"Summary measure" statistic for assessing the outcome of treatment trials in relapsing-remitting multiple sclerosis

Clarence Liu, Alain Li Wan Po, Lance D Blumhardt

Abstract

***Objectives*—To review the outcome measures commonly used in phase III treatment trials of relapsing-remitting multiple sclerosis and to introduce a method of data analysis which is clinically appropriate for the often reversible disability in this type of multiple sclerosis.**

***Methods*—The conventional end point measures for disability change are inadequate and potentially misleading. Those using the disability difference between study entry and completion do not take into account serial data or disease fluctuations. Rigid definitions of "disease progression" based on two measurements of change in disability several months apart, do not assess worsening after the defined "end point", nor the significant proportion of erroneous "treatment failures" which result from subsequent recovery from relapses that outlast the end point. Assessing attacks merely by counting their frequency ignores the variation in magnitude and duration. These problems can be largely circumvented by integrating the area under a disability-time curve (AUC), a technique which utilises all serial measurements at scheduled visits and during relapses to summarise the total neurological dysfunction experienced by an individual patient on any particular clinical scale during a study period.**

***Conclusions*—The "summary measure" statistic AUC incorporates both transient and progressive disability into an overall estimate of the dysfunction that was experienced by a patient during a period of time. It is statistically more powerful and clinically more meaningful than conventional methods of assessing disability changes, particularly for trials which are too short to expect to disclose major treatment effects on irreversible disability in patients with a fluctuating disease.**

(*J Neurol Neurosurg Psychiatry* 1998;**64**:726–729)

Keywords: multiple sclerosis; outcome measures; disability

Division of Clinical Neurology, Faculty of Medicine, University Hospital, Queen's Medical Centre, Nottingham NG7 2UH, UK
C Liu
L D Blumhardt

Centre for Evidence-Based Pharmacotherapy, Department of Pharmaceutical Sciences, University of Nottingham, Nottingham NG7 2RD, UK
A Li Wan Po

Correspondence to:
Professor L D Blumhardt, Division of Clinical Neurology, Faculty of Medicine, University Hospital, Queen's Medical Centre, Nottingham NG7 2UH, UK.

Received 19 November 1997 and in revised form 11 February 1998
Accepted 16 February 1998

Multicentre double blind placebo controlled phase III trials of various immunomodulatory agents (interferon β-1b (IFNβ-1b, Betaseron), interferon β-1a (IFNβ-1a, Avonex), copolymer-1 (Cop-1 or glatiramer acetate, Copaxone) and intravenous immunoglobulin (IVIg)) for relapsing-remitting multiple sclerosis in the past few years have all demonstrated a modest reduction in the mean number of relapses but little or no significant effects on disability.[1–4] This should hardly be surprising given the very slow average accumulation of fixed neurological deficits arising from incompletely resolved attacks[5] and the low incidence of secondary progression in the patients selected for these trials. Much criticism has been directed at the well recognised limitations of the clinical rating scales used in these trials and considerable efforts are being made to improve the methodology of assessing impairment, disability, and handicap.[6] However, we think that there is also room for improvement in the methods of analysis of both relapse and disability data. More appropriate statistical methods may play an important part in the interpretation of results, whatever the inadequacies of the rating scales used.

We firstly discuss some of the problems of analysing serial data in relapsing-remitting multiple sclerosis, and secondly suggest a simple alternative statistical method which is both more appropriate to the natural history of this condition and more likely to capture any treatment effect on the time course of this fluctuating disease.

Disability data

The validity of using rigidly defined end points to measure disease progression in relapsing-remitting multiple sclerosis is fundamentally flawed due to the frequent remissions that occur early in the course of the disease. The table summarises the clinical end points which have been used in each of the recent major phase III trials under discussion. Regular neurological examinations were carried out at either three-monthly (IFNβ-1b and Cop-1) or six-monthly (IFNβ-1a and IVIg) intervals, but the published disability outcomes were derived either from determining the numbers of patients who sustained a deterioration in their expanded disability status score (EDSS)[7] for a period of either three or six months, or from the difference in disability scores between baseline and the end of the study. Considerable amounts of the disability data collected were not utilised in these analyses.

In a disease which characteristically remits after exacerbations, a method for determining "progression" which relies on a certain deterioration of clinical ratings to be sustained for three or six months, will not incorporate any subsequent worsening that might occur after

Efficacy criteria in phase III trials for relapsing-remitting multiple sclerosis

Trials	*IFNβ-1b*	*IFNβ-1a*	*Cop-1*	*IVIg*
Primary end points	(1) Annual relapse rate (2) % Relapse free patients	(1) Time to onset of sustained (6 months) deterioration by 1.0 point on EDSS	(1) Mean No of relapses	(1) Mean change in EDSS in 2 years (2) % patients with increase/decrease by 1.0 point or no change on EDSS
Clinical secondary end points	(1) Time to 1st relapse (2) Relapse severity (3) Relapse duration (4) Change in EDSS and NRS	(1) % Patients progressing (2) Intrapatient EDSS change (3) No of relapses per patient	(1) % Relapse free patients (2) Time to 1st relapse (3) % Sustained (3 months) 1.0 point EDSS change (4) Change in EDSS and AI in 2 years	(1) Mean No of relapses (2) Annual relapse rate (3) % Relapse free patients (4) Time to 1st relapse

IFNβ-1b=interferon β-1b[1]; IFNβ-1a=interferon β-1a[2]; Cop-1=copolymer-1[3]; IVIg= intravenous immunoglobulin[4]; EDSS=expanded disability status scale[7]; NRS= Scripps neurological rating scale[19]; AI=ambulation index.[20]

this defined period. Furthermore, this type of end point will also result in a significant proportion of erroneous "treatment failures"—that is, some patients who have met the criteria for "progression" will subsequently recover to baseline levels after a period of so called "sustained deterioration". This phenomenon is common at the lower end of the EDSS where most patients recruited for these trials tend to cluster. Natural history studies on cohorts with early multiple sclerosis have shown that up to 24% of relapses last more than three months.[8 9] Similar criticisms can be applied to the Kaplan-Meier survival analysis, which has been used to study time to confirmed progression in disability,[2 10] as such a method again inappropriately assumes that the exacerbations concerned are irreversible.

The second commonly used end point, which relies on mean change over a trial period (subtracting final from initial scores), is also problematic, as it ignores any transient disability due to attacks experienced during the study, the reduction of which may be the major beneficial effect of the therapy under investigation. This reliance on the difference between assessments at the start and the end of a trial wastes all the intermediate disability data, and is not statistically or clinically meaningful.[11]

A popular way of presenting serial data is to plot mean group scores as a time series. This has not been carried out explicitly in the published phase III trials, but has been illustrated in a recent MRI study involving patients on two doses of IFNβ-1a.[12] In this paper the mean monthly number and volume of MRI enhancing lesions were plotted versus time and analysed by Student's *t* test. This type of analysis has several problems.[13] The curves joining the means may not be representative of individual disease courses, the data from each patient over time is ignored when every time point has been analysed separately, and the means at successive points are not independent as they are to some extent influenced by the values of preceding data.[14]

Another technique for evaluating serial data is the analysis of variance (ANOVA), but the assumptions for its use in treatment trials may not always be valid; complete data sets are required (analysis of ad hoc assessments associated with relapses presents difficulties) as there are problems with treating missing values[15] and the method has also been considered difficult to understand and interpret.[16]

Relapse data

The relapse data from published phase III treatment trials has been subject to many criticisms. Unresolved issues include the methods of assessing exacerbations, which are beyond the scope of this article. It has been suggested that the usefulness of relapses as a trial end point would be enhanced if a meaningful measure of the disability caused by each attack was available.[17] In three out of the four trials in question, all relapses required objective confirmation by neurological examination.[2–4] In the IFNβ-1b study, although subjectively reported events were accepted as relapses, up to 80% of attacks were verified by examination and graded according to severity.[1] However, despite this admirable but time consuming and arduous requirement of the trial investigators, the clinical end points were derived solely from the number of counted relapses (table). A substantial amount of data on the severity and the time course of attacks was not utilised. Furthermore, comparisons of "annual exacerbation rates", although widely used, are conditional on the assumption that an individual patient's attack rate during the study is independent of his baseline relapse frequency.[2]

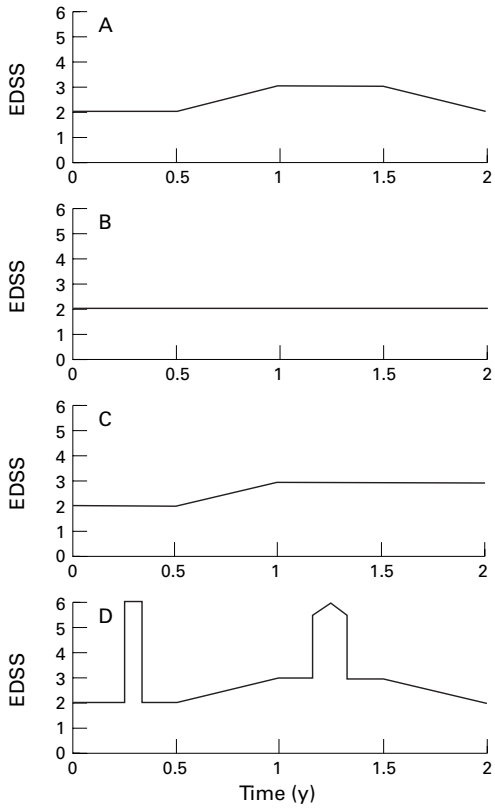
Use of "summary measure" as outcome measure

An alternative approach is to use a summary measure which captures changes in disease status over the whole study. The essential components are the sets of serial impairment or disability data, preferably with frequent sampling points for each patient, which are integrated to produce a single numerical summary of a particular dysfunction curve over the trial period. This is done by plotting an individual patient's scores serially against time. The summary measure is then obtained by calculating the area under the curve (AUC).[13] This method has the advantage that many of the statistical problems outlined above are avoided. In particular, in treatment trials of relapsing-remitting multiple sclerosis all scores acquired serially can be used and patients with missing data need not be discarded from the analysis. Different AUCs can be calculated for each clinical scale employed in the study.

In practice, there are two main methods used for calculating the AUC. The first integrates the area contained between the plot of consecutively acquired clinical rating scores during the trial and a baseline defined by the zero point on the particular clinical scale, to give a summary measure of the total disability for each patient (see appendix). The second version calculates the AUCs with respect to the baseline disability score for each patient at trial entry. This method improves the power of the summary measure, particularly if the changes in disability during the trial are small relative to the variance of the cohort at study entry. The main disadvantage is that statistical independence is lost as the

normalised AUCs depend on the stability of the baseline scores. Instability due to a recent resolving relapse may be avoided to some extent by ensuring a neurologically stable run in period. This second technique has been utilised in one study of disability in progressive multiple sclerosis to date.[18]

In the figure, EDSS scores obtained at six-monthly scheduled visits are shown for four hypothetical but typical patient courses, plotted versus time according to the trapezium rule. Note that the outcome in examples A and B would be the same—namely, "no change in disability" if the end point of an alteration in EDSS over two years was employed, whereas A and C would both become treatment failures if "sustained deterioration" of 1.0 point in EDSS was used. The calculated summary measures in these three examples are 5, 4, and 5.25 EDSS-years respectively using raw scores, and +1, 0, and +1.25 respectively, when normalised to entry baseline. If disability is also scored during relapses, then the episodes and the temporary dysfunction arising from them can be incorporated into an individual patient's summary score for the whole trial period. Unequal time intervals between data points are permitted, which is particularly useful for the unpredictable disease course of relapsing-remitting multiple sclerosis. In the figure, example D has the same "baseline disability" time course as A, but when the EDSS scores during two relapses are incorporated into the AUC, the summary measure is increased from 5 to 5.79 using raw data, and from +1 to +1.79 with data normalised to entry baseline.



Four hypothetical scenarios of disability scores (in EDSS points) plotted versus time in a two year relapsing-remitting multiple sclerosis treatment trial. The summary measures (in EDSS-years) for examples A to D are 5, 4, 5.25, and 5.79 respectively using raw data, and +1, 0, +1.25, and +1.79 respectively when normalised to baseline. Note that the plots are extrapolated between clinical assessment points according to the trapezium rule (see appendix for details).

Ideally, to improve the accuracy of any particular dysfunction curve, more frequent assessments should be made during exacerbations, to acquire more detailed information on the relapse onset and offset as well as the duration and magnitude of any transient disability. As it is cumbersome and inconvenient for patients to have repeated neurological examinations at times of increased disability, alternative scales being developed for rapid and easy administration, such as the Guy's Neurological disability scale,[19] may be particularly useful for this purpose.

For any neurological rating scale (for example, EDSS, Scripps neurological rating scale,[20] ambulation index,[21] or individual Kurtzke functional system scores[7]), the AUC obtained from serial scheduled time points can be compared with the AUC incorporating additional measurements obtained during relapses. The difference between the two values can be interpreted, to some extent, as an approximation of the short lived effects of exacerbations (for example, comparing the EDSS summary measures from examples A and D in the figure). Caution is necessary in short trials of two or three years, as fixed neurological deficits are accumulating very slowly, and an increased AUC at the end of a trial may simply represent transient disability which has either resolved or has yet to resolve. Classification into different subgroups for further analysis depending on the individual time point curves may be necessary.[13] In addition, differences in sampling frequency between patients and controls can introduce bias which may require weighting adjustment. Also, as for other statistical methods, the problems of the ordinal nature of disability scales such as the EDSS, as well as the "noise" introduced by within rater variability, remain.[22 23] Nevertheless, the summation of disability provided by the AUC, whether transient or fixed, provides a more clinically meaningful measure, particularly within the time constraints of these relatively short trials, by which to judge the effectiveness of a new therapy for relapsing-remitting multiple sclerosis.

Why has this method not been used in treatment trials of relapsing-remitting multiple sclerosis? Summary measure statistics have been in use since 1938,[24] although the technique was rarely employed in medicine until the past decade. In neurology, it has been utilised as a primary outcome variable in a headache treatment trial (using serial raw data)[25] and in a pilot study for rehabilitation in progressive multiple sclerosis (with summary measures of change of EDSS from baseline).[18] Relapsing-remitting multiple sclerosis treatment trials lasting two to three years are probably not long enough to demonstrate any meaningful effects on irreversible disability and evaluating relapses by merely counting them is an oversimplification. Summary measure statistics enable the magnitude and duration of

neurological dysfunction caused by exacerbations to be incorporated into an overall disability analysis. Moreover, the inclusion of all serial data in the AUC calculations should reduce the variance which is associated with data obtained at single time points (for example, in a comparison between the initial and final disability scores). This "variance stabilising" effect means that fewer patients should be necessary for the same power to detect a predetermined clinically significant difference. Increasingly, there is a need to consider the cost effectiveness of pharmacotherapies. From the disability-time plots of different treatments being compared, the incremental therapy benefit of the test drug relative to the control can be expressed in terms of the readily interpretable disability-year (difference in AUCs) for use in cost effectiveness studies. For all these reasons we suggest that it is appropriate to employ summary measure statistics to evaluate the effects of new treatments in patients with relapsing-remitting multiple sclerosis.

Appendix

The AUC is a summation of the areas under the graph between each pair of consecutive scores by the trapezium rule. Disability measures (y_0, y_1, y_2, ...) are plotted versus their times of assessment (t_0, t_1, t_2, ...).

The AUC using raw data is calculated as follows:

If we have n+1 measurements y_i at times t_i (i=0, ..., n), then, $AUC = 1/2 \sum_{i=o}^{n-1} (t_{i+1}-t_i)(y_i+y_{i+1})$.

1 IFNB Multiple Sclerosis Study Group. Interferon β-1b is effective in relapsing-remitting multiple sclerosis. 1. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* 1993;**43**:655–61.
2 Jacobs LD, Cookfair DL, Rudick RA, *et al.* Intramuscular interferon β-1a for disease progression in relapsing multiple sclerosis. *Ann Neurol* 1996;**39**:285–94.
3 Johnson KP, Brooks BR, Cohen JA, *et al.* Copolymer-1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: results of a phase III multicenter, double-blind, placebo-controlled trial. *Neurology* 1995;**45**:1268–76.
4 Fazekas F, Deisenhammer F, Strasser4-Fuchs S, *et al*, for the Austrian Immunoglobulin in Multiple Sclerosis Study Group. Randomised placebo-controlled trial of monthly intravenous immunoglobulin therapy in relapsing-remitting multiple sclerosis. *Lancet* 1997;**349**:589–93.
5 Runmarker B, Andersen O. Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. *Brain* 1993;**116**:117–34.
6 Rudick R, Antel J, Confavreux C, *et al.* Recommendations from the National Multiple Sclerosis Society clinical outcome assessment task force. *Ann Neurol* 1997;**42**:379–82.
7 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983;**33**:1444–52.
8 McAlpine D, Compston N. Some aspects of the natural history of disseminated sclerosis. *Q J Med* 1952;**82**:135–67.
9 Kurtzke JF, Beebe GW, Nagler B, *et al.* Studies on the natural history of multiple sclerosis. 7. Correlates of clinical change in an early bout. *Acta Neurol Scand* 1973;**49**:379–95.
10 IFNB Multiple Sclerosis Study Group and the University of British Columbia MS/MRI Analysis Group. Interferon β-1b in the treatment of multiple sclerosis. Final outcome of the randomized controlled trial. *Neurology* 1995;**45**:1277–85.
11 Whitaker JN, McFarland HF, Rudge P, *et al.* Outcome assessment in multiple sclerosis clinical trials: a critical analysis. *Multiple Sclerosis* 1995;**1**:37–47.
12 Pozzilli C, Bastianello S, Koudriavtseva T, *et al.* Magnetic resonance imaging changes with recombinant human interferon-β-1a: a short term study in relapsing-remitting multiple sclerosis. *J Neurol Neurosurg Psychiatry* 1996;**61**:251–8.
13 Matthews JNS, Altman DG, Campbell MJ, *et al*. Analysis of serial measurements in medical research. *BMJ* 1990;**300**:230–5.
14 Oldham PD. A note on the analysis of repeated measurements of the same subjects. *J Chron Dis* 1962;**15**:969–77.
15 De Klerk NH. Repeated warnings re repeated measures. *Aust N Z J Med* 1986;**16**:637–8.
16 Zolman JF. *Biostatistics. Experimental design and statistical inference*. New York: Oxford University Press, 1993.
17 Hughes RAC, Sharrack B. More immunotherapy for multiple sclerosis. *J Neurol Neurosurg Psychiatry* 1996;**61**:239–41.
18 Freeman J, Langdon DW, Hobart JC, *et al*. The long term effects of neurorehabilitation in multiple sclerosis: a longitudinal study. *J Neurol* 1997;**244**(suppl 3):S10.
19 Sharrack B, Hughes RAC, Soudain S. Guy's neurological disability scale. *J Neurol* 1996;**243**(suppl 2):S32.
20 Sipe JC, Knobler RL, Braheny SL, *et al*. A neurologic rating scale (NRS) for use in multiple sclerosis. *Neurology* 1984;**34**:1368–72.
21 Hauser SL, Dawson DM, Lehrich JR, *et al.* Intensive immunosuppression in progressive multiple sclerosis. A randomized, three-arm study of high-dose intravenous cyclophosphamide, plasma exchange, and ACTH. *N Engl J Med* 1993;**308**:173–80.
22 Francis DA, Bain P, Swan AV, *et al*. An assessment of disability rating scales used in multiple sclerosis. *Arch Neurol* 1991;**48**:299–301.
23 Goodkin DE, Cookfair D, Wende K, *et al.* Inter- and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke extended disability status scale (EDSS). *Neurology* 1992;**42**:859–63.
24 Wishart J. Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika* 1938;**30**:16–28.
25 Bendsten L, Jensen R, Olesen J. A non-selective (amitriptyline), but not a selective (citalopram), serotinin reuptake inhibitor is effective in the prophylactic treatment of chronic tension-type headache. *J Neurol Neurosurg Psychiatry* 1996;**61**:285–90.