Research

Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review
Penny Whiting, Roger Harbord, Caroline Main, Jonathan J Deeks, Graziella Filippini, Matthias Egger, Jonathan A C Sterne

Abstract

Objective To determine the accuracy of magnetic resonance imaging criteria for the early diagnosis of multiple sclerosis in patients with suspected disease.

Design Systematic review.

Data sources 12 electronic databases, citation searches, and reference lists of included studies.

Review methods Studies on accuracy of diagnosis that compared magnetic resonance imaging, or diagnostic criteria incorporating such imaging, to a reference standard for the diagnosis of multiple sclerosis.

Results 29 studies (18 cohort studies, 11 other designs) were included. On average, studies of other designs (mainly diagnostic case-control studies) produced higher estimated diagnostic odds ratios than did cohort studies. Among 15 studies of higher methodological quality (cohort design, clinical follow-up as reference standard), those with longer follow-up produced higher estimates of specificity and lower estimates of sensitivity. Only two such studies followed patients for more than 10 years. Even in the presence of many lesions (>10 or >8), magnetic resonance imaging could not accurately rule multiple sclerosis in (likelihood ratio of a positive test result 3.0 and 2.0, respectively). Similarly, the absence of lesions was of limited utility in ruling out a diagnosis of multiple sclerosis (likelihood ratio of a negative test result 0.1 and 0.5).

Conclusions Many evaluations of the accuracy of magnetic resonance imaging for the early detection of multiple sclerosis have produced inflated estimates of test performance owing to methodological weaknesses. Use of magnetic resonance imaging to confirm multiple sclerosis on the basis of a single attack of neurological dysfunction may lead to over-diagnosis and overtreatment.

Introduction

Diagnosis of multiple sclerosis is based on the principle of dissemination in both time and space. Recent criteria state that patients should experience two attacks of neurological dysfunction, such as optic neuritis, transverse myelitis, double vision, or numbness and tingling of the leg, occurring at different points in time and affecting different parts of the central nervous system—that is, signs or symptoms that cannot be attributable to a single lesion.\(^1\) Many years may elapse between first and second attacks, and not all patients who experience a first attack develop multiple sclerosis. In a study of patients with optic neuritis, a common presenting symptom of multiple sclerosis, 38% developed the disease by 10 years; of these, 50% received their diagnosis more than three years after presentation and 28% more than five years after presentation.\(^2\) In a study of patients presenting with clinically isolated syndromes (optic, spinal cord, or brain symptoms) 68% of patients had developed multiple sclerosis by 14 years, the proportions being similar for the different presenting symptoms.\(^3\)

Magnetic resonance imaging may assist in earlier diagnosis of multiple sclerosis by enabling visualisation of lesions in the brain that are clinically silent. The McDonald 2001 criteria for the diagnosis of multiple sclerosis allow an early diagnosis of multiple sclerosis to be made after one clinical attack if the patient also meets criteria for a positive result on a magnetic resonance imaging scan. The McDonald criteria have been adopted in England and Wales by the National Institute for Health and Clinical Excellence (NICE),\(^4\) but they are not universally accepted.\(^5\) Evidence shows that patients’ wellbeing is affected by early diagnosis,\(^6\) usually in a beneficial way but also occasionally in a negative way—for example, through increased insurance premiums and discrimination in the workplace.\(^7\) Earlier diagnosis of multiple sclerosis could mean the availability of earlier treatment, such as the disease modifying therapies interferon beta and glatiramer acetate, provided under the “risk sharing scheme” in the United Kingdom (www.dh.gov.uk/assetRoot/04/01/22/14/04012214.pdf).

We carried out a systematic review to estimate the accuracy of different magnetic resonance imaging criteria for the early diagnosis of multiple sclerosis in patients presenting with suspected disease, to investigate whether magnetic resonance imaging has the potential to alter diagnoses and patient management.

Methods

We identified studies, published and unpublished, by searching 12 databases from inception until September or November 2004. Search terms were “multiple sclerosis” combined with “magnetic resonance imaging” or “MRI”. No language restrictions were applied. We undertook a citation search on the article reporting the McDonald 2001 criteria,\(^4\) screened reference lists of included studies, and assessed studies included in the NICE multiple sclerosis guidelines.\(^8\)

Studies were eligible that compared magnetic resonance imaging (or diagnostic criteria incorporating such imaging) to a reference standard for the diagnosis of multiple sclerosis and reported sufficient data to enable a 2x2 table of test
performance to be constructed. If studies were reported more than once, we included the publication that provided data for the longest follow-up. We also included separate publications that reported on different criteria for magnetic resonance imaging or separate results for relevant patient subgroups.

Two reviewers independently screened titles and abstracts for relevance. Screening for inclusion, data extraction, and quality assessment were carried out by one reviewer and checked by a second. Studies were assessed for methodological quality against the QUADAS (quality assessment of diagnostic accuracy studies) criteria.11 (See bmj.com for a summary of how items were scored.) One item, the avoidance of disease progression bias, was omitted as it was not relevant to this topic. We grouped studies according to patient spectrum: prospective cohort studies that enrolled patients with suspected multiple sclerosis, and studies of other designs.

Data analysis
From each 2×2 table we computed sensitivity, specificity, and likelihood ratios, which combine data on sensitivity and specificity to give an indication of a test’s ability to rule in or rule out a condition.12 We plotted all results from all included studies on a receiver operating characteristic plot of sensitivity against specificity, with the specificity axis reversed. To compare accuracy of cohort and other studies we selected the result with the median diagnostic odds ratio (defined as the odds of positivity among people with the disease, divided by the odds of positivity among people without the disease) for each study. We used random effects meta-analysis to obtain summary diagnostic odds ratios in each group, and we carried out a permutation test13 to obtain a P value for their comparison. We restricted all further analyses to cohort studies that used a reference standard diagnosis of clinically definite multiple sclerosis, arrived at solely by clinical data.

As a final diagnosis of multiple sclerosis may be reached many years after a patient first presents with possible disease, we investigated the effect of duration of follow-up on estimates of diagnostic accuracy. We used the hierarchical summary receiver operating characteristic method proposed by Rutter and Gatsonis14 to assess the effect of duration of follow-up on overall accuracy and threshold. An association with threshold would indicate that sensitivity increased as specificity decreased, or vice versa. We drew separate receiver operating characteristic plots for each of these studies and compared areas under the curves. The statistical software package Stata release 9 was used for all analyses, except the hierarchical summary receiver operating characteristic model, which was fitted in SAS.15

Results
Figure 1 shows the flow of studies through the review. Sixty one publications met the inclusion criteria and were included in the final evidence, of which were earlier reports of included studies and were not extracted.16 Forty publications reporting the results of 29 studies (some reported results for different magnetic resonance imaging criteria, for imaging of the spine rather than the brain, or for patient subgroups) were included. Sample sizes were generally small (median 70), ranging from 15 to 1500 patients. The proportions of dropouts ranged from 0 to 58% (median 4%), increasing with length of follow-up. Table 1 provides details of the 29 publications reporting the results of 18 cohort studies. Most of these studies used clinical follow-up as the reference standard. Most used the Poser criteria,16 although some used the McDonald 1977 criteria.17 The McDonald 1977 criteria, based on clinical information alone, are not the same as the McDonald 2001 criteria, which incorporate magnetic resonance imaging.1 Table 2 provides details of the 11 studies of other designs. The studies differed according to population, quality, magnetic resonance imaging protocol, and criteria used to define a positive test result. Cohort studies varied in their inclusion criteria; some included only patients presenting with a particular clinically isolated syndrome (for example, optic neuritis or a spinal cord syndrome), whereas others included all patients being evaluated for possible multiple sclerosis. Publication dates ranged from 1986 to 2003. Over this time improvements occurred in magnetic resonance imaging technology; this is reflected in differences in scanning protocols (see table A on bmj.com).

Figure 2 summarises the results of the quality assessment (see table B on bmj.com for results of individual studies). Study quality was generally poor: only four QUADAS items were met by over 70% of studies (avoidance of partial and differential verification bias and reporting of uninterpretable results and withdrawals). Studies scored badly on three items: blinding, the use of an appropriate reference standard, and the availability of clinical information. Four publications, reporting results from three cohort studies, were susceptible to incorporation bias as magnetic resonance imaging contributed to the final diagnosis.16–18 Three of these used a combination of clinical follow-up and paraclinical tests as the reference standard,19–21 the other relied on paraclinical tests alone.18 All other cohort studies used clinical follow-up alone as the reference standard.

Figure 3 shows that cohort studies produced lower estimated sensitivity and specificity than studies of other designs. The pooled diagnostic odds ratio was 9 (95% confidence interval 5 to 16) for cohort studies and 213 (85 to 535) for studies of other designs (P<0.001, permutation test). Further analysis was restricted to the 15 cohort studies that used a diagnosis of clinically definite multiple sclerosis, arrived at by clinical information alone, as the reference standard.

The average duration of follow-up ranged from seven months to 14 years. The only criteria for which sufficient data were available to investigate the effects of duration of follow-up were presence of one or more lesions and presence of one or
Table 1  Study details and results of cohort studies recruiting patients with suspected multiple sclerosis (MS)

<table>
<thead>
<tr>
<th>Author (year), country</th>
<th>No of patients* (% female)</th>
<th>Mean (range) age (years)</th>
<th>Presenting symptoms</th>
<th>Criteria for MS</th>
<th>Mean follow-up (range) in months</th>
<th>% with MS</th>
<th>Sensitivity, specificity (%)</th>
<th>Positive, negative likelihood ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee (1991)Canada 1</td>
<td>184 (67)</td>
<td>43 (12-79) NR</td>
<td>Schumacher*</td>
<td>NR</td>
<td>25 (NR)</td>
<td>30</td>
<td>95, 49</td>
<td>1.8, 0.1</td>
</tr>
<tr>
<td>Ford (1992)Canada 2</td>
<td>15 (53)</td>
<td>33 (22-44) Spinal cord lesion (n=15)</td>
<td>Poser*</td>
<td>38.5 (NR)</td>
<td>80</td>
<td>≥1 non-clinical lesions</td>
<td>88, 50</td>
<td>2.0, 0.3</td>
</tr>
<tr>
<td>Reese (1986)Canada 3</td>
<td>47 (NR)</td>
<td>NR</td>
<td>NR</td>
<td>Diagnosed by referring neurologist (no details)</td>
<td>95</td>
<td>68</td>
<td>≥1 T2 lesions</td>
<td>5.5, 0.2</td>
</tr>
<tr>
<td>Frederiksen (1989)Denmark</td>
<td>60 (70)</td>
<td>31 (12-53) Optic lesion (n=60) NR</td>
<td>Medium 11 (1-28)</td>
<td>91, 87</td>
<td>100</td>
<td>≥1 non-clinical lesions</td>
<td>1.7, 0.1</td>
<td></td>
</tr>
<tr>
<td>Brem (2002)England 1</td>
<td>71 (62)</td>
<td>32 (13-49) Optic lesion (n=38), spinal cord lesion (n=21), brainstem lesion (n=14)</td>
<td>Poser*</td>
<td>168 (150-202)</td>
<td>68</td>
<td>≥1 non-clinical T2 lesions</td>
<td>92, 74</td>
<td>3.4, 0.1</td>
</tr>
<tr>
<td>Filipp (1994)England 1</td>
<td>84 (70)</td>
<td>31 (13-50) Optic lesion (n=40), spinal cord lesion (n=28), brainstem lesion (n=16)</td>
<td>NR</td>
<td>63 (43-84)</td>
<td>40</td>
<td>≥1 non-clinical lesions &gt;1.23 cm²</td>
<td>53, 94</td>
<td>7.7, 0.5</td>
</tr>
<tr>
<td>Miller (1988)England 1</td>
<td>53 (62)</td>
<td>33 (16-48) Optic lesion (n=69)</td>
<td>NR</td>
<td>12 (5-30)</td>
<td>28</td>
<td>≥1 non-clinical T2 lesions</td>
<td>80, 42</td>
<td>1.4, 0.5</td>
</tr>
<tr>
<td>Miller (1989)England 1</td>
<td>56 (63)</td>
<td>32 (13-49) Spinal cord lesion (n=33) Brainstem lesion (n=23)</td>
<td>NR</td>
<td>14.7 (5-31)</td>
<td>30</td>
<td>≥1 non-clinical T2 lesions</td>
<td>90, 61</td>
<td>2.2, 0.2</td>
</tr>
<tr>
<td>Sharief (1991)England 1</td>
<td>45 (62)</td>
<td>37.3 (&lt;50) Spinal cord lesion (n=20) Brainstem lesion (n=23)</td>
<td>NR</td>
<td>18 (NR)</td>
<td>49</td>
<td>Paty</td>
<td>77, 70</td>
<td>2.4, 0.3</td>
</tr>
<tr>
<td>Brem (2001)England 1</td>
<td>68 (57)</td>
<td>31 (16-50) Optic lesion (n=46), spinal cord lesion (n=6), brainstem lesion (n=16)</td>
<td>Poser*†</td>
<td>&gt;12 (NR)</td>
<td>26</td>
<td>≥1 non-clinical T2 lesions</td>
<td>69, 36</td>
<td>1.4, 0.4</td>
</tr>
<tr>
<td>Sharief (1991)England 1</td>
<td>50 (57)</td>
<td>30 (16-49) Optic lesion (n=46), spinal cord lesion (n=6), brainstem lesion (n=16)</td>
<td>Poser*†</td>
<td>12 (12-19)</td>
<td>26</td>
<td>≥1 non-clinical T2 lesions</td>
<td>77, 70</td>
<td>2.5, 0.4</td>
</tr>
<tr>
<td>Dalton (2003)England 2</td>
<td>56 (54)</td>
<td>32 (17-50) Optic lesion (n=37), spinal cord lesion (n=5), brainstem lesion (n=14)</td>
<td>Poser*†</td>
<td>NR (&gt;30)</td>
<td>34</td>
<td>New T2 lesion at three months</td>
<td>84, 89</td>
<td>7.0, 0.2</td>
</tr>
</tbody>
</table>

*No of patients includes patients recruited at the study entry who had clinical or paraclinical symptoms fulfilling the criteria for lesion in later follow-up studies.

**MRI criteria:**
- McDonald 2001 criteria at three months
- McDonald 2001 criteria or new T2 lesions at three months

---

\* Poser criteria, Paty criteria
\*† Poser criteria, paraclinical data contributed to diagnosis

---

BMJ Online First bmj.com  page 3 of 10
BMJ: first published as 10.1136/bmj.387771.583796.7C on 24 March 2006. Downloaded from http://www.bmj.com/ on 2 October 2023 by guest. Protected by copyright.
<table>
<thead>
<tr>
<th>Author (year), country</th>
<th>No of patients* (% female)</th>
<th>Mean (range) age (years)</th>
<th>Preceding symptoms</th>
<th>Criteria for MS</th>
<th>Mean follow-up (range) in months</th>
<th>% with MS</th>
<th>MRI criteria</th>
<th>Sensitivity, specificity (%)</th>
<th>Positive, negative likelihood ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dalton (2002)w15, England 2</td>
<td>50 (59)</td>
<td>31 (16-50)</td>
<td>Optic lesion (n=99), spinal cord lesion (n=22), brainstem lesion (n=12)</td>
<td>Poserw4</td>
<td>37 (29-67)</td>
<td>38</td>
<td>≥1 non-clinical lesions (brain or spinal cord)</td>
<td>100, 35</td>
<td>1.5, 0.1</td>
</tr>
<tr>
<td>Barkhof (1997)w16, Holland 1</td>
<td>74 (NR)</td>
<td>NR</td>
<td>Optic lesion (n=40), spinal cord lesion (n=22), brainstem lesion (n=12)</td>
<td>Poserw4</td>
<td>39 (23-69)</td>
<td>45</td>
<td>≥1 T2 lesions</td>
<td>97, 29</td>
<td>1.4, 0.1</td>
</tr>
<tr>
<td>Filippini (1994)w21, Italy 4</td>
<td>82 (65)</td>
<td>28 (14-51)</td>
<td>Optic lesion (n=21)</td>
<td>McDonald 1977w20</td>
<td>35 (NR)</td>
<td>34</td>
<td>≥1 MS-like abnormalities</td>
<td>96, 44</td>
<td>1.7, 0.1</td>
</tr>
<tr>
<td>Sastre-Garriga (2003)w23, Spain 1</td>
<td>51 (63)</td>
<td>29 (14-49)</td>
<td>Brainstem lesion (n=51)</td>
<td>Fazekas</td>
<td>77, 51</td>
<td>1.6, 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rio (1997)w26, Spain 2</td>
<td>35 (80)</td>
<td>31 (17-47)</td>
<td>Optic lesion (n=35)</td>
<td>NH</td>
<td>29.4 (12-66)</td>
<td>20</td>
<td>≥1 T2 lesions</td>
<td>100, 67</td>
<td>2.7, 0.1</td>
</tr>
<tr>
<td>Soderstrom (1998)w27, Sweden</td>
<td>147 (80)</td>
<td>34 (12-57)</td>
<td>Optic lesion (n=147)</td>
<td>Poserw4</td>
<td>25 (0-71)</td>
<td>41</td>
<td>≥1 lesions</td>
<td>85, 65</td>
<td>2.4, 0.2</td>
</tr>
</tbody>
</table>
resonance imaging alone.

within three years than any of the criteria based on magnetic

potential for predicting the development of multiple sclerosis

ria ranged from 2.7 to 8.7, suggesting that they have more

predicting the development of multiple sclerosis within three to

England,

BMJ

Online First bmj.com page 5 of 10

§Patients with clinically probable or possible MS were classified as having the disease.

‡Interquartile range.

†Patients with clinically probable MS were classified as having the disease.

*Total number who entered trial (includes withdrawals); this may differ from sum of number of patients with each of presenting symptoms as for most studies these were reported only for patients who completed the study.

MRI=magnetic resonance imaging; NR=not reported.

more non-clinical lesions. Figure 4 is a receiver operating characteristic plot for these criteria, with numbers showing the duration of follow-up in years. Evidence shows (P = 0.074 from hierarchical summary receiver operating characteristic analysis) that studies with longer follow-up produced higher estimated specificity and lower estimated sensitivity.

The longest average duration of follow-up was three years in studies that assessed the Barkhof, Fazekas, and McDonald 2001 criteria, and six years for studies that assessed the Paty criteria. It is therefore possible to draw conclusions regarding the ability of these criteria to predict the development of multiple sclerosis only over these relatively short periods. Figure 5 shows the receiver operating characteristic plots for these criteria. The study that developed the Barkhof criteria† showed higher estimated sensitivity and specificity than did the other studies of this criterion. The negative likelihood ratios for the Barkhof, Fazekas, and Paty criteria ranged from 0.2 to 0.5, suggesting that the curves suggested better accuracy in the English study than in the US study (P = 0.045). Estimates of the positive likelihood ratio ranged from the presence of various numbers of lesions ranged from 2.0 to 3.4. Assuming a pretest probability of multiple sclerosis of 60% this is equivalent to a post-test probability of 75%-84%, suggesting that magnetic resonance imaging is of limited utility for ruling in multiple sclerosis at any threshold. Estimates of the negative likelihood ratio ranged from 0.1 to 0.9 but were greater than 0.5 for all but one of the thresholds in the English study. This is equivalent to modifying a pretest probability of 60% to give a post-test probability of multiple sclerosis of 43%-57%, suggesting that magnetic resonance imaging is also of limited utility in ruling out a diagnosis of multiple sclerosis.
<table>
<thead>
<tr>
<th>Author (year), country</th>
<th>Study design</th>
<th>Study details and results of case-control studies and studies of other designs</th>
<th>No of patients* (% female)</th>
<th>Mean age (range) in years</th>
<th>No of patients with confirmed MS</th>
<th>No with other conditions</th>
<th>No of healthy volunteers</th>
<th>Criteria for MS</th>
<th>MRI criteria†</th>
<th>Sensitivity, specificity (%)</th>
<th>Positive, negative likelihood ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offenbacher (1993)*, Austria 1</td>
<td>Consecutive scans done for multiple purposes (9% query MS)</td>
<td>1500 (52) 46 (12-93)</td>
<td>0</td>
<td>1251</td>
<td>115</td>
<td>Poser‡</td>
<td>Fazekas</td>
<td>81</td>
<td>96</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>Fazekas (1988)*, Austria 2</td>
<td>Case-control</td>
<td>91 (NR) 46 (14-77)</td>
<td>50</td>
<td>0</td>
<td>49</td>
<td>NR†</td>
<td>Fazekas</td>
<td>88</td>
<td>100</td>
<td>73.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Ravnborg (1992)*, Denmark 2</td>
<td>Cross-sectional study of patients with &gt;1 neurological attacks</td>
<td>68 (56) 40 (18-63)</td>
<td>0</td>
<td>68</td>
<td>0</td>
<td>Poser‡‡</td>
<td>Barkhof</td>
<td>76</td>
<td>92</td>
<td>9.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Van der Eerden (1990)*, Holland 2</td>
<td>Nested case-control</td>
<td>49 (50) 45 (21-77)</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>Schumacher§§</td>
<td>Barkhof</td>
<td>95</td>
<td>65</td>
<td>2.7</td>
<td>0.1</td>
</tr>
<tr>
<td>Bot (2002)*, Holland 3</td>
<td>Case-control</td>
<td>91 (73) 43 (NR)</td>
<td>25</td>
<td>66</td>
<td>0</td>
<td>Poser‡‡</td>
<td>Barkhof</td>
<td>76</td>
<td>92</td>
<td>9.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Rovaris (2000)*, Italy 5</td>
<td>Case-control</td>
<td>123 (67) 40 (NR)</td>
<td>64</td>
<td>59</td>
<td>0</td>
<td>Poser‡‡</td>
<td>Barkhof</td>
<td>80</td>
<td>97</td>
<td>19</td>
<td>0.2</td>
</tr>
<tr>
<td>Kuroda (1995)*, Japan 1</td>
<td>Case-control</td>
<td>72 (71) 44 (13-72)</td>
<td>36</td>
<td>36</td>
<td>0</td>
<td>Schumacher§§</td>
<td>Barkhof</td>
<td>56</td>
<td>94</td>
<td>8.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

* Denotes participants from the original Poser 1987 criteria study.
† Denotes patients with confirmed MS at the end of the study.
‡ Denotes patients with other conditions at the end of the study.
§ Denotes patients with healthy volunteers at the end of the study.

MRI criteria:
- ≥3 areas of high signal intensity
- Paty A
- Paty B
- ≥1 infratentorial lesion
- ≥1 periventricular lesion
- ≥3 lesions
- ≥3 mm
- Three lesions >3 mm
- ≥1 lesions
- Any abnormality

Sensitivity and specificity ranges:
- Sensitivity: 81% - 100%
- Specificity: 73.2% - 98%
- Positive likelihood ratio: 20 - 150
- Negative likelihood ratio: 0.1 - 0.2
Discussion

Use of magnetic resonance imaging to confirm multiple sclerosis on the basis of a single attack of neurological dysfunction may lead to over-diagnosis and over-treatment. Many studies in our systematic review produced inflated estimates of test performance owing to methodological weaknesses.

Only two cohort studies on the accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis included at least 10 years’ follow-up. These suggested that the role of magnetic resonance imaging either in ruling in or ruling out a diagnosis of multiple sclerosis is limited. Studies that did not include an appropriate patient spectrum tended to overestimate both sensitivity and specificity. Studies that included shorter clinical follow-up tended to overestimate sensitivity and underestimate specificity. Specific criteria developed for the interpretation of magnetic resonance imaging scans as indicating multiple sclerosis, the Fazekas, Barkhof, and Paty criteria, have poor accuracy for predicting the development of multiple sclerosis within three to six years. The limited data on the McDonald 2001 criteria suggest that these have some potential to rule in the development of multiple sclerosis within three years. Neither the specific magnetic resonance imaging criteria nor McDonald 2001 were evaluated in studies with long term follow-up. It is therefore not possible to determine their accuracy for the diagnosis of multiple sclerosis.

### Table: Study characteristics

<table>
<thead>
<tr>
<th>Author (year), country</th>
<th>Study design</th>
<th>No of patients* (% female)</th>
<th>Mean age (range) in years</th>
<th>No of patients with confirmed MS</th>
<th>No with other conditions</th>
<th>No of healthy volunteers</th>
<th>Criteria for MS</th>
<th>MRI criteria†</th>
<th>Sensitivity, specificity (%)</th>
<th>Positive, negative likelihood ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palmer (1999)en, USA 3</td>
<td>Case-control</td>
<td>50 (18)</td>
<td>40 (18-70)</td>
<td>6</td>
<td>25</td>
<td>0</td>
<td>NR</td>
<td>Subcortical striations</td>
<td>72, 94</td>
<td>10.6, 0.3</td>
</tr>
<tr>
<td>Gean-Marton (1991)en, USA 4</td>
<td>Case-control</td>
<td>169 (46)</td>
<td>41 (22-87)</td>
<td>47</td>
<td>122</td>
<td>0</td>
<td>Poser†</td>
<td>Focal abnormalities on callosal-septal interface</td>
<td>93, 98</td>
<td>34, 0.1</td>
</tr>
<tr>
<td>Yetkin (1991)en, USA 5</td>
<td>Case-control</td>
<td>260 (NR)</td>
<td>NR (31-83)</td>
<td>92</td>
<td>68</td>
<td>100</td>
<td>Schumacher‡</td>
<td>Abnormal MRI scan: equivocal findings classed as normal</td>
<td>76, 98</td>
<td>17.1, 0.3</td>
</tr>
</tbody>
</table>

MS=multiple sclerosis; MRI=magnetic resonance imaging; NR=not reported.
*Total number of patients who entered trial (includes withdrawals); in some cases this may differ from sum of number of patients with confirmed MS, other conditions, or healthy volunteers as for most studies these were only reported for patients who completed the study.
†Patients with clinically probable MS were classified as having the disease.
‡Patients with clinically probable or possible MS were classified as having the disease.

![Fig 2](image-url) Results of quality assessment for appropriate patient spectrum studies

![Fig 3](image-url) Receiver operating characteristic plots for cohort studies and for studies of other designs
peopple known not to have the disease, either healthy controls or patients with conditions that may present with similar symptoms to multiple sclerosis. That such studies tend to exaggerate the accuracy of magnetic resonance imaging in the diagnosis of multiple sclerosis is to be expected; people with more advanced multiple sclerosis are more likely to have lesions on their magnetic resonance imaging scans than those presenting in the early stages of multiple sclerosis.

Strengths and weaknesses in relation to other studies
Although several reviews have assessed the accuracy of magnetic resonance imaging in the diagnosis of multiple sclerosis, we are unaware of any systematic review. The McDonald 2001 criteria incorporate the Barkhof criteria to define a positive MRI scan. The article reporting the McDonald 2001 criteria refers to a small number of studies to justify its selection of the Barkhof criteria for this purpose. All these had methodological weaknesses: they either used a case-control design or had an average of less than three years clinical follow-up. This paper was published before the two long term cohort studies from England and the United States.

A recently published detailed (but not systematic) review is the report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. This was limited to cohort studies and discussed in detail the problems associated with the lack of a true reference standard, an accurate method of determining whether or not a patient has multiple sclerosis that can be applied at the same time as the index test, for the diagnosis of the disease. It did not carry out any statistical synthesis and instead presents a narrative overview of the results of the English study and several other studies, also included in our review, which had relatively short clinical follow-up, and was published before the US study. It concluded, in contrast with our findings, that the presence of at least three lesions on a magnetic resonance imaging scan is a sensitive predictor of the development of multiple sclerosis in the next 7-10 years, and that normal results suggest that future development of multiple sclerosis is less likely. A more recent review article focused on the McDonald 2001 criteria but also draws on the results of the report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. It highlights the limitations of the evidence base for the McDonald 2001 criteria and draws on the results of the US and English studies to conclude, consistent with the results presented here, that presence of brain lesions does not guarantee development of multiple sclerosis over 10-14 years.

Unanswered questions and future research
The main clinical question is whether magnetic resonance imaging should be included in the work-up of patients with multiple

---

**Fig 4** Receiver operating characteristic plots for studies included in hierarchical summary receiver operating characteristic analysis. Numbers are duration of follow-up in years.

**Fig 5** Receiver operating characteristic plots (95% confidence intervals) for Barkhof, Fazekas, Paty, and McDonald 2001 criteria. In Barkhof plot, red is study that proposed Barkhof criteria. In McDonald 2001 plot, red indicates studies where McDonald 2001 criterion was applied after three months rather than 12 months.
sclerosis. Several factors need to be considered, in particular the reasons why magnetic resonance imaging is ordered. This is not simply to increase the certainty of the diagnosis: other possible reasons include ruling out differential diagnoses such as brain tumours, providing a baseline for monitoring disease progression, patient request, and patient reassurance. If magnetic resonance imaging scans are ordered to inform the diagnosis of multiple sclerosis, and if the McDonald 2001 criteria that incorporate such imaging are to be used in practice, then further research, based on long term cohort studies, is required to evaluate these criteria. A limitation consequent on the need for long term clinical follow-up in studies that evaluate the accuracy of magnetic resonance imaging is that such studies inevitably use older technology. Studies with more advanced, and hence recent, technology inevitably had much shorter periods of follow-up. Differences in estimates of sensitivity and specificity according to magnetic resonance imaging technology were therefore confounded by differences in duration of follow-up.

The two studies that included follow-up of longer than 10 years produced differing results, with the US study reporting lower estimates of sensitivity than the English study for similar thresholds for magnetic resonance imaging. It is possible that these differences reflect the smaller sample size of the English study or that the large proportion of dropouts from this study biased results. An alternative explanation is that magnetic resonance imaging may be more accurate in patients presenting with brainstem or spinal cord symptoms than in patients with optic neuritis. Future studies should assess whether the accuracy of magnetic resonance imaging varies according to presenting symptoms.

Rather than the accuracy of magnetic resonance imaging alone in diagnosing multiple sclerosis, the issue of clinical relevance is, arguably, the added value of such imaging in diagnosing the disease compared with the patient's history and clinical examination alone. None of the identified studies addressed this issue. A further limitation of published studies is that they tend to dichotomise the results of magnetic resonance imaging into positive or negative scans. The use of a scale based on features present on a scan, ranging from no lesions (in which case the probability of disease is low), to specific lesions (which may imply a greatly increased probability of disease), should be considered as an alternative to dichotomisation. This is probably consistent with how the results of magnetic resonance imaging are interpreted in practice.

Implications

In patients with clinically suspected multiple sclerosis, magnetic resonance imaging currently allows a diagnosis of the disease according to the McDonald 2001 criteria. Our results suggest that magnetic resonance imaging is a relatively poor test for both ruling in and ruling out multiple sclerosis. In clinical practice a false positive diagnosis of multiple sclerosis is potentially more dangerous than a false negative one because it implies unnecessary successive tests and treatments, or needless anxiety and psychological distress for the patient. Wrongly ruling out a diagnosis of multiple sclerosis after a first attack seems less dangerous: not all patients who experience a first attack will develop the disease and currently no treatment has been shown to delay conversion to clinically definite multiple sclerosis or impacts on long term disability. Neurologists should discuss with their patients the potential diagnosis, treatment, and ultimate effect of potential errors of false positive and false negative magnetic resonance imaging results. High quality clinical research based on improved magnetic resonance imaging techniques and measures in combination with a complete description of participants and long term clinical follow-up are needed for quantitative assessment of the clinical efficacy of magnetic resonance imaging in the diagnosis of multiple sclerosis. The disease remains a predominantly clinical diagnosis.

Contributors: PW, JACS, JJD, CM, and RH designed the study. PW carried out the literature searches. PW and CM screened the results of the searches for relevance, assessed inclusion, and carried out the data extraction and quality assessment. GF provided expert advice on magnetic resonance imaging and multiple sclerosis. RH, JJD, JACS, and PW developed the plan of analysis and RH carried out the analysis. PW, JACS, RH, JJD, and GF drafted the paper. All authors commented on drafts of the paper and approved the final manuscript. PW is guarantor for the paper.

Funding: This work was supported by the Medical Research Council Health Services Research Collaboration. The authors’ work was independent of the funders. JJD is funded by a senior research fellowship in evidence synthesis from the UK Department of Health. Competing interests: None declared.

Ethical approval: Not required.

What is already known on this topic

Magnetic resonance imaging has been recommended in the diagnosis of multiple sclerosis

The diagnostic accuracy of such imaging has been assessed but a systematic review has not previously been carried out

What this study adds

Magnetic resonance imaging is of limited utility for both ruling in and ruling out multiple sclerosis

Studies with shorter follow-up tended to produce higher estimates of sensitivity and lower estimates of specificity compared with longer term studies
Research


(Accepted 2 February 2006)

doi 10.1136/bmj.38771.583796.7C

MRC Health Services Research Collaboration, Department of Social Medicine, Bristol BS8 2PR
Penny Whiting
Roger Harbord
Jonathan A C Sterne

Centre for Reviews and Dissemination, University of York
Caroline Main
Jonathan J Deeks

Unit of Neuroepidemiology, Istituto Nazionale Neurologico “Carlo Besta,” Milan, Italy
Graziella Filippini

Department of Social and Preventive Medicine, University of Bern, Switzerland
Matthias Egger

Correspondence to: P Whiting penny.whiting@bristol.ac.uk