7.3: Summary statistics for the MS-SMART and FutureMS cohorts, used in predictive modelling based on visual ratings. All continuous/numerical variables are given as median (interquartile range), other than SDMT, given as mean ± SD. The columns headed 'n' indicate the number of subjects for whom that data was available. Medication is classified by whether participants were prescribed drugs with potentially beneficial (B) or detrimental (D) effects on cognition, or neither (N). JC lesions: juxtacortical/cortical lesions; EPVS: enlarged perivascular spaces; WMHs: white matter hyperintensities.

7.4.2.2 Construction of multiple linear regression model

As in Section 7.3.2.2, age, sex and being prescribed drugs with a potentially detrimental effect were considered important predictors of cognitive performance. Age and disease duration were significantly correlated ($r = 0.70$) and, as previously, disease duration was not used as an additional independent predictor.

7.4.2.3 Multiple linear regression model summary

A generalised linear model to predict SDMT score was constructed using the available non-imaging participant characteristics. A description of the model is presented in the first column of Table 7.4, which gave an adjusted R^2 of 0.34. This provided a significant improvement in fit to a model based on the intercept alone ($p < 0.0001$).

	Dependent variable:				
	SDMT				
	1	2	3	4	5
Constant	83.90	87.59	86.90	84.75	87.60
	$p < 0.01^{***}$	$p < 0.01^{***}$	$p < 0.01^{***}$	$p < 0.01^{***}$	$p < 0.01^{***}$
Age	−0.63	−0.51	−0.48	−0.48	−0.51
	$p < 0.01^{***}$	$p < 0.01^{***}$	$p < 0.0001^{***}$	$p < 0.01^{***}$	$p < 0.01^{***}$
Sex	−1.80	−2.17	−2.08	−1.54	−2.16
	$p = 0.40$	$p = 0.28$	$p = 0.30$	$p = 0.43$	$p = 0.29$
Detrimental	−6.97	−5.69	−5.67	−4.57	−5.67
drugs	$p = 0.002^{***}$	$p = 0.01^{***}$	$p = 0.01^{***}$	$p = 0.03^{**}$	$p = 0.01^{***}$
Global		−1.56	−1.31	−0.80	−1.51
WMHs		$p < 0.0001^{***}$	$p = 0.001^{***}$	$p = 0.04^{**}$	$p = 0.0001^{***}$
Atrophy			−1.71		
			$p = 0.22$		
Cavitation				−8.06	
				$p = 0.0003^{***}$	
JC lesions					−0.61
					$p = 0.77$
AIC	1207	1186	1187	1174	1188
BIC	1223	1204	1208	1196	1209
Observations	155	155	155	155	155
R^2	0.36	0.45	0.45	0.49	0.45
Adjusted R^2	0.34	0.43	0.43	0.48	0.43
Residual	11.66	10.86	10.84	10.42	10.89
Std. Error	(df = 151)	(df = 150)	(df = 149)	(df = 149)	(df = 149)
F Statistic	27.99^{***}	30.27^{***}	24.61^{***}	29.09^{***}	24.09^{***}
	(df = 3; 151)	(df = 4; 150)	(df = 5; 149)	(df = 5; 149)	(df = 5; 149)
Note:	*p<0.1; **p<0.05; ***p<0.01				

Table 7.4: Summary of linear models with SDMT as dependent variable and participant characteristics and imaging markers derived from visual rating as independent variables. Numbers shown in main table are model coefficients followed by associated p values. Cavitation and (juxta-)cortical (JC) lesions are binary predictors.

7.4.2.4 Interactions between variables

An alternative model to that described in Table 7.4 , column 1, using the same non-imaging predictors but allowing interactions between them, did not improve model fit given the number of parameters (AIC 1210.6, BIC 1238.0).

7.4.3 Imaging markers

7.4.3.1 Univariate relationships

Imaging-derived variables summarised in Table 7.3 were considered initially as univariate predictors of SDMT performance. These gave Spearman correlations with SDMT of $r = -0.47$ for global WMHs, $r = -0.39$ for regional WMHs, $r = -0.43$ for atrophy, $r = -0.58$ for the number of cavitated lesions, $r = -0.37$ for the number of JC lesions and $r = -0.13$ for enlarged perivascular spaces (EPVS). All of these except EPVS ($p = 0.10$) were significantly different from zero at the 1% level. For the binary variables, presence of cavitation ($p < 0.0001$) and JC lesions ($p < 0.001$), t-tests showed a significant difference in SDMT between groups.

Plots of individual imaging predictors against SDMT score are presented in Figures 7.6 and 7.7 for the combined cohort ($n = 157$). Cavitation and JC lesions are considered both by number and as binary features in Figure 7.7. The scatterplots in Figure 7.6 were open to multiple interpretations, but there was no strong evidence against the assumption of a linear relationship with SDMT. There appeared to be a more complex relationship between SDMT and the number of cavitated and JC lesions, as shown in Figure 7.7. There may be an excess of people with no or few cavitated/JC lesions, suggesting a more appropriate analysis would use binary classification.

7.4.3.2 Construction of multiple linear regression model

The global summary (‘Fazekas-style’) and summed regional WMH scores were significantly correlated ($r = 0.8$) and, as described previously, the global summary score was considered the more relevant and was used as an additional predictor to non-imaging characteristics to form a second model.

Enlarged perivascular spaces were not found to be directly related to cognitive scores in this cohort, so were not considered as part of a multiple regression model. All other imaging features of disease, including additional lesion characteristics, were assessed as independent predictors in addition to the global WMH score. The JC lesion score was significantly correlated with the global WMH score (Spearman

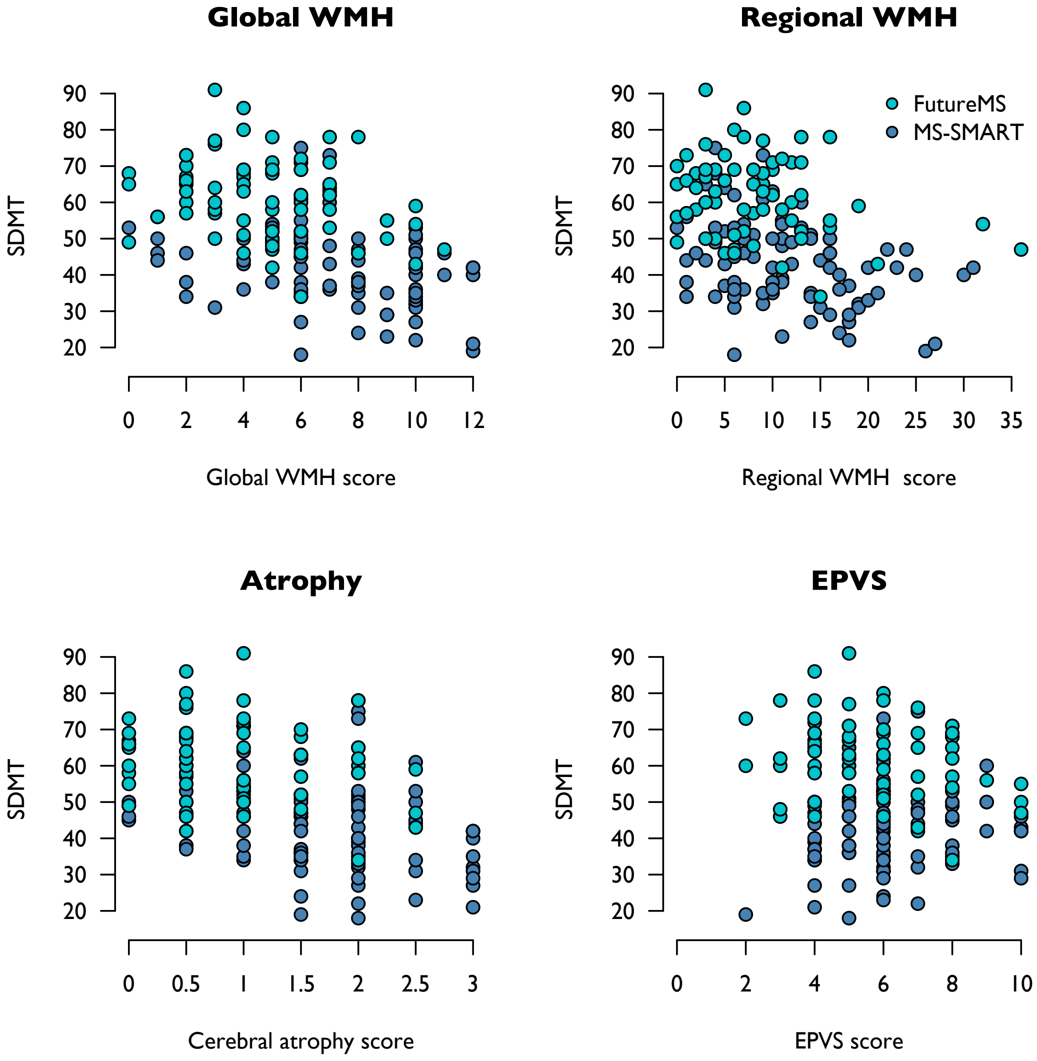


Figure 7.6: Plots of individual imaging predictors against SDMT (n = 157).

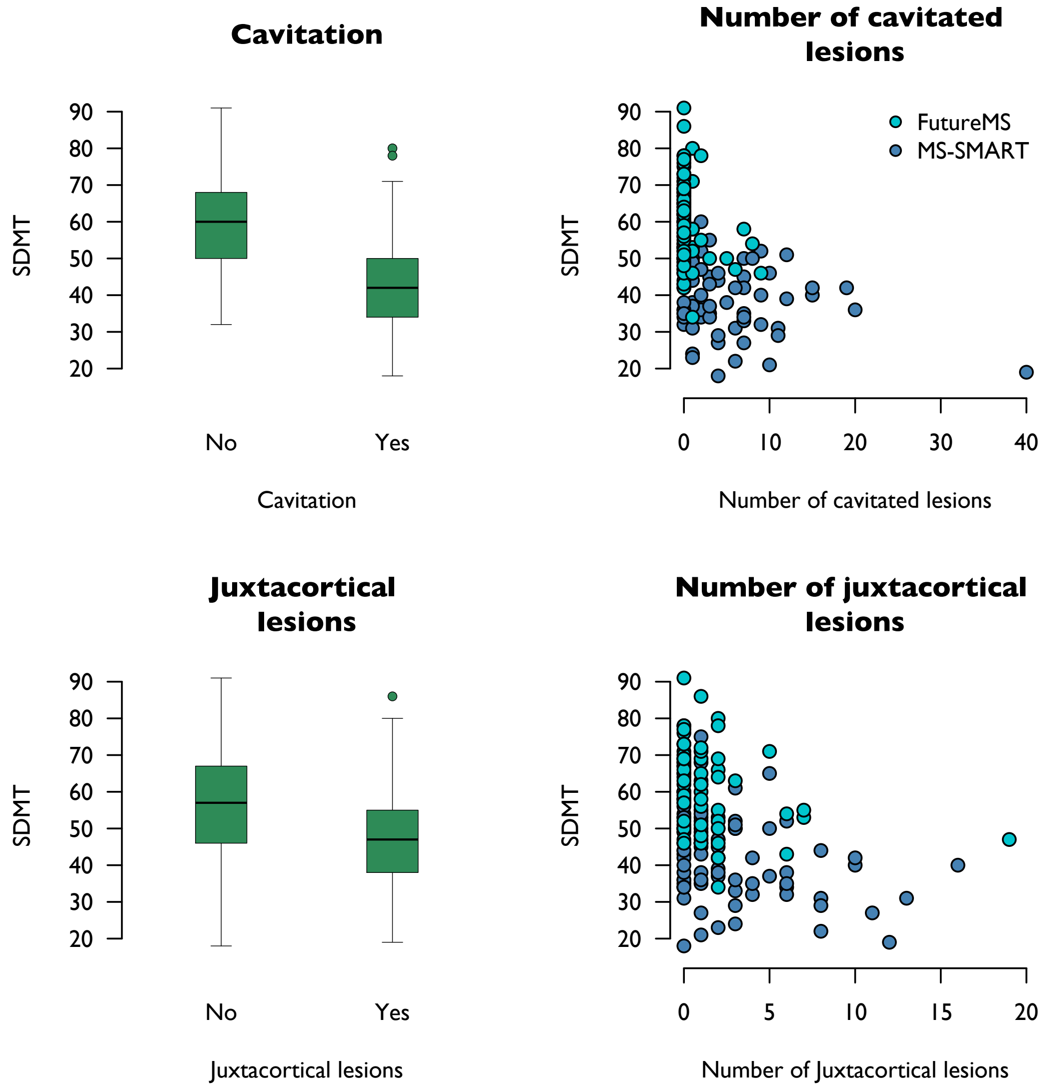


Figure 7.7: Plots of individual imaging predictors against SDMT (n = 157). JC lesions and cavitation are given both by count and as binary predictors.

$r = 0.55$). In order to avoid inflating a measure of white matter lesion burden, the involvement of (juxta-)cortical tissue was considered only as a binary predictor variable. In view of the possibly non-linear relationship between lesion cavitation and SDMT, this was also considered only as a binary predictor. The atrophy score was significantly correlated with the global WMH score ($r = 0.58$).

7.4.3.3 Multiple linear regression model summary

The addition of global summary WMH scores to a model based only on non-imaging characteristics resulted in a significant improvement in model fit ($p < 0.0001$).

Models with further imaging variables considered in turn as additions to the model are summarised in Table 7.4, columns 2 to 5. Introducing cerebral atrophy as a predictor resulted in a non-significant improvement in model fit ($p = 0.22$). Presence of lesion cavitation as a binary predictor produced a significant improvement ($p = 0.0002$) in model fit. The presence of JC lesions ($p = 0.77$) as a binary predictor did not improve model fit.

7.4.3.4 Interactions between variables

Considering a simple model with only the two imaging predictors found to be significant predictors in larger models (global WMH score and presence of cavitation), there was no improvement in model fit (increased AIC and BIC) when allowed an interaction term.

7.4.4 Model assumption checking

For the model described in column 4 of Table 7.4, a scatterplot of residuals against fitted values and a Q-Q plot are shown in Figure 7.8, with a minor suggestion that the fitted residuals deviate from normality at the tails of the distribution.

7.4.5 Linear modelling for separate cohorts

A significant correlation remained between global WMH score and SDMT when considered for each of the cohorts individually. This was stronger in the later stage cohort ($r = -0.39$) than the early stage cohort ($r = -0.24$) although this difference was not significant ($p = 0.30$).

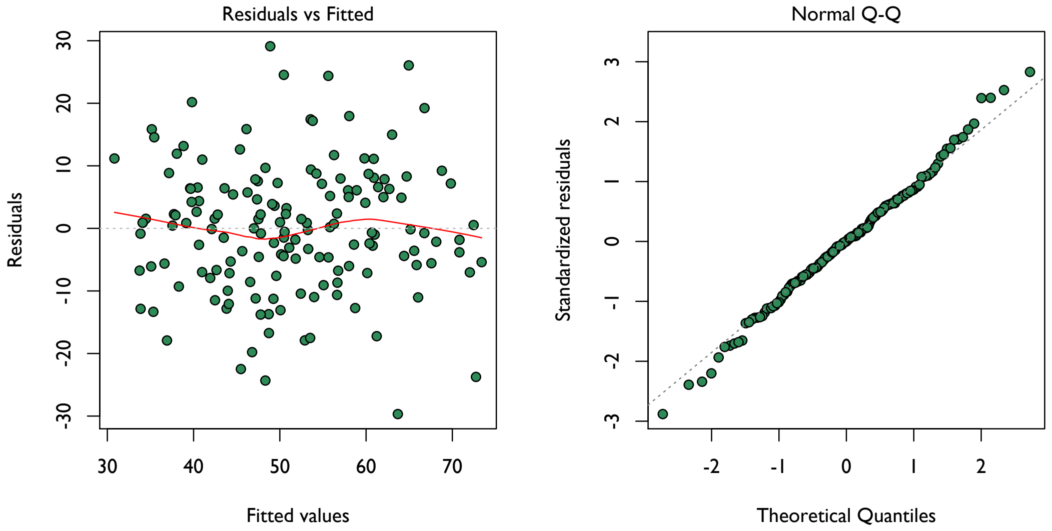


Figure 7.8: Left - plot of residuals against fitted values for linear model including participant characteristics, global WMH score and presence of cavitation as a binary predictor (column 4 of Table 7.4). Right - Q-Q plot of residuals for the same model.

The model structure described in column 4 of Table 7.4, was repeated separately for the two cohorts, to explore whether the predictor variables had different effects at different disease stages. The results are reported in Table 7.5.

In these smaller models the independent variable coefficients and their associated significance differed from the single cohort analysis. Age was a significant predictor of SDMT score in the younger, early stage disease cohort, but not in the older, later stage cohort. Medication was no longer significant for either cohort. For the imaging variables, the global WMH score only showed a trend towards being a significant predictor in the later stage group, whereas lesion cavitation remained the more significant predictor.

7.4.6 Sensitivity analyses

The addition of disease duration as a predictor variable, expressed either as a proportion of age or unadjusted, resulted in a significant improvement in model fit ($p < 0.01$) to a model based on non-imaging predictors alone. As an additional variable in a full model containing non-imaging and imaging predictors, disease duration showed a trend towards significance ($p = 0.07$) only when used in adjusted form.

	Dependent variable:	
	SDMT	
	FutureMS	MS-SMART
Constant	83.19	55.77
	p = 0.00***	p = 0.00***
Age	−0.44	−0.02
	p = 0.002***	p = 0.92
Sex	−1.44	−2.13
	p = 0.62	p = 0.39
Detrimental drugs	−3.35	−2.01
	p = 0.42	p = 0.37
Global WMHs	−0.43	−0.78
	p = 0.47	p = 0.09*
Cavitation	−5.57	−7.87
	p = 0.11	p = 0.004***
Observations	67	88
R^2	0.27	0.27
Adjusted R^2	0.22	0.22
Residual Std. Error	9.94 (df = 61)	9.81 (df = 82)
F Statistic	4.62*** (df = 5; 61)	5.94*** (df = 5; 82)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 7.5: Description of linear model from column 4, Table 7.4, recalculated for FutureMS and MS-SMART cohorts separately.

SDMT performance correlated with age only in the younger cohort, suggesting disease effects were involved rather than cognitive changes expected with ageing. Additionally, being prescribed drugs with a potentially detrimental effect on SDMT performance showed significant differences between the two cohorts (37/93 = 40% of MS-SMART; 7/67 = 10% of FutureMS) and may also have acted as a surrogate marker of disease duration and severity. To address these issues, the smaller models for the separate cohorts, as in Table 7.5, were repeated without including age as a predictor. In the FutureMS model, this resulted in a significantly worse overall fit ($p < 0.001$; adjusted $R^2 = 0.08$), but made little difference to overall fit of the MS-SMART model ($p = 0.91$; adjusted $R^2 = 0.23$). The significance levels for the remaining individual predictors in each model did not change substantially from those in the models containing age as a variable (as in Table 7.5).

7.5 Discussion

The results here confirm the relevance of white matter hyperintensities as a disease marker in multiple sclerosis. Routine imaging features contribute significantly to accurate prediction of cognitive status in people with MS, with convergent results achieved using two very different image analysis approaches. The relationship found here between cognitive outcome and both WMH volume and the semi-quantitative visual rating marker of WMH burden is stronger than in most published studies, but within the wider reported range of those using SDMT (see Chapter 3, Figure 3.8). In the model using quantitative imaging markers, WMH volume remained the most significant predictor of cognitive performance after controlling for other disease and non-disease-related variables.

There is a suggestion that this relationship is not the same at all levels of disease burden. The quantitative data shows little effect of WMH volume on cognitive performance at low levels. Similarly for the semi-quantitative data, a stronger correlation between WMH score and SDMT performance was found for the later stage participants, with greater disease burden, than those in the early disease stages. Although this was not a statistically significant finding, it is nevertheless in keeping with the reported literature (see Chapter 3, section 3.3.12.2) in which stronger effect sizes were found in cohorts with higher mean WMH volumes. Such observations are consistent with an interpretation that below a certain level of disease burden, neuronal adaptation, repair or redundancy may be able to compensate for the level of damage, with reducing capacity as the disease progresses.

The cohort in which volumetric markers were tested comprised people with later stage disease in whom accelerated atrophy was clearly present. Total brain volume, rather than WM volume, showed a trend towards significance as an additional predictor, suggesting the importance of grey matter or posterior fossa pathology. However, neither white matter nor total brain volume proved to be significant predictors in models containing other imaging-derived variables. This may reflect the dominant effect of WMHs, becoming clearer in larger cohorts, but both these measures of tissue volume may also be poor markers of neurodegeneration given their high covariance with ICV (r = 0.95/0.96). This allows the potential for more direct markers of widespread tissue integrity to contribute to measurement of the overall disease burden instead and is explored more in the following chapter. However the work here provides no strong evidence that pathology accumulation and its impact on SDMT is anything other than unidimensional.

The semi-quantitative visual rating markers provided an alternative source of information regarding the extent and degree of pathology. A simple binary marker of whether lesion cavitation was apparent proved to be a significant predictor of cognitive outcome and this remained significant when considered separately in both the individual cohorts. This could be interpreted as identifying people with more aggressive disease or those who are no longer able to adequately repair inflammatory damage. While related to the more widely used measure of T1-weighted (T1w) hypointense lesion volume, this assessment is not only simple and more selective, but also removes the confounding factor of acute inflammatory lesions also appearing hypointense on T1w imaging.

The results using the visual rating assessments must be interpreted with caution given their use of two different cohorts. The reasoning behind doing so was to use the largest possible group in which to test the assessment outcomes, however blinding as to cohort was not possible and the imaging sequences were not identical. Examining the cohorts separately suggests that different features may have greater relevance at different disease stages. In the cohort with RRMS, age was the sole significant predictor of cognitive outcome, again consistent with the proposal that initial compensation for accumulated brain pathology is possible. In the later stage group, non-imaging variables were less important than imaging ones, with presence of cavitation proving the only significant predictor at the 5% level. Although the global WMH score was a significant predictor in the larger, combined, cohort, this was not the case for the individual smaller cohorts. This may reflect a lack of sensitivity to smaller increments in disease burden related to the limited range of possible scores.

The relationship of atrophy to cognition in people with MS is established [102], although not using visual ratings, and here the atrophy scores were significantly correlated with SDMT performance. However they were not significant predictors in the multiple regression model, again possibly related to a dominant effect of WMHs and their high covariance. Although cortical lesions may be more important than white matter lesions in determining cognitive status [179], these two measures are also likely closely associated [113]. It was not possible to reliably separate purely cortical lesions on the routine sequences available for the visual rating study, and the presence of juxtacortical/cortical lesions was not found to be a significant independent predictor in the regression model of cognition. Enlarged perivascular spaces, although related in other populations [171], were not found to be directly related to cognitive scores in this cohort.

Limitations to the work here are acknowledged, not least the absence of more extensive information regarding both co-morbidities and non-disease factors which could influence cognition. The impact of a disease cannot be fully understood only in terms of the accumulated pathology, without consideration of wider factors relating to the person with the disease. This work was an opportunistic use of available data and it remains possible that imaging factors may be of lesser importance after controlling for other possible modifiers of the relationship between pathology and phenotype. This highlights a recurrent problem in the literature (see Chapter 3) with lack of consensus in recognising and recording factors that may be important and inconsistencies in data collection even for those that are accepted. No relationship between depression scores and SDMT performance was found for the cohort in which it could be tested, but this may have been influenced by the cut-off used in the study entrance criteria. Modafinil is used to treat fatigue in people with MS and has previously been associated with an improvement in SDMT performance [180]. Although only prescribed to two participants in the cohorts studied here, the results were ambiguous, possibly relating to timing of medication, which was not recorded.

As previously mentioned, a desire to use the largest possible dataset influenced the decision to test the visual assessments in a composite cohort. While recruitment strategy should not have influenced the findings from either group, the use of two cohorts with very different characteristics may have exaggerated the importance of any differences in imaging features. Relevant to the consideration of whether there exists a non-linear ‘threshold’ effect of WMHs, the use of a group with very little evidence of accumulated disease may have partly obscured identification of any threshold. Replication as well as further testing in groups at all stages of the disease is certainly required.

WMH burden, representing a history of focal neuroinflammatory damage, is clearly a useful marker of disease in people with MS, with these results suggesting that its importance may vary at different stages of the disease. Relevant to their known lack of pathological specificity, the visual rating results indicate that further simple information regarding the degree of damage represented by WMHs contributes usefully to an overall estimate of disease burden. Nevertheless, there remains unaccounted for variance in cognitive performance, which may require improved assessment of the neurodegenerative component of the disease. Tissue volume measures did not prove useful here as independent predictors, possibly related to the relatively small changes involved and their high covariance with intracranial volume. Quantitative markers related to tissue integrity in the normal-appearing white matter, such as those derived from diffusion imaging, may be more sensitive and these are explored in the following chapter.

Chapter 8

The relationship of quantitative measures of tract microstructure from diffusion tensor imaging to cognitive performance

8.1 Introduction

Variation in cognitive performance in people with multiple sclerosis (MS) is not fully accounted for by current measures of the white matter disease burden that are visible on routine structural imaging, as shown in previous chapters. This remains the case even after optimisation of white matter hyperintensity (WMH) quantification and adjustment for participant characteristics and modifiers of cognitive performance. It is however also recognised that diffuse pathological changes in white matter may not be associated with measurable changes on routine imaging sequences and that alternative techniques may be more sensitive to these processes. Diffusion tensor imaging (DTI) is an example of a quantitative magnetic resonance imaging (MRI) technique that has been shown to be sensitive to changes in the so-called 'normal-appearing' white matter (NAWM), leading to the testable hypothesis that DTI-derived biomarkers may result in stronger associations between imaging and clinical measures.

The work presented in this chapter addresses the question of how far any DTI-derived white matter metric can explain variance in cognitive performance and whether this is separate from information already available from routine structural imaging sequences. The NAWM and WMH compartments within the white matter are examined separately, with the prediction that any additional explanatory power will come from diffuse changes in the NAWM, since focal inflammatory disease burden is accounted for in WMH volume. However the possibility of DTI metrics providing additional information about the degree of tissue damage within the WMHs is also considered.

In this chapter the validity of different potential markers of tract health is first established, ensuring that they capture a range of values with the potential to be used in predictive models. Their covariance with disease and non-disease factors is then explored and the relationship of DTI-derived markers to a measure of processing speed, the Symbol Digit Modality Test (SDMT), is examined. Finally they are entered into predictive models, developed in the previous chapter, to determine whether they contribute additional information to markers already available.

Three specific hypotheses are tested: (1) that DTI metrics are a valid marker for distinguishing between WMHs and NAWM; (2) that DTI metrics provide stronger correlations with cognitive performance than volumetric measures from routine structural imaging; and (3) that in predicting cognitive performance from imaging and non-imaging features, DTI metrics increase explanatory power based on improved measurement of the neurodegenerative disease component.

8.2 Methods

8.2.1 Participants and imaging

Baseline scans from participants in the MS-SMART Advanced MRI substudy (described in Chapter 2, Section 2.1) were used in this work. The imaging and post-processing of diffusion data are described in Chapter 2, Sections 2.1.2 and 2.1.3. Averaged (mean) fractional anisotropy (FA) and mean diffusivity (MD) for the software-segmented WMH and NAWM tissue compartments, the peak width of skeletonised mean diffusivity (PSMD) and the tract-averaged FA and MD from the twelve automatically segmented major fasciculi of interest were all considered as potential predictors of cognitive status.

8.2.2 Validity of DTI-derived metrics

Plots of all potential DTI-derived predictors were considered for their validity in distinguishing between different tissues, including mean diffusion parameters derived from segmented tissue compartments and from each of the tracts extracted. Differences between DTI metrics for the segmented compartments were examined using Mann-Whitney U tests. The relationship of the novel marker PSMD to the average (mean) overall white matter MD was considered.

8.2.3 The univariate relationships between quantitative DTI measures of tract microstructure and cognitive performance

Spearman correlations between all individual predictors and SDMT scores were calculated. Plots of individual predictors against SDMT scores were examined for outliers and evidence of non-linear relationships. The participant characteristics of influential observations (individual participants) were investigated, as to whether they exhibited extreme values of other SDMT predictors.

8.2.4 Additional value of DTI-derived metrics to WMH burden in predictive models

8.2.4.1 Model construction

General linear modelling was used as in the previous chapter, with an assumption of normally distributed errors. The initial models developed in Chapter 7, based on non-imaging and routine structural imaging-derived volumetric predictors, were recalculated for the MS-SMART Advanced MRI substudy cohort only. Each DTI-derived predictor was then considered in turn as an additional independent variable in the model.

8.2.4.2 Assumption checking

Covariance of DTI metrics with age, sex and routine structural imaging-derived parameters, particularly WMH volume, was considered, in order to exclude collinearity. Correlations between all individual predictors were calculated. Cook's distances were examined for highly influential data points, with consideration of appropriate handling where necessary. For all models constructed, histograms of the residuals, Q-Q plots and plots of residuals against predicted values were assessed in order to check model assumptions and fit.

8.2.4.3 Model comparisons and fit

Parameters of goodness of fit were compared for each model, including values for adjusted R^2, Akaike information criteria (AIC) and Bayesian information criteria

(BIC). The Wald test was used to quantify statistical significance of additional model predictors.

8.3 Results

8.3.1 Participant data

Forty-three participants were recruited to the Advanced MRI substudy of MS-SMART and their baseline imaging data were used (see Chapter 2).

A score for the SDMT, taken as a marker of information processing speed, was available for all participants. Data on premorbid IQ and cognitive leisure activities were not available for this cohort. Data on educational status were unavailable for eight participants. Only one participant was prescribed medication with a potentially beneficial effect on cognitive performance (modafinil) and due to concerns over the timing of taking medication on the day of testing, as described in Chapter 7, this participant was excluded from analysis.

Tissue segmentation failed for one scan, due to issues with image registration. This participant was excluded from any further analysis based on tissue compartment metrics. Tract segmentation and skeletonisation using Tract-Based Spatial Statistics (TBSS, [75]) to derive PSMD were unaffected.

Following visual assessment of the tracts generated by probabilistic neighbourhood tractography (PNT; (http://www.tractor-mri.org.uk)) for quality control [MB], it was determined that anatomically acceptable representations of all tracts of interest were present in 34 subjects. Of the remaining nine scans, one tract in each could not be accurately extracted. Affected tracts were the splenium (n = 2), the left arcuate fasciculus (n = 2), the left (n = 4) and right (n = 1) corticospinal tracts. For these participants, their mean tract metrics were calculated using the remaining eleven tracts. The participants with missing tract data showed trends towards higher WMH volumes, higher tract MD and lower FA (p = 0.06 to 0.11), all suggesting a higher level of disease burden. Group maps of all tracts extracted are presented in Figure 8.1.

Descriptive statistics for this cohort are presented in Table 8.1, including participant characteristics and volumetric imaging markers acquired from both routine structural sequences and DTI.

	n	Summary figures
Age (years)	43	58.0 (49.9, 62.0)
Sex (F/M)	43	30/13
Disease duration (years)	43	23.4 (15.6, 27.3)
BDI	43	6 (5, 11.5)
Drugs (Beneficial/Detrimental/Neither)	43	1/14/28
Education $\leq 12/> 12$ years	35	13/22
SDMT (mean ± SD)	43	43.5 ± 12.1
ICV (ml)	42	1309 (1265, 1395)
Brain volume (ml)	42	1154 (1066, 1205)
WM volume (ml)	42	429.3 (403.4, 451.4)
WMH volume (ml)	42	31.1 (23.7, 44.4)
Mean WMH MD (μm^{2}s^{-1})	42	1122 (1055, 1191)
Mean WMH FA	42	0.309 (0.296, 0.325)
Mean NAWM MD (μm^{2}s^{-1})	42	768.8 (753.3, 802.8)
Mean NAWM FA	42	0.340 (0.308, 0.351)
Mean tract MD (μm^{2}s^{-1})	43	861.6 (833.0, 930.0)
Mean tract FA	43	0.442 (0.415, 0.464)
PSMD (μm^{2}s^{-1})	43	351.8 (309.8, 398.7)

Table 8.1: Clinical and imaging features of participants in the Advanced MRI substudy of MS-SMART cohort. All continuous/numerical variables are given as median (interquartile range) other than for SDMT. The second column 'n' indicates the number of subjects for whom that data was available. BDI: Beck Depression Index; FA: fractional anisotropy; ICV: intracranial volume; MD: mean diffusivity; PSMD: peak width of skeletonised mean diffusivity; SDMT: Symbol Digit Modality Test; NAWM: normal-appearing white matter; WM: white matter; WMH: white matter hyperintensity.

8.3.2 Validity of acquired diffusion metrics

8.3.2.1 Compartment-averaged diffusion metrics

Distributions of MD and FA for all segmented tissue compartments are shown in the boxplots of Figures 8.2 and 8.3. All differences in location and spread of values between compartments were consistent with predictions based on their

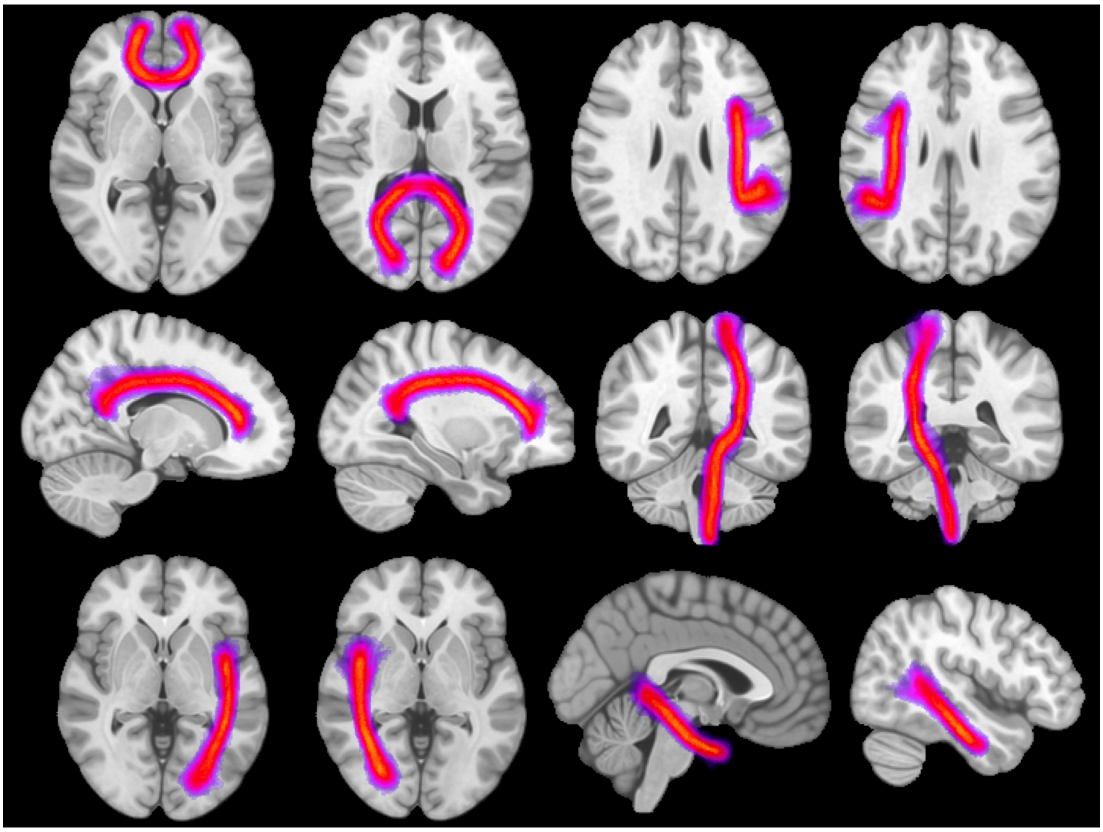


Figure 8.1: Group maps of the segmented fasciculi of interest projected onto Montreal Neurological Institute standard space T1w volume, showing consistency of tract segmentation across the cohort. Top row: genu, splenium, arcuate fasciculi (bilateral); middle row: bilateral dorsal cingulate and corticospinal tracts; bottom row: bilateral inferior longitudinal fasciculi and ventral cingulate.

known tissue characteristics, for instance cerebrospinal fluid (CSF) showed the highest water diffusivity and brainstem showed the highest directional coherence, supporting the validity of the segmentation.

Tissue labelled as abnormal (WMH) by the segmentation was associated with increased MD and decreased FA (both $p = 0.001$) values when compared with the NAWM compartment.

8.3.2.2 Peak width of skeletonised mean diffusivity

Peak width of skeletonised mean diffusivity is derived from skeletonised white matter and summarises the spread of MD values. The histogram in Figure 8.4 shows the distribution of PSMD in the MS-SMART Advanced MRI cohort, highlighting that the majority of values lie between 200 and 400 $\mu m^2 s^{-1}$ (n

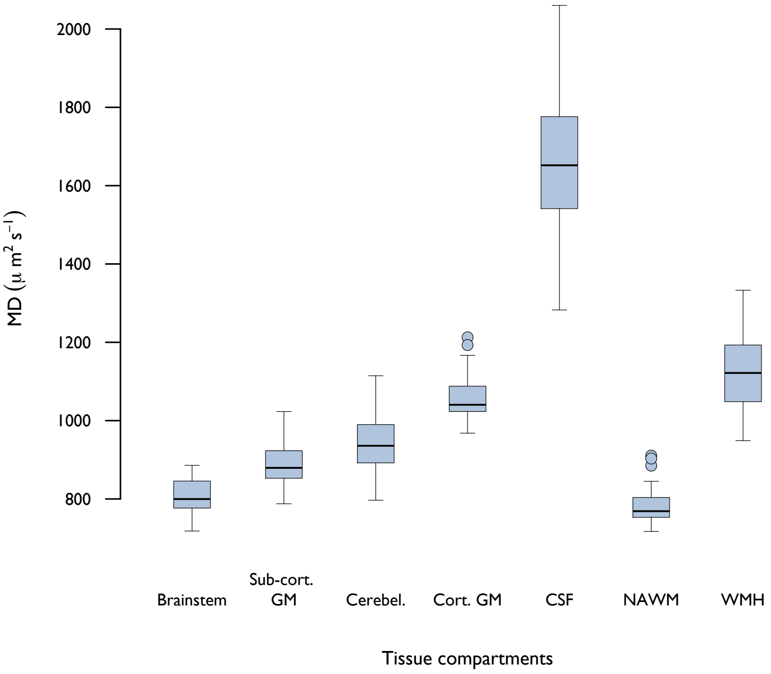


Figure 8.2: Mean diffusivity (MD) in segmented brain compartments (n = 42). From left to right, the tissue compartments are: brainstem, subcortical grey matter (GM), cerebellum, cortical GM, cerebrospinal fluid (CSF), normal-appearing white matter (NAWM) and white matter hyperintensities (WMH).

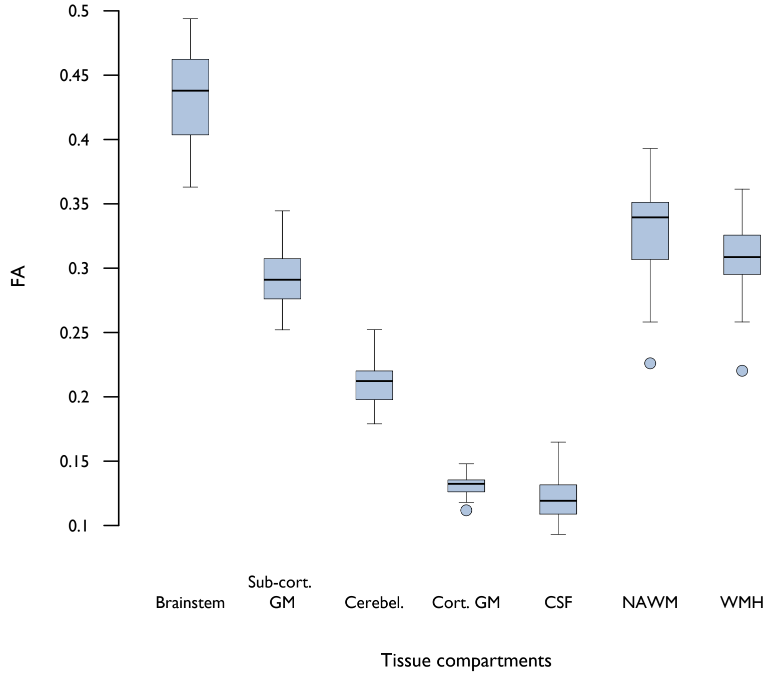


Figure 8.3: Fractional anisotropy (FA) in segmented brain compartments (n = 42). Tissue compartments as per Figure 8.2.

= 33/43, 77%). Although true normative data does not yet exist for PSMD, these values correspond approximately to the ranges found by Baykara et al [92] in healthy older populations.

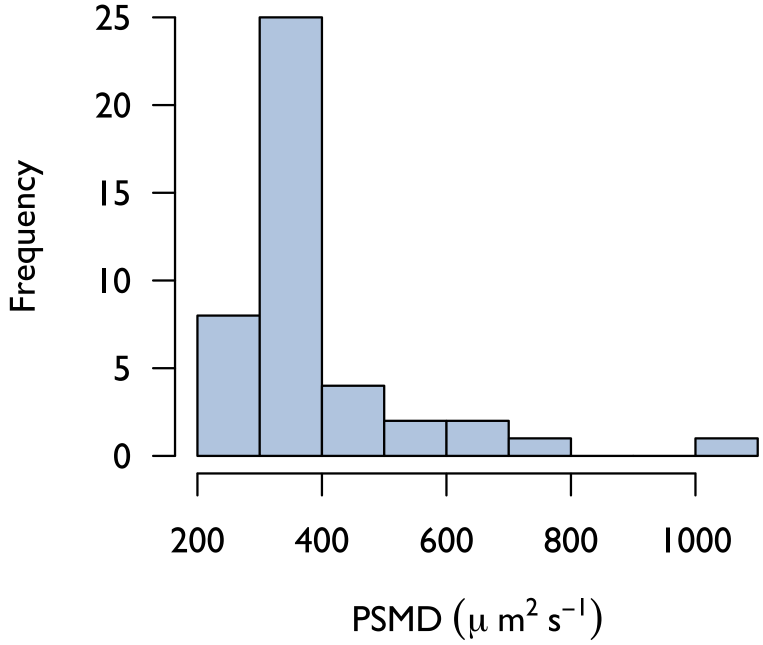


Figure 8.4: Distribution of PSMD (n = 43).

Figure 8.5 shows PSMD plotted against the mean MD for the entire white matter compartment (encompassing both WMHs and NAWM). Higher values for white matter mean MD are clearly associated with a higher PSMD (Spearman correlation: r = 0.77), i.e. a greater range of voxel MD values. The scatterplot suggests a possible non-linear relationship, with a clearly positive gradient only at higher levels, indicating that the two metrics are not supplying duplicate information.

A similar relationship is shown in Figure 8.6 where PSMD is plotted against WMH volume, with a positive, approximately linear relationship above a WMH volume threshold of around 30ml. This supports an interpretation that an increased inflammatory disease burden is associated with diffuse white matter abnormality.

8.3.2.3 Within tract diffusion metrics

Boxplots summarising the mean MD and FA in all tracts extracted are presented in Figures 8.7 and 8.8. The spread of values for each tract is partly due to its size, as is apparent with the relatively small ventral cingulate; due to its

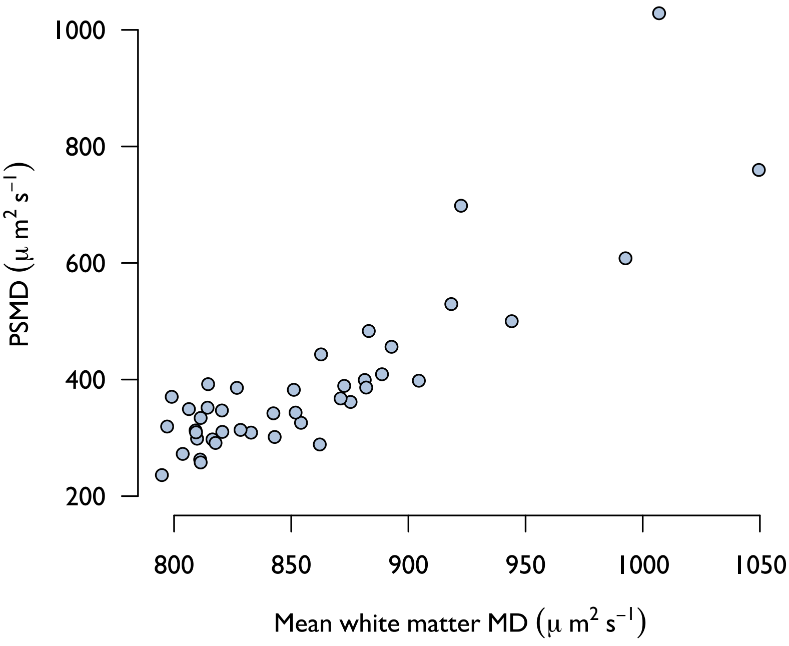


Figure 8.5: Scatterplot of PSMD against mean white matter MD (n = 42). The Spearman correlation was r = 0.77.

shape this tract has to be extracted separately from the rest of the much larger cingulate. The large spread and generally higher MD values for the splenium are related to higher inter-subject variation in anatomy and contamination from CSF proximity.

Mean tract FA was higher for each participant than the mean FA for the NAWM compartment, as shown in Figure 8.9, despite segmented tracts including tissue from both WMH and NAWM compartments. This confirmed that the most directionally coherent tissue had been extracted using the tractography approach.

8.3.3 Covariance of diffusion metrics with other variables

8.3.3.1 Age and Sex

A comparison of mean diffusion metrics for male and female participants is given in Table 8.2. The median age (years) of the female participants was 58.1 (IQR: 48.8, 61.8) and for the male participants 57.7 (53.5, 62.0). Significant ($p < 0.05$) sex differences were found for NAWM and tract-averaged MD and FA. Spearman correlations for diffusion metrics and participant age were all non-significant, with the exception of WMH FA ($r = -0.32$, $p = 0.04$).

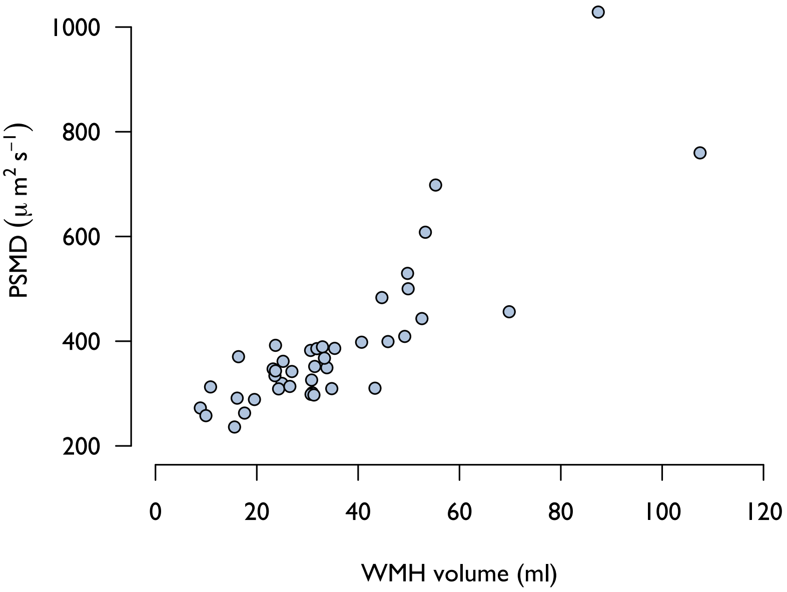


Figure 8.6: Scatterplot of PSMD against WMH volume (n = 42). The Spearman correlation was r = 0.80.

DTI metric	n	Median (female)	Median (male)	p	r (Age)	r (ICV)	r (WMH*)	r (WM*)
NAWM MD (μm^{2}s^{-1})	42	786	759	**0.02**	0.01	**-0.33**	**0.56**	−0.08
NAWM FA	42	0.331	0.348	**0.03**	0.08	**0.37**	**−0.73**	0.07
WMH MD (μm^{2}s^{-1})	42	1144	1084	0.06	0.14	−0.20	**0.61**	0.11
WMH FA	42	0.304	0.316	0.28	**−0.32**	0.13	−0.27	**−0.32**
PSMD (μm^{2}s^{-1})	43	369	334	0.10	−0.11	−0.29	**0.80**	−0.13
Mean tract MD (μm^{2}s^{-1})	43	892	834	**0.001**	0.09	−0.23	**0.64**	−0.07
Mean tract FA	43	0.425	0.451	**0.009**	0.03	**0.32**	**−0.73**	0.01

Table 8.2: Median diffusion metrics by sex, with associated p-value from Mann-Whitney U test; Spearman correlations with age, ICV, white matter (WM) volume and WMH volume. Asterisks (*) indicate that WM and WMH volumes were adjusted for ICV. Sex differences and correlations significant at the 5% level are highlighted in bold. Abbreviations as per Table 8.1.

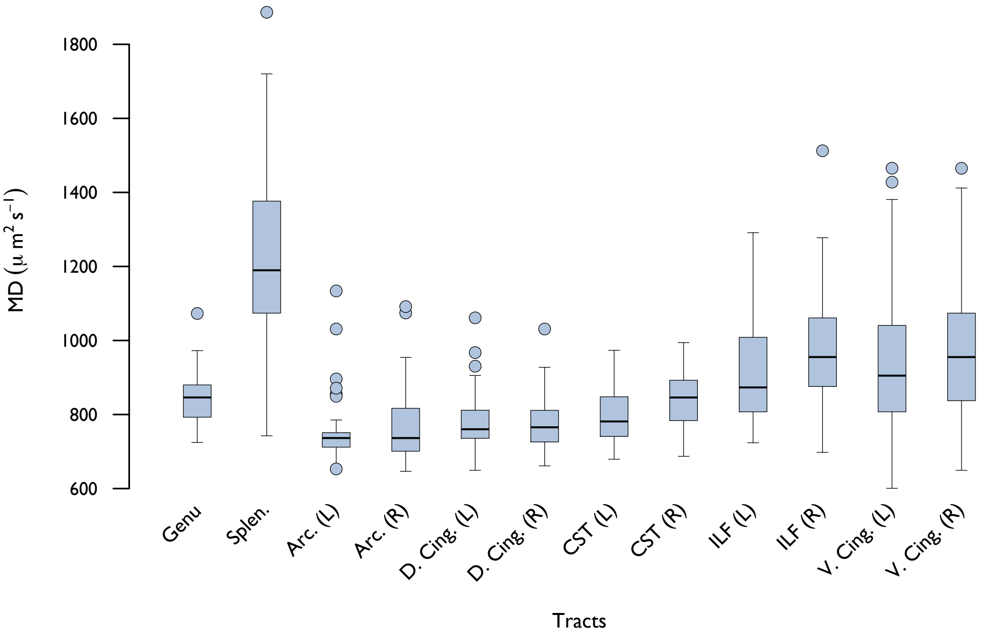


Figure 8.7: Boxplot of cohort MD for each of the 12 segmented tracts (total n = 43). From left to right the tracts are genu, splenium, arcuate fasciculi, dorsal cingulate, corticospinal tracts, inferior longitudinal fasciculi and ventral cingulate. All tracts are bilateral other than for the genu and splenium of the corpus callosum.

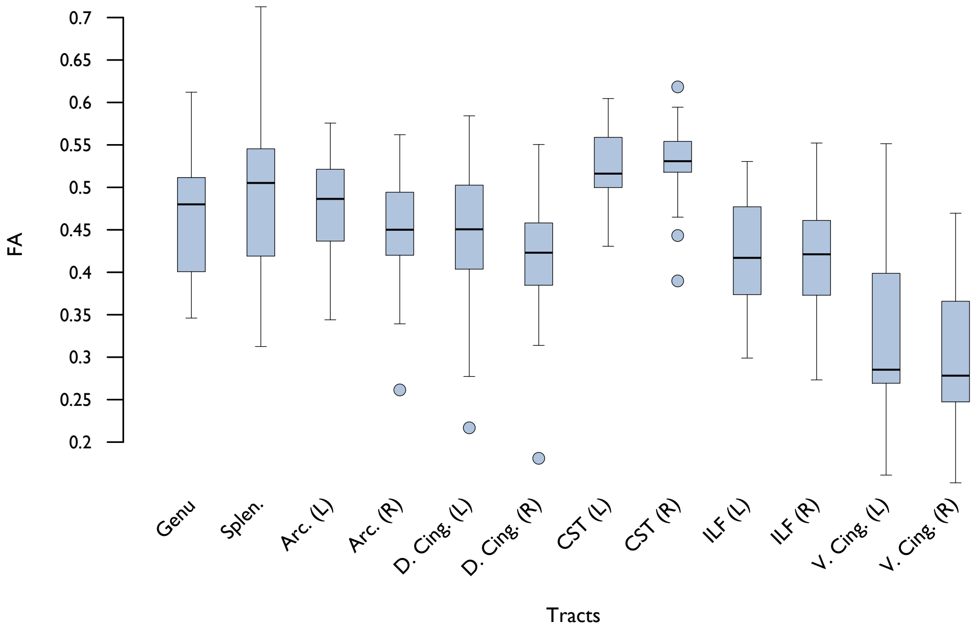


Figure 8.8: Boxplot of cohort FA for each of the 12 segmented tracts (total n = 43). The tracts are the same as those in Figure 8.7

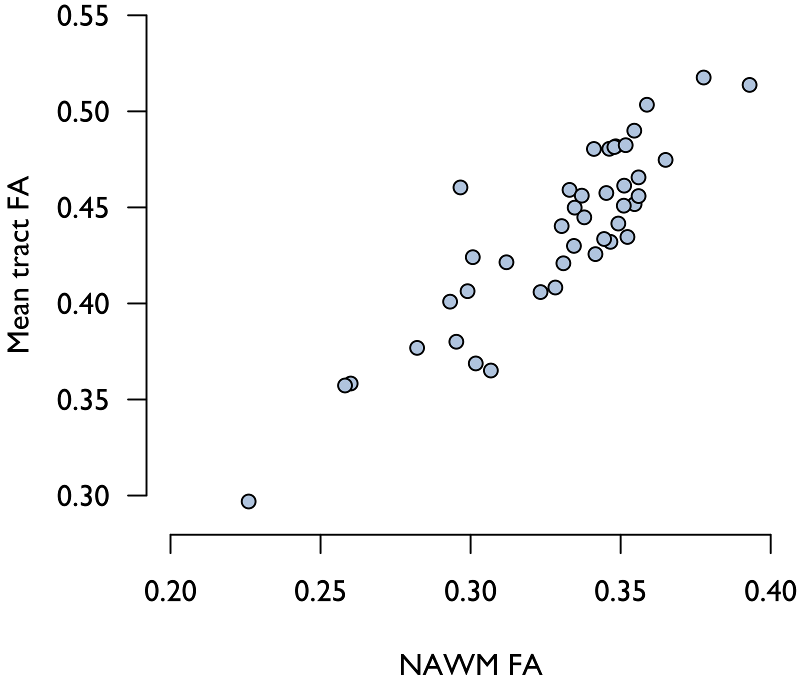


Figure 8.9: Scatterplot of mean FA for the extracted tracts against the corresponding NAWM compartment value. The Spearman correlation was r = 0.82.

8.3.3.2 Covariance with routine structural imaging markers

Intracranial volume also showed a significant correlation with several DTI-derived markers, particular NAWM MD and FA. This is a recognised feature of DTI relating to partial volume effects and the use of fixed voxel sizes [181], and may confound measurement in individuals with smaller ICV. In addition, all DTI-derived markers relating to white matter microstructure were significantly ($p < 0.001$) correlated with WMH volume, with the exception of WMH FA ($p = 0.07$). This suggests a possible unidimensional pathological pathway linking inflammation and neurodegeneration. Spearman correlation coefficients are presented in Table 8.2 and all correlations significant at the 5% level are highlighted in bold.

8.3.4 The direct relationship of quantitative measures of tract microstructure from DTI to cognitive performance

8.3.4.1 Compartment-averaged diffusion metrics

The relationships of the mean FA and MD within the NAWM and WMH tissue compartments to the SDMT score are shown in the scatterplots of Figure 8.10. These were all in the expected direction, with higher water diffusivity and lower diffusion anisotropy being associated with lower test scores, however no (Spearman) correlation was statistically significant from zero ($p > 0.1$).

8.3.4.2 Peak width of skeletonised diffusivity

The overall relationship of PSMD and SDMT score was in the expected direction, with a Spearman correlation of $r = -0.34$ ($p = 0.03$). This was similar to that found for WMH volume and SDMT in this population ($r = -0.33$; $p = 0.03$). However there appeared little relationship with SDMT score at lower values of PSMD, up to around 400 $\mu\text{m}^2\text{s}^{-1}$, suggesting that the overall association may be driven by the higher values. The participant characteristics of those individuals with high values of PSMD were examined with no notable differences seen. A scatterplot of PSMD against SDMT is shown in Figure 8.11.

8.3.4.3 Within tract diffusion metrics

Scatterplots of the mean FA and MD for all tracts against SDMT are shown in Figure 8.12. The Spearman correlations were $r = 0.37$ ($p = 0.01$) for FA and $r = -0.23$ ($p = 0.14$) for MD.

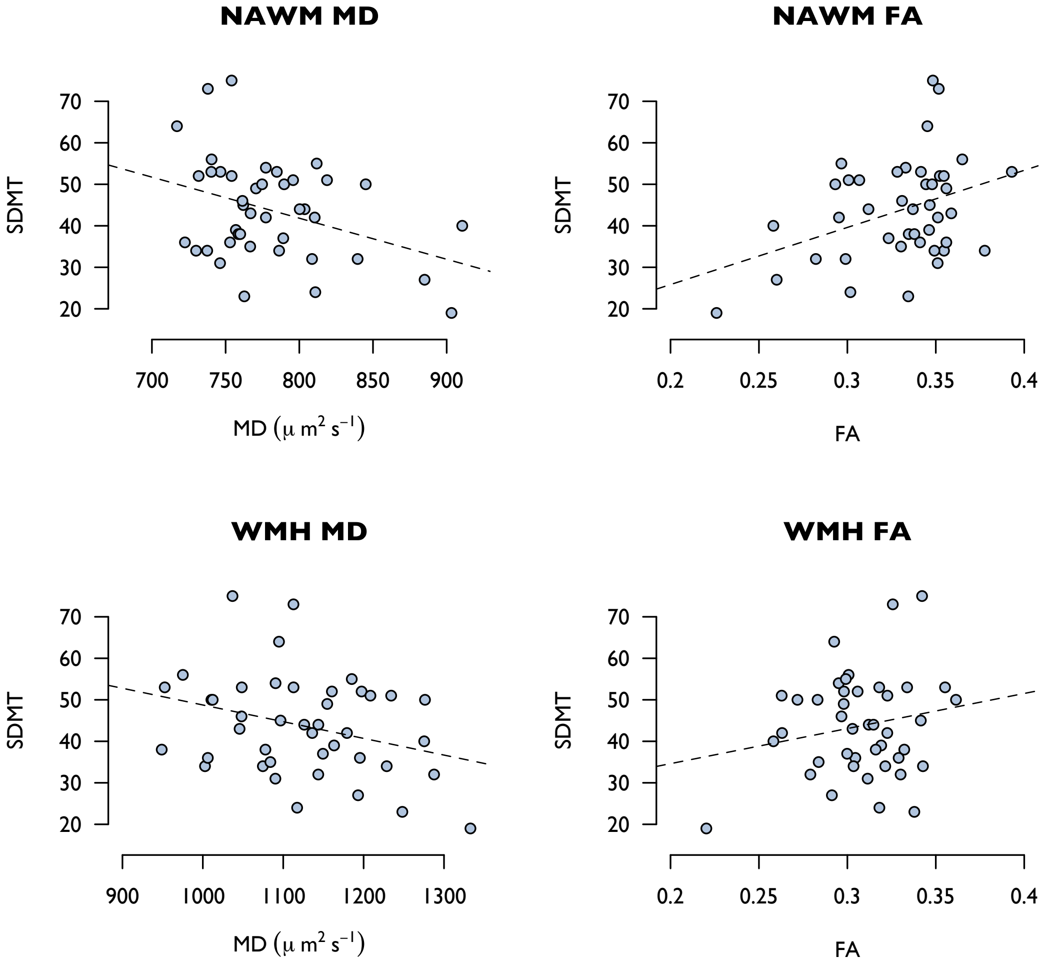


Figure 8.10: Plots of FA and MD against SDMT for NAWM and WMH tissue compartments ($n = 42$), annotated with lines of best fit. No correlation was significantly different from zero.

8.3.5 Addition of DTI-derived metrics to lesion burden in predictive models

Following recalculation in this smaller cohort of the linear model developed in Chapter Seven using participant characteristics and routine imaging markers to predict SDMT score, WMH volume was found to be the only significant predictor ($p = 0.01$) with sex being the next most significant ($p = 0.05$). This model is summarised in column 1 of Table 8.3. The addition in turn (columns 2 to 6) of each diffusion measure listed in Table 8.2, excepting those related to WMHs, did not lead to any overall improvement in model fit ($p > 0.2$).

	Dependent variable:					
	SDMT					
	1	2	3	4	5	6
Constant	−3.44	0.87	13.45	4.15	−11.64	−22.58
	p = 0.92	p = 0.99	p = 0.76	p = 0.91	p = 0.82	p = 0.60
Age (years)	0.06	0.06	0.06	0.07	0.06	0.05
	p = 0.80	p = 0.80	p = 0.81	p = 0.79	p = 0.81	p = 0.83
Sex	−10.42	−10.43	−10.69	−9.93	−10.21	−10.48
	p = 0.06*	p = 0.06*	p = 0.05**	p = 0.07*	p = 0.07*	p = 0.06*
Educ. (over	1.19	1.20	1.51	1.57	1.09	0.98
12 years)	p = 0.78	p = 0.78	p = 0.72	p = 0.71	p = 0.80	p = 0.82
Detrimental	−4.79	−4.64	−5.87	−4.67	−5.23	−3.93
drugs	p = 0.25	p = 0.33	p = 0.21	p = 0.27	p = 0.27	p = 0.37
ICV (ml)	0.04	0.04	0.05	0.04	0.04	0.04
	p = 0.10*	p = 0.12	p = 0.09*	p = 0.16	p = 0.10*	p = 0.15
WMHV (ml)	−0.31	−0.30	−0.40	−0.22	−0.33	−0.21
	p = 0.02**	p = 0.05**	p = 0.05**	p = 0.29	p = 0.07*	p = 0.24
NAWM MD		−0.01				
(μm^{2}s^{-1})		p = 0.95				
NAWM FA			−71.44			
			p = 0.55			
PSMD				−0.01		
(μm^{2}s^{-1})				p = 0.59		
Mean tract MD					0.01	
(μm^{2}s^{-1})					p = 0.84	
Mean tract FA						49.99
						p = 0.50
AIC	256.0	258.0	257.5	257.6	257.9	257.4
BIC	268.0	271.5	271.0	271.1	271.4	270.9
Observations	33	33	33	33	33	33
R^{2}	0.40	0.40	0.40	0.40	0.40	0.41
Adjusted R^{2}	0.26	0.23	0.24	0.24	0.23	0.24
Residual	10.35	10.55	10.47	10.49	10.54	10.45
Std. Error	(df = 26)	(df = 25)	(df = 25)	(df = 25)	(df = 25)	(df = 25)
F Statistic	2.83**	2.33*	2.42**	2.41**	2.34*	2.44**
	(df = 6; 26)	(df = 7; 25)	(df = 7; 25)	(df = 7; 25)	(df = 7; 25)	(df = 7; 25)
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table 8.3: Summary of linear models with SDMT as dependent variable and participant characteristics and imaging markers as independent variables. Numbers shown in the main table are model coefficients followed by associated p values. Predictor abbreviations as per Table 8.1.

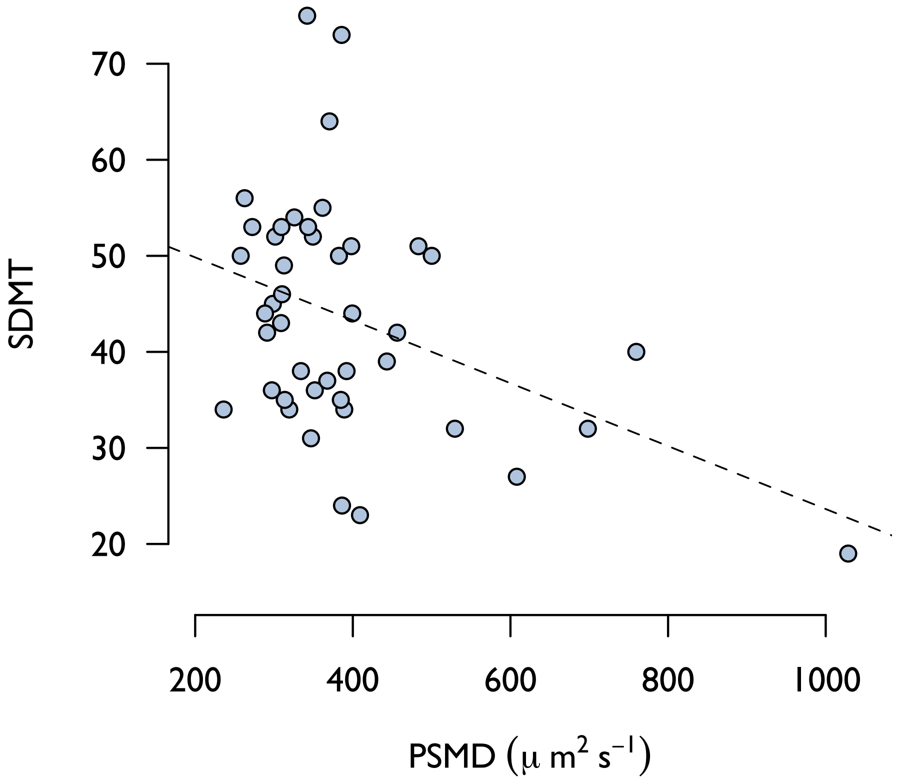


Figure 8.11: Scatterplot of PSMD against SDMT ($n = 43$), annotated with line of best fit. The Spearman correlation was $r = -0.33$.

8.4 Discussion

Using diffusion tensor imaging, it has been possible to extract valid markers of tissue microstructure in the white matter of people with multiple sclerosis. Differences found between segmented WMH and NAWM compartments were in the expected direction, supporting the tissue segmentation and the extraction of more diseased tissue in the WMH compartment. The relationship between DTI-derived markers and age found in healthy populations [182] was not found in this population, although the age range (34 to 65) was one in which this would be expected. Variation in the degree of pathological changes appears to have overpowered the normal changes from healthy ageing.

The majority of the water diffusion measures examined here were found to relate to the SDMT, a marker of cognitive ability. However they were all highly

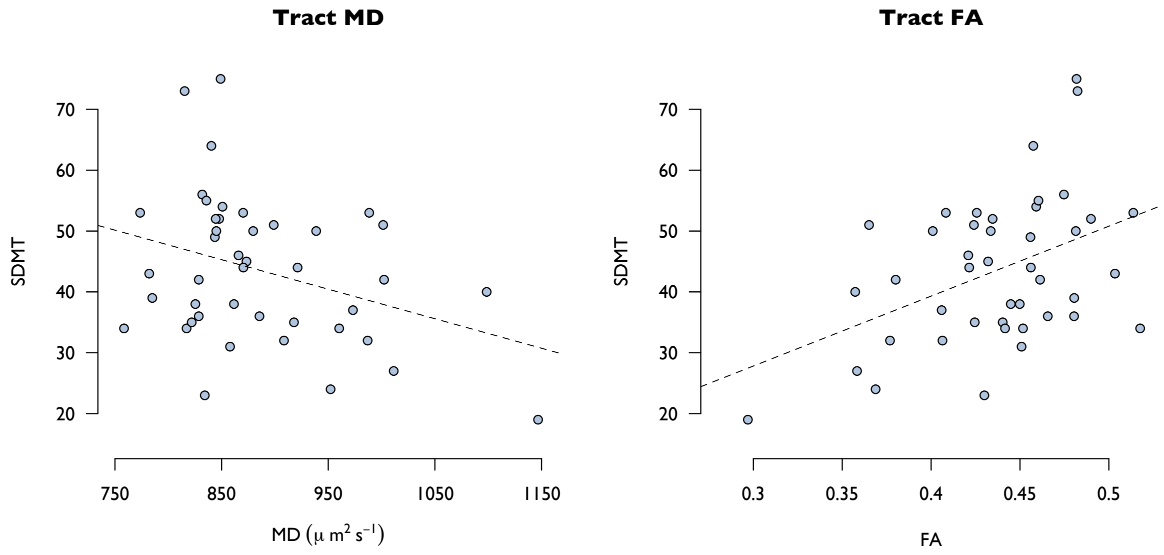


Figure 8.12: Scatterplots of mean tract FA and MD against SDMT ($n = 43$) annotated with lines of best fit.

covariant with WMH volume and it was not possible to demonstrate that DTI provided any additional predictive value once WMH volume had been accounted for. While it is tempting to interpret changes measured in normal-appearing tissue as representing a distinct, neurodegenerative, disease component, it has not been possible here to separate this from the effect of the inflammatory pathology. Further work in larger cohorts may be able to clarify this. Based on the results here, there is no evidence to reject a unidimensional disease model in which focal inflammatory damage leads to widespread abnormalities in white matter integrity.

Where the relationship appears more complex is in the suggestion of a threshold effect seen with PSMD, above which inflammatory disease burden has a deleterious effect on white matter integrity. Baykara et al [92] found similar ranges of PSMD in three separate healthy populations and no difference in PSMD between healthy controls and people with mild cognitive impairment and low WMH loads. The results in this chapter, implementing PSMD measurement for the first time in people with MS, appear remarkably similar. At levels seen in healthy populations, PSMD appeared unrelated to SDMT. Overall this new marker showed a stronger relationship to SDMT than any other measure derived from mean diffusivity values. Further work should include testing the relationship of PSMD to the more widely used marker of mean skeletonised MD and comparing the strength of their correlation with cognition.

The strongest overall correlation with SDMT for any diffusion metric tested was found for mean tract FA. Focussing on the integrity of major tracts and so minimising the confounding effect of crossing fibres may maximise the chances of uncovering a true relationship between tract integrity and cognition. One limitation to this work was using values averaged across all tracts, with equal weighting to large and small tracts, including those known to have higher variability. The tracts which could not be extracted occurred in individuals with a more severe disease burden, potentially attenuating a detectable relationship to the phenotype. A method of dimension reduction, for instance using principal component analysis to extract a general factor of tract integrity [183] might help with this issue, but the number of subjects was prohibitive in this case. While it would have been possible to select only those tracts thought likely to have cognitive functions, tractography was used here to provide a marker of global white matter microstructure and examine its relation to a marker of distributed cognition, so an inclusive approach seemed reasonable. The tracts extracted represented a wide range of projections, incorporating commissural and projection fibres, previously accurately and reproducibly segmented using PNT.

A further limitation to this work is the small study size. Interpretations discussed here will need confirmation in larger cohorts, where it may be possible to separate out cognitive effects related to both the WMH and NAWM tissue compartments. The lack of healthy control data also limits interpretation of some findings, although regression modelling was used to control for as many non-disease factors as possible. Bias may also have been introduced during study recruitment, as participants undergoing the advanced imaging protocol in MS-SMART were a self-selecting subgroup.

It is often assumed that using advanced imaging markers to quantify pathological changes in the normal-appearing white matter will provide a more accurate assessment of the total burden of disease. However these results show that using DTI in a population with fairly advanced disease has not supplied the missing link in the cognitive clinicoradiological paradox. In a population in which chronic neurodegeneration is expected to be the predominant active disease component, it has not been possible to separate out its effect from already available measures of the inflammatory disease.

Chapter 9

Discussion

The clinicoradiological paradox persists. Using optimised imaging measures of disease burden in multiple sclerosis has demonstrated a stronger relationship between white matter hyperintensities and cognitive performance than most reported in the literature, but this can still only partially account for the observed variation. Clinically relevant tests of brain function and established neuroimaging techniques cannot be made to agree, suggesting a flaw in our methods or the questions we ask of them. However it must be remembered that given the many factors, both known and unknown, affecting any psychological test and thus attenuating associations with other disease markers, the correlations found here are within the upper third of those published in psychological research [184].

Before considering the potential causes for the remaining mismatch and future directions for research, we should consider whether in fact this matters. When straightforward cognitive tests can be performed in the clinic, why should prediction of function from imaging be useful?

The great unmet need in multiple sclerosis (MS) is for truly disease-modifying drugs - those with a proven impact on longterm clinically relevant outcomes. A need to understand the pathology underlying these outcomes is clear. Currently available drugs act to reduce the shorter term impact of neuroinflammation, while carrying the risks associated with manipulation of the immune system. The development of new drugs, from initial laboratory investigations and animal studies through to large cohort trials, is time and resource intensive. The heterogeneity of MS and the difficulty in predicting individual disease outcomes is well known. For a disease with substantial heterogeneity in disease course and outcome, large numbers of people must be monitored for long periods to convincingly demonstrate success and imaging-derived surrogate outcomes are widely used to make this process more efficient. In clinical practice, with the aim of minimising delays in starting appropriate and effective treatments and in discontinuing non-effective treatments, the availability of sensitive and objective biomarkers of developing pathology becomes increasingly important.

The majority of new medications are specifically targeted against pathophysiological processes known to be relevant in MS, such as neuroinflammation, myelin repair and neurodegeneration. The need for reliable biomarkers here is twofold; firstly evidence of the effect of any intervention on these processes, in order to confirm our understanding of its action, and secondly an understanding of the relationship between developing pathology and the associated clinical phenotype. The clinicoradiological paradox exposes important gaps in our understanding of where the relevant pathology lies and without knowledge of the biological processes through which drugs act, we are restricted to a passive ‘watch-and-wait’ approach to drug development, waiting for clinically-measurable long term outcomes in response to treatments, effectively a ‘black box’ approach to neurology. With an ultimate goal of targeting drugs against pathology, the need for relevant biomarkers is clear.

With wider relevance, beyond the goal of effective treatments for MS, the gaps in our knowledge relating measurable pathology to clinical outcomes expose significant limitations in our understanding of brain function, both in health and disease, and specifically the anatomical and physiological basis for cognition. Decades of research in neuropsychology have provided evidence linking specific functions to their associated brain regions, but the continued limitations in our ability to predict clinical outcomes from neuroimaging demonstrate how much of the brain’s complexity yet remains incompletely understood or beyond the reach of current investigative tools. Studying failure of a function can tell us much about the brain in health.

Any approach to tackling the cognitive clinicoradiological paradox involves many decisions regarding what is both relevant and possible to measure. Both cognition and radiological assessment of the brain are complex areas and can be studied at multiple levels. An explicit declaration of assumptions made in addressing these areas is therefore necessary to understanding the advantages of specific approaches and potential reasons for discrepancies with other research.

Decades of psychological research have established that cognition is best considered as a multidimensional construct, composed of a number of distinct functions with some degree of shared variance. Although certain patterns of deficits are recognised as characteristic of MS, significant inter-individual variation exists, increasing the numbers needed to demonstrate the relationship of cognitive performance to any biomarker or the effect of any intervention. Impaired information processing speed is the most frequently detected deficit in MS and has been proposed as the 'core' cognitive deficit [28], mediating others through disruption to connections between critical cortical regions. A biological interpretation to this model is clear in the context of a disease known for its primary attack on the myelin sheathing of white matter axons. The most common approaches to assessing cognition in MS, either single tests incorporating processing speed or mixed 'batteries' of different tests allowing calculation of an overall cognitive index, implicitly recognise this idea of separate but linked cognitive functions. The recently proposed Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) monitoring tool [36] is consistent with this, proposing the use of up to three tests of common deficits, with priority if time is limited given to assessment of processing speed through the Symbol Digit Modality Test (SDMT).

Different approaches to cognitive assessment are appropriate in addressing different research questions. Focussed assessment of particular domains will be suitable in attempting to localise particular functions. Measures of distributed functions or summary measures from multiple separate functions, more likely to relate to the global disease burden, will have more relevance in disease monitoring and development of disease-modifying treatments. The practicalities of cognitive assessment should also not be ignored; the best evidence will come from tests that are reliable, acceptable to patients and straightforward to administer.

Developing an imaging framework of disease burden is similarly complex and can be approached from many angles. Imaging models may focus on anatomical structures of interest, putative biomarkers of particular pathological processes, or technique-based approaches may seek to capture only the sum total of abnormality using a particular imaging modality without explicit interpretation of the underlying mechanisms involved. Volumetric and semi-quantitative markers can be extracted from routine imaging sequences and fully quantitative markers are becoming available from advanced imaging techniques, such as magnetic resonance spectroscopy and diffusion tensor imaging (DTI). The psychometric performance and clinical relevance of these newer markers has yet to be established. Nevertheless at face value an attractive approach may be the use of multiple features, providing more relevant information than focussing on single markers.

For the purposes of this thesis, a decision was made to address the relationship between cognitive function and imaging features using measures relating to the total burden of brain disease. Cognitive assessments were selected from the available datasets at the highest level, using a reliable and established marker of information processing speed. Radiological assessment of the brain was primarily from a structural perspective, aiming to capture quantitative and semi-quantitative information from the largest possible brain region for which reliable metrics are available, the cerebral white matter, using commonly available imaging sequences. The potential for additional value in using quantitative markers derived from DTI to assess microstructural abnormalities beyond the resolution of routine imaging sequences was also considered.

Apparent from the systematic review process reported in Chapter 3 was the vast volume of already existing literature addressing the issue of the relationship between cognition and imaging in MS and the variability in the methodology of its investigation. Variety has its advantages - one method may yield insight where others fail - but it does not follow that all investigations are equally valid and it remains the case that individual small studies must be interpreted with caution. While many approaches may be relevant in addressing specific research aims, the investigators' model describing the hypothesised relationship between cognition and imaging, together with consideration of the appropriate level of analysis, was rarely explicitly stated. Significant gaps in the literature were identified regarding the psychometric performance of the assessment methods used and any consideration of a non-linear relationship between cognition and imaging features.

In many cases specific cognitive tests were chosen to address particular research interests. Where summary measures of cognitive performance were sought, there was some evidence of moves towards the use of common tests and similar or overlapping test batteries. This was far from the case in their interpretation and derivation of scores for use in analysis. A lack of control or normative data was not always acknowledged and even when available its interpretation was often unclear. The recording of potential modifiers of cognitive performance, such as medication use and level of education, was also highly variable, as were the methods for its inclusion, if at all, in analyses. However, acknowledging the methodological heterogeneity in cognitive assessment in the existing literature, a gradual move towards harmonisation of testing does appear to be underway, with the increased use of common tests and continuing validation work on the BICAMS monitoring tool.

Moves towards consensus in image analysis methods are less apparent. Scientific progress relies on replication of key results and dependence on 'in-house' software outputs that cannot be easily accessed by other groups is clearly not optimal, potentially prohibiting smaller research groups from contributing to the field. The most frequently found image analysis outputs were derived from software requiring some degree of observer input and data on reproducibility of these methods was extremely limited. Previously published data on the psychometric attributes of particular software may exist but needs to be clearly cited, while bearing in mind that variation is likely if applied to a new population, or used by new operators. It is unclear whether the near silence surrounding measurement error relates to an embarrassment regarding its disclosure, or confusion over the 'correct' method for its assessment. Either way it should not be allowed to hinder progress.

Moves towards greater consensus in the MS image analysis community seem vital and overdue, with clearer reporting of methodology a prerequisite. Uniformity of methodology is not necessary as long as different softwares lead to the same results. Enhanced awareness of the role of measurement error and greater transparency in reporting of reproducibility metrics is key. With an inherently highly heterogeneous disease, attenuation of potentially significant results due to limited understanding of tool performance makes inefficient use of resources. Wider availability of analysis tools would be beneficial, allowing easier comparison between methods and accumulation of data regarding their performance in different populations. This led to the recommendations [185] that the development of standardised datasets should be prioritised to facilitate method comparison and benchmarking. The equivalence of all segmentation softwares remains to be established and comparison of methodologies may highlight particular advantages. While preference should be given to use of widely applicable and accessible software methods, restricting practice to a limited set of analysis tools is not necessarily beneficial. Straightforward techniques should be favoured, but discouraging innovation would risk missing out on insights arising from exploration of newer techniques. Some features may still be best assessed using non-automated methods, such as visual scores.

The overall effect size derived from the meta-analysis of the published literatures suggested that variance in cognitive performance had not been fully accounted for by measures of WMH burden. However concerns over methodological heterogeneity meant this result was interpreted with caution. No study was found considering a non-linear relationship. Re-evaluation of this relationship between cognition and imaging measures of disease therefore seemed critical, with appropriate levels of measurement both for cognition and imaging and using metrics with clearly defined performance characteristics.

Expert manual segmentation on imaging has become the accepted reference standard for quantifying disease burden, however this should not be confused with any claims to represent ‘ground truth’. Pathological specimen examination may demonstrate certain features more clearly, but these studies will also carry measurement error, be limited by availability of samples and may not be applicable to the majority of cases of people living with MS. Other features may be better demonstrated by in vivo imaging techniques and with robust measurement will have greater potential for wider use in disease monitoring.

Manual segmentation is an imperfect tool, but remains the standard validation procedure for new techniques in tissue segmentation. While this may seem a pragmatic decision in the face of limited availability of pathological samples, it may also support a misplaced belief in the stability of its outputs. Results based on one manual segmentation may not match those performed elsewhere, at different times or in different circumstances. The investigation reported in Chapter 4 shows that significant shifts in measurements across a cohort can occur even with a single observer, and large discrepancies can be found between two observers with similar training.

There are many reasons why two manual segmentations may provide different estimates of disease burden, both in terms of disease volume and spatial location. Is the aim of segmentation to identify all areas of white matter that appear abnormal, all those that are thought to represent a previous acute inflammatory attack or simply the most hyperintense voxels? If only asked to identify voxels of a certain brightness, then a computer is clearly going to be the best choice for generating outputs, but the validity of this information is unclear. Expert human observers may provide advantages in interpreting more subtle abnormalities and normal variants and artefacts, as well as adjusting for changing background intensity. Decisions on the significance of diffusely abnormal (‘dirty’) white matter may be important in quantifying the totality of disease, but are unlikely to be straightforward. The inherent subjectivity and resulting discordant measures is hardly surprising.

Even with the most diligent approach to tackling reproducibility in manual segmentation, the consistency offered by fully automated methods may make them more appropriate for use in generating reliable biomarkers, particularly in large cohort studies, although demonstrating this over time may be more difficult. Where manual segmentation is used for direct comparison with phenotype or for benchmarking a new technique, a greater awareness of the associated measurement error is necessary. Transparency in reporting reproducibility measures, development of standard datasets and specification of observer training should be encouraged.

Several of the earliest papers relating cognition to imaging changes identified in the systematic review reported in Chapter 2 used visual rating scales in some form, but these appear to have been largely abandoned in the more recent literature in favour of volumetric outcomes. Visual rating scales have become widely accepted in other conditions, particularly so for the white matter changes associated with small vessel disease (SVD) and ageing; the data here suggest that they may be useful in MS. An opportunity for using visual ratings as an outcome of interest or a stratification tool is clear and presents many potential advantages.

A limited number of imaging appearance categories, such as in the widely used 'Fazekas' scale for white matter changes in SVD, may seem too restrictive to fully explore research findings. However given the issues surrounding reproducibility in volumetric measures, described above, a false reliance on small changes in WMH volume may be misleading. Visual ratings of white matter disease severity have been found to show a high correlation with volumetric measures in both SVD [174] and for MS in the work described in Chapter 5. The measurement tool used should be fitted to the purpose and a decision on the use of volumetric or visual rating measures may depend on the incremental change in WMH burden that is considered relevant. Whether visual ratings are suitable for use as a research outcome may depend on the sensitivity required, but clearly offer advantages, not least time efficiency and ease of collaboration across sites.

Additional value from visual assessments may come from their ability to assess more than one aspect of disease appearances. The heterogeneous manifestations of MS, in terms of both the disease and the response of the central nervous system, varying between people and over time, is unlikely to be fully captured in a unidimensional outcome. There is an apparent disconnection between imaging interest in the pattern and location of visible abnormalities in the earlier stages, and the reductive, volumetric approach in established disease. Whether this reflects a lack of confidence in the ability of any measurement tool to capture the complexity of disease appearances, or merely a lack of knowledge of alternatives, is unclear.

In the work described here, using the largest available cohort made up of people with early and late stage disease, it appeared that a measure of lesion burden and an indicator of the presence of cavitation, i.e. more severe damage, were both significant independent predictors of cognition. This work should however be interpreted with caution, given its development in two cohorts with different clinical characteristics as well as different imaging protocols, and will need testing in further groups. If confirmed, this would suggest all WMHs should not be considered equal and failure of remyelination may be a critical factor in determining disability.

A final substantial advantage for visual ratings is their potential for translation: between research and clinical work, between scanners, between centres and countries. Although the potential for widespread use of a visual rating scale in MS clearly exists, a significant obstacle will be standardisation. Individual observers will always show variation in their assignment of scores to the countless possible imaging appearances and methods must be found to both minimise and quantify this variation. More work will be needed in the development of training datasets for new raters and the effect of making more sample images available for guidance. Given the limited number of sample images provided in the work described here, the results are encouraging.

The visual rating work described in Chapter 5 was large for a reproducibility study, but clearly needs testing in new and varied cohorts, with the involvement of more observers. Larger studies will also be needed to show if any of the less common features identified have relevance in their relationship to phenotype. Additionally, the hierarchical structure for rating features as used here would allow for alternative shorter assessments, and these may be more appropriate for particular research purposes. Given the modest correlation consistently found between imaging appearances and phenotype, a straightforward assessment of WMH burden with a limited number of categories may prove sufficient for purpose where this needs to be taken into account.

Where volumetric and spatial representations of the disease burden are needed, the use of automated segmentation methods, such as the one presented here, clearly offers potential. Consistent output measures may be obtainable in large cohorts, avoiding the need for time consuming and subjective manual segmentation. However the validity of any segmentation tool is more difficult to establish and although consistent, a software output optimised to best resemble a reference segmentation will be affected by the reliability of that segmentation.

Clarity in the literature over the ‘correct’ metric for demonstrating acceptable reproducibility is not apparent. This may relate to confusion over the strengths and weaknesses of different metrics or a reflection that different aspects of reproducibility may be relevant in different situations. High sensitivity or specificity does not necessarily mean a close fit to the reference segmentation and the appropriate measurement tool may not necessarily be the most sensitive one. Work towards a consensus approach in the image analysis community on reporting metrics of agreement and reliability would be beneficial, remembering that any of these will depend on the cohort in which the technique is tested. While it would be impractical to retest all tools before each new use, an awareness of how far they have been tested and in which cohorts is necessary to understanding the validity of results gained.

In the absence of a true pathological reference for comparison, the retention of a probabilistic element to the output may be beneficial in interpreting results based on a segmentation, such as overlaid quantitative metrics. The process of manual segmentation forces a binary decision on the normality of each voxel, which may not be a realistic expectation of any imaging sequence. Without greater availability of pathological samples for comparison, proving the validity of either manual or automated segmentations will remains challenging, but the removal of any subjective element to the process is clearly advantageous.

An alternative approach to establishing validity would be optimising the segmentation process to maximise the association of its outputs with phenotype, essentially allowing test scores or clinical findings to aid interpretation of imaging features. While it may be possible to adjust segmentation parameters to increase the association between cognitive and imaging findings, this would clearly require validation in large cohorts and careful interpretation. If successful, this would lead to a reframing of the clinicoradiological paradox - a higher proportion of variance in cognitive outcomes may be explained using imaging outputs, but not those with any easily interpretable pathological significance.

The results presented here based on routine imaging sequences, using both the volumetric and semi-quantitative tools for disease burden analysis, showed a stronger relationship with cognitive performance than the overall result of the meta-analysis of the published literature. Many factors will have affected this, but the measurement tool used and its optimisation appear relevant considerations.

Nevertheless, variance in cognitive performance is far from fully explained by any measure of WMH burden. The difficulties described in defining consistent edges to WMHs suggest the possibility that changes within the surrounding white matter may also be relevant, leading to the current interest in using quantitative imaging techniques to examine them.

If consideration of the inflammatory component of the disease burden in MS cannot fully explain phenotypic changes, then a more detailed examination of the neurodegenerative component appears a logical next step in appreciating the total disease burden. If neurodegeneration progresses with at least partial independence from inflammatory damage, then biomarkers of this process may contribute additional explanatory power in predicting cognitive performance. The work described in Chapter 8 used DTI-derived biomarkers of tissue microstructure to quantify diffuse changes outwith the regions of inflammatory damage visible on routine imaging.

A variety of metrics have been derived from DTI data and a focus on particular tissue structures may be appropriate to different research questions. Disruptions to microscopic tissue architecture are inferred from these markers but the choice of the 'correct' metric to use is not always clear; all remain non-specific and must be interpreted with caution. For the purposes of this thesis, the focus was on capturing the diffuse and ill-defined changes within the white matter that could not be quantified on routine imaging sequences. Straightforward DTI metrics within the 'normal-appearing' white matter compartment, derived using automated segmentation, were considered first. Tractography was used to extract the most highly coherent white matter, recognising that tissue complexity, particularly crossing fibres, may confound measurements elsewhere, and the major white matter tracts may be highly relevant in influencing processing speed. A novel metric, peak width of skeletonised mean diffusivity (PSMD), derived using the Tract-Based Spatial Statistics (TBSS) procedure, was used as an alternative method of assessing the most coherent tissue within the white matter skeleton and summarising the spread of mean diffusivity values rather than any average measure. Where feasible, imaging tools which are as close to fully automated as possible offer clear advantages and PSMD has been proposed with a fully automated processing pipeline freely and publicly available for ongoing evaluation. The results here are thought to be its first use in the context of MS.

The DTI-derived metrics tested were found to show a range of values across the cohort with the majority of these correlating with a marker of cognitive performance. However, no correlation was stronger than that found for measures of WMH volume and no additional value was found within models predicting cognitive performance from imaging and non-imaging data.

Nowhere in this work, with the possible exception of the visual rating assessment, has it been possible to show that using two measures of white matter disease severity is ever better than one, in determining cognitive performance. The high correlations between all markers of disease, including those interpreted to represent the inflammatory and neurodegenerative components, mean that very large cohorts would be needed to show any separate effects from each. While they may have separate effects, it is possible that one pathological component is of greater importance in determining cognitive function and whichever we attempt to measure is providing a surrogate marker for it. The work here provides no evidence to reject a disease model in which inflammatory damage is the driver for diffuse white matter degeneration and together these lead to a decline in cognitive performance. In a wider context, reports of exploratory uses of advanced imaging techniques must ensure a description of their covariance with more straightforward and established measures of disease burden.

White matter pathology, as detectable by current imaging techniques, accounts for only a small proportion of the variance found in cognitive performance of people with MS. This prompts a reconsideration of whether the relationship between pathology and cognition has been addressed within the appropriate framework and using the optimal tools.

Given that many cognitive functions show regional localisation, the decision of which outcome to use may be highly relevant. All analysis presented in this thesis is based on an assumption that information processing speed is a distributed function, reliant on widespread white matter integrity. The SDMT is straightforward, quick and unlikely to be affected by physical disability or fatigue. It has recently been proposed as the single most useful test for the BICAMS. However an alternative approach could be the use of multiple cognitive tests and a method of dimension reduction, such as factor analysis, for extracting a marker of overall performance. This may be advantageous, but is likely to be a more lengthy and resource intensive procedure, requiring specialised skills and potentially limiting participation and applicability. As with all tests, the attenuating effect of poor reliability on observed correlations with other disease-related variables must be considered.

Cerebral white matter was selected for study here as the structural unit of interest relevant to distributed cognitive function and the value in disease markers derived from structural imaging techniques for this region was considered. This leaves open the question of how far white matter pathology can be considered to be fully characterised by these markers and whether it is necessary to consider other structures, for instance cerebellum, deep nuclei and cortical grey matter, in searching for pathological correlates of cognitive performance. A trend towards a significant improvement in model fit was found when adding total brain volume rather than just white matter volume as a predictor, which would support this.

The inherently multicontrast nature of magnetic resonance imaging (MRI) suggests that any unidimensional measure extracted will be limited in the representation it provides and this seems even more likely when dealing with a complex and highly variable disease; imaging features are another aspect of MS to show its characteristic heterogeneity. While summarising all this variability into practical research outcomes may seem unfeasible, the retreat to only using unidimensional WMH volume may be too simplistic. The assessment of full and partial lesion cavitation considered in Chapter 7 acknowledged that WMH volume alone is a crude marker of pathology, encompassing a range of degrees of tissue damage. The model fit improved with this addition of cavitation presence as a predictor, but this result should be interpreted with caution given its reliance on imaging data from two very different cohorts.

Additional tissue characterisation, such as that derived from advanced techniques including magnetisation transfer imaging and magnetic resonance spectroscopy, may be relevant and necessary in determining the total disease burden. Beyond the resolution of all current in vivo imaging techniques are pathological and adaptive synaptic changes, with information about these inferred on a much larger scale by functional MRI techniques. Adaptive changes are known to occur in MS and this capacity for neural plasticity may vary greatly between individuals and across the disease course.

A large number of non-disease variables, such as age, education and medication, are known to affect performance in cognitive tests. As far as possible these have been taken into account in the analyses described, but the data available was far from complete and a greater awareness of the need for considering these factors is clearly necessary. No consensus opinion on a full group of variables affecting cognitive performance is yet available and a list of variables to consider may grow as research on this topic develops. It is possible that factors such as fatigue and motivation at the time of testing that are very difficult to measure may be highly influential on cognitive outcomes. Given the incomplete assessment of cognitive modifiers here, perhaps a correlation of test scores and WMH volume of close to 0.5 is higher than might be expected, and whether this can be improved in more complete datasets should be investigated.

The most straightforward methods for examining the relationship between two numerical variables are tests of their linear association, but there is no fundamental reason regarding the neural basis for cognitive performance why this should be the case. Compensatory neural reorganisation is found in many brain diseases and there is evidence from functional imaging to support its occurrence in MS. Repair processes can take place to some extent, possibly up to the point at which recruitment of new and functional oligodendrocytes is exhausted. Network redundancy may also be built into the brain, explaining the accumulation of 'silent' inflammatory lesions without any recognised clinical event. These factors make it more plausible that a certain degree of pathology can exist without any associated deterioration of performance in skills relying on widespread neuronal integrity. The model of a 'dose-response' curve may be more appropriate than a linear relationship, with a decline in information processing speed only occurring above a certain level of disease burden.

The model of a non-linear relationship between WMHs and cognitive performance is supported by the work presented here, although further investigation of this possibility is clearly required. The varying effect sizes reported in the published literature (see Chapter 3) appeared partly related to the magnitude of the WMH burden itself, greater effect sizes being reported in cohorts with overall larger volumes of disease. In the MS-SMART cohort, a group of individuals with established disease and a wide range of WMH volumes, this also appeared to be the case, with a steady decline in cognitive performance associated with WMHs only at higher levels.

The DTI-derived data also supports the idea of a 'threshold' effect of WMHs. Using the novel marker PSMD to measure the spread of MD values across the white matter skeleton, there was tentative evidence that this spread is relatively stable, and similar to published control populations, up to a certain WMH volume. Beyond this, PSMD increased, suggesting diffuse damage to the white matter, and these values were associated with declining cognitive performance.

If this non-linear effect of disease burden on phenotype is confirmed, the consequences may be important. It would provide support for aggressive treatment approaches in the early disease stages, with an aim of preventing the disease burden reaching a threshold at which compensatory mechanisms and/or repair processes are no longer adequate to prevent disability accumulation. On a more prosaic level, a dynamic relationship between disease burden and phenotype further reinforces the need for published research to provide full descriptions of cohorts studied, particularly their WMH burden. Results should be interpreted only with reference to people at similar disease stages.

Harmonisation of standardised cognitive assessments in MS has been proposed and widespread validation studies of these tests are underway. A similar consensus in the approach to image analysis methodology and its reporting is not yet apparent. Heterogeneity in research methods for investigating this heterogeneous disease adds to the confusion in the overwhelming body of published research and is unlikely to indicate the most efficient use of resources. Encouraging the reporting of reliability metrics for all measurement tools used is a critical first step and an understanding of the role of measurement error in attenuating observed results will guide correct interpretation of future results.

In defining a basic model linking measures of pathology with cognition, further work on identifying all potential modifiers of cognitive performance, including those unrelated to disease burden, is necessary. Recording of variables already known to be relevant should become standard along with their consideration as part of any analysis. Different factors may be critical at different disease stages and results should be extrapolated to new cohorts with caution.

In considering where the remainder of the variation in imaging correlates of cognitive performance lies, optimisation of measurement tools is vital. Synthesis of markers derived using different imaging techniques, such as multimodal white and grey matter assessment, may yet prove useful, although the advanced imaging markers tested here were not shown to carry any additional benefit to routine imaging markers in terms of predicting cognitive outcomes. A major breakthrough in understanding cognitive function in MS may await advances in imaging that can quantify brain architecture at the synaptic level.

Even with optimised measurements, the relationship to cognitive outcomes is unlikely to be straightforward and system redundancy, capacity for repair and reorganisation are possibilities requiring further investigation. Sources of variation beyond the reach of current in vivo imaging techniques, such as synaptic and molecular adaptation may be necessary considerations in producing a closer approximation to the true burden of disease burden. Hidden capacities of the human nervous system may allow the cognitive clinicoradiological paradox to remain for the foreseeable future.

Appendix A

Protocol for systematic review of relationship between cognitive performance and total white matter lesion burden

AIM

To systematically review the published evidence describing the relationship between standard structural MRI measures of white matter lesion burden in people with multiple sclerosis and cognitive status.

DATABASES SEARCHED

- PubMed
- ISI Web of Knowledge
- Embase

SEARCH TERMS

- ‘multiple sclerosis’ *and*
- ‘cognitive’ *or* ‘cognition’ *and*
- ‘magnetic resonance imaging’ *or* ‘MRI’ *or* ‘MR imaging’

SEARCH FILTERS

- Articles published in the English language
- Research using human subjects
- No date restriction

SCREENING OF ABSTRACTS & PAPERS: ELIGIBILITY CRITERIA	• MS • Cognition • MRI • Primary literature (reviews and data published only in abstract form excluded) • Study of adult patients only (age $\geq$ 18 years) • Neither including nor restricted to clinically isolated syndromes • Not a duplicate publication • Not presenting previously published data • Contemporaneous capture of imaging and cognitive data • Not subsequently retracted • Primary aim of the study is to explore the relationship of MRI metrics for lesions and cognition
ADDITIONAL ASCERTAINMENT	• Screening of references from review articles identified in the initial search. • Hand search of archives of the journals Neurology, Multiple Sclerosis and the American Journal of Neuroradiology for previous ten years

EXTRACTED METRICS	• Study quality assessment (see Appendix C), based on STROBE guidelines • Study design, number of participants, interval between cognition & imaging • Participant characteristics: age, sex, disease phenotype • Cognitive testing methods: tests/batteries used; blinding, identity and training of tester; use of normative data; recording of potential confounders - age, sex, education level, premorbid IQ, cognitive leisure activities, affective disorders and drug history • Image acquisition: magnet field strength, details of sequences performed • Image analysis methods: preprocessing steps; sequence used to measure lesion burden; lesion quantification technique; softwares used; blinding, identity and training of analyst; reliability measures • Statistical analysis methods; controlling for confounders • Summary statistics for lesion burden • Main results: unadjusted and/or adjusted

Appendix B

Record of search strategy for systematic review of literature

Medline

Searched via OVID platform, on 01/07/15, with 671 references retrieved. The search strategy is shown below.

	Searches	Results
1	multiple sclerosis.mp. or Multiple Sclerosis/	56333
2	magnetic resonance imaging.mp. or Magnetic Resonance Imaging/	353245
3	mri.mp.	144054
4	mr imaging.mp.	32694
5	2 or 3 or 4	382327
6	cognitive.mp. or Cognitive Reserve/ or Delirium, Dementia, Amnestic, Cognitive Disorders/ or Cognitive Science/ or Mild Cognitive Impairment/	208670
7	Cognition Disorders/ or Cognition/ or cognition.mp.	133903
8	6 or 7	260390
9	1 and 5 and 8	742
10	limit 9 to english language	672
11	limit 10 to retracted publication	1
12	10 not 11	671

Embase

Searched via OVID platform, on 01/07/15, with 1145 references retrieved. The search strategy is shown below.

	Searches	Results
1	multiple sclerosis.mp. or multiple sclerosis/	94203
3	cognition/ or cogniti*.mp.	426368
4	1 and 2 and 3	1844
5	limit 4 to english language	1755
6	limit 5 to (conference abstract or conference paper or conference proceeding or 'conference review')	610
7	5 not 6	1145

Web of Science

Searched on 01/07/15, with 1250 references retrieved. The search strategy is shown below.

	Searches	Results
1	TOPIC: (magnetic resonance imaging) OR TOPIC: (mri) OR TOPIC: (mr imaging)	343471
2	TOPIC: (cogniti*)	426125
3	TOPIC: (multiple sclerosis)	90559
4	3 AND 2 AND 1	1396
5	3 AND 2 AND 1 Refined by: DOCUMENT TYPES: (ARTICLE OR EDITORIAL MATERIAL OR REVIEW)	1324
6	3 AND 2 AND 1 Refined by: DOCUMENT TYPES: (ARTICLE OR EDITORIAL MATERIAL OR REVIEW) AND LANGUAGES: (ENGLISH)	1250

PubMed

To ensure recent papers not yet indexed on Medline were also included, the PubMed database was also searched using the same search terms, with 816 references retrieved. The search strategy is shown below.

((((((magnetic resonance imaging) OR MRI)) OR MR imaging)) AND ((cognition) OR cognitive)) AND multiple sclerosis AND (English[lang])

Appendix C

Quality assessment criteria used in systematic review of literature

The quality assessment criteria described below were modified from the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) [112] checklist.

Where all key points are met, 1 point is awarded. Where the study meets most but not all of the applicable criteria, or only part of the relevant information is provided, a score of 0.5 is awarded.

Introduction

OBJECTIVE: *State specific objectives, including any prespecified hypotheses.*

The study should have a clearly stated objective mentioning white matter lesion volume as a metric of interest (awarded 0.5).

Full credit will only be given where the objective specifies what imaging sequence(s) will be used to measured lesion volume and what cognitive measure is used to examine the relationship between the two outcomes.

[0] [0.5] [1]

Methods

STUDY DESIGN: *Present key elements of study design early in the paper.*

The study design should be presented clearly, i.e. retrospective or prospective recruitment, case-control studies, or a sub-study of part of a larger study.

Prospective recruitment to address the study objective is considered preferable and a clear statement of this is needed for 1 point. A retrospective study design will be awarded 0.5.

Where participants are taken from a cohort being used for multiple (sub)studies, a maximum of 0.5 can be awarded, unless cognition and imaging relationships are clearly the primary aim of the overall study and cross-sectional baseline data are being used. Enough detail should be provided to ensure results are not duplications of other published work.

[0] [0.5] [1]

SETTING: *Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection.*

The dates of recruitment and testing should be provided. The delay between cognitive testing and imaging should be specified and less than 6 months. Both the above criteria are necessary for 1 point, either alone will be awarded 0.5. A description of the clinical setting (e.g. tertiary referral centre, multiple district general hospitals etc) is considered optimal, but is not necessary for full credit.

[0] [0.5] [1]

PARTICIPANTS: *Give the eligibility criteria, and the sources and methods of selection of participants.*

The authors should have clearly stipulated the criteria they used to include (and if applicable, to exclude) subjects into the study. A positive statement of who was sought for recruitment (whether any person with MS, or e.g. only people with a particular clinical phenotype) with relevant exclusion criteria is necessary for 1 mark.

Participants should not be excluded solely on the basis of higher levels of physical or cognitive disability, and where recruited subjects were unable to tolerate MRI this should be recorded.

[0] [0.5] [1]

The recruitment should be either a consecutive or random sample of eligible participants. Where this is unclear, the study will be awarded 0.

[0] [1]

VARIABLES: *Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.*

The cognitive tests performed should be specified. Whether results were interpreted relative to a control population or published norms should be clearly stated/described.

The definition of total lesion volume should be clearly defined, including the brain regions covered (whether deep grey matter included or excluded, whether posterior fossa included and how defined), the MRI sequence used for measurement and whether the results were adjusted for total (estimated) brain volume.

Clear definitions as above are required for both imaging and cognitive outcomes for a score of 1. Where one of these is unclear, a maximum of 0.5 will be awarded.

[0] [0.5] [1]

Potential confounding factors, including age, sex, education, drugs, pre-morbid IQ, pre-morbid cognitive leisure activities & affective disorders, should be measured. A score of 1 will be awarded where all these are identified, and 0.5 if $\geq$4 of them.

[0] [0.5] [1]

DATA SOURCES/MEASUREMENT: *For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.*

The person(s) performing the cognitive testing should be identified, with their level of training/experience.

Enough data should be provided to replicate the imaging sequences. This should include at least the type of sequence performed (e.g. spin echo, gradient echo), slice thickness and inter-slice interval, and preferably the pulse parameters (TE, TR, flip angle, FOV, matrix size), number of slices and magnet strength.

The method for measuring/estimating lesion volume should be clearly described, with details of the software package used if applicable. The person(s) performing the analysis should be identified with their level of training/experience. Measures of intra-/inter-observer variability should be provided.

All of the above criteria must be met for a score of 1; where $\geq$50%, but not all, of the relevant information is presented, the study will be awarded 0.5.

[0] [0.5] [1]

BIAS: *Describe any efforts to address potential sources of bias.*

Cognitive testing and image analysis should both be performed by individuals blind to the results of the other and this should be clearly stated. Where there is only a statement that the image analysis was performed blind to the cognitive results, 0.5 will be awarded, otherwise the study will be scored 0.

Ideally the image analysis and cognitive testing should be carried out blind to (as far as possible in the case of cognitive testing) all data on clinical status and confounding factors.

[0] [0.5] [1]

STUDY SIZE: *Explain how the study size was arrived at.*

A calculation of study size should be provided.

[0] [1]

QUANTITATIVE VARIABLES: *Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why.*

Ideally, the full range of cognitive scores and lesion volumes will be used for the analysis, with or without transformation to Z-scores. This should be clearly stated and correlations using the full range of values or correlations by rank will be awarded 1 point.

If participants are categorised into groups by results of cognitive status (or, less likely, lesion volumes) the justification of the group definitions should be provided and boundaries pre-specified. A maximum of 0.5 will be awarded where outcomes are dichotomised (or otherwise grouped) for analysis.

[0] [0.5] [1]

STATISTICAL METHODS: *(a) Describe all statistical methods, including those used to control for confounding. (b) Describe any methods used to examine subgroups and interactions. (c) Explain how missing data were addressed. (d) If applicable, describe analytical methods taking account of sampling strategy. (e) Describe any sensitivity analyses.*

Statistical methods should be clearly described, ideally correlations between scores of cognition and lesion volume.

Unadjusted correlations should be calculated prior to controlling for potential confounders. If either unadjusted correlations or controlling for confounders is not included, a maximum of 0.5 can be awarded.

[0] [0.5] [1]

Results

PARTICIPANTS: *(a) Report numbers of individuals at each stage of study, e.g. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed. (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram.*

Participants recruited but not completing either cognitive testing or imaging should be specified. If this is unclear, a score of 0 is awarded.

[0] [1]

DESCRIPTIVE DATA: *(a) Give characteristics of study participants (e.g. demographic, clinical, social) and information on exposures and potential confounders. (b) Indicate number of participants with missing data for each variable of interest.*

Summary statistics for basic demographic data (age, sex) and MS phenotype should be provided. If this is not given, a score of 0 will be awarded.

Information on recent steroid use and disease-modifying therapy is considered necessary for a score of 1, but not full results of all potential confounders. Ideally these would be provided in supplementary material.

If results of multiple cognitive tests are used for analysis, the number of participants with incomplete data for each test should be given. If this is unclear, a maximum of 0.5 can be awarded.

[0] [0.5] [1]

OUTCOME DATA: *Report numbers of outcome events or summary measures.*

Summary statistics should be presented for both cognitive outcomes and lesion volumes. These should include measures of the dispersion as well as central tendency. Where this is incomplete, e.g. only the numbers of participants categorised as cognitively impaired versus not impaired are provided, a maximum score of 0.5 can be awarded.

[0] [0.5] [1]

MAIN RESULTS: *(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included. (b) Report category boundaries when continuous variables were categorised. (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period.*

Unadjusted outcomes should be presented for cognitive data and, if applicable, confounder-adjusted outcomes. A measure of their precision should be provided.

[0] [0.5] [1]

Appendix D

Table of cognitive tests and scoring schemes used in individual studies reviewed

Paper	Battery	No. of tests	SDMT/ PASAT	Controls	Normative data	Definition of cognitive impairment
Sacco 2015	BRB-N, Stroop test	8	Both (PASAT 2s & 3s)	Y	Y (Italian population)	$\geq$ 2 SD below normative mean on $\geq$ 2 tests, including one memory test
Yildiz 2014	‘MUSIC’	5	N	N	N	Overall score $<$ 20/30
Laffon 2014	PASAT	1	PASAT 3s	Y	Y (French population)	$\geq$ 2 SD below normative mean
Niino 2014	BRB-N	7	Both (PASAT 2s & 3s)	Y	N	N/A
Sbardella 2013	PASAT	1	PASAT 2s &3s	Y	Y (Published reference data)	N/A
Mike 2013	‘Theory of Mind’	3	N	Y	N (Direct comparison with controls)	N/A
Francis 2013	MACFIMS	7	Both	N	Y (Published reference data)	$>$ 1.5 SD below normative mean on $\geq$ 2 tests
Rossi 2012	BRB-N	8	Both	N	Y (Italian population)	$\geq$ 2 SD below normative mean on $\geq$ 2 subtests

Paper	Battery	No. of tests	SDMT/ PASAT	Controls	Normative data	Definition of cognitive impairment
Mesaros 2012	BRB-N	8	Both	N	Y (Published reference data)	$\geq$ 2 SD below mean on 1 subtest & $\geq$ 1.5 SD on another; *or* $\geq$ 1.5SD below mean on $\geq$ 3 tests. For individual tests $\geq$ 1.5 SD below mean.
Lund 2012	Unnamed	32	N	Unclear	Y (Controls unspecified)	N/A
Bomboi 2011	MACFIMS (without PASAT)	6	SDMT	Y	N (Direct comparison with matched controls)	N/A
Mike 2011	MACFIMS	7	Both (PASAT 2s)	N	N	N/A
Akbar 2010	BRB-N (without SDMT)	4	PASAT	N	Y (From published manual)	$\leq$ 5th percentile on $\geq$ 2 subtests
Heesen 2010	SDMT (screening); then further battery	7	Both (SDMT for screening)	N	Unclear	SDMT: above mean or $\geq$ 1.5 SD below mean
Patti 2009	BRB-N, Stroop	10	Both (PASAT 2s & 3s)	N	Y(Italian population)	$\geq$ 1 SD below mean on $\geq$ 3 subtests [1]
Krause 2009	Facial expression tests (from Florida Affect Battery)	1	N	Y	Y (Controls)	$\geq$ 2 SD below control mean
Sanchez 2008	Unnamed	$>$ 20	Both	Y	Y (Controls)	Overall cognitive index $\leq$5th percentile of control data
Rovaris 2008	Unnamed	$>$ 10	PASAT 3s	N	Y (Italian population)	Scores of 0 on $\geq$ 3 subtests
Karlinska 2008	Unnamed	5	N	Y	N	Unclear

[1] Effect of using other definitions considered, e.g. 5th percentile cut-off

Paper	Battery	No. of tests	SDMT/ PASAT	Controls	Normative data	Definition of cognitive impairment
Lin 2008	PASAT	1	PASAT 3s	Y	Y (Published reference data)	≥ 2 SD below normative mean
Amato 2008	BRB-N, Stroop	6	Both	N	Y (Italian population)	≥ 2 SD below mean on ≥ 3 tests
Houtchens 2007	MACFIMS	6	Both (PASAT 2s & 3s)	Y	N	N/A
Parmenter 2007	Frontal function	2	N	Y	N	N/A
Lazeron 2006	Computer tests 'ANT'	8	N	N	Y (Normative data)	N/A
Benedict 2006	Unnamed	4	Both	Y	Y (Previously published control data)	≥ 2 SD below mean on 1 test and ≥ 1.5 SD on another, or ≥ 1.5 SD below mean on ≥ 3 tests.
Lazeron 2005	BRB-N (+/- substitution for SRT)	5	Both (PASAT 2s & 3s)	N	Y (Published reference data)	Overall scores ≥ 2 SD below reference mean
Deloire 2005	BRB-N +	10	B (PASAT 2s & 3s)	Y	Y (Controls)	≤ 5th percentile of control data
Archibald 2004	Information processing	2	N	N	Y (Previously published control data)	N/A
Benedict 2004	MACFIMS (modified)	8	Both (PASAT 2s & 3s)	Y	N	N/A
Christ-odoulou 2003	BRB-N (modified)	6	Both (PASAT 2s & 3s)	N	Y (published data and previous study subjects)	N/A
Bermel 2002	SDMT	1	SDMT	N	N	N/A
Zivadinov 2001	Unnamed	9	PASAT	Y	Y(Partly published cut-offs, partly control data)	Abnormal result on ≥ 2 tests: ≥ 2 SD below control mean or published cut-off

Paper	Battery	No. of tests	SDMT/ PASAT	Controls	Normative data	Definition of cognitive impairment
Nocentini 2001	Unnamed	20	SDMT	Y	N	Unclear
Kalkers 2001	PASAT	1	PASAT 3s	N	Y (Published reference population, partly overlapping)	N/A
Snyder 2001	PASAT	1	PASAT (4 speeds)	N	N	N/A
Comi 1999	Unnamed (frontal)	6	N	N	Y (Italian population data or previous controls)	Each test: Published cut-offs or $\leq$ 5th/10th percentile of controls. Overall groups: either ≥ 3 abnormal tests, or all normal
Camp 1999	BRB-N (modified)	6	Both	Y	Y (Controls)	N/A
Sun 1998	Dementia screening scales	2 screening scales	N	N	N	N/A
Rovaris 1998	Unnamed	10	PASAT (2s &3s)	N	N	Unclear: Abnormal results in ≥ 2 tests
Hohol 1997	BRB-N	4	SDMT	N	N (Published guidelines)	N/A
Patti 1995	Unnamed	10+	N	N	N	Unclear
Comi 1995	Unnamed	10+	N	N	Y (Italian population)	$\geq$ 3 tests $\geq$ 2 SD below mean
Moller 1994	'SIDAM' dementia screening scale	Unclear	N	N	N	$\leq$ 46/55
Swirsky-Sacchetti 1992	Unnamed	8	N	N	Y (Partly, published normative data)	N/A
Ron 1991	Unnamed	10+	N	Y	Y (Controls)	N/A
Pozzilli 1991	Unnamed	N	N	Y	Y (Controls)	$\geq$ 2 SD below control mean on $\geq$ 2 tests
Izquierdo 1991	Unnamed	5+	N	Unclear	N	N/A

Paper	Battery	No. of tests	SDMT/ PASAT	Controls	Normative data	Definition of cognitive impairment
Anzola 1990	Unnamed	9	N	N	Y (Published norms)	N/A
Franklin 1988	Unnamed	10+	SDMT	Y	Y (Previous standardisation sample)	≤16th percentile of normative data
Huber 1987	Unnamed	8	N	Y	N	≥ 1,2 or 3 SD below control mean

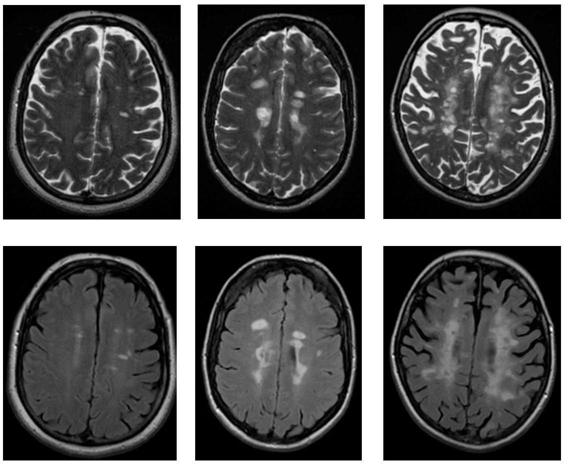
Summary of approaches to cognitive testing in all included papers in systematic review (Chapter 3). Y: Yes; N: No; N/A: Not applicable; BRB-N: Brief Repeatable Battery; MACFIMS: Minimal Assessment of Cognitive Function in Multiple Sclerosis; PASAT: Paced Auditory Serial Addition Test; SD: Standard deviations; SDMT: Symbol Digit Modality Test; SRT: Selective reminding test

Appendix E

Data collection form used in initial pilot study of visual rating scale for MS imaging features

Scan ID: Reader:

LESION BURDEN



T2/FLAIR white matter lesions
based on Fazekas scoring

Lobar scores *(images on left for guidance)*
0: None
1: Discrete lesions
2: Beginning of confluence *or* >5 non-confluent lesions
3: Confluent lesions

Periventricular scores
0: None
1: Caps / pencil-thin lining around ventricles
2: Smooth halo around ventricles
3: Irregular periventricular hyperintensities extending into deep white matter

Fazekas, F, et al. AJR 1987 **149**(2): 351-6.

Juxtacortical and cortical lesions : *Number of lesions in each lobe which abut or involve cortex.*
Scoring: **0:** None **1:** 1-2 **2:** 3-4 **3:** ≥5

Cavitated lesions: *Number of cavitated lesions. (Defined as lesions which are close to CSF signal on all sequences.)*
Scoring: **0:** None **1:** 1-2 **2:** 3-4 **3:** ≥5

		T2 / FLAIR lesion burden	Cavitation	Juxta-cortical lesions	Cortical lesions
Lobar white matter					
Frontal	R				
	L				
Parietal	R				
	L				
Temporal	R				
	L				
Occipital	R				
	L				
Insula	R				
	L				

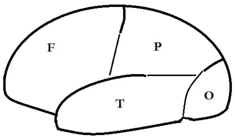
Periventricular white matter	R		
	L		
Corpus callosum			
Basal ganglia	R		
	L		
Brainstem			
Cerebellar peduncles	R		
	L		
Cerebellar hemispheres	R		
	L		

Lobes definitions:
Frontal: anterior to central sulcus

Parietal: anterior to parieto-occipital sulcus; superior to posterior extent of Sylvian fissure

Occipital: Posterior to parieto-occipital sulcus and temporo-occipital incisure

Temporal: lateral to Sylvian fissure; anterior to temporo-occipital incisure.



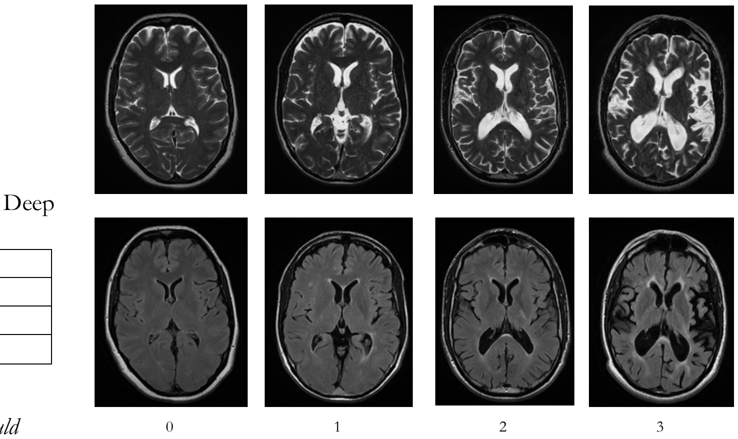
All ratings 0 – 3. Basal ganglia score to include striatum, globus pallidus, thalamus and internal capsule.

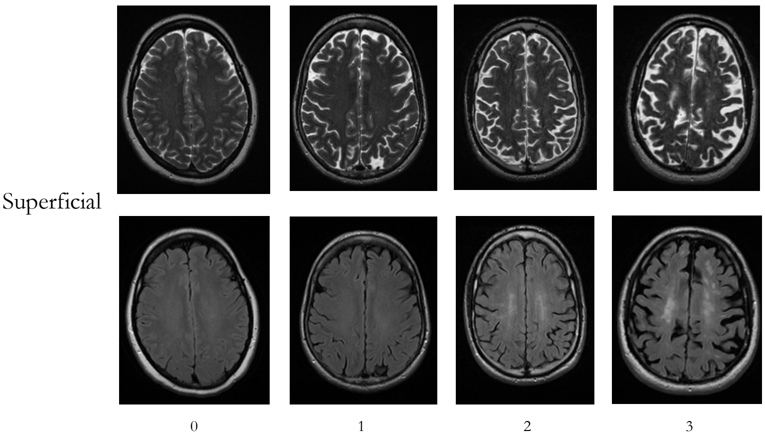
Scan ID: Reader:

ATROPHY

Cerebral	Deep	
	Superficial	
Corpus callosum		
Infratentorial		

Atrophy scoring (0 – 3). This should be rated without reference to age. Images on right for guidance.





ENLARGED PERIVASCULAR SPACES

Defined as small, sharply delineated structures of CSF intensity measuring up to 3mm and following the course of perforating vessels. All relevant slices for the anatomical area should be reviewed on T2 weighted imaging and the highest number on a slice recorded.

0: none **1:** = <10 **2:** 11-20 **3:** 21-40 **4:** >40

	Right	Left
Basal ganglia		
Centrum semiovale		

as per Potter, G.M., et al. Int J Stroke, 2015. **10**(3): 376-81

Is there focal perivascular space enlargement related to any lesion? YES/NO
If yes, where?___________________________

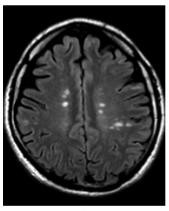
ANY OTHER FINDINGS/COMMENTS

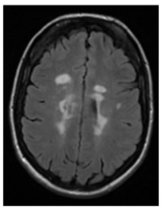
Appendix F

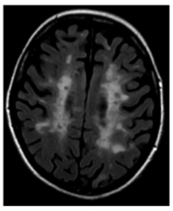
Data collection form used in second pilot study and validation of visual rating scale for MS imaging features

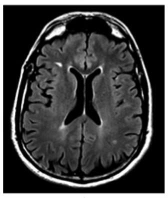
Reader: Subject ID:

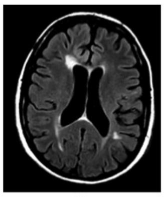
T2/FLAIR Lesion burden

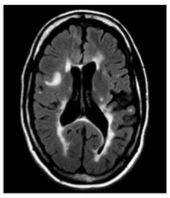

1


2


3


1


2


3

Deep white matter (global and lobar scores)

0. Absent
1. Discrete lesions
2. Intermediate appearances
3. Confluent white matter abnormality

Periventricular lesions

0. Absent
1. Caps/thin lining around ventricles
2. Intermediate appearances
3. Irregular hyperintensities extending into deep white matter.

(Based on Fazekas, F et al. AJR (1987); 149(2):351-6.)

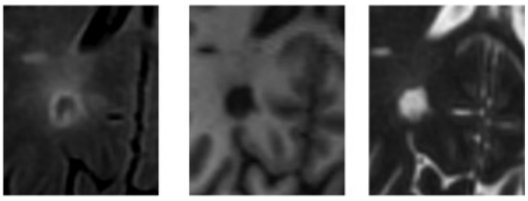
Please score all regions (other than periventricular white matter) with reference to the sample images and descriptions provided for deep white matter, choosing the category which most closely matches scan appearances.

Region	Side	Lesion score [0 - 3]	Cavitated lesions?* [Y/N]	Number/ type [F/P]	(Juxta-)cortical lesions? [Y/N]	Number/ type [J/C/both]
Example lobe	*R*	*2*	*Y*	*2P*	*Y*	*1J*
Global impression						
Deep white matter	R					
	L					
Periventricular white matter	R					
	L					
By region						
Frontal lobe	R					
	L					
Parietal lobe	R					
	L					
Temporal lobe	R					
	L					
Occipital lobe	R					
	L					
Insula	R					
	L					
Corpus callosum						
Basal ganglia	R					
	L					
Brainstem						
Cerebellar peduncles	R					
	L					
Cerebellar hemispheres	R					
	L					

*Reference pictures for lesion cavitation on page 2.

Reader: Subject ID:

Cavitation

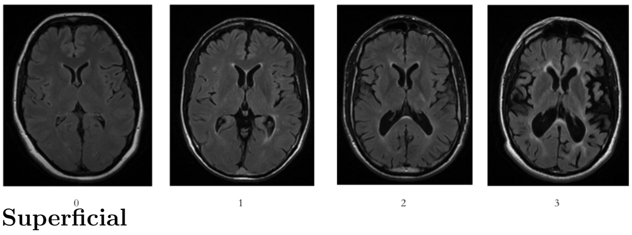


F: A lesion appears *fully* cavitated, with its major part having signal characteristics similar to CSF on all sequences.

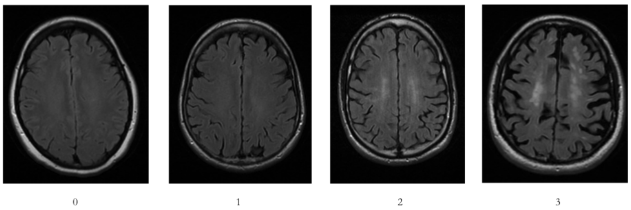
P: A lesion appears *partly* cavitated. This may include the internal structure appearing lace-like, or only a small portion (< 50%) returning CSF signal.

Atrophy

Deep



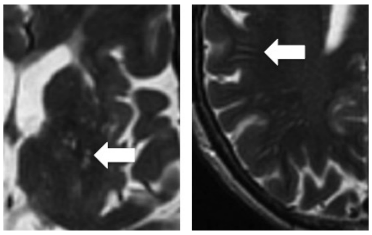
Superficial



Region		Score [0 - 3]
Cerebral hemispheres	Deep	
	superficial	
Corpus callosum		
Posterior fossa		

The images on the left are provided for guidance in rating cerebral atrophy.
The corpus callosum should be rated on the mid-sagittal image.

Enlarged perivascular spaces



Defined as small, sharply delineated structures of CSF intensity, up to 3mm in diameter and following the course of perforating vessels.
All relevant slices for the anatomical area should be reviewed on T2w imaging and the highest number on a slice recorded.

Region		Score [0 - 4]
Basal ganglia	R	
	L	
Centrum semiovale	R	
	L	

0: none **1:** < 10 **2:** 11 - 20 **3:** 21 - 40 **4:** > 40.

(Rating as per Potter, G.M., et al. Int. J. Stroke, 2015. 10(3): 376-81)

Is there focal perivascular space enlargement related to any lesion [Y/N] ?

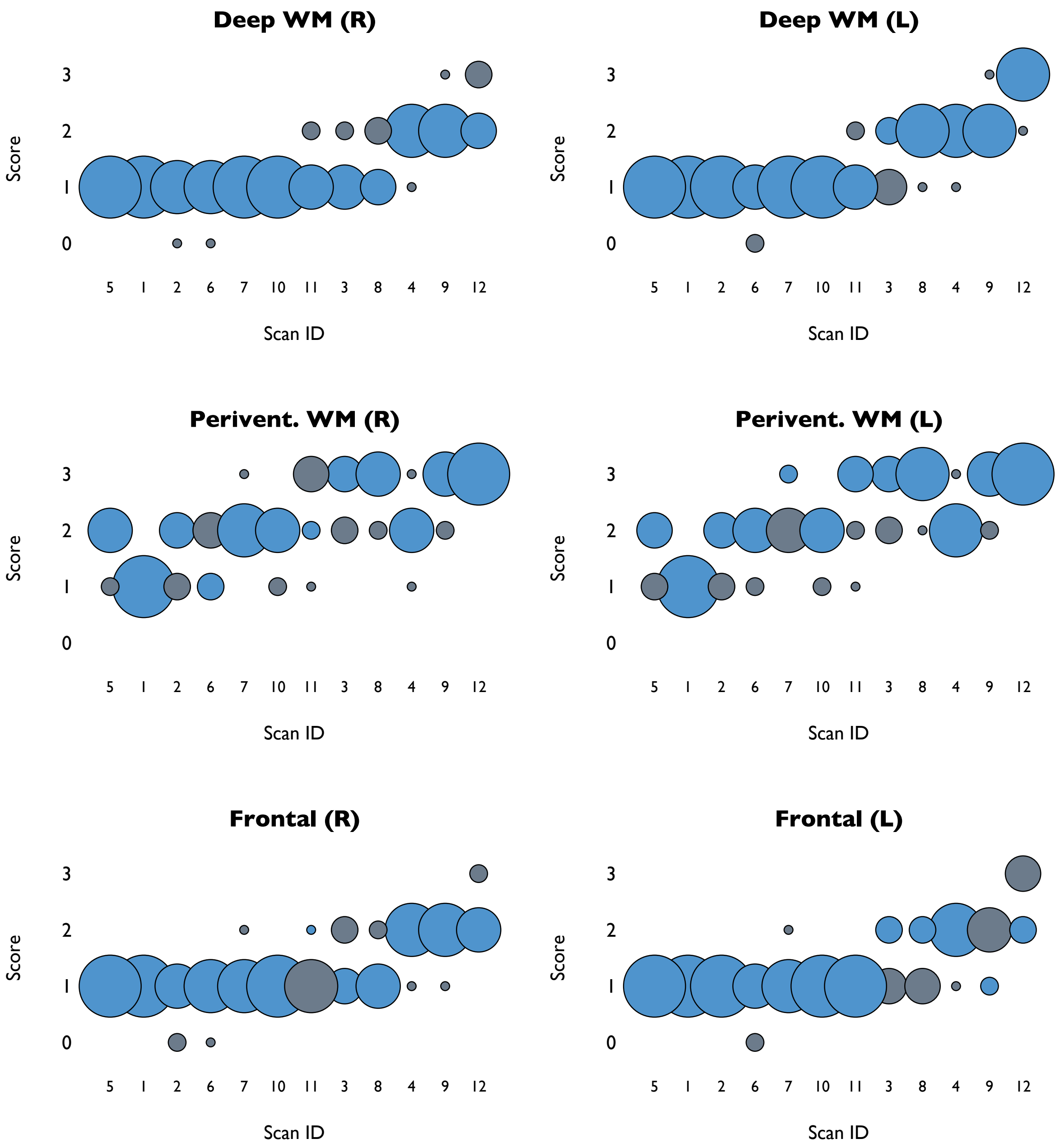
If yes, give location:

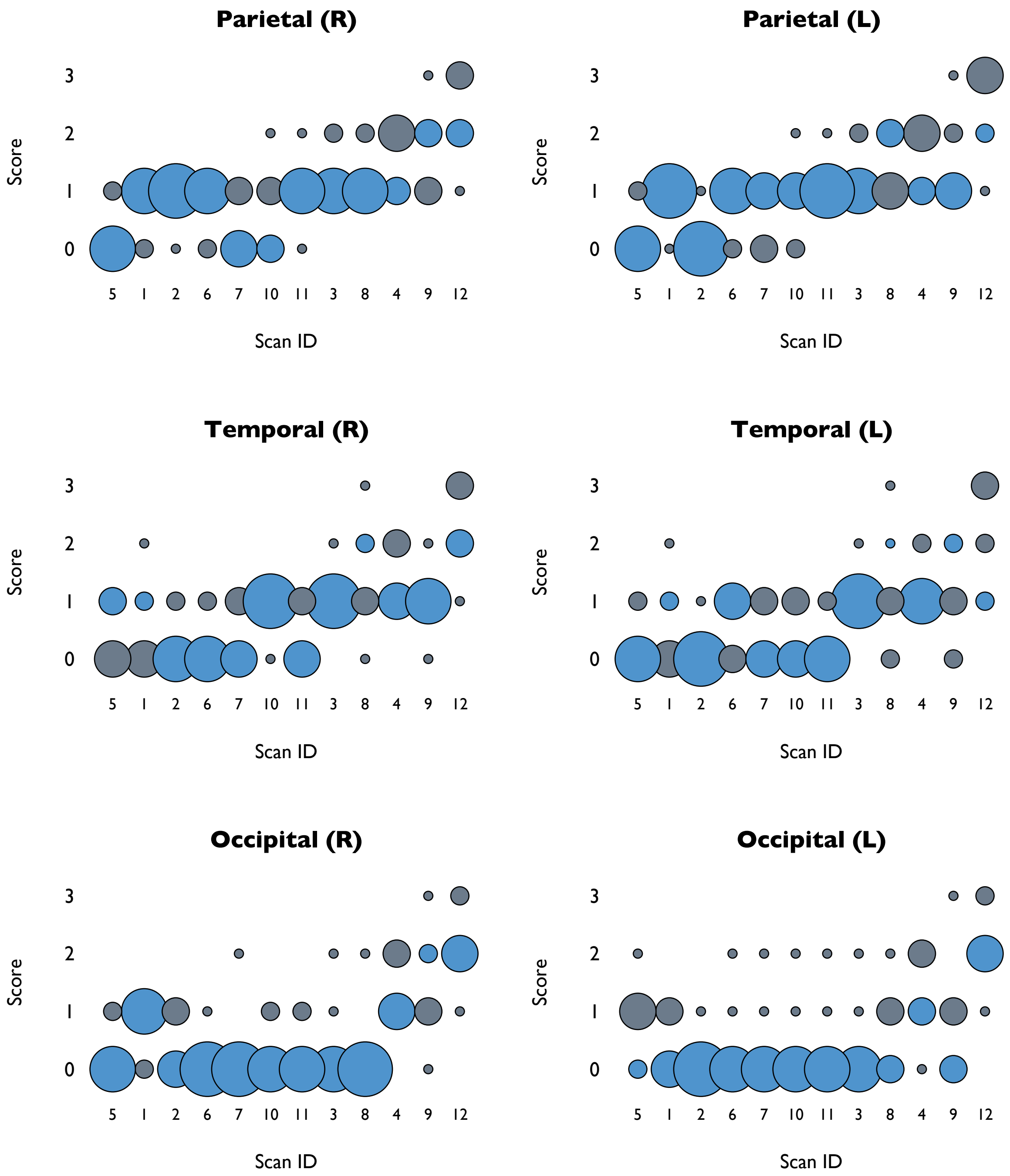
Any other findings/comments?

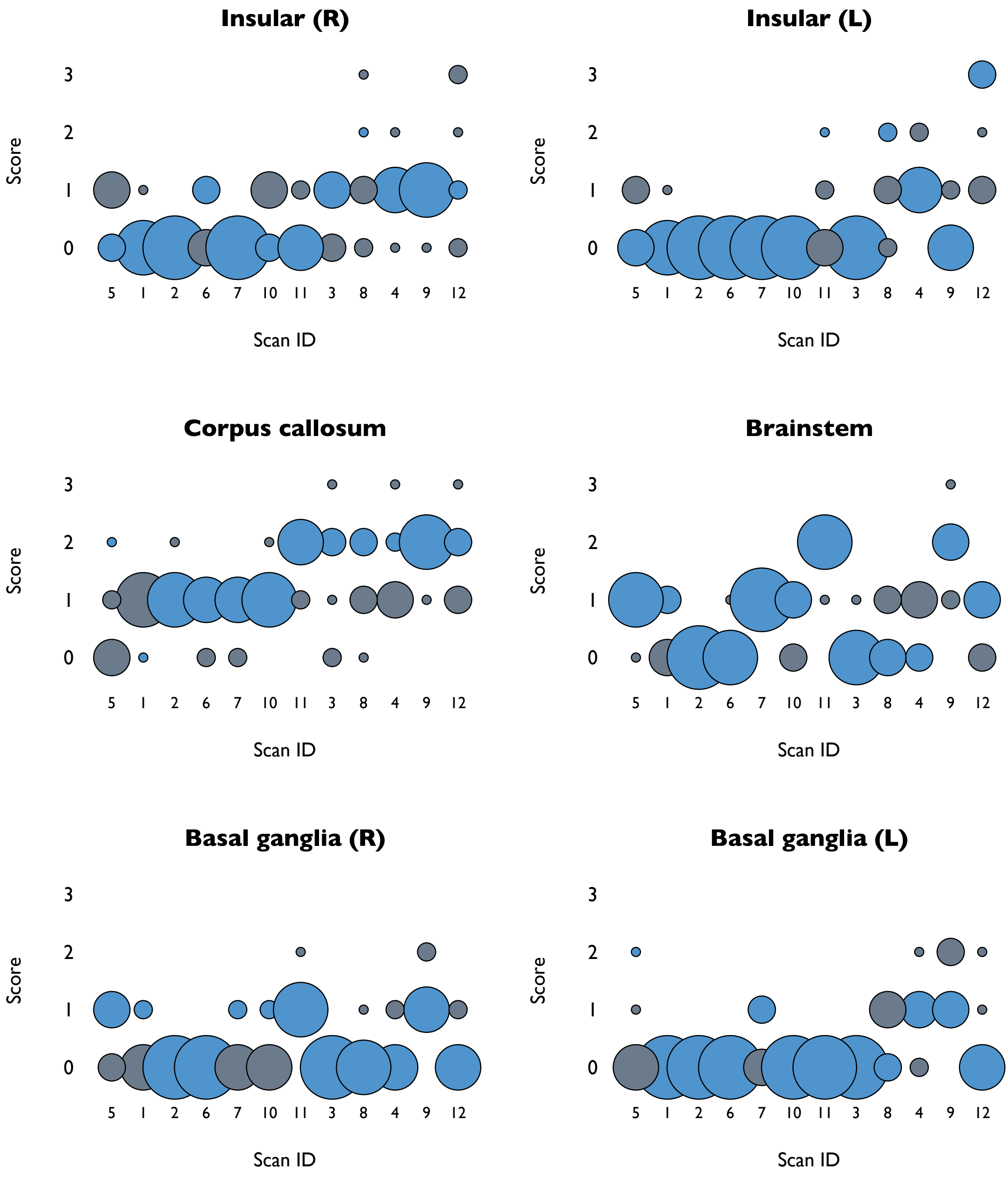
Total assessment time?

Appendix G

‘Bubble’ plots showing inter-rater agreement in second pilot study of visual rating scale for MS imaging features


Deep WM (R)
Deep WM (L)
Perivent. WM (R)
Perivent. WM (L)
Frontal (R)
Frontal (L)
Score
0
1
2
3
Scan ID
5
1
2
6
7
10
11
3
8
4
9
12


Parietal (R)
Parietal (L)
Temporal (R)
Temporal (L)
Occipital (R)
Occipital (L)
Score
0
1
2
3
5
1
2
6
7
10
11
3
8
4
9
12
Scan ID


Insular (R)

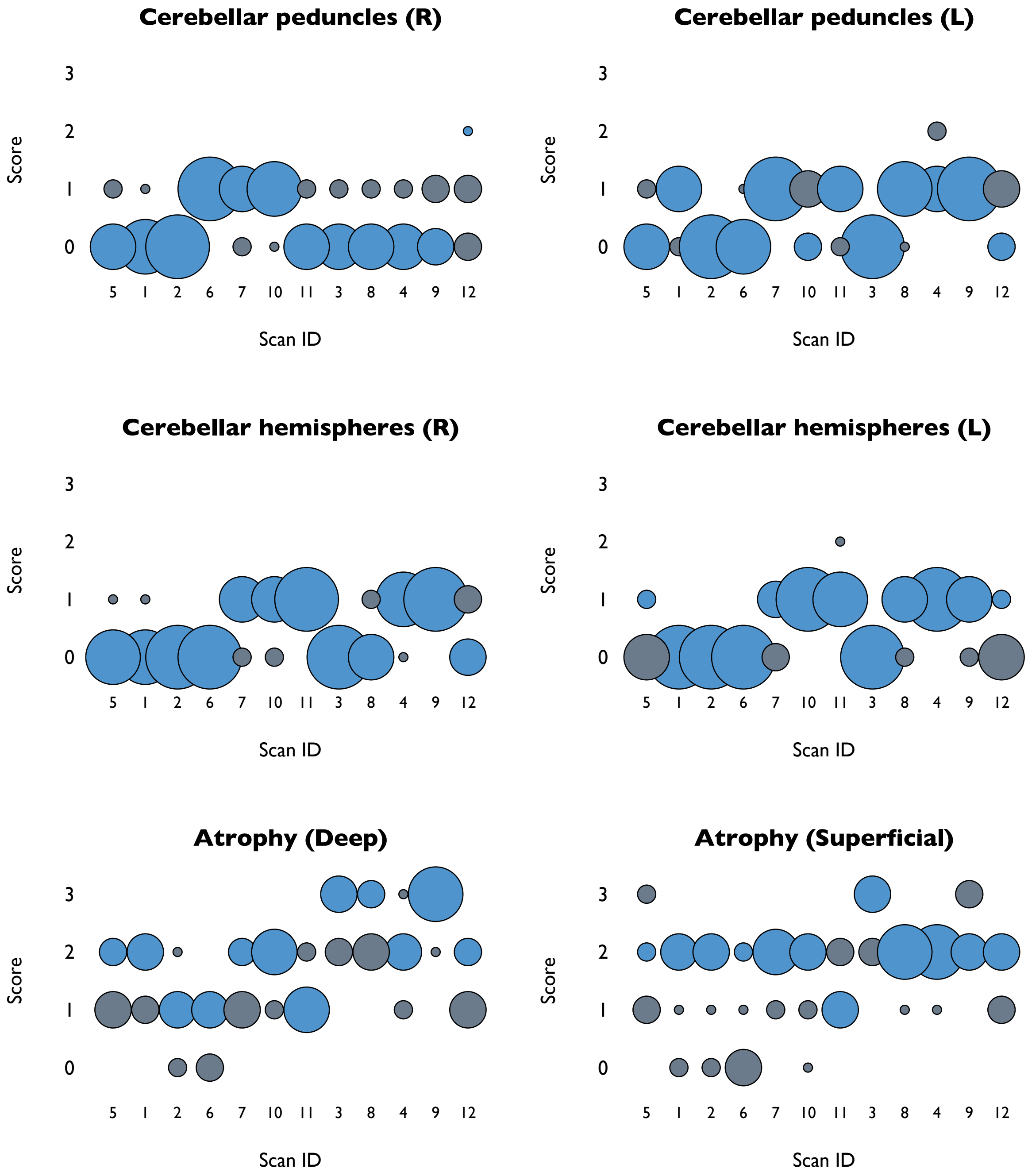
Insular (L)

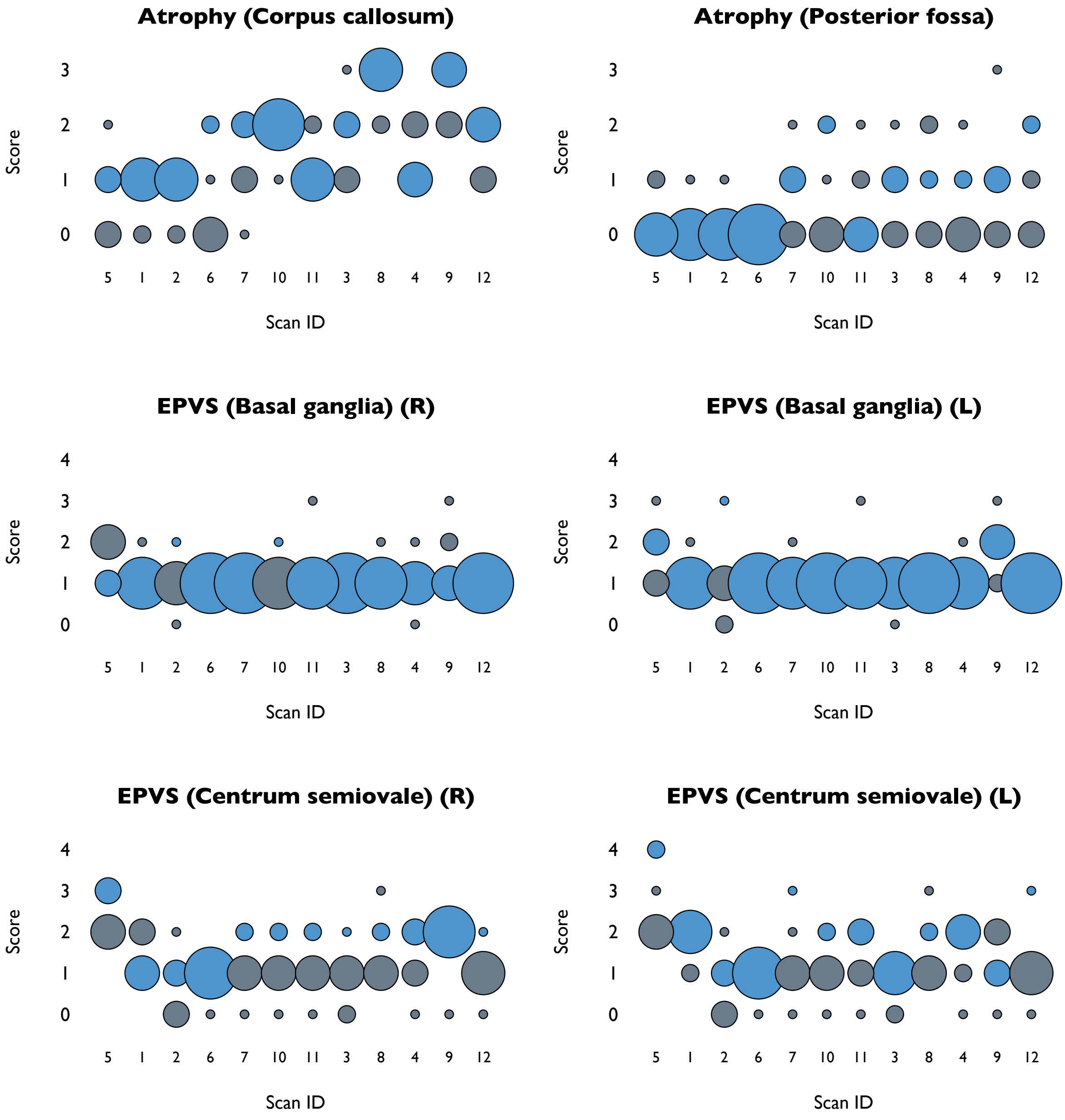
Corpus callosum

Brainstem

Basal ganglia (R)

Basal ganglia (L)


Cerebellar peduncles (R)
Cerebellar peduncles (L)
Cerebellar hemispheres (R)
Cerebellar hemispheres (L)
Atrophy (Deep)
Atrophy (Superficial)
Score
0
1
2
3
Scan ID
5
1
2
6
7
10
11
3
8
4
9
12


Atrophy (Corpus callosum)
Atrophy (Posterior fossa)
EPVS (Basal ganglia) (R)
EPVS (Basal ganglia) (L)
EPVS (Centrum semiovale) (R)
EPVS (Centrum semiovale) (L)
Score
0
1
2
3
4
Scan ID
5
1
2
6
7
10
11
3
8
4
9
12

Bibliography

[1] Charcot J. *Histologie de la sclerose en plaques.* Gaz Hop (Paris), 1868; 41:554,557–558,566.

[2] Compston A. *The 150th anniversary of the first depiction of the lesions of multiple sclerosis.* J Neurol Neurosurg Psychiatry, 1988; 51(10):1249–52.

[3] Medaer R. *Does the history of multiple sclerosis go back as far as the 14th century?* Acta Neurol Scand, 1979; 60(3):189–92.

[4] Multiple Sclerosis International Federation. *Atlas of MS 2013.* 2013.

[5] Koch-Henriksen N, Sorensen P. *The changing demographic pattern of multiple sclerosis epidemiology.* Lancet Neurol, 2010; 9(5):520–32.

[6] Mackenzie I, Morant S, Bloomfield G, MacDonald T, O'Riordan J. *Incidence and prevalence of multiple sclerosis in the UK 1990-2010: a descriptive study in the General Practice Research Database.* J Neurol Neurosurg Psychiatry, 2014; 85(1):76–84.

[7] Lublin F, Reingold S, Cohen J, Cutter G, Sorensen P, Thompson A, Wolinsky J, Balcer L, Banwell B, Barkhof F, Bebo B J, Calabresi P, Clanet M, Comi G, Fox R, Freedman M, Goodman A, Inglese M, Kappos L, Kieseier B, Lincoln J, Lubetzki C, Miller A, Montalban X, O'Connor P, Petkau J, Pozzilli C, Rudick R, Sormani M, Stuve O, Waubant E, Polman C. *Defining the clinical course of multiple sclerosis: the 2013 revisions.* Neurology, 2014; 83(3):278–86.

[8] Koch M, Kingwell E, Rieckmann P, Tremlett H. *The natural history of secondary progressive multiple sclerosis.* J Neurol Neurosurg Psychiatry, 2010; 81(9):1039–43.

[9] Filippi M, Rocca M, Ciccarelli O, De Stefano N, Evangelou N, Kappos L, Rovira A, Sastre-Garriga J, Tintore M, Frederiksen J, Gasperini C, Palace J, Reich D, Banwell B, Montalban X, Barkhof F. *MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines.* Lancet Neurol, 2016; 15(3):292–303.

[10] Compston A, Coles A. *Multiple sclerosis.* Lancet, 2008; 372(9648):1502–17.

[11] Baranzini S, Oksenberg J. *The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years.* Trends Genet, 2017; 33(12):960–970.

[12] Gourraud P, Harbo H, Hauser S, Baranzini S. *The genetics of multiple sclerosis: an up-to-date review.* Immunol Rev, 2012; 248(1):87–103.

[13] Okuda D, Srinivasan R, Oksenberg J, Goodin D, Baranzini S, Beheshtian A, Waubant E, Zamvil S, Leppert D, Qualley P, Lincoln R, Gomez R, Caillier S, George M, Wang J, Nelson S, Cree B, Hauser S, Pelletier D. *Genotype-Phenotype correlations in multiple sclerosis: HLA genes influence disease severity inferred by 1HMR spectroscopy and MRI measures.* Brain, 2009; 132(Pt 1):250–9.

[14] Reynolds R, Roncaroli F, Nicholas R, Radotra B, Gveric D, Howell O. *The neuropathological basis of clinical progression in multiple sclerosis.* Acta Neuropathol, 2011; 122(2):155–70.

[15] Dutta R, Trapp B. *Relapsing and progressive forms of multiple sclerosis: insights from pathology.* Curr Opin Neurol, 2014; 27(3):271–8.

[16] Patrikios P, Stadelmann C, Kutzelnigg A, Rauschka H, Schmidbauer M, Laursen H, Sorensen P, Bruck W, Lucchinetti C, Lassmann H. *Remyelination is extensive in a subset of multiple sclerosis patients.* Brain, 2006; 129(Pt 12):3165–72.

[17] Stadelmann C, Wegner C, Bruck W. *Inflammation, demyelination, and degeneration - recent insights from MS pathology.* Biochim Biophys Acta, 2011; 1812(2):275–82.

[18] Procaccini C, De Rosa V, Pucino V, Formisano L, Matarese G. *Animal models of multiple sclerosis.* Eur J Pharmacol, 2015; 759:182–91.

[19] Patani R, Chandran S. *Experimental and therapeutic opportunities for stem cells in multiple sclerosis.* Int J Mol Sci, 2012; 13(11):14470–91.

[20] Readhead C, Hood L. *The dysmyelinating mouse mutations shiverer (shi) and myelin deficient (shimld).* Behav Genet, 1990; 20(2):213–34.

[21] Baker D, Amor S. *Experimental autoimmune encephalomyelitis is a good model of multiple sclerosis if used wisely.* Mult Scler Relat Disord, 2014; 3(5):555–64.

[22] Praet J, Guglielmetti C, Berneman Z, Van der Linden A, Ponsaerts P. *Cellular and molecular neuropathology of the cuprizone mouse model: clinical relevance for multiple sclerosis.* Neurosci Biobehav Rev, 2014; 47:485–505.

[23] Torkildsen O, Myhr K, Bo L. *Disease-modifying treatments for multiple sclerosis - a review of approved medications.* Eur J Neurol, 2016; 23 Suppl 1:18–27.

[24] Chiaravalloti N, DeLuca J. *Cognitive impairment in multiple sclerosis.* Lancet Neurol, 2008; 7(12):1139–51.

[25] Anhoque C, Domingues S, Teixeira A, Domingues R. *Cognitive impairment in clinically isolated syndrome: A systematic review.* Dement Neuropsychol, 2010; 4(2):86–90.

[26] Achiron A, Chapman J, Magalashvili D, Dolev M, Lavie M, Bercovich E, Polliack M, Doniger G, Stern Y, Khilkevich O, Menascu S, Hararai G, Gurevich M, Barak Y. *Modeling of cognitive impairment by disease duration in multiple sclerosis: a cross-sectional study.* PLoS One, 2013; 8(8):e71058.

[27] Campbell J, Rashid W, Cercignani M, Langdon D. *Cognitive impairment among patients with multiple sclerosis: associations with employment and quality of life.* Postgrad Med J, 2017; 93(1097):143–147.

[28] Langdon D. *Cognition in multiple sclerosis.* Curr Opin Neurol, 2011; 24(3):244–9.

[29] Morrow S, Drake A, Zivadinov R, Munschauer F, Weinstock-Guttman B, Benedict R. *Predicting loss of employment over three years in multiple sclerosis: clinically meaningful cognitive decline.* Clin Neuropsychol, 2010; 24(7):1131–45.

[30] Lyall D, Cullen B, Allerhand M, Smith D, Mackay D, Evans J, Anderson J, Fawns-Ritchie C, McIntosh A, Deary I, Pell J. *Cognitive test scores in UK Biobank: Data reduction in 480,416 participants and longitudinal stability in 20,346 participants.* PLoS One, 2016; 11(4):e0154222.

[31] Mcgrew K. *Editorial: CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research.* Intelligence, 2009; 37(1):1–10.

[32] Denney D, Lynch S, Parmenter B, Horne N. *Cognitive impairment in relapsing and primary progressive multiple sclerosis: mostly a matter of speed.* J Int Neuropsychol Soc, 2004; 10(7):948–56.

[33] Rao S. *A Manual for the Brief Repeatable Battery of Neuropsychological Tests in Multiple Sclerosis.* Milwaukee,WI: Medical College of Wisconsin, 1990.

[34] Benedict R, Fischer J, Archibald C, Arnett P, Beatty W, Bobholz J, Chelune G, Fisk J, Langdon D, Caruso L, Foley F, LaRocca N, Vowels L,

Weinstein A, DeLuca J, Rao S, Munschauer F. *Minimal neuropsychological assessment of MS patients: a consensus approach.* Clin Neuropsychol, 2002; 16(3):381–97.

[35] Fischer M, Kunkel A, Bublak P, Faiss J, Hoffmann F, Sailer M, Schwab M, Zettl U, Kohler W. *How reliable is the classification of cognitive impairment across different criteria in early and late stages of multiple sclerosis?* J Neurol Sci, 2014; 343(1-2):91–9.

[36] Langdon D, Amato M, Boringa J, Brochet B, Foley F, Fredrikson S, Hamalainen P, Hartung H, Krupp L, Penner I, Reder A, Benedict R. *Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS).* Mult Scler, 2012; 18(6):891–8.

[37] Smerbeck A, Benedict R, Eshaghi A, Vanotti S, Spedo C, Blahova Dusankova J, Sahraian M, Marques V, Langdon D. *Influence of nationality on the Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS).* Clin Neuropsychol, 2017; 1–9.

[38] Sumowski J, Rocca M, Leavitt V, Riccitelli G, Comi G, DeLuca J, Filippi M. *Brain reserve and cognitive reserve in multiple sclerosis: what you've got and how you use it.* Neurology, 2013; 80(24):2186–93.

[39] Sumowski J. *Cognitive reserve as a useful concept for early intervention research in multiple sclerosis.* Front Neurol, 2015; 6:176.

[40] Kiely K, Butterworth P, Watson N, Wooden M. *The Symbol Digit Modalities Test: Normative data from a large nationally representative sample of Australians.* Arch Clin Neuropsychol, 2014; 29(8):767–75.

[41] Hinderer S, Liberty K. *Effects of baclofen on oral counting, arithmetic, and symbol decoding: An explorative multiple-baseline design across subjects.* International Journal of Rehabilitation and Health, 1996; 2(1):41–55.

[42] Gawryluk J R, Mazerolle E L, Beyea S D, D'Arcy R C. *Functional MRI activation in white matter during the Symbol Digit Modalities Test.* Front Hum Neurosci, 2014; 8:589.

[43] Thompson A J, Banwell B L, Barkhof F, Carroll W M, Coetzee T, Comi G, Correale J, Fazekas F, Filippi M, Freedman M S, Fujihara K, Galetta S L, Hartung H P, Kappos L, Lublin F D, Marrie R A, Miller A E, Miller D H, Montalban X, Mowry E M, Sorensen P S, Tintore M, Traboulsee A L, Trojano M, Uitdehaag B M J, Vukusic S, Waubant E, Weinshenker B G, Reingold S C, Cohen J A. *Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria.* Lancet Neurol, 2018; 17(2):162–173.

[44] Jonkman L, Soriano A, Amor S, Barkhof F, van der Valk P, Vrenken H, Geurts J. *Can MS lesion stages be distinguished with MRI? A postmortem MRI and histopathology study.* J Neurol, 2015; 262(4):1074–80.

[45] Wuerfel J, Haertle M, Waiczies H, Tysiak E, Bechmann I, Wernecke K, Zipp F, Paul F. *Perivascular spaces–MRI marker of inflammatory activity in the brain?* Brain, 2008; 131(Pt 9):2332–40.

[46] Corlobé A, Renard D, Goizet C, Berger E, Rumbach L, Robinson A, Dupuy D, Touzé E, Zéphir H, Vermersch P, Brochet B, Edan G, Deburghgraeve V, Crange A, Castelnovo G, Cohen M, Lebrun-Frenay C, Boespflug-Tanguy O, Labauge P. *Cavitary lesions in multiple sclerosis: multicenter study on twenty patients.* Rev Neurol (Paris), 2013; 169(12):965–9.

[47] Rovira A, Auger C, Alonso J. *Magnetic resonance monitoring of lesion evolution in multiple sclerosis.* Ther Adv Neurol Disord, 2013; 6(5):298–310.

[48] Sahraian M A, Radue E W, Haller S, Kappos L. *Black holes in multiple sclerosis: definition, evolution, and clinical correlations.* Acta Neurol Scand, 2010; 122(1):1–8.

[49] Aliaga E, Barkhof F. *MRI mimics of multiple sclerosis.* Handb Clin Neurol, 2014; 122:291–316.

[50] Seewann A, Vrenken H, van der Valk P, Blezer E L, Knol D L, Castelijns J A, Polman C, Pouwels P, Barkhof F, Geurts J J. *Diffusely abnormal white matter in chronic multiple sclerosis: imaging and histopathologic analysis.* Arch Neurol, 2009; 66(5):601–9.

[51] Vrenken H, Seewann A, Knol D L, Polman C H, Barkhof F, Geurts J J. *Diffusely abnormal white matter in progressive multiple sclerosis: in vivo quantitative MR imaging characterization and comparison between disease types.* AJNR Am J Neuroradiol, 2010; 31(3):541–8.

[52] Barkhof F. *The clinico-radiological paradox in multiple sclerosis revisited.* Curr Opin Neurol, 2002; 15(3):239–45.

[53] Palace J. *Making the diagnosis of multiple sclerosis.* J Neurol Neurosurg Psychiatry, 2001; 71 Suppl 2:ii3–8.

[54] Barkhof F, Scheltens P, Frequin S, Nauta J, Tas M, Valk J, Hommes O. *Relapsing-remitting multiple sclerosis: sequential enhanced MR imaging vs clinical findings in determining disease activity.* AJR Am J Roentgenol, 1992; 159(5):1041–7.

[55] Maranzano J, Rudko D, Arnold D, Narayanan S. *Manual segmentation of MS cortical lesions using MRI: a comparison of 3 MRI reading protocols.* AJNR Am J Neuroradiol, 2016; 37(9):1623–8.

[56] Faizy T, Thaler C, Ceyrowski T, Broocks G, Treffler N, Sedlacik J, Sturner K, Stellmann J, Heesen C, Fiehler J, Siemonsen S. *Reliability of cortical lesion detection on double inversion recovery MRI applying the MAGNIMS-Criteria in multiple sclerosis patients within a 16-months period.* PLoS One, 2017; 12(2):e0172923.

[57] Roosendaal S, Moraal B, Pouwels P, Vrenken H, Castelijns J, Barkhof F, Geurts J. *Accumulation of cortical lesions in MS: relation with cognitive impairment.* Mult Scler, 2009; 15(6):708–14.

[58] Llado X, Oliver A, Cabezas M, Freixenet J, Vilanova J, Quiles A, Valls L, Ramio-Torrenta L, Rovira A. *Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches.* Information Sciences, 2012; 186(1):164–185.

[59] Stevens J, Farlow M, Edwards M, Yu P. *Magnetic resonance imaging. Clinical correlation in 64 patients with multiple sclerosis.* Arch Neurol, 1986; 43(11):1145–8.

[60] Ormerod I, Miller D, McDonald W, du Boulay E, Rudge P, Kendall B, Moseley I, Johnson G, Tofts P, Halliday A, et al. *The role of NMR imaging in the assessment of multiple sclerosis and isolated neurological lesions. A quantitative study.* Brain, 1987; 110 (Pt 6):1579–616.

[61] Huber S, Paulson G, Shuttleworth E, Chakeres D, Clapp L, Pakalnis A, Weiss K, Rammohan K. *Magnetic resonance imaging correlates of dementia in multiple sclerosis.* Arch Neurol, 1987; 44(7):732–6.

[62] Dietemann J, Beigelman C, Rumbach L, Vouge M, Tajahmady T, Faubert C, Jeung M, Wackenheim A. *Multiple sclerosis and corpus callosum atrophy: relationship of MRI findings to clinical data.* Neuroradiology, 1988; 30(6):478–80.

[63] Franklin G, Heaton R, Nelson L, Filley C, Seibert C. *Correlation of neuropsychological and MRI findings in chronic/progressive multiple sclerosis.* Neurology, 1988; 38(12):1826–9.

[64] van Walderveen M, Barkhof F, Tas M, Polman C, Frequin S, Hommes O, Thompson A, Valk J. *Patterns of brain magnetic resonance abnormalities on T2-weighted spin echo images in clinical subgroups of multiple sclerosis: a large cross-sectional study.* Eur Neurol, 1998; 40(2):91–8.

[65] Scheltens P, Erkinjunti T, Leys D, Wahlund L, Inzitari D, del Ser T, Pasquier F, Barkhof F, Mantyla R, Bowler J, Wallin A, Ghika J, Fazekas F, Pantoni L. *White matter changes on CT and MRI: an overview of visual rating scales. European Task Force on Age-Related White Matter Changes.* Eur Neurol, 1998; 39(2):80–9.

[66] De Stefano N, Airas L, Grigoriadis N, Mattle H, O'Riordan J, Oreja-Guevara C, Sellebjerg F, Stankoff B, Walczak A, Wiendl H, Kieseier B. *Clinical relevance of brain volume measures in multiple sclerosis.* CNS Drugs, 2014; 28(2):147–56.

[67] De Stefano N, Giorgio A, Battaglini M, Rovaris M, Sormani M, Barkhof F, Korteweg T, Enzinger C, Fazekas F, Calabrese M, Dinacci D, Tedeschi G, Gass A, Montalban X, Rovira A, Thompson A, Comi G, Miller D, Filippi M. *Assessing brain atrophy rates in a large population of untreated multiple sclerosis subtypes.* Neurology, 2010; 74(23):1868–76.

[68] Moccia M, de Stefano N, Barkhof F. *Imaging outcome measures for progressive multiple sclerosis trials.* Mult Scler, 2017; 23(12):1614–1626.

[69] Nakamura K, Brown R, Araujo D, Narayanan S, Arnold D. *Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: implications for monitoring atrophy in clinical studies.* Neuroimage Clin, 2014; 6:166–70.

[70] Mukherjee P, Berman J, Chung S, Hess C, Henry R. *Diffusion tensor MR imaging and fiber tractography: theoretic underpinnings.* AJNR Am J Neuroradiol, 2008; 29(4):632–41.

[71] Rovaris M, Iannucci G, Falautano M, Possa F, Martinelli V, Comi G, Filippi M. *Cognitive dysfunction in patients with mildly disabling relapsing-remitting multiple sclerosis: an exploratory study with diffusion tensor MR imaging.* Journal of the Neurological Sciences, 2002; 195(2):103–109.

[72] Rovaris M, Riccitelli G, Judica E, Possa F, Caputo D, Ghezzi A, Bertolotto A, Capra R, Falautano M, Mattioli F, Martinelli V, Comi G, Filippi M. *Cognitive impairment and structural brain damage in benign multiple sclerosis.* Neurology, 2008; 71(19):1521–6.

[73] Akbar N, Lobaugh N, O'Connor P, Moradzadeh L, Scott C, Feinstein A. *Diffusion tensor imaging abnormalities in cognitively impaired multiple sclerosis patients.* Can J Neurol Sci, 2010; 37(5):608–14.

[74] Mesaros S, Rocca M, Kacar K, Kostic J, Copetti M, Stosic-Opincal T, Preziosa P, Sala S, Riccitelli G, Horsfield M, Drulovic J, Comi G, Filippi

M. *Diffusion tensor MRI tractography and cognitive impairment in multiple sclerosis.* Neurology, 2012; 78(13):969–975.

[75] Smith S, Jenkinson M, Johansen-Berg H, Rueckert D, Nichols T, Mackay C, Watkins K, Ciccarelli O, Cader M, Matthews P, Behrens T. *Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data.* Neuroimage, 2006; 31(4):1487–505.

[76] Roosendaal S, Geurts J, Vrenken H, Hulst H, Cover K, Castelijns J, Pouwels P, Barkhof F. *Regional DTI differences in multiple sclerosis patients.* Neuroimage, 2009; 44(4):1397–1403.

[77] Dineen R, Vilisaar J, Hlinka J, Bradshaw C, Morgan P, Constantinescu C, Auer D. *Disconnection as a mechanism for cognitive dysfunction in multiple sclerosis.* Brain, 2009; 132(Pt 1):239–49.

[78] Yu H, Christodoulou C, Bhise V, Greenblatt D, Patel Y, Serafin D, Maletic-Savatic M, Krupp L, Wagshul M. *Multiple white matter tract abnormalities underlie cognitive impairment in RRMS.* Neuroimage, 2012; 59(4):3713–3722.

[79] Warlop N, Achten E, Fieremans E, Debruyne J, Vingerhoets G. *Transverse diffusivity of cerebral parenchyma predicts visual tracking performance in relapsing-remitting multiple sclerosis.* Brain and Cognition, 2009; 71(3):410–415.

[80] Van Hecke W, Nagels G, Leemans A, Vandervliet E, Sijbers J, Parizel P. *Correlation of cognitive dysfunction and diffusion tensor MRI measures in patients with mild and moderate multiple sclerosis.* Journal of Magnetic Resonance Imaging, 2010; 31(6):1492–1498.

[81] Shu N, Duan Y, Xia M, Schoonheim M, Huang J, Ren Z, Sun Z, Ye J, Dong H, Shi F, Barkhof F, Li K, Liu Y. *Disrupted topological organization of structural and functional brain connectomes in clinically isolated syndrome and multiple sclerosis.* Sci Rep, 2016; 6:29383.

[82] Schoonheim M, Vigeveno R, Lopes F, Pouwels P, Polman C, Barkhof F, Geurts J. *Sex-specific extent and severity of white matter damage in multiple sclerosis: Implications for cognitive decline.* Human Brain Mapping, 2014; 35(5):2348–2358.

[83] Sbardella E, Petsas N, Tona F, Prosperini L, Raz E, Pace G, Pozzilli C, Pantano P. *Assessing the correlation between grey and white matter damage with motor and cognitive impairment in multiple sclerosis patients.* PLoS One, 2013; 8(5):e63250.

[84] Preziosa P, Rocca M, Pagani E, Stromillo M, Enzinger C, Gallo A, Hulst H, Atzori M, Pareto D, Riccitelli G, Copetti M, De Stefano N, Fazekas F, Bisecco A, Barkhof F, Yousry T, Arevalo M, Filippi M. *Structural MRI correlates of cognitive impairment in patients with multiple sclerosis: A multicenter study.* Human Brain Mapping, 2016; 37(4):1627–1644.

[85] Meijer K, Muhlert N, Cercignani M, Sethi V, Ron M, Thompson A, Miller D, Chard D, Geurts J, Ciccarelli O. *White matter tract abnormalities are associated with cognitive dysfunction in secondary progressive multiple sclerosis.* Mult Scler, 2016; 22(11):1429–1437.

[86] Mazerolle E, Wojtowicz M, Omisade A, Fisk J. *Intra-individual variability in information processing speed reflects white matter microstructure in multiple sclerosis.* Neuroimage-Clinical, 2013; 2:894–902.

[87] Llufriu S, Martinez-Heras E, Fortea J, Blanco Y, Berenguer J, Gabilondo I, Ibarretxe-Bilbao N, Falcon C, Sepulveda M, Sola-Valls N, Bargallo N, Graus F, Villoslada P, Saiz A. *Cognitive functions in multiple sclerosis: impact of gray matter integrity.* Multiple Sclerosis Journal, 2014; 20(4):424–432.

[88] Hulst H, Steenwijk M, Versteeg A, Pouwels P, Vrenken H, Uitdehaag B, Polman C, Geurts J, Barkhof F. *Cognitive impairment in MS: impact of white matter integrity, gray matter volume, and lesions.* Neurology, 2013; 80(11):1025–32.

[89] Genova H, DeLuca J, Chiaravalloti N, Wylie G. *The relationship between executive functioning, processing speed, and white matter integrity in multiple sclerosis.* Journal of Clinical and Experimental Neuropsychology, 2013; 35(6):631–641.

[90] Francis P, Chia T, Jakubovic R, O'Connor P, Lee L, Feinstein A, Aviv R. *Extensive white matter dysfunction in cognitively impaired patients with secondary-progressive multiple sclerosis.* American Journal of Neuroradiology, 2014; 35(10):1910–1915.

[91] Bozzali M, Spano B, Parker G, Giulietti G, Castelli M, Basile B, Rossi S, Serra L, Magnani G, Nocentini U, Caltagirone C, Centonze D, Cercignani M. *Anatomical brain connectivity can assess cognitive dysfunction in multiple sclerosis.* Multiple Sclerosis Journal, 2013; 19(9):1161–1168.

[92] Baykara E, Gesierich B, Adam R, Tuladhar A, Biesbroek J, Koek H, Ropele S, Jouvent E, Chabriat H, Ertl-Wagner B, Ewers M, Schmidt R, de Leeuw F, Biessels G, Dichgans M, Duering M. *A novel imaging marker for small vessel disease based on skeletonization of white matter tracts and diffusion histograms.* Ann Neurol, 2016; 80(4):581–92.

[93] Clayden J, Storkey A, Bastin M. *A probabilistic model-based approach to consistent white matter tract segmentation.* IEEE Trans Med Imaging, 2007; 26(11):1555–61.

[94] Kapeller P, Brex P, Chard D, Dalton C, Griffin C, McLean M, Parker G, Thompson A, Miller D. *Quantitative 1H MRS imaging 14 years after presenting with a clinically isolated syndrome suggestive of multiple sclerosis.* Mult Scler, 2002; 8(3):207–10.

[95] Rovira A, Alonso J. *1H magnetic resonance spectroscopy in multiple sclerosis and related disorders.* Neuroimaging Clin N Am, 2013; 23(3):459–74.

[96] Schmierer K, Scaravilli F, Altmann D, Barker G, Miller D. *Magnetization transfer ratio and myelin in postmortem multiple sclerosis brain.* Ann Neurol, 2004; 56(3):407–15.

[97] Mallik S, Samson R, Wheeler-Kingshott C, Miller D. *Imaging outcomes for trials of remyelination in multiple sclerosis.* J Neurol Neurosurg Psychiatry, 2014; 85(12):1396–404.

[98] Witte O. *Lesion-induced plasticity as a potential mechanism for recovery and rehabilitative training.* Curr Opin Neurol, 1998; 11(6):655–62.

[99] Reddy H, Narayanan S, Arnoutelis R, Jenkinson M, Antel J, Matthews P, Arnold D. *Evidence for adaptive functional changes in the cerebral cortex with axonal injury from multiple sclerosis.* Brain, 2000; 123 (Pt 11):2314–20.

[100] Filippi M, Rocca M. *Present and future of fMRI in multiple sclerosis.* Expert Rev Neurother, 2013; 13(12 Suppl):27–31.

[101] Mainero C, Caramia F, Pozzilli C, Pisani A, Pestalozza I, Borriello G, Bozzao L, Pantano P. *fMRI evidence of brain reorganization during attention and memory tasks in multiple sclerosis.* Neuroimage, 2004; 21(3):858–67.

[102] Rao S, Martin A, Huelin R, Wissinger E, Khankhel Z, Kim E, Fahrbach K. *Correlations between MRI and information processing speed in MS: A meta-analysis.* Mult Scler Int, 2014; 2014:975803.

[103] Spearman C. *The proof and measurement of association between two things.* . Am J Psychol, 1904; 15(1):72–101.

[104] Smith S, Jenkinson M, Woolrich M, Beckmann C, Behrens T, Johansen-Berg H, Bannister P, De Luca M, Drobnjak I, Flitney D, Niazy R, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady J,

Matthews P. *Advances in functional and structural MR image analysis and implementation as FSL.* Neuroimage, 2004; 23 Suppl 1:S208–19.

[105] Avants B, Gee J. *Geodesic estimation for large deformation anatomical shape averaging and interpolation.* Neuroimage, 2004; 23 Suppl 1:S139–50.

[106] Landman B, Ribbens A, Lucas B, Davatzikos C, Avants B, Ledig C, Ma D, Rueckert D, Vandermeulen D, Maes F, Erus G, Wang J, Holmes H, Wang H, Doshi J, Kornegay J, Manjon J, Hammers A, Akhondi-Asl A, Asman A, Warfield S. *MICCAI 2012 Workshop on Multi-Atlas Labeling.* CreateSpace Independent Publishing Platform, 2012.

[107] Dickie D. *Methods to assess changes in human brain structure across the lifecourse.* PhD Thesis, University of Edinburgh, 2014.

[108] Dice L. *Measures of the amount of ecologic association between species.* Ecology, 1945; 26(3):297–302.

[109] Cutter G R, Baier M L, Rudick R A, Cookfair D L, Fischer J S, Petkau J, Syndulko K, Weinshenker B G, Antel J P, Confavreux C, Ellison G W, Lublin F, Miller A E, Rao S M, Reingold S, Thompson A, Willoughby E. *Development of a multiple sclerosis functional composite as a clinical trial outcome measure.* Brain, 1999; 122 (Pt 5):871–82.

[110] Polman C, Reingold S, Banwell B, Clanet M, Cohen J, Filippi M, Fujihara K, Havrdova E, Hutchinson M, Kappos L, Lublin F, Montalban X, O'Connor P, Sandberg-Wollheim M, Thompson A, Waubant E, Weinshenker B, Wolinsky J. *Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria.* Ann Neurol, 2011; 69(2):292–302.

[111] Moher D, Liberati A, Tetzlaff J, Altman D. *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS Med, 2009; 6(7):e1000097.

[112] von Elm E, Altman D, Egger M, Pocock S, Gotzsche P, Vandenbroucke J. *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies.* J Clin Epidemiol, 2008; 61(4):344–9.

[113] Mike A, Glanz B, Hildenbrand P, Meier D, Bolden K, Liguori M, Dell'Oglio E, Healy B, Bakshi R, Guttmann C. *Identification and clinical impact of multiple sclerosis cortical lesions as assessed by routine 3T MR imaging.* AJNR Am J Neuroradiol, 2011; 32(3):515–21.

[114] Borenstein M. *Introduction to meta-analysis.* Chichester, U.K. : John Wiley & Sons, 2009.

[115] Higgins J, Thompson S, Deeks J, Altman D. *Measuring inconsistency in meta-analyses.* BMJ, 2003; 327(7414):557–60.

[116] Millar R. *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*, volume 111. John Wiley & Sons, 2011.

[117] Amato M, Portaccio E, Stromillo M, Goretti B, Zipoli V, Siracusa G, Battaglini M, Giorgio A, Bartolozzi M, Guidi L, Sorbi S, Federico A, De Stefano N. *Cognitive assessment and quantitative magnetic resonance metrics can help to identify benign multiple sclerosis.* Neurology, 2008; 71(9):632–8.

[118] Anzola G, Bevilacqua L, Cappa S, Capra R, Faglia L, Farina E, Frisoni G, Mariani C, Pasolini M, Vignolo L. *Neuropsychological assessment in patients with relapsing-remitting multiple sclerosis and mild functional impairment: correlation with magnetic resonance imaging.* J Neurol Neurosurg Psychiatry, 1990; 53(2):142–5.

[119] Archibald C, Wei X, Scott J, Wallace C, Zhang Y, Metz L, Mitchell J. *Posterior fossa lesion volume and slowed information processing in multiple sclerosis.* Brain, 2004; 127(Pt 7):1526–34.

[120] Benedict R, Bruce J, Dwyer M, Abdelrahman N, Hussein S, Weinstock-Guttman B, Garg N, Munschauer F, Zivadinov R. *Neocortical atrophy, third ventricular width, and cognitive dysfunction in multiple sclerosis.* Arch Neurol, 2006; 63(9):1301–6.

[121] Benedict R, Weinstock-Guttman B, Fishman I, Sharma J, Tjoa C, Bakshi R. *Prediction of neuropsychological impairment in multiple sclerosis: comparison of conventional magnetic resonance imaging measures of atrophy and lesion burden.* Arch Neurol, 2004; 61(2):226–30.

[122] Bermel R, Bakshi R, Tjoa C, Puli S, Jacobs L. *Bicaudate ratio as a magnetic resonance imaging marker of brain atrophy in multiple sclerosis.* Arch Neurol, 2002; 59(2):275–80.

[123] Bomboi G, Ikonomidou V, Pellegrini S, Stern S, Gallo A, Auh S, Evangelou I, Agarwal J, Pellicano C, Ohayon J, Cantor F, Ehrmantraut M, McFarland H, Kane R, Bagnato F. *Quality and quantity of diffuse and focal white matter disease and cognitive disability of patients with multiple sclerosis.* J Neuroimaging, 2011; 21(2):e57–63.

[124] Camp S, Stevenson V, Thompson A, Miller D, Borras C, Auriacombe S, Brochet B, Falautano M, Filippi M, Herisse-Dulo L, Montalban X, Parrcira E, Polman C, De Sa J, Langdon D. *Cognitive function in primary*

progressive and transitional progressive multiple sclerosis: a controlled study with MRI correlates. Brain, 1999; 122 (Pt 7):1341–8.

[125] Christodoulou C, Krupp L, Liang Z, Huang W, Melville P, Roque C, Scherl W, Morgan T, MacAllister W, Li L, Tudorica L, Li X, Roche P, Peyster R. *Cognitive performance and MR markers of cerebral injury in cognitively impaired MS patients.* Neurology, 2003; 60(11):1793–8.

[126] Comi G, Filippi M, Martinelli V, Campi A, Rodegher M, Alberoni M, Sirabian G, Canal N. *Brain MRI correlates of cognitive impairment in primary and secondary progressive multiple sclerosis.* J Neurol Sci, 1995; 132(2):222–7.

[127] Comi G, Rovaris M, Falautano M, Santuccio G, Martinelli V, Rocca M, Possa F, Leocani L, Paulesu E, Filippi M. *A multiparametric MRI study of frontal lobe dementia in multiple sclerosis.* J Neurol Sci, 1999; 171(2):135–44.

[128] Deloire M, Salort E, Bonnet M, Arimone Y, Boudineau M, Amieva H, Barroso B, Ouallet J, Pachai C, Galliaud E, Petry K, Dousset V, Fabrigoule C, Brochet B. *Cognitive impairment as marker of diffuse brain abnormalities in early relapsing remitting multiple sclerosis.* Journal of Neurology, Neurosurgery and Psychiatry, 2005; 76(4):519–526.

[129] Francis P, Jakubovic R, O'Connor P, Zhang L, Eilaghi A, Lee L, Carroll T, Mouannes-Srour J, Feinstein A, Aviv R. *Robust perfusion deficits in cognitively impaired patients with secondary-progressive multiple sclerosis.* AJNR Am J Neuroradiol, 2013; 34(1):62–7.

[130] Heesen C, Schulz K, Fiehler J, Von der Mark U, Otte C, Jung R, Poettgen J, Krieger T, Gold S. *Correlates of cognitive dysfunction in multiple sclerosis.* Brain Behav Immun, 2010; 24(7):1148–55.

[131] Hohol M, Guttmann C, Orav J, Mackin G, Kikinis R, Khoury S, Jolesz F, Weiner H. *Serial neuropsychological assessment and magnetic resonance imaging analysis in multiple sclerosis.* Arch Neurol, 1997; 54(8):1018–25.

[132] Houtchens M, Benedict R, Killiany R, Sharma J, Jaisani Z, Singh B, Weinstock-Guttman B, Guttmann C, Bakshi R. *Thalamic atrophy and cognition in multiple sclerosis.* Neurology, 2007; 69(12):1213–23.

[133] Izquierdo G, Campoy F J, Mir J, Gonzalez M, Martinez-Parra C. *Memory and learning disturbances in multiple sclerosis. MRI lesions and neuropsychological correlation.* Eur J Radiol, 1991; 13(3):220–4.

[134] Kalkers N, Bergers L, de Groot V, Lazeron R, van Walderveen M, Uitdehaag B, Polman C, Barkhof F. *Concurrent validity of the MS Functional*

Composite using MRI as a biological disease marker. Neurology, 2001; 56(2):215–9.

[135] Karlinska I, Siger M, Lewanska M, Selmaj K. *Cognitive impairment in patients with relapsing-remitting multiple sclerosis. The correlation with MRI lesion volume.* Neurol Neurochir Pol, 2008; 42(5):416–23.

[136] Krause M, Wendt J, Dressel A, Berneiser J, Kessler C, Hamm A, Lotze M. *Prefrontal function associated with impaired emotion recognition in patients with multiple sclerosis.* Behavioural Brain Research, 2009; 205(1):280–285.

[137] Laffon M, Malandain G, Joly H, Cohen M, Lebrun C. *The HV3 Score: A new simple tool to suspect cognitive impairment in multiple sclerosis in clinical practice.* Neurol Ther, 2014; 3(2):113–22.

[138] Lazeron R, Boringa J, Schouten M, Uitdehaag B, Bergers E, Lindeboom J, Eikelenboom M, Scheltens P, Barkhof F, Polman C. *Brain atrophy and lesion load as explaining parameters for cognitive impairment in multiple sclerosis.* Mult Scler, 2005; 11(5):524–31.

[139] Lazeron R, de Sonneville L, Scheltens P, Polman C, Barkhof F. *Cognitive slowing in multiple sclerosis is strongly associated with brain volume reduction.* Mult Scler, 2006; 12(6):760–8.

[140] Lin X, Tench C, Morgan P, Constantinescu C. *Use of combined conventional and quantitative MRI to quantify pathology related to cognitive impairment in multiple sclerosis.* J Neurol Neurosurg Psychiatry, 2008; 79(4):437–41.

[141] Lund H, Jonsson A, Andresen J, Rostrup E, Paulson O, Sorensen P. *Cognitive deficits in multiple sclerosis: correlations with T2 changes in normal appearing brain tissue.* Acta Neurol Scand, 2012; 125(5):338–44.

[142] Mike A, Strammer E, Aradi M, Orsi G, Perlaki G, Hajnal A, Sandor J, Banati M, Illes E, Zaitsev A, Herold R, Guttmann C, Illes Z. *Disconnection mechanism and regional cortical atrophy contribute to impaired processing of facial expressions and theory of mind in multiple sclerosis: a structural MRI study.* PLoS One, 2013; 8(12):e82422.

[143] Moller A, Wiedemann G, Rohde U, Backmund H, Sonntag A. *Correlates of cognitive impairment and depressive mood disorder in multiple sclerosis.* Acta Psychiatr Scand, 1994; 89(2):117–21.

[144] Niino M, Mifune N, Kohriyama T, Mori M, Ohashi T, Kawachi I, Shimizu Y, Fukaura H, Nakashima I, Kusunoki S, Miyamoto K, Yoshida K, Kanda T, Nomura K, Yamamura T, Yoshii F, Kira J, Nakane S, Yokoyama K, Matsui M, Miyazaki Y, Kikuchi S. *Association of cognitive impairment with magnetic resonance imaging findings and social activities in patients*

with multiple sclerosis. Clinical and Experimental Neuroimmunology, 2014; 5(3):328–335.

[145] Nocentini U, Rossini P, Carlesimo G, Graceffa A, Grasso M, Lupoi D, Oliveri M, Orlacchio A, Pozzilli C, Rizzato B, Caltagirone C. *Patterns of cognitive impairment in secondary progressive stable phase of multiple sclerosis: correlations with MRI findings.* Eur Neurol, 2001; 45(1):11–8.

[146] Parmenter B, Zivadinov R, Kerenyi L, Gavett R, Weinstock-Guttman B, Dwyer M, Garg N, Munschauer F, Benedict R. *Validity of the Wisconsin card sorting and Delis-Kaplan executive function system (DKEFS) sorting tests in multiple sclerosis.* J Clin Exp Neuropsychol, 2007; 29(2):215–23.

[147] Patti F, Di Stefano M, De Pascalis D, Ciancio M, De Bernardis E, Nicoletti F, Reggio A. *May there exist specific MRI findings predictive of dementia in multiple sclerosis patients?* Funct Neurol, 1995; 10(2):83–90.

[148] Patti F, Amato M, Trojano M, Bastianello S, Goretti B, Caniatti L, Di Monte E, Ferrazza P, Brescia Morra V, Lo Fermo S, Picconi O, Luccichenti G, Vecchio R, Maimone D, Messina S, Gasperini C, Orefice V, Florio C, Portaccio E, Zipoli V, Bertolotto A, Bramanti P, Sessa E, Centonze D, Cottone S, Salemi G, Falcini M, Gallo P, Perini P, Gigli G, Giuliani G, Grimaldi L, Murri L, Lugaresi A, Monaco F, Montanari E, Motti L, Neri S, Paciello M, Provinciali L, Ragno M, Rosati G, Ruggieri S, Tola M, Tonali P, Batocchi A, De Caro M, Ghezzi A, Zaffaroni M, Zolo P, Zorzon M, Signorino M, Scarpini E, Durelli L, Carolei A, Todaro M, Spitaleri D, Tartaglione A. *Cognitive impairment and its relation with disease measures in mildly disabled patients with relapsing-remitting multiple sclerosis: Baseline results from the Cognitive Impairment in Multiple Sclerosis (COGIMUS) study.* Multiple Sclerosis, 2009; 15(7):779–788.

[149] Pozzilli C, Passafiume D, Bernardi S, Pantano P, Incoccia C, Bastianello S, Bozzao L, Lenzi G, Fieschi C. *SPECT, MRI and cognitive functions in multiple sclerosis.* J Neurol Neurosurg Psychiatry, 1991; 54(2):110–5.

[150] Ron M, Callanan M, Warrington E. *Cognitive abnormalities in multiple sclerosis: a psychometric and MRI study.* Psychol Med, 1991; 21(1):59–68.

[151] Rossi F, Giorgio A, Battaglini M, Stromillo M, Portaccio E, Goretti B, Federico A, Hakiki B, Amato M, De Stefano N. *Relevance of brain lesion location to cognition in relapsing multiple sclerosis.* PLoS One, 2012; 7(11):e44826.

[152] Rovaris M, Filippi M, Falautano M, Minicucci L, Rocca M, Martinelli V, Comi G. *Relation between MR abnormalities and patterns of cognitive impairment in multiple sclerosis.* Neurology, 1998; 50(6):1601–8.

[153] Sacco R, Bisecco A, Corbo D, Della Corte M, d'Ambrosio A, Docimo R, Gallo A, Esposito F, Esposito S, Cirillo M, Lavorgna L, Tedeschi G, Bonavita S. *Cognitive impairment and memory disorders in relapsing-remitting multiple sclerosis: the role of white matter, gray matter and hippocampus.* J Neurol, 2015; .

[154] Sanchez M, Nieto A, Barroso J, Martin V, Hernandez M. *Brain atrophy as a marker of cognitive impairment in mildly disabling relapsing-remitting multiple sclerosis.* Eur J Neurol, 2008; 15(10):1091–9.

[155] Snyder P, Cappelleri J. *Information processing speed deficits may be better correlated with the extent of white matter sclerotic lesions in multiple sclerosis than previously suspected.* Brain Cogn, 2001; 46(1-2):279–84.

[156] Sun X, Tanaka M, Kondo S, Okamoto K, Hirai S. *Clinical significance of reduced cerebral metabolism in multiple sclerosis: a combined PET and MRI study.* Ann Nucl Med, 1998; 12(2):89–94.

[157] Swirsky-Sacchetti T, Field H, Mitchell D, Seward J, Lublin F, Knobler R, Gonzalez C. *The sensitivity of the Mini-Mental State Exam in the white matter dementia of multiple sclerosis.* J Clin Psychol, 1992; 48(6):779–86.

[158] Yildiz M, Tettenborn B, Radue E, Bendfeldt K, Borgwardt S. *Association of cognitive impairment and lesion volumes in multiple sclerosis–a MRI study.* Clin Neurol Neurosurg, 2014; 127:54–8.

[159] Zivadinov R, De Masi R, Nasuelli D, Monti Bragadin L, Ukmar M, Pozzi-Mucelli R, Grop A, Cazzato G, Zorzon M. *MRI techniques and cognitive impairment in the early phase of relapsing-remitting multiple sclerosis.* Neuroradiology, 2001; 43(4):272–278.

[160] Sterne J, Sutton A, Ioannidis J, Terrin N, Jones D, Lau J, Carpenter J, Rucker G, Harbord R, Schmid C, Tetzlaff J, Deeks J, Peters J, Macaskill P, Schwarzer G, Duval S, Altman D, Moher D, Higgins J. *Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials.* BMJ, 2011; 343:d4002.

[161] Traboulsee A, Simon J, Stone L, Fisher E, Jones D, Malhotra A, Newsome S, Oh J, Reich D, Richert N, Rammohan K, Khan O, Radue E, Ford C, Halper J, Li D. *Revised recommendations of the consortium of MS centers task force for a standardized MRI protocol and clinical guidelines for the*

diagnosis and follow-up of multiple sclerosis. AJNR Am J Neuroradiol, 2016; 37(3):394–401.

[162] Rovira A, Wattjes M, Tintore M, Tur C, Yousry T, Sormani M, De Stefano N, Filippi M, Auger C, Rocca M, Barkhof F, Fazekas F, Kappos L, Polman C, Miller D, Montalban X. *Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-clinical implementation in the diagnostic process.* Nat Rev Neurol, 2015; 11(8):471–82.

[163] Ahmed I, Sutton A, Riley R. *Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey.* BMJ, 2012; 344:d7762.

[164] Rocca M, Amato M, De Stefano N, Enzinger C, Geurts J, Penner I, Rovira A, Sumowski J, Valsasina P, Filippi M. *Clinical and imaging assessment of cognitive dysfunction in multiple sclerosis.* Lancet Neurol, 2015; 14(3):302–17.

[165] Filippi M, Rocca M, Barkhof F, Bruck W, Chen J, Comi G, DeLuca G, De Stefano N, Erickson B, Evangelou N, Fazekas F, Geurts J, Lucchinetti C, Miller D, Pelletier D, Popescu B, Lassmann H. *Association between pathological and MRI findings in multiple sclerosis.* Lancet Neurol, 2012; 11(4):349–60.

[166] Bland J, Altman D. *Measuring agreement in method comparison studies.* Stat Methods Med Res, 1999; 8(2):135–60.

[167] Fleiss J L, Cohen J. *The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability.* Educational and Psychological Measurement, 1973; 33(3):613–619.

[168] Valdes Hernandez Mdel C, Ferguson K J, Chappell F M, Wardlaw J M. *New multispectral mri data fusion technique for white matter lesion segmentation: method and comparison with thresholding in flair images.* Eur Radiol, 2010; 20(7):1684–91.

[169] Fazekas F, Chawluk J, Alavi A, Hurtig H, Zimmerman R. *MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging.* AJR Am J Roentgenol, 1987; 149(2):351–6.

[170] Filippi M, Rocca M. *MRI and cognition in multiple sclerosis.* Neurological Sciences, 2010; 31(SUPPL. 2):S231–S234.

[171] Maclullich A, Wardlaw J, Ferguson K, Starr J, Seckl J, Deary I. *Enlarged perivascular spaces are associated with cognitive function in healthy elderly men.* J Neurol Neurosurg Psychiatry, 2004; 75(11):1519–23.

[172] Potter G, Chappell F, Morris Z, Wardlaw J. *Cerebral perivascular spaces visible on magnetic resonance imaging: development of a qualitative rating scale and its observer reliability.* Cerebrovasc Dis, 2015; 39(3-4):224–31.

[173] Ingle G, Thompson A, Miller D. *Magnetic resonance imaging in primary progressive multiple sclerosis.* J Rehabil Res Dev, 2002; 39(2):261–71.

[174] van Straaten E, Fazekas F, Rostrup E, Scheltens P, Schmidt R, Pantoni L, Inzitari D, Waldemar G, Erkinjuntti T, Mantyla R, Wahlund L, Barkhof F. *Impact of white matter hyperintensities scoring method on correlations with clinical data: the LADIS study.* Stroke, 2006; 37(3):836–40.

[175] Garcia-Lorenzo D, Francis S, Narayanan S, Arnold D, Collins D. *Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging.* Med Image Anal, 2013; 17(1):1–18.

[176] Taha A, Hanbury A. *Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool.* BMC Med Imaging, 2015; 15:29.

[177] Griffanti L, Zamboni G, Khan A, Li L, Bonifacio G, Sundaresan V, Schulz U, Kuker W, Battaglini M, Rothwell P, Jenkinson M. *BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities.* Neuroimage, 2016; 141:191–205.

[178] Cleveland W. *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, 1979; 74(365):829.

[179] Calabrese M, Agosta F, Rinaldi F, Mattisi I, Grossi P, Favaretto A, Atzori M, Bernardi V, Barachino L, Rinaldi L, Perini P, Gallo P, Filippi M. *Cortical lesions and atrophy associated with cognitive impairment in relapsing-remitting multiple sclerosis.* Arch Neurol, 2009; 66(9):1144–50.

[180] Moller F, Poettgen J, Broemel F, Neuhaus A, Daumer M, Heesen C. *HAGIL (Hamburg Vigil Study): a randomized placebo-controlled double-blind study with modafinil for treatment of fatigue in patients with multiple sclerosis.* Mult Scler, 2011; 17(8):1002–9.

[181] Vos S, Jones D, Viergever M, Leemans A. *Partial volume effect as a hidden covariate in DTI analyses.* Neuroimage, 2011; 55(4):1566–76.

[182] Hsu J, Van Hecke W, Bai C, Lee C, Tsai Y, Chiu H, Jaw F, Hsu C, Leu J, Chen W, Leemans A. *Microstructural white matter changes in normal aging: a diffusion tensor imaging study with higher-order polynomial regression models.* Neuroimage, 2010; 49(1):32–43.

[183] Penke L, Munoz Maniega S, Murray C, Gow A, Hernandez M, Clayden J, Starr J, Wardlaw J, Bastin M, Deary I. *A general factor of brain white matter integrity predicts information processing speed in healthy older people.* J Neurosci, 2010; 30(22):7569–74.

[184] Hemphill J. *Interpreting the magnitudes of correlation coefficients.* Am Psychol, 2003; 58(1):78–9.

[185] Mollison D, Sellar R, Bastin M, Mollison D, Chandran S, Wardlaw J, Connick P. *The clinico-radiological paradox of cognitive function and MRI burden of white matter lesions in people with multiple sclerosis: A systematic review and meta-analysis.* PLoS One, 2017; 12(5):e0177727.