The developing role of Magnetic Resonance Imaging in Phase III multiple sclerosis treatment trials: technical considerations and results of a large multicentre study.

PD Molyneux

Thesis submitted for the degree of doctor of medicine to the University of London.

NMR Research Unit
Institute of Neurology
Queen Square
London, UK
Date of Submission: March 1999
Abstract

Magnetic resonance imaging (MRI) provides a powerful tool for both monitoring disease evolution in multiple sclerosis (MS) and assessing the efficacy of therapeutic intervention. A monthly MRI protocol is now routinely used to screen potential MS therapies, with the effect of treatment on MR activity providing the primary outcome measure. Phase III definitive MS treatment trials also generally incorporate an MR protocol. However, in view of the uncertain relationship between MR measures and clinical outcome, the primary endpoints of phase III studies are currently based on clinical indices.

The first chapter of this thesis provides a critical review of the available MRI tools for monitoring disease progression. The limited extent of the relationship between clinical outcome and standard MR measures is stressed, and potential factors contributing to this dissociation are reviewed.

The thesis is then presented in three parts, the first of which looks at recent developments in MRI acquisition and analysis methodology. In particular, the utility of three established techniques for measuring brain lesion volume (global thresholding, manual outlining and contouring) is examined. Work is presented that suggests a limited role for global thresholding-based lesion segmentation in serial MS studies. A comparison of a semi-automated local thresholding algorithm (contouring) with the currently recognised gold standard segmentation method (manual outlining) is presented, highlighting the advantages in terms of precision and reliability with the contour technique. The effect of slice thickness on MS lesion detection and brain lesion volume is examined, in order to define an optimal slice thickness for MS treatment trials.
The second part examines the relationship between different MRI parameters and assesses the strength of clinical/MRI correlations in a substantial cohort of patients. The extent to which change in annual $T_2$ lesion volume is dictated by ongoing inflammatory activity (assessed with monthly gadolinium enhanced imaging) is investigated using this cohort. Furthermore, the relationship between standard clinical indices of disease activity and progression on the one hand, and MR activity on monthly and annual imaging on the other, is defined. Using an extension of this database, power calculations are presented based on an annual imaging protocol where the change in $T_2$ lesion volume provides the outcome measure. The aim is to guide efficient trial design by providing sample sizes for a typical multicentre parallel group placebo-controlled design. The results suggest that annual $T_2$ lesion volume quantification provides a powerful and robust tool for monitoring the effect of treatment, and demonstrate that the current practice of including several hundred patients in an MR protocol produces substantial overpowering.

Part Three describes the MRI results of a phase III trial using interferon beta-1b in patients with secondary progressive MS. The relationships between clinical and MR indices are also presented, and their implications for future trial design are discussed.
Abbreviations

ANOVA Analysis of variance
BBB Blood brain barrier
CNR Contrast to noise ratio
CoR Coefficient of repeatability
CoV Coefficient of variation
CR Contrast ratio
CSE Conventional spin echo
DSS Disability status scale
EAE Experimental allergic encephalomyelitis
EDSS Expanded Disability Status Scale
FLAIR Fluid attenuated inversion recovery
FOV Field of view
FSE Fast spin echo
FSS Functional Status Score
Gd-DPTA Gadolinium-diethylene-triaminepenta-acetic acid
ICC Intraclass correlation coefficient
IFN Interferon
ITT Intention to treat
MHC Major histocompatibility complex
MMP matrix metalloproteinase
MRI Magnetic resonance imaging
MS Multiple sclerosis
NAA N-acetyl aspartate
NAWM Normal appearing white matter
PP Primary progressive
RR Relapsing remitting
SD Standard deviation
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEM</td>
<td>Standard error of the mean</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to noise ratio</td>
</tr>
<tr>
<td>SP</td>
<td>Secondary progressive</td>
</tr>
<tr>
<td>SRCC</td>
<td>Spearman rank correlation coefficient</td>
</tr>
<tr>
<td>TLV</td>
<td>Total lesion volume</td>
</tr>
<tr>
<td>TNF</td>
<td>Tumour necrosis factor</td>
</tr>
<tr>
<td>VCAM</td>
<td>Vascular cell adhesion molecule</td>
</tr>
<tr>
<td>VEP</td>
<td>Visual evoked potential</td>
</tr>
</tbody>
</table>
Tables

Table 1.1 Schumacher criteria for diagnosis of MS 1
Table 1.2 Poser Committee criteria 3
Table 1.3 MS symptoms at presentation and during disease course 5
Table 1.4 The distribution of brain MRI abnormalities in 200 MS patients 24
Table 1.5 Diagnostic MRI criteria strongly suggestive of MS 25
Table 1.6 Requirements for MS clinical outcome measures 33
Table 1.7 Clinical outcome measures in patients with MS 36
Table 1.8 Studies showing a reduction in MRI activity 39
Table 1.9 Significant clinical/MRI correlations from natural history and other treatment trials 43

Table 3.1 Lesion volume measurements on baseline scans and for changes in lesion volume with the two techniques 58
Table 3.2 Intra and inter rater precision and reliability for baseline TLV 59
Table 3.3 Intra and inter rater precision and reliability for change in TLV 60
Table 5.1 Lesion volume according to sequence/slice thickness 82
Table 5.2 The effects of progressive reduction in slice thickness 84
Table 5.3 Contrast to noise and contrast ratios for each sequence 85
Table 6.1 Clinical and MRI characteristics of the patient sample 96
Table 6.2 Cross sectional and longitudinal clinical/MRI correlations 98
Table 6.3 Correlations between the MRI parameters 101
Table 7.1 Clinical and MRI characteristics of the patient sample 114
Table 7.2 Variance-covariance matrices for RR and SP MS subgroups 114
Table 7.3 Sample size according to effect size and study duration 115
Table 7.4 The effect of added noise on sample size requirements 117
Table 8.1 The effect of IFN-β-1b on brain TLV 132
Table 8.2 The effect of IFN-β-1b on MRI activity 133
Table 8.3 Gadolinium enhanced lesion activity 136
Table 8.4 The effect of treatment on MRI results

Table 9.1 Baseline clinical characteristics

Table 9.2 Proportion of patients with confirmed EDSS progression

Table 9.3 Secondary/tertiary outcome variables

Table 9.4 Baseline TLV according to treatment arm

Table 9.5 Absolute change in TLV (cm³) according to treatment arm

Table 9.6 Absolute change in TLV (cm³) from previous year by treatment arm

Table 9.7 Percentage change in TLV from baseline by treatment arm

Table 9.8 Cumulative number of active T2 lesions by treatment arm

Table 9.9 The proportion of active scans per patient by treatment arm

Table 9.10 The proportion of active patients by treatment arm

Table 9.11 Correlations between TLV change and active T2 lesions

Table 9.12 Correlations between baseline TLV and subsequent MRI activity

Table 9.13 Cross sectional/longitudinal correlations between EDSS and TLV

Table 9.14 Correlations between clinical indices and T2 activity
Figures

Figure 1.1 Disease subtypes according to time course 4
Figure 1.2. (a) Random alignment of Hydrogen nuclei in absence of a 15
magnetic field and (b) net magnetic field ($M_o$) created by $B_o$.
Figure 1.3 Hydrogen nucleus precessing in the presence of $B_o$ 15
Figure 1.4 The $T_1$ relaxation curve 17
Figure 1.5 The $T_2$ relaxation curve 17
Figure 1.6 The Spin Echo Sequence 18
Figure 1.7 Comparison between CSE and fast-FLAIR sequences 21
Figure 4.1 Lesion volumes on baseline scans according to technique 70
Figure 4.2 Axial brain MRI at (a) baseline and (b) follow up 71-72
Figure 4.3 The variable incorporation of non-lesion regions over time 73
Figure 5.1 Normal findings with the 3D fast-FLAIR sequence 80
Figure 5.2 The effect of sequence and slice thickness on lesion volume 83
Figure 5.3 3D fast-FLAIR images at slice thickness 5, 3 and 1 mm 86
Figure 6.1 Relapse rate vs new enhancing MRI activity for RRMS 99
Figure 6.2 Relapse rate vs new enhancing MRI activity for SPMS 99
Figure 6.3 Lesion volume change vs new enhancing lesion rate in RRMS 102
Figure 6.4 Lesion volume change vs new enhancing lesion rate in SPMS 102
Figure 7.1 Sample size per arm for RRMS cohort 116
Figure 7.2 Sample size per arm for SPMS cohort 116
Figure 7.3 The impact of noise addition on sample size for RRMS cohort 118
Figure 7.4 The impact of noise addition on sample size for SPMS cohort 118
Figure 8.1 The effect of IFN-β-1b on extended follow up 134
Figure 9.1 The mean percentage change in TLV by treatment arm 154
Figure 9.2 Mean cumulative $T_2$ activity by treatment arm 159
Publications arising from this thesis


Acknowledgements

This thesis would not have been possible without the help and support of a large number of people. First of all, I would like to thank Professor DH Miller for supervising my research in such a constructive and supportive manner. His ability to make time to discuss any aspect of my work despite his multitude of other commitments has been a revelation.

I am also grateful to the many other members of the NMR Research Unit who have provided moral and scientific support throughout my research. Particular thanks go to Niall Tubridy, Catherine Noctor and Peter Brex for providing a voice of reason when things were not going according to plan.

The lesion volume quantification performed in Chapters 3 and 9 was performed by a large number of dedicated raters, notably Elizabeth Gunn, Alison Fletcher, Paula Robinson, Sara Lewis, Sarah Gregory, Claire Middleditch and Caraline Fogg. Professor DH Miller and Dr IF Moseley checked the hard copy marking that I performed for the work presented in Chapter 4, with Dr Ming Lai acting as the second rater. Niall Tubridy acted as the second hard copy marker in Chapter 5.

The database presented in Part 2 was acquired as part of an EEC funded European collaboration (MAGNIMS). The clinical and MRI data from the various sites was provided as follows: Amsterdam data provided by Dr F Barkhof, Milan by Dr M Filippi, Munich by Dr T Yousry, Basel by Dr L Kappos, and Rome by Dr C Gasperini. The lesion volume quantification from the non-London patients was performed locally by Drs MA Rocca, L Truyen and K Schweikert. The monthly imaging was analysed by Drs F Barkhof and IF Moseley and Dr H Ader helped in the
statistical analysis for the power calculations presented in Chapter 7.

I am also grateful to Beatrice Gomez-Ansen, Peter Brex, Valerie Stevenson and Niall Tubridy for assisting in the hard copy analysis of the data presented in Part 3, together with the physicists and radiographers who were such an important part of this project. I would also like to acknowledge the substantial number of personnel who performed the clinical evaluation across the European centres.

Finally, I would like to thank Schering for funding my research and Sheilagh for her continued and unfailing support.
Chapter 1. Introduction and background

1.1 Multiple Sclerosis

1.1.1 Clinical subgroups based on disease time course and severity

1.1.2 Symptomatology

1.1.3 Neuropathology

1.1.3.1 White matter components

1.1.3.2 The acute lesion

1.1.3.3 The chronic lesion

1.1.4 Immunology

1.1.5 Epidemiology

1.1.6 Natural History

1.1.7 Diagnosis of MS

1.2 MRI in multiple sclerosis

1.2.1 Technical background

1.2.1.1 Spin echo imaging

1.2.1.2 Spatial localisation

1.2.1.3 Techniques for T₂ weighted imaging

1.2.2 MRI in diagnosis

1.2.3 MRI natural history studies in understanding disease evolution

1.2.3.1 Gadolinium enhanced imaging
### 5.2.3 Image Analysis

### 5.2.4 Statistical analysis

### 5.3 Results

### 5.4 Discussion

### Part 2. An assessment of $T_2$ lesion volume as a surrogate outcome

#### Chapter 6. The relationship between $T_2$ lesion volume change and monthly MRI activity

##### 6.1 Introduction

##### 6.2 Methods

- **6.2.1 Patients**
- **6.2.2 MRI**
- **6.2.3 Statistical methods**

##### 6.3 Results

- **6.3.1 Clinical characteristics**
- **6.3.2 MRI characteristics**
- **6.3.3 Clinical/MRI correlations for baseline MRI data**
- **6.3.4 Clinical/MRI correlations for longitudinal MRI data**
- **6.3.5 Correlations between the MRI parameters**

##### 6.4 Discussion

### Chapter 7. Sample size calculations for phase III trials based on lesion volume

##### 7.1 Introduction

##### 7.2 Methods

- **7.2.1 Patients**
- **7.2.2 MRI**
- **7.2.3 Statistical methods**

##### 7.3 Results

##### 7.5 Discussion
Chapter 1. Introduction and Background

1.1 Multiple Sclerosis

Multiple sclerosis (MS) is an acquired primary demyelinating disease of the central nervous system (CNS) in which myelin is the target of an autoimmune inflammatory process. It represents the most common idiopathic inflammatory disease of the CNS. Although there had been earlier partial descriptions, multiple sclerosis was first identified as a distinctive disease by Charcot in 1868, who named it 'sclérose en plaques' (Charcot, 1868). His great contribution was in linking careful observation of symptoms and signs of disease in life with pathological findings in the nervous system after death. Several other clinicians and investigators have since made important contributions in characterising the disease.

The onset of first symptoms is typically in early adult life; for 70% of patients the onset is between 20 and 40 years of age. The hallmarks of the disease are temporal and anatomical dissemination involving the CNS white matter. In 90%, the initial temporal profile comprises neurological disturbances occurring in multiple episodes (also termed attacks, exacerbations or relapses), followed by recovery (remission). A relapse has been defined as a neurological event lasting more than 24 hours that is not attributable to another cause (Poser et al., 1983). Remission usually occurs more slowly than onset of a relapse, and may be complete or incomplete. The second hallmark is the anatomical dissemination of lesions within the CNS.

Different diagnostic criteria have been developed over the years. The Schumacher criteria link natural history and physical signs for the diagnosis of MS (Schumacher 1965) (Table 1.1). These criteria are based exclusively on clinical features. With the advent of various investigative techniques (paraclinical tools) with the potential to aid diagnosis, more recent criteria have now
become established that incorporate the results of such investigations (Poser et al., 1983). The Poser Committee criteria (Table 1.2) are now routinely followed for research purposes, combining both clinical and paraclinical information to define the level of diagnostic certainty. It is self evident that there must be no other cause for the symptoms, signs or abnormal results of investigations.

---

Neurologic examination reveals objective abnormalities of CNS function
History indicates involvement of two or more parts of the CNS
CNS disease predominantly reflects white matter involvement
Involvement of CNS follows one or two patterns:
- Two or more episodes, each lasting at least 24 hours, and ≥ one month apart
- Slow or stepwise progression of signs and symptoms over at least 6 months

Patient 10-50 years old at onset
Signs and symptoms cannot be better explained by other disease process

Table 1.1. Schumacher criteria for diagnosis of MS (from Schumacher et al., 1965)

1.1.1 Clinical subgroups based on disease time course and severity
MS is a disease with a wide range in clinical expression in terms of relapse frequency, rate of progression and severity. Categorisation of MS cases into different subtypes is important in the design of, and recruitment for, multicentre clinical trials that are based on defined patient groups and require narrow entrance criteria. Following a survey among clinicians involved in MS, a standardised classification was proposed that has now achieved widespread recognition (Lublin & Reingold 1996). The commonest initial presentation (90%) is a relapsing remitting disease pattern. Relapsing remitting (RR) MS is characterised by clearly defined relapses with or without
full recovery (Figure 1.1); between relapses there is no disease progression. Secondary progressive (SP) MS is defined as an initial relapsing remitting course followed by progression with or without occasional relapses, minor remissions and plateaus. Secondary progressive MS can be seen as a long-term outcome of relapsing remitting disease, where the baseline between relapses has begun to worsen. Primary progressive (PP) MS is characterised as disease progression from onset with occasional plateaus and temporary minor improvements allowed. Such a course is seen in 10% of MS patients. Benign MS is defined as a relapsing remitting time course with minimal or no disability 15 years after disease onset. Other terms such as relapsing progressive and progressive relapsing have been less clearly defined (Lublin & Reingold 1996) and are not commonly used to define patients for inclusion into clinical trials.

<table>
<thead>
<tr>
<th>Category</th>
<th>Attacks</th>
<th>Clinical evidence</th>
<th>Paraclinical evidence</th>
<th>CSF oligoclonal bands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinically definite MS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>and 1</td>
<td></td>
</tr>
<tr>
<td>Clinically probable MS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>and 1</td>
<td></td>
</tr>
<tr>
<td>Laboratory supported definite MS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>or 1</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>and</td>
<td>+</td>
</tr>
<tr>
<td>Laboratory supported probable MS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

Table 1.2. Poser Committee criteria (from Poser et al., 1983)
(i) Relapsing remitting MS

(ii) Secondary Progressive MS

(iii) Primary Progressive MS

(iv) Benign MS

Figure 1.1 Four different disease subtypes are depicted according to time course
1.1.2 Symptomatology

MS is a complex clinical disorder with a wide range of expression, often related to the severity and form of the disease. Typically, negative symptoms and signs predominate, examples including loss of vision, strength or sensation (Table 1.3). The commonest symptoms of a relapse are weakness, optic neuritis or sensory disturbance, often occurring in isolation; ataxia usually occurs in combination with a cluster of symptoms such as vertigo, diplopia, weakness and sensory disturbance, reflecting an infratentorial focus. In primary progressive MS, there is most often an insidious development of a spastic and ataxic paraparesis; very rarely there may be progressive optic neuropathy. Less commonly, positive symptoms such as trigeminal neuralgia, Lhermittes symptom or tonic spasms may occur.

<table>
<thead>
<tr>
<th>Reported deficit</th>
<th>Presenting</th>
<th>During disease course</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual/Oculomotor</td>
<td>49%</td>
<td>100%</td>
</tr>
<tr>
<td>Paresis</td>
<td>43%</td>
<td>88%</td>
</tr>
<tr>
<td>Paraesthesia</td>
<td>41%</td>
<td>87%</td>
</tr>
<tr>
<td>Incoordination</td>
<td>23%</td>
<td>82%</td>
</tr>
<tr>
<td>Genito-urinary/bowel</td>
<td>10%</td>
<td>63%</td>
</tr>
<tr>
<td>Cerebral</td>
<td>4%</td>
<td>39%</td>
</tr>
</tbody>
</table>

Note. Since some patients had multiple symptoms, the total is greater than 100%.

Table 1.3 MS symptoms at presentation and during disease course (Poser et al., 1979)
1.1.3 The Neuropathology of MS

1.1.3.1 White matter components

Myelin is the major structural component in normal white matter, accounting for about 25% of the dry weight of the brain. Oligodendrocytes are responsible for the elaboration and maintenance of myelin within the CNS. Each myelin sheath is elaborated around a segment of axon by the flattening and spiral wrapping of a single oligodendrocyte cell process. This represents an internode of myelin and each end is demarcated by a node of Ranvier. Oligodendrocytes are heavily outnumbered by the myelin sheaths they produce, and it has been estimated that each can maintain 30-50 internodes (Raine, 1997). They do not respond actively after injury and frequently degenerate, a key feature in MS plaque formation.

Separating and penetrating the bundles of myelinated nerve fibres are astrocytes, the predominant supporting and structural elements of white matter. Unlike oligodendrocytes, astrocytes respond readily to injury by proliferating and synthesising glial fibrils. Such astrogliosis leads to a state of sclerosis (scarring), a key hallmark of the MS plaque. The other major white matter components are microglia, ependymal cells lining the ventricles, and blood vessels.

The interaction between these white matter structural components and the immune system influences the evolution of demyelinating lesions, which over time progress to leave scar tissue containing bundles of naked axons.

1.1.3.2 The acute MS lesion

The major pathological hallmarks of the acute MS lesion are oedema, inflammation and acute myelin breakdown. The margin of an acute lesion is typically indistinct due to ongoing demyelination. Oedema at the centre of such a lesion is found, associated with an expanded
extracellular space. One of the most striking histological features is the intense inflammatory response and clear association of lesion activity with small venules. Diffuse parenchymal infiltration by macrophages and perivascular lymphocyte cuffing are common. The previously well myelinated parenchyma is reduced to a loosely packed, oedematous zone containing an abundance of myelin debris and fat laden macrophages. Myelin destruction is effected by phagocytosis involving investing macrophages, although hypertrophic astrocytes may be involved to a lesser degree. Signs of axonal transection and degeneration may also be seen in acute, early lesions (Trapp et al., 1998).

1.1.3.3 The Chronic MS lesion

Chronic MS plaques are the most frequent type found at autopsy and reveal variable degrees of demyelination, axonal loss and gliosis. Myelin staining confirms chronic plaques as areas of demyelination, demarcated from adjacent white matter by a sharp edge to give a punched out appearance. Most of the demyelinated plaque parenchyma is replaced by a network of astroglial scar tissue, and axonal loss is prominent (Lassmann et al., 1994), particularly at the centre of the plaque. Astrocytes are well separated, often multi-nucleated, and sharply demarcate the plaque from surrounding normal white matter. In the most chronically demyelinated plaques, fibrous astrogliosis is intense, with glial processes forming parallel rows; there is no evidence of ongoing inflammation or remyelination, and oligodendrocytes are absent.

The edge of a chronic lesion typically displays evidence of ongoing activity, with reactive astrocytes and evidence of proliferation. Furthermore, it is not unusual to find a rim of proliferated oligodendrocytes around the edge of a lesion, sometimes in association with remyelination. More extensive remyelination may be seen, especially in patients with a short disease duration (Prineas et al., 1993). In more chronically active lesions, there is a superimposed
prominent inflammatory component upon a previously demyelinated plaque, with astrogial hypertrophy, oligodendrocyte hyperplasia and ongoing demyelination.

1.1.4 Immunology of MS

The aetiology of MS is still unknown, but it is widely accepted that a T-cell autoimmune response is involved in the demyelinating process. An autoimmune pathogenesis is supported by a number of findings; (i) the composition of CNS white matter infiltrates consisting primarily of lymphocytes and monocytes, (ii) the association with genes relevant to immune responses, and (iii) the response to both immunosuppressive and immunomodulatory treatments (Martin et al., 1997). A greater understanding of the immunopathological basis for the inflammatory demyelinating process has resulted from the study of an animal model, experimental allergic encephalomyelitis (EAE), which is induced by injection of myelin components into susceptible animal strains. Collectively, studies in patients and animals indicate that MS is likely to reflect the result of a primarily T-cell driven aberrant immune response to several myelin and non-myelin CNS antigens. Epidemiological and immunological data suggest that viral infection may play a pivotal role in induction of disease in genetically susceptible individuals. Several viruses such as measles virus, herpes virus, paramyxovirus and various retroviruses have been implicated in the aetiology of MS. However, a universal role for a single virus has yet to be demonstrated. Viruses may cause demyelination by direct lysis of oligodendrocytes, immune mediated damage of virus infected oligodendrocytes, virus induced autoimmune demyelination or bystander demyelination (Martin & McFarland 1997). While the precise mechanism of viral induction is at present elusive, an autoimmune pathogenesis directed against myelin antigens is likely. T-cells reactive to a number of myelin sheath antigens such as myelin basic protein, proteolipid protein and myelin oligodendrocyte protein have been isolated from both the blood and cerebrospinal fluid (CSF) of MS patients (Hohlfeld et al., 1995). A variety of mechanisms such as molecular
mimicry, bystander activation and cytokine release have been postulated for this process (Hartung 1997). Penetration of the blood brain barrier (BBB) by such autoreactive circulating T-cells then occurs.

Once within the CNS, activated T-cells undergo proliferation and start a cascade of inflammatory reactions through release of cytokines such as tumour necrosis factor α (TNFα) and interferon-γ (Beck et al., 1988). Further myelin damage may occur via reactive oxygen species, proteases and complement activation.

Several mechanisms may operate to terminate the acute inflammatory episode; examples include increased generation of down regulatory cytokines, apoptosis of autoreactive lymphocytes (Schmeid et al., 1993) and enhancement of T-cell suppressor activity (Hartung 1997).

1.1.5 Epidemiology of MS

The varying susceptibility according to race (more common in northern Europeans, less common in Orientals) implicates genetic factors. This is supported by the increased incidence of MS in close blood relatives, especially in the monozygotic twins of patients with MS (Mumford et al., 1994). Familial MS studies suggest that more than one gene is likely to confer susceptibility to the disease. However, at present little is known about the genes determining genetic susceptibility, although the common Caucasian MHC class II HLA-DR2 haplotype appears to play a role (Compston & Sadovnick 1992).

The aetiological role of environmental factors is supported by the distinctive geographic distribution of the disease. The worldwide distribution can be described within three frequency bands: (1) high frequency areas such as Northern America, Northern Europe and New Zealand
with prevalence rates of greater than 30 per 100 000; (2) medium frequency areas such as Australia, the Southern United States and parts of Russia, with prevalence rate of 5-29 per 100 000; and (3) low frequency areas, such as equatorial Africa, Alaska and Greenland, with prevalence rates less than 5 per 100 000 (Kurtzke 1997). Furthermore, countries such as Australia and New Zealand that exhibit relative ethnic homogeneity among the northern European resident population, show a North-South gradient. This latitude effect has been used to support the role of environmental factors in MS aetiology (Hammond et al 1988; Miller et al., 1990). This has been further supported by studies on migrants. Some prevalence studies on groups of North Americans and Northern Europeans who had moved from high to low risk areas have demonstrated that those migrating after 15 years of age tend to maintain a higher level of risk. In contrast, those moving before this age acquired the risk of the host country (Alter & Okihiro 1971). However, the Australian migrant studies showed no age cut off.

The existing data supports the hypothesis that both genetic and environmental factors are important. Since genetic susceptibility is not sufficient by itself to cause MS, and with evidence that no single environmental factor in isolation can explain its development, it is possible that several factors are involved. MS predisposition may be induced by polygenes, and various environmental factors may trigger the disease in genetically susceptible individuals. However, one still cannot exclude the possibility that a single infectious agent triggers the disease, resulting in a heterogeneous phenotype because of differences in the hosts immunogenetic response.

1.1.6 Natural History of MS

In 90% of MS patients there is an initial relapsing and remitting course, but subsequent conversion to secondary progressive disease occurs in over 50% within 10-15 years (Weinshenker et al., 1989; Runmarker & Andersen 1993). Disability can be accumulated by both
failure of remission and secondary progression of disease.

The average relapse frequency in relapsing remitting disease is 0.1-1 per year (Weinshenker & Ebers 1987). Most (British and Dutch MS azathioprine Trial Group 1988; Paty et al., 1993), but not all (Goodkin et al., 1989) studies have found that relapse frequency declines over time, although this may in part reflect regression to the mean (Weinshenker 1994). Relapse rate also declines with increasing age (Weinshenker 1994). Several factors modestly predict the outcome of MS relapses. The rapidity of onset is associated with recovery; the more rapid the onset, the higher the level of recovery. The relapse severity may also be a factor (Weinshenker 1994). While recovery from a relapse may be protracted, the probability of recovery drops substantially after one month following relapse (Kurtzke et al., 1973).

Once the progressive phase is entered, the average rate of progression is approximately 0.5 EDSS points per year (Weinshenker et al., 1989), but this is highly variable. Using data acquired from the placebo group of MS treatment trials, it has been shown that 30-50% deteriorate by one EDSS point over 2-3 years (Weinshenker & Sibley 1992). However, the EDSS is non linear, with differences in duration of stay at a given level. Thus, the mean duration of stay at EDSS 4.0 to 5.0 is short (1.22-1.25 years) compared to that at EDSS 6.0 and 7.0 (3.06-3.77 years) (Weinshenker et al., 1991).

While the outcome of MS is notoriously difficult to predict, a number of clinical and demographic factors are, to an extent, predictors of clinical outcome. Favourable factors include an age of onset less than 40 years, female sex, optic neuritis or sensory symptoms at onset, a relapsing remitting course and low relapse frequency. Unfavourable variables are age of onset more than 40 years, male sex, motor/cerebellar features at onset, progressive course and high
relapse frequency (Wynn et al., 1990; Weinshenker et al., 1991; Runmarker et al., 1994). Disability status at 5 years is moderately predictive of the future prognosis (Kurtzke et al., 1977; Miller et al., 1992). Multivariate mathematical models have been designed to identify those factors that are predictive for various clinical endpoints (Weinshenker et al., 1991; Runmarker et al., 1994), but even with such multifactorial modelling, predictive accuracy at an individual level is poor.

### 1.1.7 Diagnosis of MS

MS is diagnosed primarily on clinical grounds, rather than from the results of paraclinical investigation in isolation. Confident diagnosis relies on demonstration of clinically disseminated CNS white matter lesions in both time and space, in the absence of features suggesting an alternative diagnosis. The differential diagnosis includes other inflammatory CNS disorders that can produce multifocal CNS lesions, examples including systemic lupus erythematosis, polyarteritis nodosa, isolated angiitis of the CNS, sarcoidosis, Sjögren’s syndrome and Lyme disease.

Where clinical assessment does not suggest an alternative diagnosis but is insufficient to prove temporal and spatial dissemination, paraclinical studies may be performed in order to increase diagnostic confidence. The major paraclinical tools are CSF examination, evoked potentials and neuroimaging. CSF examination can be used to demonstrate evidence of CNS inflammation. The measurement of CSF IgG alone is of limited usefulness, since its level may be influenced by serum IgG concentration, blood brain barrier dysfunction and CSF turnover. In order to correct for such variables, an IgG index is used. The demonstration of unique CSF oligoclonal bands is also used to increase diagnostic confidence. The diagnostic sensitivity of the IgG index and unique oligoclonal banding respectively is 92% and 97% (Tourtellotte & Tumani 1997).
Evoked potentials can provide evidence of spatial dissemination, where lesions are clinically silent. Demonstration of such lesions can be used to confirm a diagnosis of MS according to the Poser criteria (Table 1.2). Visual evoked potentials (VEPs) are abnormal in most patients with a clear history of optic neuritis (Chiappa 1990). Even where there is no clinical evidence for optic neuritis, VEPs are abnormal in over 50% of cases in established MS (Chiappa 1990). Somatosensory evoked potentials and brain stem auditory evoked potentials are also used to provide evidence of more widespread dissemination, and multimodal evoked potentials may increase the sensitivity compared to single evoked potential studies (Chiappa 1990).

However, it is MRI that is now established as the most valuable investigation in supporting and making more certain the diagnosis of MS. The demonstration of disseminated lesions in space can upgrade the diagnosis from clinically probable to clinically definite MS using the Poser criteria (Table 1.2).

After a basic review of MRI physics, the roles of MRI in diagnosis, as a prognostic tool and in developing a greater understanding of the natural history of the disease will be reviewed. The remainder of this chapter will then focus on the utility of MRI as an outcome measure in MS clinical trials.

### 1.2 MRI in multiple sclerosis

#### 1.2.1 Technical background

The basic principles of nuclear magnetic resonance (NMR) were first described in 1946 (Bloch et al., 1946; Purcell et al., 1946). Initial applications were confined to the realms of chemistry and physics, and it wasn't until 1973 that the first image of an object was published (Lauterbur 1973). The first clinically useful images were published in 1980 (Edelstein et al., 1980), and
since then the role of MRI has expanded explosively.

NMR depends on an interaction between an 'MR active' atomic nucleus, an external magnetic field and an applied radio frequency electromagnetic field. Such nuclei are characterised by possession of an odd number of protons, and the most abundant of these in organic matter is hydrogen, predominantly as constituent atoms of water and lipid molecules. The hydrogen nucleus comprises a single proton. The combination of the proton's positive charge and spin results in a nuclear magnetic moment. In the absence of an applied external magnetic field, the axes of the protons' spins are randomly aligned and there is no net magnetic moment. However, when exposed to a static applied external magnetic field (called \(B_0\)), hydrogen nuclei align themselves either parallel and antiparallel to that field. A tiny excess of hydrogen nuclei aligns parallel and this results in a net magnetic moment (\(M_o\)), aligned along the axis of \(B_0\) (Figure 1.2).

The spin axes of the hydrogen nuclei rotate around the field lines of the applied magnetic field (\(B_0\)); this is termed precession (Figure 1.3).

The precession frequency is governed by the Larmor equation:

\[
\omega_0 = \gamma \times B_0
\]

where: \(\omega = \) Larmor frequency

\(\gamma = \) gyromagnetic ratio, which is a constant for a given nucleus

If a radiofrequency (RF) pulse is applied that is polarised to \(B_0\) at an angle of 90° with the correct Larmor frequency, it will induce resonance, a process termed excitation. The first result of
Figure 1.2 (a) Random alignment of Hydrogen Nuclei in absence of a magnetic field and (b) net magnetic field ($M_o$) created by $B_o$.

Figure 1.3 Hydrogen nucleus precessing in the presence of $B_o$. 
resonance is that the net magnetic vector moves out of alignment away from $B_o$. The angle to which the net magnetic vector moves out of alignment is termed the flip angle. A 90° RF pulse will tip the net vector through an angle of 90°, causing the hydrogen nuclei to precess in phase (phase coherence). The component of the rotating net magnetic vector in the transverse plane induces an electric current in a receiver coil or antenna. Therefore, as the net magnetic vector precesses at the Larmor frequency in the transverse plane (at 90° to the $B_o$ direction), a voltage is induced in the coil, which in turn produces the signal from which MR images or spectra may be formed.

When the RF pulse is turned off, the protons return to their equilibrium position aligned with $B_o$. Two events occur in this process; the first is an increase in "longitudinal" magnetisation as hydrogen nuclei return to realign with $B_o$. This process is known as longitudinal relaxation and requires transfer of energy from the hydrogen nuclei to the immediate environment or lattice. The rate of recovery is an exponential process, with a recovery time constant called $T_1$ (Figure 1.4). The second process is a decrease in the transverse net magnetic vector as hydrogen nuclei begin to precess out of phase with one another. This decay occurs via an interaction with adjacent atoms and is termed spin spin relaxation. This also occurs as an exponential process, defined by the $T_2$ relaxation time (Figure 1.5).

Both $T_1$ and $T_2$ relaxation times depend on the chemical environment of the hydrogen nuclei. $T_1$ is typically much longer than $T_2$ (300-2000 ms versus 30-150 ms) at typical clinical field strengths. For water, both $T_1$ and $T_2$ are long, whereas lipids have much shorter $T_1$ and $T_2$. Such differences are used to provide signal contrast between tissues.
Figure 1.4 The $T_1$ relaxation curve after a $90^\circ$ pulse

Figure 1.5 The $T_2$ relaxation curve
1.2.1.1 Spin Echo imaging

The frequency, power and duration of RF pulses can be variously manipulated to produce specific effects on protons. The combination of a 90° pulse followed by a 180° pulse is termed a spin echo sequence (Figure 1.6), and this is the most commonly used sequence. Following a 90° pulse, protons are tipped into the xy plane and they precess in phase. This generates a detectable transverse magnetic field. However, signal immediately begins to decline due to a combination of T₂ decay and dephasing due to external magnetic field inhomogeneity, the so-called T₂* or free induction decay (FID). After a delay, termed the echo time (TE), a 180° pulse is applied. This rotates the protons’ magnetic moments through 180°, reversing the direction of precession. Thus there is a return of all the transverse magnetisation that was lost due to T₂* effects, and an “echo” is generated. Therefore, the amplitude of the resulting echo is dependent primarily on T₂ relaxation. Generation of an image requires several repetitions of the spin echo sequence, with the time between successive 90° pulses termed the repetition time (TR).

![Figure 1.6 The Spin echo sequence](image-url)
Manipulation of both the TR and TE allows images to be generated with contrast dependant on varying degrees of $T_1$, $T_2$ or proton density (PD) characteristics. Weighting refers to the type of contrast that dominates an image. With a short TR, those protons with longer $T_1$ will not have returned to align along $B_0$ before the next excitation. The combination of a short TR with a short TE produces an image that is $T_1$ weighted, those tissues with a long $T_1$ appearing dark and those with a shorter $T_1$ appearing bright. The combination of a long TR with a short TE generates a PD weighted image, where contrast is less dependent on either $T_1$ or $T_2$. With both a long TR and TE, those protons with a short $T_2$ dephase to a greater degree than those with a longer $T_2$, yielding a $T_2$ weighted image, where tissues with short $T_2$ appear dark relative to those of a longer $T_2$.

1.2.1.2 Spatial localisation

Localisation of the MR signal is required for image generation. This is achieved through the modification of the external magnetic field by the application of magnetic field gradients. The application of a slice selection gradient along the $z$ axis allows selective excitation of a particular $xy$ slice according to the frequency of the RF pulse used. This allows localisation of signal along the $z$ axis.

A similar approach is used to allow localisation along the $x$ axis. Thus, a frequency encoding gradient is applied perpendicular to the slice select gradient, allowing spatial localisation along the $x$ axis. To encode spatial information along the $y$ axis requires a phase encoding gradient. This gradient is applied only briefly to alter the phase, but not the frequency, of the derived signal. The stronger the phase encoding gradient, the greater the difference in phase angle along the $y$ direction. The phase of different components of the signal therefore identifies their origin.
1.2.1.3 Techniques for $T_2$ weighted imaging

With conventional spin echo (CSE) imaging, a two-feature approach is adopted, using the early and late echoes to generate proton density (PD) and $T_2$ weighted images, respectively. One line of k-space is acquired per TR, which is usually within the range of 2000 to 3000 msec. K-space is where the raw data of spatially encoded MR signals is collected during application of magnetic field gradients. In CSE imaging, one line of k-space is encoded per TR interval, and the pulse sequence is repeated typically 128 or 256 times (phase encodings) per image. Typically, TEs of 25 to 50 msec for PD weighting and 80 to 120 msec for $T_2$ weighting are used. The exact parameters will depend to an extent on scanner field strength, since while $T_2$ is virtually field strength independent, $T_1$ is positively correlated with field strength. Therefore, to obtain similar CSF suppression across different Tesla imagers (a $T_1$ effect on PD weighted images), the TR is lower for low-field strength machines (Filippi et al., 1997a). MS lesions appear hyper-intense relative to background white matter on both PD and $T_2$ weighted images (Figure 1.7) by virtue of their higher water content.

More recently, fast spin echo (FSE) imaging has become increasingly used. The FSE sequence is based on the rapid acquisition with relaxation enhancement (RARE) sequence (Hennig et al., 1986; Hennig et al., 1988). With FSE imaging, multiple phase encodings are performed in each TR and multiple echoes per TR are acquired. Therefore, instead of a single line of k-space per TR interval, from 2 to 16 or more lines are encoded per TR, resulting in a considerable reduction in acquisition time in proportion to the number of echoes collected.

Although the FSE sequence produces images that are broadly similar to a corresponding CSE image, they are not identical.
Figure 1.7 Axial images of a patient with clinically definite MS; PD (Figure 1.7 a) and T$_2$ weighted (Figure 1.7 b) images from a dual echo CSE sequence, and corresponding fast FLAIR image (Figure 1.7 c). Several periventricular and discrete white matter lesions are demonstrated with all three sequences. Note the CSF suppression with the fast FLAIR sequence.
The FSE sequence may produce increased signal at tissue interfaces (for example, in periventricular regions), making detection of MS lesions in this region more difficult (Bastienello et al., 1997; Thorpe et al., 1994). It may also be less sensitive in detecting small lesions due to point spread function effects (Constable & Gore, 1992). This latter effect describes the blurring that can occur in the phase encoding direction due to the collection of different lines of k-space at different times, obscuring detection of small lesions.

While long TR long TE sequences offer a higher level of contrast between lesions and white matter (determined by $T_2$ mechanisms), an undesirable consequence is reduced lesion-to-CSF contrast, making identification of periventricular and subcortical lesions more difficult (Figure 1.7). The high CSF signal can be suppressed by application of a 180° inversion pulse, with an appropriately long inversion time (TI) before each excitation pulse, thereby allowing longer TE while at the same time suppressing CSF signal (De Coene et al., 1992; Thorpe et al., 1994). This is known as the fluid attenuated inversion recovery (FLAIR) sequence, and increases the number of MR visible lesions (White et al., 1992; De Coene et al., 1993). However, the long TIs and TRs required demand longer acquisition times with a standard FLAIR sequence. More recently, the inversion pulse has been combined with an FSE pulse sequence to produce the fast FLAIR sequence (Figure 1.7). This sequence can acquire 36 slices of 5 mm slice thickness in just over 5 minutes (Rydberg et al., 1994).

1.2.2 MRI in diagnosis

The major utility of MRI in the diagnosis of MS is the ability to visualise a large extent of clinically silent pathology; as a tool for demonstrating spatial dissemination of disease, MRI is unrivalled. In clinically definite MS, brain lesions are seen in up to 99% of patients (Ormerod et al., 1987). However, the MRI signal characteristics of MS lesions are not specific to MS; the
vast majority of MS lesions are hyper-intense on both PD and T₂ weighted images. Such signal characteristics may also be found in ischaemia and neoplastic disease. Furthermore, small foci of hyper-intensity are a common incidental MRI finding in normal subjects, especially with increasing age. Various features of lesion morphology and distribution have been identified as potentially more specific to MS. The lesions seen in MS are typically irregular in shape, although an ovoid appearance is also characteristic (Horowitz et al., 1989). This has been attributed to perivascular extension of the pathological processes along straight medullary venules. Lesions have a periventricular predominance (Table 1.4), with a tendency to be orientated perpendicular to the ventricular walls, corresponding to the pathologists eponym of Dawson's fingers (Fazekas et al., 1998). Common locations are adjacent to the body and trigone of the lateral ventricles, the floor of the fourth ventricle and cerebellar peduncles.

Discrete white matter lesions are also typically seen, and a cortical/subcortical or juxtracortical location is another characteristic of MS plaques (Barkhof et al., 1997a). Involvement of the subcortical U-fibres is less commonly identified in subcortical arteriosclerotic encephalopathy (Révész et al., 1989). Sagittal imaging has emphasised the specificity of lesions located in the corpus callosum (Simon et al., 1986; Gean-Marton et al., 1991), and thin section fast FLAIR may further improve the detection of abnormalities in the corpus callosum (Hashemi et al., 1995). Lesions are less commonly seen in the basal ganglia in MS than cerebrovascular disease, a feature of some value in differential diagnosis (Ormerod et al., 1987).
<table>
<thead>
<tr>
<th></th>
<th>No. of patients</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral ventricle</td>
<td>196</td>
<td>98</td>
</tr>
<tr>
<td>Body</td>
<td>194</td>
<td>97</td>
</tr>
<tr>
<td>Trigone</td>
<td>171</td>
<td>86</td>
</tr>
<tr>
<td>Occipital horn</td>
<td>149</td>
<td>75</td>
</tr>
<tr>
<td>Temporal horn</td>
<td>132</td>
<td>66</td>
</tr>
<tr>
<td>Frontal horn</td>
<td>117</td>
<td>59</td>
</tr>
<tr>
<td>Third Ventricle</td>
<td>31</td>
<td>16</td>
</tr>
<tr>
<td>Fourth Ventricle</td>
<td>106</td>
<td>53</td>
</tr>
<tr>
<td>Discrete Cerebral White Matter</td>
<td>185</td>
<td>93</td>
</tr>
<tr>
<td>Cortico-medullary junction</td>
<td>130</td>
<td>65</td>
</tr>
<tr>
<td>Internal Capsule</td>
<td>83</td>
<td>42</td>
</tr>
<tr>
<td>Cerebral Cortex</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>Basal Ganglia</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>Brain Stem</td>
<td>132</td>
<td>66</td>
</tr>
<tr>
<td>Pons</td>
<td>103</td>
<td>52</td>
</tr>
<tr>
<td>Midbrain</td>
<td>72</td>
<td>36</td>
</tr>
<tr>
<td>Medulla</td>
<td>29</td>
<td>15</td>
</tr>
<tr>
<td>Cerebellum</td>
<td>113</td>
<td>57</td>
</tr>
</tbody>
</table>

**Table 1.4** The distribution of brain MRI abnormalities in 200 patients with clinically definite MS (from Miller 1988)

In order to improve the specificity of MR in the diagnosis of MS, two research groups have developed criteria based on the pattern of lesion distribution (Table 1.5, Paty et al 1988; Fazekas et al., 1988). A study to validate these criteria demonstrated higher specificity for the criteria of Fazekas et al (96%) than those of Paty et al (92%) at the expense of lower sensitivity (81% versus 87% respectively) (Offenbacher et al., 1993). Both sets of criteria were found to be less
specific in older patients due to difficulties in differentiating MS lesions from those due to subcortical arteriosclerotic encephalopathy.

| Paty’s (from Paty et al., 1988) | Four or more lesions, or three lesions, one of which is periventricular |
| Fazeka’s (from Fazekas et al., 1988) | Three or more lesions, with at least two of the following properties: (i) infratentorial (ii) periventricular (iii) ≥ 6 mm |

Table 1.5. Diagnostic MRI criteria strongly suggesting MS

The potential role of gadolinium-diethylene-triaminepenta-acetic acid (gd-DPTA, or gadolinium) enhanced imaging as a diagnostic tool has also been suggested. Since gadolinium enhancement is a temporary phenomenon, the co-existence of longstanding (non enhancing) lesions with acute (enhancing) lesions may improve diagnostic specificity (Tas et al., 1995). A recent study examined the contribution of various MRI features in predicting conversion to MS after a clinically isolated syndrome (Barkhof et al., 1997a). Of the various parameters included, only gadolinium enhancement and the presence of juxtacortical lesions provided independent information of predictive value.

Spinal cord involvement may also be useful in supporting a diagnosis of MS. Using phased array coils and a slice thickness of 3 mm, intramedullary lesions have been detected in 70-80% of
patients with MS (Kidd et al., 1993). Furthermore, a high proportion of patients with suspected
MS but without cerebral lesions will have abnormalities on spinal MRI (Thorpe et al., 1996a;
O'Riordan et al., 1998). It has been suggested that the combination of brain and spinal cord MRI
might increase the sensitivity for MS to nearly 100% (Thorpe et al., 1996a).

1.2.3 MRI natural history studies in understanding disease evolution

Serial studies of MS patients have provided important insights into the dynamics and
pathogenesis of the disease process. Early studies using T\textsubscript{2} weighted imaging showed that new
areas of hyper-intensity were frequently observed in the absence of clinically expressed activity,
occurring up to five times as frequently as relapses (Isaac et al., 1988; Willoughby et al., 1989).

1.2.3.1 Gadolinium enhanced imaging

Studies incorporating gadolinium enhanced sequences have demonstrated that enhancement is
an early event in most new T\textsubscript{2} lesions in patients with both relapsing remitting and secondary
progressive MS, often preceding the appearance of a T\textsubscript{2} lesion (Kermode et al., 1990a; Kermode
et al., 1990b).

Gadolinium does not normally cross the blood brain barrier in cerebral white matter. The results
of studies of both EAE (Hawkins et al., 1990; Hawkins et al., 1991) and post mortem work (Katz
et al., 1993) have suggested that gadolinium enhancement is precipitated by alteration of the
blood brain barrier in association with an acute inflammatory response. The duration of
enhancement of MS lesions is typically 3-6 weeks; more than 75% of enhancing lesions show
enhancement for ≤1 month (Smith et al., 1993). About 20% of lesions enhance beyond one
month, with about 5% showing persistent enhancement up to 3-4 months. Early in the period of
enhancement, the area of signal change on the corresponding unenhanced T\textsubscript{2} image typically
expands to involve an area larger than the region of gadolinium enhancement, probably reflecting
the presence of surrounding vasogenic oedema. As gadolinium enhancement ceases, the size of
the T2 abnormality decreases to leave a smaller T2 'scar'. Early work suggested that many such
T2 lesions disappear altogether (Willoughby et al., 1989), but studies at higher field strength and
with thinner slices indicate that a residual, albeit small T2 abnormality almost invariably persists
(Thompson et al., 1991). Therefore, with ongoing enhancing lesion activity, the total area of
abnormal hyper-intensity on T2 scans increases over time.

Gadolinium enhancement occurs with about ten times the frequency of clinical relapse in both
relapsing remitting and secondary progressive MS (Thompson et al., 1991). Enhancement in a
clinically eloquent site such as the optic nerve has been correlated with reversible clinical deficit
(Youl et al., 1991). Furthermore, enhancing lesions occur more often during clinical relapse than
remission, and correlations between gadolinium enhanced MRI and clinical activity have also
been demonstrated in several studies (Grossman et al., 1986; Willoughby et al., 1989;
Bastienello et al., 1990; Kermode et al., 1990a; Miller et al., 1993; Smith et al., 1993; Thorpe
et al., 1996b). Monosymptomatic relapses are frequently associated with multiple new areas of
gadolinium enhancement.

A study using gadolinium enhanced MRI and evoked potentials in patients presenting with optic
neuritis has provided important insights into the mechanisms underlying relapse and remission
(Youl et al., 1991). Ten patients presenting with optic neuritis were studied within two weeks
of symptom onset. All of the affected nerves showed initial gadolinium enhancement. However,
one month later, all but two had ceased to enhance. A marked reduction in VEP amplitude and
increase in latency was found at the time of gadolinium enhancement, signifying both conduction
block and demyelination. After enhancement ceased, conduction delay persisted while both the
amplitude of the evoked potential and visual acuity improved. These observations suggest that;
(i) the clinical features of optic neuritis are associated with inflammation as assessed by MRI,
(ii) clinical recovery can occur without remyelination, as suggested by functional recovery in the
face of persistent conduction delay, and (iii) inflammation plays an important role in production
of symptoms. These observations are supported by the often rapid clinical improvement
associated with administration of corticosteroids. The mechanisms through which conduction
block is induced by acute inflammation are as yet not well defined, but one possibility is that
cytokine release may compromise sodium channel function (Brosnan et al., 1989).

1.2.3.2 $T_2$ weighted imaging

While gadolinium enhancement is widely recognised to be associated with BBB breakdown and
inflammation, the conventional $T_2$ weighted sequence has poor specificity in discriminating the
different elements of the pathological process (acute inflammation with oedema, demyelination,
axonal loss, gliosis, remyelination). With such a sequence, areas of high signal white matter are
a consequence of increased water content and change in $T_2$ relaxation time, that can potentially
result from a number of pathological processes. Chronic MS lesions are pathologically
heterogenous, ranging from those with relative preservation of axons (gliotic lesions) to those
with an expanded extracellular space and extensive axonal loss. Serial $T_2$ weighted imaging can
be used to document the accumulation of high signal abnormality over time, but the sequence
gives no indication about the integrity of myelin or axons. Other MRI techniques with greater
pathological specificity are needed in order to elucidate the mechanisms underlying the
development of disability further.
1.2.3.3 Serial Studies in MS subgroups

Several longitudinal studies, detailed below, have incorporated a frequent MRI scanning protocol, providing important insights into the dynamics of different MS subgroups and highlighting marked differences between subgroups in terms of activity profile.

Serial studies of patients with benign MS have shown that this subgroup is less active on MRI than patients with early relapsing remitting MS (Thompson et al., 1992; Kidd et al., 1994). Furthermore, only a third of new lesions display enhancement with gadolinium, suggesting that those lesions that do occur may be less inflammatory (Thompson et al., 1992).

Studies of patients with relapsing remitting MS have demonstrated that MRI activity on both enhanced and unenhanced sequences occurs more frequently than clinical relapse (Isaac et al., 1988; Willoughby et al., 1989; Harris et al., 1991; Barkhof et al., 1992; Smith et al., 1993; Frank et al., 1994; Kidd et al., 1996; Thorpe et al., 1996b). However, there are wide inter individual differences in the level of activity demonstrated on serial MRI. Furthermore, there is substantial month to month variation in the number of enhancing lesions detected. Bursts of MRI activity have been found in association with new clinical activity (Frank et al., 1994), but no predictable pattern of enhancement has been identified.

The degree of MRI activity in secondary progressive patients is also subject to substantial differences between individuals. Serial studies have indicated that the level of enhancement in this subgroup is broadly similar to that of relapsing remitting patients (Thompson et al., 1991; Tubridy et al., 1998a). However, a lower level of MRI activity has been identified in those secondary progressive patients without ongoing relapses in comparison with those continuing to relapse (Kidd et al., 1996; Tubridy et al., 1998a ).
MRI studies investigating primary progressive multiple sclerosis have shown clear differences in comparison with the other MS subgroups (Thompson et al., 1997). Cross sectional studies have demonstrated that the level of cerebral abnormality in this subgroup is smaller than that identified in benign and secondary progressive MS (Thompson et al., 1990; Filippi et al., 1995a), despite marked disability. Furthermore, longitudinal studies have shown that (i) fewer new lesions occur over time in primary progressive MS and (ii) those lesions that do occur are less likely to demonstrate gadolinium enhancement (Thompson et al., 1991). Even using a triple dose of gadolinium to increase sensitivity, few enhancing lesions are seen (Silver et al., 1997). This suggests that lesions in primary progressive MS may be of a less inflammatory nature than those in other MS subgroups, a finding further supported by post mortem work (Révész et al., 1994).

1.2.4 The prognostic utility of MRI in clinically isolated syndromes

In more than 90% of patients who subsequently develop MS, clinical presentation is with an acute neurological disturbance with subsequent remission. This initial episode typically involves the optic nerves, brain stem or spinal cord. Several studies have reported the MRI findings of patients presenting with such clinically isolated syndromes and defined their predictive value for subsequent development of multiple sclerosis (Beck et al., 1993; Morissey et al., 1993). Where first presentation is with optic neuritis, conventional MRI reveals clinically silent white matter lesions in 50-70% of patients. Where an alternative diagnosis has been excluded, a similar proportion of patients presenting with isolated syndromes of the brain stem and spinal cord also demonstrate such lesions. Based on five years of follow up, one study found that if four or more lesions are identified at presentation, the positive predictive value of conversion to definite MS was 65%, while the negative predictive value of a normal brain scan at presentation is 97% (Morrissey et al., 1993). Another study on patients presenting with optic neuritis (Beck et al., 1993) found that patients fulfilling the Paty criteria (see table 1.5) had a 35% likelihood of
developing definite MS at two years; more recent work has confirmed these findings (Tas et al., 1995). This predictive value may be further improved by incorporating information from gadolinium enhanced imaging (Barkhof et al., 1997a).

It is therefore apparent that MRI has a powerful role in predicting the risk of conversion to clinically definite MS from a clinically isolated syndrome. This is important in counselling individual patients. MRI information can also be used to select patients for trials of therapies aimed at preventing evolution from an isolated syndrome to clinically definite MS (Miller 1998).

1.3 MS treatment trials

1.3.1 Trial design

In recent years there has been much excitement with the advent of MS therapies that have demonstrated a positive impact on disease course (Paty et al., 1993; Johnson et al., 1995; Jacobs et al., 1996; PRISMS Study Group 1998). A substantial number of putative treatments have been advocated over the years, but it was the development of appropriate clinical trial design that laid the foundations for a rational and objective approach to assessing efficacy. By convention, clinical trials can be divided into four phases:

1) Phase I studies involve the initial introduction of a putative treatment into humans, with the primary aims of establishing its pharmacokinetics, identifying potential side effects, and establishing an appropriate dosage strategy.

2) Phase II studies (also called exploratory or preliminary) aim to further define dosing strategy in patients with the disease under study. Phase II trials should also provide information on treatment efficacy and side effects. They should be well controlled and
closely monitored, involving no more than several hundred patients.

3) Phase III studies (also called definitive, advanced or pivotal) include expanded controlled and uncontrolled trials, and are undertaken only after preliminary work has suggested treatment efficacy. Since up to several thousand patients may be entered into such studies, a multicentre design is typical. The aim is to establish the overall risk-benefit ratio. The gold standard for phase III studies is the randomised, placebo-controlled and double-blind trial that uses an intention to treat design. Sample size estimations should be performed as part of study design, in order to ensure an adequate power to detect treatment effect. Clear, objective outcome measures need to be defined, together with a specific primary endpoint. The design of the study, sample size estimations and statistical methods chosen should reflect the primary outcome measure.

4) Phase IV studies. These begin after drug licensing and consist of post marketing accumulation of data on safety and efficacy in clinical practice.

1.3.2 Clinical outcome measures
The major goal of any trial must be to effect clinical outcome favourably. This requires valid and objective measures of clinical status with which to monitor treatment efficacy. An ideal clinical scoring system should have the characteristics listed in Table 1.6. In addition, for statistical purposes, the ideal system should be interval (not ordinal) and linear in nature. In view of the typical clinical patterns of early relapsing remitting and later progressive disease, measures of both relapse frequency/severity and impairment/disability are generally adopted.
Criterion | Explanation
--- | ---
Sensitivity | The measure should be sensitive to disease worsening over a relatively short period
Reliability | The score should be derived from objective criteria and have high intra and inter rater reproducibility
Validity | The test instrument measures impairment caused directly by the disease and disability that is clinically relevant
Measure contains components that reflect the independent dimensions of MS | The test instrument measures the principal independent dimensions of the disease, but contains minimally redundant components
Measure is applicable to a range of MS impairments | Available scores should allow classification of all patients and avoid ceiling effects
Ease of administration | The measurement instrument should be quick and easy to administer
Cost effectiveness and efficiency | The measurement instrument should be conservative of time and resources

Table 1.6 Requirements for MS clinical outcome measures (from Whitaker et al., 1995)

1.3.2.1 Relapse frequency

The relapse frequency is a seemingly straightforward clinical outcome. A relapse is generally defined as new or worsening MS symptoms lasting longer than 24 hours. However, difficulties exist in formally defining what is meant by a relapse, and quantifying its severity. Furthermore, relapse frequency has been shown to decrease over time, and the relationship between relapse rate and long term disability is uncertain (Weinshenker et al., 1989; Runmarker & Andersen 1993). New scales are needed that can quantify relapse severity, duration and recovery.
1.3.2.2 The expanded disability status scale and functional systems

The Kurtzke disability status scale (DSS) (Kurtzke 1955) was designed to measure dysfunction in several different neurological systems, incorporating functional status scores (FSS) to provide a score for disability. The score was subsequently refined to the expanded disability status scale (EDSS) (Kurtzke 1983) in which disability is measured on a scale of zero (normal) to 10 (dead), in increments of 0.5 points. While it has been criticised with respect to several limitations (Willoughby & Paty 1988), currently, the EDSS is still the most widely used clinical outcome instrument. The advantages of the EDSS are its relative ease of administration, familiarity to investigative clinicians and the fact that it allows a simple comparison between patients or within a single patient over time (Whitaker et al., 1995). However, it has several major limitations as listed here:

- It lacks linearity, such that incremental steps are unequal. For example, a change from EDSS 1 to 2 is not equivalent to a change from 5 to 6. It is therefore a rank order, ordinal scale, to which parametric statistics cannot be applied.

- In any population of patients, the distribution of EDSS scores is clustered at levels 3-4 and 6-7, with fewer patients at EDSS 4.0-4.5. This reflects the finding that different time periods are spent at each level (Weinshenker et al., 1991).

- The scale is heavily weighted towards ambulation, with less emphasis on upper limb and cognitive function. For example, upper limb dysfunction contributes to the EDSS score only between grades 8-9.

- Certain elements of the FSS are subjective, with descriptors such as mild, moderate and severe.

- Inter rater reproducibility is sub optimal (Noseworthy et al., 1990), and varies according to the range. In the lower range, from EDSS 1-3.5, inter rater
agreement within 1.0 EDSS point is only 60-70% (Goodkin et al., 1992).

- The EDSS score has a low level of responsiveness, with several years often spent at a given EDSS level.

With these limitations in mind, a number of other rating scales have been developed for use either on their own, or in conjunction with the EDSS (Table 1.7). However, none of these scales has yet replaced the EDSS as the established outcome measure in clinical trials, and none meets all the criteria listed in Table 1.6.
Impairment scales
   Expanded Disability Status Scale
   Functional systems*
   Disability status scale*
   Disease Steps

Quantified neurological examination scales
   Scripps Neurologic rating scale*

Quantitative neuroperformance tests
   Quantitative evaluation of Neurologic Function
   Biochemical isometric strength testing
   Box-and-block and nine-hole peg tests*

Combination Scales
   Ambulation index*
   Cambridge basic multiple sclerosis score

Disability scales
   Incapacity status scale
   Environmental status scale*
   Functional independence measure

Cognition Scales
   Standard neuropsychometric batteries
   Screening tests

Global opinions

Self-rating scales

Quality of life scales
   Farmer quality of life index
   Multiple Sclerosis Quality of Life Instrument

*Used primarily with the EDSS; * Essentially replaced by the EDSS

Table 1.7. Clinical outcome measures in patients with MS (from Wingerchuk et al., 1997)
1.3.3 Surrogate outcome measures

The above limitations, combined with the unpredictable and highly variable clinical course, dictate that phase III studies of several years duration, involving hundreds of patients are required. Consequently, efforts have been directed towards the identification and use of alternative outcome measures to monitor efficacy. The term 'surrogate' is used to define a non clinical measure that can be used in place of, or as a substitute for, clinical outcome indices as a trial endpoint. The Food and Drug Administration (FDA) uses the term more rigorously, to indicate a non clinical test that can predict ultimate clinical change in a long term disease (Food and Drug Administration 1992). Potential candidates for surrogate markers in MS include serial evoked potentials, clinical laboratory measurement and MRI. Previous work using evoked potentials has generally failed to demonstrate a relationship with clinical activity (Bednarik & Kadanka 1992), although a recent small study has demonstrated a significant cross sectional correlation between evoked potential abnormalities and EDSS (O'Connor et al., 1998). At present, no laboratory markers are suitable candidates as surrogate markers (Whitaker et al., 1995), and MRI is currently the most established candidate surrogate.

1.4 MRI as an outcome measure

MRI outcomes are now routinely used in both phase II and phase III MS treatment trials. There are several perceived advantages of MRI outcome measures over clinical indices:

- Objectivity. Blinding is easily maintained since analysis can be performed remote from the patient.

- Retrievability. In contrast to clinical assessment, MRI data forms a retrievable document, lending itself to re-analysis. This is important for data audit, in which,
for example, other investigators may duplicate the analysis in order to confirm original findings (Filippi et al., 1998a).

- **Sensitivity.** Serial gadolinium enhanced monthly MRI detects 5-10 new lesions for every clinical relapse in both relapsing remitting and secondary progressive disease. Natural history studies have demonstrated frequent enhancing lesions occurring even during periods of clinical remission (Grossman et al., 1988; Miller et al., 1988; McFarland et al., 1992). Several strategies have more recently been developed to improve sensitivity still further, notably weekly imaging (Lai et al., 1996), magnetisation transfer imaging (Silver et al., 1997), triple dose gadolinium and delayed imaging (Filippi et al., 1996a). Serial gadolinium enhanced scanning allows an assessment of treatment effect much more quickly and in smaller numbers of patients than would be possible based on clinical monitoring alone.

- **Cost effectiveness.** While MRI is expensive, its ability to demonstrate treatment effect rapidly, using smaller cohorts of patients, does render it cost effective.

### 1.4.1 MRI in exploratory (phase II) studies

Given the marked sensitivity of MRI to asymptomatic disease activity, serial monthly contrast enhanced MRI has now become the primary outcome measure of choice in exploratory (phase II) studies in relapsing remitting and secondary progressive MS (Miller et al., 1996). This approach is supported by the relationship between gadolinium enhancement and relapse. Gadolinium enhanced imaging is now routinely used for screening new therapies where the proposed mechanism of action is in suppressing new pathological activity and inflammation.
A positive treatment effect can be shown with fewer patients and in a shorter time than is required to demonstrate an effect on relapse rate. Lesion activity is generally defined in terms of new or persistent enhancement on T₁ weighted images and new/enlarging lesions on T₂ weighted images (Miller et al., 1996). While the majority of new T₂ lesions demonstrate gadolinium enhancement, the concurrent use of a T₂ weighted image is recommended to optimise sensitivity and confirm equivocal activity seen on the enhanced images.

<table>
<thead>
<tr>
<th>Therapy</th>
<th>Design</th>
<th>Effect (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta Interferon 1b</td>
<td>Parallel groups (RR)</td>
<td>60-75</td>
<td>Paty et al., 1993</td>
</tr>
<tr>
<td>Alpha interferon</td>
<td>Parallel groups (RR)</td>
<td>95</td>
<td>Durelli et al., 1994</td>
</tr>
<tr>
<td>Beta Interferon 1b</td>
<td>Baseline crossover (RR)</td>
<td>75</td>
<td>Stone et al., 1995a</td>
</tr>
<tr>
<td>Beta Interferon 1a</td>
<td>Parallel group (RR)</td>
<td>50</td>
<td>Jacobs et al., 1996</td>
</tr>
<tr>
<td>Beta Interferon 1a</td>
<td>Baseline crossover (RR)</td>
<td>64</td>
<td>Pozzilli et al., 1996</td>
</tr>
<tr>
<td>Campath-1H</td>
<td>Baseline crossover (SP)</td>
<td>90</td>
<td>Moreau et al., 1994</td>
</tr>
<tr>
<td>Cladribine</td>
<td>Parallel groups (SP)</td>
<td>90</td>
<td>Sipe et al., 1994</td>
</tr>
<tr>
<td>Mitoxantrone</td>
<td>Parallel groups (RR/SP)</td>
<td>80</td>
<td>Edan et al., 1997</td>
</tr>
<tr>
<td>*Linomide</td>
<td>Parallel group (RR)</td>
<td>70</td>
<td>Andersen et al., 1996</td>
</tr>
<tr>
<td>*Linomide</td>
<td>Parallel group (SP)</td>
<td>55</td>
<td>Karussis et al., 1996</td>
</tr>
<tr>
<td>Intravenous</td>
<td>Double crossover (RR)</td>
<td>70</td>
<td>Sorensen et al., 1998</td>
</tr>
<tr>
<td>immunoglobulin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copolymer</td>
<td>Baseline crossover</td>
<td>57</td>
<td>Mancardi et al., 1998</td>
</tr>
</tbody>
</table>

*Phase III trials of linomide were subsequently discontinued due to unexpected side effects.

Table 1.8 Studies showing a reduction in MRI activity
Both crossover and parallel group designs have been used, although a crossover design is more powerful (Nauta et al., 1994). Using such an approach, a significant treatment effect has been demonstrated in only 7 patients (Moreau et al., 1994). However, a baseline crossover design is susceptible to regression to the mean, where patients are entered into a study during a particularly active clinical and MRI phase with subsequent improvement not attributable to treatment effect; a parallel group design is therefore more robust. Sample size estimations have recently been performed for both relapsing remitting and secondary progressive MS subgroups, based on a placebo-controlled, parallel group design over six months, using a monthly gadolinium enhanced imaging protocol (Tubridy et al., 1998b). With a single pre treatment scan, detection of a 70% reduction in MRI activity on treatment with adequate power would require for relapsing remitting and secondary progressive MS subgroups respectively, only 2 x 30 and 2 x 50 patients. With an additional pre treatment scan, sample sizes are even smaller at, respectively, 2 x 20 and 2 x 30 for relapsing remitting and secondary progressive MS patients.

1.4.2 MRI in definitive (Phase III) treatment trials

While there are now a number of MRI techniques available to monitor disease progression in MS, none has yet replaced clinical indices as the primary outcome measure in phase III trials. The major reason for this is the lack of strong relationship between clinical and MRI markers of progression. The most important requirement of a surrogate marker is that it should be predictive of long term clinical outcome (Whitaker et al., 1995). The most established MR surrogate for disease progression is serial brain T₂ total lesion volume (TLV) measurement. In practice, it is often a shorter TE, PD weighted sequence that is used in lesion volume measurement rather than a strictly T₂ weighted sequence, but by convention, terms such as T₂ lesion volume and T₂ lesion load are still used.
Several natural history studies and clinical trials have demonstrated significant correlations between EDSS and T₂ lesion volume both cross sectionally and longitudinally. However, the level of correlation is typically modest (Table 1.9). This may be in part due to the limitations of the EDSS, but several other factors are also likely to be important:

- Increased signal on a conventional T₂ weighted image results from a change in the local water environment and T₂ relaxation time. Any stage of MS lesion development and evolution, from early inflammation and oedema, to chronic, destructive pathology with demyelination and axonal loss, can therefore be represented as an area of high signal that contributes to the measured brain lesion volume. Considering this lack of specificity for the more destructive elements of MS pathology, it is not surprising that stronger correlations have not been demonstrated.

- Standard measures of brain lesion volume give equal weighting to all brain lesions, regardless of their anatomical location. Since the majority of MS lesions occur in regions that do not directly contribute to locomotor disability, it is not at all surprising that the association between the EDSS, with its emphasis on locomotor function, and total brain lesion volume is poor. When corticospinal tract lesion volume is considered in isolation, a slightly stronger correlation has been demonstrated (Riahi et al., 1998). Furthermore, measurement of brain lesion volume ignores the impact of spinal cord lesions that may substantially contribute to gait disturbance. However, little correlation between cord lesion volume and EDSS has been observed (Kidd et al., 1993), suggesting that lesion site per se is not decisive in determining disability.

- The contribution of measurement error in attenuating the level of demonstrated
correlation may be substantial. A number of potential sources of measurement error exist in both image acquisition and data analysis. The anticipated annual change in brain lesion volume is small, typically 5-10%, and if a measuring instrument is to be responsive to such a change, a high level of accuracy and reproducibility are prerequisites. Efforts are now being directed towards identifying and resolving the many possible sources of measurement error, and this may yet lead to the demonstration of stronger clinical/MRI correlations.

- There is growing evidence that the lesion volume visible on a conventional T₂ weighted sequence does not provide a full measure of the total disease burden. Postmortem work has revealed microscopic changes in the normal appearing white matter (NAWM) (Allen et al., 1979). Furthermore, MR studies using magnetisation transfer imaging (Dousset et al., 1992; Filippi et al., 1995b; Loevner et al., 1995a), MR spectroscopy (Arnold et al., 1992; Husted et al., 1994; Davie et al., 1997; Rooney et al., 1997), and T₁ and T₂ measurement (Barbosa et al., 1994; Gasperini et al., 1996) have demonstrated that the contribution of such MR occult pathology to the total burden of abnormality may be important. The presence of microscopic pathology in the NAWM may well contribute to the lack of clinical/MRI correlation using T₂ lesion volume measurements alone.

- Cortical lesions, poorly seen on T₂ weighted MRI, and cortical adaptation to subcortical pathology (measurable using functional MRI) may both affect disability independent of T₂ lesion volume.

Due to this uncertain relationship between MRI and long term clinical outcome, it is appropriate that the primary endpoints in a phase III study are clinical (Miller et al., 1996). This situation
may yet change if improvements in clinical and MR measuring instruments reveal stronger
correlations. At present, however, the role of MRI is as a secondary outcome measure, providing
important additional information on the long-term effect of treatment on the pathological process.

Its utility in this regard was first demonstrated by the North American trial of interferon beta-1b
in relapsing remitting MS, where the effect of treatment on stabilisation of brain $T_2$ lesion
volume and the development of new $T_2$ lesions provided powerful support for treatment efficacy
(Paty et al., 1993).

<table>
<thead>
<tr>
<th>Study</th>
<th>Clinical/MRI parameter</th>
<th>r-value for correlation</th>
<th>No. of patients</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyclosporine (SP) Parallel group</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.22</td>
<td>163</td>
<td>Zhao et al., 1997</td>
</tr>
<tr>
<td>Cyclosporine (SP) Parallel group</td>
<td>Change in EDSS vs change in lesion volume</td>
<td>0.19</td>
<td>163</td>
<td>Zhao et al., 1997</td>
</tr>
<tr>
<td>Natural History (RR and SP)</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.33</td>
<td>46</td>
<td>Truyen et al., 1996</td>
</tr>
<tr>
<td>Natural History (RR, SP, benign and PP)</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.33</td>
<td>43</td>
<td>Gass et al., 1994</td>
</tr>
<tr>
<td>Interferon beta 1a (RR) Parallel group</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.22</td>
<td>237</td>
<td>Simon et al., 1998</td>
</tr>
<tr>
<td>Natural History (RR, SP, benign and PP)</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.46</td>
<td>41</td>
<td>Gasperini et al., 1996</td>
</tr>
<tr>
<td>Natural History (RR, SP, benign and PP)</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.30</td>
<td>130</td>
<td>Mammi et al., 1996</td>
</tr>
<tr>
<td>Natural History (RR, SP, benign and PP)</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.49</td>
<td>56</td>
<td>Gawne-Cain et al., 1998</td>
</tr>
<tr>
<td>Natural History (RR)</td>
<td>Baseline EDSS vs baseline lesion volume</td>
<td>0.60</td>
<td>39</td>
<td>Riahi et al., 1998</td>
</tr>
</tbody>
</table>

Table 1.9 Significant clinical/MRI correlations from recent studies
1.4.3 Putative MR Markers of Disease Progression

Two major limitations of the conventional T₂ weighted sequence are; (i) the lack of pathological specificity, and (ii) the inability to detect microscopic focal or diffuse pathology beyond the resolution of standard MR. Other MRI techniques offering the prospect of greater pathological specificity are now being investigated. Examples include quantification of T₁ hypo-intense lesions, magnetisation transfer (MT) imaging, the serial measurement of brain and spinal cord atrophy, and MR spectroscopy. These approaches may offer the potential to selectively identify the more destructive pathological elements of the disease, demyelination and axonal loss. MR spectroscopy and MT imaging are also able to identify subtle abnormalities in the NAWM. There is emerging evidence that such MRI tools may be more strongly predictive of long term outcome than conventional T₂ weighted imaging. Several such putative markers are now being assessed in a clinical trial context (Polman et al., 1995); this will allow an opportunity to validate them with respect to evolving clinical course.

1.4.3.1 Magnetisation transfer imaging

MT imaging provides a means of investigating macromolecular protons that are essentially invisible on conventional MR because of their rapid T₂ relaxation. This 'bound' proton pool can be indirectly imaged by application of a saturation pre-pulse to saturate it selectively. Subsequent transfer of saturation to the 'free' pool occurs via cross-relaxation, spin diffusion and chemical exchange (Wolff & Balaban 1989; Wolff & Balaban 1994). The result of this transfer is reduction in the intensity of the free proton signal; this is termed the MT effect. The magnitude of this effect, the MT ratio (MTR) can be quantitatively measured by acquisition of two sets of images, with and without a saturation pulse, all other parameters remaining identical.
The MTR of a given region of interest is then given by the following equation:

\[ \text{MTR} = \frac{(M_0 - M_s)}{M_0} \]

where \( M_0 \) and \( M_s \) represent the mean pixel intensity of a region without and with saturation respectively. The MTR therefore quantifies the fractional signal loss due to saturation of the bound pool, providing a reflection of the amount of structure in a given region. In healthy brain white matter, the predominant contribution to MTR values is likely to be produced by the presence of myelin (Fralix et al., 1991). While not specific to demyelination, there is evidence that demyelination, either with or without concurrent axonal loss, can cause MTR reduction (Lexa et al., 1994; Dousset et al., 1994; Dousset et al., 1995; Thorpe et al., 1995; Filippi et al., 1998b). Serial studies incorporating gadolinium enhanced imaging have shown an initial reduction in MTR, with subsequent complete or incomplete recovery (Dousset et al., 1995; Lai et al., 1997; Silver et al., 1998a; Filippi et al., 1998b). While this may in part reflect development and resolution of oedema, the contribution of demyelination and subsequent remyelination may also be important (Lai et al., 1997).

MT imaging can be used to characterise MS lesions visible with conventional imaging (Gass et al., 1994), but also offers the potential for investigating microscopic changes in the NAWM (Dousset et al., 1992; Filippi et al., 1995b; Loevner et al., 1995a). While investigating the MT characteristics of a particular region of interest can provide important data on finite areas of lesion or NAWM, such an assessment does not describe the total burden of abnormality. Histogram analysis of MTR pixel distributions may be able to provide whole brain quantitative data, offering a global measure of disease burden in MS (van Buchem et al., 1996; van Buchem et al., 1997). Comparison of MT histograms between controls and MS patients has revealed a
decreased height of the peak of the normalised distribution of pixels (van Buchem et al., 1996; van Buchem et al., 1997), indicating an increase in the proportion of tissue with low MTR values in the MS patients. A number of MTR histogram measures are now revealing stronger clinical/MRI correlations than conventional T2 weighted MR indices (Silver et al., 1998b; van Waesberghe et al., 1998; Iannucci et al., 1998). Furthermore, analysis techniques are emerging with the potential for fast, automated assessment. MT imaging may therefore provide a powerful tool for monitoring disease progression in MS, and longitudinal studies are now underway in order to define its utility in a trial context further. However, more work is needed to address issues of both standardisation of acquisition parameters across multiple centres and quality control (Berry et al., 1996; Barker et al., 1997).

1.4.3.2 Hypo-intensity on T1 weighted images (black holes)

A proportion of MS lesions visible on T2 weighted sequences can also be seen as areas of chronic hypo-intensity on a corresponding T1 image (Uhlenbrock & Sehlen 1989). There are several reasons to suggest that such areas are a more specific marker for severe tissue destruction than T2 hyper-intensity, reflecting demyelination and axonal loss. First, low MTRs have been demonstrated in hypo-intense lesions, suggesting severe structural loss (Gass et al., 1994; Hiehle et al., 1995; Loevner et al., 1995b). Second, loss of MT has been inversely correlated with T1 relaxation time (van Waesberghe et al., 1997). Third, a relationship between accumulation of T1 hypo-intense lesion volume and progression of disability has been identified (Truyen et al., 1996). Fourth, secondary progressive MS patients have a larger proportion of T2 hyper-intensity regions associated with corresponding T1 hypo-intensity than relapsing remitting patients, perhaps suggesting a deficit in available repair mechanisms in the former group (Truyen et al., 1996). Furthermore, more enhancing lesions remain persistently hypo-intense after six months follow up in secondary progressive than relapsing remitting MS, suggesting that enhancing
lesions are more destructive in secondary progressive disease (van Waesberghe et al., 1998). Finally, preliminary work with postmortem MRI on unfixed brain has demonstrated a relationship between $T_1$ hypo-intensity and tissue destruction, with features of axonal loss and an expanded extracellular space (van Walderveen et al., 1998).

Several of the above studies have confirmed that $T_1$ hypo-intense lesion volume provides a stronger level of correlation with disability than $T_2$ hyper-intense lesion volume. However, the incorporation of quantitative measurement of $T_1$ hypo-intense lesion volume in a multicentre study will require both standardisation of imaging parameters between centres and clear guidelines defining what constitutes hypo-intensity.

1.4.3.3 Spinal cord imaging

It has been suggested that the lack of strong correlation between $T_2$ lesion volume and EDSS may be in part attributable to the presence of spinal cord disease (McDonald 1992). However, the number and extent of spinal cord lesions have not been shown to correlate with EDSS (Kidd et al., 1993; Kidd et al., 1996; Nijeholt et al., 1998). Furthermore, serial studies have demonstrated a low frequency of new spinal cord lesions, although those that do occur are clinically expressed more often than brain lesions (Kidd et al., 1996; Thorpe et al., 1996b). Therefore, while standard spinal cord imaging is of proven diagnostic value (Thorpe et al., 1996a), its utility in a trial context is less certain.

The quantification of spinal cord cross sectional area may be a more promising tool for monitoring disease progression. Several groups have investigated the potential of spinal cord atrophy measurement as a marker of axonal loss, and a significant relationship between EDSS
and cord area has been confirmed (Kidd et al., 1993; Kidd et al., 1996; Filippi et al., 1997b, Losseff et al., 1996a; Stevenson et al., 1998). Longitudinal studies are now being performed to evaluate its predictive value over time and utility in clinical trials (Polman et al., 1995; Leary et al., 1997; Stevenson et al., 1998). While earlier studies had methodological limitations and suboptimal reproducibility, even small changes in spinal cord cross sectional area can now be reliably measured (Stevenson et al., 1998). The technique may be particularly useful as a prognostic marker in monitoring primary progressive patients, since this group have a lower rate of new lesion development than is seen in relapsing remitting and secondary progressive cohorts (Thompson et al., 1997). A large longitudinal study is underway to investigate this further (Leary et al., 1997).

1.4.3.4 Brain volume quantification

The serial quantification of brain white matter volume may offer a useful tool for monitoring the development and progression of cerebral atrophy. The most potent pathological contributor to atrophy is likely to be axonal loss (Losseff et al., 1996a; Losseff et al., 1996b); such changes may therefore be highly relevant in functional terms. A recent study has demonstrated that cerebral volume can be measured with a high level of scan-rescan reproducibility, sufficient to allow detection of as little as 1.1% change within 95% confidence limits (Losseff et al., 1996a). In this study, significant progression in atrophy was demonstrated in more than half the patients imaged over 18 months. Furthermore, a significant relationship between progression in atrophy and sustained increase in EDSS was found. By combining such techniques with thin slice 3D acquisition (Filippi et al., 1996c) and accurate image registration/subtraction (Fox et al., 1996), it should prove possible in future to monitor the progression of atrophy with a high level of sensitivity.
1.4.3.5 Proton MR Spectroscopy

MR spectroscopy can be used to assess metabolite concentrations in both visible MS lesions and NAWM. In particular, the serial study of N-acetyl aspartate (NAA) concentration may provide a marker for axonal loss (Arnold et al., 1990; Davie et al., 1995). The concentration of NAA has been correlated with functional impairment in cerebellar white matter (Davie et al., 1995). MR spectroscopy has also identified abnormalities of metabolite concentrations in NAWM (Arnold et al., 1992; Husted et al., 1994). However, transient decreases in NAA have been observed in acute MS lesions, followed by partial recovery (Davie et al., 1994; De Stefano et al., 1995). This suggests that NAA concentration may not simply measure axonal density alone, since this level of cellular recovery would not be expected. This may in part reflect a degree of reversible axonal functional impairment. The role of MR spectroscopy in monitoring therapeutic efficacy has yet to be determined. Current limitations in terms of time consumption, reproducibility and low signal-to-noise ratio (SNR) will need to be addressed, but in the future this technique may provide a powerful means to assess the impact of therapeutic intervention (Miller 1998).

1.4.3.6 Other potential putative markers

Several other MR parameters have been identified as potentially useful in a trial context. Examples include $T_1$ magnetisation decay curve analysis, perfusion and diffusion imaging. At present, the application of these techniques is restricted to a few highly specialised centres; wider implementation to multicentre trials will require more developmental work.
Part 1. Technical considerations for $T_2$ lesion volume quantification

Chapter 2. Background

To provide a measure of brain lesion volume requires both lesion identification and delineation. A substantial number of techniques are now available to perform such an analysis, and the following criteria can be used to judge their performance: accuracy, reproducibility, reliability, efficiency and stability over time.

2.1 Accuracy

This refers to the extent to which a technique measures the truth. The accuracy of lesion volume measurement is hard to establish since no perfect gold standard measure exists with which to compare measured values. Approaches to determining the accuracy of a technique include; (i) use of realistic phantoms where a true value is known (Tofts et al., 1997), (ii) the use of a simulated MR dataset (Evans et al., 1997), and (iii) comparison against a gold standard measure. At present, the recognised gold standard is manual outlining, performed in consensus by a group of experts or experienced raters (Filippi et al., 1998a).

2.2 Reproducibility

Reproducibility (or precision) refers to the degree with which repeated measurements on the same object are in agreement. The assessment of reproducibility is important, since it can define the extent of random measurement error that can be anticipated in an MRI study. There are multiple potential sources of variability in the measurement of $T_2$ lesion volume (Plante & Turkstra 1991; Clarke et al., 1995), reflecting inconsistencies in both image acquisition and
analysis. Examples include repositioning errors (Gawne-Cain et al., 1996; Simon et al., 1997; Filippi et al., 1997c), inconsistent scanner performance (Wang et al., 1997a), variation in scanner model and field strength (Filippi et al., 1997a), variable motion and flow artifact, and variability in the segmentation technique itself where any human interaction is required (intra and inter observer measure-remeasure variability). With automated techniques, repeated measurements on an individual MR dataset should necessarily be highly reproducible. However, the robustness of any segmentation technique can only be fully assessed by the analysis of multiple scans that are acquired within a short period, with the patients leaving the scanner between measurements (scan-rescan reproducibility) (Simon et al., 1997; Filippi et al., 1998a). This forms an important part of technique validation.

2.3 Reliability

Reliability can be defined as a measurement technique’s ability to discriminate between the different members of a sample population (Fleiss 1985; Streiner & Norman 1995). It describes the proportion of the variance in repeated measurements on a patient sample that is attributable to differences between patients. For a measurement technique to have perfect reliability, all the variance in repeated measurements must arise from systematic differences between subjects. Since the aim of serial $T_2$ lesion volume quantification is to discriminate between subjects and identify trends, reliability is an important consideration. An evaluation of a technique's reliability also allows the impact of measurement error on sample size requirements to be calculated (Fleiss 1985).

2.4 Efficiency

This describes the relative ease with which any technique can be applied in a treatment trial, particularly in terms of cost, computational and human resources. Many available segmentation
techniques require a substantial level of human intervention, in either delineating lesion boundaries, or in the process of reviewing the effectiveness of more complex algorithm-derived lesion volumes. With MR protocols for phase III studies typically requiring analysis of many hundreds (or even thousands) of scans, efficiency is clearly at a premium. For more operator-dependant techniques, analysis of the entire dataset by a single observer may not be possible within a reasonable time frame, thereby introducing inter observer inconsistency as a further potential source of variability. Considerable efforts are now being directed towards the development of more efficient quantitative techniques.

2.5 Stability over time

Measurement stability is a further important consideration. While it is imperative that the image acquisition methodology remains stable, consistency in the application of the segmentation technique is also essential. The likely occurrence of operator drift over time has already been demonstrated in two phase III clinical trials that used T2 lesion volume quantification to provide an outcome measure (Paty et al., 1993; IFNB MS Study Group 1995; Jacobs et al., 1996; Simon et al., 1998). In one study (Paty et al., 1993), a step decrease in lesion volume of about 9% was seen in the placebo arm after three years. This was attributed to a change of strategy by the single observer over the course of the study. Subsequent re-analysis confirmed that this step change was artificial (IFNB Study Group 1995). In a second, more recent study (Jacobs et al., 1996; Simon et al., 1998), a similar paradoxical reduction in T2 lesion volume in the placebo group was also identified. This was also attributed to measurement drift, disappearing on re-analysis with an improved quantification method based on manual outlining (Simon et al., 1998). The risk of such a step change in analysis technique application might be reduced by rigorous operator training and regular consistency checks on a representative dataset, to identify and address operator drift where this occurs. Furthermore, all the scans from an individual patient should be analysed by
the same observer, ideally within a single session (Filippi et al., 1998a).

2.6 Manual outlining

Until recently, manual outlining remained the gold standard technique for quantifying T₂ lesion volume. The process involves lesion identification by experienced observers, and subsequent computer-assisted delineation of lesion boundaries by manual tracing with a mouse or tracker-ball. The technique has been used in several clinical trials (Kastrukoff et al., 1990; Paty et al., 1993; Koopmans et al., 1993; Zhao et al., 1997; Simon et al., 1998). Manual outlining utilises human pattern recognition capabilities for discriminating lesion from artefact/normal anatomy. However, it suffers from two major disadvantages; (i) sub optimal reproducibility (Filippi et al., 1995d; Grimaud et al., 1996; Filippi et al., 1998c), and (ii) substantial time consumption. While comprehensive operator training has been shown to improve reproducibility to an extent (Filippi et al., 1998c), the high level of operator intervention results in intra and inter rater variability even for experienced observers. Furthermore, the analysis time is long at 30-60 minutes per scan. Therefore, many hundreds of hours of operator time are needed to analyse the large MRI datasets generated in phase III treatment trials. These limitations have led to attempts to develop quantitative techniques with a higher level of automation; the contour technique is one such approach. A study comparing the contour with the manual outlining technique is now presented.
Chapter 3. The Contour technique; a comparison with manual outlining

3.1 Introduction
The contour method of lesion segmentation (Plummer 1992; Grimaud et al., 1996) is based on a computer algorithm that can detect the intensity gradient across a lesion boundary. It offers the potential for better reproducibility than manual outlining, since the algorithm imparts a degree of automation to the process of lesion delineation. A cross sectional study comparing contouring against the manual outlining method demonstrated a higher level of both intra and inter rater reproducibility with the former approach (Grimaud et al., 1996). In treatment trials, however, it is not the absolute lesion volume, but its change over serial studies that provides an outcome measure; the precision and reliability of lesion volume quantification in identifying such a change has not previously been defined. Furthermore, the measurement of reliability also provides a means for assessing the impact of measurement error on sample size requirements using $T_2$ weighted lesion volume as an outcome measure. The present study addresses these issues by evaluating the performance of two quantitative techniques, the manual outlining and contours methods, in measuring change in $T_2$ lesion volume over time.

3.2 Patients and Methods
The baseline and year two scans of 16 patients with clinically definite MS according to Poser criteria (Poser et al., 1983) were used for this study. These patients represented part of a larger cohort of patients that participated in the North American interferon beta-1b study on patients with relapsing remitting MS (Paty et al., 1993); eight patients were randomly selected from the placebo arm, and a further eight from the group treated with high dose (8 million IU) interferon beta-1b, thus providing a representative set of images that had been already used in a treatment trial. The EDSS scores of the patients ranged from 1.0 to 5.5.
3.2.1 MRI sequences

All MRI scans were performed on a 1.5 T Signa system (General Electric, Milwaukee, WI). Twenty-two contiguous axial images were obtained from foramen magnum to vertex using a slice thickness of 5 mm. A dual echo CSE sequence (TR; 2000 ms, TE; 30 and 70 ms, field of view (FOV); 200 mm, matrix; 128×256) was acquired on all patients at baseline, and then again after two years.

3.2.2 Lesion identification

The author identified and marked the MS lesions on the hard copies, with a final consistency review provided by a neuroradiologist. The baseline and year two studies of each patient were assessed together in order to allow consistent decisions to be made on whether or not to include equivocal lesions on the serial scans. Lesion identification and subsequent delineation by the three raters (see below) were performed on the shorter TE images.

3.2.3 Quantification of lesion volume

Three experienced raters performed the lesion volume quantification on Sun workstations (Sun Microsystems, Mountain View, Calif, USA), and only those lesions that had been marked on the hard copy were segmented. Analysis of all 32 scans was performed twice, independently by each of the three raters, using both techniques. This provided a means of assessing both intra (within) rater and inter (between) rater precision and reliability. The potential for any memory of the images to introduce systematic bias was minimised by randomising the scan order and ensuring a delay of at least one week between repeated measurements on the same scan.

The manual outline technique was performed on the computer display by tracing the lesion outline with a mouse-controlled cursor.
The contour method incorporates a local thresholding algorithm to trace the lesion boundary and runs as part of the Dispimage package (Grimaud et al., 1996; Plummer 1992). A point on the lesion edge is identified by the rater. The algorithm then finds the lesion edge by searching for the strongest local intensity gradient. The lesion is delineated by following the contour of iso-intensity, which is then displayed to allow expert review. Manual editing of part of the lesion boundary is sometimes necessary, in order to delete regions of increased signal not corresponding to lesion, particularly where the lesion-to-background contrast is poor.

Lesion volumes were calculated automatically for both techniques as the summed lesion areas on all slices multiplied by the slice thickness.

3.2.4 Statistics

Several statistical methods can be used to define the precision of a measurement technique. The coefficient of variation (CoV) has been used as a measure of precision in several previous cross-sectional studies (Paty et al., 1993; Grimaud et al., 1996; Filippi et al., 1996d; Gawne-Cain et al., 1996; Filippi et al., 1998d). This statistic was therefore calculated for the repeated measurements on the baseline scans in order to allow comparison with other recent studies. The CoV was calculated as the standard deviation (SD) of the replicated measurements divided by their mean (Zar 1984). The intra rater CoV averaged across the three raters was calculated for all sixteen baseline scans with each technique. The inter rater CoV for each baseline scan was averaged across the two repeats performed by each rater.

However, the CoV is limited as a measure of precision, due a tendency for its dependence on the magnitude of the measured value; an inverse relationship often exists between the lesion volume and the CoV of replicated measurements. Where this is the case, a single mean or median value
for the CoV cannot fully describe measurement precision across a wide range of lesion volumes. Therefore, CoVs were not used for assessing precision in measuring the change in lesion volume, as their value was found to be heavily dependent on the size of the measured change. Thus, repeatability coefficients were used to describe precision for measurements of the change in lesion volume (Bland & Altman 1986; Bland & Altman 1996). The difference between two measurements for the same subject is expected to be less than the repeatability coefficient in 95% of observations; precision is expressed in terms of the unit of measurement. The assumptions inherent in the repeatability coefficient are; (i) that there should be no systematic bias between replicated measurements and, (ii) no relationship exists between the standard deviation and the mean of the replicated measurements. For the baseline measurements, the second of these criteria were not met (the standard deviation was positively correlated with the magnitude of the lesion volumes) and repeatability coefficients were therefore not calculated. However, for the replicated measurements of the change in lesion volume, both criteria were fulfilled by the data in this study; this statistic was therefore used to describe precision in measurement of change in lesion volume.

An intraclass correlation coefficient (ICC) was calculated as a measure of intra and inter rater reliability for both absolute lesion volumes and the change in lesion volume (Fleiss 1985; Streiner & Norman 1995). Analysis of variance (ANOVA) was used to calculate the ICC using a model that treated raters as a fixed factor. The ICC gives the proportion of the total variance in the repeated measurements from several subjects that arises from the true variance between the subjects. It varies from zero (no reliability) to one (perfect reliability). An ICC was also used as a measure of the agreement between the results obtained with the two techniques (Armitage & Berry 1994).
Differences between lesion volumes and CoVs obtained with the two techniques were evaluated with the Wilcoxon signed ranks test. All calculations were performed using the SPSS package.

3.3 Results

3.3.1 Lesion volumes obtained by the two techniques

The baseline lesion volumes (Table 3.1) showed excellent agreement between the two techniques (ICC = 0.996), but the mean volume obtained with the manual outlining method was slightly higher (p = 0.01) with a bias of 3%. Agreement between the techniques for the change in lesion volume was also high (ICC = 0.910).

<table>
<thead>
<tr>
<th>Technique</th>
<th>Mean</th>
<th>Median</th>
<th>Min ~ Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Lesion Volume (cm³)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual outlining</td>
<td>18.8</td>
<td>12.2</td>
<td>2.0 ~ 77.1</td>
</tr>
<tr>
<td>Contour method</td>
<td>18.2</td>
<td>12.1</td>
<td>2.1 ~ 74.7</td>
</tr>
<tr>
<td>Change in Lesion Volume (cm³)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual outlining</td>
<td>+2.5</td>
<td>+0.8</td>
<td>−3.4 ~ +14.9</td>
</tr>
<tr>
<td>Contour method</td>
<td>+1.8</td>
<td>+0.5</td>
<td>−4.4 ~ +14.8</td>
</tr>
</tbody>
</table>

Table 3.1 Lesion volume measurements with the two techniques
3.3.2 Repeated measurements on the baseline scans by the three observers

The intra and inter rater performances are given in Table 3.2. The median intra rater CoV averaged across the three raters for the contour and manual outlining methods were 3.2% and 7.6% respectively (p < 0.005). The median inter rater CoV for the contour and manual outlining methods were 3.8% and 6.1% (p < 0.01). There was no significant difference between intra and inter rater CoV for the manual outlining (p = 0.1) or contour methods (p = 0.2) The intra and inter rater reliability values for both techniques were greater than 0.99 (Table 3.2).

<table>
<thead>
<tr>
<th>Manual outlining</th>
<th>Contouring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CoV (%)</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Intra rater</td>
<td>8.6</td>
</tr>
<tr>
<td>Inter rater</td>
<td>7.2</td>
</tr>
</tbody>
</table>

ICC = Intra class correlation coefficient; CoV = coefficient of variation

Note: The intra rater precision and reliability was calculated by pooling data acquired from the repeated measurements of all three raters on the 16 patients. The inter rater rater precision and reliability were averaged across the two sets of measurements with each technique.

Table 3.2 Intra and inter rater precision (CoV) and reliability (ICC) for absolute lesion volumes
3.3.3 Repeated measurements of change in lesion volume

The values for precision (repeatability coefficients) and reliability (ICC) for the change in lesion volume are given in Table 3.3. The intra and inter rater precision and reliability were better for the contour method than the manual outlining technique.

3.3.4 Time consumption for each technique

The time consumption of the two techniques was similar, requiring 20-40 minutes per scan once sufficient experience in their application had been gained.

<table>
<thead>
<tr>
<th></th>
<th>Manual outlining</th>
<th>Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Repeatability coefficient</td>
<td>ICC</td>
</tr>
<tr>
<td>Intra rater</td>
<td>3.3 (cm³)</td>
<td>0.939</td>
</tr>
<tr>
<td>Inter rater</td>
<td>3.7 (cm³)</td>
<td>0.921</td>
</tr>
</tbody>
</table>

Table 3.3 Intra and inter rater precision and reliability for lesion volume change

3.4 Discussion

Several cross sectional studies have confirmed a high level of precision in lesion volume quantification with newer techniques (Wicks et al., 1992; Filippi et al., 1995d; Grimaud et al., 1996). However, it is not the absolute lesion volume, but the difference between serial estimates that is measured to provide an endpoint in definitive treatment trials; no previous work has
defined the precision and reliability of such techniques in measuring this change. Clearly, measurement of any change requires a technique with a high level of precision, since random errors in measuring lesion volume at each time point will have a cumulative effect in compromising the identification of differences over serial MRI investigations. This is particularly important given that changes in T2 lesion volume measured on annual MRI are frequently small (typically 5-10% per annum).

In this study, the efficacy of two quantitative techniques for measuring lesion volume was examined. Lesion segmentation is only one of many potential sources of measurement error; the overall accuracy and precision of measurement is affected by errors at each stage. These results therefore ignore the impact of variable scanner performance. Furthermore, the effects of suboptimal repositioning (Gawne-Cain et al., 1996; Simon et al., 1997) and inconsistency in lesion identification have also not been considered, because the aim of this study was to define and compare the precision and reliability of the quantitative techniques themselves.

The values for intra and inter rater CoV obtained in this study confirm a higher level of precision with the contour technique than for manual outlining. A high level of agreement was also found between the lesion volumes obtained with the two segmentation techniques. The manual outline technique is currently regarded as a gold standard measure, but this study has shown that the contour method produces very similar lesion volumes to those derived with manual outlining, with the higher precision afforded by computer-assisted lesion delineation. The results also confirm that the contour method has a higher level of precision than the manual outlining technique in identifying differences in lesion volume between serial studies. This implies that, being less subject to random error, it represents a more powerful technique for identifying any effect of treatment on change in lesion volume.
Reliability provides an index of the ability of a measurement technique to discriminate between the different members of a sample population (Fleiss 1985; Streiner & Norman 1995). Even a technique that is highly precise may not be able to distinguish between patients if the population range of the measured value is narrow. The very high values of reliability for measurements of baseline lesion volumes are perhaps not surprising, given the wide range of lesion volumes on these scans. More significantly, however, the reliability for measuring the relatively small changes in lesion volume over two years was excellent with both techniques. This suggests that the variance due to random measurement error is small, compared with that due to the wide biological variability in lesion volume changes across the patient population. To exclude the possibility that sample variability had been increased by including eight patients treated with interferon beta-1b, the variance between patients for the change in lesion volume in the placebo group and for the group as a whole was subsequently analysed. The variance between patients was actually reduced by inclusion of the treated group; the values obtained for reliability were not therefore elevated by the sample choice.

The impact of less than perfect reliability on sample size estimations for treatment trials is illustrated by the following equation (Fleiss 1985);

\[ n = n*/R \]

where \( n^* \) is the sample size per group based on a perfect measurement technique, \( R \) is the reliability defined as the ICC and \( n \) is the sample size per group after incorporating the effects of measurement error. With values for reliability of greater than 0.9, the effect of measurement error on sample size requirements is clearly small for both segmentation techniques (measure-
ment error would necessitate an increase sample size in each arm of less than 11%). This reflects the wide distribution within the sample for the change in lesion volume, and might suggest that optimal precision may not be an imperative. However, in a more homogenous population, or with a shorter interval between serial studies, the higher precision of the contour method might be reflected in a more substantial difference in reliability between the two techniques, and it is clearly appropriate to use the more reliable segmentation method.

A major disadvantage of both quantitative techniques used in this study is the high level of human interaction that they require. Phase III treatment trials may require analysis of thousands of images; lesion identification and segmentation can therefore take months or even years to perform. Several automated quantitative techniques have recently been developed using multi-parametric approaches to perform lesion segmentation (Mitchell et al., 1994; Simmons et al., 1994; Udupa et al., 1996; Udupa et al., 1997; Evans et al., 1997; Simon et al., 1997). These offer the potential for considerably greater efficiency. However, the significant presence of motion artefacts, field inhomogeneity within images and partial volume effects can potentially lead to errors in lesion classification. Any such inconsistency over serial MR images will result in the inaccurate assessment of any change in lesion volume. These techniques should therefore be validated by demonstrating that they can remain responsive to real changes in lesion volume over time, in the presence of such artefacts. Despite the considerable time requirements of the contour technique, human intervention in lesion identification reduces the risk of mis-classification. Furthermore, if serial images are assessed together, consistent decisions can be made on whether or not to classify equivocal areas of high intensity as lesions. The contour method utilises both the ability of an experienced observer to discriminate between lesion, artefact and normal anatomy, and the higher degree of precision in lesion delineation than is possible with the fully manual technique.
Although the contour technique is more precise than manual tracing of the lesion boundary, the algorithm still requires an observer to place the cursor at a point on the lesion edge. Lesions may have poorly defined edges due to the effects of volume averaging. Several possible boundaries can be produced by the contour algorithm for less discrete lesions, depending on the exact position of cursor placement, and this substantially contributes to the residual inconsistency in derived lesion volumes. Two approaches may further improve precision in serial studies. The first is to optimise the lesion-to-background contrast and therefore reduce the amount of manual editing required. The fast FLAIR sequence suppresses CSF signal intensity and is reported in some (Filippi et al., 1996d; Bastienello et al., 1997; Filippi et al., 1998d) but not all (Gawne-Cain et al., 1998) cross sectional studies to improve precision with the contour method. The second approach is to use a smaller slice thickness in order to reduce partial volume effects. One effect of finite slice thickness is to cause tissue mixing at the perimeter of lesions and produce loss of edge definition. As slice thickness is reduced, volume averaging effects should be less apparent and this may improve precision in quantification of lesion volume. This issue will be addressed in Chapter 5.

In summary, this study has demonstrated that the contour technique represents a major improvement over manual outlining for lesion volume quantification in terms of precision and reliability; these results support the further use of the contour method of $T_2$ lesion volume quantification on serial MRI. Indeed, based on the data in this chapter, it was decided to use the contour technique in assessment of the $T_2$ lesion volume in the therapeutic trial of interferon beta-1b in secondary progressive MS, as described in Part 3 of this thesis.
Chapter 4. The performance of global thresholding with a histogram matching correction

4.1 Introduction

The global thresholding method represents another quantitative technique developed in an attempt to provide better reproducibility and speed than with manual outlining (Wicks et al., 1992). The method has previously required three observer dependant stages:

1. Identification of the optimum intensity threshold that segments hyper-intense lesions from background.
2. Deletion of those areas not corresponding to a lesion.
3. Addition of any lesions that were of lower intensity than the threshold (and had therefore not been segmented).

The necessity to perform all these steps inevitably leads to loss of automation, reproducibility and speed. The operator-dependant choice of an intensity threshold is a subjective process and represents a major determinant of the derived lesion volume (Filippi et al., 1996e). Reproducibility has been shown to improve when a single threshold is chosen by consensus (Filippi et al., 1995d). However, a different threshold must be chosen for each scan in serial studies due to the variations in scanner sensitivity that occur over time (arising, for example from differences in coil loading, receiver attenuation setting and scanner pre-amplifier gain). A more objective method for determination of the threshold is therefore a priority.

The histogram matching algorithm corrects for changes in scanner sensitivity over time, by matching the intensity distribution histogram of different MRI studies (Wang et al., 1997a). This
process involves two stages; (i) windowing of the image histograms to eliminate the influence of background signal, and (ii) minimising the windowed squares of the residual differences between the two matched histograms (Wang et al., 1997a).

The aim of this study was to determine whether histogram matching would allow a single global threshold to be applied across multiple scans. Furthermore, post threshold editing is currently required to both delete non-lesion areas and add those lesions missed by the threshold. This process also leads to a loss of reproducibility and speed. The proportion of spurious non-lesion regions increases with a lower threshold choice, but the use of a relatively low threshold has been advocated since it reduces the requirement for adding lesions missed by the threshold (Filippi et al., 1996e). Without such post threshold editing, the calculated lesion volumes will inevitably include areas of hyper-intensity not corresponding to a lesion. However, if these areas were consistently included over serial studies, any change in lesion volume over time should still be detected.

These issues were therefore addressed by comparing T2 lesion volumes derived by the histogram corrected global threshold method against contour-derived lesion volumes, both cross sectionally and longitudinally. The contour method was used as a gold standard as it yields lesion volumes similar to manual outlining with better reproducibility, as discussed in chapter 3.

4.2 Patients and Methods

The baseline and month 9 MRI scans of 8 patients with clinically definite relapsing remitting or secondary progressive MS fulfilling Poser criteria (Poser et al., 1983) were selected. All MRI scans were obtained on a 1.5 T imager with 28 contiguous 5 mm axial slices from foramen magnum to vertex. A CSE sequence was used with the following imaging parameters; TR 2000
ms, TEs 34 and 90 ms, matrix 256x256, FOV 25 cm. Non-uniformity corrections were applied to all of the images using an oil-phantom based correction (Wicks et al., 1993).

Image analysis was performed on a Sun workstation (Sun Microsystems, Mountain View, CA). The contouring method (Plummer, 1992; Grimaud et al., 1996) has been described in Chapter 3. The author performed the lesion quantification on all the images using the contour method. The baseline and follow up scans were analysed on separate occasions and without reference to each other. In order to measure intra-rater reproducibility, the author repeated the lesion quantification on the same MRI dataset after a delay of at least one week. Analysis was also performed by a second rater, to obtain a measure of inter-rater reproducibility. Lesions were identified independently by the two raters from the electronic data and segmented from the shorter TE images.

The histogram matching correction was first applied to the baseline short TE images using a randomly chosen scan as reference. For each patient, the follow up scan was then histogram matched to its corresponding corrected baseline scan in order to obtain the best possible match between the serial scans. Assessment of the effectiveness of the correction was performed by identifying and comparing the mean signal intensity of an area of NAWM in the deep left frontal lobe before and after the correction for the baseline and serial scans.

The global thresholding technique was then applied to the reference scan. This involves two stages; (i) application of an automated algorithm that performs cranial extra cerebral tissue extraction using thresholding and a knowledge of the 3-D structure of the brain, and (ii) identification of the intensity threshold that best separates hyper-intense MS lesions from surrounding brain (Wicks et al., 1992). The intensity threshold was set such that most MR
visible lesions were above the threshold and therefore included in the total lesion volume. This same threshold was then applied to the other 15 histogram matched scans.

Finally, the regions included by the global threshold were reviewed and manual editing was performed to delete any segmented regions that did not represent lesion. This involved both complete removal of non-lesion regions, and manual editing to amend regions that contained both lesion and non-lesion areas. After a delay of at least a week, this process was repeated on the baseline scans in order to provide a measure of reproducibility.

4.2.1 Statistics

The inter rater reproducibility for the contour method was assessed by comparison of the first set of measurements of the author and the single set of volumes obtained by the second observer, using CoVs, having confirmed that the size of individual CoVs was independent of the mean lesion volume. The intra-rater reproducibility was similarly described using CoVs for the two sets of measurements of observer one.

In order to assess the agreement between the lesion volumes obtained with global thresholding and contouring methods, the mean of the three contour derived lesion volumes on each scan was compared with the volume obtained using the global threshold (before and after manual editing). Agreement between the techniques was calculated for both baseline lesion volumes and change over time as an ICC (Armitage & Berry, 1994; Streiner & Norman, 1995). The differences between the volumes measured with these techniques were also used to calculate the limits of agreement and bias. In 95% of the measurements, there should be agreement to within mean ±2 standard deviations of the differences (Bland & Altman, 1986).
4.3 Results

4.3.1 Effectiveness of the histogram matching correction

The mean absolute percentage difference in signal intensity of the NAWM area between the reference scan and the other seven baseline studies was 15.2% before histogram matching and only 3.8% after the correction. For the serial scans the mean absolute differences in NAWM intensity between the baseline and follow up scans before and after the correction were 11.9% and 2.1% respectively.

4.3.2 Reproducibility and time consumption of the contour method

The median intra and inter-rater CoVs for measurements on the baseline scans were 2.4% and 7.5% respectively. The median time per person taken to perform the contour method was 44 minutes.

4.3.3 Global threshold volumes before editing

The ICC for agreement between absolute lesion volumes derived by the global thresholding and contouring was only 0.46 and the limits of agreement were wide at $46.6cm^3 \pm 118.9cm^3$. This bias of $46.6cm^3$ reflects the larger volumes obtained with the global threshold and the agreement is clearly poor. The differences between the two techniques are displayed graphically in Figure 4.1. The ICC for the changes in lesion volumes detected with the two techniques was only 0.17 and the limits of agreement were again wide at $25.5cm^3 \pm 88.8 cm^3$. 
4.3.4 Global threshold volumes after manual editing

The median intra-rater CoV for the process of manual editing was 6.2%, with a median time requirement of 48 minutes. The volume of non-lesion regions varied widely between the eight baseline scans, from 9.8 cm$^3$ to 188.5 cm$^3$ (Figures 4.2 and 4.3). For the serial scans, the mean difference between the volume of non-lesion regions between the two time points was 29.5 cm$^3$ with a range of 2.4 cm$^3$ to 127.5 cm$^3$. After manual editing to delete hyper-intense regions not corresponding to a lesion, the global threshold derived volumes were all smaller than those obtained with the contour method (Figure 4.1). Agreement between the techniques was substantially improved by manual editing, with the ICC for the baseline scans of 0.89 and for the change in lesion volume of 0.58.
Figure 4.2 A: Axial MRI image at baseline demonstrating hyper-intense regions included by the global threshold before manual editing. Several lesions have been missed by the threshold and four non-lesion regions have been included.
Figure 4.2B: Follow up scan of same patient. The threshold has successfully incorporated a new lesion. However, fewer non-lesion regions have been included and two lesions that were present on the baseline scan are no longer segmented.
4.4 Discussion

Prior to manual editing, the global threshold derived lesion volumes on the baseline scans were all higher than those identified with the contour method. This was anticipated, since the intensity threshold was chosen to include most MR visible lesions, inevitably leading to the inclusion of other hyper-intense regions, notably non-lesion brain (especially grey matter) and choroid plexus (Filippi et al., 1996e). If the volumes of these non-lesion regions had been similar between patients, and had not varied substantially over time, this bias should not have adversely influenced agreement between the techniques when assessing change in lesion volume. However, this study has demonstrated that the volume of non-lesion regions varies widely, both between patients and over serial scans. Such inconsistency is sufficiently large to mask any real change in lesion volume that may have occurred. It is therefore clear that the global threshold method cannot be applied without the manual deletion of non-lesion regions, even after the histogram.
Three potential factors are likely to have contributed to the substantial variability in the volume of non-lesion regions included by the threshold. First, even the small residual differences in intensity between scans after histogram matching may have been sufficiently large to preclude the application of a single threshold, since the global threshold method is highly sensitive to the choice of threshold (Filippi et al., 1996e). The residual difference in NAWM intensity over serial studies may reflect a number of factors; (i) difficulty in identifying the same area of NAWM over time due to repositioning errors, (ii) genuine biological change over time, and (iii) the imperfect nature of the correction itself. Secondly, repositioning errors on serial scans represent a significant source of measurement error in lesion volume quantification (Plante & Turkstra 1991; Gawne-Cain et al., 1996; Rovaris et al., 1997). Any alteration in partial volume effect due to repositioning may have altered the local signal intensity sufficiently to cause inconsistency in the inclusion of high intensity areas. Finally, motion and flow artefacts may have been more apparent on some scans than others, leading to inconsistency in the global threshold derived volumes over the serial scans.

It is therefore clear that the global threshold method cannot be applied without manual editing, even after correcting for both coil non-uniformity and variation in scanner sensitivity. The process of manual editing to remove non-lesion regions did improve agreement for the lesion volumes obtained with the two techniques. However, even with a threshold set sufficiently low to segment the majority of lesions, the derived lesion volumes were still smaller with the global threshold technique (Figure 4.1). This reflects the fact that the intensity of some MR visible lesions was still below the set threshold. Furthermore, the process of deleting spurious areas of high signal was more time consuming than the contour method, due to the large number of non-
lesion regions that had to be deleted or amended, and it was also less reproducible. With this requirement for manual editing, the technique therefore offers no advantages over the contour method, which is both faster and more reproducible. Future improvements, such as the incorporation of image registration techniques to adjust for any repositioning errors, and the use of smaller slice thickness to reduce volume averaging, may yet reduce variability over serial studies with the global thresholding technique. However, variability of motion and flow artefact will remain a problem, and the inconsistent presence of non-lesion regions over serial studies currently precludes the use of the global threshold technique without manual editing.

In conclusion, the global threshold technique, as applied in this study, was both fully automated and reproducible, and yet it failed to yield meaningful results. Other fully automated lesion segmentation techniques are currently being developed, with the potential for optimal reproducibility and speed. This study highlights the importance of carefully validating new automated techniques for lesion volume quantification prior to their application to treatment trials.
Chapter 5. The Effect of slice thickness on MRI lesion detection and quantification in MS

5.1 Introduction

It has become clear that a substantial proportion of white matter pathology is not accessible to a standard two-dimensional (2D) CSE sequence with 5 mm thick slices (Dousset et al., 1992; Barbosa et al., 1994; et al., Filippi et al., 1995a; Filippi et al., 1995c; Filippi et al., 1996d; Gasperini et al., 1996). This undetected lesion volume might have an important impact in functional terms, and recent efforts have being directed towards increasing the sensitivity of MRI to white matter pathology.

One approach to increasing lesion detection is to increase spatial resolution in the slice selection direction by reducing the slice thickness (Filippi et al., 1995c; Hashemi et al., 1995). Where a lesion occupies only part of a voxel, its contrast relative to background tissue is dependent on both the signal from the lesion and the proportion of the voxel that it occupies (Bradley & Glenn 1987). With a standard slice thickness of 5 mm, small low-contrast lesions occupying only part of a voxel can go undetected. A further effect of this volume averaging is blurring of the apparent border of a lesion if its surface is not along the slice select direction, even if a biologically sharp boundary exists (Filippi et al., 1995c; Tofts et al., 1997). This loss of edge definition leads to difficulty in defining lesion boundaries in a reproducible manner.

A reduction in slice thickness from 5 mm to 3 mm has been shown to increase the MR-visible lesion volume by about 9% with a standard CSE sequence (Filippi et al., 1995c), and it may also increase the precision of volume measurement (Filippi et al., 1996b). However, it has not yet been possible to obtain adequate 2D CSE images down to very thin slices, since the necessary
scan time becomes prohibitive and the poor SNR unacceptable. Three dimensional (3D) MR sequences allow acquisition of thinner slices with acceptable acquisition times and SNR. When such a sequence was applied to the detection of contrast enhancing lesions, detection rates were shown to be 12% higher at a 1 mm slice thickness than at a 3 mm thickness (Filippi et al., 1996c).

An alternative approach to increasing lesion detection is to increase lesion-to-background contrast at a given slice thickness. The 2D fast-FLAIR sequence produces higher contrast and therefore greater conspicuity of cortical and subcortical lesions than a standard CSE sequence (Filippi et al., 1996d; Bastienello et al., 1997; Gawne-Cain et al., 1998), albeit with a reduced sensitivity to lesions in the posterior fossa (Gawne-Cain et al., 1997). It may also yield a higher overall brain lesion volume than a standard CSE sequence at the same 5 mm slice thickness (Filippi et al., 1996d). The reproducibility of lesion volume quantification with fast-FLAIR is as good as (Gawne-Cain et al., 1998) or better than with CSE (Bastienello et al., 1997; Filippi et al., 1998d), probably owing to increased lesion-background contrast and better edge definition.

Initial experimentation with a 3D FSE sequence revealed that this was unsatisfactory because of high signal CSF and increased signal at CSF-tissue interfaces. However, a 2mm 2D fast-FLAIR sequence has enabled detection of more brain lesions than 1.5 mm 3D FSE (Tubridy et al., 1998c). With 2D fast-FLAIR, a slice thickness of less than 2mm is precluded by limitations in terms of SNR and acquisition time. A 3D fast-FLAIR sequence was therefore developed for this study, combining the greater lesion conspicuity of the FLAIR sequence with the higher spatial resolution and SNR per unit time possible with 3D imaging (Barker, MRI In Press). This provided an opportunity to study the impact of increasing resolution down to a slice thickness
of 1 mm on MR derived lesion volumes. In this study 3D fast-FLAIR images were acquired at three slice thicknesses (5, 3 and 1 mm). The aim was to study the impact of decreasing slice thickness on both the MR visible lesion volume and reproducibility of volume measurement.

5.2 Patients and Methods

5.2.1 Patients

Eight patients (four men and four women) with clinically definite MS (Poser et al., 1983) were studied. Their mean age was 44.8 years (range, 31-56 years), mean disease duration was 16 years (range, 8-26 years) and mean EDSS score was 4.9 (range, 1.0-8.5). Three patients had relapsing remitting and five secondary progressive MS. Written informed consent was obtained before entry into the study.

5.2.2 MRI

Patients were imaged during a single visit, in two sessions separated by an interval of 5 minutes. A 1.5-T imager was used to acquire the images with contiguous interleaved slabs/slices in the axial plane. During the first session, an oblique axial dual echo CSE sequence was used with contiguous 5 mm slices and the following parameters: TR 2000ms, TE 34 and 90ms, matrix 256×256, FOV 25cm. The 3D fast-FLAIR images were then obtained with axial slices in decreasing slice thickness order (5 mm, 3 mm, and 1 mm). To allow meaningful comparison between the 3D fast-FLAIR sequences, the following parameters were used at all slice thicknesses: TR 4600, TE 140, TI 1740, echo train length 24, FOV 25cm, matrix size 256×192. For the 5 mm and 3 mm sequences, acquisition time was 12 minutes and 48 slices were acquired. The 1 mm sequence required 18 minutes and produced 144 slices.

Next, the patient was removed from the imager for 5 minutes and then, during a second session,
5.2.3 Image Analysis

The images obtained at each slice thickness for each patient were first assessed in isolation, in randomized order, and with a delay of at least one week between assessing any set of images of the same patient. Lesions were identified and marked on hard copy by the author, in consensus with a second observer, over a three-month period. Prior to this study, 10 normal control subjects had been imaged with the 3D fast-FLAIR sequence using a 1.5 mm slice thickness, in order to identify areas of hyper-intensity that could potentially be confused with lesions in the patient group. The normal findings (Figure 5.1) were:

- Increased signal around the occipital poles of the lateral ventricles
- Small areas of increased signal around the temporal horns
- A thin rim of high signal around the remainder of the third and lateral ventricles
- Areas of high signal at the frontal poles of the lateral ventricles
- Symmetrical areas of increased signal in the posterior limbs of the internal capsules
- Slightly increased signal in the posterior centrum semiovale.

When identifying lesions on the 3D fast-FLAIR images, such “normal” areas of higher signal were excluded unless they were particularly prominent and asymmetrical. The aim of this conservative approach in these regions was to minimise the potential for tissue misclassification. Lesion volumes were quantified by the author using the contour algorithm described in Chapter 3. The time taken to perform the quantitative analysis at each slice thickness was recorded. Only lesions marked on the hard copy were included in the analysis.
Figure 5.1 Normal findings with the 3D fast-FLAIR sequence (4600/136/1740 (TR/TE/TI), echo train length of 24, slice thickness 1.5 mm) from two normal subjects. Note the increased signal around the occipital horns of the lateral ventricles (Figure 5.1 a), the thin rim of high signal around the rest of the ventricle and the areas of increased signal at the frontal poles of the lateral ventricles (Figure 5.1 b).
The 3D fast-FLAIR images of each patient obtained during the first session were then reviewed side by side by the same two observers. The images with 5 mm slice thickness were compared against those with a 3 mm thickness, and all lesions marked on just one image were identified. This same procedure was adopted for comparison of the 3 mm and 1 mm images. The volume of these "missed" lesions was quantified from measurements previously obtained using the contour technique.

Finally, 32 lesions more than 1 cm in diameter were randomly selected to provide an estimate of both contrast to noise ratio (CNR) and contrast ratio (CR) at each slice thickness. The CNR was calculated according to: 
\[
\text{CNR} = \frac{\text{SI}_{\text{lesion}} - \text{SI}_{\text{NAWM}}}{\text{noise}},
\]
where SI represents signal intensity and NAWM indicates an area of normal appearing white matter adjacent to the lesion. Noise was estimated from a region of the field of view not containing tissue and not contaminated by phase-encoding artefacts. The CR was calculated as a measure of lesion conspicuity according to the formula: 
\[
\text{CR} = \frac{\text{SI}_{\text{lesion}} - \text{SI}_{\text{NAWM}}}{\text{SI}_{\text{NAWM}}},
\]

5.2.4 Statistical analysis

The mean of the two lesion volumes at each slice thickness was used to compare the effects of slice thickness, as this approach should reduce the impact of random measurement error. Friedman two way ANOVA was used to assess the significance of differences between lesion volumes between the images. Post hoc comparisons were performed using the Wilcoxon signed ranks test. The same statistical approach was used to assess differences in time taken to quantify lesion volumes at each slice thickness. Scan-rescan reproducibility was assessed as the percentage agreement between the lesion volumes obtained from the first and second sessions.
5.3 Results

The lesion volumes obtained with the different images (Table 5.1 and Figure 5.2) were significantly different ($p < 0.001$). The mean lesion volumes obtained with the 5 mm 3D fast-FLAIR sequence were 22% greater (range, 10 to 48%) than those identified for the corresponding 5 mm CSE sequence ($p = 0.01$). Reduction in the slice thickness from 5 mm to 3 mm resulted in an increase in the derived lesion volumes for all but one patient, with a mean increase of 8.1% (range, -2.9 to 19.2%); this difference was statistically significant ($p < 0.05$). In contrast, reducing the slice thickness further from 3 mm to 1 mm did not yield an additional increase in total lesion volume ($p = 0.9$). The mean volume obtained with the 1 mm 3D fast-FLAIR sequence was on average 31% higher ($p = 0.01$), and with the 3 mm 3D fast-FLAIR sequence 32% higher ($p = 0.01$) than with the standard 5 mm CSE sequence.

<table>
<thead>
<tr>
<th>Sequence/slice thickness</th>
<th>Mean</th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 mm CSE</td>
<td>19.2</td>
<td>17.8</td>
<td>5.5-47.3</td>
</tr>
<tr>
<td>5 mm fast-FLAIR</td>
<td>23.4</td>
<td>21.7</td>
<td>8.1-52.2</td>
</tr>
<tr>
<td>3 mm fast-FLAIR</td>
<td>25.3</td>
<td>22.6</td>
<td>9.7-60.4</td>
</tr>
<tr>
<td>1 mm fast-FLAIR</td>
<td>25.1</td>
<td>22.6</td>
<td>10.1-63.6</td>
</tr>
</tbody>
</table>

Note: For the fast-FLAIR sequences, the above values represent the mean of the measurements obtained at the two imaging sessions.

**Table 5.1** Lesion volume according to sequence and slice thickness
Figure 5.2 The effect of sequence and slice thickness on derived lesion volume
The number of additional lesions identified by progressive reduction in slice thickness. A total of 11 lesions were marked on the 5 mm sequence and not on the 3 mm sequence, and a further 11 lesions were marked at 3 mm and not at 1 mm.

* The number of additional lesions identified by progressive reduction in slice thickness. A total of 11 lesions were marked on the 5 mm sequence and not on the 3 mm sequence, and a further 11 lesions were marked at 3 mm and not at 1 mm.

* The volume contribution of these additional lesions to total lesion volume per patient.

@ The individual lesion volume of these additional lesions.

**Table 5.2** The effects of progressive reduction in slice thickness

<table>
<thead>
<tr>
<th></th>
<th>*Number of extra lesions</th>
<th>#Total volume contribution (cm(^3)) of extra lesions. Mean (range)</th>
<th>@Individual lesion volume of extra lesions (cm(^3)). Mean (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 mm vs 3 mm</td>
<td>170</td>
<td>2.2 (0.6-6.2)</td>
<td>0.13 (0.012-0.31)</td>
</tr>
<tr>
<td>3 mm vs 1 mm</td>
<td>185</td>
<td>0.6 (0.2-1.4)</td>
<td>0.027 (0.006-0.085)</td>
</tr>
</tbody>
</table>

The effects of progressive reduction in slice thickness on lesion detection are shown in Table 5.2. Comparing the 5 mm and 3 mm images side by side, 170 lesions were identified on only the 3 mm images, whereas just 11 were marked exclusively on the 5 mm images. When only lesions seen at both 5 mm and 3 mm slice thickness were included in the lesion volume, thereby excluding the additional contribution of lesions identified only at 3 mm, the mean lesion volumes at 5 mm and 3 mm were 23.0 cm\(^3\) and 23.1 cm\(^3\) respectively (p = 0.7). A similar side by side comparison of the 3 mm and 1 mm images revealed that 185 lesions were seen on only the 1 mm sequence, whereas just 11 lesions were identified exclusively on the 3 mm sequence. Considering only those lesions identified on both 3 mm and 1 mm sequences, lesion volumes were 25.3 cm\(^3\) and 24.7 cm\(^3\) respectively; a difference that approached statistical significance (p = 0.07).
The values for CNR and CR are given in Table 5.3. The 3D fast-FLAIR sequence demonstrated significantly higher CR than the 2D 5 mm CSE sequence at each slice thickness (p < 0.001). There was no significant difference in contrast ratio at the different slice thicknesses for the 3D fast-FLAIR sequence (p = 0.4). However, the CNR became progressively worse as slice thickness was reduced (p < 0.001), being approximately proportional to slice thickness (as would be anticipated).

<table>
<thead>
<tr>
<th></th>
<th>CNR mean (SD)</th>
<th>CR mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 mm CSE</td>
<td>13.8 (2.6)</td>
<td>0.27 (0.05)</td>
</tr>
<tr>
<td>5 mm fast-FLAIR</td>
<td>30.2 (7.5)</td>
<td>0.83 (0.19)</td>
</tr>
<tr>
<td>3 mm fast-FLAIR</td>
<td>20.4 (3.3)</td>
<td>0.86 (0.15)</td>
</tr>
<tr>
<td>1 mm fast-FLAIR</td>
<td>7.4 (3.3)</td>
<td>0.81 (0.15)</td>
</tr>
</tbody>
</table>

Table 5.3 Contrast to noise (CNR) and Contrast ratios (CR) for each sequence

The mean scan-rescan agreement for derived lesion volumes was 91% (range, 87 to 97%) at 5 mm slice thickness, 95% (range, 92 to 98%) at 3 mm and 98% (range, 93 to 100%) at 1 mm. The mean operator times required to perform the quantification on the 3D fast-FLAIR images at 5, 3 and 1 mm respectively were 31 minutes (SD, 9.9), 58 minutes (SD, 16.2) and 154 minutes (SD, 30.9). These differences were significant overall (p < 0.001) and for each decrement in slice thickness (p < 0.05).
Figure 5.3 3D fast-FLAIR images of a patient at slice thickness 5 mm (Fig 5.2 a), 3 mm (Fig 5.2 b) and 1 mm (Fig 5.2 c). The SNR deteriorates as slice thickness is reduced. However, despite this, there is an increase in resolution of periventricular lesions with decreasing slice thickness. Lesions that appear confluent at 5 mm can be seen to be comprised of smaller, discrete areas of high signal at 1 mm.
5.4 Discussion

Several previous studies have addressed the impact of slice thickness on both the sensitivity and precision of MS lesion volume quantification (Filippi et al., 1995c; Hashemi et al., 1995; Filippi et al., 1996b). Progressive reduction in slice thickness from 15 mm to 3 mm has been shown in one study to increase the number of detected lesions and lesion volume (Filippi et al., 1995c). In that study, a gain in derived lesion volume of 9% was produced by reducing the slice thickness from 5 mm to 3 mm, and the study authors postulated a linear relationship between slice thickness and MR-visible lesion volume, estimating that a further reduction in slice thickness might produce up to a 20% increase in MR-detectable lesion volume.

The results of this current study are similar in showing an increase in detected lesion volume of 8% with a reduction in slice thickness from 5 mm to 3 mm. This increase was found to be a consequence of the greater sensitivity of the 3 mm sequence to small, generally low-contrast lesions obscured by volume averaging at a 5 mm slice thickness. Comparing the 3 mm and 1 mm slice thickness images side by side, a further similar increase was found in the number of detectable lesions. However, the mean individual lesion volume of lesions detected by decreasing the slice thickness from 5 to 3 mm was more than four times that observed when by decreasing from 3 to 1 mm (Table 5.2); thus the extra contribution of the small lesions detected at 1 mm to total lesion volume was extremely small. It is therefore not surprising that reducing the slice thickness from 3 mm to 1 mm did not increase the overall lesion volume. Indeed, further analysis of the data shows that when considering only those lesions identified at both 3 mm and 1 mm slice thicknesses, the derived lesion volumes at the 1 mm slice thickness were generally smaller.

Several possible factors may have contributed to these findings. First, the lower signal and contrast-to-noise ratios at the lesser slice thickness may have altered the observer's perception of lesion boundaries. A previous study has suggested that as lesion contrast is reduced, the
reported volume decreases (Tofts et al., 1997). Second, differences in volume averaging may be important. For a round lesion of similar diameter to the slice thickness, volume averaging will tend to cause an overestimation of volume (it is assumed to be a cylinder rather than a sphere), while at smaller thicknesses this effect would diminish. The impact of this will depend on the size distribution of lesions for an individual patient, but this averaging may have been more apparent at the 3 mm slice acquisitions. Third, with the increased spatial resolution at 1 mm, several lesions that appeared as confluent areas of high signal at 3 mm could be seen with a slice thickness of 1 mm to be comprised of several smaller, discrete lesions (Figure 5.3). This effect would result in smaller volumes being measured for such regions at a very small slice thicknesses. Since such confluent regions can make a substantial contribution to the total lesion volume, even if this change in perception of lesion boundary at a 1 mm slice thickness occurred only rarely, it might be sufficient to produce an apparent difference in total measured lesion volume.

It might be concluded from these results that lesion volumes derived at 3 mm and 1 mm are not substantially different, and that therefore no advantage in terms of sensitivity is conferred by reducing the slice thickness beyond 3 mm. However, although total lesion volumes were not different, several small, low contrast lesions were identified only on the thinnest slice acquisitions. Such small lesions might be important in functional terms: it is conceivable that gradual accumulation of a large number of these small lesions in a strategic pathway could result in a slow progression in disability (Wang et al., 1997b). A particular advantage of increased slice resolution may be the ability to identify and localise smaller lesions in pathologically eloquent areas, such as the pyramidal tracts. This may yield stronger correlations between MR measures and functional systems scales. Furthermore, it is conceivable, if unlikely, that new treatments may predominantly alter the development of lesions of a particular size or contrast. If a putative drug were to preferentially inhibit the development of small, low contrast lesions, such a treatment effect might only be seen at high slice resolution.
These results also demonstrate that despite the decreasing CNR as slice thickness decreased, the scan-rescan reproducibility improved. Several sources of error contribute to scan-rescan variability (Barkhof et al., 1997b; Plante & Turkstra 1991), notably differences in the extent of motion and flow artefact, inconsistent patient positioning (Gawne-Cain et al., 1996; Simon et al., 1997) and observer inconsistency in application of the quantitative technique (measure-remeasure variability) (Wicks et al., 1992; Grimaud et al., 1996). Scan-rescan reproducibility is a more meaningful reflection of the variability that might be found in a longitudinal study than assessment of measurement error on a single scan alone. As slice thickness decreases, the impact of suboptimal patient repositioning on lesion volume measurements should become less important, since the reduction in volume averaging should allow more consistent identification of lesion boundaries.

Against the increased detection of small lesions and greater reproducibility achieved by reducing slice thickness, must be weighed the increase in both acquisition time and loss of SNR (Figure 5.3). Furthermore, there is a substantial increase in operator time needed to perform the quantification at very thin slices with the contour technique used in this study. The exact relationship between sensitivity, reproducibility and sample size requirements for treatment trials is not yet known. It may be possible to reduce the number of patients and examinations necessary to show a significant treatment effect with MR imaging by increasing the sensitivity and reproducibility of quantification with very thin slices. Therefore, the increasing analysis time with thin slice thickness acquisition could be at least partially offset by the requirement for fewer patients. This issue will only be resolved when the interaction between measurement error and sample size is better defined. However, the contour technique used in this study required a mean analysis time of more than 150 minutes per MRI scan at a slice thickness of 1 mm. For serial studies involving large numbers of patients, this level of analysis time will almost certainly preclude the routine incorporation of such a protocol. The much more feasible analysis times for
3 mm thick slices suggest that this slice resolution is more appropriate for studies where a semi-automated segmentation technique such as contouring is used. However, several automated techniques are now being developed with the potential for minimal operator intervention. Once such methods are fully validated, they offer the potential for application to 1 mm slice thickness images, with the benefits of better sensitivity and less susceptibility to repositioning errors.

In summary, this study shows that reducing slice thickness increases detection of small lesions and increases reproducibility, at the expense of increasing operator time. Furthermore, the relationship between slice thickness and MS lesion volume is not linear at very small slice thicknesses. More work is needed to define the impact of these improvements on both clinical/MRI correlations and sample size requirements for MS treatment trials. If the very small gains in sensitivity and precision do not achieve these goals, the continuation of 3 mm slice thickness as the gold standard will be confirmed.
Part 2. An assessment of the utility of $T_2$ lesion volume quantification as a surrogate outcome

Chapter 6. The relationship between enhancing lesion activity and annual change in brain $T_2$ lesion volume

6.1 Introduction

Recent consensus guidelines have defined the current role of MRI in exploratory and definitive MS treatment trials (Miller et al., 1996). Exploratory (phase II) screening trials utilise monthly enhanced imaging, with disease activity assessed by gadolinium enhancement serving as the primary outcome measure. Such a monthly enhanced imaging protocol undoubtedly provides a sensitive marker of disease activity, with 5-10 new lesions identified for every clinically expressed relapse (Grossman et al., 1986; Willoughby et al., 1989; Bastienello et al., 1990; Kermode et al., 1990a; Miller et al., 1993; Smith et al., 1993). However, the time expenditure and costs of such a protocol render it difficult to apply to all patients in the context of large cohort treatment trials. Definitive (phase III) studies rely on clinical assessment as the primary outcome, with annual change in $T_2$ lesion volume serving as a secondary outcome measure.

The relationship between disease activity demonstrated on monthly enhanced imaging and an annual $T_2$ weighted imaging protocol has not previously been defined. Were such a relationship to be identified, this would confirm that changes in $T_2$ lesion volume are due, at least in part, to ongoing blood brain barrier breakdown and inflammation. This would support the use of annual lesion volume assessment in treatment trials. A study was therefore performed to establish the extent of such a relationship in a cohort of MS patients. In view of the possibility that immunomodulatory treatment might have a differential effect on the two MRI parameters,
patients on such treatment were excluded from this study. Furthermore, the dynamics of disease activity in primary progressive MS are fundamentally different from other MS subgroups (Thompson et al., 1991; Thompson et al., 1997), and this study was therefore restricted to patients with relapsing remitting and secondary progressive MS.

6.2 Methods

6.2.1 Patients

Seventy-three patients (52 women and 21 men) with clinically definite MS according to the Poser criteria were selected from five European Centres (12 from Amsterdam, 36 from London, 6 from Milan, 9 from Munich and 10 from Rome). The cohort consisted of 46 patients with relapsing remitting MS, as defined by a history of relapses and remissions without gradual deterioration, and 27 patients with secondary progressive MS, as defined by an initial relapsing and remitting course with subsequent progressive deterioration for at least 6 months, with or without superimposed relapses. The median age at entry was 33 years (range, 15 to 61 years) and median disease duration was 5 years (range, 1 to 28 years). Patients were either involved in natural history studies (45 patients) or formed the placebo arms of treatment trials (28 patients). To be included, patients had to have had serial monthly gadolinium enhanced T1 weighted images for at least 9 months with a T2 weighted scan coinciding with the first and last enhanced scan (study entry and exit). A full neurological assessment was performed at study entry and exit in all cases, with disability quantified using the EDSS. Details of any relapses over the study duration were also available. Patients taking immunosuppressive drugs other than infrequent short courses of corticosteroids during relapses were excluded, and no patient with relapsing remitting MS entered the secondary progressive phase during the study.

6.2.2 MRI

Serial T2 weighted CSE or FSE images were acquired at study entry and exit. In addition, T1 weighted imaging 5-15 minutes after injection of gadolinium was performed monthly throughout
the study period. Conventional dose gadolinium (0.1 mmol/kg) was given in 64 of 73 patients, the other nine patients (Munich cohort) received 0.2 mmol/kg. Gadolinium enhanced images were not acquired within one week of corticosteroid treatment (to exclude the possibility that corticosteroids might have a transient suppressive effect on enhancement). In London, all MRI scans were acquired on a Signa GE 1.5 Tesla scanner with either CSE (18 patients, SE 2000/34 at entry and exit, SE 640/14 for enhanced scans, 5 mm contiguous axial slices) or FSE (18 patients, SE 3500/18 at entry and exit, SE 579/19 or 570/13 for enhanced scans, 4 mm contiguous axial slices). In Amsterdam, CSE images were obtained on a 0.6 Tesla Technicare (Teslacon II) machine (SE 2755/60 at entry and exit, SE 400/28 or 450/28 for enhanced scans, 5 mm axial slices with an interslice gap of 1.25 mm). In Milan, a Siemens 1.5 Tesla machine was used to obtain CSE images (SE 2000/50 at entry and exit, SE 768/15 for enhanced scans, 5 mm contiguous axial slices). In Rome, a Toshiba 0.5 Tesla machine was used to obtain CSE images (SE 2500/30 at entry and exit, SE 400/18 for enhanced scans, 5 mm axial slices with an interslice gap of 1.0 mm). In Munich, images were obtained on a Magneton scanner at 1.0 Tesla (SE 3000/40 at entry and exit, SE 600/28 for enhanced scans, 5 mm contiguous axial slices). Scanners were not changed or upgraded over the study duration and image acquisition parameters were not modified between entry and exit.

Lesion volume quantification for the 36 London patients was performed by the author. For the other patients, analysis was performed locally by two other observers blinded to the clinical data, using one of two similar semi-automated local intensity-based segmentation techniques, the seed growing method in Amsterdam (van Walderveen et al., 1995), and contouring (see Chapter 3) at all other centres. The number of new and persistently gadolinium enhancing lesions were identified on all monthly scans by experienced observers.

6.2.3 Statistical methods

The relapse rate, change in EDSS, and both absolute and percentage change in lesion volume
over the study duration were all adjusted for the length of study for each patient. Adjusted annual values for all the parameters were then entered into the analysis. The new enhancing and total (new and persistently enhancing) lesion rates were also adjusted for the study duration and for missing data points to provide monthly enhancing lesion rates for each patient (lesions/month/patient). Since most of the clinical and MRI data were not normally distributed, medians rather than means were used to describe the data. Site by site differences between the MRI outcome measures were assessed using the Kruskal-Wallis test; differences in clinical characteristics between the relapsing remitting and secondary progressive subgroups were assessed by means of the Mann Whitney Test. All clinical and MRI correlations were evaluated using the Spearman’s Rank Correlation Coefficient (SRCC). To reflect the large number of statistical comparisons, a significance level of $p \leq 0.01$ was considered significant, a $p$ value between 0.01 and 0.05 as a trend, and values greater than 0.05 were not considered significant.

6.3 Results

6.2.1 Clinical characteristics

Seventy-three patients were followed up for a median duration of 11 months (range, 9 to 14 months). The clinical characteristics over the study duration are given in Table 6.1. There were significant differences between the two clinical subgroups in terms of age, disease duration at entry and EDSS at entry to the study. Relapse rates during the study period were also significantly different between the two subgroups. However, there was no significant difference between the two groups for change in EDSS (exit minus entry) during the study. There was an increase in EDSS during the study duration in 29 of the 73 patients, a decrease in EDSS in 14 and no change in EDSS in 30 patients. Only 19 patients changed EDSS by $\geq$ 1.0 point during the study.
6.3.2 MRI characteristics

All patients had serial monthly enhanced MRI for at least 9 months. A total of 761 scans were acquired and only 19 time points were missed within the continuous series of scans (12 in secondary progressive and 7 in relapsing remitting MS patients). The MRI characteristics are given in Table 6.1. There were no significant site by site differences in terms of baseline lesion volume ($p = 0.08$), change in lesion volume ($p = 0.6$) or enhancing lesion rates ($p = 0.09$). Neither the baseline T$_2$ lesion volumes or change in lesion volume over the study duration were significantly different between the two subgroups. For the group as a whole, the median change in lesion volume was +1.5 cm$^3$ (range: -5.6 to +18.8 cm$^3$), with a median increase of 8.9% (range: -21.6 to +190%). The T$_2$ lesion volume increased in 52 patients (33 in the relapsing remitting and 19 in the secondary progressive subgroup) and decreased in 21 patients (13 in the relapsing remitting and 8 in the secondary progressive MS subgroup). Although the median new and total enhancing lesion rates were substantially higher in the relapsing remitting compared to the secondary progressive MS patients, no statistically significant difference between the two subgroups was found (possibly due to the high inter-patient variability). Sixty-three patients had at least one enhancing lesion over the study duration; in 10 patients (four with relapsing remitting and six with secondary progressive MS) there were no enhancing lesions on any of the serial scans.
<table>
<thead>
<tr>
<th></th>
<th>All patients</th>
<th>RRMS (n=46)</th>
<th>SPMS (n=27)</th>
<th><em>p</em>-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median (range)</td>
<td>Median (range)</td>
<td>Median (range)</td>
<td></td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>33 (15 - 61)</td>
<td>30.5 (15 - 44)</td>
<td>42 (28 - 61)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Disease duration (yrs)</td>
<td>5 (1 - 28)</td>
<td>3 (1 - 12)</td>
<td>10 (2 - 28)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>EDSS at entry</td>
<td>3.5 (0 - 8)</td>
<td>2.5 (0 - 6.5)</td>
<td>5.5 (3.5 - 8)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>EDSS at exit</td>
<td>4 (0 - 8)</td>
<td>2.5 (0 - 6.5)</td>
<td>6.0 (2.5 - 8)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>#Change in EDSS per year</td>
<td>0 (-2.0 +4.0)</td>
<td>0 (-2.0 +4.0)</td>
<td>0 (-1.0 +3.3)</td>
<td>NS</td>
</tr>
<tr>
<td>Yearly Relapse rate</td>
<td>1.3 (0 - 6.7)</td>
<td>2.0 (0 - 5.5)</td>
<td>0 (0 - 6.7)</td>
<td>0.01</td>
</tr>
<tr>
<td>T₂ lesion volume at entry (cm³)</td>
<td>14.7 (1.4 - 97.2)</td>
<td>12.9 (1.4 - 72.8)</td>
<td>23.4 (3.7-97.2)</td>
<td>NS</td>
</tr>
<tr>
<td>#Change in lesion volume (cm³)</td>
<td>+1.5 (-5.6 +18.8)</td>
<td>+1.3 (-4.4 +18.8)</td>
<td>+1.5 (-5.6 +8.9)</td>
<td>NS</td>
</tr>
<tr>
<td>#Percentage change in lesion volume</td>
<td>+8.9 (-21.6 +190)</td>
<td>+10.4 (-19.6 +190)</td>
<td>+8.2 (-21.6 +84.9)</td>
<td>NS</td>
</tr>
<tr>
<td>†New enhancing lesion rate</td>
<td>1.0 (0 - 13.5)</td>
<td>1.2 (0 - 13.5)</td>
<td>0.4 (0 - 11.7)</td>
<td>NS</td>
</tr>
<tr>
<td>‡Total enhancing lesion rate</td>
<td>1.4 (0 - 17.0)</td>
<td>1.6 (0 - 14.5)</td>
<td>0.8 (0 - 17.0)</td>
<td>NS</td>
</tr>
</tbody>
</table>

† New enhancing lesion rate represents number of new enhancing lesions per month.
‡ Total enhancing lesion rate represents the number of new and persistently enhancing lesions per month.
# Annualised change, adjusted for length of study in each patient.

*Mann Whitney test significance levels for differences between the subgroups: Significance level set at p ≤ 0.01; trend, 0.01< p < 0.05; not significant (NS) p ≥ 0.05.

Table 6.1 Clinical and MRI characteristics of patients
6.3.3 Clinical/MRI correlations for the baseline MRI data

There was a significant correlation between EDSS and lesion volume at study entry (SRCC = 0.31, p = 0.007) although this was not statistically significant for the individual subgroups (Table 6.2). Furthermore, the baseline lesion volume was also predictive for the change in EDSS over the study duration in the relapsing remitting MS group (SRCC = 0.44, p = 0.002) but not for the secondary progressive MS subgroup (SRCC = 0.10, p = 0.6). The number of enhancing lesions at study entry was predictive of the subsequent relapse rate over the study duration for the group as a whole (SRCC = 0.47, p < 0.001) and for the secondary progressive MS group (SRCC = 0.56, p = 0.002), but this was only a trend in the relapsing remitting MS subgroup (SRCC = 0.35, p = 0.02).

6.3.4 Clinical/MRI correlations for the longitudinal MRI data

Significant correlations (Table 6.2; Figures 6.1 and 6.2) were identified between the relapse rate and both new and total enhancing lesion rate over the study duration for the group as a whole (SRCC = 0.52, p < 0.001 and SRCC = 0.51, p < 0.001 respectively); this relationship was particularly strong in the secondary progressive MS subgroup (SRCC = 0.73, p < 0.001 for both new and total enhancing lesion rate). No significant correlation was demonstrated between new enhancing lesion rate and change in EDSS. There was also no significant correlation between change in lesion volume over the study duration and relapse rate. Furthermore, no significant correlation was found in either subgroup between absolute or percentage change in lesion volume and change in EDSS. Comparing the change in EDSS between the 52 patients with an increase in lesion volume and the 21 whose lesion volumes decreased, no significant difference was found.
<table>
<thead>
<tr>
<th></th>
<th>All patients SRCC (<em>p-value</em>)</th>
<th>RRMS patients SRCC (<em>p-value</em>)</th>
<th>SPMS patients SRCC (<em>p-value</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDSS at entry vs baseline lesion volume (cm$^3$)</td>
<td>0.31 (0.007)</td>
<td>0.34 (0.02)</td>
<td>0.17 (NS)</td>
</tr>
<tr>
<td>Change in EDSS vs baseline lesion volume</td>
<td>0.30 (0.01)</td>
<td>0.44 (0.002)</td>
<td>0.10 (NS)</td>
</tr>
<tr>
<td>Annual relapse rate vs number of enhancing lesions at entry</td>
<td>0.47 (&lt; 0.001)</td>
<td>0.35 (0.02)</td>
<td>0.56 (0.002)</td>
</tr>
<tr>
<td>EDSS change vs change in lesion volume (cm$^3$)</td>
<td>0.09 (NS)</td>
<td>0.28 (NS)</td>
<td>-0.15 (NS)</td>
</tr>
<tr>
<td>EDSS change vs percentage change in lesion volume</td>
<td>0.09 (NS)</td>
<td>0.18 (NS)</td>
<td>-0.14 (NS)</td>
</tr>
<tr>
<td>Annual relapse rate vs change in lesion volume (cm$^3$)</td>
<td>0.10 (NS)</td>
<td>0.06 (NS)</td>
<td>0.21 (NS)</td>
</tr>
<tr>
<td>Annual relapse rate vs new enhancing lesion rate</td>
<td>0.52 (&lt; 0.001)</td>
<td>0.33 (0.02)</td>
<td>0.73 (&lt; 0.001)</td>
</tr>
<tr>
<td>Annual relapse rate vs total enhancing lesion rate</td>
<td>0.51 (&lt; 0.001)</td>
<td>0.32 (0.03)</td>
<td>0.73 (&lt; 0.001)</td>
</tr>
<tr>
<td>EDSS change vs new enhancing lesion rate</td>
<td>0.15 (NS)</td>
<td>0.15 (NS)</td>
<td>0.26 (NS)</td>
</tr>
</tbody>
</table>

*Significance levels: Significance level set at $p \leq 0.01$; trend, $0.01 < p < 0.05$; not significant (NS)

$p \geq 0.05$.

**Table 6.2** Cross sectional and longitudinal clinical/MRI correlations
Figure 6.1 Scatterplot of relapse rate vs new enhancing MRI activity for RRMS

Figure 6.2 Scatterplot of relapse rate vs new enhancing MRI activity for SPMS
6.3.5 Correlations between the MRI parameters

There was a strong relationship (see Table 6.3; Figures 6.3 and 6.4) between both new and total enhancing lesion rate over the study duration and change in lesion volume for the data as a whole (SRCC = 0.53, p < 0.001 and SRCC = 0.50, p < 0.001 respectively); this was particularly apparent in the relapsing remitting MS subgroup (SRCC = 0.60, p < 0.001 and SRCC = 0.58, p < 0.001 respectively). However, the number of enhancing lesions at study entry did not predict subsequent change in lesion volume, and only modest correlations were found between enhancing lesion activity in the first three months of the study and change in lesion volume over the study duration. In contrast, enhancing lesion activity in the last three months and at study exit alone correlated substantially better with change in lesion volume, to a level similar to values for enhancing lesion rates over the entire study duration (Table 6.3).

For the 52 patients whose lesion volume increased and 21 patients whose lesion volume decreased, the median new lesion rates were, respectively, 1.9 and 0.3 lesions/month/patient and this difference was significant (p = 0.005).

The number of enhancing lesions at study entry was predictive of subsequent enhancing lesion activity (SRCC = 0.60, p < 0.001); this relationship was particularly apparent in the secondary progressive MS subgroup (SRCC = 0.78, p < 0.001).
<table>
<thead>
<tr>
<th></th>
<th>All patients SRCC (*p-value)</th>
<th>RRMS patients SRCC (*p-value)</th>
<th>SPMS patients SRCC (*p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. enhancing lesions at entry vs new enhancing lesion rate over study duration</td>
<td>0.60 (&lt; 0.001)</td>
<td>0.46 (0.002)</td>
<td>0.78 (&lt; 0.001)</td>
</tr>
<tr>
<td>Baseline lesion volume vs change in lesion volume (cm³)</td>
<td>0.17 (NS)</td>
<td>0.19 (NS)</td>
<td>0.17 (NS)</td>
</tr>
<tr>
<td>No. enhancing lesions at entry vs change in lesion volume (cm³)</td>
<td>0.20 (NS)</td>
<td>0.19 (NS)</td>
<td>0.23 (NS)</td>
</tr>
<tr>
<td>New enhancing lesion rate in first 3 months of study vs change in lesion volume (cm³)</td>
<td>0.37 (0.001)</td>
<td>0.43 (0.003)</td>
<td>0.28 (NS)</td>
</tr>
<tr>
<td>New enhancing lesion rate in last 3 months of study vs change in lesion volume (cm³)</td>
<td>0.54 (&lt; 0.001)</td>
<td>0.54 (0.001)</td>
<td>0.53 (0.001)</td>
</tr>
<tr>
<td>New enhancing lesion rate at study exit vs change in lesion volume (cm³)</td>
<td>0.53 (&lt; 0.001)</td>
<td>0.48 (0.001)</td>
<td>0.64 (&lt; 0.001)</td>
</tr>
<tr>
<td>New enhancing lesion rate over entire study duration vs change in lesion volume (cm³)</td>
<td>0.53 (&lt; 0.001)</td>
<td>0.60 (&lt; 0.001)</td>
<td>0.46 (0.02)</td>
</tr>
<tr>
<td>Total enhancing lesion rate over entire study duration vs change in lesion volume (cm³)</td>
<td>0.50 (&lt; 0.001)</td>
<td>0.58 (&lt; 0.001)</td>
<td>0.44 (0.02)</td>
</tr>
</tbody>
</table>

*Significance levels: Significance level set at p ≤ 0.01; trend, 0.01 < p < 0.05; not significant (NS) p ≥ 0.05.

**Table 6.3 Correlations between the MRI parameters**
Figure 6.3  Scatterplot of change in lesion volume (cm$^3$) vs new enhancing lesion rate in RRMS

Figure 6.4  Scatterplot of change in lesion volume (cm$^3$) vs new enhancing lesion rate in SPMS
6.4 Discussion

This study demonstrates the existence of a moderate relationship between enhancing lesion rate and change in $T_2$ lesion volume in patients with relapsing remitting and secondary progressive MS. While several previous studies have highlighted the sensitivity of these MRI parameters to disease activity and evolution, previous work has not directly compared enhancing lesion rates with changes in annual lesion volume. These results provide evidence that serial lesion volume quantification is sensitive to gadolinium enhancing disease activity over relatively short periods. This supports the current practice of performing an annual lesion volume quantification protocol as a surrogate marker of disease progression in definitive treatment trials (Miller et al., 1996; Filippi et al., 1998a) and suggests that lesion volume is strongly influenced by the extent of ongoing inflammatory lesion activity.

The degree of this correlation might be lessened by several factors. First, a proportion of newly enhancing lesions are not apparent on $T_2$ weighted images (Kermode et al., 1990b; Miller et al., 1993), and therefore will not be reflected as a change in lesion volume. Such lesions may reflect the reactivation of established confluent lesions, particularly in periventricular areas, where any subtle change in appearance on $T_2$ weighted images is obscured in large regions of abnormal signal (Harris et al., 1991; Capra et al., 1992; Miller et al., 1993). The larger lesion volumes in the secondary progressive MS patients in this study might explain why the correlations were weaker in this subgroup.

Secondly, in studies at monthly intervals, a proportion of new or enlarging $T_2$ lesions are not visible on enhanced $T_1$ weighted images with conventional doses of gadolinium (Thompson et al., 1991; Thompson et al., 1992; Miller et al., 1993). Therefore, $T_2$ weighted lesion volume can potentially increase in the absence of gadolinium enhancement. Recently described strategies to
optimise the sensitivity to gadolinium enhancement (Filippi et al., 1996a; Lai et al., 1996; Silver et al., 1997) may further diminish the number of new $T_2$ lesions not demonstrating enhancement, thereby improving the strength of this relationship. The possibility that some new $T_2$ lesions form without any phase of enhancement also cannot be excluded (Lee et al., 1998).

Thirdly, measurement errors in both identification of enhancing lesions (Barkhof et al., 1997c) and measuring serial lesion volume are likely to have attenuated the strength of the observed relationship (Fleiss 1985). Even among experienced observers, there are inevitable differences in interpretation of gadolinium enhanced scans, and in this study several observers were involved in lesion identification. Furthermore, there are several well-recognised sources of measurement error in serial lesion volume quantification in both image acquisition (Filippi et al., 1997c, Gawne-Cain et al., 1996) and post-processing. Nevertheless, lesion volume quantification, as performed in this study, was still responsive to disease activity determined by gadolinium enhancement on the monthly scans, supporting the use of such segmentation techniques in longitudinal studies.

Finally, the natural history of evolving $T_2$ lesions is also likely to be important. Initially, the area of increased signal in a new $T_2$ lesion tends to be larger than the area of gadolinium enhancement on the corresponding $T_1$ weighted image (Kermode et al., 1990b; Harris et al., 1991). The larger area of $T_2$ lesions early in their evolution is likely to reflect increased signal due to oedema extending beyond the region of BBB breakdown (Isaac et al., 1989; Willoughby et al., 1989). Subsequent $T_2$ scans typically show decreasing lesion size over a period of months as oedema resolves (Isaac et al., 1988; Kermode et al., 1990b; Harris et al., 1991). A smaller $T_2$ abnormality generally persists, although MRI lesions may even disappear (Willoughby et al., 1989; McDonald et al., 1992). The change in lesion volume between two time points is therefore a
reflection of a dynamic process of initial lesion enlargement, and subsequent shrinkage or disappearance, occurring in multiple white matter foci. Marked month to month fluctuations in lesion volume have indeed been shown in a study where serial imaging was performed in patients with relapsing remitting MS (Stone et al., 1995b). The amount of gadolinium enhancement in that study predicted a simultaneous increase in lesion volume, suggesting that transient increases in $T_2$ lesion volume reflect periods of increased BBB breakdown. An important consequence of this process is that the change in lesion volume on annual MRI may be influenced more by disease activity shortly before the exit scan than earlier activity. This concept is supported by the results of the present study. The correlations between enhancing lesion rate detected in the last three months of the study and change in lesion volume were almost as good as that incorporating enhancing lesion activity over the entire study duration. This suggests that changes in lesion volume on annual $T_2$ weighted images may predominantly be a reflection of activity shortly before the exit scan.

The implication of such a finding for definitive treatment trials using annual lesion volume assessment alone is that a non-sustained treatment effect on disease activity determined by gadolinium enhancement may not be detected on annual MRI. This supports the use of monthly enhanced imaging in at least a subgroup of patients for six months after treatment has been initiated (Miller et al., 1996). The waxing and waning of new $T_2$ lesions may also explain why lesion volumes decreased in a few patients despite ongoing disease activity, as measured on the monthly enhanced scans (Figures 6.3 and 6.4). If these patients had developed a number of new lesions immediately before study entry, their initial $T_2$ lesion volumes may have been heavily influenced by such lesions and a subsequent decrease in lesion volume could be a reflection of the evolution of these lesions. It is therefore entirely possible that lesion volumes could decrease between two time points, despite ongoing activity.
Several other important clinical and MRI correlations were revealed in this study. The number of enhancing lesions at study entry predicted subsequent relapse rate and also significantly correlated with subsequent activity on serial monthly imaging. Therefore, MRI activity at a single time point can predict both future MRI and clinical activity. This study also confirms previous work (Grossman et al., 1986; Willoughby et al., 1989; Bastienello et al., 1990; Kermode et al., 1990a; Miller et al., 1993; Smith et al., 1993) demonstrating a significant longitudinal correlation between clinical exacerbations and enhancing lesion rates. This correlation of MRI activity with relapses provides support for short term MRI as a clinically relevant surrogate marker in screening new treatments.

A modest correlation between lesion volume at study entry and baseline EDSS has been shown, and baseline lesion volume was also predictive of subsequent change in EDSS in the relapsing remitting MS subgroup. However, no significant longitudinal correlations were found between progression of disability and either change in lesion volume or monthly enhancing activity. Therefore, while relapse rate was predicted by MRI activity at a single time point, neither of the MRI parameters used in this study were predictive of progression of disability. This is perhaps not surprising, given the relatively small numbers of patients, sample heterogeneity and short duration of follow up. Several smaller studies with a longer period of follow up have demonstrated significant longitudinal correlations between MRI activity and progression of disability (Table 1.9), albeit of a modest degree.

In conclusion, these results confirm that lesion volume changes in patients with relapsing remitting and secondary progressive MS are correlated with the rate of appearance of new enhancing lesions on monthly enhanced MRI. The existence of this relationship suggests that accumulation of abnormal signal measured on T₂ scans provides a measure of disease activity
over time. Yearly MRI provides only a snapshot of a dynamic process of lesion formation and evolution; this snapshot may be more weighted towards activity near the end of the interval between serial scans. This suggests that for definitive treatment trials, monthly enhanced imaging immediately after initiation of therapy may be advisable in a cohort of patients, to reduce the potential for a non-sustained treatment effect going undetected.
Chapter 7. The use of MRI in MS treatment trials: power calculations for annual lesion volume measurement

7.1 Introduction

In definitive phase (III) treatment trials, change in clinical status remains the accepted primary outcome measure (Whitaker et al., 1995; Rudick et al., 1996a). Consequently, sample size estimations for phase III studies are appropriately based on definitive clinical endpoints such as progression of disability and relapse rate. However, these measures are not without limitations. Potential difficulties exist in both defining what constitutes a relapse and quantifying its severity. Furthermore, the relationship between clinical relapse and long term outcome is less than clear. Several criticisms of the EDSS have been highlighted earlier. These limitations, together with the variable and highly unpredictable natural history of MS, dictate that therapeutic trials based on clinical endpoints generally require hundreds of patients and long study durations to ensure sufficient power. Large cohort multicentre studies over several years' duration are required to ensure sufficient power to detect treatment effect using clinical outcome measures.

MRI outcomes are also almost universally used in phase III trials, providing information that supplements and extends the clinical data. The accepted strategy to date has been to apply an MRI protocol at annual intervals in order to measure T_2 lesion volume changes across the entire study population, without making specific power calculations for this outcome (Paty et al., 1993; Jacobs et al., 1995; Polman et al., 1995; Miller et al., 1996; Simon et al., 1998; Filippi et al., 1998a, PRISMS Study Group 1998). However, the costs and logistic demands of MRI data acquisition, transfer to a coordinating centre and subsequent analysis with such a protocol are substantial. Image acquisition across multiple centres demands a consistently high image quality.
from all involved sites, standardisation of imaging characteristics between sites and rigorous quality control (Filippi et al., 1998a). Furthermore, until fully automated techniques for measuring brain lesion volume across serial studies are comprehensively validated, human intervention in the process of image analysis will be required and this can be extremely time consuming when applied to such large datasets. For these reasons, any strategy to measure the impact of treatment on lesion volume should be as efficient as possible.

Several studies have published power calculations for exploratory phase II treatment trials in which frequent gadolinium enhanced MRI provides the primary outcome measure (Nauta et al., 1994; Truyen et al., 1997; Tubridy et al., 1998b). However, it is less clear how many patients are required to provide sufficient power to detect treatment effect on MRI lesion volume accumulation in definitive phase III treatment trials. This information would promote efficient design of imaging protocols in such studies, preventing them from being unnecessarily overpowered, while at the same time avoiding Type I errors (inability to detect efficacy due to insufficient sample size). The aim of this study was to estimate the sample sizes needed to detect an effect of treatment on T2 lesion volume accumulation, in order to plan an optimal imaging strategy in definitive MS treatment trials.

7.2 Methods

7.2.1 Patients

The power calculations were based on analysis of a database of MS patients fulfilling Poser Criteria for clinically definite MS (Poser et al., 1983). The cohort comprised 128 patients (37 men and 91 women) from six European Centres: Amsterdam, Basel, London, Milan, Munich, and Rome. All of the patients were either involved in natural history studies or formed the placebo arms of treatment trials. A number of these patients were recruited from the database
described in Chapter 6. The median study duration was 12 months (range 11-13 months). Patients taking immunosuppressive drugs other than infrequent courses of corticosteroids for relapses were excluded from the study. There were 63 patients fulfilling criteria for relapsing remitting MS, and a further 65 had secondary progressive MS. Full neurological history and examination were performed at study entry and exit, with disability assessment using the EDSS.

7.2.2 MRI

T₂ weighted CSE or FSE sequences were acquired at study entry and exit, with an inter-scan interval of 12 months ± 30 days. The imaging parameters of the patients from London, Amsterdam, Milan, Rome and Munich are described in Chapter 6. For the patients imaged in Basel, several scanners were used with field strengths ranging from 0.5 to 1.5 T (SE 2500/40, 5 mm contiguous slices). For individual patients, scanners were not changed or upgraded over the study duration and image acquisition characteristics were not modified between entry and exit.

Lesion volume quantification was performed by several observers using either of two similar semi-automated local intensity-based segmentation techniques, the seed growing method in Amsterdam (van Walderveen et al., 1995) and contouring at all other centres. All observers were blinded to the clinical status of patients.

7.2.3 Statistical Methods

First, the basic data set was explored: analysis of covariance was used to identify any interactions between clinical parameters and entry and exit T₂ lesion volume. This model revealed differences in lesion volume between the relapsing remitting and secondary progressive MS subgroups, indicating that separate power calculations were necessary for both patient cohorts.
This basic data set was then used to estimate means and variances between subsequent lesion volumes. The power calculations were performed based on the MRI dynamics of the basic data set, under the assumption that the cohort typified a placebo treated control group.

Note that the basic dataset provided lesion volumes for only entry and year one (exit). Values for subsequent years (years two and three) were extrapolated for each patient by assuming a linear increase in lesion volume over time. The basic estimates were used to construct variance-covariance matrices for longer follow up durations in such a manner that a linear increase in variance was assumed and compound symmetry (equal off-diagonal correlations) was obeyed.

All calculations were based on a parallel group design, with equal numbers of patients in placebo and treatment arms. The effect of treatment on the distribution of the response variable is not known, and therefore it was assumed that the variance in lesion volume changes and correlations between timepoints would be the same in the placebo and treatment groups. A further assumption was that treatment effect is sustained and unchanged throughout the study duration.

Sample sizes for each given study duration were calculated for a 100% reduction in the rate of increase in MRI lesion volume (ie complete stabilisation of lesion volume in the treated cohort). Further sample size estimations were based on less marked effect sizes down to 30% reductions in lesion volume increase over the various study durations.

Finally, the effect of adding 'noise' to the data was investigated by artificially increasing the between-patient variance in lesion volumes for each study time point. Progressive percentage increments in the variance were entered into the power calculations to simulate increased levels
of measurement error.

For all the sample size estimations, a power of 0.80 and two-sided significance level of 5% was assumed. The STAT-POWER package (Bavry JL, Statistical design analysis system, Scientific Software, 1984) was used to perform the power calculations.

7.3 Results

The clinical and MRI characteristics of the patients are given in Table 7.1. There was a wide range of disease duration and disability between patients, but clinical characteristics were typical of both disease subgroups. As expected, there were significant differences between the relapsing remitting MS and secondary progressive MS subgroups in terms of age, disease duration, and EDSS at study entry (Mann Whitney U test, p < 0.001 for all three clinical variables).

The analysis of covariance model revealed that age did not influence lesion volume ($F_{(1,112)} = 0.317, p = 0.6$) or sex ($F_{(1,111)} = 0.04, p = 0.8$). However, there was a significant difference in lesion volume between the two MS subgroups ($F_{(1,122)} = 16.8, p < 0.001$) and a significant increase in lesion volume across all 128 patients, ($F_{(1,126)} = 19.31, p < 0.001$). There was no significant difference between the MS subgroups for the absolute change in lesion volume (exit lesion volume minus entry lesion volume) over the study duration ($F_{(1,113)} = 0.15, p = 0.7$).

The median percentage changes in lesion volume in the relapsing remitting and secondary progressive subgroups were 12.5% and 8.8% respectively. There was a strong correlation between lesion volume at entry and exit for both the relapsing remitting (Pearson correlation coefficient, $r = 0.96$) and secondary progressive ($r = 0.93$) MS subgroups. The corresponding variance-covariance matrices used in the power estimations are given in Table 7.2.
The results of the power calculations for a parallel group, placebo-controlled design with change in lesion volume as response variable are given in Table 7.3 and Figures 7.1 and 7.2. Sample sizes are presented for different treatment efficacies and durations of follow up. The sample sizes per arm show substantial differences between the two MS subgroups for all three study durations. First, the number of patients per arm required for a given treatment effect are more than double for secondary progressive than relapsing remitting patients. Secondly, there is a very marked decrease in sample size with increasing follow up, especially with progression from one to two years of follow up, but also moving to a three year follow up period. Thirdly, while it is relatively easy to show complete stabilisation of lesion volume even with only one year of follow up, considerably larger sample sizes and duration of follow up are needed for efficacies of 50% and less.

The impact on sample size requirements of artificially adding noise to the data is shown in Table 7.4 and Figures 7.3 and 7.4. Added noise significantly increases the sample sizes. The effect of added noise is slightly less pronounced for secondary progressive than relapsing remitting patients. For both groups, the increase becomes less marked with larger periods of follow up.
## Table 7.1 Clinical and MRI characteristics of the patient sample

<table>
<thead>
<tr>
<th></th>
<th>All patients (n = 128)</th>
<th>RRMS (n = 63)</th>
<th>SPMS (n = 65)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yrs)</td>
<td>Median: 33, Mean: 34.8, SD: 9.1</td>
<td>Median: 30, Mean: 30.6, SD: 6.4</td>
<td>Median: 37, Mean: 38.8, SD: 9.5</td>
</tr>
<tr>
<td>Disease duration (yrs)</td>
<td>Median: 5, Mean: 6.8, SD: 5.5</td>
<td>Median: 3, Mean: 4.2, SD: 3.0</td>
<td>Median: 8, Mean: 9.7, SD: 6.3</td>
</tr>
<tr>
<td>EDSS at entry</td>
<td>Median: 3.5, Mean: 3.8, SD: 1.9</td>
<td>Median: 2, Mean: 2.3, SD: 1.3</td>
<td>Median: 6, Mean: 5.2, SD: 1.3</td>
</tr>
<tr>
<td>EDSS at exit</td>
<td>Median: 3.5, Mean: 3.9, SD: 2.3</td>
<td>Median: 2, Mean: 2.3, SD: 1.4</td>
<td>Median: 6, Mean: 5.7, SD: 1.6</td>
</tr>
<tr>
<td>Entry T&lt;sub&gt;2&lt;/sub&gt; lesion volume (cm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>Median: 13.6, Mean: 21.6, SD: 19.9</td>
<td>Median: 7.1, Mean: 14.1, SD: 16.7</td>
<td>Median: 26.3, Mean: 29.0, SD: 20.2</td>
</tr>
<tr>
<td>Exit T&lt;sub&gt;2&lt;/sub&gt; lesion volume (cm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>Median: 15.9, Mean: 24.2, SD: 21.2</td>
<td>Median: 9.8, Mean: 16.6, SD: 18.7</td>
<td>Median: 40.0, Mean: 31.6, SD: 21.1</td>
</tr>
<tr>
<td>% change in lesion volume</td>
<td>Median: 10.4, Mean: 22.6, SD: 41.3</td>
<td>Median: 12.5, Mean: 32.2, SD: 51.6</td>
<td>Median: 8.8, Mean: 13.3, SD: 25.0</td>
</tr>
</tbody>
</table>

Note: This table shows the variance-covariance matrices used for the power calculations based on a one year study duration and under the assumption of compound symmetry.

### Table 7.2 Variance-covariance matrices for RR and SP MS subgroups

<table>
<thead>
<tr>
<th></th>
<th>RR MS subgroup</th>
<th>SP MS Subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry Lesion volume</td>
<td>2.78x10&lt;sup&gt;8&lt;/sup&gt;</td>
<td>4.07x10&lt;sup&gt;8&lt;/sup&gt;</td>
</tr>
<tr>
<td>Exit Lesion volume</td>
<td>2.99x10&lt;sup&gt;8&lt;/sup&gt;</td>
<td>3.48x10&lt;sup&gt;8&lt;/sup&gt;</td>
</tr>
<tr>
<td>Entry T&lt;sub&gt;2&lt;/sub&gt; lesion volume</td>
<td>3.96x10&lt;sup&gt;8&lt;/sup&gt;</td>
<td>4.45x10&lt;sup&gt;8&lt;/sup&gt;</td>
</tr>
<tr>
<td>Treatment Effect (%)</td>
<td>After 1 year</td>
<td>After 2 years</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td>RR MS</td>
<td>SP MS</td>
</tr>
<tr>
<td>100</td>
<td>60</td>
<td>137</td>
</tr>
<tr>
<td>90</td>
<td>73</td>
<td>167</td>
</tr>
<tr>
<td>80</td>
<td>92</td>
<td>211</td>
</tr>
<tr>
<td>70</td>
<td>119</td>
<td>274</td>
</tr>
<tr>
<td>60</td>
<td>161</td>
<td>373</td>
</tr>
<tr>
<td>50</td>
<td>232</td>
<td>535</td>
</tr>
<tr>
<td>40</td>
<td>366</td>
<td>834</td>
</tr>
<tr>
<td>30</td>
<td>638</td>
<td>1483</td>
</tr>
</tbody>
</table>

Note: *Treatment effect is based on the percentage reduction in the rate of increase in lesion volume. Hence an effect size of 100% represents complete stabilization of lesion volume (ie no increase in the treatment arm).

Table 7.3 Sample sizes per arm according to disease type, effect size and study duration
Figure 7.1 Sample size per arm for RR subgroup by study duration and effect size (%)

Figure 7.2 Sample size per arm for SP subgroup by study duration and effect size (%)

116
The impact on sample size requirements of artificially adding noise to the data is demonstrated in Table 7.4 and Figures 7.3 and 7.4. Noise addition substantially increased the sample sizes, in a linear fashion. The effect of this added noise is less pronounced for the secondary progressive than the relapsing remitting patients. For both groups, the increase becomes less marked with larger periods of follow up.

<table>
<thead>
<tr>
<th>% Variance Added</th>
<th>Sample size per arm for 100% treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>After 1 year</td>
</tr>
<tr>
<td></td>
<td>RR MS</td>
</tr>
<tr>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>119</td>
</tr>
<tr>
<td>6</td>
<td>149</td>
</tr>
<tr>
<td>8</td>
<td>179</td>
</tr>
<tr>
<td>10</td>
<td>209</td>
</tr>
<tr>
<td>20</td>
<td>358</td>
</tr>
<tr>
<td>40</td>
<td>656</td>
</tr>
</tbody>
</table>

Note: Progressive % increments of between-patient variance were added at each time point to simulate the effect of higher levels of measurement error. These data assume a 100% treatment effect.

Table 7.4 The effect of increasing noise on required sample size
Figure 7.3 The impact of noise addition on sample size, by study duration, for the RR cohort

Figure 7.4 The impact of noise addition on sample size, by study duration, for the SP cohort
7.4 Discussion

The aim of this study was to determine sample size requirements for definitive MS treatment trials, based on an annual assessment of MRI brain lesion volume. Calculations were based on natural history and placebo data derived from several centres, under the assumption that this cohort is representative of the behaviour of the placebo arm of a definitive phase III trial. There are several reasons to support this assumption. First, the clinical characteristics of the patients used in this study are consistent with those of patients who might be entered into a treatment trial. Second, the changes in lesion volume observed in this cohort at 12 months are of a similar magnitude to previously published data (Paty et al., 1993; Zhao et al., 1997; Stone et al., 1995b). Third, the variance estimates used in the power calculations include the multiple sources of 'noise' typical of such studies. The total between patient variance for the change in lesion volume observed in a sample is a composite of true biological differences between patients (sample heterogeneity) and the contribution of measurement error, as discussed in Chapter 4. Site by site differences in derived lesion volumes due to inter-scanner variation (Filippi et al., 1997a), differences in pulse sequence, slice thickness, and repositioning technique (Gawne-Cain et al., 1996; Simon et al., 1997; Filippi et al., 1997c) are all likely to increase observed variance. The database for this study therefore reflects the multiple sources of variation due to measurement error that contribute to 'noise' in serial lesion volume assessment, as well as 'true' biological variance between patients.

While several studies have reported power calculations for phase II exploratory treatment trials, where the primary outcome measure is based on MRI activity, only one study has attempted size estimations based on annual lesion volume quantification (Stone et al., 1995b). Its authors used a bootstrapping procedure on an MRI dataset that comprised the serial estimates of total brain white matter abnormality in seven relapsing remitting MS patients. It was estimated that 353
patients with relapsing remitting MS would need to be scanned to detect stabilisation of MRI lesion volume after 12 months. In contrast, the results of this current study suggest that only 120 patients (60 per arm) would be required to provide sufficient power in detecting this endpoint. This difference probably reflects the fact that the results presented here were based on a different and much larger cohort of patients, but differences in the image analysis techniques for the two studies may also have contributed.

Much recent work has focussed on the importance of reducing measurement error in serial lesion volume quantification. By adding 'noise' as increased variance in lesion volume change, this study has shown that sample sizes are indeed sensitive to the effect of increased measurement error (Table 7.3; Figures 7.3 and 7.4). However, even with addition of 20% extra 'noise', with a three year study duration, a 100% treatment effect should still be detected with fewer than 100 patients per arm for both MS subgroups. The impact of measurement error is probably less of a determinant of sample size than the inherent biological variability in the study population. With a typical core annual imaging protocol for all patients in a definitive MS trial, incorporating several hundred patients per arm, the power to detect MRI treatment effect should be sufficiently high to tolerate substantial added noise without risking a Type I error. Thus, while the drive to improve precision is clearly important, the impact of modest gains in precision on sample size requirements is unlikely to be substantial. These results support the data presented in Chapter 4, where the relative lack of impact of measurement error in image post processing was discussed.

The marked differences in sample size requirements between the relapsing remitting and secondary progressive MS subgroups justify the decision to perform separate sample size estimations for the two disease categories; differences (albeit less marked) have also recently
been shown for exploratory phase II treatment trials that evaluate enhancing lesion activity (Tubridy et al., 1998b). Over twice as many patients are required to detect a given treatment effect on T<sub>2</sub> lesion volume with a population of patients with secondary progressive compared with relapsing remitting MS. This reflects the greater variance between patients for both the baseline lesion volume and the change in lesion volume in this cohort.

One of the most challenging aspects of sample size planning is deciding what represents a clinically important effect. This is a particularly complex issue when powering for a surrogate marker such as T<sub>2</sub> lesion volume accumulation, since any MRI treatment effect would always be reviewed in conjunction with the results of therapeutic impact on the primary clinical endpoints. Furthermore, the magnitude of the effect size has a dramatic impact on the sample sizes required (Table 7.2). Thus the sample sizes needed to adequately power for a 30% reduction in lesion volume accumulation are an order of magnitude higher than that required to detect a 100% reduction (complete stabilisation of lesion volume). It is by no means certain that the extra cost and logistic demands of powering the MRI aspect of a trial for detection of a modest treatment effect are justified. It is probably unnecessary to power for detection less than a 50% reduction in the rate of lesion volume accumulation. Any change less than this is of doubtful clinical relevance, particularly in light of the known disparity between MRI and clinical outcome measures (Paty et al., 1993; IFNB MS Study Group 1995). A caveat is that it is conceivable that a therapy may be effective in reducing the accumulation of disability (for example via remyelination or neuro-protection) without preventing the development of new T<sub>2</sub> abnormalities.

The power calculations for longer study durations of two and three years should be interpreted with caution since these rely on extrapolated values (assuming a linear trend in lesion volume
over time) rather than actual values for change in lesion volume, and also assume a sustained
treatment effect over this time period. However, there is evidence that lesion volume changes
are stable over time, at least over the typical duration of a phase III treatment trial (Paty et al.,
1993; IFNB MS Study Group 1995). Furthermore, the effect of a longer duration of follow up
on study power is dramatic and probably more than can be accounted for by the limitations of
this model alone. This reflects the fact that the baseline and one year lesion volumes are highly
correlated and the fact that at a given time point, the inter patient variability markedly exceeds
intra patient variability. The longer periods of follow up strongly benefit from this discrepancy
using the repeated measures analysis (Frison & Pocock 1992). With a follow up period of three
years, sample sizes of only $2 \times 12$ and $2 \times 33$ are needed to show stabilisation of MRI lesion
volume in the relapsing remitting and secondary progressive MS subgroups respectively.

There are three reasons to suggest that these results may even be conservative estimates. First,
it was assumed that between patient variance for lesion volume change is the same in placebo
and treatment arms. In reality, treatment may reduce such variance, as demonstrated in Chapter
3, and this would have the effect of increasing study power. Second, stratification for MRI
centre may allow a correction for between centre differences, thereby reducing variance and
increasing study power still further. By not correcting for between centre differences, this has
sought to ensure that these sample size estimations are conservative. Third, the patient cohort
may be more heterogenous in terms of MR activity than a clinical trial population, since there
were no MR or clinical entry criteria based on recent activity, and the range of scanner field
strengths used for this study is probably wider than that which might now be used in a Phase III
trial (Filippi et al., 1998a). In a more homogenous population, sample sizes are likely to be even
smaller. These factors are likely to combine to reduce sample sizes still further, although the
relative stability of sample sizes to added noise simulations suggests that any further reduction
is unlikely to be substantial.

The current practice of performing an annual MRI protocol over two or three years on the entire study group therefore provides sufficient power to detect even very small MRI treatment effects of doubtful clinical relevance when a cohort of several hundred patients is involved. Such an approach is clearly overpowered in terms of MRI outcome. It may however provide a useful tool for subgroup analysis in large clinical trials, where this is beyond the limited power of clinical outcome measures. For example, it may be possible to stratify treatment effect according to baseline EDSS and lesion volume, the presence or absence of drug antibodies (Rudick et al., 1998b), or the presence of enhancing lesions at baseline (Simon et al., 1998). Furthermore, the power of this MRI outcome should be sufficiently high to allow planned or unplanned interim analysis.

Alternatively, given the substantial demands of a multicentre MRI protocol, the number of scans obtained during a study could be reduced to a minimum level compatible with ensuring detection of a clinically relevant treatment effect. This implies that the number of centres where MRI data is collected could be restricted to those that can best provide high quality serial images without sacrificing power, as long as the MRI patient subset is clinically representative of the trial population. By concentrating resources at fewer centres, quality control should be better. This approach could allow exclusion of those sites where a major MRI upgrade is planned over the duration of a study, to prevent a potential step change. Furthermore, restriction of sites to those operating at 1.5 tesla will reduce the inter-scanner variation in measured lesion volumes and improve reproducibility (Filippi et al., 1997a). A further advantage of reducing scan numbers is a potential impact on image analysis, since the potential for operator fatigue and drift over time may be reduced. Furthermore, if fewer personnel were required to perform image analysis,
inter-observer differences in measurement reliability would be diminished. Finally, the costs of MR acquisition and analysis would be markedly reduced.

While there is a current consensus that the main MR outcome in definitive treatment trials should be change in lesion volume on T\textsubscript{2} weighted images (Miller \textit{et al.}, 1996; Filippi \textit{et al.}, 1998a), several other MR outcome measures are also being applied to such studies, especially with a view to improving the prediction of disability. Monthly gadolinium enhanced imaging provides a highly sensitive measure of short term disease activity, and its use in a subgroup of patients in long term trials provides valuable data on the degree and persistence of treatment effect on this feature. However, in common with T\textsubscript{2} lesion volume, the predictive value of enhancing lesions for disability appears at best modest, as demonstrated in Chapter 6.

Other MRI techniques offering the prospect of greater pathological specificity have been reviewed earlier. MR tools such as MT imaging, MR spectroscopy, serial measurement of brain and spinal cord atrophy and the quantification of T\textsubscript{1} hypo-intense lesion volume offer the potential to selectively identify the more destructive pathological elements of the disease and predict disability. Such putative markers are now being assessed in a clinical trial context (Polman \textit{et al.}, 1995). In the future, sample size calculations will be required for those techniques that are validated as useful surrogate markers.
Part Three. A multicentre phase III trial of interferon-β-1b in secondary progressive MS

Chapter 8. Background to the Trial

The interferons (IFNs) are a class of naturally occurring proteins produced by a large variety of vertebrate species in response to infections by viruses, bacterial components, natural and synthetic double stranded RNA, mitogens, and antigens in sensitized animals. They interfere with viral replication, and also have both antiproliferative and immunomodulatory effects. Four major serotypes of IFNs, α, β, γ and ω have been described (DeMaeyer & DeMaeyer-Guignard 1988). Interferons α, β, and ω are classified as type I IFNs since, although differing in both amino acid sequence and immunological properties, they all compete for binding to the same receptor and share similar biological properties (DeMaeyer & DeMaeyer-Guignard 1988). Type I IFN genes are located on the short arm of chromosome 9. Interferon-γ is classified as a type II IFN, differing from the type I IFNs in terms of amino acid sequence and cell-surface receptor. While more than 20 human IFN-α subtypes and several IFN-ω subtypes are produced by lymphocytes, only one IFN-β subtype, produced by fibroblasts, has been identified.

The IFNs are part of a large and growing number of cytokines, other examples of which include the interleukins, numerous growth factors and other immunoregulatory products. IFN-α is synthesised by lymphocytes and macrophages in response to exposure to viruses or viral proteins. In contrast, IFN-β is predominantly produced by fibroblasts. Unlike most IFN-α subtypes, IFN-β is a glycoprotein. IFN-γ is produced by activated T cells and natural killer cells, shares no sequence homology with IFN-α or IFN-β, and is localised to a different chromosome.
Both IFN-β and IFN-γ have antiviral properties, mediated through syntheses of several host cell proteins. Furthermore, they both increase expression of Class I major histocompatibility complex (MHC) molecules, thereby enhancing the ability of virus infected cells to present peptides to CD8+ T cells (Hohlfeld 1997).

8.1 Immunoregulation by the interferons

The rationale for the initial tests of IFNs in patients with MS was the belief that the disease might be caused by a persistent or latent viral CNS infection in genetically susceptible individuals, and that IFNs have antiviral properties (Dunnick & Galasso 1979; Cook & Dowling 1980). Three other observations also supported the early IFN clinical trials:

1) Reduced synthesis of IFN-α and IFN-γ by leukocytes in patients with MS had been reported in response to both viral and mitogen challenges (Neighbour et al., 1981; Vervliet et al., 1983; Vervliet et al., 1984). It was hypothesised that exogenous IFN might replace this defective synthesis or stimulate lymphocytes to produce normal amounts of IFN (Neighbour et al., 1981; Neighbour 1984).

2) Both IFN-α and IFN-β influence immunoglobulin (Ig) synthesis via an effect on plasma cells, and might potentially influence IgG synthesis, which is known to be increased within the BBB in MS patients (Tourtellotte et al., 1984).

3) It was thought that the reduced natural killer cell activity seen in some patients with MS could potentially be corrected by IFNs that were known to increase natural killer cell cytotoxic activity (Neighbour et al., 1981; Neighbour 1984; Jacobs & Johnson 1994).
8.2 Early clinical trials

A pilot study to assess the safety of IFN-γ in MS was reported in 1987 (Panitch et al., 1987). Treatment with IFN-γ significantly increased the relapse rate in the 18 patients participating in the study. A dose dependent induction of HLA-DR was observed during treatment, together with increased natural killer cell cytotoxicity. This study provided a powerful rationale for clinical trials of other agents, notably IFN-α and IFN-β, that could inhibit the effects of IFN-γ on the immune system.

The first significant study of IFN-β therapy in MS was performed unblinded in 20 patients and suggested that intrathecal administration of natural IFN-β reduced relapse frequency (Jacobs et al., 1982). This was followed by a larger, blinded, placebo-controlled study on 69 MS patients that demonstrated a significant reduction in relapse rate over a follow up period of 2 years in the group treated with intrathecal IFN-β (Jacobs et al., 1986; Jacobs et al., 1987). However, despite the success of this study, intrathecal administration of IFN-β did not become an accepted treatment modality, due to the need for multiple lumbar punctures and concerns over possible side effects, particularly arachnoiditis. Furthermore, a more recent study using a different natural IFN-β preparation showed a non-significant increase in relapse rate with intrathecal administration (Milanese et al., 1990).

Early work with IFN-α was performed using subcutaneous administration and a crossover design (Knobler et al., 1984) with a 6 month washout period. A non-significant trend for reduced relapse rate on treatment was identified, and treatment was well tolerated. Other early studies included a trial of lymphoblastoid IFN (predominantly IFN-α) in 100 patients with chronic progressive MS (Kastrukoff et al., 1990). This study was double-blind, with 50 patients given placebo and 50 receiving lymphoblastoid IFN subcutaneously. A trend favouring treatment was
seen for both clinical outcomes and $T_2$ lesion volume measurements, but the authors felt that the results were not sufficiently promising to recommend lymphoblastoid IFN treatment in chronic progressive MS.

8.3 Systemic administration of recombinant IFN

The development of recombinant DNA technology in the 1980s allowed the production of highly purified IFNs from bacterial and mammalian cells, facilitating further clinical trials. In the first study of recombinant IFN in MS, 98 patients were enrolled in a placebo-controlled trial of IFN-α-2b, administered at a dose of 2 MU subcutaneously three times per week for one year (Camenga et al., 1986). No therapeutic effect was observed, possibly because the dose used was too low. A further randomised, placebo-controlled study with IFN-α also failed to demonstrate a significant treatment effect (AUSTIMS Research Group. 1989). The most recent study using IFN-α investigated the efficacy of IFN-α-2a given the intramuscular route on alternate days. In contrast to the previous studies, a striking reduction in both clinical and MRI activity was demonstrated with treatment in this small study (Durelli et al., 1994) resuming after stopping treatment (Durelli et al., 1996), although side effects were relatively severe.

IFN-β-1b (betaseron, Berlex Laboratories, Schering AG) was developed as a non glycosolated recombinant IFN-β, produced in a bacterial cell line (Escherichia coli) through insertion of a genetically engineered plasmid. The native gene was obtained from human fibroblasts and altered in a manner that substitutes serine for the cysteine residue at position 17, since this was found to improve stability (IFNB MS Study Group 1993). It also lacks the N-terminal methionine of native human IFN-β. A pilot study was performed on 30 patients with relapsing remitting MS who were randomised to receive various doses of IFN-β-1b or placebo, by subcutaneous injection three times weekly (Johnson et al., 1990). A dose-related reduction in
relapse rate was observed. The optimum dose was identified as 8 MIU, and subsequently all patients were switched to this dose. Ten of 24 (42%) of IFN-β-1b patients, but only one of six (17%) placebo patients remained relapse free over 3 years.

8.4 Mechanism of action of IFN-β

Three large phase III studies have now confirmed that IFN-β can favourably alter disease outcome (IFNB MS Study Group 1993; IFNB MS Study Group 1995; Jacobs et al., 1996; Simon et al., 1998; PRISMS Study Group 1998). These are described below, but is appropriate at this point to review recent work which sheds further light on the potential mechanisms of action underlying this treatment efficacy. These studies suggest that the predominant mechanism of action is immunomodulatory, with alteration of the cytokine profile towards an anti-inflammatory phenotype, together with a reduction in transmigration of T lymphocytes into the CNS (Hohlfield et al., 1997; Yong et al., 1998).

The induction of MHC class II surface molecules on antigen presenting cells (APCs) seems to play a role in MS. It is now recognised that IFN-γ is pro inflammatory, with effects including activation of both monocytes and macrophages, and up-regulation of several adhesion cell molecules that regulate entry of T-cells across the BBB (Yong et al., 1998). It also induces expression of class II MHC molecules on astrocytes, microglia and endothelial cells, which can then present antigens to T cells (Barna et al., 1989).

In contrast, IFN-β inhibits the IFN-γ induced upregulation of of MHC class II expression by human astrocytes (Barna et al., 1989; Jiang et al., 1995). The exact mechanism of IFN-β induced MHC class II inhibition is unclear, but it appears to involve reduction in activity of the class II transactivator CIITA, a factor necessary for IFN-γ induced MHC class II transcription (Lu et al.,
1995). IFN-β is also now recognised to inhibit the proliferation of T lymphocytes and reduce their production of IFN-γ (Noronha et al., 1993). These studies support the hypothesis that a predominant mechanism of action of IFN-β in MS is through inhibition of class II MHC expression, thereby inhibiting activation by IFN-γ (Hohlfeld et al., 1997; Yong et al., 1998).

There are a number of other potential sites of action of IFN-β. One potential mechanism is alteration towards an anti-inflammatory cytokine profile, with reduced production of both TNF-α (Chabot et al., 1997) and IL-12 (Cousens et al., 1997), and increased production of IL-10 (Rep et al., 1996; Rudick et al., 1996b; Rudick et al., 1998a).

IFN-β may also have an effect to prevent migration of leukocytes across the BBB. This is supported by the reduction in gadolinium enhanced lesion activity seen with IFN-β in crossover studies in relapsing remitting MS (Stone et al., 1995a; Pozzilli et al., 1996). Passage of leukocytes into the CNS requires both transmigration across endothelial cells and passage through the basement membrane. There is emerging evidence that IFN-β can down-regulate expression of vascular cell adhesion molecules (VCAMs) such as VCAM-1 (Calabresi et al., 1997), and hinder the adherence of leukocytes to endothelial cells (Corsini et al., 1997). A further effect of IFN-β on maintaining basement membrane integrity has also been proposed. Matrix degrading enzymes such as the matrix metalloproteinases (MMPs), particularly MMP-9, may be important in mediating migration of leukocytes through the basement membrane (Yong et al., 1998). IFN-β has been shown to reduce T-cell migration across basement membrane in vitro, possibly mediated by reduced production of MMP-9 by lymphocytes (Stuve et al., 1996; Leppert et al., 1996; Stuve et al., 1997). This may further contribute to the BBB protective effect of IFN-β.
8.5 The IFNB-1b study in relapsing remitting MS

Based on the findings of the pilot study, a randomised phase III, double-blind, placebo-controlled trial of IFNB β-1b was initiated in 1988. The trial was conducted using 372 patients recruited from seven American and four Canadian centres (IFNB MS Study Group 1993). Entry criteria included a diagnosis of relapsing remitting MS, EDSS scores of 0-5.5 and at least two relapses in the preceding two years. Patients were randomised to receive either placebo, or one of two doses of IFN-β-1b, 1.6 MIU or 8 MIU. The trial was designed to last for 2 years, but patients were given the option to continue for an additional year at the end of years 2, 3 and 4. The primary endpoint was relapse rate, with the proportion of relapse-free patients, time to first, second and third relapse and EDSS assessment providing secondary endpoint measures. In addition, serial brain MRI was performed annually on the entire cohort, with blinded measurement of brain T₂ lesion volume providing an additional endpoint. A subgroup of 52 patients also underwent unenhanced brain MRI every six weeks for 2 years, with the number of new or enlarging lesions on visual assessment providing a further MRI outcome (Paty et al., 1993).

In late 1992, the results were analysed by an external advisory committee. Statistically significant results favouring high dose IFN-β-1b were found for both clinical and MRI outcomes. In the 8 MIU group, the relapse rate was reduced by about a third (0.84 per year versus 1.27 per year in the placebo group, p = 0.0001). There were also significant increases in both the proportion of relapse-free patients (p = 0.007) and delay in the time to first relapse on treatment (p = 0.015). However, no significant therapeutic effect on EDSS was identified. At the end of 3 years, 28% (34/123) of patients in the placebo group had confirmed disability progression, compared with 20% (25/124) of the 8 MIU treated patients, although this difference did not reach statistical significance. The only side effects significantly associated with treatment were
injection site reactions, fever, chills, myalgia, sweating and malaise.

The 3 year results of the annual MRI lesion volume quantification are given in Table 8.1. The MRI data were analysed for 327 of the 372 patients enrolled in the study. Highly significant treatment effects on lesion volume change were observed for the 8 MIU treated group, with a mean increase in lesion volume of 17.1% compared with a mean decrease of 6.2% in the 8 MIU treated group.

<table>
<thead>
<tr>
<th>Time Point</th>
<th>Measurement</th>
<th>Placebo</th>
<th>Treatment</th>
<th>Placebo vs 8 MIU p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.6 MIU</td>
<td>8 MIU</td>
</tr>
<tr>
<td>1 year</td>
<td>N</td>
<td>110</td>
<td>110</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>Mean Group Change</td>
<td>12.2%</td>
<td>4.1%</td>
<td>-1.1%</td>
</tr>
<tr>
<td></td>
<td>Median change</td>
<td>10.9%</td>
<td>3.0%</td>
<td>-6.2%</td>
</tr>
<tr>
<td>2 years</td>
<td>N</td>
<td>98</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Mean Group Change</td>
<td>20.0%</td>
<td>10.5%</td>
<td>-0.1%</td>
</tr>
<tr>
<td></td>
<td>Median change</td>
<td>16.5%</td>
<td>11.4%</td>
<td>-0.8%</td>
</tr>
<tr>
<td>3 years</td>
<td>N</td>
<td>111</td>
<td>112</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>Mean Group Change</td>
<td>17.1%</td>
<td>1.1%</td>
<td>-6.2%</td>
</tr>
<tr>
<td></td>
<td>Median change</td>
<td>15.0%</td>
<td>0.2%</td>
<td>-9.3%</td>
</tr>
</tbody>
</table>

Table 8.1 The effect of IFN-β-1b on brain TLV (from Paty et al., 1993)
A dramatic treatment effect was also seen in the frequent MRI subgroup (Tables 8.2). The median percentage of active scans (containing at least one new or enlarging lesion) was 29.4% in the placebo group and 5.9% in the 8 MIU treated group. Similar striking effects were seen for the annual rate of new lesions, with a median reduction of 75% in the rate of new lesion formation.

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>1.6 MIU</th>
<th>8 MIU</th>
<th>p value placebo vs 8 MIU</th>
<th>p value placebo vs 1.6 MIU</th>
<th>p value 1.6 vs 8 MIU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median % of active scans</td>
<td>29.4</td>
<td>11.8</td>
<td>5.9</td>
<td>0.006</td>
<td>0.035</td>
<td>0.469</td>
</tr>
<tr>
<td>Median number of active lesions/yr</td>
<td>3</td>
<td>1</td>
<td>0.5</td>
<td>0.009</td>
<td>0.041</td>
<td>0.507</td>
</tr>
<tr>
<td>Median number of new lesions/yr</td>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.003</td>
<td>0.032</td>
<td>0.321</td>
</tr>
</tbody>
</table>

**Table 8.2** The effect of IFN-β-1b on MRI activity (from Paty et al., 1993)

The analysis of the final results of the blinded study was reported in 1995 (IFNB in MS Study Group 1995). Although the annual relapse rate for the further years of follow up remained about a third lower in the 8 MIU group than the placebo group, this did not reach statistical significance for individual years after year 2. This loss of significance is likely to reflect; (i) the reduced sample sizes for years 2-5 (ii) the drop in relapse rate over time observed in all arms of the study, and (iii) the fact that the dropouts with the highest relapse rates were in the placebo group.
In view of the previously noted step change in lesion volumes at year 3, all quantitative MRI analysis was repeated. The percentage change in lesion volume from baseline was measured in the 217 patients that had a year 4 or year 5 scan after baseline (Figure 8.1). The yearly increases in lesion volume in the placebo and 1.6 MIU treated group were significant, but there was no significant increase in the 8 MIU treated group. The differences between the placebo and 8 MIU treated group remained significant at all time points.

Based on the results of the above study, the FDA licensed IFN-β-1b for treatment of ambulatory patients with relapsing remitting MS.

Figure 8.1 The effect of IFN-β-1b treatment on $T_2$ lesion volume over extended follow up (adapted from IFNB MS Study Group 1995)
8.6 Recombinant IFN-β-1a (Avonex)

Recombinant IFN-β-1a (Avonex, Biogen) is produced from a mammalian cell line: the Chinese hamster ovary cell. It differs from IFN-β-1b in that it is both glycosolated and identical in its amino acid sequence to natural human IFN-β. The Multiple Sclerosis Collaborative Research Study Group (MSCRG) conducted a multicentre, randomised, double-blind, placebo-controlled study to assess its efficacy in relapsing remitting MS (Jacobs et al., 1996). Entry criteria included EDSS scores of 1-3.5 and two or more relapses in the previous 3 years. It is notable that in this study is the fact that the primary endpoint was the time to confirmed worsening on EDSS of ≥1 point, persisting for two successive scheduled visits ≥6 months apart, rather than relapse rate.

In the trial, 301 patients were randomly allocated to placebo or 6 MIU IFN-β-1a, given intramuscularly once weekly for the study duration. The MRI protocol for the study comprised gadolinium enhanced imaging at each annual time point, together with quantification of brain lesion volume.

Due to an unexpectedly low dropout rate (4%), the study was terminated before all patients had completed 2 years, once the primary endpoint was achieved. Estimated two-year progression rates using the experimental model were 22.6% for the treated patients, compared to 36.3% in the placebo group, a reduction of nearly 40%. Based on this data, the estimated median time to confirmed progression was 3.1 years for the placebo cohort and 5.4 years for the treated group (p = 0.024). The side effect profile was broadly similar to that seen in the IFN-β-1b study.

The MRI results from the annual gadolinium enhanced imaging protocol are shown in Table 8.3. Treatment was associated with a significant reduction in gadolinium enhancement at both years 1 and 2.
<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>IFN- β -1a</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of patients</td>
<td>132</td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>Mean no. Lesions (SEM)</td>
<td>2.32 (0.37)</td>
<td>3.17 (0.62)</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>0-23</td>
<td>0-56</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Year 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of patients</td>
<td>123</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>Mean no. Lesions (SEM)</td>
<td>1.59 (0.31)</td>
<td>1.04 (0.28)</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>0-22</td>
<td>0-28</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Year 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of patients</td>
<td>82</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Mean no. Lesions (SEM)</td>
<td>1.65 (0.48)</td>
<td>0.80 (0.22)</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>0-34</td>
<td>0-13</td>
<td>0.051</td>
</tr>
</tbody>
</table>

**Gadolinium enhanced Lesion Volume**

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>IFN- β -1a</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of patients</td>
<td>132</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>Mean vol. (mm³) (SEM)</td>
<td>219 (36.2)</td>
<td>255 (45.1)</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>0-2752</td>
<td>0-2858</td>
<td>0.97</td>
</tr>
<tr>
<td><strong>Year 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of patients</td>
<td>123</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>Mean vol. (mm³) (SEM)</td>
<td>96.5 (21.1)</td>
<td>70.0 (24.9)</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>0-1977</td>
<td>0-2797</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Year 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of patients</td>
<td>82</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>Mean vol. (mm³) (SEM)</td>
<td>122.4 (48.5)</td>
<td>74.1 (38.3)</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>0-3791</td>
<td>0-2847</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 8.3 Gadolinium enhanced lesion activity (from Jacobs et al., 1996)
<table>
<thead>
<tr>
<th>Patients who had 1 year MRI assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
</tr>
<tr>
<td>No. of patients</td>
</tr>
<tr>
<td>Enhancing lesion number</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Median T&lt;sub&gt;2&lt;/sub&gt; lesion volume (mm&lt;sup&gt;3&lt;/sup&gt;)</td>
</tr>
<tr>
<td>1 year</td>
</tr>
<tr>
<td>Median actual change in lesion volume (mm&lt;sup&gt;3&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Median % change in lesion volume (mm&lt;sup&gt;3&lt;/sup&gt;)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patients who had 2 year MRI assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
</tr>
<tr>
<td>No. of patients</td>
</tr>
<tr>
<td>Enhancing lesion number</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Median T&lt;sub&gt;2&lt;/sub&gt; lesion volume (mm&lt;sup&gt;3&lt;/sup&gt;)</td>
</tr>
<tr>
<td>2 years</td>
</tr>
<tr>
<td>Median actual (%) change in lesion volume (mm&lt;sup&gt;3&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Median % change in lesion volume (mm&lt;sup&gt;3&lt;/sup&gt;)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Placebo</th>
<th>IFN-β -1a</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>113</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>3.0</td>
<td>0.57</td>
</tr>
<tr>
<td>12,075</td>
<td>9,238</td>
<td>0.01</td>
</tr>
<tr>
<td>455</td>
<td>152</td>
<td>0.14</td>
</tr>
<tr>
<td>12.0</td>
<td>3.0</td>
<td>0.057</td>
</tr>
<tr>
<td>80</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>3.0</td>
<td>0.55</td>
</tr>
<tr>
<td>13,620</td>
<td>10,210</td>
<td>0.12</td>
</tr>
<tr>
<td>1,410</td>
<td>628</td>
<td>0.22</td>
</tr>
<tr>
<td>19.0</td>
<td>17.3</td>
<td>0.80</td>
</tr>
</tbody>
</table>

**Table 8.4** The effect of treatment on MRI results (adapted from Simon *et al.*, 1998)
The initial quantitative MRI analysis of $T_2$ lesion volume used a quantitative technique that had not been appropriately validated. There was an apparent decrease in lesion volume in both treated and placebo groups that was attributed to measurement drift (the initial analysis had been performed over a 3-4 year period). Analysis was subsequently repeated using a manual outlining approach (Simon et al., 1998), together with visual identification of new/enlarging $T_2$ lesions appearing over the 2 year study period (Table 8.4).

As shown in Table 8.4, the treatment and placebo groups were not well matched according to baseline $T_2$ lesion volume. While re-analysis revealed the expected annual increase in lesion volume in the placebo group, no significant treatment effect on lesion volume was identified. In contrast, a significant treatment effect on development of new $T_2$ lesions was seen (mean number of new $T_2$ lesions over 2 years in placebo and treated groups 3.2 and 2.0 respectively, $p = 0.006$).

Linear regression analysis was performed to identify those factors predictive of change in $T_2$ lesion volume and number of new $T_2$ lesions in the placebo group. The model revealed that the number of baseline gadolinium enhanced lesions was positively predictive for change in lesion volume ($r^2 = 0.368$, $p < 0.001$) and new lesion formation ($r^2 = 0.147$, $p < 0.001$).
8.7 Phase III study of IFN-β-1a (Rebif)

A second study using IFN-β-1a has recently been reported (PRISMS Study Group 1998). This used a formulation (Rebif, Ares-Serono) that facilitated subcutaneous rather than intramuscular injection. Entry criteria included a diagnosis of relapsing remitting MS, EDSS of 0-5.0 and at least 2 relapses within the previous 2 years. A total of 560 patients were recruited to either placebo (187 patients) or one of two doses of IFN-β-1a (22µg for 189 patients and 44µg for 184 patients), administered 3 times per week. The primary outcome measure was relapse rate. Other clinical efficacy measures included times to first and second relapse and progression in disability (defined as a sustained increase in EDSS of at least 1 point). All patients were also subjected to an MRI protocol involving 6 monthly imaging to measure T2 lesion volume and identify lesion activity.

Treatment was associated with a significant reduction in relapse rate of 27-33% at both doses (p < 0.005). Significant reductions in the number of moderate/severe relapses (p < 0.005) and delay in time to both first and second relapses were also seen. Furthermore, significant beneficial treatment effects with both dosage regimes were found for both time to sustained progression of EDSS and EDSS change.

Brain T2 lesion volume showed a progressive median increase in the placebo group of 10.9% over two years, whereas the 22µg treated group showed a median decrease of 1.2% and the 44µg group 3.8% (p < 0.0001 compared with placebo for both doses). The number of active T2 lesions identified 6 monthly was also significantly lower (67% and 78%) in the low and high dose groups respectively (p < 0.0001). A dose effect was also observed in favour of the high dose (p = 0.0003).
8.8 Summary of the IFN-β studies in relapsing remitting MS

The results of these phase III studies confirm a consistent beneficial effect of treatment on relapse rate. However, the data on disability progression is less secure. No significant effect on disability was demonstrated in the IFN-β-1b study. Furthermore, while a significant delay in EDSS progression has been shown with both studies using IFN-β-1a, the Avonex study has been criticised for the premature termination and extrapolated nature of a number of outcome measures, the relatively small number of patients and lack of effect on T₂ lesion volume. The results of the PRISMS study are reassuring in confirming an effect of IFN-β-1a on disability progression.

A direct comparison of the MRI outcomes is largely precluded by variations in the imaging protocols adopted in the 3 studies. However, a powerful effect of treatment on MRI activity in terms of new lesion appearance on T₂ weighted or enhanced imaging has been demonstrated, providing important data to supplement the clinical outcomes. A dosage effect on both clinical and MRI outcomes is also suggested by both of the studies that incorporated different dosage regimes. High dose was favoured in the IFN-β-1b study for change in T₂ lesion volume, and in the Rebif study for new lesion activity (no p values being provided for dosage effect on lesion volume in the latter study). A crossover study with IFN-β-1a that incorporated a monthly gadolinium enhanced protocol also suggested a dosage effect on MRI activity (Pozzilli et al., 1996). However, more work is needed to define the optimum dosage/route of administration.

8.9 The IFN-β-1b study in secondary progressive MS

The above studies were all performed on patients in the relapsing remitting phase of the disease. A need to assess the utility of IFN-β in secondary progressive MS was apparent, and a phase III study was therefore designed to address this issue (Polman et al., 1995). A major concern of
patients with secondary progressive MS and their physicians is the gradual development of
disability. The time to confirmed progression in disability was therefore chosen as the primary
endpoint of this study. An extensive MRI protocol was also incorporated, details of which are
given in the Chapter 9.
Chapter 9. The IFN-β-1b study: Results of the annual MRI protocol

9.1 Trial methodology

Between August 1994 and June 1995, 718 patients with secondary progressive MS were recruited from 30 European centres into a 3 year randomised, double-blind, placebo-controlled trial of IFN-β-1b. The study was designed to have a 36 month period of treatment, followed by a drug free follow up for 3 months (Polman et al., 1995). The key inclusion criteria were as follows:

- Age 18-55 years
- Clinically definite of laboratory supported MS for ≥ 1 year
- Disease in SP phase
- Clinically active MS with either deterioration of at least 1 EDSS point or ≥ 2 relapses within the preceding 24 months
- No relapse within 30 days of entry into study
- EDSS score of 3.0-6.5

The primary efficacy measure for the study was time to confirmed progression in EDSS on two consecutive visits, 3 months apart. A one point change in EDSS was required to confirm progression for EDSS scores 3.0-5.5. However, half point changes between EDSS 6.0 and 7.0 were scored as one point changes, to improve the sensitivity and reliability of assessment (Goodkin 1991; Polman et al., 1995).
Patients were randomised to receive either placebo or IFN-β-1b, 4 MIU subcutaneously, subsequently increasing after 2 weeks to 8 MIU or placebo on alternate days. Treatment was discontinued in cases of severe side effects or clinically relevant laboratory deviations, pregnancy, use of immunomodulatory medication other than steroids (given in a standardised regime), or if the code was broken.

Sample size estimations were based on the primary clinical outcome (time to confirmed EDSS progression). Assuming 50% of placebo patients would progress within 3 years and a 20% dropout rate, with 360 patients per arm, a difference of -12.5% would be detected with a power of 0.80 in a two sided log rank test with $\alpha = 0.029$ (Polman et al., 1995). An interim analysis was planned after all patients had been in the study for at least 24 months and all statistical analysis was performed on an intention to treat (ITT) basis.

The other clinical outcome measures included relapse rate and time to becoming wheelchair bound. Full details of the trial design, inclusion and exclusion criteria of the patients have been described elsewhere (Polman et al., 1995; European Study Group on IFN-β-1b in SPMS 1998).

9.1.1 MRI acquisition protocol

A core imaging protocol was performed for all patients entered into the study. This comprised an annual dual echo CSE PD/T1 weighted brain MRI sequence. Imaging was performed locally at each site and data were then transferred to the MRI Analysis centre (National Hospital, Queen Square, London), where analyses were performed by staff blinded to the clinical details (see below).
The following MRI endpoints were evaluated from the core protocol:

**Primary MRI outcome**

Change in brain $T_2$ lesion volume

**Secondary MRI outcomes**

Number of new or enlarging lesions developing over the study duration

Number of active $T_2$ weighted scans, ie containing any new/enlarging $T_2$ lesions

Number of active patients, ie patients with any active scans

The MR machine manufacturer, field strength and imaging parameters varied according to site, but within the parameters listed as follows: 0.6-1.5 tesla machine; TR = 2-2.5 s; short TE = 30-40 ms; long TE = 80-100 ms; 28 axial oblique, contiguous, interleaved, 5 mm thick slices; matrix 256 x 256; FOV 25 cm; 1 or 2 excitations. Repositioning was achieved using a protocol based on identification of standardised anatomical landmarks (Gallagher et al., 1997). This involved the acquisition of planning scans in the following order: a single axial scan was performed from which a coronal was planned using an oblique projection where necessary to compensate for patient misalignment; a sagittal planning scan was then prescribed from the coronal, again compensating for misalignment. Finally, the main scan series was prescribed from the sagittal scan, the orientation of this running from a line joining the infra anterior and infra posterior parts of the corpus callosum.

Prior to entry into the trial, each MRI centre sent a representative MRI dataset to the central analysis centre. The imaging protocol for this comprised two consecutive imaging sessions separated by an interval of a few minutes, between which the patient was removed from the
scanner. This dataset was assessed for each site to confirm that the imaging parameters, image quality and repositioning strategy were acceptable.

Throughout the duration of the trial, hard copy images of all MRI studies were checked by the central analysis centre within a few days of image acquisition to confirm adherence to the acquisition and repositioning guidelines. Where deviations from these guidelines occurred, imaging was repeated within a few days.

A subgroup of 125 patients from 7 centres also underwent monthly gadolinium enhanced imaging at months 0-6 and 19-24, to assess the effect of treatment on frequent MRI activity. Furthermore, a smaller subgroup of 95 patients from 5 centres had an additional protocol using several putative markers of demyelination and axonal loss; (i) measurement of $T_1$ hypo-intense lesion volume on pre and post gadolinium enhanced $T_1$-weighted SE scans (ii) quantification of atrophy in the brain and spinal cord, and (iii) magnetisation transfer imaging. However, this thesis will be restricted to analysis of the annual MRI protocol, since this was the major task of work undertaken or supervised by the author.

9.1.2 MRI Analysis

All MRI scans at each site were archived on to hard copy film and electronic media and transported to the MRI analysis site. The hard copy was analysed by the author, supported by a number of other experienced clinical raters, and analysis was performed in pairs. The author himself marked over 70% of the scans, and of those marked by other clinicians, the author checked the overwhelming majority. The first ten serial MRI datasets were marked by the author and checked by an experienced neuroradiologist to ensure an appropriate standard of lesion
identification. The same neuroradiologist also regularly reviewed a small proportion of marked studies throughout the trial to ensure a consistently high standard of lesion marking.

Every lesion visible on the PD-weighted scan was marked and roughly outlined on an overlaid transparency. First, lesions were identified and marked on the baseline scan. The month 12 study was then compared side by side with the baseline scan and lesions were then marked on the month 12 study.

In parallel with this process, all new and enlarging lesions that had developed between these time points were identified and marked by consensus. This same approach was used to assess the month 24 against month 12, and month 36 against month 24 studies. Finally, a review of the entire scan series was performed by the author, to ensure a high standard of consistency in lesion delineation across the whole scan data set. This approach therefore identified all the lesions to be included in the T2 lesion volume quantification and also documented the number of new and enlarging lesions seen on each annual follow up scan.

The electronic data were received from the scanning sites on their usual archive media, and in their particular scanner's data format. They were first transferred to a 'JukeBox' system, attached to a network of Sun Workstations (Sun Microsystems, Palo Alto, [A]), and then converted to a single common format understood by all the subsequent display and analysis tools. The data were then sorted by site, patient identification number and scan date, and duplicate images (from scans repeated due to artefacts or patient motion, or because of multiple archiving, or other problems) were discarded. Throughout both the data conversion and sorting, the original hard copy (received from the site immediately after a scan was completed) was used to ensure that
both the image data themselves, and the associated patient data (initials, date of birth, scan date etc) were correctly maintained.

A group of raters, previously trained to a high level of reproducibility (see below), then performed the T₂ lesion volume analysis as follows: (i) the hard copy with all lesions outlined was placed alongside the identical computer generated image; (ii) all lesions marked on the hard copy were outlined on the computer image using the semi-automated contour technique (see Chapter 4); (iii) after all lesions had been outlined, another observer checked the consistency of lesion identification from hard copy to computer image, and any lesions missed in the latter were outlined by the original rater; (iv) a computer programme then summed all the individual lesion volumes (calculated as the surface area of each lesion multiplied by the slice thickness [5 mm in all cases]) and a final T₂ lesion volume was generated and stored in a specially constructed database.

Before starting to analyse the computerised images from the trial patients, each rater underwent a period of comprehensive training in the use of the contour technique. Once a rater had developed a sufficient level of expertise, performance was evaluated using a standardised MR data set, comprising the PD/T₂ weighted brain images of 16 MS patients with a representative range of lesion volumes. Each rater measured the brain lesion volume on two separate occasions, separated by an interval of at least one week. The minimum acceptance criteria for passing this validation procedure was considered to be a median intra-rater CoV of less than 5%. Only once this had been demonstrated was a rater allowed to commence work on the trial data set.
9.1.3 Statistical analysis of annual MRI data

A team of external statisticians performed the analysis on an intention to treat (ITT) basis and data up to the final termination of the study was evaluated. The baseline MRI characteristics for the placebo and treatment groups were compared using the Wilcoxon rank sum test. Non parametric statistical methods were employed for the analysis of all MRI endpoints and p-values for two sided statistical test were provided.

9.1.3.1 Primary MRI outcome; change in $T_2$ lesion volume

Only patients with a valid baseline scan and at least one subsequent on-study scan were included in the analysis; no correction was made for missing values. Treatment groups were compared with respect to percentage and absolute changes in $T_2$ lesion volume from baseline to endpoint, based on a non parametric method of analysis of covariance. Stratification adjustment was performed for centre using the Extended Mantel-Haenszel test for covariance.

9.1.3.2 Number of new/enlarged lesions

The cumulative sums of active (new and enlarging) lesions was determined for all timepoints and all patients; i.e., 12 months, 12 months plus 24 months for 24 months, 12 months plus 24 months plus 36 months for 36 months. Non parametric analysis of covariance was again used for comparison of treatments, with standardisation across centres using the Extended Mantel-Haenszel test.

9.1.3.3 Number of active scans and number of active patients

The number of active scans was analysed in terms of the proportion of active scans among those performed for a given patient, analysed as a categorical variable. Comparison between treatment
groups for both the number of active scans and patients was again performed using the Extended Mantel-Haenszel test, with stratification adjustment for centre.

9.1.3.4 Clinical/MRI correlations

Rank correlation coefficients were used throughout for assessing the strength of relationship between clinical and MRI activity measures, using Goodman-Kruskal coefficients. Correlations were calculated for treatment and placebo groups in isolation, and overall (for the entire trial population). The significance of correlation was determined using a normal approximation based on the pooled treatment mean correlation and its standard error.

9.2 Clinical Results

A total of 718 patients of 768 screened in 32 European centres were randomly assigned to IFN-β-1b (n = 360) or placebo (n = 358). Based on the results of the prospective interim analysis of the 2 year data, the independent advisory board recommended early termination of the study (European Study Group 1998). The mean follow up period at interim cut-off was 892 days in the placebo arm and 901 days in the IFN-β-1b treated group. Treatment groups were comparable for all baseline variables (Table 9.1).

A total of 57 patients (31 placebo, 26 IFN-β-1b) dropped out of the study, and a further 130 patients (66 placebo, 64 IFN-β-1b) stopped treatment but were followed up.

For the primary efficacy variable, time to confirmed neurological deterioration in EDSS, the non parametric analysis of covariance showed a significant effect favouring treatment (p = 0.0008). Of the 358 patients on placebo, 178 (49.8%) had confirmed progression, while in the treated
group of 360 patients, 140 (38.9%) had progressed. Piecewise logistic regression analysis demonstrated an odds ratio of 0.65 (95% CI 0.52-0.83).

The estimated probability of remaining progression free (estimated survival rates) was also calculated for each 3 month study period. Group differences became significant at 12 months and remained significant throughout the remainder of the study. The treatment effect on progression was similar, irrespective of baseline EDSS or superimposed relapses before, or during the study (Table 9.2).

<table>
<thead>
<tr>
<th></th>
<th>Placebo (n = 358)*</th>
<th>IFN-β-1b (n = 360)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age in years (SD)</td>
<td>40.9 (7.2)</td>
<td>41.1 (7.2)</td>
</tr>
<tr>
<td>Women</td>
<td>64.2%</td>
<td>58.1%</td>
</tr>
<tr>
<td>Mean Disease Duration in years (SD)</td>
<td>13.4 (7.5)</td>
<td>12.8 (6.6)</td>
</tr>
<tr>
<td>Mean time since evidence of progressive deterioration in years (SD)</td>
<td>2.1 (2.2)</td>
<td>2.2 (2.4)</td>
</tr>
<tr>
<td>Mean EDSS at baseline (SD)</td>
<td>5.2 (1.1)</td>
<td>5.1 (1.1)</td>
</tr>
<tr>
<td>EDSS by category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 3.5</td>
<td>47 (13.1%)</td>
<td>67 (18.6%)</td>
</tr>
<tr>
<td>4.0-5.5</td>
<td>142 (39.7%)</td>
<td>140 (38.9%)</td>
</tr>
<tr>
<td>≥6.0</td>
<td>169 (47.2%)</td>
<td>153 (42.5%)</td>
</tr>
<tr>
<td>Patients without relapses in 2 years before study</td>
<td>101 (28.2%)</td>
<td>115 (31.9%)</td>
</tr>
</tbody>
</table>

* No significant differences between treatment groups (p > 0.05)

Table 9.1 Baseline clinical characteristics
<table>
<thead>
<tr>
<th></th>
<th>Placebo (n = 358)</th>
<th>IFN-β-1b (n = 360)</th>
<th>*Relative difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline EDSS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 3.5</td>
<td>47</td>
<td>67</td>
<td>-19.5%</td>
</tr>
<tr>
<td>4.0-5.5</td>
<td>142</td>
<td>140</td>
<td>-20.8%</td>
</tr>
<tr>
<td>≥6.0</td>
<td>169</td>
<td>153</td>
<td>-22.1%</td>
</tr>
<tr>
<td><strong>Relapse during study</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With</td>
<td>224</td>
<td>194</td>
<td>-19.9%</td>
</tr>
<tr>
<td>Without</td>
<td>134</td>
<td>166</td>
<td>-22.0%</td>
</tr>
<tr>
<td><strong>Relapse 2 years before study</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With</td>
<td>257</td>
<td>245</td>
<td>-22.9%</td>
</tr>
<tr>
<td>Without</td>
<td>101</td>
<td>115</td>
<td>-19.2%</td>
</tr>
</tbody>
</table>

*Relative difference in proportion of patients with confirmed progression between placebo and treated groups.

**Table 9.2** Proportion of patients with confirmed progression by baseline EDSS and occurrence of relapse before/during study
Other clinical outcome measures including the effect of treatment on relapse rate are given in Table 9.3. The mean annual relapse rate was reduced by about 30% overall in the treatment group. The rates dropped annually in both groups (Table 9.3), maintaining treatment effect over time (although not significant at year 3).

<table>
<thead>
<tr>
<th>Efficacy Variable</th>
<th>Placebo (n = 358)</th>
<th>IFN-β-1b (n = 360)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to becoming wheelchair bound</td>
<td></td>
<td></td>
<td>0.0133</td>
</tr>
<tr>
<td>Estimated probability of not becoming wheelchair bound</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year 1</td>
<td>0.90</td>
<td>0.96</td>
<td>0.0129</td>
</tr>
<tr>
<td>Year 2</td>
<td>0.81</td>
<td>0.89</td>
<td>0.0094</td>
</tr>
<tr>
<td>Year 3</td>
<td>0.66</td>
<td>0.77</td>
<td>0.0133</td>
</tr>
<tr>
<td>Mean EDSS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At endpoint</td>
<td>5.84</td>
<td>5.57</td>
<td>0.0750</td>
</tr>
<tr>
<td>Change at endpoint (endpoint minus baseline)</td>
<td>0.60</td>
<td>0.47</td>
<td>0.0299</td>
</tr>
<tr>
<td>Mean annual relapse rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.64</td>
<td>0.44</td>
<td>0.0002</td>
</tr>
<tr>
<td>Year 1</td>
<td>0.82</td>
<td>0.57</td>
<td>0.0095</td>
</tr>
<tr>
<td>Year 2</td>
<td>0.47</td>
<td>0.35</td>
<td>0.0201</td>
</tr>
<tr>
<td>Year 3</td>
<td>0.35</td>
<td>0.24</td>
<td>0.1624</td>
</tr>
<tr>
<td>Median time to first relapse (days)</td>
<td>403</td>
<td>644</td>
<td>0.0030</td>
</tr>
<tr>
<td>Proportion of patients with moderate/severe relapses</td>
<td>53.1%</td>
<td>43.6%</td>
<td>0.0083</td>
</tr>
</tbody>
</table>

Table 9.3 Secondary/tertiary outcome variables
9.3 Results of annual MRI analysis

The placebo and IFN-β-1b groups were well matched with respect to their baseline T₂ lesion volume (Table 9.4): the median T₂ lesion volume was 23.8 cm³ in the placebo and 21.6 cm³ in the IFN-β-1b group (p = 0.41). In the placebo group, there was a highly significant increase in T₂ lesion volume at each yearly time point when compared to baseline T₂ lesion volume (Table 9.5), and also when compared to the previous year's T₂ lesion volume (Table 9.6). The absolute and percentage increase in T₂ lesion volume (Tables 9.5 and 9.6; Figure 9.1) was somewhat larger in the third year than in the preceding two years. By the end of the third year, the mean and median percent increases in T₂ lesion volume compared to baseline were 16% and 11% respectively.

<table>
<thead>
<tr>
<th>Treatment arm</th>
<th>No. patients</th>
<th>Median</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>344</td>
<td>23.82</td>
<td>28.35</td>
<td>22.46</td>
<td>0.6</td>
<td>135.4</td>
</tr>
<tr>
<td>IFN-beta-1b</td>
<td>346</td>
<td>21.60</td>
<td>26.62</td>
<td>21.17</td>
<td>0.3</td>
<td>128.8</td>
</tr>
<tr>
<td>Overall</td>
<td>690</td>
<td>22.14</td>
<td>27.48</td>
<td>21.82</td>
<td>0.3</td>
<td>135.4</td>
</tr>
</tbody>
</table>

Wicoxon rank-sum test two sided p = 0.412 for comparison between treatment arms

Table 9.4 Baseline T₂ lesion volume (cm³) according to treatment arm

In the IFN-β-1b group, there was a significant decrease in T₂ lesion volume at year 1 compared with baseline with mean and median decreases were 4 and 5% respectively (Table 9.5 and 9.6; Figure 9.1). There was no significant change in T₂ lesion volume between years 1 and 2 (p = 0.25), but a significant increase in T₂ lesion volume occurred during the third year (p = 0.0001), although to a lesser degree than that seen in the placebo group (Table 9.6). The net effect of the initial decrease and subsequent increase in T₂ lesion volume in the treated group was such that
there was no net significant difference in $T_2$ lesion volume in the IFN-beta-1b treated group (Table 9.5) between baseline and year 3 ($p = 0.15$).

When the placebo and IFN-β-1b groups were compared, a highly significant difference in the $T_2$ lesion volume change was identified (in both absolute and percentage terms) in favour of IFN-β-1b at all three annual follow up periods compared to baseline (Tables 9.5 and 9.7), and also over each yearly interval i.e including year 2 versus year 1 and year 3 versus year 2 (Table 9.6).

![Graph showing mean percentage annual change in $T_2$ lesion volume according to treatment arm]

**Figure 9.1** The mean percentage annual change in $T_2$ lesion volume according to treatment arm
<table>
<thead>
<tr>
<th>Visit</th>
<th>Treatment</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>p-value (1)</th>
<th>p-value (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>Placebo</td>
<td>321</td>
<td>1.31</td>
<td>0.30</td>
<td>-12.4</td>
<td>35.3</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>329</td>
<td>-1.22</td>
<td>-0.77</td>
<td>-17.6</td>
<td>33.1</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Year 2</td>
<td>Placebo</td>
<td>302</td>
<td>2.30</td>
<td>0.40</td>
<td>-12.6</td>
<td>60.9</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>308</td>
<td>-1.53</td>
<td>-1.06</td>
<td>-26.8</td>
<td>51.9</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Year 3</td>
<td>Placebo</td>
<td>274</td>
<td>4.26</td>
<td>1.79</td>
<td>-14.4</td>
<td>71.6</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>293</td>
<td>-0.61</td>
<td>-0.73</td>
<td>-31.1</td>
<td>73.2</td>
<td>0.1530</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Note: p-value (1) represents t-test for significance of the within group change from baseline; p-value (2) represents between group significance using extended Mantel-Haenszel test with stratification adjustment for centre and covariance adjustment for both baseline T₂ lesion volume and centre.

Table 9.5 Absolute change in T₂ lesion volume (cm³) from baseline according to treatment
<table>
<thead>
<tr>
<th>Visit</th>
<th>Treatment</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>p-value (1)</th>
<th>p-value (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 0-1 Placebo</td>
<td>321</td>
<td>1.31</td>
<td>0.30</td>
<td>-12.4</td>
<td>35.3</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFN-β-1b</td>
<td>329</td>
<td>-1.22</td>
<td>-0.77</td>
<td>-17.6</td>
<td>33.1</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>Year 1-2 Placebo</td>
<td>306</td>
<td>1.04</td>
<td>0.15</td>
<td>-21.4</td>
<td>31.7</td>
<td>0.0005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFN-β-1b</td>
<td>318</td>
<td>-0.23</td>
<td>-0.20</td>
<td>-18.5</td>
<td>24.8</td>
<td>0.2464</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td>Year 2-3 Placebo</td>
<td>272</td>
<td>2.32</td>
<td>1.04</td>
<td>-26.0</td>
<td>27.6</td>
<td>&lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFN-β-1b</td>
<td>296</td>
<td>0.88</td>
<td>0.37</td>
<td>-25.7</td>
<td>21.3</td>
<td>0.0001</td>
<td>0.0002</td>
<td></td>
</tr>
</tbody>
</table>

Note: p-value (1) represents t-test for significance of the within group change from previous year; p-value (2) represents between group significance using extended Mantel-Haenszel test with stratification adjustment for centre and covariance adjustment for baseline $T_2$ lesion volume and centre.

**Table 9.6** Change in absolute $T_2$ lesion volume ($cm^3$) from previous year according to treatment
<table>
<thead>
<tr>
<th>Visit</th>
<th>Treatment</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>p-value (1)</th>
<th>p-value (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>Placebo</td>
<td>321</td>
<td>3.60</td>
<td>1.64</td>
<td>-30.7</td>
<td>111.6</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>329</td>
<td>-3.71</td>
<td>-4.94</td>
<td>-48.5</td>
<td>202.3</td>
<td>0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Year 2</td>
<td>Placebo</td>
<td>302</td>
<td>7.77</td>
<td>2.42</td>
<td>-35.9</td>
<td>123.5</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>308</td>
<td>-4.77</td>
<td>-6.92</td>
<td>-57.7</td>
<td>317.6</td>
<td>0.0009</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Year 3</td>
<td>Placebo</td>
<td>274</td>
<td>16.01</td>
<td>10.98</td>
<td>-44.0</td>
<td>226.2</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>293</td>
<td>-1.61</td>
<td>-5.24</td>
<td>-53.9</td>
<td>234.1</td>
<td>0.2847</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Note: p-value (1) represents t-test for significance of the within group change from baseline and p-value (2) represents between group significance using extended Mantel-Haenszel test, adjusted for baseline lesion volume and centre, stratified for centre.

Table 9.7 Percentage change in T₂ lesion volume from baseline according to treatment
<table>
<thead>
<tr>
<th>Visit</th>
<th>Treatment</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
<th>p-value (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>Placebo</td>
<td>345</td>
<td>3.76</td>
<td>2.00</td>
<td>0.0</td>
<td>45.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>350</td>
<td>1.48</td>
<td>0.00</td>
<td>0.0</td>
<td>32.0</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Year 2</td>
<td>Placebo</td>
<td>345</td>
<td>6.67</td>
<td>4.00</td>
<td>0.0</td>
<td>58.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>350</td>
<td>2.65</td>
<td>1.00</td>
<td>0.0</td>
<td>68.0</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Year 3</td>
<td>Placebo</td>
<td>345</td>
<td>8.82</td>
<td>5.00</td>
<td>0.0</td>
<td>87.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IFN-β-1b</td>
<td>350</td>
<td>3.77</td>
<td>1.50</td>
<td>0.0</td>
<td>71</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Note: p-value (1) represents between group significance using extended Mantel-Haenszel test with stratification adjustment for centre.

Table 9.8 Cumulative number of new/enlarging lesions according to treatment arm
There was also a highly significant reduction in the cumulative number of active (new or enlarging) lesions in the IFN-β-1b group at all annual time points, apparent even at year 1 (Table 9.8 and Figure 9.2). The mean and median reduction compared to placebo were 57% and 80% respectively over the three years of follow up.

Figure 9.2 Mean cumulative number of active T₂ lesions according to treatment arm
Significant differences were identified in the proportion of active scans per patient, again favouring treatment (Table 9.9). Furthermore, there was a significant beneficial treatment effect on the proportion of active patients (Table 9.10), with no active lesions seen in 16.2% of placebo and 35.7% of IFN-β-1b patients over the study period (p < 0.0001, extended Mantel-Haenszel test).

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>IFN-β-1b</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 50%</td>
<td>112 (32.3)</td>
<td>225 (64.3%)</td>
<td>337 (48.5%)</td>
</tr>
<tr>
<td>&gt; 50%</td>
<td>233 (61.7%)</td>
<td>125 (35.7%)</td>
<td>358 (51.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>345 (100.0%)</td>
<td>350 (100.0%)</td>
<td>695 (100.0%)</td>
</tr>
</tbody>
</table>

Extended Mantel-Haenszel test with stratification adjustment for centre: p < 0.0001

**Table 9.9** The proportion of active scans per patient according to treatment group

<table>
<thead>
<tr>
<th></th>
<th>Placebo</th>
<th>IFN-β-1b</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>56 (16.2%)</td>
<td>125 (35.7%)</td>
<td>181 (26.0%)</td>
</tr>
<tr>
<td>*Yes</td>
<td>289 (83.8%)</td>
<td>225 (64.3%)</td>
<td>476 (74.0%)</td>
</tr>
<tr>
<td>Total</td>
<td>345 (100.0%)</td>
<td>350 (100.0%)</td>
<td>695 (100.0%)</td>
</tr>
</tbody>
</table>

*Patient had at least one scan with activity

Extended Mantel-Haenszel test with stratification adjustment for centre: p < 0.0001

**Table 9.10** The proportion of active patients according to treatment group
9.4 Correlations between different MRI outcomes

In the annual MRI analyses, both absolute and percentage change in T$_2$ lesion volume correlated significantly with T$_2$ lesion activity (Table 9.11) over three years ($r = 0.36$, $p < 0.0001$ for absolute T$_2$ lesion volume change). The year on year correlations were also significant for all patients combined and for both treatment arms considered separately, except for the IFN-β-1b group during the first 12 months ($r = 0.09$, $p > 0.05$). The baseline T$_2$ lesion volume correlated with the cumulative number of active (new/enlarging) T$_2$ lesions over the study period ($r = 0.16$, $p < 0.0001$); the strength of this correlation was similar in the placebo ($r = 0.18$, $p < 0.0001$) and IFN-β-1b ($r = 0.16$, $p = 0.0003$) groups.

The relationship between baseline T$_2$ lesion volume and absolute change in T$_2$ lesion volume was more complex, differing according to treatment group (Table 9.12). In the placebo group, there was a positive correlation ($r = 0.18$, $p < 0.0001$), and in the IFN-β-1b group a negative correlation ($r = -0.16$, $p < 0.0001$) over the whole study period. In a year by year analysis, the negative correlation in the IFN-β-1b group was most clearly evident during the first year of treatment ($r = 0.22$, $p < 0.0001$).
<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Placebo</th>
<th>IFN-β-1b</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r</td>
<td>p-value</td>
<td>r</td>
</tr>
<tr>
<td>ear 0-1 TLV change (cm³)</td>
<td>Year 1 T₂ lesions</td>
<td>0.255</td>
<td>&lt; 0.0001</td>
<td>0.093</td>
</tr>
<tr>
<td>ear 1-2 TLV change (cm³)</td>
<td>Year 2 T₂ lesions</td>
<td>0.189</td>
<td>&lt; 0.0001</td>
<td>0.151</td>
</tr>
<tr>
<td>ear 2-3 TLV change (cm³)</td>
<td>Year 3 T₂ lesions</td>
<td>0.342</td>
<td>&lt; 0.0001</td>
<td>0.341</td>
</tr>
<tr>
<td>ear 0-3 TLV change (cm³)</td>
<td>Year 0-3 T₂ lesions</td>
<td>0.357</td>
<td>&lt; 0.0001</td>
<td>0.153</td>
</tr>
<tr>
<td>ear 0-1 TLV % change</td>
<td>Year 1 T₂ lesions</td>
<td>0.286</td>
<td>&lt; 0.0001</td>
<td>0.133</td>
</tr>
<tr>
<td>ear 1-2 TLV % change</td>
<td>Year 2 T₂ lesions</td>
<td>0.223</td>
<td>&lt; 0.0001</td>
<td>0.179</td>
</tr>
<tr>
<td>ear 2-3 TLV % change</td>
<td>Year 3 T₂ lesions</td>
<td>0.349</td>
<td>&lt; 0.0001</td>
<td>0.337</td>
</tr>
<tr>
<td>ear 0-3 TLV% change</td>
<td>Year 0-3 T₂ lesions</td>
<td>0.351</td>
<td>&lt; 0.0001</td>
<td>0.211</td>
</tr>
</tbody>
</table>

Table 9.11 Correlations between change in T₂ lesion volume and the cumulative number of active T₂ lesions

162
<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Placebo</th>
<th></th>
<th>IFN-β-1b</th>
<th></th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r</td>
<td>p-value</td>
<td>r</td>
<td>p-value</td>
<td>r</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 0-1 TLV change (cm³)</td>
<td>0.116</td>
<td>0.0067</td>
<td>-0.220</td>
<td>&lt; 0.0001</td>
<td>-0.052</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 1-2 TLV change (cm³)</td>
<td>-0.014</td>
<td>0.7428</td>
<td>-0.092</td>
<td>0.0257</td>
<td>-0.050</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 2-3 TLV change (cm³)</td>
<td>0.184</td>
<td>&lt; 0.0001</td>
<td>0.124</td>
<td>0.0024</td>
<td>0.155</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 0-3 TLV change (cm³)</td>
<td>0.182</td>
<td>&lt; 0.0001</td>
<td>-0.160</td>
<td>&lt; 0.0001</td>
<td>0.021</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 1 T₂ lesions</td>
<td>0.230</td>
<td>&lt; 0.0001</td>
<td>0.162</td>
<td>0.0008</td>
<td>0.192</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 2 T₂ lesions</td>
<td>0.144</td>
<td>0.0009</td>
<td>0.153</td>
<td>0.0039</td>
<td>0.153</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 3 T₂ lesions</td>
<td>0.041</td>
<td>0.3659</td>
<td>0.066</td>
<td>0.2216</td>
<td>0.055</td>
</tr>
<tr>
<td>Baseline TLV</td>
<td>Year 0-3 T₂ lesions</td>
<td>0.180</td>
<td>&lt; 0.0001</td>
<td>0.155</td>
<td>0.0003</td>
<td>0.158</td>
</tr>
</tbody>
</table>

Table 9.12 Correlations between baseline T₂ lesion volume and subsequent annual MRI activity
9.5 Clinical/MRI correlations

There were significant, albeit modest, cross sectional correlations between EDSS and T₂ lesion volume demonstrated at baseline and each follow up year (Table 9.13), with r-values in the range of 0.09 to 0.15 for the trial population as a whole.

Significant correlations were also identified for the change in EDSS against both absolute and percentage change in T₂ lesion volume for the whole trial population (Table 9.13) between each annual timepoint and for the entire study duration, with an r-value for change over the total study period of 0.17 (p < 0.0001).

Correlations of a similar magnitude were found for the change in EDSS against the number of active T₂ lesions at each annual timepoint and for the entire study duration (Table 9.14), with an r-value for change over the total study period of 0.18 (p < 0.0001). A somewhat stronger relationship was found between the relapse rate and activity on the annual T₂ scans (Table 9.14), with an r-value over the study period of 0.25 for the total trial population (p < 0.0001).

Finally, a significant (albeit modest) correlation was identified between baseline T₂ lesion volume and subsequent change in EDSS over 3 years (r = 0.067, p = 0.020), implying that T₂ lesion volume is, to a minor extent, predictive of clinical outcome.
## Table 9.13 Cross sectional and longitudinal correlations between EDSS and T<sub>2</sub> lesion volume according to treatment group

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Placebo</th>
<th>IFN-β-1b</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r</td>
<td>p-value</td>
<td>r</td>
</tr>
<tr>
<td>Baseline EDSS</td>
<td>Baseline TLV</td>
<td>0.065</td>
<td>0.1124</td>
<td>0.104</td>
</tr>
<tr>
<td>Year 1 EDSS</td>
<td>Year 1 TLV</td>
<td>0.065</td>
<td>0.1334</td>
<td>0.125</td>
</tr>
<tr>
<td>Year 2 EDSS</td>
<td>Year 2 TLV</td>
<td>0.097</td>
<td>0.0258</td>
<td>0.095</td>
</tr>
<tr>
<td>Year 3 EDSS</td>
<td>Year 3 TLV</td>
<td>0.150</td>
<td>0.0011</td>
<td>0.126</td>
</tr>
<tr>
<td>Year 0-1 EDSS change</td>
<td>Year 0-1 TLV change (cm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>0.096</td>
<td>0.0333</td>
<td>0.065</td>
</tr>
<tr>
<td>Year 1-2 EDSS change</td>
<td>Year 1-2 TLV change (cm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>0.094</td>
<td>0.0646</td>
<td>0.095</td>
</tr>
<tr>
<td>Year 2-3 EDSS change</td>
<td>Year 2-3 TLV change (cm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>0.160</td>
<td>0.0053</td>
<td>0.123</td>
</tr>
<tr>
<td>Year 0-3 EDSS change</td>
<td>Year 0-3 TLV change (cm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>0.140</td>
<td>0.0018</td>
<td>0.179</td>
</tr>
<tr>
<td>Year 0-1 EDSS change</td>
<td>Year 0-1 TLV % change</td>
<td>0.133</td>
<td>0.0048</td>
<td>0.074</td>
</tr>
<tr>
<td>Year 1-2 EDSS change</td>
<td>Year 1-2 TLV % change</td>
<td>0.117</td>
<td>0.0234</td>
<td>0.075</td>
</tr>
<tr>
<td>Year 2-3 EDSS change</td>
<td>Year 2-3 TLV % change</td>
<td>0.106</td>
<td>0.0544</td>
<td>0.166</td>
</tr>
<tr>
<td>Year 0-3 EDSS change</td>
<td>Year 0-3 TLV % change</td>
<td>0.140</td>
<td>0.0015</td>
<td>0.187</td>
</tr>
<tr>
<td>Variable 1</td>
<td>Variable 2</td>
<td>Placebo</td>
<td>IFN-β-1b</td>
<td>Overall</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------</td>
<td>-----------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r</td>
<td>p-value</td>
<td>r</td>
</tr>
<tr>
<td>Year 1 relapse rate</td>
<td>Year 1 T₂ lesions</td>
<td>0.305</td>
<td>&lt; 0.0001</td>
<td>0.120</td>
</tr>
<tr>
<td>Year 2 relapse rate</td>
<td>Year 2 T₂ lesions</td>
<td>0.198</td>
<td>0.0145</td>
<td>0.319</td>
</tr>
<tr>
<td>Year 3 relapse rate</td>
<td>Year 3 T₂ lesions</td>
<td>0.254</td>
<td>0.0031</td>
<td>0.238</td>
</tr>
<tr>
<td>Year 0-3 relapse rate</td>
<td>Year 0-3 T₂ lesions</td>
<td>0.240</td>
<td>&lt; 0.0001</td>
<td>0.203</td>
</tr>
<tr>
<td>Year 1 EDSS change</td>
<td>Year 1 T₂ lesions</td>
<td>0.130</td>
<td>0.0095</td>
<td>-0.040</td>
</tr>
<tr>
<td>Year 2 EDSS change</td>
<td>Year 2 T₂ lesions</td>
<td>0.141</td>
<td>0.0100</td>
<td>0.121</td>
</tr>
<tr>
<td>Year 3 EDSS change</td>
<td>Year 3 T₂ lesions</td>
<td>0.211</td>
<td>0.0005</td>
<td>0.124</td>
</tr>
<tr>
<td>Year 0-3 EDSS change</td>
<td>Year 0-3 T₂ lesions</td>
<td>0.208</td>
<td>&lt; 0.0001</td>
<td>0.118</td>
</tr>
</tbody>
</table>

Table 9.14 Correlations between clinical indices and activity on the annual T₂ scans
9.6 Discussion

This is the first placebo-controlled trial to report the effect of a IFN-β-1b preparation on the evolving pathological process in secondary progressive MS. It is also by far the largest MS study to date in which MR analysis has been performed. It reveals that IFN-β-1b has a substantial effect on the pathological evolution of disease in patients with secondary progressive MS, and that this effect was sustained for the duration of the trial (up to 3 years). This treatment effect was readily apparent both for the total volume of brain pathology visible on the T₂ weighted images, and for the numbers of active T₂ lesions appearing during the study period. Furthermore, statistically significant, though generally modest, correlations were observed between a number of MRI and clinical measures of disease activity and progression. The data strongly support the clinical evidence for efficacy demonstrated in this trial (European Study Group on IFN-β-1b in SPMS 1998). The results are now discussed in more detail, first considering the MRI findings per se and secondly, the clinical/MRI correlations.

9.6.1 MRI findings

The baseline T₂ lesion volume was similar and generally large in both treatment arms of the study; mean and median T₂ lesion volume values were approximately twice those reported in previous trials involving cohorts with relapsing remitting disease. This is consistent with the results presented in Chapter 6 (where analysis of a multicentre database revealed very similar differences between the MS subgroups) and elsewhere (Gawne-Cain et al., 1998). The enormous range of individual T₂ lesion volumes is of course also evident (Table 9.4), as is the modest correlation of baseline T₂ lesion volume with EDSS (r = 0.09, p = 0.002), confirming that, for individual patients with secondary progressive MS, the T₂ lesion volume is not a reliable indicator of clinical status.
In the placebo group, median $T_2$ lesion volume increased significantly year upon year. The overall median percentage increases from baseline (1.6%, 2.4% and 11% at 1, 2 and 3 years respectively) are somewhat lower than the 5-10% per annum changes typically reported in relapsing remitting MS placebo cohorts (Paty et al., 1993; IFNB MS Study Group 1995; PRISMS 1998; Simon et al., 1998). This largely reflects the much higher baseline $T_2$ lesion volume in the secondary progressive cohort, and the absolute increase in $T_2$ lesion volume is similar to that found in early relapsing remitting cohorts. A more or less linear increase in $T_2$ lesion volume from year to year might have been anticipated, but in fact a greater increase was seen in the third year (Table 9.6). Between years 2 and 3, several sites changed or upgraded their MRI scanner, with an increase in field strength and a noticeable improvement in image quality. Such changes can certainly have an important impact on measured $T_2$ lesion volume (Filippi et al., 1997a), and this probably explains the year 2-3 result. However, any such effect would apply equally to both treatment arms (due to the randomisation schedule, treatments were balanced within centres), and the difference in outcome between the treatment groups was still clearly evident: thus during year 3 the median change in $T_2$ lesion volume was an increase of 1.05 cm$^3$ in the placebo group and 0.37 cm$^3$ in the IFN-β-1b group ($p = 0.0002$).

In the treated group, there was a significant and substantial reduction in $T_2$ lesion volume seen at the first follow up. The same phenomenon has been reported in two previous studies in relapsing remitting MS (Paty et al., 1993; PRISMS et al., 1998). This is likely to reflect both; (i) the striking resolution on $T_2$ weighted images that is often seen in inflammatory/oedematous lesions of recent origin, even without therapeutic intervention, and (ii) the effect of treatment on inhibiting new lesion development. It is also possible that IFN-β-1b treatment facilitates a more rapid and complete resolution of such active lesions, by suppressing inflammation and creating...
a more favourable environment for effective repair, although, as yet there is no evidence for this. In the IFN-β-1b group, the effect of resolution of existing inflammatory/oedematous lesions clearly outweighed the increase in T<sub>2</sub> lesion volume due to the small number of new lesions that were identified despite treatment. However, this reduction in T<sub>2</sub> lesion volume was not seen beyond the first annual follow up, which indicates a more transient mechanism, most likely the natural resolution of pre-existing inflammatory lesions.

The treatment effect on the annual assessment of T<sub>2</sub> lesion activity was significant and sustained throughout the study. IFN-β-1b treatment was associated with a 57% reduction in the mean and 80% reduction in the median number of new/enlarging lesions compared to placebo over the duration of the study. A similar level of treatment effect was apparent in each of the three years of the study and this effect is at least as powerful as that previously demonstrated in relapsing remitting MS cohorts (Paty et al., 1993; Simon et al., 1998; PRISMS Study Group 1998). This suggests that the inhibition of new lesion development with IFN-β-1b treatment is common to both relapsing remitting and secondary progressive MS subgroups.

The correlation between T<sub>2</sub> lesion volume and T<sub>2</sub> lesion activity in the placebo group was significant but modest (r = 0.36 over the whole study period). This is not perhaps surprising, given that the net change in T<sub>2</sub> lesion volume represents a composite of increase due to new lesion formation, together with decrease due to lesion resolution. Furthermore, the impact of new lesions on the change in T<sub>2</sub> lesion volume is not linear over time, with activity shortly before the exit scan having a greater impact than earlier activity (see Chapter 6). A recent study has also confirmed that the net change in T<sub>2</sub> lesion volume substantially underestimates volume of new T<sub>2</sub> lesions that have developed, due to resolution in other areas of T<sub>2</sub> hyper-intensity (Lee et al., 1998). This
effect, together with the inevitable attenuation of correlations due to measurement error (both in qualitative and quantitative analysis), probably accounts for the only modest overall correlation between the changes in $T_2$ lesion volume and $T_2$ lesion activity. It is also notable that the correlation between these outcomes for the whole study period is weaker in the IFN-$\beta$-1b group ($r = 0.15$, $p = 0.001$), and this was most apparent when the first 12 months of change is considered ($r = 0.09$, $p = 0.067$); this dissociation is likely to reflect the initial treatment effect of reducing $T_2$ lesion volume in the face of ongoing (albeit diminished) new lesion activity.

When the relationship between baseline $T_2$ lesion volume and subsequent $T_2$ lesion volume change was studied, an interesting difference was found between the placebo and treated groups (Table 9.12). In the placebo group, there was a positive correlation between baseline $T_2$ lesion volume and both absolute change in $T_2$ lesion volume over the next 3 years ($r = 0.18$, $p < 0.0001$) and the cumulative number of new/enlarging lesions ($r = 0.18$, $p < 0.0001$). This may be a reflection of the fact that those patients with a larger $T_2$ lesion volume at study entry had accumulated abnormality more rapidly than those with a lower $T_2$ lesion volume prior to study entry, and this correlation may therefore merely be a manifestation of this effect continuing on into the study.

In contrast, in the IFN-$\beta$-1b group there was a negative correlation between baseline $T_2$ lesion volume and subsequent $T_2$ lesion volume change ($r = -0.16$, $p = 0.0003$). It is likely that patients with higher $T_2$ lesion volumes will also have a higher load of inflammatory/oedematous lesions with the potential to resolve, and that this accounts for the greater decrease in $T_2$ lesion volume, which was in fact fully apparent at the first year of follow up.
9.6.2 Clinical/MRI correlations

A significant, albeit modest, correlation was found between the change in T₂ lesion volume and EDSS over the study period and this was equally apparent in both the placebo and IFN-β-1b groups (Table 9.13). Thus, treatment did not result in a dissociation such that the profound MRI related treatment effect was no longer concordant with the clinical benefit. Slightly stronger correlations were apparent between the annual T₂ lesion activity and EDSS change. These correlations support the use of such MRI measurements as clinically relevant markers of disease evolution in secondary progressive MS. Notwithstanding, it is also apparent that the magnitude of the treatment effect on MRI and clinical outcomes is quantitatively different (the more pronounced effects being seen on MRI), and the modest degree of clinical/MRI correlation suggests that other pathological processes than those studied using T₂ weighted brain imaging contribute to the evolution of disability. The major factors likely to contribute to this clinical/MRI dissociation have been discussed in Chapter 1.

The nature of the EDSS scale also may contribute to the limited relationship between this parameter and T₂ lesion volume. It is noticeable that in a year by year cross-sectional analysis of the correlation between T₂ lesion volume and EDSS, the level correlation increased with each year of follow up (at baseline: r = 0.09, p = 0.002; at 3 year follow up: r = 0.17, p < 0.0001). The initial EDSS range at entry was relatively narrow - from 3 to 6.5 - but increased during follow up as some, but not all, patients accrued more disabilities. As has been previously noted (Gawne-Cain et al., 1998), a greater range of disabilities as indicated by a widened range of EDSS measurements provides a better opportunity for appreciating the MRI-EDSS relationship.

Another factor which theoretically might determine subsequent progression in disability is the
extent of pre-existing irreversible pathology; areas of chronic persistent demyelination might be an unsuitable environment for maintaining axonal viability; axonal degeneration may eventually occur even in the absence of an ongoing active pathological process. In this study, there was a weak overall correlation between baseline T₂ lesion volume and EDSS change over the next 3 years (r = 0.067, p = 0.020), implying that T₂ lesion volume is modestly predictive of subsequent clinical progression in established MS. This suggests that existing disease load in secondary progressive MS may contribute, to a minor extent, to future clinical evolution. The situation is somewhat different earlier in the disease course, where T₂ weighted imaging can be powerfully predictive of clinical evolution (Filippi et al., 1994; O'Riordan et al., 1998). This discrepancy supports the concept that pathological changes other than those defined by T₂ weighted imaging become increasingly important contributors to clinical progression later in the disease course.

Significant correlations were also identified between relapse rate and T₂ lesion activity for all annual timepoints and over the whole study period (Table 9.14). This is not perhaps surprising, given; (i) the well known relationship between gadolinium enhancing lesion activity and relapse rate, and (ii) the fact that a substantial proportion of gadolinium enhanced lesions are also seen on corresponding T₂ weighted images. It is clear, however, that a monthly gadolinium enhanced imaging protocol is more sensitive to short term disease activity than monthly T₂ weighted imaging (Miller et al., 1993). Furthermore, given the natural resolution of new T₂ lesions over a period of weeks to leave a small residual abnormality, an additional proportion of new T₂ lesions will not be visible with an inter-scan interval of 12 months. Therefore, visual analysis of annual T₂ weighted imaging has a limited role as a marker of short term disease activity. However, the results suggest an important role as a marker of disease progression. The fact that treatment effect has been demonstrated so powerfully with simple visual assessment of serial T₂ weighted images,
with clinical/MRI correlations that are least as strong as for T\textsubscript{2} lesion volume quantification, raises a question of the need for T\textsubscript{2} lesion volume analysis in such a study. The logistic requirements of electronic data acquisition, transfer of electronic data to a coordinating centre, and subsequent quantitative analysis are enormous. In contrast, simple visual assessment on hard copy is vastly less logistically demanding, requiring only 10-15 minutes per pair of MRI studies. Given that the most important criteria of any surrogate are objectivity, efficiency and clinical predictive value, it could readily be argued that qualitative annual T\textsubscript{2} activity analysis is a more appropriate outcome measure than T\textsubscript{2} lesion volume quantification. This is further supported by the MRI results of the Avonex IFN-\(\beta\)-1a study (Simon et al., 1998), which found a powerful treatment effect on T\textsubscript{2} lesion activity (supporting the clinical outcome data), without any significant effect on T\textsubscript{2} lesion volume.

In conclusion, the MRI results of this study support and supplement the clinical efficacy endpoint measures. Highly significant and powerful treatment effects have been identified for both qualitative and quantitative analysis of the annual MRI studies. The results also suggest that the existence of significant clinical/MRI correlations is common to both secondary progressive and relapsing remitting MS, extending previous observations on relapsing remitting cohorts (Paty et al., 1993; IFNB MS Study Group 1995) to the later secondary progressive phase of the disease course. These results may also guide future phase III trial design, by highlighting the utility of simple visual assessment of T\textsubscript{2} weighted MR studies, and suggesting a greater role for this outcome in future phase III studies. However, the modest nature of the clinical/MRI correlations confirm the appropriate role of clinical endpoints as primacy efficacy variables, and highlight the need for newer MRI tools.
Conclusions

The results of recent phase III MS trials offer the exciting prospect of altering long term outcome in MS. These trials have consistently used MR outcome measures to both support and supplement the primary clinical endpoints, and highly significant MR treatment effects have been demonstrated with several different IFN preparations, with further studies currently underway.

The results presented in Part One of this thesis suggest that contouring is the current gold standard quantitative technique, offering a substantial advantage over manual outlining in terms of precision. Furthermore, contouring has now been validated in the context of the Phase III trial, as detailed in Part Three of this thesis. At present, the global thresholding technique is not sufficiently robust to be applied to serial MRI datasets without substantial manual editing, and therefore its use in a treatment trial context cannot be recommended. The results presented in Chapter 5 suggest that for serial MRI studies, a slice thickness of 3 mm provides an optimal compromise between reproducibility and lesion detection on the one hand, and time consumption on the other, when using semi-automated techniques such as contouring. However, even with a less sensitive, standard 5 mm slice thickness, the sample size estimations presented in Chapter 7 highlight the fact that routine incorporation of the entire trial population to an annual imaging protocol results in substantial overpowering. It could be argued that a more efficient and appropriate design would be to scan only a representative subgroup of patients, restricting MR imaging to those centres able to provide the highest imaging quality and where upgrades will not occur over the study duration.

The significant correlations between change in $T_2$ lesion volume and monthly enhanced imaging presented in Part Two are reassuring, confirming that annual change in $T_2$ lesion volume does...
provide a reasonable reflection of ongoing disease activity in the short term. However, it also appears that the timing of such activity is important in influencing the magnitude of the measured change, with activity shortly before the exit scan having the greatest impact. This supports the current practice of performing monthly imaging in a subgroup of patient in Phase III studies, to avoid the possibility of a non-sustained treatment effect going undetected.

The MR results of the Phase III study of IFN-β-1b provide powerful evidence of therapeutic effect, the predominant mechanism of which is likely to be BBB protection, as evidenced by the effect of treatment on annual T₂ lesion activity. However, the study again highlights the modest nature of the clinical/MR relationship, re-affirming the status of clinical indices as primary endpoint measures. It could also be argued that simple visual assessment of annual T₂ datasets provides a similar level of clinical/MR correlation to that found with annual T2 lesion volume assessment, with substantially greater efficiency, although current consensus guidelines still highlight the importance of quantitative measurement.

A major limitation of T₂ weighted MRI is its lack of pathological specificity. A number of putative MR tools with greater specificity for tissue destruction are now demonstrating stronger clinical/MRI correlations than standard T₂ weighted imaging, and they are likely to play an increasing role in future studies. However, at present serial T₂ weighted MRI is the most established and validated MR tool in Phase III MS studies.
References


Barbosa S, Blumhardt LD, Roberts N., et al. Magnetic resonance relaxation time mapping in


Capra R, Marciano N, Vignolo LA., et al. Gadolinium-pentetic acid magnetic resonance imaging

Chabot S, Williams G, Yong VW. Microglial production of TNF-alpha is induced by activated T lymphocytes. Involvement of VLA-4 and inhibition by interferon beta-1b. J Clin Invest 1997;100:604-612

Charcot JM. Histologie de la sclérose en plaques. Gaz Hôp 1868;41:554-566


Constable RT, Gore JC. The loss of small objects in variable TE imaging: implications for FSE, RARE and EPI. Mag Reson Imaging 1992;28:9-24


Corsini E, Gelati M, Dufour., et al. Effects of beta-IFN-1b treatment in MS patients on adhesion between PBMNCs, HUVECs and MS-HBECs: an in vivo and in vitro study. J Neuroimmunol 1997;79:76-83

180
Cousens LP, Orange JS, Su HC, Biron CA. Interferon-alpha/beta inhibition of interleukin 12 and interferon-gamma production in vitro and endogenously during viral infection. Proc Nat Acad Sci USA 1997;94:634-9


181


Fleiss JL, ed. The design and analysis of clinical experiments. New York: John Wiley and Sons, 1985


186


187

Goodkin DE. EDSS reliability (letter). Neurology 1991;41:332


Implications. Brain 1990a;113:1477-1489


Lauterbur PC. Image formation by induced local interactions: examples employing nuclear


Martin R & McFarland HF. Immunology of Multiple Sclerosis and experimental allergic encephalomyelitis. In Multiple Sclerosis. Clinical and pathogenetic basis (eds Raine RC,
McFarland HF and Tourtellotte WW), London, Chapman and Hall 1997;221-242


Miller DH, Hammond SR, McLeod JG., et al. Multiple sclerosis in Australia and New Zealand: are the determinants genetic or environmental? J Neurol Neurosurg Psychiatry 1990;53:903-905


198


Ormerod IE, Miller DH, McDonald WL., et al. The role of NMR imaging in the assessment of


Paty DW, Li DKB, the UBC MS/MRI Study Group, the IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing remitting MS. MRI analysis results of a multi-center, randomised, double blind, placebo controlled trial. Neurology 1993;43:662-667

Plante E, Turkstra L. Sources of error in the quantitative analysis of MRI scans. Mag Res Imaging 1991;9:589-595

Plummer DL. Dispimage: a display and analysis tool for medical images. Rivista di Neuroradiologia 1992;5:489-495


203


Simmons A, Arridge SR, Barker GJ, Williams SCR. Simulation of MRI cluster plots and application to neurological segmentation Mag Reson Imag 1996;14:73-92


Streiner DL & Norman GR. Reliability; edited by Streiner DL and Norman GR. 1995; p104-127


Thorpe JW, Barker GJ, Jones SJ., et al. Magnetisation transfer ratios and transverse magnetisation
decay curves in optic neuritis: correlation with clinical findings and electrophysiology. J Neurol Neurosurg Psychiatry 1995;59:487-492


Truyen L, van Waesberghe JH, van Walderveen., et al. Accumulation of hypointense lesions
('black holes') on T1 spin-echo MRI correlates with disease progression in multiple sclerosis.
Neurology 1996;47:1469-1476


Uhlenbrock D & Sehlen S. The value of T1 weighted images in the differentiation between MS, white matter lesions, and subcortical arteriosclerotic encephalopathy (SAE). Neuroradiology 1989;31:203-212


210


Wingerchuk DM, Noseworthy JH, Weinshenker BG. Clinical outcome measures and rating scales

Wolff SD, Balaban RS. Magnetisation transfer contrast (MTC) and tissue water proton relaxation in vivo. Magn Reson Med 1989;10:135-144

Wolff SD, Balaban RS. Magnetization transfer imaging: practical aspects and clinical applications. Radiology 1994;192:593-599


212