Sample sizes for brain atrophy outcomes in trials for secondary progressive multiple sclerosis

D.R. Altmann, DPhil
B. Jasperse, MD
F. Barkhof, PhD
K. Beckmann, MSc
M. Filippi, MD
L.D. Kappos, MD
P. Molyneux, MD
C.H. Polman, PhD
C. Pozzilli, MD
A.J. Thompson, FRCP
K. Wagner, MD
T.A. Yousry, FRCR
D.H. Miller, FRCP

Address correspondence and reprint requests to Dr. Dan R. Altmann, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK
daniel.altmann@lshtm.ac.uk

ABSTRACT

Background: Progressive brain atrophy in multiple sclerosis (MS) may reflect neuroaxonal and myelin loss and MRI measures of brain tissue loss are used as outcome measures in MS treatment trials. This study investigated sample sizes required to demonstrate reduction of brain atrophy using three outcome measures in a parallel group, placebo-controlled trial for secondary progressive MS (SPMS).

Methods: Data were taken from a cohort of 43 patients with SPMS who had been followed up with 6-monthly T1-weighted MRI for up to 3 years within the placebo arm of a therapeutic trial. Central cerebral volumes (CCVs) were measured using a semiautomated segmentation approach, and brain volume normalized for skull size (NBV) was measured using automated segmentation (SIENAX). Change in CCV and NBV was measured by subtraction of baseline from serial CCV and SIENAX images; in addition, percentage brain volume change relative to baseline was measured directly using a registration-based method (SIENA). Sample sizes for given treatment effects and power were calculated for standard analyses using parameters estimated from the sample.

Results: For a 2-year trial duration, minimum sample sizes per arm required to detect a 50% treatment effect at 80% power were 32 for SIENA, 69 for CCV, and 273 for SIENAX. Two-year minimum sample sizes were smaller than 1-year by 71% for SIENAX, 55% for CCV, and 44% for SIENA.

Conclusion: SIENA and central cerebral volume are feasible outcome measures for inclusion in placebo-controlled trials in secondary progressive multiple sclerosis. *Neurology*® 2009;72:595-601

GLOSSARY

ANCOVA = analysis of covariance; **CCV** = central cerebral volume; **FSL** = FMRIB Software Library; **MNI** = Montreal Neurological Institute; **MS** = multiple sclerosis; **NBV** = normalized brain volume; **PBVC** = percent brain volume change; **RRMS** = relapsing-remitting multiple sclerosis; **SPMS** = secondary progressive multiple sclerosis.

Definitive clinical trials of potential new disease-modifying agents in multiple sclerosis (MS) often evaluate disability as the primary outcome measure. Because MS is characterized by a variable but generally slow clinical evolution, controlled studies with disability endpoints require large numbers of patients (several hundreds) to be studied over several years. Accordingly, there is considerable interest in developing surrogate laboratory markers of disease progression that, if more sensitive than disability, would enable trials to be performed more quickly and with fewer patients.

Irreversible and progressive disability in MS is likely due to neuroaxonal loss and demyelination, which occur in focal white matter lesions¹ and also in normal-appearing white^{2,3} and gray

Supplemental data at
www.neurology.org

Editorial, page 586

e-Pub ahead of print on November 12, 2008, at www.neurology.org.

From the Medical Statistics Unit (D.R.A.), London School of Hygiene and Tropical Medicine, UK; Nuclear Magnetic Resonance Research Unit (D.R.A., D.H.M.), Department of Neuroinflammation, Institute of Neurology, University College London, UK; Multiple Sclerosis Centre Amsterdam and Image Analysis Centre (B.J., F.B., C.H.P.), VU University Medical Centre, The Netherlands; Bayer Schering Pharma AG (K.B., K.W.), Berlin, Germany; Neuroimaging Research Unit (M.F.), Department of Neurology, Scientific Institute and University Ospedale San Raffaele, Milan, Italy; Neurology and Department of Biomedicine (L.D.K.), University Hospital, Basel, Switzerland; Addenbrookes Hospital (P.M.), Cambridge, UK; Department of Neurological Science (C.P.), University of Rome "La Sapienza," Italy; and Department of Brain Repair and Rehabilitation (A.J.T., T.A.Y.), Institute of Neurology, University College London, UK.

The Nuclear Magnetic Resonance Research Unit is partly supported by The Multiple Sclerosis Society of Great Britain and Northern Ireland. The Multiple Sclerosis Centre Amsterdam is supported by the Dutch Foundation for MS Research (grant 05-538c).

Disclosure: Bayer Schering Pharma AG supported the data collection for this study. F.B., M.F., P.M., C.H.P., and D.H.M. have received honoraria from Bayer Schering Pharma AG (less than \$10,000). K.W. and K.B. are current employees of Bayer Schering Pharma AG.

matter.⁴ MRI-measured brain atrophy has been proposed as a marker of progressive axonal and myelin loss,⁵ and it is now often acquired as an outcome measure in phase III trials.⁶⁻⁸ If brain atrophy is to be used as a reliable outcome measure in clinical trials, power calculations are required not only to determine the sample sizes needed to show therapeutic efficacy, but also to help identify the most suitable atrophy outcome measures, which is our primary aim here. In this report, based on data acquired in a multicenter sample of placebo-treated subjects with secondary progressive MS (SPMS), we calculate and compare sample sizes required in a parallel-group, placebo-controlled trial for SPMS subjects, using three brain atrophy outcome measures: a semiautomated measure of a regional (central) cerebral volume that has previously been used in MS cohorts⁹⁻¹¹ and two whole-brain automated measures—SIENA and SIENAX—also used extensively.^{7,12,13} Two secondary aims are to contrast the sample sizes required for different trial durations and analyses and to examine the relationships between the three atrophy outcomes.

METHODS Patients. A substudy¹⁰ from five centers in a placebo-controlled trial of interferon beta-1b in SPMS acquired 6-monthly T1-weighted brain MRI over 3 years. There were 46 placebo-treated patients from the five centers (20 women, 26 men), 43 of which provided usable data. The mean age at entry was 40.9 years (SD 7.9 years), the mean disease duration was 13.4 years (SD 7.5 years), the mean time since evidence of progression was 3.8 years (SD 3.4 years), and the mean Expanded Disability Status Scale score was 5.2 (SD 1.1, range 3–6.5). These patients underwent 6-monthly T1-weighted spin echo MRI (repetition time 500–700 msec, echo time 5–25 msec, 256 × 256 matrix, 24-cm field of view) for 3 years with 5-mm-thick contiguous axial slices acquired through the brain on each occasion.

Brain atrophy measures. Central cerebral volume (CCV) was measured using an automated technique that segments cerebral tissue from surrounding scalp and other extracerebral tissue using a four-step algorithm. The details of the methodology are described elsewhere.^{9,10} The slices were chosen with the most caudal being at the level of the velum interpositum cerebri. Four contiguous, axial, 5-mm-thick slices were studied. This region of the cerebral hemispheres was chosen because in a previous study 1) there had been substantial atrophy seen over an 18-month period in subjects with SPMS⁹ and 2) the measure–reposition–rescan–remeasure coefficient of variability of the method was 0.56%.⁹

SIENAX was used to measure normalized brain volume.¹⁴ SIENAX automatically segments brain from nonbrain matter, calculates the brain volume, and applies a normalization factor to

correct for skull size. The normalization factor is obtained by registering the subject's scan to the Montreal Neurological Institute (MNI) 152 standard image using the skull to normalize spatially. Percentage brain volume change (PBVC) for each time point relative to baseline was measured using SIENA.¹⁴ SIENA registers the baseline and follow-up magnetic resonance image using the skull as scale and skew constraint, and then estimates the displacement of the brain edge for each point of the brain edge between these two scans. The brain edge displacements of all edge points are used to calculate the “overall” PBVC, which is expressed as a single value. Because not all scans included the full brain, the SIENAX and SIENA analyses were restricted to a pre-specified interval along the z-axis, ranging from –52 to +60 mm in standard MNI152 space. When necessary, errors in brain extraction were corrected manually by a single experienced observer; this has been shown previously¹³ to reduce unwanted variability in SIENA and SIENAX results without materially introducing interobserver/intercenter variability; all scans required manual correction to a varying extent. SIENAX and SIENA are part of the FMRIB Software Library (FSL).¹⁵ All SIENAX and SIENA analyses were performed using FSL version 3.1.

Statistical methods and issues. Sample size estimates were calculated for trial durations of 12, 24, and 36 months to detect treatment effects of 30%, 40%, 50%, and 60% at 80% and 90% power, all with a two-tailed α (significance level) of 5%. Treatment is assumed to have an immediate and constant effect, and in the absence of a healthy control group treatment effects assume zero atrophy in healthy subjects, 100% equating with zero volume loss. For each duration, three standard statistical analysis methods were considered for the comparisons between active and placebo trial groups: 1) comparison of the mean change from baseline, using a t test; 2) comparison of baseline adjusted mean change from baseline, using analysis of covariance (ANCOVA)¹⁶; and 3) comparison of mean rates of change estimated from longitudinal linear mixed models,¹⁷ using either 6-monthly or annual time points. Relative efficiencies are used to summarize comparisons: the relative efficiency of procedure A vs B is the inverse of the ratio of the corresponding sample sizes required to achieve the same power. These methods are discussed further below, but technical details of the statistical models and calculations are given in appendix e-1 on the *Neurology*[®] Web site at www.neurology.org.

A number of issues are relevant to the comparisons we present and to their potential impact on trial design. Chiefly, these relate to the choice of sample required to obtain valid comparisons between outcomes or between different trial durations or statistical analyses, and issues regarding outcome type.

Choice of samples for comparison. For the primary comparison, between atrophy measures, best estimates come from subjects with all three measures available at a given time point, “all-three” samples. This ensures that differences between measures are not due to different subjects. For these comparisons, at different time points, sample sizes were calculated just for a 50% treatment effect (because the relative efficiency of the volume measures is approximately constant over different treatment effects for a given analysis method and duration). For any given trial duration and analysis method, this gives a valid comparison across the atrophy measures. For the simplest analysis method, the t test of changes, the nonparametric bias-corrected bootstrap¹⁸ (1,000 replicates), was used to assess the statistical significance of sample size differences between the measures: standard errors for the differences in sample size estimates are not theoret-

Table 1 Volumes and changes from baseline for the three measures, by month, with numbers of patients contributing (maximum n = 43)

Month	Measure	CCV, cm ³ n; Mean (SD)	SIENAX, cm ³ n; Mean (SD)	SIENA, % n; Mean (SD)	No. with all three measures
0	Volumes	43; 307.87 (269.77)	41; 1,460.57 (63.15)	—	41*
6	Volumes	34; 302.97 (26.11)	40; 1,446.88 (55.04)	—	32
	Changes	34; -2.73 (4.17)	40; -11.23 (31.97)	—	
	[As % [†]]	[-0.89 (1.35)]	[-0.73 (2.14)]	40; -0.63 (0.78)	
12	Volumes	38; 304.95 (27.08)	40; 1,452.79 (58.82)	—	37
	Changes	38; -4.37 (5.26)	39; -7.68 (38.25)	—	
	[As % [†]]	[-1.40 (1.69)]	[-0.48 (2.59)]	39; -1.24 (1.05)	
18	Volumes	38; 301.27 (25.13)	37; 1,443.88 (66.52)	—	34
	Changes	38; -6.44 (9.11)	35; -12.19 (44.86)	—	
	[As % [†]]	[-2.02 (2.83)]	[-0.79 (3.09)]	35; -1.73 (1.19)	
24	Volumes	33; 300.36 (24.24)	36; 1,431.92 (61.82)	—	30
	Changes	33; -8.82 (10.66)	34; -25.55 (50.36)	—	
	[As % [†]]	[-2.76 (3.30)]	[-1.68 (3.40)]	34; -2.47 (1.57)	
	Correlation [§]	0.91	0.65	—	
30	Volumes	29; 298.99 (24.59)	29; 1,433.86 (68.52)	—	24
	Changes	29; -8.76 (10.64)	27; -22.39 (42.46)	—	
	[As % [†]]	[-2.74 (3.38)]	[-1.49 (2.93)]	27; -2.43 (1.64)	
36	Volumes	31; 293.73 (27.18)	30; 1,431.57 (75.41)	—	28
	Changes	31; -11.96 (10.96)	29; -26.28 (50.03)	—	
	[As % [†]]	[-3.86 (3.53)]	[-1.77 (3.50)]	29; -3.24 (2.23)	
	Correlation [§]	0.92	0.76	—	

*Unless otherwise indicated.

[†]Baseline implicit for SIENA in subjects with at least one later observation.

[‡]As percentage of baseline: mean (SD).

[§]Pearson correlation coefficients between absolute volume at time point, and baseline (all $p < 0.0001$).

CCV = central cerebral volume.

ically available, but in this context the bootstrap method gives a valid test, estimating confidence intervals for the differences empirically by multiple resampling (replicates) of the data. (p value ranges are given because of the computationally intensive nature of the bootstrap).

For best results within each individual measure and also for the secondary comparison between analysis methods and trial durations using a given measure, optimal estimates are given for each volume measure separately by fitting a longitudinal model using an “all-data” sample: the 36-month duration 6-monthly longitudinal model, which uses every available time point for that measure. Because the “all-three” samples have to drop a subject at a given time point if one of the three measures is missing, the “all-data” sample gives additional information on the robustness of the “all-three” comparisons to missing data. The estimated slope and variance parameters for the “all-data” model were then used to deduce the parameters relevant to the different statistical analyses and time points and thus generate the appropriate sample sizes. Thus, from the single set of “master” 36-month parameters, we obtain a valid comparison of the different analysis methods and durations in each measure, assuming constant atrophy over the period. Under this assumption, these parameters also allow estimation of the effect of altering observation times. It has been shown¹⁹ that the timing of observations is relevant to gains in power, e.g., adding a third observation midway between baseline and final follow-up provides no additional information with which to estimate linear

change. Though our primary aim is to compare the volume measures rather than establish optimal design, for interest we report some efficiency gains from a theoretically more efficient concentration of observations toward the trial period extremes.

Volume measures. The methodology of SIENA, calculating the percentage brain volume change (PBVC), is a “direct”²⁰ measure of change, with theoretically less measurement error compared to indirect measures of change obtained by numerical subtraction between volumes calculated at separate time points, as is required for CCV and SIENAX. The superior precision of SIENA compared with indirect volume measures has been noted previously in cohorts with relapsing–remitting MS (RRMS).^{21–23} However, direct difference methods have a different error structure than absolute measures, and this was taken account of in constructing the longitudinal models to estimate SIENA parameters.²⁰

To examine the concordance between the three measures, the “all-three” sample was used, with CCV and SIENAX converted into PBVC units using $100 \times (\text{volume at time point} - \text{baseline volume})/\text{baseline volume}$. Pearson correlation coefficients and Bland–Altman plots²⁴ were obtained, and the standard deviations of the measures were statistically compared using the Pitman test²⁵ for paired variances.

RESULTS Of the 46 patients available, a maximum of 43 patients were used in the analyses: 2 subjects were excluded having only SIENAX baseline and no other valid measurements (both dropped out at 6 months), and 1 subject with only baseline measures in CCV and SIENAX (6-month scan electronic data rejected and then dropped out at 12 months) was also excluded. The patients provided a maximum of 246 data points for the analyses. From a theoretical maximum of $43 \times 7 = 301$ observations, 55 were missing: 25 because of patient dropout, 3 because of scan nonacquisition, 17 because of electronic data rejection, 1 because of hard copy (and therefore electronic data) rejection, and 9 because of unavailable electronic data. Table 1 shows the number of patients with all three measures available at any one time point, along with summary statistics of changes in volume from baseline and, for CCV and SIENAX only, absolute volumes and correlations between baseline and later volumes.

Concordance between the volume measures. There was in general much better agreement between SIENA and CCV percentage changes than with SIENAX (table 1; figure). Concordance between the three measures is further detailed in appendix e-2; figure e-1, A–C; and figure e-2, A–C.

Comparison of sample size estimates between the measures. Table e-1 gives the parameter estimates on which the sample size calculations for the “all-three” comparisons are based. (Details of the longitudinal parameters are given in appendix e-1.) Longitudinal model residuals did not show any serious nonnormality. Table 2 shows sample size estimates for 50%

treatment effect across the three measures, but the sample size ratios (relative efficiencies) within any single row would be the same for other treatment effects. SIENA has relative efficiencies between 2 (36-month *t* test) and 2.5 (24-month *t* test) compared with CCV and between 6.8 (36-month longitudinal) and 31.8 (12-month *t* test) compared with SIENAX. CCV has relative efficiency between 3.2

(36-month longitudinal) and 15.2 (12-month *t* test) compared with SIENAX. Bootstrap inference, for the pairwise differences in *t* test sample sizes between measures, showed that all sample size differences were $p < 0.05$: in particular, SIENA vs SIENAX gave $p < 0.001$ at all three durations; SIENA vs CCV gave $0.03 < p < 0.04$ at 12 months, $0.004 < p < 0.005$ at 24 months, and $0.01 < p < 0.02$ at 36 months; and CCV vs SIENAX gave $0.001 < p < 0.002$ at 12 months, $0.02 < p < 0.03$ at 24 months, and $0.01 < p < 0.02$ at 36 months.

Comparison of sample size estimates between analysis methods and trial durations. Table e-2 gives the parameter estimates underlying these sample size calculations. Table e-3 shows the sample size estimates across the different analysis methods and trial durations, for each volume measure separately, allowing valid comparisons within the columns. For all measures, the most influential factor in determining sample sizes is trial duration. Minimum 2-year sample sizes per arm for 50% treatment effect at 80% power were 32 for SIENA, 69 for CCV, and 273 for SIENAX and were 71%, 55%, and 44% lower than corresponding 1-year sizes. Detailed comparisons between analysis methods and trial durations are presented in appendix e-3. Key points are that adding an observation at the midpoint of the follow-up period does not add relevant information to the baseline and

Table 2 Comparison* of the three measures for 50% treatment effect: n per trial arm

Statistical analysis	Patients [†] contributing to all three measures	Volume measure					
		CCV power		SIENAX power		SIENA power	
		80%	90%	80%	90%	80%	90%
t Test of change							
12 m–baseline	37	98	130	1,494	2,000	47	63
24 m–baseline	30	70	93	250	335	28	37
36 m–baseline	28	60	80	235	315	30	40
ANCOVA							
12 m	37	97	129	1,318	1,764	—	—
24 m	30	68	91	199	267	—	—
36 m	28	57	76	207	277	—	—
Longitudinal, with time points							
0, 6, 12	41:226	139	186	1,137	1,521	56	74
0, 6, 12, 18, 24	41:226	69	92	316	423	31	42
0, 6, 12, 18, 24, 30, 36	41:226	58	77	184	246	27	36

*Because different numbers of patients contributed to the estimates for different analyses and time points, comparisons are valid horizontally across the measures but not generally across analyses or time points.

†For change and analysis of covariance (ANCOVA), each patient contributed two data points toward estimating the parameters from which the sample sizes are calculated. For the longitudinal models, parameters were estimated from the largest sample, with 41 patients contributing a total of 226 data points.

CCV = central cerebral volume.

final scans, while the effect of additional informative (noncentral) time points for a given duration is greater the more variable the measure. Thus, additional informative time points have an impact for SIENAX, with its greater variability and lower correlation between times; but for CCV, and particularly for SIENA, adding time points between baseline and last follow-up gives little theoretical gain, even if the scans are clustered at the period extremes, provided there is negligible patient dropout.

DISCUSSION Sample sizes based on four volume measures including SIENA²¹ and SIENA precision²³ have been estimated previously in RRMS cohorts, reporting the superior precision of SIENA compared with indirect measures of volume change.

Our results show generally better agreement between CCV and SIENA than between either of these and SIENAX. Differences between CCV and SIENA may be because the latter is a registration-based method directly measuring brain volume changes, whereas the former involves numerical subtraction. Additionally, these differences may be due to using a greater portion of the brain for SIENA. Nevertheless, there was good agreement between these two measures, particularly regarding longitudinal trajectory.

Comparing the three measures for the same analyses/durations gives highest sample sizes for SIENAX, followed by CCV and then SIENA, with the advantage of SIENA more pronounced at shorter durations. These results are explained by the comparative standard deviations of the three measures, relative to treatment effects. Although the variability of SIENAX absolute volumes, as a percentage of the volume, is actually lower than for CCV, the SIENAX *changes* have much higher variability than the other two measures, leading to higher SIENAX sample sizes for the analyses of changes. For the longitudinal models, sample sizes over shorter durations are dominated by the within-subject standard deviation, which was highest relative to treatment effect for SIENAX and lowest for SIENA. Over longer durations, sample sizes are influenced more by the between-subject atrophy rate standard deviation, which was again highest for SIENAX and lowest for SIENA. Although some patients were lost to the “all-three” sample underlying direct between-measure comparisons, the general similarity in sample sizes from the “all-three” and the “all-data” samples suggest the between-measure comparison is robust to patient loss.

Although in theory analyzing CCV with adjustment for baseline intracranial volume would only reduce the variability between subjects at baseline rather than of atrophy rates and, therefore, may not

greatly enhance power in longitudinal studies, further work is required to assess the potential gains from such adjustment. Further work is also required to assess any change in power from calculating SIENA direct changes between consecutive time points, rather than from baseline as in these data; or from using ANCOVA to adjust SIENA for baseline SIENAX, though our data suggest little gain from this because ANCOVA results tend to approach but not improve on the corresponding longitudinal analysis with annual time points.

Detecting smaller treatment effects, or increasing test power, naturally increased the required sample sizes. Comparing analyses and durations, for all three measures, increasing the duration or the number of informative (i.e., not midway) time points reduced the required sample size, with increased duration generally having greater impact than number of time points. In general, “noisier” measures gain more than precise measures from an increase in the number of informative data points: thus, SIENAX gains the most from increasing the intrinsic power of the analysis by extending duration or adding points (particularly points toward the period extremes), followed by CCV, with the least gains for SIENA.

SIENA sample sizes for different trial durations have previously been estimated²¹ as 69 (1 year), 44 (2 years), and 40 (3 years), based on an RRMS cohort to be analyzed with *t* tests of change at 90% power and 50% treatment effect, close to our corresponding 77, 45, and 39 in an SPMS cohort (table e-3). This might suggest that—despite of the use of different T1-weighted sequences on which atrophy was measured (three-dimensional in the RRMS group, two-dimensional in the SPMS group)—the average rate of brain atrophy and its variance between subjects may be similar in RRMS and SPMS cohorts.²⁶ The SPMS cohort in our European trial of interferon beta-1b had more ongoing relapses and a shorter disease duration than the SPMS cohort that took part in a North American trial of interferon beta-1b,²⁷ and further work might investigate sample sizes in a longer-disease-duration nonrelapsing SPMS cohort.

One assumption that may exaggerate the study power is that 100% treatment effect equates to zero volume loss. However, healthy controls experience some brain volume loss (0.1%–0.3% per year), and if disease-specific treatment effects do not affect the “normal” atrophy associated with aging, a larger sample size will be required to show the same disease-specific effect. If 0.1% “healthy” annual loss is assumed, the SIENA sample size of 28 required for a 50% treatment effect, 80% power 3-year longitudinal analysis increases to 33; if 0.3% is assumed, the new sample size is 50. This effect might be allowed

for in analysis models where healthy controls are scanned using the same protocol.

Determining optimal trial design has to take careful consideration of issues such as dropout rate and scanning burden on patients, and is outside the scope of this article; we can here only highlight relevant factors. It is important to note that the relatively small gain in power for SIENA and CCV shown by multi-time point longitudinal analyses compared with *t* tests and ANCOVA conceals an important advantage of the more sophisticated models: missing one data point at either baseline or final follow-up will remove a subject from the simpler analyses, whereas the longitudinal models can use all available data points efficiently and thus minimize the impact of missing data, in terms of both power and potential bias from differential dropout. Possible dropout toward the end of follow-up may also limit the power gains from timing scans near the trial end rather than spacing them regularly.¹⁹

We assumed a linear volume change over time. Testing for nonlinearity, we found weak evidence of trajectories leveling off over time, consistent with a proportionate change, which is linear on a logarithmic volume scale. As a precaution, we repeated the sample size calculations on the log outcomes, but obtained sizes almost identical to those we report for SIENAX and SIENA and around 10% greater for CCV (probably because the changes tend to be larger as a proportion of absolute volumes for CCV than for the other measures). Further work on larger data sets would be required to assess possible nonlinearity satisfactorily.

For CCV and particularly for SIENA, extending the trial duration from 2 to 3 years reduces sample sizes relatively modestly. In contrast, extending the duration from 1 to 2 years can roughly halve the sample sizes required for these outcomes. A further disadvantage of 1-year duration is the possible short-term effect of biologic confounds tending to undermine sample size calculations, which, as here, assume immediate onset and constancy of treatment effect. First, any wallerian degeneration from axonal injury before the commencement of treatment may continue to evolve, and thus cause atrophy, for several months after the start of treatment, possibly delaying any treatment benefit from manifesting as reduced atrophy rate. Second, if the therapy has an anti-inflammatory as well as a neuroprotective effect, it may cause an initial decrease in brain volume due to resolution of inflammation. Such an effect has been proposed to contribute to decreases in brain volume seen after treatment with IV methylprednisolone,²⁸ beta interferon,^{6,29} and natalizumab.⁸ To avoid these confounds, baseline for analysis could be taken after

an initial treatment “burn in” period. The appropriate interval is uncertain, but 3 or 6 months might be considered reasonable.²⁹

AUTHOR CONTRIBUTIONS

Statistical analysis was conducted by D.R.A.

ACKNOWLEDGMENT

The authors thank Stenmar van Steenbrugge for assisting in the SIENA and SIENAX analyses and Chris Frost and Jonathan Bartlett for their statistical advice.

Received May 19, 2008. Accepted in final form August 20, 2008.

REFERENCES

1. Trapp BD, Peterson J, Ransohoff RM, et al. Axonal transection in the lesions of multiple sclerosis. *N Engl J Med* 1998;338:278–285.
2. Evangelou N, Esiri MM, Smith S, Palace J, Matthews PM. Quantitative pathological evidence for axonal loss in normal appearing white matter in multiple sclerosis. *Ann Neurol* 2000;47:391–395.
3. Kutzelnigg A, Lucchinetti CF, Stadelmann C, et al. Cortical demyelination and diffuse white matter injury in multiple sclerosis. *Brain* 2005;128(pt 11):2705–2712.
4. Peterson JW, Bö L, Mörk S, Chang A, Trapp BD. Transected neuritis, apoptotic neurons, and reduced inflammation in cortical multiple sclerosis lesions. *Ann Neurol* 2001;50:389–400.
5. Miller DH, Barkhof F, Frank JA, Parker GJM, Thompson AJ. Measurement of atrophy in multiple sclerosis: pathological basis, methodological aspects and clinical relevance. *Brain* 2002;125:1676–1695.
6. Rudick RA, Fisher E, Lee JC, Simon J, Jacobs L. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. Multiple Sclerosis Collaborative Research Group. *Neurology* 1999;53:1698–1704.
7. Filippi M, Rovaris M, Inglesse M, et al. Interferon beta-1a for brain tissue loss in patients at presentation with syndromes suggestive of multiple sclerosis: a randomised, double-blind, placebo-controlled trial. *Lancet* 2004;364:1489–1496.
8. Miller DH, Soon D, Fernando KT, et al. MRI outcomes in a placebo-controlled trial of natalizumab in relapsing MS. *Neurology* 2007;68:1390–1401.
9. Losseff NA, Wang L, Lai HM, et al. Progressive cerebral atrophy in multiple sclerosis: a serial MRI study. *Brain* 1996;119:2009–2019.
10. Molyneux PD, Kappos L, Polman C, et al. The effect of interferon beta-1b treatment on MRI measures of cerebral atrophy in secondary progressive multiple sclerosis. *Brain* 2000;123:2256–2263.
11. Stevenson VL, Smith SM, Matthews PM, Miller DH, Thompson AJ. Monitoring disease activity and progression in primary progressive multiple sclerosis using MRI: sub-voxel registration to identify lesion changes and to detect cerebral atrophy. *J Neurol* 2002;249:171–177.
12. Smith SM, De Stefano N, Jenkinson M, Matthews PM. Normalized accurate measurement of longitudinal brain change. *J Comput Assist Tomogr* 2001;25:466–475.
13. Jasperse B, Valsasina P, Neacsu V, et al. Intercenter agreement of brain atrophy measurement in multiple sclerosis patients using manually-edited SIENA and SIENAX. *J Magn Reson Imaging* 2007;26:881–885.

14. Smith SM, Zhang YY, Jenkinson M, et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* 2002;17:479–489.
15. Smith SM, Jenkinson M, Woolrich MW, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 2004;23(suppl 1):208–219.
16. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med* 1992;11:1685–1704.
17. Goldstein H. *Multilevel Statistical Models*. Kendall's Library of Statistics Series 3. London: Hodder Arnold; 1995.
18. Carpenter JR, Bithall JF. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000;19:1141–1164.
19. Schott JM, Frost C, Whitwell JL, et al. Combining short interval MRI in Alzheimer's disease: implications for therapeutic trials. *J Neurol* 2006;253:1147–1153.
20. Frost C, Kenward MG, Fox NC. The analysis of repeated "direct" measures of change illustrated with an application in longitudinal imaging. *Stat Med* 2004;23:3275–3286.
21. Anderson VM, Bartlett JW, Fox NC, Fisniku L, Miller DH. Detecting treatment effects on brain atrophy in relapsing remitting multiple sclerosis: sample size estimates. *J Neurol* 2007;254:1588–1594.
22. Anderson VM, Fernando KT, Davies GR, et al. Cerebral atrophy measurement in clinically isolated syndromes and relapsing remitting multiple sclerosis: a comparison of registration-based methods. *J Neuroimaging* 2007;17:61–68.
23. Sormani MP, Rovaris M, Valsasina P, Wolinsky JS, Comi G, Filippi M. Measurement error of two different techniques for brain atrophy assessment in multiple sclerosis. *Neurology* 2004;62:1432–1434.
24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–310.
25. Armitage P, Berry G, Matthews J. *Statistical Methods in Medical Research*, 4th ed. Oxford: Blackwell Science; 2002.
26. Kalkers NF, Ameziene N, Bot JC, Minneboo A, Polman CH, Barkhof F. Longitudinal brain volume measurement in multiple sclerosis: rate of brain atrophy is independent of the disease subtype. *Arch Neurol* 2002;59:1572–1576.
27. Panitch H, Miller A, Paty D, et al. Interferon beta-1b in secondary progressive MS: results from a 3-year controlled study. *Neurology* 2004;63:1788–1795.
28. Rao AB, Richert N, Howard T, et al. Methylprednisolone effect on brain volume and enhancing lesions in MS before and during IFNbeta-1b. *Neurology* 2002;59:688–694.
29. Hardmeier M, Wagenpfeil S, Freitag P, et al. Rate of brain atrophy in relapsing MS decreases during treatment with IFNbeta-1a. *Neurology* 2005;64:236–240.

Learn. Earn. Network.

2009 AAN Annual Meeting: An Excellent Value

- Learn about the latest scientific advances in neurology
- Earn valuable CME credit and fulfill Maintenance of Certification requirements
- Network with your peers at exciting social events all week long
- Enjoy the convenience and value of all this and more—in just one meeting

Early registration and hotel deadline is March 20, 2009.

Register today at www.aan.com/AM2009.

Support Research at the AAN Foundation Wine Tasting & Auction

You are cordially invited to sample some of the finest wines in the Pacific Northwest. Enjoy an array of food, music, and fun. This exclusive 2009 Annual Meeting event takes place Wednesday, April 29, from 7:00 p.m. to 9:30 p.m. in the Grand Ballroom at the Sheraton Hotel in Seattle. Bid on a variety of silent auction prizes including exquisite wine, entertainment, gifts, and more. The evening culminates with a live auction of exciting prizes. Bid high and often to support research. Proceeds benefit the AAN Foundation Research Program. Tickets are \$100 per guest. Tickets are limited, so go to www.aan.com/wine and register now!