Multiple Sclerosis Journal
http://msj.sagepub.com/

Trends in annualized relapse rates in relapsing–remitting multiple sclerosis and consequences for clinical trial design

Richard Nicholas, Sebastian Straube, Heinz Schmidli, Simon Schneider and Tim Friede

Mult Scler published online 17 May 2011
DOI: 10.1177/1352458511406309

The online version of this article can be found at:
http://msj.sagepub.com/content/early/2011/05/13/1352458511406309

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

Research Paper

MULTIPLE SCLEROSIS JOURNAL | MSJ

Multiple Sclerosis Journal
0(00) 1–7

DOI: 10.1177/1352458511406309
msj.sagepub.com
SAGE

Trends in annualized relapse rates in relapsing–remitting multiple sclerosis and consequences for clinical trial design

Richard Nicholas[1], Sebastian Straube[2], Heinz Schmidli[3], Simon Schneider[4] and Tim Friede[4]

Abstract

Background: Sample size calculation is a key aspect in the planning of any trial. Planning a randomized placebo-controlled trial in relapsing–remitting multiple sclerosis (RRMS) requires knowledge of the annualized relapse rate (ARR) in the placebo group.

Objectives: This paper aims (i) to characterize the uncertainty in ARR by conducting a systematic review of placebo-controlled, randomized trials in RRMS and by modelling the ARR over time; and (ii) to assess the feasibility and utility of blinded sample size re-estimation (BSSR) procedures in RRMS.

Methods: A systematic literature review was carried out by searching PubMed, Ovid Medline and the Cochrane Register of Controlled Trials. The placebo ARRs were modelled by negative binomial regression. Computer simulations were conducted to assess the utility of BSSR in RRMS.

Results: Data from 26 placebo-controlled randomized trials were included in this analysis. The placebo ARR decreased by 6.2% per year ($p < 0.0001$; 95% CI (4.2%; 8.1%)) resulting in substantial uncertainty in the planning of future trials. BSSR was shown to be feasible and to maintain power at a prespecified level also if the ARR was misspecified in the planning phase.

Conclusions: Our investigations confirmed previously reported trends in ARR. In this context adaptive strategies such as BSSR designs are recommended for consideration in the planning of future trials in RRMS.

Keywords

adaptive design, annualized relapse rate, blinded sample size review, computer simulations, relapsing–remitting multiple sclerosis, systematic review

Date received: 12th December 2010; revised: 13th February 2011; accepted: 15th March 2011

Introduction

Sample size calculation is a key aspect in the planning of any clinical trial due to the impact of numbers of subjects recruited on costs, trial duration, and number of recruiting centres. In the planning of a randomized controlled trial in relapsing–remitting multiple sclerosis (RRMS) with clinical relapses as the primary efficacy outcome, sample size calculations need to consider the following: the assumed annualized relapse rate (ARR) in the control group (or alternatively overall ARR across both treatment groups), a factor accounting for between-patient heterogeneity (extra-Poisson variation), the clinically relevant effect size, the length of

[1]Imperial College Healthcare NHS Trust, London, UK.
[2]Department of Occupational, Social and Environmental Medicine, University Medical Center Göttingen, Göttingen, Germany.
[3]Statistical Methodology, Novartis Pharma AG, Basel, Switzerland.
[4]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany.

Corresponding author:
Professor T Friede, University Medical Center Göttingen, Department of Medical Statistics, Humboldtallee 32, D-37073 Göttingen, Germany
Email: tim.friede@med.uni-goettingen.de

follow-up (commonly 1 or 2 years), the targeted power (usually 80 or 90%), and the significance level (usually 5%). Recently Inusah et al.[1] identified a downward trend in ARR for trials published between 1980 and 2008 and considerable variation between studies around this trend line, which results in considerable uncertainty in the planning of future trials in RRMS. This is further substantiated by very recent trials conducted to the highest standards that observed relapse rates very much lower than the relapse rates assumed in their sample size calculations.[2,3] For instance, Kappos et al.[2] report that a placebo ARR of 0.7 was expected in the sample size calculation whereas the observed ARR was 0.4. The draft FDA guidance on adaptive designs[4] recommends that 'sample size adjustment using blinded methods to maintain desired study power should generally be considered for most studies' (FDA Guidance document, Section V.B)[4]. Recently, blinded sample size adjustment strategies for recurrent events such as relapses were proposed.[5–7] The present paper aims to:

- characterize and quantify the uncertainty in ARR by conducting a systematic review of placebo-controlled, randomized trials in RRMS and by modelling the ARR over time
- assess the feasibility of blinded sample size re-estimation (BSSR) procedures in RRMS, and
- investigate the utility of BSSR procedures in RRMS by means of computer simulations of virtual trials in a scenario motivated by a recently published trial.

Finally, the advantages and disadvantages of BSSR procedures are discussed by comparing BSSR designs with unblinded group-sequential designs.

Methods

Literature review

We updated a recent systematic review[8] by searching PubMed, Ovid Medline and the Cochrane Register of Controlled Trials from 2008 to April 2010. The details of the search strategy are described in Appendix A provided as supplementary material.

Statistical methods

Modelling time trends in ARR. The total numbers of relapses in the studies were reconstructed by multiplying the ARR with the number of patients and follow-up per patient in years. With the number of relapses of individual patients during follow-up in a study following a negative binomial distribution[9] (assuming random withdrawal) with expectation μ and overdispersion parameter k, the total sum of relapses of that study follows a negative binomial distribution with parameters $n\times\mu$ and k/n.[10] This means that, for sufficiently large sample sizes, the total numbers of relapses in the studies follow Poisson distributions. Fitting a negative binomial regression on the total number of relapses with the logarithm of the total follow-up time as offset models potential between-study heterogeneity by a gamma random effect. For the estimated parameters and the predicted model values 95% Wald confidence intervals (CIs) were computed.

Sample size calculation. In a study with a fixed follow-up time t per patient and a 1:1 randomization to placebo and experimental treatment, the sample size n per group to achieve a power 1-β with a significance level α is given by

$$n = \sigma^2 \frac{1}{t\lambda}\left(1+\frac{1}{\theta}\right)\frac{(Z_\alpha + Z_\beta)^2}{(\log\theta)^2}$$

assuming a placebo ARR λ, a rate ratio θ and an extra-Poisson variation σ^2.[6,11] We assume that the primary analysis aims to show superiority of the experimental arm against placebo, and is based on a Poisson regression allowing for overdispersion (extra-Poisson variation). Similar sample size formulas are obtained when the primary analysis is based on a negative binomial model.[7] However, the dependence of the variance on the mean is different for the overdispersed Poisson distribution and the negative binomial distribution, which means that changing ARRs can result in changes in overdispersion in one model but not the other.

Blinded sample size re-estimation. The procedure consists of three steps: (1) initial sample size calculation; (2) blinded sample size review; and (3) final analysis.[12] The initial sample size calculation is carried out using the sample size formula given above based on initial estimate of the ARR, the smallest clinically relevant effect size and the extra-Poisson variation, which might be obtained from previous studies. Multiplying the initial sample size with the follow-up time in years gives an initial estimate of the required total length of follow-up (so-called person-years). Once a certain proportion of the initially estimated total follow-up data becomes available, the overall ARR across the two treatment groups and the extra-Poisson variation are estimated in a blinded fashion, i.e. without knowledge of the patients' treatment allocations. The required sample size is then reestimated based on these estimates and the prespecified clinically relevant effect size again using the sample size formula given above. Details of this procedure are given in Friede and Schmidli.[6]

Computer simulations

To evaluate the utility of the BSSR design described above in RRMS we adopted the recently proposed clinical scenario evaluation framework[13] in its refined version.[14,15] Information on the disease-specific features, such as ARR, were gathered through the systematic review described above. Two design options were considered: a conventional design with fixed sample size, and the described BSSR design with a sample size review once half of the follow-up data are available. The ARR data of the virtual trials were simulated by sampling from negative binomial distributions, which are considered appropriate statistical models to describe ARR data.[9] To assess the operating characteristics of the designs (e.g. statistical power) 100,000 trials per scenario were simulated. In order to conduct simulations close to clinical practice, the two parameters of the negative binomial distribution were estimated from published frequency tables (with potentially censored observations) by the maximum likelihood method.

Results

We identified 18 abstracts that were not included in the review by Sormani et al.[8] Of these, three studies[2,3,16] were suitable for inclusion in this review and were analysed together with the 23 trials previously identified by Sormani et al.,[8] giving a total of 26 randomized, placebo-controlled trials in RRMS. The placebo ARR of the 26 trials decreased by 6.2% per year ($p < 0.0001$; 95% CI (4.2%; 8.1%)), which translates to a reduction of almost 50% over 10 years. The between-study heterogeneity was very small with the negative binomial overdispersion k being 0.0466 (95% CI (0.0105, 0.0827)). This corresponds to an extra-Poisson variation of $\sigma^2 = 1.0466$ (1.0233) for ARR = 1 (0.5), i.e. a variance inflation by less than 5% by differences between studies. In Figure 1 the ARR is plotted against the year of publication. It is apparent that there is not only a pronounced downward trend over the years, but also considerable variation around this trend. This results in substantial uncertainty in the planning of a new trial regarding the sample size. As illustrated by Figure 2, the required number of person-years per treatment group strongly depends on the ARR in the placebo group and the extra-Poisson variation, which are both hard to predict in the planning phase of a trial. The assumed treatment effects of 25% and 40% were used in the planning of two recently completed trials[2,3] and serve as examples here. As detailed below, extra-Poisson variations between 1 and 2 are not untypical for ARR data. In all trials included in our review apart from one small trial, placebo ARR were in the range of 0.3–1.5 which corresponds to a factor of five in sample

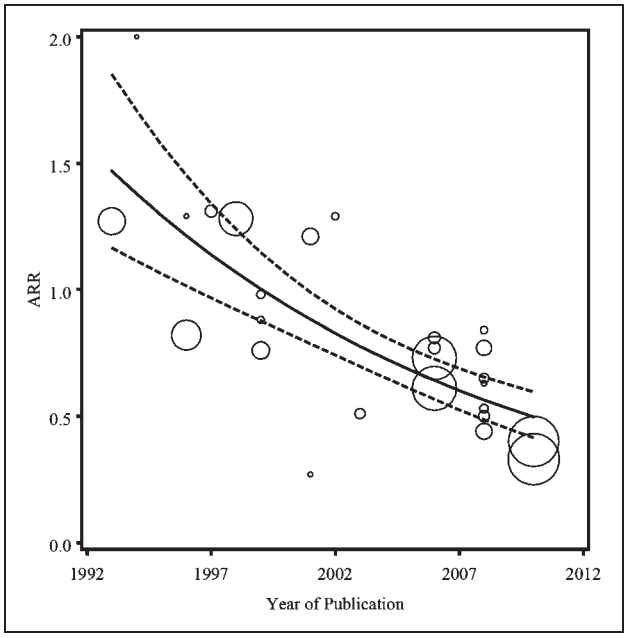

Figure 1. Annualized relapse rates observed in the 26 trials identified in the literature review against the calendar year in which the papers were published. The size of the bubble indicates the size of the trial and is proportional to the sample size. The solid line shows the model values of the negative binomial regression and the dashed lines indicate 95% confidence intervals.

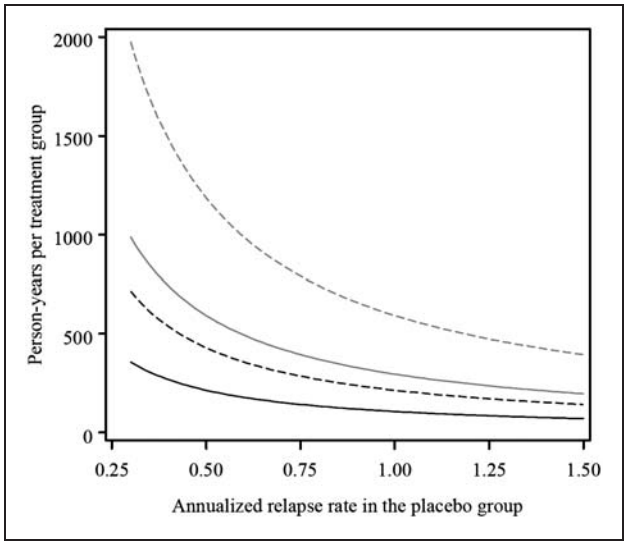

Figure 2. Required number of person-years per treatment group to achieve a power of 90% at the usual significance level of 5% depends on the annualized relapse rate in the placebo control group. The required person-years per treatment group increase for lower treatment effects (treatment effect: 25% (rate ratio of 0.75; black lines) and 40% (rate ratio of 0.60; grey lines)), but also with increased extra-Poisson variation (factor $\sigma^2 = 1$ (no extra variation; solid lines) and 2 (dashed lines)).

size, as can be seen from Figure 2. Even for trials published within a short span of time some considerable variation in the ARR in the placebo group remains. For instance, the placebo ARR for the seven trials published in 2008 varied between 0.44 and 0.84 (mean (SD) 0.62 (0.15); predicted model value 0.56) corresponding to sample size estimates that vary almost by a factor of two.

In the 26 trials identified in the literature review, the follow-up times were between 6 and 24 months with a median follow-up of 9 months. Commonly, the information on clinical relapses is collected regularly during follow-up. This means that also data from patients that have not fully completed follow-up yet can be included in an interim sample size review. Hence blinded sample size reviews are feasible in RRMS trials even if the follow-up times per patient are longer than the recruitment period. However, it should be noted that the BSSR procedure relies on the assumption of constant relapse rates during follow-up in this situation. If the recruitment period is actually longer than the follow-up period, the BSSR design becomes even more attractive in that the sample size re-estimation does not rely that heavily on this assumption. The utility of BSSR in RRMS is demonstrated in panel (a) of Figure 3, which shows the power of the BSSR procedure in comparison with the power of the standard fixed sample size design depending on the placebo ARR for scenarios motivated by the study described in Giovannoni et al.[3]

In the virtual BSSR design the interim review was carried out when 50% of the primary endpoint data were available. In the planning of the trial reported by Giovannoni et al.[3] a treatment effect of 25%, a placebo ARR of 1.14 and an extra-Poisson variation of about 2.3 were assumed, resulting in a sample size of 430 patients per group (after adjustment for withdrawals) to achieve a power of 90% at a Bonferroni adjusted significance level of 2.5%. From Table 2 in Giovannoni et al.[3] we estimated the actual extra-Poisson variation to be about 1.3, i.e. considerably lower than initially assumed. The observed placebo ARR was 0.33 (Table 2 in Giovannoni et al.[3]), also substantially lower than planned. The effects of these deviations from the planning assumptions on power in the conventional fixed sample size design can be observed in Figure 3. With the ARR decreasing from 1.14 to 0.33 the power dropped from 90% to 37% under the assumed extra-Poisson variation. Since the actual extra-Poisson variation was considerably lower than assumed, some of the drop in power was compensated with a power of 61% under the observed values for ARR and extra-Poisson variation. However, this was still substantially below the target power of 90%. For an ARR larger than 0.65 the drop in ARR would have even been overcompensated by the lower than assumed extra-Poisson variation leading to studies with power well above 90%, i.e. unnecessarily large studies. In contrast, the power of the BSSR design was always close to the target value of 90% for all values of the ARR and extra-Poisson variation. This is achieved by allocating resources where they are needed. This is shown in Figure 3(b), which displays the average sample size in the BSSR design against the placebo ARR. When the planning assumptions are correct (placebo ARR of 1.14, extra-Poisson variation of 2.3) the average sample size is with 388 patients per group, only slightly higher than the sample size of the fixed design ($n = 383$ without adjustment for withdrawal). In the BSSR design, however, the sample size is variable, which makes individual trials less predictable, for example, with regard to budget and drug supply. When the ARR is overestimated or the extra-Poisson variation was underestimated in the planning phase, the BSSR procedure will allocate additional patients to the study arms so that the desired power is maintained, whereas in situations where the ARR was underestimated or the extra-Poisson variation was overestimated, the BSSR design can lead to savings in sample size. Figure 3 demonstrates that the BSSR procedure can be applied in RRMS to minimize the risks of too small and thereby inconclusive studies on the one hand, and unnecessary large studies on the other hand.

Discussion

The advent of therapies in MS has revolutionized its management, but have yet to impact on long-term outcomes.[17] As a result, current FDA and EMA regulations continue to support the use of placebo in RRMS trials of MS therapies.[18] However, the development of new therapies based on a placebo approach has become increasingly difficult, as patients have access to an increasingly diverse range of treatments. Although trialists have used a number of strategies to maintain recruitment into placebo studies, such as inequality of access to treatment in different healthcare systems and countries around the world, here we have demonstrated a prominent downward trend in ARR, the outcome of interest to regulators, with time in the placebo group. After adjusting for year of publication, no substantial between-study heterogeneity in ARR was observed. This is consistent with findings by Inusah et al.,[1] who did not identify any further predictors apart from possibly mean age at baseline when adjusting for year of publication. Although different search criteria (search date: April 2010 vs. November 2008 in Inusah et al.;[1] magnetic resonance imaging (MRI) data required vs.

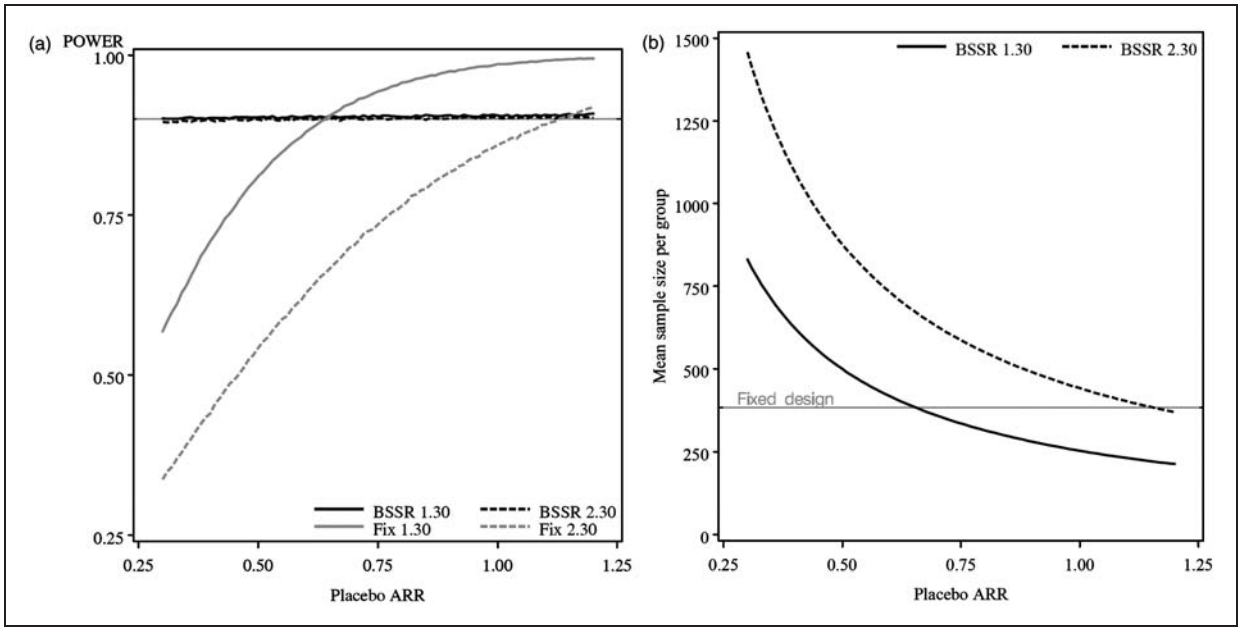

Figure 3. Power (a) and mean sample size (b) per group of the blinded sample size reestimation (BSSR) procedure (black lines) in comparison with the power of the standard fixed sample size design (grey curves in (a) and grey reference line in (b)) depending on the placebo annualized relapse rate (ARR) for scenarios motivated by the Giovannoni et al.[3] study (i.e. treatment effect of 25%, target power 90%, extra-Poisson variation factor $\sigma^2 = 1.3$ (solid line) and 2.3 (dotted line)). In the BSSR design the interim review was carried out when 50% of the primary endpoint data were available.

no MRI data required in Inusah et al.[1]) resulted in differing numbers of included studies (26 vs. 32 in Inusah et al.,[1] 17 shared) and different statistical models for data analysis were used in Inusah et al.[1] and in the present study, the identified trends in ARR are very similar, which provides strong support for the findings. By narrowing the search criteria to trials that reported MRI data the intention was to focus on trials that used an additional efficacy endpoint (Committee for Medicinal Products for Human Use,[18] Section 3.2.2), giving potentially higher quality and reliability. Given the significance of early relapses in MS on MS prognosis,[19] the drop in ARR makes it unclear whether placebo groups in early studies are comparable with placebo groups in later studies, where the ARR can be threefold lower. As a result, this trend makes it increasingly difficult to compare new products with established competitors both for regulators and organizations developing evidence-based guidelines (e.g. National Institute of Health and Clinical Excellence in the UK) and individual clinicians.

This trend means that the sample size required in studies must be increased to compensate for any potential lowering in the ARR that may take place during the timeframe of the study. This adds to the expense of investigating new potential therapies but can be predicted based on the model used here. However, a further issue could arise with the sample size calculation as a result of the variation in ARR around this downward trend in ARR. This variability could result in a trial being over or under-powered if fixed at the planning stage, increasing the risk of failure of the study. Although the BSSR design can be used to achieve savings in sample size by reducing the sample size below an initially calculated sample size if this is indicated by the sample size review, in some situations one might prefer to consider only potential increases of sample size, as overall costs might not be substantially reduced by a decrease in sample size due to fixed infrastructure costs. Other considerations influencing sample size include the assessment of safety, where a number of patients may be required. In these settings, the BSSR procedure can be used as a 'safety net' that adds a robustness to the trial by raising the sample size if further evidence arises that the (conservative) planning assumptions were incorrect.

We have demonstrated using computer simulations that this technique can preserve power as the trial progresses. Power is maintained at the required target level for a clinically relevant effect size independent of ARR, reducing any uncertainty in the outcome. The BSSR technique utilizes a transparent methodology that maintains statistical rigour and optimizes costs. It is relatively simple, requiring one further assessment at the midpoint of a study that can be implemented

without changing current recruitment strategies, making it a feasible addition to a study.

Regulatory guidelines require that blinding and control of the type I error rate are maintained when applying adaptive designs in late phase clinical trials.[4,20,21] The adaptation in the BSSR procedure is based on blinded data only and therefore maintains the trial's integrity. Furthermore, it was demonstrated elsewhere that the type I error rate is not affected by the BSSR procedure.[6,7] Hence, the BSSR fulfils the requirements set out by the international guidelines.

The BSSR technique does have limitations; there is no early stopping for futility or with rejection of the null hypothesis as in unblinded (adaptive) group-sequential designs.[22] In two recent trials, the observed treatment effects were substantially larger than the effects assumed in the sample size calculations. Giovannoni et al.[3] report that a 55% reduction was observed although 25% reduction was expected, and Kappos et al.[2] assumed a 40% reduction and actually observed a decrease of 60%. In these situations, efficacy of the treatments might have been demonstrated with smaller numbers of patients if a group-sequential design had been applied. However, the implementation of a group-sequential design is slightly more complex in that it requires an independent Data Monitoring Committee. Also the group-sequential design might not be attractive if complete follow-up for each patient is required, as recruitment time is typically shorter than follow-up time. As any adaptive design also the BSSR design requires careful planning including clinical but also operational, regulatory and statistical considerations to assess benefits and costs against a traditional fixed design.[23]

As the MS environment changes, active controlled trials might become more common.[24] Although not considered in this paper, the statistical procedures underlying the BSSR procedure are also applicable to active control trials planned to demonstrate superiority or non-inferiority.[7]

Acknowledgements

RN is grateful for support from the NIHR Biomedical Research Centre.

Funding

Work undertaken at Imperial College Healthcare NHS Trust received a proportion of funding from the Department of Health's NIHR Biomedical Research Centers funding scheme.

Conflict of interest statement

The authors have no commercial interests that might pose a conflict of interest in connection with the submission of this manuscript.

References

1. Inusah S, Sormani MP, Cofield SS, Aban IB, Musani SK, Srinivasasainagendra V, et al. Assessing changes in relapse rates in multiple sclerosis. *Mult Scler* 2010; 16: 1414–1421.
2. Kappos L, Radue EW, O'Connor P, Polman C, Hohlfeld R, Calabresi P, et al. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *N Engl J Med* 2010; 362: 387–401.
3. Giovannoni G, Comi G, Cook S, Rammohan K, Rieckmann P, Soelberg Sørensen P, et al. A placebo-controlled trial of oral cladribine for relapsing multiple sclerosis. *N Engl J Med* 2010; 362: 416–26.
4. Food and Drug Administration (FDA). *Guidance for industry – Adaptive design clinical trials for drugs and biologics. Draft guidance, 2010*. Food and Drug Administration, 2010.
5. Cook RJ, Bergeron PJ, Boher JM and Lie Y. Two-stage design of clinical trials involving recurrent events. *Stat Med* 2009; 28: 2617–2638.
6. Friede T and Schmidli H. Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Stat Med* 2010; 29: 1145–1156.
7. Friede T and Schmidli H. Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. *Methods Inf Med* 2010; 49: 618–624.
8. Sormani MP, Laura Bonzano L, Roccatagliata L, Cutter GR, Mancardi GL and Bruzzi P. Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: a meta-analytic approach. *Ann Neurol* 2009; 65: 268–275.
9. Wang YC, Meyerson L, Tang YQ and Qian N. Statistical methods for the analysis of relapse data in MS clinical trials. *J Neurol Sci* 2009; 285: 206–211.
10. Furman E. On the convolution of the negative binomial random variables. *Stat Probab Lett* 2007; 77: 169–172.
11. Ng HKT and Tang M-L. Testing the equality of two Poisson means using the rate ratio. *Stat Med* 2005; 24: 955–965.
12. Friede T and Kieser M. Sample size recalculation in internal pilot study designs: a review. *Biom J* 2006; 48: 537–555.
13. Benda N, Branson M, Maurer W and Friede T. Aspects of modernizing drug development using scenario planning and evaluation. *Drug Inf J* 2010; 44: 299–315.
14. Chataway J, Nicholas R, Todd S, Parsons N, Todd S, Miller D, et al. A novel adaptive design strategy increases the efficiency of clinical trials in secondary progressive multiple sclerosis. *Mult Scler* 2011; 17: 81–88.
15. Friede T, Nicholas R, Stallard N, Todd S, Parsons N, Valdés-Márquez E, et al. Refinement of the Clinical Scenario Evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug Inf J* 2010; 44: 713–718.
16. Kappos L, Gold R, Miller DH, MacManus DG, Havrdova E, Limmroth V, et al. Efficacy and safety of oral fumarate in patients with relapsing-remitting multiple sclerosis: a multicentre, randomised, double-blind, placebo-controlled phase IIb study. *Lancet* 2008; 372: 1463–1472.

17. Ebers GC, Traboulsee A, Li D, Langdon D, Reder AT, Goodin DS, et al. Analysis of clinical outcomes according to original treatment groups 16 years after the pivotal IFNB-1b trial. *J Neurol Neurosurg Psychiatry* 2010; 81: 907–912.
18. Committee for Medicinal Products for Human Use. *Guideline on clinical investigation of medicinal products for the treatment of multiple sclerosis. Doc. Ref. EMEA/CHMP/EWP/561/98 Rev 1*. London: CHMP, 2005.
19. Scalfari A, Neuhaus A, Degenhardt A, Rice GP, Muraro PA, Daumer M, et al. The natural history of multiple sclerosis, a geographically based study 10: relapses and long-term disability. *Brain* 2010; 133: 1914–1929.
20. ICH Harmonised Tripartite Guideline E9: Statistical principles for clinical trials. *Stat Med* 1999; 18: 1905–1942.
21. European Medicines Agency (EMA). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. CHMP/EWP/2459/02, 2007*. Available from: http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf (accessed 12 February 2011).
22. Jennison C and Turnbull BW. *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC, 2000.
23. Gaydos B, Anderson K, Berry DA, Burnham N, Chuang-Stein C, Dudinak J, et al. Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Inf J* 2009; 43: 539–556.
24. Polman CH, Reingold SC, Barkhof F, Calabresi PA, Clanet M, Cohen JA, et al. Ethics of placebo-controlled clinical trials in multiple sclerosis: a reassessment. *Neurology* 2008; 70: 1134–1140.