Statistical power of MRI monitored trials in multiple sclerosis: new data and comparison with previous results

M P Sormani, P D Molyneux, C Gasperini, F Barkhof, T A Yousry, D H Miller, M Filippi

Neuroimaging Research Unit, Department of Neuroscience, Scientific Institute Ospedale San Raffaele, University of Milan, Milan, Italy
M P Sormani
M Filippi

Unit of Clinical Epidemiology and Trials, National Institute for Cancer Research, Genoa, Italy
M P Sormani

NMR Research Unit, Institute of Neurology, London, UK
P D Molyneux
D H Miller

Department of Neurology, University "La Sapienza", Rome, Italy
C Gasperini

Dutch MS/MR Center, Free University Hospital, Amsterdam, The Netherlands
F Barkhof

Department of Neuroradiology, University of Munich, Munich, Germany
T A Yousry

Correspondence to:
Dr Massimo Filippi, Neuroimaging Research Unit, Department of Neuroscience, Scientific Institute Ospedale San Raffaele, Via Olgettina 60, 20132 Milan, Italy.
Telephone 0039 2 2643 3033; Fax 0039 2 2643 3031; email filippi.massimo@hsr.it.

Received 7 May 1998 and in final form
17 September 1998
Accepted 14 October 1998

Abstract

***Objectives*—To evaluate the durations of the follow up and the reference population sizes needed to achieve optimal and stable statistical powers for two period cross over and parallel group design clinical trials in multiple sclerosis, when using the numbers of new enhancing lesions and the numbers of active scans as end point variables.**

***Methods*—The statistical power was calculated by means of computer simulations performed using MRI data obtained from 65 untreated relapsing-remitting or secondary progressive patients who were scanned monthly for 9 months. The statistical power was calculated for follow up durations of 2, 3, 6, and 9 months and for sample sizes of 40–100 patients for parallel group and of 20–80 patients for two period cross over design studies. The stability of the estimated powers was evaluated by applying the same procedure on random subsets of the original data.**

***Results*—When using the number of new enhancing lesions as the end point, the statistical power increased for all the simulated treatment effects with the duration of the follow up until 3 months for the parallel group design and until 6 months for the two period cross over design. Using the number of active scans as the end point, the statistical power steadily increased until 6 months for the parallel group design and until 9 months for the two period cross over design. The power estimates in the present sample and the comparisons of these results with those obtained by previous studies with smaller patient cohorts suggest that statistical power is significantly overestimated when the size of the reference data set decreases for parallel group design studies or the duration of the follow up decreases for two period cross over studies.**

***Conclusions*—These results should be used to determine the duration of the follow up and the sample size needed when planning MRI monitored clinical trials in multiple sclerosis.**

(*J Neurol Neurosurg Psychiatry* 1999;**66**:465–469)

Keywords: multiple sclerosis; magnetic resonance imaging; power calculations; treatment trials

Brain MRI is the optimal technique for the diagnosis of multiple sclerosis. It also provides an objective and sensitive measure for monitoring disease evolution, either natural or modified by treatment.[1 2] Monthly brain MRI detects active multiple sclerosis lesions five to 10 times more often than the occurrence of clinical relapses in patients with early relapsing-remitting and secondary progressive multiple sclerosis.[3] This high sensitivity of MRI is mainly based on its high sensitivity in detecting gadolinium (Gd) enhancing lesions.[4] This is why to evaluate the efficacy of an experimental treatment in multiple sclerosis, the most widely used end points are the counts of enhancing lesions or scans with at least one of such lesions.[5 6]

The costs of clinical trials based on MRI derived outcome measures are significantly affected by the numbers of scans needed, which in turn depends on the number of patients enrolled, the duration of the trial, and the frequency of MRI sampling. It was shown that with a given treatment efficacy and sample size, the power of the trial increases with increasing duration of the follow up.[6] However, this increase is not linear and, after a certain number of months, the gain in power increases more slowly.[6] Thus the advantage of having a more powerful study might not counterbalance the increased costs. At present, no formal study evaluating the appropriate durations of such trials has been performed. In addition, previous statistical simulations[6–8] were based on relatively small data sets and this may result in an overestimation of the calculated power due to the resampling methodology used. In the present study, we performed simulations using a larger patient data set to calculate the gain in statistical power of two period cross over and parallel group design trials with increasing trial durations. In this patients' cohort, we also evaluated the number of patients and the duration of follow up period needed to obtain relatively stable power estimates. The ultimate goal of this study was to provide a rational basis to calculate the duration of clinical trials in multiple sclerosis when using enhanced MRI derived end points.

Patients and methods

PATIENTS

Sixty five patients (47 women and 18 men) with clinically definite multiple sclerosis[9] were selected from five European centres (eight patients were recruited in Amsterdam, 32 in London, six in Milan, nine in Munich, and 10 in Rome). According to recently published criteria,[10] the cohort consisted of 43 patients with relapsing-remitting multiple sclerosis and 22 patients with secondary progressive

Table 1 Estimated statistical powers required for parallel group trials using different sample sizes and durations of follow up for different treatment effects using the number of new enhancing lesions (A.III) and the number of active scans (A.II) as the end point measure

			Duration of follow up (months)			
Sample size	*Effect (%)*	*Response variable*	*2*	*3*	*6*	*9*
2×20	50	A.II	13	13	14	15
		A.III	14	17	19	21
	60	A.II	20	21	24	25
		A.III	28	35	37	37
	70	A.II	31	35	40	43
		A.III	37	54	57	57
2×30	50	A.II	16	18	21	22
		A.III	22	27	30	31
	60	A.II	27	31	36	39
		A.III	43	48	50	54
	70	A.II	45	51	58	60
		A.III	55	71	74	75
2×40	50	A.II	21	23	27	29
		A.III	26	33	38	40
	60	A.II	36	39	46	48
		A.III	54	62	62	67
	70	A.II	57	64	71	72
		A.III	69	83	85	87
2×50	50	A.II	24	27	32	34
		A.III	31	41	46	46
	60	A.II	43	48	55	58
		A.III	63	67	74	74
	70	A.II	69	73	80	81
		A.III	76	90	92	92

multiple sclerosis. The median age at entry was 33 years (range 15 to 61 years) and median disease duration was 4 years (range 1 to 28 years). Patients were either involved in natural history studies (43 patients) or formed the placebo arms of previous treatment trials (22 patients). The overall patients' characteristics at entry were similar to those of patients usually recruited for MRI monitored trials in multiple sclerosis. To be included, patients had to have had serial monthly gadolinium enhanced T1 weighted scans for at least 9 months with a T2 weighted scan at study entry and exit. Patients taking immunosuppressive drugs other than infrequent courses of corticosteroids during relapses were excluded. No patients with relapsing-remitting multiple sclerosis entered the secondary progressive phase of the disease during the follow up.

MRI

Serial T2 weighted conventional spin echo (CSE) or fast spin echo (FSE) MR images were acquired at study entry and exit. T1 weighted imaging 5–15 minutes after injection of gadolinium was also performed monthly throughout the study period. Conventional dose gadolinium-DTPA (0.1 mmol/kg) was given in 56 of 65 patients, the other nine patients, all recruited in Munich, received 0.2 mmol/kg. Gadolinium enhanced images were not performed within 1 week of corticosteroid treatment.

In Amsterdam, CSE images were obtained using a 0.6 Tesla Technicare (Teslacon II) scanner (SE 2755/60 at entry and exit, SE 450/28 for enhanced scans, 5 mm contiguous axial slices with an interslice gap of 1.25 mm). In London, all MRI was performed using a GE 1.5 Tesla scanner with either CSE (14 patients, 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 580/13 for enhanced scans, 4 mm contiguous axial slices). 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 Munich, images were obtained on a Siemens Impact scanner operating at 1.0 Tesla (SE 3000/40 at entry and exit, SE 600/28 for enhanced scans, 5 mm contiguous axial slices). In Rome, a 0.5 T Toshiba machine was used to obtain CSE images (SE 2500/30 at entry and exit, 400/18 for enhanced scans, 5 mm contiguous axial slices with an interslice gap of 1.0 mm). The scanners used are representative of those typically used in multicentre multiple sclerosis studies. The intercentre variability in the MRI acquisition parameters was slightly larger than that usually allowed in MRI monitored clinical trials, and reflected the post hoc nature of the present study (this might have resulted in a slight overestimation of the power estimates). Scanners were not changed or upgraded over the duration of the study and image acquisition parameters were not modified between entry and exit. The numbers of total and new enhancing lesions were counted on monthly scans by experienced observers at each individual centre.

COMPUTER SIMULATIONS

The results of MRI monitored clinical trials are usually analysed by means of non-parametric tests.[6] In such a context, computer simulations are needed to calculate sample sizes. In the present study, computer simulations were performed according to the method used by Nauta *et al*,[6] Truyen *et al*,[7] and Tubridy *et al*,[8] considering only the case of a homogeneous response.[8] The number of new enhancing lesions and the number of active scans were considered the end point variables. According

Table 2 Estimated statistical powers required for two period cross over trials using different sample sizes and durations of follow up for different treatment effects using the number of new enhancing lesions (A.III) and the number of active scans (A.II) as the end point measure

			Duration of each treatment period (months)			
Sample size	*Effect (%)*	*Response variable*	*2*	*3*	*6*	*9*
2×10	30	A.II	3	7	16	19
		A.III	22	35	62	76
	40	A.II	3	7	16	21
		A.III	37	55	83	91
	50	A.II	10	19	40	49
		A.III	56	73	93	98
2×20	30	A.II	27	36	43	52
		A.III	51	69	91	98
	40	A.II	31	39	47	55
		A.III	77	91	92	100
	50	A.II	54	70	82	89
		A.III	92	98	100	100
2×30	30	A.II	49	50	58	70
		A.III	69	87	98	100
	40	A.II	50	52	63	73
		A.III	92	98	100	100
	50	A.II	79	83	93	98
		A.III	99	100	100	100
2×40	30	A.II	52	57	68	79
		A.III	81	95	100	100
	40	A.II	57	60	73	85
		A.III	97	100	100	100
	50	A.II	85	91	98	100
		A.III	100	100	100	100

Table 3 Estimated statistical powers required for parallel group trials for a 70% treatment effect using variable numbers of patients for the reference data set. The duration of follow up was 9 months and the tested response variable was the number of new enhancing lesions

	Number of patients used for simulations			
Sample size	*10*	*25*	*40*	*65*
2×20	74	72	62	57
2×30	89	88	78	75
2×40	96	93	88	87
2×50	99	94	96	92

Table 4 Estimated statistical powers required for two period cross over trials for a 20% treatment effect using variable durations of follow up for the reference data set. The simulated duration of follow up was 3 months for each period and the tested response variable was the number of new enhancing lesions

	Number of months used for simulations		
Sample size	*3*	*6*	*9*
2×10	27	25	23
2×20	67	47	41
2×30	77	71	56
2×40	89	73	64

to previous studies,[6–8] we indicated as response variable A.II the number of active scans and as response variable A.III the number of new enhancing lesions. The 65 patients studied for 9 months were considered representative of the "untreated" multiple sclerosis population. The first stage of the simulation process consisted in simulating the treatment effect. For each lesion observed in the "untreated" group (and for each active scan) a Bernoulli trial with a probability for success set to the desired treatment effect was performed; a success consisted of the disappearance of the lesion (or of the entire activity on individual scans). In this way, a treated group was created for each treatment effect. It is assumed that the experimental treatment becomes effective immediately and its efficacy does not change during follow up. The second stage of the simulation consisted of generating 1000 trials for each study design. The trials were simulated by drawing random sets of different size from the data (sampling with replacement). The power for parallel group design and the two period cross over design for follow up durations of 2, 3, 6, and 9 months was calculated as the proportion of trials which yielded a significant result. To evaluate the number of patients and the duration of follow up periods needed to achieve relatively stable estimates of the study power, we applied the same statistical approach to random subsets of the original data set (10, 25, and 40 patients with 9 months of follow up for the parallel group design and all the available patients with 3 and 6 months of follow up for the two period cross over design). The statistical tests used in simulations were the Wilcoxon signed rank test for the two period cross over design and the Wilcoxon rank sum test for the parallel group design.

Results

A total of 585 scans were obtained during the study period: 316 (54%) had one or more enhancing lesions (active scans). Nine patients had no enhancement for the entire study period, whereas 10 patients had always active scans. The total number of enhancing lesions was 1436: the mean number of enhancing lesions/patient/scan was 2.4 (range= 0–53).

The power of the parallel group design trials is presented in table 1. For each treatment effect, the statistical power of the trial increased with the duration of the follow up until 3 months duration, when using the number of new enhancing lesions as the end point (A.III). For instance, with an expected treatment effect of 70%, depending on the sample size, the absolute increase in statistical power of the trial was 14%–17% from 2 to 3 months of follow up duration and was 2% to 3% from 3 to 6 months duration, whereas no further gain in power was obtained from 6 to 9 months of follow up duration. When simulations were performed using the proportions of active scans as the end point (A.II), a similar trend was found, but the gain in statistical power was lower. For instance, for a treatment effect of 70% we found, depending on the sample sizes, an increase in statistical power of 4%–7% from 2 to 3 months of follow up duration, by 5%–7% from 3 to 6 months of follow up duration and by 1%–3% from 6 to 9 months of follow up duration.

The power of two period cross over trials are presented in the table 2. For this study design, the statistical power increased considerably with increasing duration of the follow up for both the variables studied (A.II and A.III). For instance, for an expected treatment effect of 30% and a sample size of 2×10 patients, the power increased by 13% from 2 to 3 months of follow up duration, by 27% from 3 to 6 months of follow up durations, and by 14% from 6 to 9 months of follow up duration. However, for higher treatment effects, no further gain in power was seen after 6 months of scanning, whereas, using active scans as the end point variable, the statistical power steadily increased until 9 months.

To test the simulation method performance and the stability of power estimates, we performed the same simulations by applying the resampling procedure on subsets of data randomly selected from the original data set and using A.III as the response variable. In table 3, the results for the parallel group design are reported. Firstly, we extracted random subsets of 10, 25, and 40 patients, each with 9

Table 5 Comparison of the results obtained in four studies using the same simulation approach in computing the statistical power for a parallel group design with follow up duration of 6 months and for expected treatment effects of 60% and 70%. The response variable studied was the number of new enhancing lesions

Sample size		*Truyen et al[7] (n=12)*	*Nauta et al[6] (n=23)*	*Tubridy et al[8] (n=31)*	*Present study (n=65)*
2×10	60	-	-	-	35
	70	84	71	61	57
2×20	60	80	70	63	50
	70	93	88	81	74
2×30	60	89	78	72	62
	70	99	95	89	85
2×40	60	94	87	83	74
	70	99	100	95	92

months of follow up, for the reference data set used to calculate the powers of parallel group trials. The power was calculated for an expected treatment effect of 70% and for different sample sizes. The power was increasingly overestimated when sampling from smaller data sets. Secondly, for the two period cross over design (table 4), subsets of 3 and 6 months of follow up were selected for the reference data set. The power was calculated for an expected treatment effect of 20% for different sample sizes with 3 months of follow up. Again, power steadily decreased with increasing the durations of follow up periods. These findings were confirmed by comparing the power estimates of the present study with those obtained by previous studies,[6–8] based on smaller sample sizes (table 5).

Discussion

In this study, we aimed to find out the optimal duration of follow up of MRI monitored clinical trials, with the minimum number of scans required to achieve a desired statistical power, when the end point variables are the number of new enhancing lesions and the number of active scans. We also investigated the performance of the computer simulation procedure adopted if applied to smaller data sets. The main result of the study is that when the end point variable is the number of new enhancing lesions as detected by monthly MRI, there is little gain in collecting more than three monthly scans for each patient for parallel group design trials, whereas six scans per patient for each treatment period are more desirable when planning two period cross over studies.

For the parallel group design, the statistical power increases very slowly when collecting more than three scans per patient. When such a study design is used, the factor that mainly affects the statistical power is the between subject variability, which is not substantially reduced by increasing the follow up duration. As already pointed out by Nauta *et al*,[6] adding up scans for a patient participating in a trial may just result in adding correlated information, which may be redundant to existing information. Thus as our simulations indicate, when planning parallel group design trials, it is more desirable to enroll a larger number of patients to be followed up for a period no longer than 3 months rather than enrolling fewer patients and following them up for longer periods.

On the contrary, for two period cross over design studies, the duration of the follow up clearly has a higher influence on the statistical power. In this case, the within patient variability is less of an issue, whereas patients with no enhancing lesions in the pretreatment period have a crucial role in determining the power of the study. These patients can, in fact, either increase the number of enhancing lesions or, at least, continue to be inactive during the treatment period and they do not have any possibility of "doing better" on treatment. Therefore, increasing the follow up durations increases the likelihood of finding an active scan in the pretreatment period, and, as a consequence, increases the likelihood of observing a treatment effect.

At present, computer simulations are the approach used to calculate sample sizes for multiple sclerosis trials, when the end point is lesion counting on MRI. The method we used, firstly proposed by Nauta *et al*,[6] has the advantage of being independent from any assumption about the distribution of lesion counts. In fact, the simulated treatment effect is not performed on population parameters (for example, a lesion appearance rate) and their related distributions, but directly on the observed lesion number of each patient in the data set. The set of "treated" patients is, in fact, the same as the "untreated" set, but with reduced lesion numbers. Therefore, by simulating the treatment as described above, a patient can only "do better" after the treatment initiation, whereas this is not the case in the daily life situation even in the case of very effective treatments. Thus it is not surprising that such a method works better when using large data sets from which performing the repeated sampling procedure**.**

The statistical power of trials is likely to be overestimated by this approach, when resampling from small patient cohorts for parallel group design trials or larger cohorts with short follow up periods for cross over design trials. For parallel group design trials, resampling from the same small cohort of patients would inevitably lead to the formation of two groups (treated and untreated) made up almost by the same subjects. Thus the between subject variability would be artificially lost with, as a consequence, a power overestimation. The same is true for crossover design trials when resampling from groups of patients with short follow up periods; in this case, the artificial reduction of the within subject variability would also result in power overestimation. A parametric simulation might be a preferable approach to sample size calculations for MRI monitored trials in multiple sclerosis. This might be obtained by resampling from a theoretical population using a fitted model. At present, however, no parametric model has been proposed to describe MRI lesion counts in multiple sclerosis. Therefore, we consider this issue a desirable future development in this area.

Table 5, which shows a comparison with previous studies,[6–8] confirms this issue. For this comparison, we chose the data calculated for the parallel group design with a duration of follow up of 6 months, as this is analysed by all the authors. The statistical power decreases with the size of the patients' reference data set used for resampling. For instance, the statistical power computed by Truyen *et al*,[7] who used a data set of 12 patients, is 30% higher than the power computed in the present study. Although we cannot exclude that these differences may be due, at least partially, to the different samples used (for example, different frequency of enhancement and proportions of active scans, different patient clinical subgroups studied, etc), the trend observed suggests that there is a lower limit in the number of patients and in the duration of follow up needed to perform an adequate resampling procedure.

The power estimates we obtained in the present study are clearly dependent on the characteristics of the patients' sample used as the reference data set and the MRI procedures used. Therefore, it would be desirable to develop an ad hoc computer program that could be used when setting up clinical trials to calculate the needed sample sizes on the basis of the local data. Nevertheless, the clinical characteristics of the patients we studied are similar to those of patients usually recruited for MRI monitored trials in multiple sclerosis. Similarly, the range of the MRI scanners used in this study was comparable with that of previously performed multicentre clinical trials in multiple sclerosis.[11] On the contrary, the MRI acquisition parameters used in the present study were somewhat more flexible than those allowed in clinical trials, and reflected the post hoc nature of the present study. This increased heterogeneity in data acquisition should, however, make our power estimates more conservative.

Recent studies have shown that the use of different MR techniques may result in a higher detection of enhancing lesions and active scans,[2] which might further increase the power of MR monitored clinical trials or require even smaller sample sizes or shorter follow up durations. An important caveat is that the duration of the follow up period should not only be determined by statistical considerations, but should also be based on the characteristics of the experimental treatment used. In fact, although studies with small sample sizes or short follow up periods may have enough statistical power, it is more likely that a small sample size will not be as fully representative of the response subsequently found in the general multiple sclerosis population, and that very short term follow up periods may miss a treatment effect that takes several months to develop. This is particularly true when considering that the simulation procedure is based on the assumption that an experimental treatment becomes immediately effective and has a constant efficacy over time. Therefore, in the case of treatments known to become effective after a certain period of time, it is advisable either to ignore the first scans collected or to delay the scanning period after treatment initiation. Small, short term studies will also have less opportunity for detecting relevant side effects. Finally a very short term study may fail to detect an earlier failure of treatment—for example, the number of enhancing lesions rises again after a few months due to the development of an antidrug antibody.

In conclusion, this study provides stable estimates of the optimal durations of the follow up needed for two period cross over and parallel group design studies. It suggests that, contrary to conventional wisdom, which suggests that at least six scans per patient should be obtained, only three scans per patient are needed when performing parallel group studies.

This study was supported by an EC funded program called MAGNIMS (ERBCHRXCT940684). Dr MP Sormani is supported by a grant from TEVA Italy.

1 Miller DH, Albert PS, Barkhof F, *et al.* Guidelines for the use of magnetic resonance techniques in monitoring the treatment of multiple sclerosis. *Ann Neurol* 1996;**39**:6–16.
2 Filippi M, Miller DH. MRI in the differential diagnosis and monitoring the treatment of multiple sclerosis. *Curr Opin Neurol* 1996;**9**:176–86.
3 Miller DH, Barkhof F, Berry I, *et al.* Magnetic resonance imaging in monitoring the treatment of multiple sclerosis: Concerted Action Guidelines. *J Neurol Neurosurg Psychiatry* 1991;**54**:683–8.
4 Miller DH, Barkhof F, Nauta JJP. Gadolinium enhancement increased the sensitivity of MRI in detecting disease activity in MS. *Brain* 1993;**116**:1077–94.
5 McFarland HF, Frank JA, Albert PS, *et al.* Using gadolinium-enhanced magnetic resonance imaging to monitor disease activity in multiple sclerosis. *Ann Neurol* 1992;**32**:758–66.
6 Nauta JJP, Thompson AJ, Barkhof F, *et al.* Magnetic resonance imaging in monitoring the treatment of multiple sclerosis patients: statistical power of parallel-groups and crossover designs. *J Neurol Sci* 1994;**122**:6–14.
7 Truyen L, Barkhof F, Tas M, *et al.* Specific power calculations for magnetic resonance imaging (MRI) in monitoring active relapsing-remitting multiple sclerosis (MS): implications for phase II therapeutic trials. *Multiple Sclerosis* 1997;**2**:283–90.
8 Tubridy N, Ader HJ, Barkhof F, *et al.* Exploratory treatment trials in multiple sclerosis using MRI: sample size calculations for relapsing remitting and secondary progressive subgroups using placebo controlled parallel groups. *J Neurol Neurosurg Psychiatry* 1998;**64**:50–55.
9 Poser CM, Paty DW, Scheinberg L, *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol* 1983;**13**:227–31.
10 Lublin FD, Reingold SC, the National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Defining the clinical course of multiple sclerosis: results of an international survey. *Neurology* 1996;**46**:907–11.
11 Filippi M, Horsfield MA, Ader HJ, *et al.* Guidelines for using quantitative measures of brain magnetic resonance imaging abnormalities in monitoring the treatment of multiple sclerosis. *Ann Neurol* 1998;**43**:499–506.