PATIENT'S CHOICE

RESEARCH PAPER

Relationship between early clinical characteristics and long term disability outcomes: 16 year cohort study (follow-up) of the pivotal interferon β-1b trial in multiple sclerosis

Douglas S Goodin,[1] Anthony Traboulsee,[2] Volker Knappertz,[3,4] Anthony T Reder,[5] David Li,[2] Dawn Langdon,[6] Christian Wolf,[7] Karola Beckmann,[3] Andreas Konieczny,[8] George C Ebers,[9] for the 16-Year Long Term Follow-up Study Investigators

[1]University of California, San Francisco, California, USA
[2]University of British Columbia, Vancouver, British Columbia, Canada
[3]Bayer HealthCare, Montville, New Jersey, USA, Berlin, Germany
[4]Heinrich-Heine-Universität, Düsseldorf, Germany
[5]Department of Neurology, University of Chicago, Chicago, Illinois, USA
[6]Department of Psychology, Royal Holloway, Egham, Surrey, UK
[7]Lycalis, Bruxelles, Brussels, Belgium
[8]Lampe Konieczny and Company, Berlin, Germany
[9]University Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK

Correspondence to
Professor D S Goodin, University of California, San Francisco, 505 Parnassus Ave, Rm 794M, San Francisco, CA 94143-0114, USA; douglas.goodin@ucsf.edu

Received 16 August 2011
Revised 12 November 2011
Accepted 14 November 2011
Published Online First 21 December 2011

ABSTRACT

Background Evaluating the long term benefit of therapy in multiple sclerosis (MS) is challenging. Although randomised controlled trials (RCTs) demonstrate therapeutic benefits on short term outcomes, the relationship between these outcomes and late disability is not established.

Methods In a patient cohort from the pivotal interferon β-1b trial, the value of clinical and MRI measures were analysed, both at baseline and during the RCT, for predicting long term physical and cognitive outcome.

Results Baseline disability correlated with both physical (R^2=0.22; p<0.0001) and cognitive (R^2=0.12; p<0.0001) outcome after 16 years. Accrual of disability during the RCT (R^2=0.12; p<0.0001) and annualised relapse rates during the trial correlated with physical outcome (R^2=0.12; p<0.0001) but not with cognition. In contrast, baseline MRI measures of atrophy and lesion burden correlated with cognitive (R^2=0.21; p<0.0001), but not with physical, outcome. Accumulation of plaque burden measured by MRI did not correlate with late physical disability or with cognitive outcome. Multivariate regression analysis using stepwise elimination demonstrated that baseline variables contributed independently to predicting long term outcomes while trial outcome variables contributed little. Overall, and considerably dependent on baseline measures, the models developed by this method accounted for approximately half of the variance in long term cognitive and disability outcome.

Conclusions Although on-trial change in some short term clinical measures correlated with long term physical and disability outcomes, the proportion of the variance explained by single commonly employed on-study variables was often small or undetectable. Better correlations were observed for several baseline measures, suggesting that long term outcome in MS may be largely determined early in the disease course.

Trial registration number http://Clinical Trials.gov, study registration NCT00206635.

INTRODUCTION

The efficacy of disease modifying therapies in multiple sclerosis (MS) has generally been evaluated by monitoring selected clinical and paraclinical outcomes in relatively short (1–3 years) randomised controlled clinical trials (RCTs). However, despite the widespread adoption of both MRI and clinical markers for use in clinical trials, the relationship between these short term outcomes and longer term outcomes is unclear. Demonstration of the value of short term measures for predicting long term outcome in MS, in particular disability, would help in projecting longer term impacts of therapy on social, economic and medical costs of the untreated versus treated disease.

Formal validation of surrogate outcomes serving as putative predictors in clinical trials involves more than simply establishing correlations between candidate surrogates and trial outcomes.[1–5] Nevertheless, demonstration of a correlation between short term outcomes in an RCT and long term outcomes is a crucial starting point in the development of a surrogate for hard long term outcomes, which have unassailable clinical significance.

The first RCT of interferon β-1b (IFNβ-1b) in MS was begun more than two decades ago.[6–8] This trial randomised 372 patients to three different treatment arms (IFNβ-1b 250 μg, IFNβ-1b 50 μg and placebo). Unequivocal treatment benefit for the higher dose arm was seen at 2 and 3 years for several short term clinical outcomes, including relapse rate, relapse free interval, time to first relapse and categorical change (ie, change of ≥1 point at the end of the study) on the Kurtzke Expanded Disability Status Scale (EDSS).[9] Patients in the higher of the two dose arms also demonstrated benefits on MRI measures of T2 disease burden and new active T2 lesions.[6–8] Nevertheless, because few patients reached hard disability outcomes such as EDSS ≥6 by trial completion, the study could not suitably address questions about the effects on unremitting long term disability in MS. Patients participating in the original RCT have been followed since its conclusion, and after more than 16 years from RCT onset and almost 6000 patient-years of follow-up, we can now address important questions about the relationships between the short term clinical and MRI measures used in the RCT and long term disability outcome.

Here we assess the predictive validity[10] of several clinical and MRI measures from the pivotal IFNβ-1b RCT for change in physical and cognitive

outcomes at the 16 year follow-up. The effect of therapy on long term outcomes will be reviewed elsewhere.[11]

METHODS

Patients

The design and methods of the original RCT and the 16 year follow-up study have been described in detail previously.[6–8 12 13] These other papers were descriptive in nature and did not explore the predictive validity of the different clinical trial outcomes. Briefly, patients participating in the original IFNβ-1b pivotal trial were re-contacted in 2005 (approximately 12 years after completion of the pivotal trial) and asked to participate in the follow-up study.[12] Of the 373 patients in the original phase III study, 328 (88.2%) were identified. Among these, 293 were still alive and 260 (70%) consented to detailed assessment of their interim disease course (medical record review and personal interview) and current physical, imaging and cognitive assessments. Physical disability was measured both by the EDSS and the MS Severity Score (MSSS) at the start of the RCT and by EDSS during the RCT and also during the long term follow-up (LTF) period. Cognitive outcome was assessed at LTF by a battery of five neuropsychological tests, consisting of the Paced Auditory Serial Addition Task (PASAT), the Symbol Digit Modality Task (SDMT), the California Verbal Learning Test II (CVLT-II), the Controlled Oral Word Association Task (COWAT) and the Delis–Kaplan Executive Function System (D-KEFS) test. In addition, the Wechsler Test of Adult Reading was used to estimate the premorbid IQ.[14]

Patients who agreed to participate were assessed over the course of 1–3 clinic visits. When unable to participate in person, patients were offered a home visit by study investigators for their assessment. A comparison of baseline RCT data between those patients who did and those who did not participate in the LTF study showed that the two groups were very similar for all baseline measures for on-trial behaviour (table 1). As a result, our sample is likely representative of the entire RCT population. Ethics approval for the follow-up study was obtained from the institutional review boards or independent ethics committees of the participating centres. All subjects gave written informed consent.

Of relevance to any long term follow-up study of this type, it is important to recognise that ability to contact patients (at least in the USA) is severely restricted due to the Health Insurance Portability and Accountability Act (HIPAA) regulations. Thus the HIPAA regulations do not permit any patient contact after a failure (by the patient) to respond to two written letters requesting their participation. We cannot simply call them, even when we know their address and phone number.

Study procedures

Several clinical and MRI variables were determined, both at the start and over the course of the pivotal trial.[6–8] On-RCT variables were defined as those assessed during the first 2 or 3 years of the study. Because the 2 and 3 year analyses led to essentially identical conclusions, only data from the 2 year analysis are presented because this data set contained a more complete ascertainment of patients enrolled into the RCT. Thus after completion of the original 2 year protocol, the RCT was extended for an additional year at the request of the US Food and Drug Administration but several patients (fulfilling their commitment to take part in a 2 year study) withdrew their consent (see figure 1).

Candidate predictor variables evaluated for their relationship to long term outcome categorically included relapse related, disability, MRI and other variables (table 2). Relapse related variables included pre-trial and on-trial relapse rates. Pre-RCT attacks were determined historically by patient interview and from medical records whereas on-RCT relapses were determined by study investigators every 3 months during the pivotal trial. Disability variables included baseline EDSS, baseline MSSS, a 1 point change in EDSS sustained for 3 months, a categorical change (≥1 point) on the EDSS scale measured from baseline to trial end and the measured EDSS change over the course of the RCT. MRI variables consisted of baseline T2 burden of disease (BOD), defined as the volume (measured in cm^2 per slice) of

Table 1 Baseline and on-randomised control trial clinical characteristics of the patients included (and those not included) in the detailed long term follow-up evaluation after 16 years*

	Patients included in LTF population	Patients not included in LTF population	p Value†
No of patients	260‡	112‡	–
% Women	69%	71%	0.7125
Age at disease onset (years)	27.3 (6.8)	27.7 (7.3)	0.5361
Age at start of RCT (years)	35.4 (7.4)	35.8 (6.7)	0.5220
Baseline EDSS	2.9 (1.3)	2.9 (1.3)	0.8373
Baseline EDSS ≥3 (% of population)	138 (53)	62 (55)	0.7343
Disease duration (years)	8.0 (6.2)	8.1 (6.3)	0.9950
Baseline MSSS	4.3 (2.3)	4.4 (2.2)	0.7118
Baseline MRI T2 BOD (cm^2)	19.6 (20.1)	23.1 (23.8)	0.0699
Baseline third ventricular width (mm)	4.86 (2.28)	5.19 (2.42)	0.1893
Annualised relapse rate (2 years prior to RCT)	1.68 (0.77)	1.67 (0.85)	0.5964
Annualised relapse rate (on-RCT)	1.2 (1.2)	1.6 (2)	0.0849
EDSS change (on-RCT)	0.05 (1.3)	0.3 (1.6)	0.2415
No of T2 CAL (on-RCT)	2.4 (3.3)	3.0 (4.0)	0.1613
MRI T2 BOD change (on-RCT) (cm^2)	1.3 (6.1)	2.2 (10.2)	0.0729
Change, third ventricular width (on-RCT) (mm)	0.62 (0.97)	0.63 (1.14)	0.7321
On IFNβ-1b (250 μg) during RCT (%)	37	25	0.0178

*Values are means (SDs) or number.
†p Value derived from Fisher's exact test for rates, z score for percentages and Wilcoxon's rank sum test for all others.
‡Seven deceased patients included in the LTF population; 28 deceased patients not included in LTF population.
BOD, burden of disease; CAL, combined active lesions; EDSS, Expanded Disability Status Scale score; IFNβ, interferon β; LTF, long term follow-up; MSSS, Multiple Sclerosis Severity Score; RCT, randomised controlled trial.

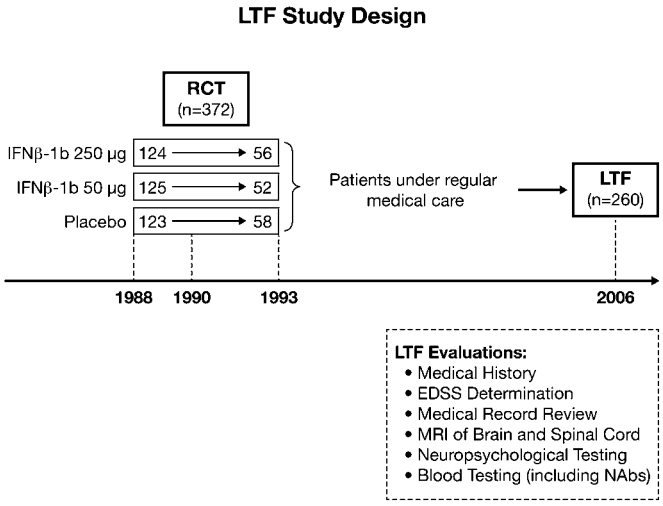

Figure 1 Flow diagram for the long term follow-up (LTF) study. Patient recruitment for the original randomised control trial (RCT) was begun in 1988 and was complete by 1990. By 1993, the last person in the trial had completed 3 years on-study and the trial was ended; at which time 166 patients (45%) had completed 5 years. These 5 year completers were split, as indicated, 56, 52 and 58, between the treatment arms. Following the RCT, patients were not part of a trial but, beginning in 2005, patients were re-contacted and asked to take part in the LTF study evaluations. EDSS, Expanded Disability Status Scale score; IFNβ, interferon β; NAbs, neutralising antibodies.

hyperintense lesions seen on T2 weighted images, on-RCT change in the T2 BOD, third ventricular width (measured in mm) and numbers of new and newly enlarging lesions seen on annual T2 weighted images during the RCT. Other variables consisted of duration of disease, age, gender, treatment history, randomisation group in the pivotal trial and the development of neutralising antibodies (NAbs) measured in neutralising units/ml (NU/ml). NAbs were considered present when found at titres ≥20 NU/ml on two or more consecutive occasions during the RCT.

Because of the possibility of a relationship between baseline variables and ultimate outcome, we included all of these variables as candidates for the model. In our analysis, in addition to these baseline variables, we also included the variables of treatment during the pivotal study, total exposure to IFNβ-1b and also changes (during the RCT) in third ventricular width, BOD, T2 activity, EDSS and relapse rate. No variable was forced into the model. In this way, we have addressed all baseline as well as on-trial predictors of importance over the first 2 years under randomised treatment allocation.

Long term outcome

For analysis of LTF physical outcomes, a dichotomous, composite, 'negative' measure was used. A 'negative' physical outcome was defined when a patient either converted to secondary progressive (SP) MS or reached an EDSS ≥6.0. These outcomes were chosen because they were clinically important and considered unlikely to remit once sustained. SPMS was defined prospectively as a progressive increase in disability (following a relapsing–remitting course), evolving over ≥12 months and not relapse associated. In addition, SPMS patients had to experience an increase of ≥1 point on the EDSS scale over the previous 2 years (or a 0.5 point increase from an EDSS score of 6.0 or 6.5) with or without superimposed exacerbations. To

Table 2 Univariate regressions of the relationship between 2 year outcome measured in the original pivotal interferon β-1b study and the 16 year outcome for physical and cognitive function*

	Physical outcome (logistic regression)			Cognitive outcome (linear regression)		
	OR	R^2	p Value†	Slope	R^2	p Value†
Baseline variables						
Baseline EDSS	2.07	0.22	<0.0001	−1.12	0.12	<0.0001
MSSS at trial onset	1.23	0.07	0.0004	−0.25	0.02	0.09
Baseline MRI T2 BOD (cm^2)	1.03	0.07	0.001	−0.11	0.21	<0.0001
Duration of MS (y)	1.07	0.05	0.003	−0.15	0.05	0.004
Third ventricular width (mm)	1.18	0.04	0.011	−0.94	0.21	<0.0001
Age at trial start	1.03	0.01	NS	−0.19	0.00	NS
Age at MS onset	0.98	0.01	NS	0.12	0.04	0.02
Annual relapse rate prior to trial (2 years)	1.11	0.00	NS	0.07	0.00	NS
Premorbid IQ	0.99	0.00	NS	0.18	0.14	<0.0001
Gender	0.91	0.00	NS	0.77	0.00	NS
On-RCT variables						
Annual relapse rate	1.82	0.12	<0.0001	−0.47	0.02	NS
Actual EDSS change—baseline to 2 years	1.59	0.11	<0.0001	−0.32	0.01	NS
Categorical EDSS change (≥1 point)	2.71	0.06	0.002	−1.68	0.02	0.05
Confirmed 1 point EDSS progression	1.84	0.02	0.05	0.81	0.00	NS
Change, third ventricular width (mm)	1.06	0.00	NS	−1.27	0.07	0.003
50 μg treatment group during RCT‡	0.74	0.01	NS	1.42	0.02	0.09
250 μg treatment group during RCT‡	1.07	0.01	NS	1.41	0.02	0.09
Total IFNβ-1b exposure (years) (on LTF)	0.98	0.00	NS	−0.08	0.01	NS
No of new T2 lesions	0.96	0.01	NS	−0.10	0.01	NS
NAbs (≥20 NU/ml)	0.83	0.00	NS	−0.49	0.00	NS
Change, MRI T2 BOD (cm^2)	1.01	0.00	NS	−0.01	0.01	NS

*Physical outcome=either SPMS or EDSS=6; cognitive outcome=Cognitive Performance Index (see text).
†NS=p>0.1.
‡Comparison versus placebo.
BOD, burden of disease; EDSS, Expanded Disability Status Scale score; IFNβ, interferon β; MS, multiple sclerosis; MSSS, Multiple Sclerosis Severity Score; NAbs, neutralising antibodies; RCT, randomised controlled trial; SPMS, secondary progressive multiple sclerosis.

reach EDSS ≥6.0, this level of disability had to be confirmed by two consecutive evaluations (at least 3 months apart) and sustained for the remaining follow-up period. Secondary analyses explored EDSS and SPMS as individual (not composite) physical outcomes at LTF evaluation.

For analysis of LTF cognitive outcomes, a continuous measure was used. This was the so-called 'Cognitive Performance Index' which represented the sum of the patient's z scores on the PASAT, SDMT, CVLT-II, COWAT and D-KEFS tests.

Statistical analysis

Relationships between candidate predictors and hard physical outcomes at LTF (eg, EDSS ≥6 or SPMS, as well as the composite 'negative' physical outcome of both outcomes) were explored using logistic regression modelling. We also looked at models including death as a 'negative' outcome. However, because this did not change any of the results and because we only had permission to review the records for seven of the 35 deceased patients,[12] we felt it was preferable to exclude this outcome from our composite measure. Relationships with the continuous 'Cognitive Performance Index' were explored with linear regression models. Two methods of analysis were undertaken. In the first, we explored 'univariate' relationships, in which regression analyses were run with each candidate predictor considered individually. In the second, we developed multivariate regression models using stepwise elimination procedures for model selection, in which all candidate variables were allowed to enter if their coefficient had a significance of $p<0.5$. A candidate predictor was eliminated from the model as soon as it failed to contribute to the overall R^2 for the model at a significance level of $p<0.10$, with the p values derived from t tests (linear regression) or Wald χ^2 tests (logistic regression).

RESULTS

Clinical and MRI disability outcomes during the RCT and at LTF

The mean EDSS score for the entire group, at baseline, was 2.89. By the end of 2 years the mean EDSS had increased by 0.05 points (SD 1.33) and by the LTF it had increased by 2.28 points (SD 2.04). Thus the mean EDSS at the LTF was 5.17 (SD 2.43). The EDSS was available in all 260 patients. At the 2 year point, 21.2% of patients (55/260) had sustained a 1 point confirmed EDSS change from their baseline. At the LTF, 43.5% (113/260) of patients had reached an EDSS of 6.0, 40.0% (104/260) had reached SPMS and 53.8% (140/260) had reached either negative outcome. Cognitive assessment at the LTF was completed in 58.5% (152/260) of the patients and the Cognitive Performance Index had a mean summed z score of −4.52 (SD 4.22). At baseline, the third ventricular width was 4.86 mm (SD 2.28), and by year 2 this had increased by 0.644 mm (SD 0.972). BOD at baseline was 1.96 cm^2 (SD 2.02) and by 2 years this had increased by 0.13 cm^2 (SD 0.61).

Univariate and multivariate analyses

The exploratory univariate analyses for the relationship of candidate predictors with respect to physical and cognitive function 16 years later are shown in table 2. In these univariate explorations (table 2), several baseline and on-RCT variables (but not others) were significantly correlated with long term disability outcome (either physical or cognitive). Baseline disability correlated significantly with both physical ($R^2=0.22$; $p<0.0001$) and cognitive ($R^2=0.12$; $p<0.0001$) outcome after 16 years. Accrual of disability during the RCT ($R^2=0$ 0.11; $p<0.0001$) and annualised relapse rates during the trial ($R^2=0.12$; $p<0.0001$) correlated significantly with physical outcome but not with cognition. In contrast, baseline measures of third ventricular width ($R^2=0.21$; $p<0.0001$), MRI lesion burden ($R^2=0.21$; $p<0.0001$) and premorbid IQ ($R^2=0.14$; $p<0.0001$) were correlated with cognitive, but not with physical, outcome. Notably, with the exception of the measure of third ventricular width, a change in MRI over the course of the trial did not correlate with late disability—either cognitive or physical. The actual change in EDSS over the course of the trial was a superior predictor of physical outcome compared with more commonly used measures such as sustained or categorical 1 point EDSS change. Moreover, neither the sustained nor the categorical 1 point EDSS change remained in the multivariate model. These disability measures, however, were all poor predictors of cognitive outcome in the univariate analysis (table 2). Finally, the occurrence of NAbs during the RCT had no relationship to outcome (table 2).

In the principal multivariate analysis, the contribution of each potential predictor variable was tested using a stepwise elimination procedure to estimate a final model for predicting both physical and cognitive outcome (table 3). The most significant predictor of both physical and cognitive outcome after 16 years was baseline EDSS (table 3). Similarly, in the final regression model, the change in EDSS score over the first 2 years of the RCT was an independent predictor of cognitive and (especially) physical outcome. In contrast, MRI measures such as T2 BOD (at baseline) and third ventricular width (both at baseline and change during the RCT) contributed largely (or only) to cognitive outcome. Annualised relapse rates during the RCT contributed only to predicting physical outcome (table 3).

In both multivariate models, explained variance was approximately half of the total variance in long term outcome (table 3) but more so from baseline measures than from on-study surrogates. The amount of the variance explained by any single variable was generally quite small (table 2).

Table 3 Multiple regression model for outcome at long term follow-up derived with stepwise model selection procedure: fitted regression model including predictors with p≤0.5 to enter; p<0.1 to stay in the model

	Estimate	SE	p Value
Physical outcome* model fit (logistic regression): $R^2=0.51$			
Baseline variables			
Intercept	−5.3	0.91	<0.0001
EDSS at baseline	1.20	0.22	<0.0001
MRI T2 BOD at baseline (cm^2)	0.05	0.02	0.001
Gender	0.93	0.47	0.045
On-RCT variables			
Actual EDSS change from baseline	0.86	0.21	<0.0001
Annualised relapse rate	0.52	0.23	0.025
Cognitive outcome† model fit (linear regression): $R^2=0.49$			
Baseline variables			
Intercept	−11.2	3.98	0.006
EDSS at baseline	−0.99	0.25	<0.0001
Premorbid IQ	0.12	0.035	0.0007
MRI T2 BOD at baseline (cm^2)	−0.05	0.02	0.018
Third ventricular width at baseline (mm)	−0.41	0.16	0.014
On-RCT variables			
Actual EDSS change from baseline	−0.67	0.24	0.007
Change, third ventricular width (mm)	−0.87	0.33	0.009

*Physical outcome=either SPMS or EDSS=6.
†Cognitive outcome=Cognitive Performance Index (see text).
BOD, burden of disease; EDSS, expanded disability status scale score; LTF, long term follow-up; RCT, randomised controlled trial; SPMS, secondary progressive multiple sclerosis.

DISCUSSION

In many chronic disabling diseases it is difficult to establish long term efficacy for any specific therapy and MS is no exception. The protracted observation times necessary for patients to reach hard disability outcomes contrast with the relatively short term formal RCT trial periods that have been successfully executed. RCT designs have not allowed for sufficient time to reach these outcomes. Furthermore, once a drug has been shown to improve patient outcome on measures thought by even a substantial minority to be clinically relevant to the disease process, impediments to continuation arise. Patients may not agree to prolonged exposure to placebo and many clinicians will likely discourage it.[15 16] For these reasons, establishing long term efficacy at present rests on the analysis of non-randomised longitudinal data with best available adjustments for the many biases that impact such studies.[11]

In addition to such alternative analysis strategies, however, it is also imperative to establish that these short term outcome measures correlate with (and predict) long term outcome as an essential first step towards establishing surrogacy. The present study is the first to evaluate the predictive value of short term outcome measures used in MS clinical trials for hard disability endpoints. Much of the predictive power of the final regression models (table 3) came from single *baseline* measures rather than from on-study changes. Observations of simple correlations between short term measures and long term outcomes fall well short of proving that these measures are true surrogates for long term efficacy.[1] Although the short term measures we explored (clinical attacks, disability and MRI lesions) are generally believed to be reflections of the pathological processes (ie, episodic inflammation, demyelination and axonal injury) which underlie permanent disability in MS,[17] none (individually) was strongly associated with disability or cognitive outcome, and some widely used and previously influential trial outcomes were completely disassociated.

A therapy might either alter disease course without affecting all of these processes or alter them without affecting outcome. For example, a neuroprotective agent could limit axonal injury or promote oligodendrocyte survival without affecting inflammation. Similarly, an immune suppressant might impact destructive inflammation to a lesser extent than reparative inflammation and thus lead to less (but more disabling) inflammation. Finally, it is also possible that because of either functional redundancy or plasticity within the CNS, the correlation between a particular short term measure and long term outcome may be weak or non-existent.[18] Nevertheless, even though true surrogate markers have not been established, any therapy that successfully interrupts one or another of these basic pathogenic mechanisms, which are correlated with long term physical and cognitive outcome in MS, has potential for limiting long term disability. The findings should stimulate efforts to find better surrogate markers.

This study was necessarily restricted to those who agreed to current physical and cognitive assessments and review of their interim disease course (n=260; 70% of the 328 patients identified from the original cohort). The baseline characteristics of those who participated and those who did not are detailed elsewhere.[11] In brief, no significant differences were observed between these groups for any clinical and outcome related features. Notably, the percentage of women (69% vs 71%), age of onset (27.3 vs 27.7 years), duration of disease (8.0 vs 8.1 years), entry EDSS (2.9 vs 2.9) and mean on-trial change in EDSS (0.0 vs 0.3) were nearly identical.

Outcomes focused on physical signs arising from CNS inflammation (ie, clinically evident relapses and less clearly disability progression on the EDSS scale) seem to be better predictors of physical rather than cognitive outcome. In contrast, outcomes thought to better measure clinically silent pathology within the CNS (ie, T2 lesions, BOD and atrophy) were better predictors of cognitive than physical outcome but were weak nevertheless. Such disassociation has been suggested previously, based on the belief that spinal cord pathology has a greater impact on physical function whereas intracerebral pathology is more likely to impact cognitive function.[19] This suggests that more attention should be paid in future studies to this differential impact.

Although on-RCT behaviour for some outcomes correlated with long term outcome (tables 1 and 2), the strongest associations were actually with simple and single baseline functions as measured by the EDSS (for physical outcome) or the BOD and third ventricular width (for cognitive outcome). This result is consistent with several other reports in the literature.[20–23] Thus these baseline measures effectively provide a type of integrated assessment of disease activity that had occurred up to the point of evaluation. In contrast, on-RCT measures provide an estimate of disease activity over a shorter time frame.

Baseline EDSS was significantly related to development of both physical disability and cognitive decline and based on the R^2 for univariate explorations, and better predicted these outcomes than did MSSS (table 2). Similarly, in the multiple logistic regression model for outcome derived with stepwise selection, the measure of MS severity consistently selected for inclusion in the model was baseline EDSS, not MSSS (table 3). This observation is also consistent with our recursive partitioning analysis.[10] Indeed, EDSS may be a much better measure of MS severity than its detractors might otherwise suggest.[24 25]

Notably, the actual change in EDSS over the 2 year course of the trial data analysis was a better predictor of long term outcome than either the sustained 1 point EDSS change or the categorical 1 point change at trial end. This observation, combined with the fact that the 1 point change definition of treatment failure was found no more likely to occur than improvement to the same degree in placebo arms, strongly suggests that the current practice of using sustained EDSS change with only 3–6 months as the time of confirmation of worsening as a primary disability outcome should be revised.[26]

Neither MRI T2 burden change nor accumulation of new MRI lesions during relapses was predictive of disability or cognitive change. On the clinical side, on-study attack rates were not predicted by rates prior to entry, confirming the findings of Young and colleagues.[27] This raises considerable doubt about the validity of using pre-trial relapse frequency as a criterion for trial entry. In addition, the observation that the pre-study attack rate did not contribute to predicting physical outcome (tables 1 and 2) is consistent with the reported lack of relationship between number of attacks during the relapsing–remitting phase and time to cane, bedridden status or death from MS.[28]

Many adjustments are necessary to control for biases that can contaminate non-randomised observational studies.[11] No relationship between therapy during the RCT and either physical or cognitive outcome emerged in our regression analysis (table 2). Nevertheless, these observations are not definitive because the difference between the three arms (in intent to treat terms) consists of the few years during which therapeutic exposure differed between groups. Post-trial, all participants were offered and encouraged to take the active treatment.

Unlike our longitudinal data for physical disability, our cognitive data are cross sectional. Consequently, these data could not be analysed using the same bias mitigating statistical methods that we applied to the physical disability data.[11] Therefore, any data regarding the effect of treatment on cognition will remain contaminated by the natural tendency for patients who are doing well to stay on therapy and for those who are doing poorly to switch or to stop therapy—a source of bias that can adversely affect any non-randomised study.[11]

In summary, this is the first study to assess the predictive validity of a variety of short term outcome measures for very long term physical and cognitive outcomes of patients with MS. These included the key outcomes used ubiquitously in trials for determination of efficacy. We found that baseline measurement of disability and MRI, and the on-RCT measures of clinical attacks, disability change from entry to exit and atrophy, modestly but independently correlated with physical or cognitive long term outcomes after 16 years. Although nearly half of the long term physical and cognitive outcome was predictable, much of this came from single and simple baseline measures (table 3). In general, the amount of the variance explained by any single variable was quite small. Importantly, the previously influential, expensive and widely used on-trial change in MRI plaque burden (as measured by T2 lesion volume) did not correlate independently with either physical or cognitive outcome.

Acknowledgements The authors acknowledge the assistance of Ray Ashton and Maria Bell from PAREXEL MMS (Worthing, UK) with manuscript preparation.

Funding The study was sponsored by Bayer HealthCare Pharmaceuticals. PAREXEL MMS received payment from Bayer HealthCare Pharmaceuticals for editorial support.

Competing interests DSG, GE, AR, DLi, DL, AK, CW, VK and KB have support from Bayer HealthCare for the submitted work, AT has a specified relationship with Bayer HealthCare and might have an interest in submitted work from the previous 3 years. Both the sponsors and the independent investigators were intimately involved in the study design. The researchers DSG, GE, AR, DLi, AT and DL were independent of the funders. The researchers AK, CW, VK and KB either work or previously worked for the funders.

Ethics approval The study obtained ethics approval from the institutional review boards or independent ethics committees of the participating centres before long term follow-up planning, which began in 2004.

Contributors All authors had full access to all of the data in the study and all authors take full responsibility for the integrity of the data and accuracy of the data analysis. DSG contributed to the design/conceptualisation of the study, analysis/interpretation of the data and drafting/revising the manuscript for intellectual content. AT contributed to the analysis/interpretation of the data and drafting/revising the manuscript for intellectual content. VK contributed to the design/conceptualisation of the study, analysis/interpretation of the data and drafting/revising the manuscript for intellectual content. AR contributed to the analysis/interpretation of the data and drafting/revising the manuscript for intellectual content. DLi contributed to the analysis/interpretation of the data and drafting/revising the manuscript for intellectual content. DL contributed to drafting/revising the manuscript for intellectual content. CW contributed to the design/conceptualisation of the study, analysis/interpretation of the data and drafting/revising the manuscript for intellectual content. KB contributed to the design/conceptualisation of the study, analysis/interpretation of the data and drafting/revising the manuscript for intellectual content. AK contributed to the design/conceptualisation of the study and analysis/interpretation of the data. GE contributed to the design/conceptualisation of the study, analysis/interpretation of the data and drafting/revising the manuscript for intellectual content.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Statistical and data tables are available from the corresponding author (douglas.goodin@ucsf.edu). Participants gave written informed consent for data acquisition and data sharing.

REFERENCES

1. **Baker SG,** Kramer BS. A perfect correlate does not a surrogate make. *BMC Med Res Methodol* 2003;**3**:16.
2. **Fleming TR,** DeMets DL. Surrogate endpoints in clinical trials: Are we being misled? *Ann Intern Med* 1996;**125**:605—13.
3. **Bucher HC,** Guyatt GH, Cook DJ, *et al*. Users guide to the medical literature XIX: applying clinical trial results: a. how to use an article measuring the effect of an intervention on surrogate endpoints. *JAMA* 1999;**282**:771—8.
4. **Fleming TR.** Surrogate endpoints in cardiovascular disease trials. *Am Heart J* 2000;**139**:S193—6.
5. **Sormani MP,** Bonzano L, Roccatagliata L, *et al*. Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: a meta-analytical approach. *Ann Neurol* 2009;**65**:268—75.
6. **The IFNB Multiple Sclerosis Study Group.** Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I. clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* 1993;**43**:655—61.
7. **Paty DW,** Li DK. Interferon beta-lb is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial. UBC MS/MRI Study Group and IFNB Multiple Sclerosis Study Group. *Neurology* 1993;**43**:662—7.
8. **The IFNB Multiple Sclerosis Study Group and the University of British Columbia MS/MRI Analysis Group.** Interferon beta-1b in the treatment of multiple sclerosis: final outcome of the randomized controlled trial. *Neurology* 1995;**45**:1277—85.
9. **Kurtzke JF.** Rating neurologic impairment in multiple sclerosis: an Expanded Disability Status Scale (EDSS). *Neurology* 1983;**33**:1444—52.
10. **Graves RM.** *Survey errors and survey costs*. New York: John Wiley and Sons, 1989.
11. **Goodin DS,** Jones J, Li D, *et al;* for the investigators of the 16-Year Long-Term Follow-Up Study. Establishing long-term efficacy in chronic disease: the use of recursive partitioning and propensity score adjustment to estimate long-term outcome in multiple sclerosis. *PLoS One* 2011;**6**:e22444.
12. **Ebers G,** Reder A, Traboulsee A, *et al*; for the investigators of the 16-Year Long-Term Follow-Up Study. Long-term follow-up of the original interferon-β1b trial in multiple sclerosis: Design and lessons from a 16-year observational study. *Clin Ther* 2009;**31**:1724—36.
13. **Ebers GC,** Traboulsee A, Li D, *et al*; for the investigators of the 16-Year Long-Term Follow-Up Study. Analysis of clinical outcomes according to original treatment groups 16 years after the pivotal IFNB-1b trial. *J Neurol Neurosurg Psychiatry* 2010;**81**:907—12.
14. **Wechsler D.** Wechsler Test of Adult Reading™(WTARTM) Scoring and Administration Manual. San Antonio, Texas: Psychological Corporation, 2001.
15. **Lublin FD,** Reingold SC. Placebo-controlled clinical trials in multiple sclerosis: ethical considerations. National Multiple Sclerosis Society (USA) Task Force on placebo-controlled clinical trials in MS. *Ann Neurol* 2001;**49**:677—81.
16. **Polman CH,** Reingold SC, Barkof F, *et al*. Ethics of placebo-controlled clinical trials in multiple sclerosis: a reassessment. *Neurology* 2008;**70**:1134—40.
17. **Trapp BD,** Ransohoff R, Rudick R. Axonal pathology in multiple sclerosis: relationship to neurologic disability. *Curr Opin Neurol* 1999;**12**:295—302.
18. **Goodin DS.** MRI as a surrogate outcome measure of disability in multiple sclerosis: Have we been overly harsh in our assessment? *Ann Neurol* 2006;**59**:597—605.
19. **Summers M,** Swanton J, Fernando K, *et al*. Cognitive impairment in multiple sclerosis can be predicted by imaging early in the disease. *J Neurol Neurosurg Psychiatry* 2008;**79**:955—8.
20. **Furby J,** Hayton T, Anderson V, *et al*. Magnetic resonance imaging measures of brain and spinal cord atrophy correlate with clinical impairment in secondary progressive multiple sclerosis. *Mult Scler* 2008;**14**:1068—75.
21. **Kappos L,** Traboulsee A, Constantinescu C, *et al*. Long-term subcutaneous interferon beta-1a therapy in patients with relapsing-remitting MS. *Neurology* 2006;**67**:944—53.
22. **Fisher E,** Rudick RA, Simon JH, *et al*. Eight-year follow-up study of brain atrophy in patients with MS. *Neurology* 2002;**59**:1412—20.
23. **Fisher E,** Rudick RA, Cutter G, *et al*. Relationship between brain atrophy and disability: an 8-year follow-up study of multiple sclerosis patients. *Mult Scler* 2000;**6**:373—7.
24. **Fischer JS,** Rudick RA, Cutter GR, *et al*. The multiple sclerosis functional composite measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. *Mult Scler* 1999;**5**:244—50.
25. **Roxburgh RH,** Seaman SR, Masterman T, *et al*. Multiple sclerosis severity score: using disability and disease duration to rate disease severity. *Neurology* 2005;**64**:1144—51.
26. **Ebers GC,** Heigenhauer L, Daumer M, *et al*. Disability as an outcome in MS clinical trials. *Neurology* 2008;**71**:624—31.
27. **Young PJ,** Lederer C, Eder K, *et al;* Sylvia Lawry Centre for Multiple Sclerosis Research. Relapses and subsequent worsening of disability in relapsing-remitting multiple sclerosis. *Neurology* 2006;**67**:804—8.
28. **Scalfari A,** Neuhaus A, Degenhardt A, *et al*. The natural history of multiple sclerosis, a geographically based study 10: relapses and long-term disability. *Brain* 2010;**133**:1914—29.