The usefulness of evaluative outcome measures in patients with multiple sclerosis


Departments of †Rehabilitation Medicine, §Neurology, ∥Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands and ¶EMGO Institute, VU University Medical Center, Amsterdam, The Netherlands

Correspondence to: Vincent de Groot, Department of Rehabilitation Medicine, VU University Medical Center, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands
E-mail: v.degroot@vumc.nl

To select the most useful evaluative outcome measures for early multiple sclerosis, we included 156 recently diagnosed patients in a 3-year follow-up study, and assessed them on 23 outcome measures in the domains of disease-specific outcomes, physical functioning, mental health, social functioning and general health. A global rating scale (GRS) and the Expanded Disability Status Scale (EDSS) were used as external criteria to determine the minimally important change (MIC) for each outcome measure. Subsequently, we determined whether the outcome measures could detect their MIC reliably. From these, per domain the outcome measure that was found to be most sensitive to changes (responsive) was identified. At group level, 11 outcomes of the domains of physical functioning, mental health, social functioning and general health could reliably detect the MIC. Of these 11, the most responsive measures per domain were the Medical Outcome Study 36 Short Form sub-scale physical functioning (SF36pf), the Disability and Impact Profile (DIP) sub-scale psychological, the Rehabilitation Activities Profile sub-scale occupation (RAPocc) and the SF36 sub-scale health, respectively. Overall, the most responsive measures were the SF36pf and the RAPocc. In individual patients, none of the measures could reliably detect the MIC. In sum, in the early stages of multiple sclerosis the most useful evaluative outcome measures for research are the SF36pf (physical functioning) and the RAPocc (social functioning).

Keywords: multiple sclerosis; evaluative outcome measures; responsiveness; minimally important change; smallest real change

Abbreviations: DIP = Disability and Impact Profile; EDSS = Expanded Disability Status Scale; GAS = Graphic Assessment Scale; MIC = minimally important change; MSFC = Multiple Sclerosis Functional Composite Measure; NHPT = nine-hole peg test; RAPocc = Rehabilitation Activities Profile sub-scale occupation; SaGAS = Short and Graphic Assessment Scale; SF36pf = Medical Outcome Study 36 Short Form sub-scale physical functioning; TWT = timed-walk test


Introduction

The Expanded Disability Status Scale (EDSS) is a frequently used and well-known outcome measure for multiple sclerosis. However, it is criticized because it has unsatisfactory validity, and its reliability is poor (Noseworthy, 1994; Sharrack and Hughes, 1999; Hobart et al., 2000). In response to this situation, the National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force reviewed a large number of data sets to determine which outcome measures would adequately reflect the consequences of the disease and are capable of reliably assessing these consequences. (Cutter et al., 1999; Fischer et al., 1999). This led to the development of the Multiple Sclerosis Functional Composite Measure (MSFC), which consists of the 25-foot timed-walk test (TWT), the nine-hole peg test (NHPT) and the paced auditory serial addition test (PASAT). Originally, the Task Force intended to include a measure of visual acuity, but no reliable measure could be found. The MSFC is intended to replace the EDSS as outcome measure in current and future trials (Cutter et al., 1999; Miller et al., 2000; Cohen et al., 2001). The interpretation of the scores of the individual components of the MSFC is straightforward. However, the total score, which results from a relatively complex formula to combine the component scores, is more difficult to interpret. An adaptation of the MSFC, the short and graphic assessment scale (SaGAS), (Vaney et al., 2004) uses only the TWT and the NHPT. Through specific transformation,
a score is obtained that should be easier to interpret. Other newly developed disease-specific outcomes are the multiple sclerosis impact scale (Hobart et al., 2001a) and the Guy’s Neurological Disability Scale (Sharrack and Hughes, 1999). In addition to these new, disease-specific, measures, several other disability and quality of life measures have been used in research into this illness (Granger et al., 1990; Kidd et al., 1995; Jonsson et al., 1996; Lankhorst et al., 1996; Ottenbacher et al., 1996; Cohen et al., 1999; Pfennings et al., 1999a; Van der Putten et al., 1999; Freeman et al., 2000; Hobart et al., 2001b).

Responsiveness is an important clinimetric property. It represents the ability to measure change, and is particularly relevant when outcome measures are to be used in longitudinal studies, such as clinical trials (De Vet et al., 2001; Terwee et al., 2003). In connection with multiple sclerosis, however, it has been studied much less extensively than validity and reliability (Koziol et al., 1999; Sharrack and Hughes, 1999; Schwid et al., 2000; Hoogervorst et al., 2001a; Patzold et al., 2002; Uitdehaag et al., 2002; Riazi et al., 2003; Hobart et al., 2004; McGuigan and Hutchinson, 2004). Moreover, in the literature there is no consensus about the exact definition of responsiveness (Terwee et al., 2003). Consequently, there are many currently available methods that have been developed to assess responsiveness (Terwee et al., 2003; Crosby et al., 2003; Husted et al., 2000). It has been shown that applying different methods leads to different conclusions about the absolute responsiveness of an outcome measure (Terwee et al., 2003). However, conclusions about the relative responsiveness, i.e. how do different measures perform in relation to each other, are less dependent on the method used (Terwee et al., 2003). To assess the relative responsiveness, several outcome measures of interest should be included, and parallel assessments should be made at the same points in time.

The methods that can be used to assess whether scores have changed can be sub-divided into distribution-based and anchor-based methods (Lydick and Epstein, 1993; Cella et al., 2002a, b; Schmitt and Di Fabio, 2004) Distribution-based methods, using standardized metrics, focus on the ability of an outcome measure to reliably determine change, and aim to quantify the noise, i.e. the variability of the score changes in the absence of a relevant change. Anchor-based methods focus on the correspondence of the change on the outcome measure of interest with the change on an external criterion (Cella et al., 2002a; Schunemann et al., 2003) and aim to quantify the signal, i.e. the size of the score change when there is a relevant change. The results of anchor-based methods depend on the external criterion and the cut-off point chosen (Cella et al., 2002a). The usefulness of an evaluative outcome measure depends on whether score changes associated with a relevant change can reliably be distinguished from the variability of score changes in absence of a relevant change (Guyatt et al., 1987).

In this study, 23 (sub-)scales of outcome measures were compared. The aim was to select the most useful evaluative outcome measures for the early stages of multiple sclerosis.

**Material and methods**

**Patients**

All consecutive potentially eligible patients visiting the participating neurology outpatient clinics were invited to participate. A cohort of 156 recently (<6 months previously) diagnosed patients, aged 16–55 years, was recruited and followed prospectively for 3 years. Diagnosis was based on the Poser criteria for definite multiple sclerosis (Poser et al., 1983) Patients with other neurological disorders, or systemic or malignant neoplastic diseases, were excluded. The measurements took place at baseline, and 6 months, and after 1, 2 and 3 years. In the case of a relapse, the measurements were postponed for a few weeks until the relapse had subsided. The patients were visited at home in order to minimize drop-out. Four well-trained raters were responsible for the scoring.

**Outcome measures**

We studied the (sub-)scales of the EDSS (Kurtzke, 1983; Whitaker et al., 1995; Rudick et al., 1996), the MSFC (Cutter et al., 1999; Fischer et al., 1999; Cohen et al., 2000; Kalkers et al., 2000, 2001; Miller et al., 2000; Hoogervorst et al., 2001b), the SaGAS (Vaney et al., 2004), the Action Research Arm Test (ARAT) (Lyle, 1981; Van der Lee et al., 2001), the Disability and Impact Profile (DIP) (Laman and Lankhorst, 1994; Jonsson et al., 1996; Lankhorst et al., 1996; Cohen et al., 1999; Pfennings et al., 1999a), the Functional Independence Measure (FIM) (Granger et al., 1990; Kidd et al., 1995; Marolf et al., 1996), the Rehabilitation Activity Profile (RAP) (Van Bennemol et al., 1995, 1996), the Rivermead Mobility Index (RMI) (Collen et al., 1991; Forlander and Bohannon, 1999; Hsieh et al., 2000; Antonucci et al., 2002) and the Medical Outcome Study Short Form 36 (SF36). (Vickrey et al., 1995; Brunet et al., 1996; Freeman et al., 2000; Hobart et al., 2001b). The 23 (sub-)scales covered 5 domains: 3 disease-specific measures, 10 physical functioning measures (5 mobility measures, 3 self-care measures and 2 upper limb function measures), 4 mental health measures (2 cognitive function measures and 2 emotional well-being measures), 5 social functioning measures and 1 general health measure. Of these, 11 outcome measures were questionnaires, 7 were (parts of) measures that required physical examination or testing procedures and 5 outcome measures were based on semi-structured interviews. When possible, outcome measures were transformed into a scale ranging from 100 (best) to 0 (worst). Scores on the NHPT, the 10-m TWT, the MSFC, and the SaGAS could not be transformed in this way, because these continuous scales do not have defined end-points for best or worst scores. Table 1 presents an overview of the outcome measures and the baseline scores (standard deviation).

**Analysis of responsiveness**

To determine whether a patient’s score had changed, we applied two external criteria: (i) a 7-point Likert-type patient rated global rating scale (GRS) of change, using the situation at diagnosis as reference point, (Jaeschke et al., 1989; Juniper et al., 1994; Liang, 1995; Stucki et al., 1995; Besette et al., 1998; Cella et al., 2002b; Guyatt et al., 2002) emphasizing the perspective of the patient, and

**Brain (2006), 129, 2648–2659**

2649
(ii) a change on the EDSS, representing the perspective of the clinician. The GRS question asked was: ‘How would you rate your current health when compared with your health at the time of diagnosis?’ The answering categories were: very much improved, much improved, slightly improved, stable, slightly deteriorated, much deteriorated, and very much deteriorated. The EDSS is a single-scale measure that ranges from \(0 = \) a normal neurological examination, to \(10 = \) death due to multiple sclerosis.

To assess the relative responsiveness, that is relatively independent of the method used to assess the responsiveness (Terwee et al., 2003) we calculated the area under the receiver operating characteristic (ROC) curve with its 95% confidence interval (AUC, 95% CI) for every outcome measure, using score changes since baseline at 3 years (Beurskens et al., 1996; Van der Windt et al., 1998; De Vet et al., 2001; Mancuso and Peterson, 2004). We used a non-parametric method which does not make any assumptions about the distributions to compute the AUC. Figure 1 shows an example of two ROC curves. The relative responsiveness was assessed separately for deterioration and improvement. For both external criteria the scores were dichotomized, using the category stable (no change) as reference category.

The minimally important change score of an outcome measure (MIC) is calculated as the mean change score in patients who showed a minimally important change according to an external criterion (Wyrwich et al., 1999). For the GRS of the patient’s perspective we used the categories of slightly improved or slightly deteriorated to identify the patients who showed a minimally important change.

### Table 1 Outcome measures studied and baseline scores of 156 multiple sclerosis patients

<table>
<thead>
<tr>
<th>Outcome measure</th>
<th>Sub-scale</th>
<th>Type</th>
<th>Transformed baseline score [0–100% (SD)]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease-specific</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDSS</td>
<td>Expanded Disability Status Scale</td>
<td>pt</td>
<td>74.9 (11.2)</td>
</tr>
<tr>
<td>MSFC</td>
<td>Multiple Sclerosis Functional Composite</td>
<td>pt</td>
<td>0.0 (0.7)*</td>
</tr>
<tr>
<td>SaGAS</td>
<td>Short and Graphic Assessment Scale</td>
<td>pt</td>
<td>7.0 (0.4)*</td>
</tr>
<tr>
<td><strong>Physical functioning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPmob</td>
<td>Disability and Impact Profile</td>
<td>Mobility</td>
<td>q</td>
</tr>
<tr>
<td>RAPmob</td>
<td>Rehabilitation Activities Profile</td>
<td>Mobility</td>
<td>i</td>
</tr>
<tr>
<td>RMI</td>
<td>Rivermead Mobility Index</td>
<td>q</td>
<td>95.7 (8.7)</td>
</tr>
<tr>
<td>SF36pf</td>
<td>Medical Outcome Study Short Form 36</td>
<td>Physical functioning</td>
<td>q</td>
</tr>
<tr>
<td>TVT</td>
<td>10-m timed-walk test</td>
<td>pt</td>
<td>6.4 (3.2) *</td>
</tr>
<tr>
<td><strong>Self-care</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPself</td>
<td>Disability and Impact Profile</td>
<td>Self-care</td>
<td>q</td>
</tr>
<tr>
<td>FIMmf</td>
<td>Functional Independence Measure</td>
<td>Motor function</td>
<td>i</td>
</tr>
<tr>
<td>RAPself</td>
<td>Rehabilitation Activities Profile</td>
<td>Self-care</td>
<td>i</td>
</tr>
<tr>
<td><strong>Upper limb function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARAT</td>
<td>Action Research Arm Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHPT</td>
<td>Nine-hole peg test</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mental health</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cognitive function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIMcf</td>
<td>Functional Independence Measure</td>
<td>Cognitive function</td>
<td>i</td>
</tr>
<tr>
<td>PASAT3</td>
<td>Paced serial addition test</td>
<td>3-second version</td>
<td>pt</td>
</tr>
<tr>
<td><strong>Emotional well-being</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPpsy</td>
<td>Disability and Impact Profile</td>
<td>Psychological</td>
<td>q</td>
</tr>
<tr>
<td>SF36mh</td>
<td>Medical Outcome Study Short Form 36</td>
<td>Mental health</td>
<td>q</td>
</tr>
<tr>
<td><strong>Social functioning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIPsoc</td>
<td>Disability and Impact Profile</td>
<td>Social functioning</td>
<td>q</td>
</tr>
<tr>
<td>RAPsoc</td>
<td>Rehabilitation Activities Profile</td>
<td>Occupation</td>
<td>i</td>
</tr>
<tr>
<td>SF36re</td>
<td>Medical Outcome Study Short Form 36</td>
<td>Role emotional</td>
<td>q</td>
</tr>
<tr>
<td>SF36rp</td>
<td>Medical Outcome Study Short Form 36</td>
<td>Role physical</td>
<td>q</td>
</tr>
<tr>
<td>SF36sf</td>
<td>Medical Outcome Study Short Form 36</td>
<td>Social functioning</td>
<td>q</td>
</tr>
<tr>
<td><strong>General health</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF36gh</td>
<td>Medical Outcome Study Short Form 36</td>
<td>General Health</td>
<td>q</td>
</tr>
</tbody>
</table>

pt = performance test; q = questionnaire; i = interview by professional.
*Not transformed into a 100 (best) to 0 (worst) scale.

![ROC curves](image.png)

**Fig. 1** ROC curves. In a ROC curve the sensitivity is plotted against 1–specificity. The AUC is a measure of the responsiveness of the outcome measure. An AUC = 0.5 (diagonal line) indicates that the outcome measure is not responsive. The more the ROC curve approaches the upper left corner the more responsive the outcome measure is.
important change. Figure 2 illustrates graphically where the MIC is located on the spectrum of change-scores. The next possible categories, namely much improved or much deteriorated, were not used, because they indicate substantial improvement or deterioration. For EDSS of the clinician’s perspective we used an improvement or deterioration of one point since baseline, because a change of one EDSS point is frequently used in trials and is the lowest EDSS change that can reliably be detected in the lower EDSS ranges (Noseworthy et al., 1990; Goodkin et al., 1992). The MIC was calculated from the patient’s perspective (MIC-P-improvement and MIC-P-deterioration), and the clinician’s perspective (MIC-C-improvement and MIC-C-deterioration). Because the longitudinal study design had five repeated measurements, we used generalized estimating equations (GEE) to estimate the MIC. This regression analysis technique for longitudinal data makes optimal use of the available data and reduces the standard error of the estimates, while at the same time correcting for the dependence between subsequent measurements (Zeger and Liang, 1986). The correlation structure was chosen on the basis of the correlation matrix of the outcome measures, and set at exchangeable (i.e. correlation coefficients between the first and successive measurements are progressively smaller). Scores on the outcome measures were used as dependent variable \( Y(t) \), and time \( t \), in years) and four dummy variables based on the external criteria (deteriorated, slightly deteriorated, slightly improved, improved) were used as independent variables. The stable group was used as reference. Because the GRS used the time of diagnosis as reference point, we used an autoregression formula that also includes the score for the outcome measure at baseline \( Y(t_0) \) as independent variable. In the formula:

\[
Y(t) = \alpha + \beta_1 \times Y(t_0) + \beta_2 \times t + \beta_3 \times \text{deteriorated} + \beta_4 \times \text{slightly deteriorated} + \beta_5 \times \text{slightly improved} + \beta_6 \times \text{improved}
\]

\( \beta_4 \) is interpreted as the mean score change on the outcome measure for patients who were slightly deteriorated, and provides an estimate for the MIC\(_{\text{deterioration}}\). \( \beta_5 \) is interpreted as the mean score change on the outcome measure for patients who were slightly improved, and provides an estimate for the MIC\(_{\text{improvement}}\).

To assess the reliability of two scores on each outcome measure, we used the smallest real change (SRC) (Pfenning et al., 1999b; Beckerman et al., 2001; De Vet et al., 2001). The SRC is more often referred to as the smallest real difference, but since our main focus

---

**A. MIC > SRC**

![Graph A](image1)

**B. MIC < SRC**

![Graph B](image2)

**Fig. 2** Relationship between SRC and MIC. (A) Shows the distribution of change scores for the categories (stable, slightly deteriorated and deteriorated) of the external criterion. There is minimal overlap between scores and the MIC is much larger than the SRC. This outcome measure is useful. (B) Shows again the distribution of change-scores for each category of the external criterion. There is much overlap between the scores and the MIC is smaller than the SRC. This outcome measure is not useful.
is on intra-individual changes, we prefer to use the term smallest real change. For each external criterion the SRC was calculated in the sub-group of patients who did not change, according to the external criterion during the first 6 months after inclusion. The SRC takes two sources of variability into account: (i) the reliability of the outcome measure, and (ii) the naturally occurring variability in stable patients. The SRC offers the opportunity to calculate a measure for comparisons at group level (SRCgroup) and at individual level (SRCindividual) (Pfennings et al., 1999b). The SRCindividual was calculated as 1.96 × SD of the score changes in stable patients. Figure 2 shows graphically where the SRC is located on the spectrum of change-scores. The SRCgroup was calculated as SRCindividual / n.

The selection of the most useful evaluative outcome measure was based on the relative responsiveness (highest AUC), whether the MIC > SRCindividual or SRCgroup, (see Fig. 2) and whether the results were comparable for both external criteria. For each outcome measure we calculated the sample sizes (patients per group) needed to show differences between independent samples in future studies. We used the formula 2 × [(Zα + Zβ) × (SRCgroup/ 1.96)]^2 (Guyatt et al., 1987), where α is set at 0.05 (Zα = 1.96) and β is set at 0.20 (Zβ = 0.84), in order to achieve a power of 0.80.

The statistical analyses were performed with SPSS version 11.5 for Windows. GEE analyses were performed with the Statistical Package for Interactive Data Analysis (SPIDA) version 6.05 from the Statistical Computing Laboratory.

**Results**

A total of 156 patients were included in the cohort between January 1998 and January 2001. Table 2 shows the baseline characteristics of these patients. Most characteristics comply with the expected pattern: more females than males in the relapsing–remitting group, more males than females in the primary progressive group, and more severe neurological deficits in the primary progressive group. Seven patients were lost to follow-up (three after 1 year, one after 2 years and three after 3 years), and 15 measurements were missing. The baseline scores on the outcome measure are presented in Table 1.

Table 3 shows the distribution of GRS and EDSS scores for each measurement. The distributions are remarkably different. The GRS scores are more equally spread across the categories, and according to the GRS fewer patients were stable, and more patients had improved. Over time there is a tendency for both external criteria to change towards deterioration. The percentage of patients that deteriorated (taking categories deteriorated and slightly deteriorated together) according to the patient’s and clinician’s perspective, respectively, is 36 and 22% at 6 months, 46 and 33% at 1, 50 and 46% at 2, and 60 and 44% at 3 years. The agreement between the patient’s and clinician’s perspective to classify patients as deteriorated, stable or improved is 36 and 22% at 6 months, 46 and 33% at 1, 50 and 46% at 2, and 60 and 44% at 3 years.

Tables 4 and 5 show that the AUCs range from 0.50 to 0.75 and have wide CIs. For five (patient’s perspective) and seven (clinician’s perspective) outcome measures the AUC does not significantly differ from 0.50. For a substantial number of outcome measures the MIC does not significantly differ from zero, which means that the MIC cannot be

<table>
<thead>
<tr>
<th>Table 2 Baseline characteristics of patients with multiple sclerosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>n (%)</td>
</tr>
<tr>
<td>Age (SD)</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Female (%)</td>
</tr>
<tr>
<td>Time since diagnosis (years)</td>
</tr>
<tr>
<td>Time since symptoms (years)</td>
</tr>
<tr>
<td>Number of exacerbations</td>
</tr>
<tr>
<td>EDSS</td>
</tr>
</tbody>
</table>

n (percentage), mean (SD) or median (IQR). RR = relapsing–remitting multiple sclerosis; SP = secondary progressive multiple sclerosis; PP = primary progressive multiple sclerosis. EDSS, original score.

<table>
<thead>
<tr>
<th>Table 3 Distribution (n, %) of the GRS (patient’s perspective) and EDSS (clinician’s perspective) based external criteria for each measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>External criteria</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>6 months 1 year 2 years 3 years 6 months 1 year 2 years 3 years</td>
</tr>
<tr>
<td>Deteriorated</td>
</tr>
<tr>
<td>Slightly deteriorated</td>
</tr>
<tr>
<td>Stable</td>
</tr>
<tr>
<td>Slightly improved</td>
</tr>
<tr>
<td>Improved</td>
</tr>
</tbody>
</table>

For deterioration and improvement the categories were ‘very much’ and ‘much’ have been combined.
detected beyond chance for these outcome measures in this population. It also means that these outcome measures are not suitable to evaluate change in this population. Furthermore, none of the outcome measures has an MIC$_{\text{deterioration}}$ which makes the outcome measures unsuitable to detect an minimally important change in an individual patient. However, several measures have an MIC$_{\text{deterioration}}$ which makes them suitable for research purposes. The final columns in the tables show a large variation in required sample sizes. The unrealistically high estimates of the sample sizes are caused by large estimates of the SRC$_{\text{individual}}$ relative to the estimate of the MIC.

The results for deterioration from the patient’s perspective can be found in Table 4. Of the disease-specific outcome measures, the EDSS has the highest AUC [0.70 (95% CI 0.62–0.79)] and an MIC$_{\text{deterioration}}$ (−8.58) that exceeds the SRC$_{\text{group}}$ (−4.38). Of the outcome measures related to mental health, the FIM sub-scale cognitive function (FIMcf) and the DIP sub-scale psychological (DIPpsy) have approximately the same AUCs [0.65 (95% CI 0.55–0.74) and 0.64 (95% CI 0.55–0.73), respectively]. For the DIPpsy the MIC$_{\text{deterioration}}$ (−2.88) exceeds the SRC$_{\text{group}}$ (−2.80), but for the FIMcf the MIC$_{\text{deterioration}}$ (−1.47) is smaller than the SRC$_{\text{group}}$ (−1.66). Of the outcome measures related to social functioning, the RAP sub-scale occupation (RAPocc) has the highest AUC [0.73 (95% CI 0.64–0.81)] and an MIC$_{\text{deterioration}}$ (−7.74) exceeding the SRC$_{\text{group}}$ (−4.24).

Table 5 shows the results for deterioration from the clinician’s perspective. Because information from the EDSS is used to obtain the external criterion, results for the EDSS cannot be calculated. The two disease-specific outcome measures...
measures have a very similar AUC [0.72 (95% CI 0.63–0.81)] for the SaGAS and 0.71 (95% CI 0.62–0.80) for the MSFC], and for both the MIC-C_{deterioration} was small and did not significantly differ from zero. Of the outcome measures related to physical functioning, SF36pf has the highest AUC [0.72 (95% CI 0.63–0.80)] and an MIC-C_{deterioration} (−8.52) that amply exceeds the SRC_{group} (−2.81). Of the outcome measures related to mental health, the DIPpsy and the PASAT3 (test 3-second version) have an AUC of 0.60 (95% CI = 0.50–0.70 and 0.50–0.69, respectively). For both outcome measures the MIC-C_{deterioration} is small and does not significantly differ from zero. Of the outcome measures related to social functioning, the RAPocc has the highest AUC [0.69 (95% CI 0.61–0.78)] and an MIC-C_{deterioration} (−8.40) that amply exceeds the SRC_{group} (−2.69).

Regardless of the domain of the outcome measures, the five most responsive (AUC) outcome measures to detect deterioration from the patient’s perspective are the SF36pf [0.75 (0.67–0.84)], the DIP sub-scale mobility [DIPmob; 0.73 (0.65–0.82)], the RAPocc [0.73 (0.64–0.81)], the DIP sub-scale self-care [DIPSelf; 0.70 (0.62–0.79)] and the EDSS [0.70 (0.62–0.79)]. Of these, only the EDSS does not fulfil the criterion MIC-P_{deterioration} > SRC_{group}. The five most responsive outcome measures to detect deterioration (AUC) from the clinician’s perspective are the SaGAS [0.72 (0.63–0.81)], the SF36pf [0.75 (0.67–0.84)], the DIP sub-scale mobility [DIPmob; 0.73 (0.65–0.82)], the RAPocc [0.73 (0.64–0.81)], the DIP sub-scale self-care [DIPSelf; 0.70 (0.62–0.79)] and the EDSS [0.70 (0.62–0.79)]. Of these, only the EDSS does not fulfil the criterion MIC-P_{deterioration} > SRC_{group}.

The results for improvement are less clear, because of the small percentage of patients in the slightly improved groups (data not shown). The MIC was either very small or did not significantly differ from zero. Therefore, it was not possible to compare the results with the SRC. Consequently, we can
only look at the relative responsiveness by comparing the AUCs. From the patient’s perspective, the highest AUCs were found for the EDSS [0.78 (95% CI 0.70–0.87)], the DIPmob [0.73 (95% CI 0.64–0.85)], the FIM sub-scale motor function [FIMmf; 0.71 (0.63–0.80)], the SF36pf [0.71 (95% CI 0.62–0.80)] and the RAPOcc [0.71 (95% CI 0.62–0.82)]. From the clinician’s perspective, the highest AUCs were found for the RAPOcc [0.79 (95% CI 0.63–0.95)], the SF36pf [0.77 (95% CI 0.64–0.90)], the FIMmf [0.74 (95% CI 0.62–0.86)], the FIMcf [0.74 (95% CI 0.59–0.90)] and the RAPmob [0.72 (95% CI 0.58–0.87)]. Irrespective of the external criterion that is applied, the most responsive outcome measures to detect improvement are the FIMmf, the SF36pf, the RAPOcc and the EDSS. However, the criterion MIC > SRC could not be evaluated for any of the measures.

**Discussion**

In the early stages of multiple sclerosis, the two most useful evaluative outcome measures to detect deterioration, and that perform well irrespective of the external criterion that is applied, are the SF36pf for the physical functioning domain (mobility), and the RAPOcc for the social functioning domain. Both measures have an MIC > SRCgroup which makes them suitable for application in clinical research. However, none of the outcome measures that we studied had an MIC > SRCindividual which means that the reliability demands that warrant application at individual patient level are not met.

The selection of an outcome measure is not only guided by its responsiveness. It is also important to select an outcome measure that really measures the phenomena of interest. Therefore, we categorized the outcome measures that we have studied into five domains and five sub-domains, which should guide their selection. Before the final selection of an outcome measure, one should study the content of an outcome measure to make sure it measures the variable one is interested in. The measures that perform best in the other domains are the DIPpsy (mental health domain, emotional well-being) and the SF36gh (general health domain), but none of the disease-specific outcome measures fulfilled our selection criteria.

We were looking for an outcome measure with a performance that did not depend on the required perspective. Finding such an outcome measure would increase our confidence in this measure, because it would imply that the results obtained with this measure have the same meaning for both the clinician and the patient. However, it might be very legitimate to emphasize one or both perspectives depending on the research aim. For more basic research purposes reliance on examiner-driven outcomes might be fully acceptable. But for more clinically oriented research questions, i.e. studies that are interested in the effects on patients, such as clinical trials, reliance on examiner-driven assessments only is not sufficient. In these studies one should also include patient-driven outcome measures, because that is the only way to show benefit for patients. For the evaluation of this kind of clinically oriented research it would be very valuable to have a (primary) outcome measure available which evaluative ability is independent of the chosen perspective (patient versus examiner), because only then the MIC is the same for the patient and the examiner, which facilitates the interpretation of this research.

An important strength of this study is the simultaneous evaluation of several outcome measures that are frequently used in multiple sclerosis research. Scores were collected for 23 (sub-scales of) outcome measures in the same patients and in the same way. This enables a direct comparison of the results. Information about the responsiveness of outcome measures is often derived from several studies with different designs, different populations, different anchors, and different outcome measures. This hampers the selection of the most responsive outcome measure, because no direct comparison can be made.

The relative responsiveness is quite independent of the particular approach to the evaluation of responsiveness (Terwee et al., 2003). We chose the approach presented in this article for two reasons. First of all, we aimed to identify the most responsive outcome measures by comparing the outcome measures on the basis of the AUC (relative responsiveness). Second, we tried to obtain data that would facilitate the interpretation of score changes in future studies. The interpretation depends on two aspects of the score change: (i) what is a minimally important change, and (ii) is the instrument capable of measuring this change? We have used the MIC as a measure of minimally important change, and the SRC to estimate the ability of a measure to detect this change. From our results we conclude that our strategy worked well for the analysis of changes in the direction of deterioration, because we were able to clearly show the relative responsiveness, and provide clear data that facilitate the interpretation of score changes. However, the results with regard to changes in the direction of improvement are inconclusive, due to the small number of patients in this category.

Another aspect of this study that deserves some attention is the analysis of repeated measures. We made optimal use of the longitudinal data by applying longitudinal data-analysis techniques, which reduces the standard error of our estimates. Moreover, we constructed a regression model that enabled us to estimate the MIC for deterioration and improvement in one model. The possibility of this study to show improvement is limited by its design, because recruiting recently diagnosed patients, who are only mildly disabled, implies a limitation in the possibility to improve. Therefore, our results for improvement are not as clear as those for deterioration. However, despite this limitation, the study does provide some preliminary evidence that the MICdeterioration and the MICimprovement are not necessarily equal (Cella et al., 2002b).
A well-known problem in studies of anchor-based responsiveness is the choice of the external criterion to define change (Cella et al., 2002a). Norman et al. (1997) compared two methods to assess responsiveness with each other: (i) an effective therapy as construct for change, and (ii) a retrospective method to assess change using a GRS. In this direct comparison the GRS performs worse than the effective therapy as external criterion. The problem with the generalization of these results is that there is often not an effective therapy available. Particularly in longitudinal cohort studies, such as ours, we cannot rely on an effective therapy. There are ways to use effective therapy as construct for change in multiple sclerosis by applying outcome measures in patients that were treated for a relapse with corticosteroids. A major problem in these studies is that one is looking at improvements. It is absolutely not certain that these results can subsequently be used in studies that look at deterioration.

Because a gold standard for change is lacking, we had to rely on other methods to define change. We decided not to rely on one method, because the chosen method to define change influences the results of the analyses. Furthermore, we carefully sought for sensible external criteria. Roughly speaking, there are three constructs for the evaluation of change in multiple sclerosis: data obtained from repeated MRI studies, the EDSS as the most frequently used clinical outcome measure, and a GRS which emphasizes the perspective of the patient. Our main focus in this study was on disability and quality of life. Therefore, using MRI data as a construct for change is not appealing, since it only offers information at the level of pathological changes, which are only remotely related to disability and even less related to quality of life. The EDSS has limitations with regard to its validity and reliability, which might make it relatively unsuitable as an external criterion for change. However, despite this criticism, it is a scale that is very well known among clinicians. It is, in fact, so well-known that a description of a study population is not complete without EDSS data. Therefore, we used the EDSS to determine important change from a clinician’s point of view. Because the first question of a clinician during a visit often is a global rating: ‘How are you doing since the last visit’, and because the first question of a clinician during a visit often is a global rating: ‘How are you doing since the last visit’, and because the latter two categories are often used by patients who are reluctant to classify themselves as stable, while their situation would justify this classification (Ostelo and De Vet, 2005). We performed a sensitivity analysis (data not shown), with the category ‘much deteriorated’ as cut-off, and compared the MIC-P and the MIC-P estimates obtained in this sensitivity analysis (MIC-Psens) with the MIC-C. For 17 outcome measures the MIC-P was closer to the MIC-C than the MIC-Psens indicating that there is a greater correspondence between the MIC-P and the MIC-C than between the MIC-Psens and the MIC-C, which supports the use of the category ‘slightly deteriorated’ as cut-off in this sample. In future studies it might be useful to add extra categories to the GRS between ‘slightly’ and ‘much’, for example by using ‘deteriorated’ and ‘improved’ on their own, and to use these categories to determine the MIC. This might lessen the (semantic) gap between ‘slightly’ and ‘much’, and might aid patients who are reluctant to use the category ‘stable’, without influencing the estimation of the MIC.

Recently, Solari et al. (2005) studied the practice effects of the MSFC and suggested that, to improve efficiency, one prebaseline administration of TWT, three of PASAT and four of NHPT are needed. Their study consisted of repeated administrations of the tests in 1 day. What their results mean for repeated MSFC measurements with intervals of 6 months or longer, such as our study, is not immediately clear. Will you never lose your ability to perform the PASAT or NHPT once you have mastered it, or do you again need some prebaseline administrations after you have not been performing the PASAT or NHPT for some time? For the components of the MSFC and the SaGAS we used the same test protocol at each measurement. The NHPT and the TWT were conducted twice. For the TWT this is sufficient, for the
NHPT two additional administrations would have been better. The PASAT was always administered once, but in any case after at least one practice trial, as described in the MSFC manual. Although the interval between subsequent measurements was at least 6 months, we cannot rule out a practice effect. Ignoring a possibly present practice effect will lead to inflated measures of responsiveness in the direction of deterioration for the NHPT and PASAT, because the measured change in cognitive or upper limb function is smaller than the real change. The opposite would occur for the measures of responsiveness in the direction of improvement, because the measured improvement in cognitive function is larger than the real improvement.

Although we were able to identify the most responsive outcome measures and to show, for several of these outcome measures, that the signal (MIC) exceeds the noise (SRC\text{group}), it should be noted that our results are not automatically applicable to all patients with multiple sclerosis. In general, our population was only mildly disabled, had a disease duration of just over 3 years at the end of the study, and was treated with disease modifying treatment if indicated (44 patients were on disease modifying treatment at the end of the study). Because this treatment will influence the outcomes and the external criteria in the same direction, it will probably not significantly alter our results. The results of this study can therefore be used in early intervention studies. With the positive effects of disease modifying treatments, patients will be mildly disabled for a longer period. Future trials will have to compare newly developed treatments with the current disease modifying treatments. Showing differences in effectiveness in these studies will increasingly suffer from power problems. In comparative studies an outcome measure should be able to show differences between longitudinal changes of two (or more) groups (arms of a trial), which is probably more difficult than showing changes within one group only. In our opinion this is a requirement that can only be fulfilled when an outcome measure is already capable of detecting longitudinal changes. Our results clearly show that some of the outcome measures that we have studied, and that are not regularly used in trials, are more suitable to evaluate changes than others. In the early stages of multiple sclerosis a reduction of the walking distance is more often a problem than a reduction in walking speed. The SF36pf probably performs well because it also contains items about walking distance, whereas the regularly used TWT only measures walking speed. The RAPocc and, to a lesser extent, the DIPsoc, probably perform well because they measure social functioning. Although social functioning is seriously affected in the early stage of multiple sclerosis, it is not part of the measures that are regularly used in trials. Future responsiveness studies should focus on more severely disabled populations and populations with a longer duration of the disease.

None of the outcome measures used in this study could detect important change in individual patients. Outcome measures that might be useful should have a relatively low SRC\text{individual}. This point has already been acknowledged in relation to the MSFC. Several authors have stated that a change of 20% for the components of the MSFC is required to exceed measurement error (Kaufman et al., 2000; Schwid et al., 2002) and that changes for the MSFC and SаGАS should be >0.5 (Hoogervorst et al., 2004; Vaney et al., 2004). Depending on the external criterion used, we found that in our sample a change of 2.6–3.0 s (40% of baseline) for the TWT and 2.8–5.3 s (13% of baseline) for the NHPT is required to exceed measurement error. In our sample, changes in MSFC and SаGАS should exceed 0.54–0.72 and 0.25–0.44, respectively, in order to indicate significant change. However, MSFC scores should be interpreted with caution, because it is not evident from the total score which component contributes most to the total score. The differences between results reported in the literature (Kaufman et al., 2000; Schwid et al., 2002; Hoogervorst et al., 2004; Vaney et al., 2004) and our results might be explained by our study design. We recruited recently diagnosed patients, whereas in the other studies the patients had the disease for various lengths of time. Furthermore, we used a fixed interval of 6 months between visits to identify the stable patients, whereas the other studies used a 5-day or a variable interval. The design of the present study matches usual patient care, which increases the validity of our results, but, unfortunately, leads to the conclusion that the outcome measures in this study are not suitable for detecting change within a few years in individual, recently diagnosed, patients.

**Acknowledgements**

The Netherlands Organization for Scientific Research (NWO 940-33-009) supported this study. It has been performed on behalf of the Functional Prognostication and Disability (FuPro) Study Group: G.J.L., J. Dekker, A. J. Dallmeijer, M. J. IJzerman, H.B., V.d.G.: VU University Medical Center Amsterdam (project co-ordination); A. J. H. Prevo, E. Lindeman, V. P. M. Schepers: University Medical Center, Utrecht; H. J. Stam, E. Odding, B. van Baalen: Erasmus Medical Center, Rotterdam; A. Beelen, I. J. M. de Groot: Academic Medical Center, Amsterdam. We would like to thank the neurologists in the participating hospitals (VU University Medical Center, Academic Medical Center Amsterdam, Sint Lucas Andreas Hospital Amsterdam, OLVG Hospital Amsterdam, Erasmus Medical Center Rotterdam) for recruiting the patients, and M. van der Bruggen, M. Schothorst, and T. Wedding for performing the measurements.

**Contribution of authors**


Acquisition of data: V.deG., H.B., B.M.J.U., C.H.P.

Drafting of the manuscript: V.deG., H.B.

Conflict of interest

There are no conflicts of interest. The corresponding author (V.deG.) had full access to all the data used in the study, and had the final responsibility for the decision to submit the manuscript for publication.

References


Evaluative outcome measures in multiple sclerosis


Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. J Clin Epidemiol 2004; 57: 1008–18.


