Multiple Sclerosis

http://msj.sagepub.com/

Improving the translational hit of experimental treatments in multiple sclerosis

Hanna M Vesterinen, Emily S Sena, Charles ffrench-Constant, Anna Williams, Siddharthan Chandran and Malcolm R Macleod

Mult Scler published online 4 August 2010
DOI: 10.1177/1352458510379612

The online version of this article can be found at:
http://msj.sagepub.com/content/early/2010/07/29/1352458510379612

Published by:
SAGE
http://www.sagepublications.com

Mult Scler OnlineFirst, published on August 4, 2010 as doi:10.1177/1352458510379612

Multiple Sclerosis

Multiple Sclerosis
0(00) 1–12
© The Author(s) 2010
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1352458510379612
msj.sagepub.com
SAGE

Improving the translational hit of experimental treatments in multiple sclerosis

Hanna M. Vesterinen, Emily S. Sena,
Charles ffrench-Constant, Anna Williams,
Siddharthan Chandran and Malcolm R. Macleod

Abstract

Background: In other neurological diseases, the failure to translate pre-clinical findings to effective clinical treatments has been partially attributed to bias introduced by shortcomings in the design of animal experiments.
Objectives: Here we evaluate published studies of interventions in animal models of multiple sclerosis for methodological design and quality and to identify candidate interventions with the best evidence of efficacy.
Methods: A systematic review of the literature describing experiments testing the effectiveness of interventions in animal models of multiple sclerosis was carried out. Data were extracted for reported study quality and design and for neurobehavioural outcome. Weighted mean difference meta-analysis was used to provide summary estimates of the efficacy for drugs where this was reported in five or more publications.
Results: The use of a drug in a pre-clinical multiple sclerosis model was reported in 1152 publications, of which 1117 were experimental autoimmune encephalomyelitis (EAE). For 36 interventions analysed in greater detail, neurobehavioural score was improved by 39.6% (95% CI 34.9–44.2%, $p < 0.001$). However, few studies reported measures to reduce bias, and those reporting randomization or blinding found significantly smaller effect sizes.
Conclusions: EAE has proven to be a valuable model in elucidating pathogenesis as well as identifying candidate therapies for multiple sclerosis. However, there is an inconsistent application of measures to limit bias that could be addressed by adopting methodological best practice in study design. Our analysis provides an estimate of sample size required for different levels of power in future studies and suggests a number of interventions for which there are substantial animal data supporting efficacy.

Keywords
Animal models, EAE, meta-analysis, systematic review

Date received: 9th February 2010; revised: 21st April 2010; 2nd July 2010; accepted: 4th July 2010

Introduction

Multiple sclerosis (MS) is a neurological disease second only to trauma as a cause of disability in young adults. It is a multi-focal and multi-phasic immune-mediated disorder characterized pathologically by inflammatory demyelination, neuronal injury and partial remyelination. Clinically the disease has two distinct phases reflecting inter-related pathological processes; inflammation is dominant during relapse and neurodegeneration is the substrate of progression. Several decades of laboratory research, most commonly using experimental autoimmune encephalomyelitis (EAE) where an inflammatory disease of the central nervous system (CNS) is created by generating immune activity targeted at myelin, has led to major insights into disease evolution as well as the development of five licensed interventions (two forms of interferon beta-1α,

Centre for Clinical Brain Sciences, Department of Clinical Neurosciences, University of Edinburgh, Western General Hospital, UK.

Corresponding author:
Dr Malcolm R. Macleod, University of Edinburgh, Bramwell Dott Building, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK
Email: malcolm.macleod@ed.ac.uk

interferon beta-1ß, glatiramer acetate, and natalizumab). Collectively these treatments reduce relapse frequency, although effects on disease progression are less certain,[1–4] and have altered clinical practice. However, there are no proven treatments for progressive disease highlighting the great unmet need to identify new therapies. Many putative interventions have been tested in animal models of MS and either failed in clinical trial or not taken forward to clinical trial. Against this background it is important to develop strategies to better understand existing data from animal studies to inform selection of interventions and the design of the next generation of clinical trials in MS.

Systematic review can provide both an overview of a field of research and of the efficacy of individual interventions; and insights into prevalence and consequences of any shortcomings regarding internal (measures to avoid bias, e.g. random allocation to group and a blinded assessment of outcome) and external validity (e.g. the effect of publication bias or the relevance of the various conditions of testing in animals) of a field of research. In animal models of focal ischaemia, we have shown that low reported study quality is a potent source of bias, resulting in a substantial overstatement in the reported efficacy of, among others, NXY059,[5] Tirilazad[6] and FK506.[7] In addition, we showed that the design of clinical trials has not always reflected the circumstances under which maximum efficacy is observed in animal models.[8]

Here we have set out to systematically describe the published literature reporting the efficacy of interventions tested in animal models of MS. We have focused on EAE, for which several individual models exist, with varying disease courses and pathologies.[9] Since the present review is primarily concerned with study quality and design characteristics, the different EAE models have been grouped together. Using a systematic approach we: (1) describe the breadth of interventions tested in EAE; (2) evaluate the impact of study design characteristics (sample size, time to treatment/assessment and outcomes measured); (3) for interventions tested five or more times, use meta-analysis to report summary estimates of efficacy; and (4) use stratified meta-analysis to assess the impact of study quality and study design characteristics on the reported effect sizes of interventions in EAE.

Materials and methods

Search strategy

Studies of interventions tested in animal models of MS were identified from PubMed (to 17 April 2008). Our search strategy used the terms: 'multiple sclerosis' OR 'experimental allergic encephalomyelitis' OR 'experimental autoimmune encephalomyelitis' OR 'experimental allergic EAE' OR 'experimental autoimmune EAE' OR 'autoimmune demyelinating disease'; limited to animals, with no language restrictions. Abstracts were screened independently by two investigators (HV and MM) to identify those meeting our inclusion criteria (see the following section), with differences clarified by discussion with a third investigator (ES).

Inclusion criteria

We included publications testing an intervention in an *in vivo* animal model of MS where the outcome was measured as a change in neurobehavioural score, axonal loss and/or demyelination/remyelination. Where interventions had been tested five or more times, we conducted meta-analyses to provide summary estimates of efficacy. For this we extracted data from controlled studies which reported the mean outcome and the reported variance. Where data were expressed graphically or were missing we attempted to contact authors for further information. Where it was not clear whether variance was expressed as a standard deviation (SD) or standard error of the mean (SEM) we recorded reported variance as SD, as for the purposes of meta-analysis, this is a more conservative approach.

Extracted data

For each publication, information on quality (see the following section) and experiments (animal species, strain and intervention tested) were entered into a centralized Microsoft Access Database. Neurobehavioural scores were the most commonly reported outcome measure and were used to evaluate efficacy. Where different neurobehavioural outcomes (e.g. a mean clinical severity score and a mean maximal severity score) were reported from the same cohort of animals, we combined these, using fixed effects meta-analysis as described, and used this aggregate figure for further analysis. Extracted data for both a treatment and control group included the number of animals, the neurobehavioural score, the SD or SEM, and the treatment protocol (dose, route of administration, time and number of administrations, and the duration of assessment).

Quality score

The methodological quality of individual studies was assessed using a five-item checklist. This was derived from the consensus statement 'Good laboratory practice' in the modelling of stroke,[10] and encompasses the reporting of measures to reduce bias: blinded assessment of outcome, random allocation to group and a sample size calculation. In addition, we included the

Table 1. The five items in the quality checklist and the percentages reported to be met in the experimental autoimmune encephalomyelitis (EAE) and focal cerebral ischemia (FCI) literature[20]

Item	*EAE*	*FCI*
Random allocation to group	9%	36%
Blinded assessment of outcome	16%	29%
Sample size calculation	<1%	3%
Compliance with animal welfare regulations	32%	57%
Statement of a potential conflict of interest	6%	23%

reporting of compliance with animal welfare regulations and a statement of a potential conflict of interest to provide an approximate measure of overall quality (Table 1).

Meta-analysis

For each comparison we expressed the mean outcome for the treatment group and the SD in treatment and control groups as a proportion of the outcome in the control group. The effect size (the normalized difference between the treatment and control groups) and its standard error were then calculated. Data were aggregated using a weighted mean difference method. To account for anticipated heterogeneity we used the random effects model of DerSimonian and Laird,[11] in which the weighting given to individual comparisons depends on the variance within those comparisons and on overall heterogeneity. This is a more conservative technique than fixed effects meta-analysis.

The significance of differences between n groups was assessed by partitioning heterogeneity (stratified meta-analysis) and by using the χ^2 distribution with $n - 1$ degrees of freedom (df). To allow for multiple comparisons we adjusted our significance level to $p < 0.002$ for neurobehavioural scores using Bonferroni correction.

When a control group served more than one experimental group, the number of observations in that control group was, for the purpose of the meta-analysis, divided by the number of experimental groups served.

Results

Scope of intervention testing in experimental multiple sclerosis studies

Experiments testing 1717 interventions were identified from 1152 papers published on PubMed between 1961 and May 2008 (Figures 1 and 2, Supplementary Material 1). EAE was reported in 1117 publications; active EAE being the most commonly used model (1038 publications), with 191 reporting adoptively transferred EAE (Figure 2). Experiments were conducted in mice (494 publications), rats (481), guinea pigs (126), marmosets (8), cynomolgus monkeys (5), rhesus monkeys (10), rabbits (16), chickens (1) and ewes (1).

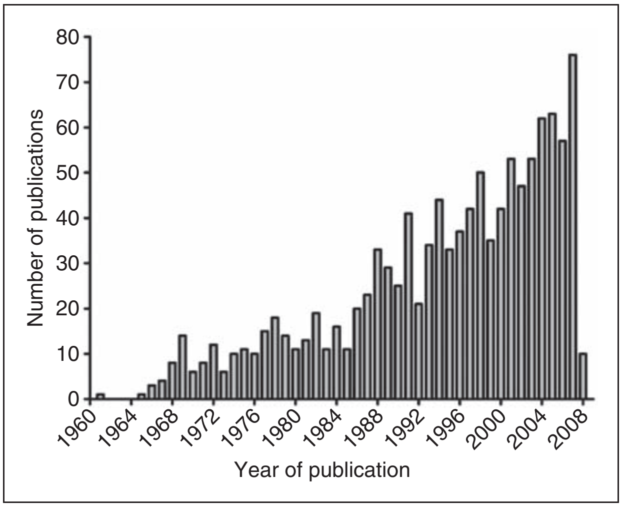

Figure 1. Time course of publications: histogram showing the number of included articles published by year.

Measures of efficacy of interventions used

Efficacy was measured using neurobehavioural outcomes in 1110 publications, quantitative histological data for the extent of demyelination and/or remyelination in 240 publications and axon loss in 82 publications (Figure 3A).

We therefore focused our analyses on studies reporting neurobehavioural outcomes in EAE, and found 38 interventions (Table 2) in 388 publications which were tested five or more times (Supplementary Material 2). The most common neurobehavioural outcomes were the mean maximal severity score (the mean of the maximum severity of EAE for each animal in the group, calculated over the total number of animals in the entire group (96/126 publications)) and the mean clinical score (mean combined score of each animal over the duration of the experiments (39/126)) (Figure 3B). Thirty-six interventions had reported mean maximal severity and/or mean clinical score and were thus used in the meta-analysis. The odds of developing EAE were reported in 170 publications, but because this was likely to be exquisitely sensitive to the interval (if any) between the induction of EAE and the initiation of treatment, we excluded these data from further analysis.

For the 36 interventions included here, using data from 126 publications, the mean maximal severity

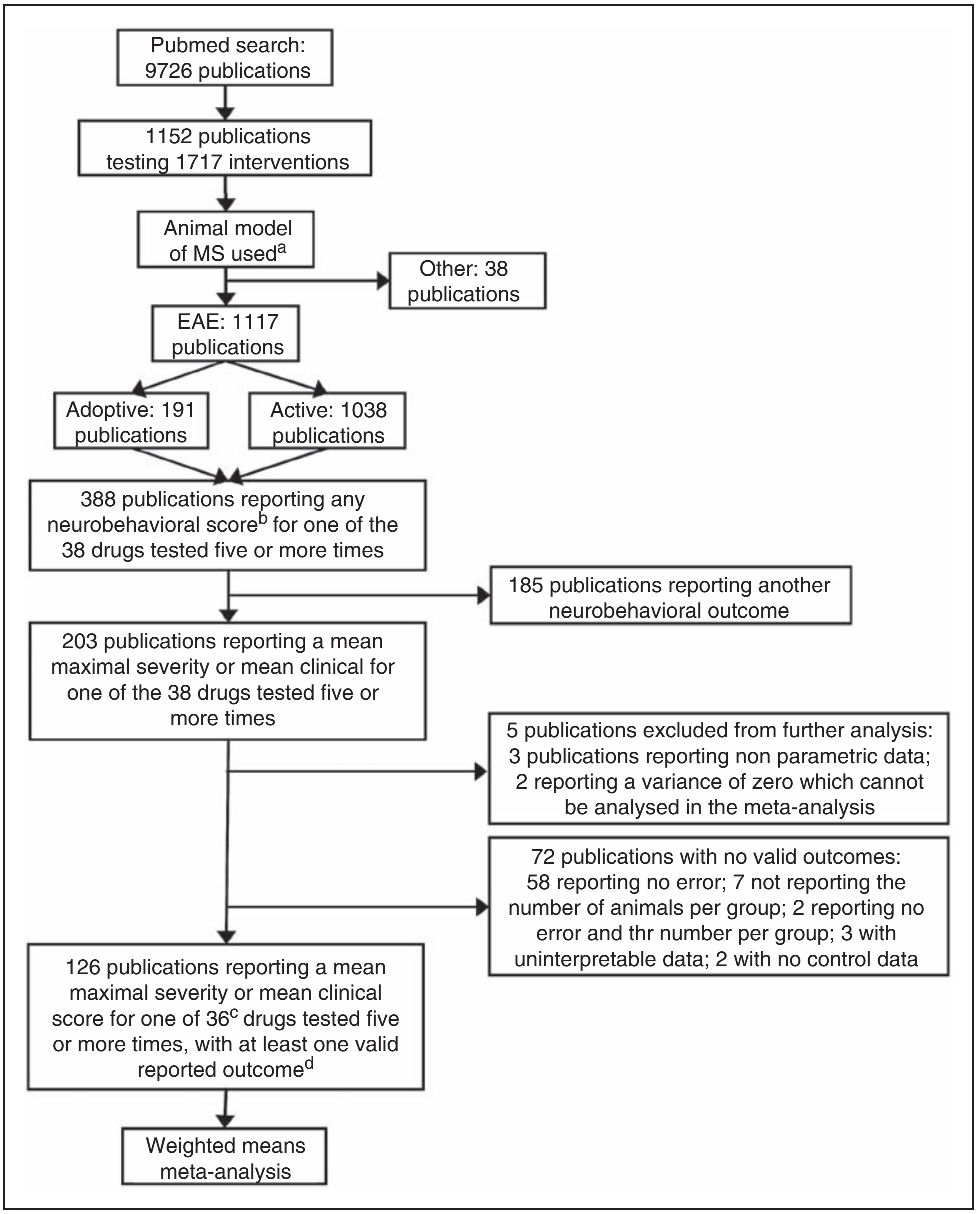
Pubmed search: 9726 publications
1152 publications testing 1717 interventions
Animal model of MS used[a]
Other: 38 publications
EAE: 1117 publications
Adoptive: 191 publications
Active: 1038 publications
388 publications reporting any neurobehavioral score[b] for one of the 38 drugs tested five or more times
185 publications reporting another neurobehavioral outcome
203 publications reporting a mean maximal severity or mean clinical for one of the 38 drugs tested five or more times
5 publications excluded from further analysis: 3 publications reporting non parametric data; 2 reporting a variance of zero which cannot be analysed in the meta-analysis
72 publications with no valid outcomes: 58 reporting no error; 7 not reporting the number of animals per group; 2 reporting no error and thr number per group; 3 with uninterpretable data; 2 with no control data
126 publications reporting a mean maximal severity or mean clinical score for one of 36[c] drugs tested five or more times, with at least one valid reported outcome[d]
Weighted means meta-analysis

Figure 2. Quorum chart showing fate of identified studies: [a]The number of publications reporting these models of multiple sclerosis (MS). Some publications are reported in more than one category. [b]Any characteristic relating to physical signs of disease (weight loss, relapse rate etc). [c]Two drugs had no valid outcomes reported for mean maximal severity score or mean clinical score. [d]Where it was not clear whether variance was expressed as standard deviation or standard error (26 publications) we recorded reported variance as standard deviation.

score and mean clinical score were assessed by no fewer than 82 unique scoring systems. These are generally subjective ordinal scales used to measure the level of disability, typically from 0 (no clinical signs of disease) to 5 (moribund or dead).

Efficacy of interventions

One hundred and twenty-six publications reported a 39.6% (95% CI 34.9–44.2%) improvement in neurobehavioural outcome for 36 interventions tested in 450 experiments using 7258 animals. There was significant between-study heterogeneity for neurobehavioural score ($\chi^2 = 17{,}866.5$, $\mathrm{df} = 35$, $p < 0.002$), reflecting the anticipated differences between interventions, models used and study design.

For the various interventions used, improvement in neurobehavioural outcome ranged from 100% for linomide (95% CI 99.0–101.0%, two publications) to a 21.7% worsening for proteolipid protein (−36.6

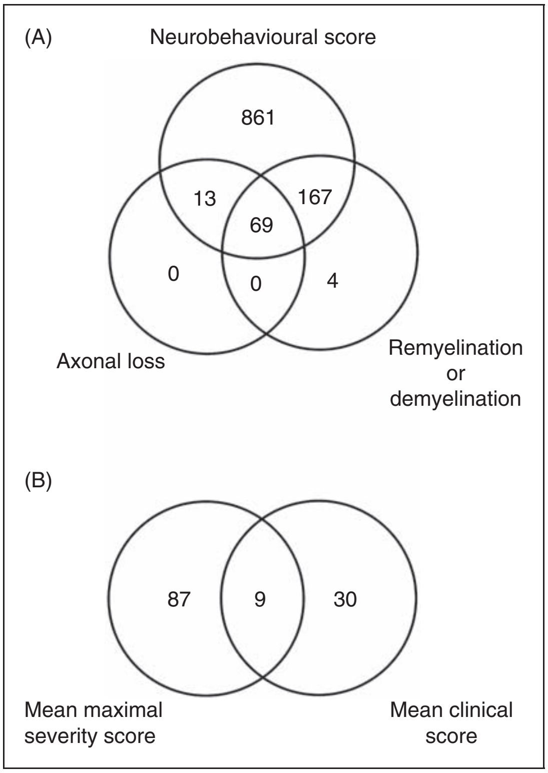

Figure 3. Outcomes reported in 1117 publications in EAE. Venn diagrams of (A) the reporting of neurobehavioural scores, axonal loss and remyelination or demyelination from 1117 publications on EAE; three publications reported either the degree of inflammation or lesions without one of the above; and (B) the reporting of mean clinical score and mean maximal severity score from the 126 publications which reported the use of a drug tested more than four times with at least one valid result, in a model of EAE.

to −6.8%, two publications). Glatiramer acetate improved outcome by 50.6% (34.5–66.7%, nine publications), and taken together, myelin basic proteins improved outcome by 44.2% (36.2–52.2%, 33 publications); see Table 2 and Figure 4.

Measures to avoid bias

Measures to reduce bias were reported in few of the 1117 included publications (Table 1 and Supplementary Material 3). The median quality score was 0 (interquartile range [IQR] 0–1). Randomization was reported by 106 publications (9%), blinded assessment of outcome by 178 publications (16%), a power calculation by 2 publications (< 1%), compliance with animal welfare regulations was reported in 357 publications (32%) and a potential conflict of interest by 63 publications (6%). Interestingly, the only study quality item for which reporting changed substantially over time was compliance with animal welfare regulations, where consistent reporting began in the late 1990 s, reached 50% in 2001, and by 2007 was reported in over 70% of publications.

These quality items appeared to have an effect on reported outcome. For the 36 interventions analysed in the greatest detail, non-randomized studies reported significantly higher efficacy (41.6%, 95% CI 36.7–46.5%) than randomized studies (20.6%, 95% CI 11.4–29.7%; $\chi^2 = 1797.5$, df = 1, $p < 0.002$). Similarly, studies which did not blind the assessment of outcome reported higher estimates of efficacy (41.0%, 95% CI 36.2–45.8%) than blinded studies (29.8%, 95% CI 19.8–39.8%, $\chi^2 = 2602.0$, df = 1, $p < 0.002$); see Figure 5. Too few studies reported a sample size calculation to allow stratified meta-analysis.

Study design

Timing of intervention. We defined the day of EAE induction as day zero. Of 450 included individual outcomes from the 126 selected publications, the intervention was delivered before the induction of EAE in 48%; on the day of induction in 22%; and only 30% administered after the day of induction. Treatment was commenced more than 2 weeks after the induction of EAE in 15 experiments (4%) and more than 3 weeks after induction in 4 (< 1%) experiments. The median time to treatment was 0 days (interquartile range –11 to 4). One per cent of outcomes were from publications that did not report the day of intervention administration (Table 3). Effect sizes were significantly lower with longer delays to treatment ($\chi^2 = 8568.1$, df = 5, $p < 0.002$, Figure 6A). We attempted to categorize studies into those modelling acute, chronic or chronic relapsing disease, but unfortunately this was not possible because of differences in the way models with particular attributes were employed, for instance using pretreatment[12] and early measurement of outcome[13] in models considered to have attributes of chronic disease.

Timing of assessment of outcome. One per cent of outcomes were assessed before day 10, 22% between days 10 and 20, 21% between days 21 and 30, 20% between 31 and 40, and 19% after day 40, with a median of 30 days (IQR 20–40). A further 17% did not report the day of assessment (Table 3).

Sample size and power calculation. The median sample size was eight for the treatment groups and five for the control groups, (IQR 5–10 and 3–8, respectively). Stratifying heterogeneity according to the mean number of animals in both groups accounted for a significant proportion of observed heterogeneity ($\chi^2 = 2106.2$, df = 3, $p < 0.002$) with effect size lower in larger studies (Figure 6B).

Table 2. A summary of the 38 drugs used five or more times in experimental autoimmune encephalomyelitis (EAE); 36 interventions had point estimates of efficacy calculated

			Number Reporting			Neurobehavioural Score[a]				
Intervention	Number of Publications	Mean Quality Score	Neurobehavioural Score	Axonal Loss	Demyelination/ Remyelination	Number of Experiments	Number of Animals	Effect Size	95% CI Lower	95% CI Upper
6-Mercaptopurine	8	0.25	8	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Aminoguanidine	5	1	5	1	2	8	122	24.7	−9.4	58.9
Anti-Lymphocyte Serum	8	0.13	8	n/a	n/a	1	15	12.6	−18	43.1
Azathioprine	12	0.08	12	n/a	1	3	44	17.8	−9.3	44.9
Bone Marrow	6	1.17	6	4	5	12	202	26.6	20	33.3
Cobra Venom Factor	7	0.14	7	n/a	3	8	119	45.6	18.8	72.5
Complete Freunds Adjuvant	9	0	9	n/a	n/a	6	68	52.8	17.7	88
Cyclophosphamide	32	0.16	32	n/a	3	10	469	−12.7	−64.7	39.4
Cyclosporin	24	0.21	24	n/a	n/a	14	251	36.5	9.2	63.7
Dexamethasone	23	0.74	22	n/a	n/a	4	44	24.4	1.3	47.5
Oestrogen Hormone	10	0.8	10	n/a	4	44	685	48.3	39.5	57
FTY720	6	1	6	1	1	9	128	87.6	77.1	98.1
Glatiramer Acetate	25	0.68	25	2	6	20	457	50.6	34.5	66.7
Immunoglobulin	9	0.78	9	n/a	1	5	209	31.1	9.2	53
Indomethacin	7	0.43	7	0	1	3	42	34.2	12.6	55.7
Insulin Like Growth Factor	5	1.4	5	2	n/a	4	76	21	18.4	23.6
Interferon Beta	15	1.13	15	3	5	9	125	12.2	−30.1	54.4
Interleukin 4	6	0.5	6	n/a	1	9	114	32.1	10.9	53.4
Interleukin 10	9	0.67	9	n/a	1	27	371	1	−10.7	12.7

(continued)

Table 2. Continued

Intervention	Number of Publications	Mean Quality Score	Number Reporting			Neurobehavioural Score[a]				
			Neurobehavioural Score	Axonal Loss	Demyelination/ Remyelination	Number of Experiments	Number of Animals	Effect Size	95% CI Lower	95% CI Upper
Linomide	8	0.38	8	n/a	1	4	146	100	99	101
Lovastatin	7	1.43	7	1	5	7	107	59	37.4	80.6
Methotrexate	11	0.27	11	n/a	n/a	2	21	33.1	−64.9	131.1
Methylprednisolone	6	1.5	6	1	1	10	254	7	−4.2	18.2
Minocycline	5	0.8	5	2	4	3	44	64.7	39.4	90
Mitoxantrone	9	1	9	n/a	1	30	342	59.1	46.2	71.9
Myelin Oligodendrocyte Glycoprotein	7	0.86	7*	n/a	1	n/a	n/a	n/a	n/a	n/a
Myelin	6	0.17	6	n/a	2	4	71	−2.2	−29.9	25.5
Myelin Basic Protein	108	0.44	108	1	10	135	1758	44.2	36.2	52.2
Phenytoin	5	1	5	4	n/a	2	22	56.5	43.5	69.4
Prednisolone	8	0.75	8	n/a	1	3	30	48.5	27.6	69.5
Proteolipid Protein	12	1.08	12	n/a	2	2	47	−21.7	−36.6	−6.8
Reserpine	5	0.4	5	n/a	n/a	5	30	62.2	39.9	84.6
Retinoic Acid	5	0.6	5	1	1	1	18	29.7	13.4	45.9
Rolipram	7	1	6	1	2	6	166	14.1	−13.1	41.4
Spinal Cord Protein	12	0	12	n/a	n/a	5	49	46.4	−21.7	114.5
Transforming Growth Factor	9	0.56	9	n/a	2	8	118	19.2	−6.3	44.7
V Beta 8	13	0.23	13	n/a	n/a	14	111	50.2	21.8	78.7
Vitamin D	15	0.87	15	1	1	13	383	48.5	36.8	60.3

[a] 6-mercaptopurine and myelin oligodendrocyte glycoprotein had no valid outcomes reported for mean maximal severity score or mean clinical score.

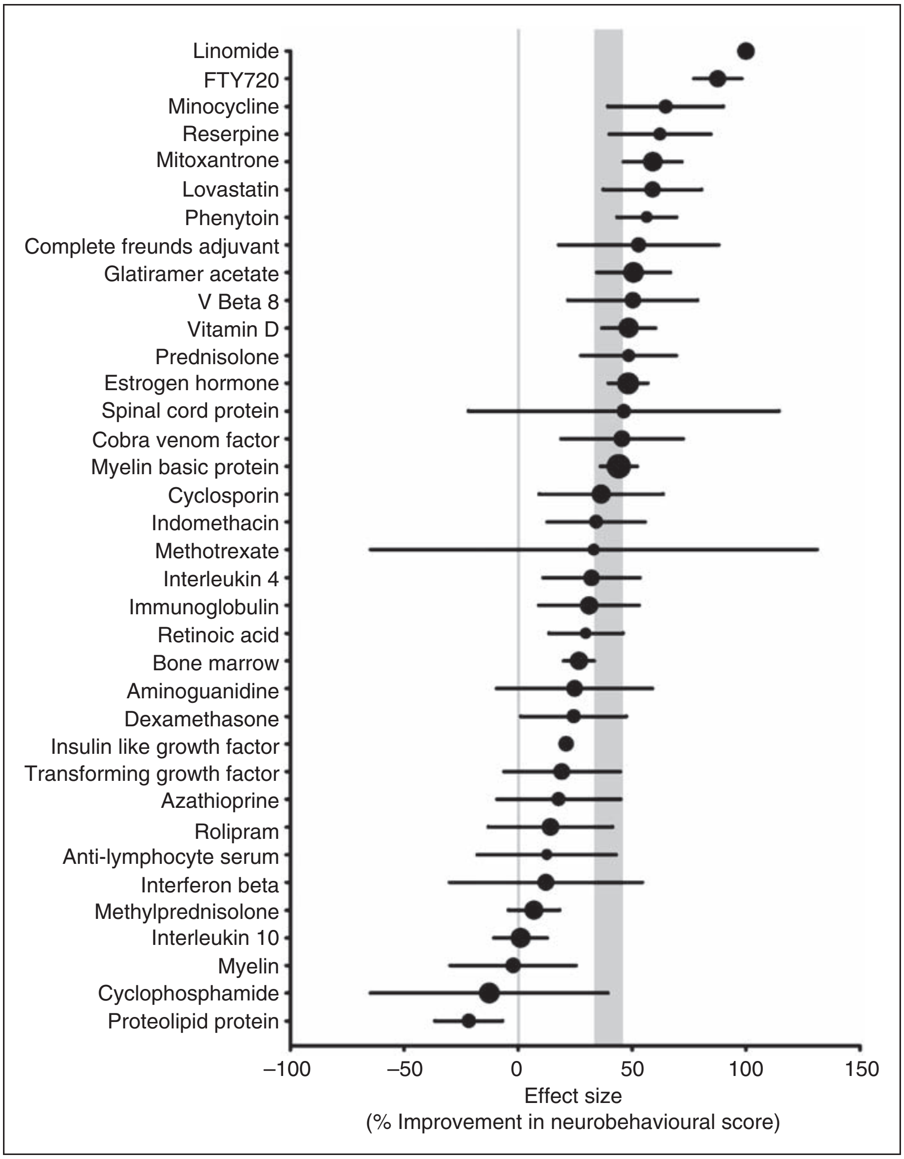

Figure 4. Effect of various interventions on neurobehavioural outcome. Point estimates of efficacy (% improvement over control) for neurobehavioural score for 36 interventions. The horizontal error bar represents the 95% confidence interval (CI) for that intervention. The vertical grey bar represents the global estimate of efficacy and its 95% CI. The size of the symbols represents the log of the number of animals for that intervention.

Post hoc power calculations have limited validity, but assuming an improvement in mean maximal severity of 40% (from our global efficacy analysis), a median standard error of 28% and a median of five animals in the control group and eight animals in the treatment group, the typical EAE study included here is powered at 63%. Based on these data we have provided estimates for the number of animals required to achieve different levels of power, the power of a study for different observed effect sizes and the number of animals required to observe a specified effect size[14] (Figure 7).

Discussion

The present study undertook systematically to review pre-clinical experiments testing the efficacy of interventions in animal models of MS. The vast majority of these (1117/1152) used EAE, the focus of this study. Our goals were to summarize the design of studies and the reporting of measures to avoid bias, and to provide some estimates of effect size. We had to exclude from our analysis data from a substantial number of studies because they did not report fundamental aspects of their data such as variance (i.e. SD or SEM) or the

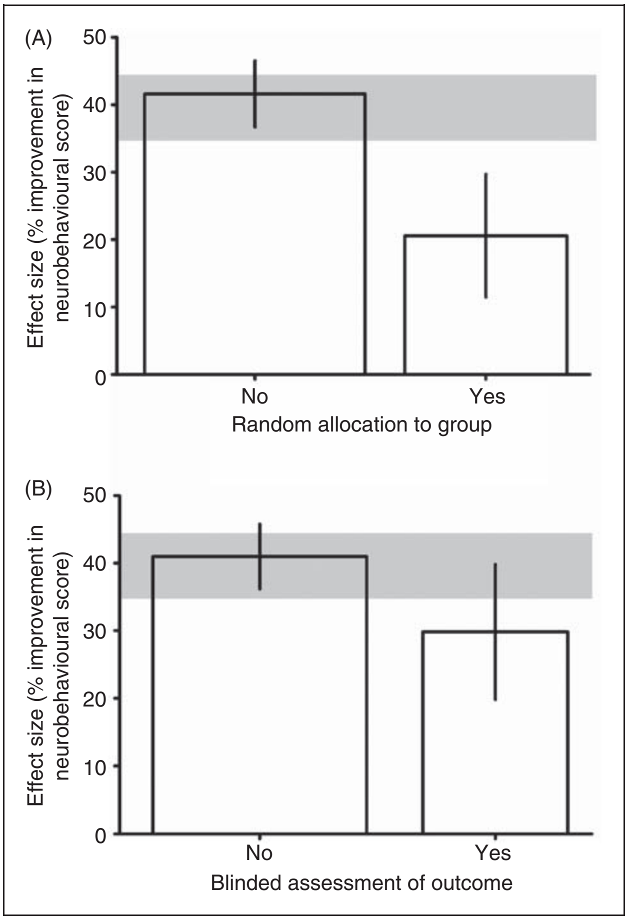

Figure 5. Impact of reporting of measures on avoiding study quality bias. The effect of random allocation to group (A) and blinded assessment of outcome (B) on the estimates of effect size for neurobehavioural score. The horizontal grey bar represents the 95% confidence limits of the global estimate of efficacy. The vertical error bars represent the 95% confidence intervals for the individual estimate. Bar widths represent the log of the number of animals contributing to that comparison.

number of animals in their experimental groups. Data from such publications should be interpreted with caution. In those we could analyse, the calculated summary efficacy of the 36 interventions investigated in most detail varied considerably, with linomide performing most favourably. However, these summaries should be treated with caution, due to differences in study design characteristics, and cannot be used to give a rank order of potency.

Methods to avoid bias: Internal validity

We have shown that studies reporting measures to avoid bias (random allocation to group and blinded assessment of outcome, both important indicators of internal validity) give substantially lower estimates of efficacy than studies that do not report such measures. This finding is comparable to a similar impact of study quality on estimates of efficacy in the experimental stroke literature.[5–8] Importantly, these measures are reported by only a minority of EAE studies, and substantially fewer than in the stroke literature. Of course, some studies may have taken such measures but not reported them, although a survey of actual versus reported study quality in experimental stroke suggests that these are broadly similar.[15] In addition it is possible that studies taking measures to avoid bias with negative or less-impressive results might remain unpublished and therefore unknown to this analysis as is the case with the clinical trials literature.[16] We have also previously estimated that publication bias leads to an overstatement of efficacy of around 30% in the experimental stroke literature.[17] The estimates of efficacy reported here are therefore likely to represent overestimates of true efficacy.

Table 3. The number and percentage of experiments administering their intervention at various time points, and the time of assessment of outcome for the 450 experiments measuring neurobehavioural score

Time of Administration	Number of Experiments	%
< Day 0	215	48
Day 0	101	22
> Day 0	131	30
(Day of Symptom Onset	7	2)
Unknown	3	1%
Time of Assessment	**Number of Experiments**	**%**
< Day 10	6	1
Days 10–20	98	22
Days 21–30	96	21
Days 31–40	88	20
> Day 40	86	19
Unknown	76	17

External validity

In clinical trials of MS the primary clinical outcome measures commonly used are relapse frequency and disease progression. Here we chose the endpoint most frequently reported in animal studies: the severity of the initial illness as determined by a neurobehavioural outcome measure. Taking into account the period over which outcome is assessed in EAE it appears that these studies are most closely aligned with the initial inflammatory phase of MS rather than with established MS, where additional neurodegenerative processes result in axon loss that causes progressive disease.

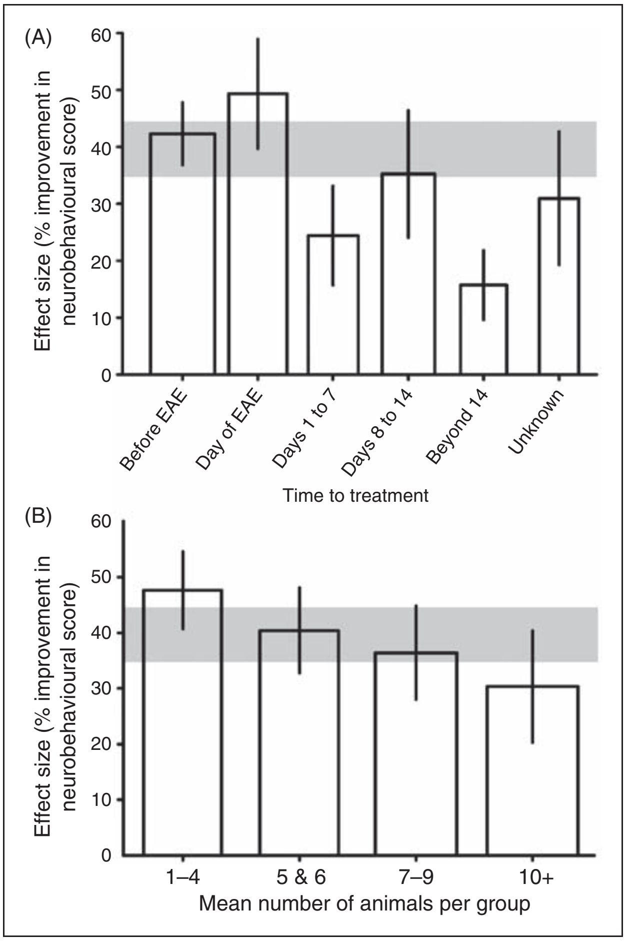

Figure 6. Impact on efficacy of delay to treatment and size of experiment. The effect of (A) time to treatment and (B) mean sample size on the estimate of effect size for neurobehavioural score. The horizontal grey bar represents the 95% confidence limits of the global estimate of efficacy. The vertical error bars represent the 95% confidence intervals for the individual estimate. Bar widths represent the log of the number of animals contributing to that comparison.

Furthermore, interventions were most commonly administered either before or on the day of disease induction (i.e. some days before the development of neurological impairment), an observation also made of the SOD1 mouse model of motor neuron disease.[18] In EAE, interventions may be efficacious by blocking the induction of disease (where mechanisms such as sequestering immunogen or inhibiting the initial immune response may be most relevant) rather than through an effect on the primary pathophysiology of neuronal and glial injury or the evolution of neurological impairment over time. The relevance of such studies to the development of interventions for established relapsing–remitting disease, primary and secondary progressive disease is not clear. It could be argued that efficacy in EAE studies might only be predictive of efficacy in clinical trial if treatment were started after the onset of symptoms (some days after induction), as we presently have no way of identifying patients prior to the onset of the disease. Importantly, treatments given after the immunization which leads to the induction of EAE were much less effective than earlier treatment. Finally, disease burden in MS reflects a complex interplay between inflammation, demyelination, remyelination and neurodegeneration with a temporal shift in pathological emphasis from inflammation to neurodegeneration; further work is required to describe the characteristics of different EAE models in these different domains.

Thus, it may be that some of the difficulties in translating efficacy from animals to man arises because data from appropriately designed studies modelling MS pathophysiology do not provide sufficient insights to likely efficacy in human disease, where different endpoints might be considered key, and longer delays to the initiation of treatment are unavoidable.

Power calculations for future studies

Only two publications reported a sample size calculation, which may reflect the difficulty of performing a meaningful power calculation without a prior meta-analysis such as we have performed here. We have therefore used these data to produce some guidance as to sample sizes required to give various levels of power. Our *post hoc* power calculation suggests that half the experiments included in this analysis are powered at less than 63%.

While we believe these calculations are the best that can be achieved at the present time, they should be interpreted with caution, along with the other data presented here, due to a number of unavoidable limitations. First, in our view the available data are of inconsistent and sometimes poor quality; data from a further 72 publications reporting a mean maximal severity score or mean clinical score outcome could not be included in this analysis because basic information such as the number of animals per group or the variance of presented data were not reported. Second, a meta-analysis is by its nature a *post hoc* analysis and should therefore be considered to be only hypothesis generating. Third, our search strategy was broad but lacked depth, since only one electronic database was searched; and in addition, our search terms are more likely to have identified publications describing EAE rather than other less commonly used models of MS, and these may be at least as valid as EAE. Fourth, by grouping all models of EAE together, this analysis will not have taken into account the strengths and weaknesses of each one individually. A systematic review of the animal models, identifying their common ground with human MS, will give us greater information with

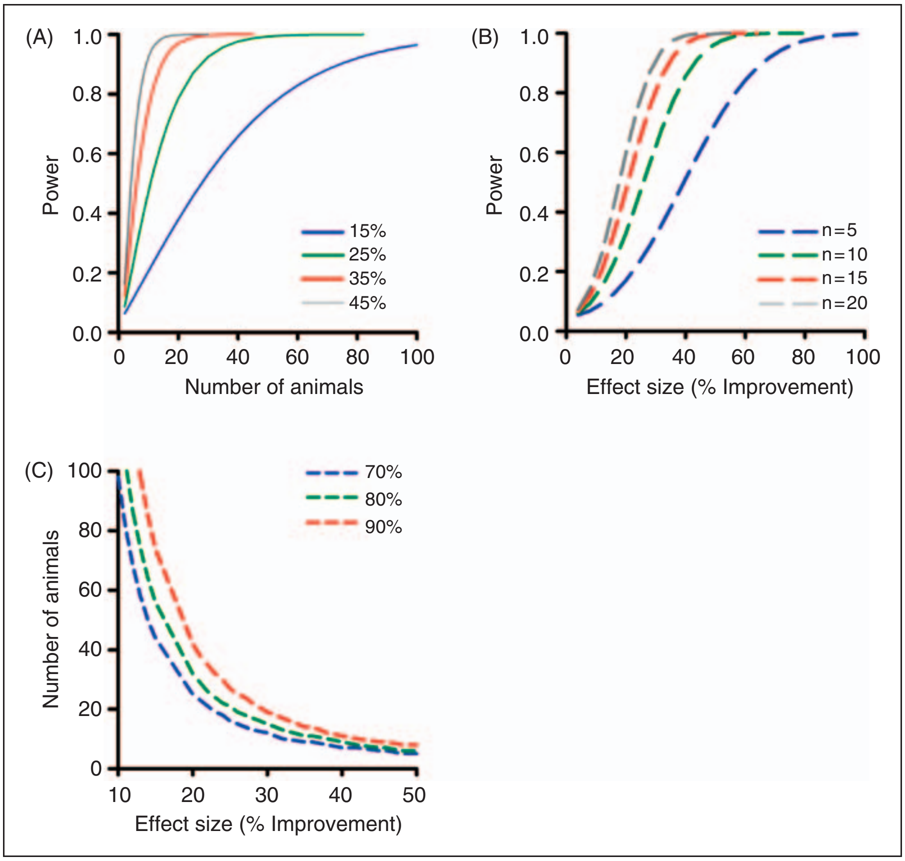

Figure 7. Indicative power calculations. Models of (a) the power of a study using *n* number of animals per group when looking for an effect size of 15%, 25%, 35% or 45% (solid blue, green, red and grey lines, respectively), (b) the power of a study when observing a specific effect size, using 5, 10, 15 or 20 animals per group (long dashed blue, green, red and grey lines, respectively) and (c) the number of animals required to observe a specific effect size powered at 70%, 80% and 90% (short dashed blue, green and red lines, respectively).

which to interpret these results and those of future systematic reviews of interventions in experimental MS. Finally, we have only been able to assess information available in scientific publications.

Recommendations for studies modelling multiple sclerosis

Our findings are based on observation rather than hypothesis testing and therefore should be interpreted with caution. However, we do believe that there is now sufficient evidence, from both here and elsewhere,[19] to recommend that studies testing the efficacy of candidate drugs in animal models of MS should take (and report) measures to improve their internal validity, such as randomization, allocation concealment and the blinded assessment of outcome; that sample size calculations should be performed and reported; and that the preclinical testing should focus on testing efficacy under clinically relevant conditions including the initiation of treatment at some time after the induction of injury, using models which are specifically designed to reflect the complexities of the human disease.

Conclusions

MS is a disabling condition for which only moderately effective treatments are available. EAE has proven immensely valuable in modelling particularly inflammatory aspects of MS and has led to many insights into disease mechanism as well as several licensed treatments for early MS. However, there remains a great need to identify the next generation of therapeutics that will particularly target the unmet need of treatment for the progressive phase of disease. While EAE has proved to be a valuable model of inflammatory myelin disease for studies of mechanism, our data suggest that to date the testing in animals of candidate interventions for MS has potentially been confounded by limited internal validity (with little reported use of randomization, blinding and power calculations) and by limited external validity (with few treatments given at clinically appropriate time points) Analyses such as ours can we hope provide insights into the strengths and limitations of existing animal data. Further work is required to identify which aspects of experimental design are the most powerful determinants of bias, to allow continuing improvements in the use of animal models and an evidence-based approach to the design of clinical trials in MS.

Funding

HMV is funded by the Multiple Sclerosis Society of Great Britain. No other funds were used to support this review process; however, ESS is funded by The University of Edinburgh PhD Scholarship and MRM was supported in part through the MRC Trials Methodology Hub. AW is a Wellcome Trust Intermediate Fellow.

Conflict of interest statement

None declared.

Acknowledgements

We are grateful to Professor David Baker for helpful discussions.

References

1. Munari L, Lovati R and Boiko A. Therapy with glatiramer acetate for multiple sclerosis. *Cochrane Database Syst Rev* 2004; CD004678.
2. Clegg A and Byrant J. Immunomodulatory drugs for multiple sclerosis: a systematic review of clinical and cost effectiveness. *Expert Opin Pharmacother* 2001; 2: 623–639.
3. Schiess N and Calabresi PA. Natalizumab: bound to rebound? *Neurology* 2009; 72: 392–393.
4. Rojas JI, Romano M, Ciapponi A, Patrucco L and Cristiano E. Interferon beta for primary progressive multiple sclerosis. *Cochrane Database Syst Rev* 2009; CD006643. DOI: 10.1002/14651858.
5. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U and Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* 2008; 39: 2824–2829.
6. Sena E, Wheble P, Sandercock P and Macleod M. Systematic review and meta-analysis of the efficacy of tirilazad in experimental stroke. *Stroke* 2007; 38: 388–394.
7. Macleod MR, O'Collins T, Horky LL, Howells DW and Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. *J Cereb Blood Flow Metab* 2005; 25: 713–721.
8. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* 2007; 334: 197–202.
9. Baker D and Jackson S. Models of multiple sclerosis. *Adv Clin Neurosci Rehabil* 2007; 6(6): 10–12.
10. Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, et al. Reprint: Good laboratory practice: preventing introduction of bias at the bench. *J Cereb Blood Flow Metab* 2008; 29: 221–223.
11. DerSimonian R and Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; 7: 177–188.
12. Gilgun-Sherki Y, Panet H, Holdengreber V, Mosberg-Galili R and Offen D. Axonal damage is reduced following glatiramer acetate treatment in C57/bl mice with chronic-induced experimental autoimmune encephalomyelitis. *Neurosci Res* 2003; 47: 201–207.
13. Brundula V, Rewcastle NB, Metz LM, Bernard CC and Yong VW. Targeting leukocyte MMPs and transmigration: minocycline as a potential therapy for multiple sclerosis. *Brain* 2002; 125: 1297–1308.
14. Length RV. Java Applets for Power and Sample Size, 2006. http://www.stat.uiowa.edu/~rlenth/Power (accessed 10 January 2010)
15. Samaranayake S. Study quality in experimental stroke. CAMARADES Monograph No 2, 2006. Available at: www.camarades.info/index_files/CM2.pdf (accessed 10 January 2010).
16. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009; DOI:10.1002/14651858. MR000006.pub3.
17. Sena ES, van der Worp HB, Bath PMW, Howells DW and Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8e1000344.
18. Benatar M. Lost in translation: treatment trials in the SOD1 mouse and in human ALS. *Neurobiol Dis* 2007; 26: 1–13.
19. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? *PLoS Med* 2010; 7e1000245.
20. Sena E, van der Worp HB, Howells D and Macleod MR. How can we improve the pre-clinical development of drugs for stroke? *Trends in Neurosciences* 2007; 30: 433–439.