# Chapter 4.2

# Outcome measurement in multiple sclerosis: Detection of clinically relevant improvement

Lisa ML van Winsen MD, Jolijn J Kragt MD , Erwin LJ Hoogervorst PhD,
Chris H Polman PhD, Bernard MJ Uitdehaag PhD

*Submitted Multiple Sclerosis*

## Abstract

Detecting clinically relevant changes in patients with MS has become more important with the availability of effective disease modifying therapy (DMT). The ability of detecting improvement is sparsely studied. To evaluate the responsiveness of the EDSS and two quantitative tests separately and in combination to detect improvement after IVMP. EDSS, the T25FW and the 9-Hole Peg Test (9-HPT) were assessed in 101 MS patients before and 6 weeks after IVMP. In addition patients were asked to rate their change as an anchor to evaluate the performance of the tests.

Combining the T25FW and the 9-HPT turned out to be the optimal combination of measures to predict patient perceived improvement (likelihood ratio of positive test 2.84 (95%CI 1.28 – 6.31). In the higher EDSS range (EDSS 4.5 and higher) for all measures a significant change was more often perceived as clinically relevant than in the lower disability range.

The EDSS is not the preferred outcome of choice to detect patient perceived improvement in MS, especially in the lower EDSS range. Combining T25FW and 9-HPT can improve the sensitivity to detect clinically relevant changes without conceding with respect to specificity.

## Introduction

Outcome measures used to evaluate therapeutic interventions should be valid, reliable and responsive. Responsiveness has been defined as an instrument's ability to detect change over time. In MS the EDSS is the most widely used outcome measure in clinical trials. It is a physician-oriented outcome scale that scores the findings of the neurological examination on eight functional systems (23). It has been shown that the EDSS is not very responsive to change, at least when worsening of the disease is considered (143, 144, 157). Over the past two decades, due to the introduction of disease modifying therapy (DMDs) and the awareness of the importance of patient-oriental outcomescales, new scales for MS have been introduced. These aim to be more responsive. The use of these scales is restrained because there is doubt whether changes of these scales are associated with clinically relevant changes.

As a result, most randomized clinical trials keep using change in EDSS as an outcome for disability progression. In addition, most trials focus on the ability of interventions to prevent worsening. We have recently shown in primary progressive MS that a 20% worsening in the T25FW, a quantative test of ambulatory function (24), has a higher event rate than the EDSS. Combining EDSS and T25FW resulted in a further increase (158).

Assessing improvement in MS might different from deteriovation. Unfortunately, so far there are only few situations in which clinical neurological improvement can be studied in MS. Recovery from exacerbations offers such a possibility. Corticosteroids are widely used to accelerate recovery from functional disability due to acute exacerbations of MS (40). High doses of IVMP given for 3-5 days are recommended as first-line treatment of MS relapses (159) (160). In our hospital IVMP is also occasionally offered to progressive patients with recent deterioration.

Studies investigating relative responsiveness of different outcome measures after IVMP treatment are sparse. Percentages of patients with a clinically response, i.e. a decrease in disability, as measured by the EDSS varied (25 up to 93%) and was usually, although not always, much lower than improvement as perceived by the patient. Varying results were obtained using quantitative tests like the T25FW or the 9-Hole Peg Test (9-HPT) (161-163),

In the present study, we evaluated the responsiveness of two quantitative tests (T25FW and 9-HPT) and the EDSS in a population of MS patients treated with high dose IVMP. Responsiveness was evaluated using an anchor-based approach. This requires an external independent standard to 'anchor' the meaning of clinical importance. For this we used the transition question to evaluate the patient's perceived change over a period of six weeks. In transition questions are asked to compare their prior health state and compare it to how they are feeling currently (164). The aim was to evaluate the responsiveness of the EDSS and the quantitative tests separately and when applied in combination.

## Patients and methods

### Patients

MS patients who were treated with IVMP at our MS clinic underwent EDSS, T25FW and 9-HPT examinations and were evaluated for their own perception of the treatment effect. Examinations were done prior to IVMP and 6 weeks after treatment, when treatment effect was likely to have been occurred (165). The treatment consisted of a daily dose of 1000 mg IVMP for 3 consecutive days or 500 mg IVMP daily for 5 consecutive days. The indication for IV-MP treatment in PPMS patients was subacute deterioration in the absence of relapses. Examinations were performed in the same visit under standardized conditions by well-trained medical doctors as described previously (150). No selection for age, gender, MS subtype or disability was applied. However, patients, who were unable to perform T25FW and 9-HPT during both visits, were excluded from analysis to avoid ceiling effects, which adversely influence responsiveness (143). The study was approved by the Medical Ethical Committee of the VU University Medical Centre.

### Outcome measures

The patient perceived change was taken as external criterion. Patients rated the change in one of four categories: no recovery at all, little recovery, moderate recovery or complete recovery. For the analysis we dichotomized patients perception in clinical relevant improvement (including categories moderate recovery and complete recovery) and no clinical relevant improvement (including categories little recovery and no recovery). Little recovery was included in the latter to obtain a more robust measure of clinical relevant improvement.

For a significant change in EDSS we applied the following definition: EDSS change 1.0 or more for baseline EDSS < 5.5 and EDSS change 0.5 or more for baseline EDSS ≥ 5.5 (153). For both the T25FW and the 9-HPT a  20% change was used as a threshold for a real change (166). For the 9-HPT the score of the dominant hand was used.

When combining tests, an improvement was only granted when there were no 'opposing changes', i.e. when there was no simultaneous significant worsening on one of the other tests. Because we would like to evaluate the different tests in patients with and without walking difficulties we divided patients according to baseline EDSS (4.0 and lower versus 4.5 and higher).

### Statistical analysis

To determine the performance of the measure, each result was categorized as true positive (TP), false negative (FN), true negative (TN), and false positive (FP) according to the dichotomized patient perception. The following predictive characteristics were calculated: sensitivity (TP/ [TP + FN]), specificity (TN/ [TN + FP]), positive predictive value (PPV) (TP/ [TP + FP]), and

negative predictive value (NPV) (TN/ [TN + FN]). Furthermore, we determined likelihood ratios (LRs). In this study the LR is the ratio of the probability of the specific result in patients experiencing an improvement to the probability in patients not experiencing an improvement. We performed the following calculations: LR of a positive test (LR+) = sensitivity/ (1 - specificity), LR of a negative test (LR-) = (1 - sensitivity)/specificity. A LR greater than 1 indicated that the test result was associated with improvement, whereas a LR less than 1 indicated that the test result was associated with the absence of improvement. The significance of the association between the different outcomes and the patient perception was evaluated using chi square or Fisher's exact statistics.

## Results

### Patient characteristics

A total of 116 MS patients were treated with IVMP and underwent all tests. Of these patients three patients were unable to perform both 9-HPT and T25FW before and after IVMP treatment. Another nine patients were unable to perform the T25FW before and after IVMP. Furthermore, 1 baseline T25FW, 1 follow-up 9-HPT and 1 patient perception value was missing. This resulted in 101 patients eligible for analysis. The patient characteristics are listed in table 1.

### Outcome measures

Table 2 shows the proportions of patient's perception of change and significant changes on the separate tests. Forty-five patients (45%) judged themselves as improved while only 25 patients (25%) improved significantly on the EDSS score. The association between EDSS changes and the patient's perception is shown in table 3. This association was not significant (p = 0.157).

**Table 1.** Patient characteristics

| N | 101 |
|---|---|
| age[1] | 44 (10) |
| Female[2] | 58 (57%) |
| Disease duration[1] | 10.5 (7.4) |
| MS subtype | |
| RRMS[2] | 62 (61%) |
| SPMS[2] | 30 (30%) |
| PPMS[2] | 8 (8%) |
| CIS[2] | 1 (1%) |
| EDSS[3] | 4.0 (3.5 to 6.0) |
| 9-HPT[4] | 22.5 (19.3 to 27.1) |
| T25FW[4] | 6.3 (4.7 to 8.4) |

EDSS = Expanded Disability Status Scale, 9-HPT = 9 Hole Peg Test, T25FW = Timed 25-foot Walk
[1] in years, mean (standard deviation) [2] number (%) [3] median (interquartile range) [4] in seconds, median (interquartile range)

RR = relapsing remitting, SP = secondary progressive, PP = primary progressive, CIS = clinically isolated syndrome.

**Tabel 2.** Outcome measures

| | | |
|---|---|---|
| Patient's perception of improvement | Not at all | 30 (30%) |
| | A little | 26 (26%) |
| | Moderate | 39 (39%) |
| | Complete | 6 (6%) |
| EDSS | Worsened | 5 (5%) |
| | No change | 71 (70%) |
| | Improved | 25 (25%) |
| 9-HPT | Worsened | 7 (7%) |
| | No change | 84 (83%) |
| | Improved | 10 (10%) |
| T25FW | Worsened | 7 (7%) |
| | No change | 75 (74%) |
| | Improved | 19 (20%) |

Patient's perception of improvement: proportion of patients who experienced no, little, moderate or complete improvement using absolute numbers (n) and percentages (%). EDSS = Expanded Disability Status Scale, 9-HPT = 9 Hole Peg Test, T25FW = Timed 25-foot Walk

**Table 3.** Patient's perception of improvement versus EDSS scores

| | Patient's perception of improvement | | | |
|---|---|---|---|---|
| | complete | moderate | a little | not at all |
| significant improvement on EDSS | 3 (12%) | 11 (44%) | 4 (16%) | 7 (28%) |
| no change | 3 (4%) | 27 (38%) | 22 (31%) | 19 (27%) |
| significant worsening on EDSS | 0 (0%) | 1 (20%) | 0 (0%) | 4 (80%) |

EDSS = Expanded Disability Status Scale. Proportion of patients who improved, did not change or worsened on EDSS versus patients who experienced complete, moderate, a little or no improvement at all using absolute numbers (n) and percentages (%).

**Prediction of perceived improvement by the outcome measures.**

Table 4 shows the characteristics of the individual outcomes and combinations of outcomes when compared to the dichotomized patient's perception as external anchor for the total group. All measures and combinations are characterized by a low sensitivity and a much higher specificity to detect improvement perceived as important to the patient. The PPV was lowest for the EDSS score (56%) and highest for the 9-HPT (70%). The optimal and most accurate combination of tests was T25FW and 9-HPT with a PPV of 70% and a NPV of 63%, and a positive likelihood ratio of 2.84.

**Stratification for baseline EDSS.**

When stratifying the group by ambulatory abilities at baseline (EDSS ≤ 4.0, i.e. ambulatory without aid or rest for ≥ 500 m and EDSS > 4.0, i.e. not able to walk 500 m), clear differences were observed (table 5). In the higher baseline EDSS range sensitivity considerably increased without affecting the specificity for all measures under study. In the lower baseline EDSS range, the sensitivity dramatically dropped without a clear increase in specificity. Both PPV and NPV

**Table 4.** Outcome measures and their validation

| Sign improvement (without opposing changes) in | Sensitivity (%) | Specificity (%) | PVV % | NPV (%) | LR+ | LR- | p-value |
|---|---|---|---|---|---|---|---|
| EDSS | 31 (20-46) | 80 (68-89) | 56 (36-73) | 59 (48-70) | 1.58 (0.80-3.14) | 0.86 (0.68-0.08) | 0.137 |
| 9-HPT | 16 (8-29) | 95 (85-98) | 70 (40-89) | 58 (48-68) | 2.90 (0.8-10.59) | 0.89 (0.78-1.03) | 0.090 |
| T25FW | 29 (18-43) | 89 (79-95) | 68 (46-85) | 61 (50-71) | 2.70 (1.11-6.53) | 0.80 (0.65-0.98) | 0.020 |
| EDSS and/or T25FW | 42 (29-57) | 77 (64-86) | 59 (42-74) | 62 (51-73) | 1.82 (1.01-3.27) | 0.75 (0.56-1.00) | 0.041 |
| EDSS and/or 9-HPT | 38 (25-52) | 79 (66-87) | 59 (41-74) | 61 (50-72) | 1.76 (0.94-3.3) | 0.79 (0.61-1.03) | 0.057 |
| T25FW and/or 9-HPT | 36 (23-50) | 88 (76-94) | 70 (49-84) | 63 (52-73) | 2.84 (1.28-6.31) | 0.74 (0.58-0.93) | 0.006 |
| EDSS and/or T25FW and/or 9-HPT | 44 (31-59) | 75 (62-84) | 59 (42-74) | 63 (51-73) | 1.78 (1.02-3.11) | 0.74 (0.55-1.00) | 0.033 |

T25FW = Timed 25-foot Walk, 9-HPT = 9 Hole Peg Test, EDSS = Expanded Disability Status Scale, PVV = positive predictive value, NPV=negative predictive value, LR+ = positive likelihood ratio, LR- = negative likelihood ratio. 95% confidence interval between brackets. p-value of chi square test.

**Table 5a.** Outcome measures and their validation divided in EDSS ≤ 4.0

| Sign improvement (without opposing changes) in | Sensitivity (%) | Specificity (%) | PVV (%) | NPV (%) | LR+ | LR- | p-value |
|---|---|---|---|---|---|---|---|
| EDSS | 10 | 83 | 43 | 43 | 0.62 | 1.08 | 0.499 |
| | (4 - 26) | (64 - 93) | (16 - 75) | (30 - 58) | (0.15 - 2.51) | (0.87 - 1.34) | |
| 9-HPT | 7 | 92 | 50 | 45 | 0.83 | 1.02 | 0.844 |
| | (2 - 22) | (74 - 98) | (15 - 85) | (32 - 59) | (0.13 - 5.44) | (0.87 - 1.19) | |
| T25FW | 14 | 92 | 67 | 47 | 1.66 | 0.94 | 0.532 |
| | (5 - 31) | (74 - 98) | (30 - 90) | (33 - 61) | (0.33 - 8.27) | (0.78 -1.14) | |
| EDSS and/or T25FW | 17 | 75 | 45 | 43 | 0.69 | 1.10 | 0.488 |
| | (8 - 35) | (55 - 88) | (21 - 72) | (29 - 58) | (0.24 - 1.98) | (0.83 - 1.47) | |
| EDSS and/or 9-HPT | 17 | 75 | 45 | 43 | 0.69 | 1.10 | 0.488 |
| | (8 -35) | (55 - 88) | (21 - 72) | (43 - 58) | (0.24 - 1.98) | (0.83 - 1.47) | |
| T25FW and/or 9-HPT | 21 | 88 | 67 | 48 | 1.66 | 0.91 | 0.429 |
| | (10 - 38) | (69 - 96) | (35 - 88) | (34 - 62) | (0.46 - 5.93) | (0.71 - 1.15) | |
| EDSS and/or T25FW and/or 9-HPT | 24 | 71 | 50 | 44 | 0.83 | 1.07 | 0.679 |
| | (12 - 42) | (51 - 85) | (27 - 73) | (29 - 59) | (0.34 - 2.03) | (0.77 - 1.49) | |

**Table 5b.** Outcome measures and their validation divided in EDSS > 4.0.

| Sign improvement (without opposing changes) in | Sensitivity (%) | Specificity (%) | PVV (%) | NPV (%) | LR+ | LR- | p-value |
|---|---|---|---|---|---|---|---|
| EDSS | 69 | 78 | 61 | 83 | 3.14 | 0.4 | 0.002 |
| | (44 - 86) | (61 - 89) | (39 - 80) | (66 - 93) | (1.51 - 6.54) | (0.19 - 0.85) | |
| 9HPT | 31 | 97 | 83 | 74 | 10.0 | 0.71 | 0.005 |
| | CI14 - 56) | (84 - 99) | (44 - 97) | (59 - 85) | (1.27 - 78.58) | (0.51 - 0.99) | |
| T25FW | 56 | 88 | 69 | 80 | 4.50 | 0.5 | 0.001 |
| | (33 - 77) | (72 - 95) | (42 - 87) | (64 - 90) | (1.63 - 12.4) | (0.28 - 0.88) | |
| EDSS and/or T25FW | 88 | 78 | 67 | 93 | 4.00 | 0.16 | 0.114 |
| | (64 - 97) | (61 - 89) | (45 - 83) | (77 - 98) | (2.03 - 7.9) | (0.04 - 0.59) | |
| EDSS and/or 9-HPT | 75 | 81 | 67 | 87 | 4.00 | 0.31 | < 0.001 |
| | (51 - 90) | (65 - 91) | (44 - 84) | (70 - 95) | (1.84 - 8.68) | (0.13 - 0.73) | |
| T25FW and/or 9-HPT | 63 | 88 | 71 | 82 | 5.00 | 0.43 | < 0.001 |
| | (39 - 82) | (72 - 95) | (45 - 88) | (66 - 92) | (1.85 - 13.49) | (0.22 - 0.82) | |
| EDSS and/or T25FW and/or 9-HPT | 81 | 78 | 65 | 89 | 3.71 | 0.24 | < 0.001 |
| | (57 - 93) | (61 - 89) | (43 - 82) | (73 - 96) | (1.85 - 7.45) | (0.09 - 0.68) | |

T25FW = Timed 25-foot Walk, 9-HPT = 9 Hole Peg Test, EDSS = Expanded Disability Status Scale, PVV = positive predictive value, NPV=negative predictive value, LR+ = positive likelihood ratio, LR- = negative likelihood ratio. 95% confidence interval between brackets. p-value of chi square test

were low in the lower baseline EDSS stratum with results for the EDSS alone being the worst. Although not as good as for the whole group, also in the lower baseline EDSS stratum the combination of T25FW and 9-HPT turned out to be optimal. In the higher baseline EDSS stratum the combination of EDSS and T25FW obtained optimal characteristics.

## Discussion

In this study we evaluated different physician oriented outcome measures to detect a reliable improvement after IVMP treatment, using patient's own perception as an external standard. Combining T25FW and 9-HPT using a cut-off of 20% as significant change turned out to be the optimal combination of measures to predict patient's perceived improvement.

So far, there has been strong persistence among scientists, clinical trial designers and regulatory authorities in adhering to the EDSS to document disability progression in MS clinical trials and reluctance to accept alternative outcome measures, which is partially due to concerns about the clinical relevance of certain changes on the latter. This study shows that this is not justified. The accuracy of a 20% change on the T25FW or 9-HPT, in terms of being perceived as clinically relevant by the patient, was even better in comparison with the traditional one step change of the EDSS. In the study population the PPV, which points to the likelihood that the significant change is relevant to the patient, is clearly higher for both the T25FW and the 9-HPT when compared to the EDSS. Previous studies have also shown that the MSFC, a measure which contains the T25FW and 9-HPT, was more sensitive to detect improvement after IVMP than the EDSS (161, 162), however this could not be confirmed by another study, in which the relative sensitivity for improvement of the MSFC was low (163).

Although the EDSS is believed to be an impairment scale in its lower range and a disability scale in its upper range (167), in the lower range it was not very sensitive to detect improvement after IVMP according to the patient's perception. It was a consistent trend in this study, for all measurements included, that a significant change in the higher disability range (for this study defined as baseline EDSS 4.5 or higher) was more often perceived as clinically relevant than in the lower disability range. The studies included in the Cochrane review in which the value of corticosteroids for acute exacerbations in MS was confirmed (30), were fully based on EDSS scores. In these studies baseline EDSS scores ranged from 4 to 6.

Although previous studies have shown that the scores for deterioration and improvement are not necessarily equal (168, 169), it is interesting to elaborate on the results of our study in the light of MS clinical trials on DMT. Especially in patients with an EDSS below 4.5, who were by far the largest population in the pivotal trials for the currently approved DMT, the accuracy of the EDSS change is disappointing. Recently there has been pointed to specific EDSS related problems in MS clinical trials, stating that small changes on EDSS in RRMS patients are invalid, even if confirmed at 3 or 6 months (170). Our data do not allow us to address the question of confirmation for the several outcome measures applied in our study.

The use of a transition question as an anchor is a critical issue in our study design. It is

required that patients are able to recall their prior health status when they return, which depends on the patient's cognitive status and the time between the two visits. A drastic grounding point, such as a neurological worsening needing intervention, increases the ability of the patient to compare his health status after treatment with the baseline visit (164). It is also important to realize that the anchor we used is not measuring the same construct as the other outcome measures. Therefore, we cannot exclude that some of the perceived improvements were actually due to mood or reflect changes in life-events rather than disease changes (171). However, conversely also clear functional improvements apparent for both patient and physician not always translate into EDSS changes.

Our study also shows that it is possible to develop combined outcome measures. In these a clinically significant change is determined on the basis of a change in one out of two (or even three) scales) that allow to detect a higher number of patients who show a clinically significant change, without leading to a concomitant loss of accuracy in terms of being relevant to the patient. This approach has recently been recommended by us based on a different type of analysis in a completely separate patient population (158). Such an increase in patients who have experienced a clinically significant change during a certain observation period does have a favourable impact on the number of patients enrolled, which nowadays is especially critical in relation to the acceptance of a placebo arm in such trials.

In conclusion, the present study shows that the EDSS seems not the preferred outcome measure of choice to detect improvement in MS as noted by the patient, especially in the lower EDSS range. Combining T25FW and 9-HPT can improve the sensitivity to detect clinically relevant changes without conceding with respect to specificity. The use of combinations of outcome measures in MS should be further explored.

## Acknowledgement

The MS centre VUmc is partially funded by a program grant of the Dutch MS Research Foundation. We take responsibility for the integrity of the data and the accuracy of the data analysis.