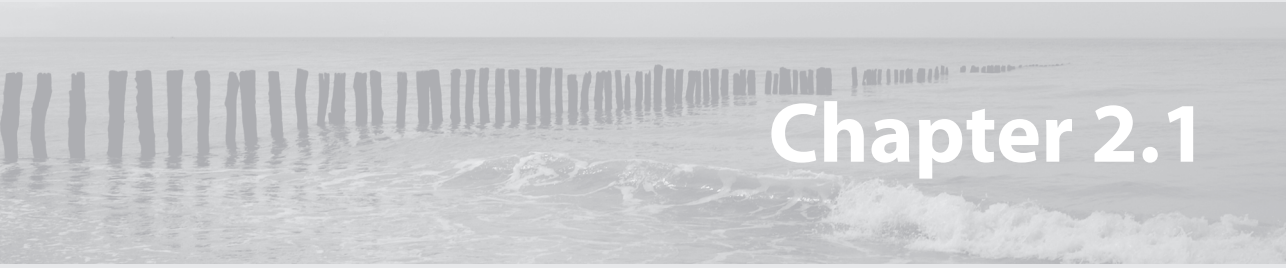

Chapter 2.1

The search for responsive clinical endpoints in primary progressive multiple sclerosis

LVAE Bosma, JJ Kragt, L Brieva, Z Khaleeli, X Montalban,
CH Polman, AJ Thompson, M Tintoré, BMJ Uitdehaag

Multiple Sclerosis 2009;15:715-720

ABSTRACT

Objective: To determine whether in primary progressive multiple sclerosis (PPMS) combining scores of Expanded Disability Status Scale (EDSS) with data from Timed 25-foot Walk (T25FW) and 9-hole Peg Test (9HPT) would produce a clinical endpoint that has a higher event rate than the EDSS alone.

Methods: In a group of 161 PPMS-patients EDSS, T25FW and 9HPT were performed at three time points over 2 years. We calculated how many patients showed clinically meaningful deterioration (or improvement) on individual and combined scales. We defined improvements on one scale with deterioration on the other as "opposing changes". We investigated the possible effect of baseline disability on the definition of our endpoint by dividing the population into two subsets of patients determined by baseline EDSS level.

Results: On individual scales, event rates were highest on the T25FW: 34% and 46% 1 year and 2 years after baseline. On a combination of two scales, at 1 year the event rate was highest on T25FW/9HPT (46%; with a high rate of opposing changes) and at 2 years on T25FW/EDSS (57%; with a lower rate of opposing changes). In both subsets, event rates were highest on the T25FW and (at 2 years) on the combination of T25FW/EDSS.

Conclusions: T25FW has the highest event rate as a single scale, independent of the baseline disability level. A term of 2 years turned out to be more meaningful to observe than 1 year. 'Worsening on either T25FW or EDSS' is the most appropriate composite endpoint in this patient group.

INTRODUCTION

Primary progressive multiple sclerosis (PPMS) patients experience progressive disease from onset and a clinical course without relapses. Evidence from magnetic resonance imaging (MRI) and histopathology suggests that disease progression is caused by progressive axonal degeneration and neuronal loss and is associated with less active inflammation than in relapsing multiple sclerosis (MS).[1-3]

There have been few treatment trials in PPMS. A small Phase II, 2-year, placebo-controlled treatment trial with interferon beta (50 PPMS patients) showed no treatment effect on Expanded Disability Status Scale (EDSS), Multiple Sclerosis Functional Composite (MSFC), or any MRI-marker.[4] Another Phase II, 2-year trial with interferon beta (73 patients) showed no effect on the primary outcome measure, EDSS, but significant improvements were reported for the MSFC score and some MRI measures compared with placebo.[5] In the only Phase III study, 943 PPMS patients were treated with glatiramer acetate or placebo.[6] No treatment effect could be shown on the primary outcome measure, the EDSS, partially due to a slow movement on the EDSS. The MSFC also failed to show a treatment effect, but data for individual MSFC components have not been presented to date.

PPMS represents a major unmet medical need because of progressive disability and absence of effective therapy. Trial design poses particular problems, mainly in the selection of a responsive, appropriate clinical outcome measure. To contribute to resolving this issue, we analyzed clinical data from a cohort of PPMS patients followed as part of a natural history study. Our objective was to analyze whether – combining data from T25FW and 9HPT with EDSS scores – an endpoint can be defined that has a higher event rate than the EDSS alone. T25FW and 9HPT were selected for this purpose because PPMS is especially a disease of extremity motor dysfunction.

METHODS

Patients and test procedures

Data of patients with MS[7] from three European MS centers were retrospectively selected for analysis. The main selection criteria were: PPMS,[8] baseline EDSS ranging from 2.0 to 6.5, and data from at least 3 assessments of EDSS and MSFC: a baseline assessment followed by a first follow-up visit (FU1) after 1 year (≥ 260 days) and a second follow-up visit (FU2) after 2 years. No selection criteria for age or gender were applied. All assessments were obtained as part of

routine outpatient care during regular visits. Both EDSS[9] and MSFC[10] were performed in the same visit under standardized conditions. The three tests of the MSFC were practised at least once before baseline assessments were completed. For this study, however, we only investigated T25FW and 9HPT data and no data of the Paced Auditory Serial Addition Test were included. Therefore, we did not create a MSFC sum score by calculating Z-scores. If patients were unable to perform a test due to MS-related symptoms, the maximum allowed time for this test was assigned.

Analyses and statistics

Because we wanted the endpoint to be clinically interpretable and applicable, we decided to use certain predefined changes in T25FW and 9HPT[11] rather than mean change scores. We have previously shown that 20% changes in T25FW and 9HPT are clinically meaningful.[12,13] Concerning definitions of progression on the EDSS, we used the earlier defined change of 1 point on this scale in patients with baseline EDSS < 5.5 or 0.5 point in patients with an entry EDSS of ≥ 5.5 as a clinically meaningful change.[14]

We calculated absolute differences in the scores of the EDSS and absolute as well as relative differences in the scores of the T25FW and 9HPT over 1 year and over 2 years. We then determined event rates: percentages (with 95% CI) of patients who showed clinically meaningful deterioration on these scales, and we investigated how these event rates changed when combining the different clinical scales. For the 9HPT, we looked at worsening of the function of, at least, one of both hands. We also calculated percentages of patients who showed clinically meaningful improvement on the separate scales, referred to as "improvement".

To get an impression of the extent of inconsistency one measures when using combinations of clinical scales, we also took into account the percentages of patients who showed clinically meaningful improvement on one scale while showing clinically meaningful deterioration on the other scale(s), referred to as "opposing changes". Because real improvement over 1 or 2 years is unlikely to occur in PPMS, both "improvement" on a scale and "opposing changes" on combinations of scales should be minimal and were considered as "false" in this study. Note that "opposing changes" are part of the percentage of worsened patients (opposing changes are within patients on different scales), whereas "improvement" indicates another, separate percentage (in other patients on the same scale).

Furthermore we looked into a possible effect of baseline disability – as measured by the baseline EDSS-score – on the event rate (and with this on the definition of the most appropriate clinical endpoint), by dividing the PPMS population into two subsets: patients with baseline EDSS 2-4.5 (subset A) and patients with baseline EDSS 5-6.5 (subset B).

All statistical analyses were performed by using the Statistical Package for Social Sciences (SPSS) version 12.0.

RESULTS

Patient characteristics

A total of 161 patients fulfilled the selection criteria. During the observation period none of our patients converted to progressive relapsing MS. There were no missing data. Ages at baseline visit ranged from 31 to 67 years, with a mean age of 49 years (SD 8.9). Baseline EDSS scores ranged from 2.0 to 6.5 with a median of 5.0. The mean T25FW score at baseline was 16.5 seconds (SD 21.2). Mean 9HPT scores at baseline were 30.5 seconds (SD 38.5) and 33.6 seconds (SD 38.5) concerning the dominant respectively the non-dominant hand.

After a follow up of approximately 1 (FU1) and 2 (FU2) years, the median EDSS score had changed to 5.5 (range 2.0-7.0) and to 6.0 (range 2.0-8.0). Mean T25FW scores were 18.9 seconds (SD 27.3) at FU1 and 24.3 seconds (SD 35.6) at FU2. Mean 9HPT scores concerning the dominant hand were 30.6 seconds (SD 38.6) at FU1 and 31.2 seconds (SD 39.3) at FU2. With regard to the non-dominant hand, the mean scores were 37.9 seconds (SD 49.2) at FU1 and 38.1 seconds (SD 45.9) at FU2. All patient characteristics are shown in **Table 2.1.1**, also showing median values and interquartile ranges (IQRs) for the T25FW and 9HPT.

Event rates on separate clinical scales

As **Table 2.1.2** shows, most deterioration of functioning was observed on the T25FW: on this test 34% of all patients showed meaningful deterioration at FU1 and 46% at FU2. As for the EDSS, the event rates were 17% at FU1 and 32% at FU2. On the 9HPT, 20% of all patients showed worsening at FU1 and 24% at FU2. For all three clinical scales, a substantial rate of improvement could be seen, especially during the first year (also shown in **Table 2.1.2**).

Event rates on combinations of clinical scales

As one would expect by making different combinations of T25FW, 9HPT and EDSS, the event rates increased. **Table 2.1.3** shows event rates for all possible combinations. The event rate was, obviously, highest when we looked at deterioration on (at least) one of three scales, both at FU1 and at FU2: 52% and 63%. When we looked at deterioration on a combination of two scales, on the combination of 9HPT/EDSS event rates were low (34% at FU1 and 43%

Table 2.1.1 Patient characteristics at baseline and follow-up

Characteristics	Baseline	At FU1	At FU2
Age			
mean (SD), y	49.4 (8.9)		
EDSS			
median (range)	5.0 (2.0-6.5)	5.5 (2.0-7.0)	6.0 (2.0-8.0)
T25FW			
mean (SD), s	16.5 (21.2)	18.9 (27.3)	24.3 (35.6)
median (IQR), s	9.7 (7-17)	9.9 (7-18)	11.4 (7-21)
9HPT(dom)			
mean (SD), s	30.5 (38.5)	30.6 (38.6)	31.2 (39.3)
median (IQR), s	23.0 (20-28)	23.0 (20-28)	23.0 (20-29)
9HPT(n-dom)			
mean (SD), s	33.6 (38.5)	37.9 (49.2)	38.1 (45.9)
median (IQR), s	25.0 (22-31)	25.1 (21-33)	26.0 (22-35)

EDSS: Expanded Disability Status Scale, T25FW: Timed 25-foot Walk, IQR: Interquartile Range, 9HPT(dom): 9-hole Peg test dominant hand, 9HPT(n-dom): 9-hole Peg test non-dominant hand, FU1: first follow-up (1 year), FU2: second follow-up (2 years).

Table 2.1.2 Percentages [with 95% CI] of patients who showed clinically meaningful deterioration on separate clinical scales (*with improvements on separate scales*), total population (n=161)

	Event rates at FU1			**Event rates at FU2**		
Clinical scale	Deterioration	95% CI	*Improvement*	Deterioration	95% CI	*Improvement*
T25FW	34%	27-41	*17% (n=28)*	46%	38-54	*11% (n=18)*
9HPT	20%	14-27	*11% (n=17)*	24%	17-30	*11% (n=18)*
EDSS	17%	12-23	*12% (n=20)*	32%	24-39	*9% (n=15)*

T25FW: Timed 25-foot Walk, 9HPT: 9-hole Peg test, EDSS: Expanded Disability Status Scale, FU1: first follow-up (1 year), FU2: second follow-up (2 years).

at FU2). At FU1, the event rate was highest for deterioration on either T25FW or 9HPT: 46% (vs. 41% for deterioration on either T25FW or EDSS). At FU2 event rates were approximately equal for deterioration on either T25FW or EDSS and for deterioration on either T25FW or 9HPT (57% and 56%).

When examining the worsened patients who showed improvement on one scale while deteriorating on the other(s), the opposing changes, we found – again as one would expect

Table 2.1.3 Percentages [with 95% CI] of patients who showed clinically meaningful deterioration on combinations of clinical scales (*with percentages of opposing changes*), total population (n=161)

	Event rates at FU1			**Event rates at FU2**		
Combination of clinical scales	Deterioration	95% CI	*Opposing changes*	Deterioration	95% CI	*Opposing changes*
Either T25FW or 9HPT	46%	38-54	*9%*	56%	48-64	*6%*
Either T25FW or EDSS	41%	33-49	*4%*	57%	49-65	*4%*
Either 9HPT or EDSS	34%	27-41	*6%*	43%	35-51	*5%*
Either T25FW or 9HPT or EDSS	52%	44-60	*14%*	63%	56-71	*11%*

T25FW: Timed 25-foot Walk, 9HPT: 9-hole Peg test, EDSS: Expanded Disability Status Scale, FU1: first follow-up (1 year), FU2: second follow-up (2 years).

– that these percentages were highest on the combination of all three scales: 14% at FU1 and 11% at FU2. When looking at a combination of two scales, opposing changes were highest on the combination of T25FW/9HPT: 9% at FU1 and 6% at FU2. Opposing changes were lowest on the combination of T25FW/EDSS: 4% both at FU1 and at FU2 (**Table 2.1.3**).

Effect of baseline disability on event rates

In both subsets of patients, event rates on separate clinical scales were highest on the T25FW. In general, at FU1 and FU2, event rates were higher in subset B (EDSS 5-6.5) than in subset A (EDSS 2-4.5): on the T25FW 55 vs. 36% worsening, on the 9HPT 27 vs. 20%, and on the EDSS (less clear) 33 vs. 31%, all at FU2. Also the proportion of patients improving on the separate scales was higher in subset B (data not shown).

When combining clinical scales, the event rates were again, in both subsets, highest on a combination of all three scales. When looking at a combination of two scales, in subset A event rates were approximately equal for deterioration on either T25FW or EDSS and for deterioration on either T25FW or 9HPT (33% at FU1, 48 and 49% at FU2). Rates of opposing changes were, again, highest on the combination of T25FW/9HPT, at FU1 and FU2. On the combinations of T25FW/EDSS and 9HPT/EDSS the opposing changes were lower and equal (**Table 2.1.4**).

In subset B, event rates at FU1 were highest for deterioration on either T25FW or 9HPT (57 vs. 48% for deterioration on either T25FW or EDSS). Event rates at FU2 were approximately equal for deterioration on either T25FW or EDSS and for deterioration on either T25FW or 9HPT

Table 2.1.4 Percentages [with 95% CI] of patients who showed clinically meaningful deterioration on combinations of clinical scales (*with percentages of opposing changes*), subset A and B

Subset A (EDSS 2-4.5) n=75	**Event rates at FU1**			**Event rates at FU2**		
Combination of clinical scales	Deterioration	95% CI	*Opposing changes*	Deterioration	95% CI	*Opposing changes*
Either T25FW or 9HPT	33%	23-44	*4%*	48%	37-59	*1%*
Either T25FW or EDSS	33%	23-44	*1%*	49%	38-61	*0%*
Either 9HPT or EDSS	32%	21-43	*1%*	39%	28-50	*0%*
Either T25FW or 9HPT or EDSS	41%	30-52	*5%*	56%	45-67	*1%*
Subset B (EDSS 5-6.5) n=86	**Event rates at FU1**			**Event rates at FU2**		
Combination of clinical scales:	Deterioration	95% CI	*Opposing changes*	Deterioration	95% CI	*Opposing changes*
Either T25FW or 9HPT	57%	47-67	*13%*	63%	53-73	*10%*
Either T25FW or EDSS	48%	37-58	*6%*	64%	54-74	*7%*
Either 9HPT or EDSS	36%	26-46	*9%*	47%	36-57	*9%*
Either T25FW or 9HPT or EDSS	62%	51-72	*21%*	70%	60-79	*19%*

T25FW: Timed 25-foot Walk, 9HPT: 9-hole Peg test, EDSS: Expanded Disability Status Scale, FU1: first follow-up (1 year), FU2: second follow-up (2 years).

(64 and 63%). Both at FU1 and at FU2, opposing changes were highest on the combination of T25FW/9HPT and lowest on the combination of T25FW/EDSS (**Table 2.1.4**).

Tables 2.1.2-2.1.4 show that on all (combinations of) scales, at FU2 rates of clinically meaningful deterioration were higher than at FU1, whereas rates of improvement and opposing changes were lower.

CONCLUSIONS

First, it is clear that in the PPMS population of this study, the T25FW has the highest event rate as a single scale, independent of the baseline disability level.

Second, in our study population, a period of 2 years turns out to be much more meaningful to look at than just 1 year. Over 2 years rates of clinically meaningful worsening are higher

(as expected in PPMS[15-17]), whereas rates of clinically meaningful improvement (on separate scales) and opposing changes (measured when using combinations of two or three clinical scales), which in PPMS probably do not indicate true improvement of functioning, are lower. The fact that these last rates are lower at 2 years than at 1 year is completely in line with the progressive nature of the disease in PPMS and suggest that these "improvements" at 1 year are of very short duration or even are caused by measurement error. If we interpret the difference between deterioration rates and improvement/opposing changes as a measure of "signal-to-noise" we see that 2 year data are much more reliable.

Third, by combining data from T25FW and 9HPT with EDSS scores, it is indeed possible to define a valuable endpoint that has much higher event rates than the EDSS alone. Mainly combinations with T25FW (T25FW/9HPT or T25FW/EDSS) lead to high event rates. However, our data showed that when combining different clinical scales, a substantial amount of improvement is measured, especially when using combinations of all three clinical scales or the combination of T25FW/9HPT (i.e., only the MSFC-components without the EDSS). For T25FW and 9HPT, this might be due to the known learning curve in the MSFC-tests, although all tests were practiced at least once before baseline assessment. Because of this amount of measured concomitant improvement and hence inconsistency, it is important to select the right combination of different scales. In line with the aforementioned, FU2 should be the decisive time point in this selection rather than FU1.

A combination of 9HPT and EDSS seems less desirable because this combination produces low event rates, at FU2 even lower than T25FW alone. Also, the combination of all three scales seems less attractive because, as already mentioned, this combination picks up too much inconsistency. This leaves two options to add to the T25FW: 9HPT or EDSS. At 1 year, the event rate is higher on a combination of T25FW with 9HPT. However, at the more decisive time point of 2 years, the event rate on a combination of T25FW and EDSS is (slightly) higher and, maybe even more important, this combination is more consistent. From this, we can conclude that a combination of T25FW and EDSS is the most desirable combination of different clinical scales: we define the endpoint as "worsening on either T25FW or EDSS".

When adding up the event rates on separate scales (Table 2.1.2) and comparing this with the event rates on combinations of scales (Table 2.1.3), one can conclude that there is an overlap when we look at deterioration on one of two (or more) clinical scales. But, this overlap (which means that one worsens on both scales) only forms a minor part of the event rate.

There are multiple reasons to support the usefulness of including deterioration on *only one of several* different clinical scales. First, this is most obvious for scales that measure motor function

in different parts of the body, for example a patient who is wheelchair-bound in whom a change in arm function can obviously occur without concomitant change in leg function. Also, in the case of a composite endpoint consisting of T25FW and EDSS, different properties of T25FW and EDSS lead to different selections of worsened patients and consequently enlarge the chance of detecting a change of function. For instance, looking at the more disabled patients, the T25FW is much more likely to change than the EDSS, which probably can be explained by ceiling effects of the EDSS and longer "staying times" at this level.[18] Vice versa, starting the use of a stick is a significant hallmark of the EDSS, whereas this aid might add stability so that the T25FW does not significantly worsen. We strongly feel, based on our data, that a combination of the two measures gives complementary information in PPMS patients. Different aspects of walking function, with different measurement consequences, were also found when developing and using the MS Walking Scale.[19]

The baseline disability level has no influence on selection of "worsening on either T25FW or EDSS" as the most desirable clinical endpoint: this applies to both subsets of patients. In general, in the subset of patients at higher baseline disability levels, event rates were higher than in the subset of patients with milder baseline disability. The more disabled patients showed more worsening on both EDSS and MSFC-components but also more improvement.

Finally, a number of limitations of this study need to be addressed. First, the reported EDSS and MSFC changes were not confirmed by repeated measurements after 3 or 6 months. Second, our study population was a clinical population and not a trial cohort. Third, the data were retrospectively selected for analysis, and it might be possible that some patients came to the hospital because of worsening, although most were seen during regular, scheduled visits to their neurologists. However, there might be some selection bias favoring more progressive patients.

Taken together, in this study, we defined a new clinical endpoint, with the goal to improve outcome measures for future trials in PPMS. We hope that this will contribute to designing trials with enough power to identify a treatment effect of future therapeutic agents.

REFERENCES

1. Revesz T, Kidd D, Thompson AJ, Barnard RO, McDonald WI. A comparison of the pathology of primary and secondary progressive multiple sclerosis. Brain 1994; 117: 759-765.
2. Thompson AJ, Kermode AG, MacManus DG, et al. Patterns of disease activity in multiple sclerosis: clinical and magnetic resonance imaging study. BMJ 1990; 300: 631-634.
3. Thompson AJ, Kermode AG, Wicks D, et al. Major differences in the dynamics of primary and secondary progressive multiple sclerosis. Ann Neurol 1991; 29: 53-62.
4. Leary SM, Miller DH, Stevenson VL, Brex PA, Chard DT, Thompson AJ. Interferon beta-1a in primary progressive MS: an exploratory, randomized, controlled trial. Neurology 2003; 60: 44-51.
5. Montalban X. Overview of European pilot study of interferon beta-Ib in primary progressive multiple sclerosis. Mult Scler 2004; 10 Suppl 1: S62-S64.
6. Wolinsky JS, Narayana PA, O'Connor P, et al. Glatiramer acetate in primary progressive multiple sclerosis: results of a multinational, multicenter, double-blind, placebo-controlled trial. Ann Neurol 2007; 61: 14-24.
7. Poser CM, Paty DW, Scheinberg L, et al. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983; 13: 227-231.
8. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 1996; 46: 907-911.
9. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33: 1444-1452.
10. Cutter GR, Baier ML, Rudick RA, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999; 122: 871-882.
11. Schwid SR, Goodman AD, McDermott MP, Bever CF, Cook SD. Quantitative functional measures in MS: what is a reliable change? Neurology 2002; 58: 1294-1296.
12. Hoogervorst EL, Kalkers NF, Cutter GR, Uitdehaag BM, Polman CH. The patient's perception of a (reliable) change in the Multiple Sclerosis Functional Composite. Mult Scler 2004; 10: 55-60.
13. Kragt JJ, van der Linden FA, Nielsen JM, Uitdehaag BM, Polman CH. Clinical impact of 20% worsening on Timed 25-foot Walk and 9-hole Peg Test in multiple sclerosis. Mult Scler 2006; 12: 594-598.
14. Goodkin DE. EDSS reliability. Neurology 1991; 41:332.
15. Cottrell DA, Kremenchutzky M, Rice GP, Hader W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: a geographically based study. 6. Applications to planning and interpretation of clinical therapeutic trials in primary progressive multiple sclerosis. Brain 1999; 122: 641-647.
16. Cottrell DA, Kremenchutzky M, Rice GP, et al. The natural history of multiple sclerosis: a geographically based study. 5. The clinical features and natural history of primary progressive multiple sclerosis. Brain 1999; 122: 625-639.
17. Ingle GT, Stevenson VL, Miller DH, Thompson AJ. Primary progressive multiple sclerosis: a 5-year clinical and MR study. Brain 2003; 126: 2528-2536.

18. Kragt JJ, Thompson AJ, Montalban X, et al. Responsiveness and predictive value of EDSS and MSFC in primary progressive MS. Neurology 2008; 70: 1084-1091.

19. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. Measuring the impact of MS on walking ability: the 12-Item MS Walking Scale (MSWS-12). Neurology 2003; 60: 31-36.