The usefulness of evaluative outcome measures in patients with multiple sclerosis

V. de Groot,[1,4] H. Beckerman,[1,4] B. M. J. Uitdehaag,[2,3] H. C. W. de Vet,[4] G. J. Lankhorst,[1,4]
C. H. Polman[2] and L. M. Bouter[4]

Departments of [1]Rehabilitation Medicine, [2]Neurology, [3]Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands and [4]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

Received February 10, 2006. Revised July 21, 2006. Accepted July 25, 2006. Advance Access publication September 6, 2006.

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.*, 2001*a*) 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.*, 1999*a*; Van der Putten *et al.*, 1999; Freeman *et al.*, 2000; Hobart *et al.*, 2001*b*).

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.*, 2001*a*; 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.*, 2002*a*, *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.*, 2002*a*; 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.*, 2002*a*). 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.*, 2001*b*), 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.*, 1999*a*), the Functional Independence Measure (FIM) (Granger *et al.*, 1990; Kidd *et al.*, 1995; Marolf *et al.*, 1996), the Rehabilitation Activities Profile (RAP) (Van Bennekom *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.*, 2001*b*). 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; Bessette *et al.*, 1998; Cella *et al.*, 2002*b*; Guyatt *et al.*, 2002) emphasizing the perspective of the patient, and

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

Outcome measure		Sub-scale	Type	Transformed baseline score [0–100% (SD)]
Disease-specific				
EDSS	Expanded Disability Status Scale		pt	74.9 (11.2)
MSFC	Multiple Sclerosis Functional Composite		pt	0.0 (0.7)*
SaGAS	Short and Graphic Assessment Scale		pt	7.0 (0.4)*
Physical functioning				
Mobility				
DIPmob	Disability and Impact Profile	Mobility	q	86.9 (10.5)
RAPmob	Rehabilitation Activities Profile	Mobility	i	85.7 (14.1)
RMI	Rivermead Mobility Index		q	95.7 (8.7)
SF36pf	Medical Outcome Study Short Form 36	Physical functioning	q	71.3 (23.5)
TWT	10-m timed-walk test		pt	6.4 (3.2) s*
Self-care				
DIPself	Disability and Impact Profile	Self-care	q	94.3 (8.6)
FIMmf	Functional Independence Measure	Motor function	i	95.2 (5.4)
RAPself	Rehabilitation Activities Profile	Self-care	i	92.3 (11.1)
Upper limb function				
ARAT	Action Research Arm Test		pt	99.1 (4.0)
NHPT	Nine-hole peg test		pt	21.1 (4.0) s*
Mental health				
Cognitive function				
FIMcf	Functional Independence Measure	Cognitive function	i	95.2 (5.2)
PASAT3	Paced serial addition test	3-second version	pt	76.9 (18.3)*
Emotional well-being				
DIPpsy	Disability and Impact Profile	Psychological	q	79.4 (12.3)
SF36mh	Medical Outcome Study Short Form 36	Mental health	q	72.1 (17.5)
Social functioning				
DIPsoc	Disability and Impact Profile	Social functioning	q	87.0 (10.2)
RAPocc	Rehabilitation Activities Profile	Occupation	i	75.0 (20.6)
SF36re	Medical Outcome Study Short Form 36	Role emotional	q	74.1 (37.0)
SF36rp	Medical Outcome Study Short Form 36	Role physical	q	51.9 (42.0)
SF36sf	Medical Outcome Study Short Form 36	Social functioning	q	77.9 (23.2)
General health				
SF36gh	Medical Outcome Study Short Form 36	General Health	q	52.6 (19.8)

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

(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

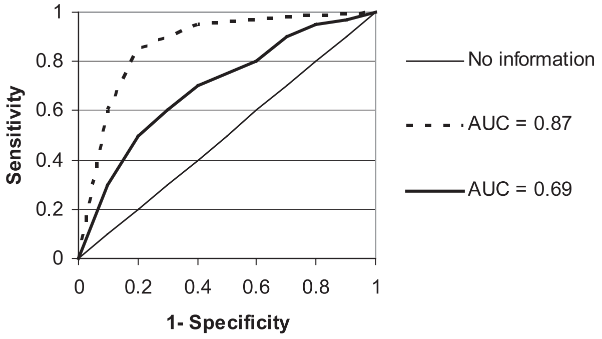

Fig. 1 ROC curves. In a ROC curve the sensitivity is plotted against 1–specificity. The AUC is a measure of the responsiveness of the outcome measure. An AUC ⩽0.5 (diagonal line) indicates that the outcome measure is not responsive. The more the ROC curve approaches the upper left corner the more responsive the outcome measure is.

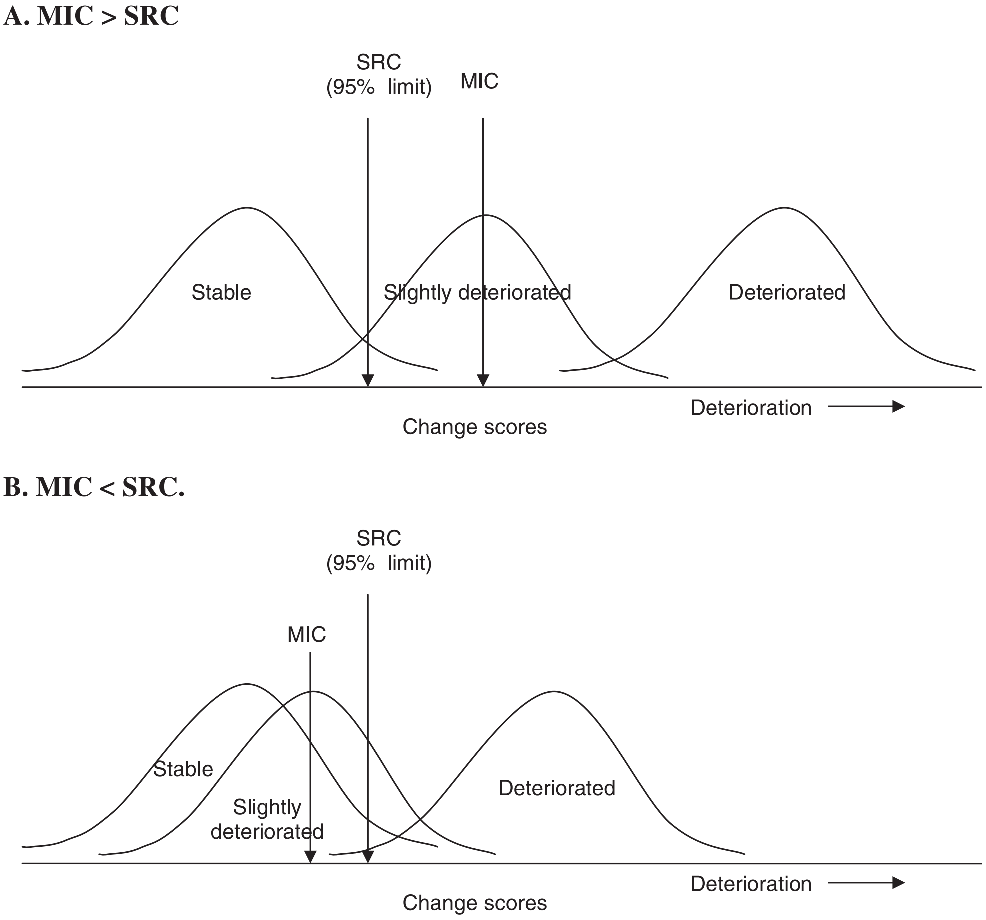

Fig. 2 Relationship between SRC and MIC. (**A**) Shows the distribution of change scores for the categories (stable, slightly deteriorated and deteriorated) of the external criterion. There is minimal overlap between scores and the MIC is much larger than the SRC. This outcome measure is useful. (**B**) 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.

important change. Figure 2 illustrates graphically were 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-$P_{\text{improvement}}$ and MIC-$P_{\text{deterioration}}$), and the clinician's perspective (MIC-$C_{\text{improvement}}$ and MIC-$C_{\text{deterioration}}$). 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 approximately equal) for all outcomes except the cognitive sub-scale of the FIM that was set at 4-dependence (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:

$$\begin{aligned}Y(t) = \alpha + \beta_1 * Y(t_0) + \beta_2 \times t + \beta_3 \times \text{deteriorated}\\ + \beta_4 \times \text{slightly deteriorated} + \beta_5 \times \text{slightly improved}\\ + \beta_6 \times \text{improved}\end{aligned}$$

β_4 is interpreted as the mean score change on the outcome measure for patients who were slightly deteriorated, and provides an estimate for the $\text{MIC}_{\text{deterioration}}$. β_5 is interpreted as the mean score change on the outcome measure for patients who were slightly improved, and provides an estimate for the $\text{MIC}_{\text{improvement}}$.

To assess the reliability of two scores on each outcome measure, we used the smallest real change (SRC) (Pfennings *et al.*, 1999*b*; 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 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 ($\mathrm{SRC}_{\mathrm{group}}$) and at individual level ($\mathrm{SRC}_{\mathrm{individual}}$) (Pfennings *et al.*, 1999*b*). The $\mathrm{SRC}_{\mathrm{individual}}$ 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 $\mathrm{SRC}_{\mathrm{group}}$ was calculated as $\mathrm{SRC}_{\mathrm{individual}} /\sqrt{n}$.

The selection of the most useful evaluative outcome measure was based on the relative responsiveness (highest AUC), whether the MIC > $\mathrm{SRC}_{\mathrm{individual}}$ or $\mathrm{SRC}_{\mathrm{group}}$, (*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 $2 \times \{[(Z_{\alpha} + Z_{\beta}) \times (\mathrm{SRC}_{\mathrm{group}}/1.96)]/\mathrm{MIC}\}^2$ (Guyatt *et al.*, 1987), where α is set at 0.05 ($Z_{\alpha} = 1.96$) and β is set at 0.20 ($Z_{\beta} = 0.84$), 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 35% ($\kappa = 0.10$) at 6 months, 42% ($\kappa = 0.14$) at 1, 40% ($\kappa = 0.07$) at 2, and 45% ($\kappa = 0.13$) 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 2 Baseline characteristics of patients with multiple sclerosis

Characteristics	RR	SP	PP	Not yet known	Total
n (%)	120 (77)	8 (5)	25 (16)	3 (2)	156 (100)
Age (SD)	35.5 (8.9)	48.2 (6.7)	43.2 (8.9)	45.5 (6.9)	37.6 (9.5)
Gender					
Female (%)	84 (70.0)	3 (37.5)	11 (44)	3 (100)	101 (64)
Time since diagnosis (years)	0.26 (0.15–0.41)	0.33 (0.24–0.48)	0.28 (0.15–0.33)	0.14 (0.14–0.17)	0.26 (0.15–0.40)
Time since symptoms (years)	1.83 (0.67–4.40)	7.50 (3.35–14.51)	2.10 (1.07–3.15)	3.62 (3.53–4.63)	2.15 (0.79–4.36)
Number of exacerbations	2.0 (1.0–3.0)	2.0 (1.0–7.0)	0 (0–0)	0 (0–0)	2.0 (1.0–3.0)
EDSS	2.0 (2.0–3.0)	3.0 (2.5–3.9)	3.0 (2.5–4.0)	2.5 (2.0–4.0)	2.5 (2.0–3.0)

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 3 Distribution (*n*, %) of the GRS (patient's perspective) and EDSS (clinician's perspective) based external criteria for each measurement

External criteria	Patient's perspective (%)				Clinician's perspective (%)			
	6 months (*n* = 113)	1 year (*n* = 130)	2 years (*n* = 141)	3 years (*n* = 145)	6 months (*n* = 153)	1 year (*n* = 47)	2 years (*n* = 145)	3 years (*n* = 146)
Deteriorated	11 (10)	15 (12)	19 (13)	28 (19)	12 (8)	24 (16)	40 (28)	41 (28)
Slightly deteriorated	29 (26)	44 (34)	52 (37)	60 (41)	21 (14)	25 (17)	26 (18)	24 (16)
Stable	26 (23)	30 (23)	29 (21)	22 (15)	100 (65)	79 (54)	66 (46)	69 (47)
Slightly improved	14 (12)	11 (8)	19 (13)	10 (7)	11 (7)	11 (7)	11 (8)	8 (5)
Improved	33 (29)	30 (23)	22 (16)	25 (17)	9 (6)	8 (5)	2 (1)	4 (3)

For deterioration and improvement the categories were 'very much' and 'much' have been combined.

Table 4 AUC, MIC-P and SRC for deterioration using the patient's perspective as external criterion

Outcome measure	AUC		$\mathrm{MIC}_{\mathrm{deterioration}}$	$\mathrm{SRC}_{\mathrm{individual}}$	$\mathrm{SRC}_{\mathrm{group}}$	Sample size*
	AUC	95% CI				
Disease-specific						
EDSS	0.70	0.62–0.79	−1.50 (ns)	−16.04	−3.15	467
MSFC**	0.62	0.53–0.72	−0.05 (ns)	−0.54	−0.11	476
SaGAS**	0.65	0.56–0.75	−0.05 (ns)	−0.25	−0.05	102
Physical functioning						
Mobility						
DIPmob	0.73	0.65–0.82	−4.25	−8.99	−1.80	18
RAPmob	0.66	0.57–0.76	−3.42	−19.88	−3.90	138
RMI	0.67	0.58–0.76	−0.88 (ns)	−5.91	−1.16	184
SF36pf	0.75	0.67–0.84	−8.58	−21.91	−4.38	27
TWT**	0.65	0.56–0.74	1.15 (ns)	2.56	0.50	20
Self-care						
DIPself	0.70	0.62–0.79	−2.11	−9.54	−1.91	83
FIMmf	0.68	0.59–0.76	−1.45	−5.74	−1.13	64
RAPself	0.65	0.56–0.74	−2.41	−11.96	−2.35	101
Upper limb function						
ARAT	0.53	0.43–0.63	−0.06 (ns)	−1.61	−0.32	2939
NHPT**	0.59	0.49–0.69	0.30 (ns)	2.82	0.55	361
Mental health						
Cognitive function						
FIMcf	0.65	0.55–0.74	−1.47	−8.47	−1.66	136
PASAT3	0.50	0.40–0.60	2.56	19.62	4.18	240
Emotional well-being						
DIPpsy	0.64	0.55–0.73	−2.88	−14.01	−2.80	97
SF36mh	0.56	0.46–0.66	−4.45	−28.13	−5.63	163
Social functioning						
DIPsoc	0.68	0.59–0.77	−2.84	−8.08	−1.62	33
RAPocc	0.73	0.64–0.81	−7.74	−21.63	−4.24	32
SF36re	0.50	0.40–0.59	−8.13	−67.26	−13.45	279
SF36rp	0.60	0.51–0.69	−21.69	−92.24	−18.45	74
SF36sf	0.68	0.59–0.77	−11.15	−41.17	−8.23	56
General health						
SF36gh	0.66	0.57–0.75	−9.86	−26.61	−5.32	30

AUC = area under the ROC curve at 3 years after baseline with 95% CIs; $\mathrm{MIC}_{\mathrm{deterioration}}$ = minimally important change; $\mathrm{SRC}_{\mathrm{individual}}$ = smallest real change at individual level; $\mathrm{SRC}_{\mathrm{group}}$ = smallest real change at group level, based on 26 stable patients at 6 months; ns = not significantly different from 0.
*Patients per group, calculation based on a significance level of 0.05 and a power of 0.8.
**MSFC, SaGAS, NHPT and TWT data not transformed into a 0–100 point scale.

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 > $\mathrm{SRC}_{\mathrm{individual}}$, which makes the outcome measures unsuitable to detect an minimally important change in an individual patient. However, several measures have an MIC > $\mathrm{SRC}_{\mathrm{group}}$, 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 $\mathrm{SRC}_{\mathrm{individual}}$ 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)]. For all three disease-specific outcome measures the $\mathrm{MIC\text{-}P}_{\mathrm{deterioration}}$ is small, and does not significantly differ from zero. Of the outcome measures related to physical functioning, the SF36pf has the highest AUC [0.75 (95% CI 0.67–0.84)] and an $\mathrm{MIC\text{-}P}_{\mathrm{deterioration}}$ (−8.58) that exceeds the $\mathrm{SRC}_{\mathrm{group}}$ (−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 $\mathrm{MIC\text{-}P}_{\mathrm{deterioration}}$ (−2.88) exceeds the $\mathrm{SRC}_{\mathrm{group}}$ (−2.80), but for the FIMcf the $\mathrm{MIC\text{-}P}_{\mathrm{deterioration}}$ (−1.47) is smaller than the $\mathrm{SRC}_{\mathrm{group}}$ (−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 $\mathrm{MIC\text{-}P}_{\mathrm{deterioration}}$ (−7.74) exceeding the $\mathrm{SRC}_{\mathrm{group}}$ (−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

Table 5 AUC, MIC-C, and SRC for deterioration using the clinician's perspective as external criterion

Outcome measure	AUC		$MIC_{deterioration}$	$SRC_{individual}$	SRC_{group}	Sample size*
	AUC	95% CI				
Disease-specific						
EDSS						
MSFC**	0.71	0.62–0.80	0.08 (ns)	−0.72	−0.08	331
SaGAS**	0.72	0.63–0.81	−0.06 (ns)	−0.44	−0.04	220
Physical functioning						
Mobility						
DIPmob	0.66	0.57–0.75	−2.56	−10.52	−1.06	69
RAPmob	0.67	0.58–0.76	−5.62	−16.26	−1.63	34
RMI	0.65	0.56–0.75	−1.30	−7.53	−0.75	137
SF36pf	0.72	0.63–0.80	−8.52	−27.99	−2.81	44
TWT**	0.69	0.59–0.78	0.34 (ns)	3.03	0.30	324
Self care						
DIPself	0.65	0.55–0.74	−2.16	−8.70	−0.87	66
FIMmf	0.68	0.59–0.77	−1.70	−6.43	−0.64	58
RAPself	0.62	0.52–0.72	−1.33 (ns)	−14.79	−1.48	505
Upper limb function						
ARAT	0.55	0.45–0.65	−0.14 (ns)	−5.27	−0.53	5784
NHPT**	0.67	0.58–0.76	0.51 (ns)	5.32	0.53	444
Mental health						
Cognitive function						
FIMcf	0.54	0.44–0.64	−1.41	−6.26	−0.63	80
PASAT3	0.60	0.50–0.69	−0.77 (ns)	26.45	2.77	4816
Emotional well-being						
DIPpsy	0.60	0.50–0.70	−1.11 (ns)	−16.68	−1.68	922
SF36mh	0.55	0.45–0.65	−2.48 (ns)	−28.44	−2.86	537
Social functioning						
DIPsoc	0.64	0.55–0.74	−2.16	−10.27	−1.03	92
RAPocc	0.69	0.61–0.78	−8.40	−26.89	−2.69	42
SF36re	0.53	0.43–0.63	−4.79 (ns)	−74.24	−7.46	980
SF36rp	0.61	0.51–0.71	−12.29	−89.05	−8.95	214
SF36sf	0.60	0.51–0.70	−5.04	−40.71	−4.09	266
General health						
SF36gh	0.51	0.42–0.61	−3.15 (ns)	−30.71	−3.09	388

AUC = area under the ROC curve at 3 years after baseline with 95% CIs; $MIC_{deterioration}$ = minimally clinically important change; $SRC_{individual}$ = smallest real change at individual level; SRC_{group} = smallest real change at group level, based on 100 stable patients at 6 months; ns = not significantly different from 0.
*Patients per group, calculation based on a significance level of 0.05 and a power of 0.8.
**MSFC, SaGAS, NHPT and TWT data not transformed into a 0–100 point scale.

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 [0.72 (0.63–0.80)], the MSFC [0.71 (0.62–0.80)], the RAPocc [0.69 (0.61–0.78)] and the TWT [0.69 (0.59–0.78)]. Of these, only the SF36pf and the RAPocc have an MIC-$C_{deterioration}$ > 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 > $\text{SRC}_{\text{group}}$, which makes them suitable for application in clinical research. However, none of the outcome measures that we studied had an MIC > $\text{SRC}_{\text{individual}}$, 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 outcome measures, and facilitates interpretation 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 $\text{MIC}_{\text{deterioration}}$ and the $\text{MIC}_{\text{improvement}}$ are not necessarily equal (Cella *et al.*, 2002*b*).

A well-known problem in studies of anchor-based responsiveness is the choice of the external criterion to define change (Cella *et al.*, 2002*a*). 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 a stronger external criterion is lacking, we used a GRS to emphasize the perspective of the patient. Because all outcomes were compared with these two sensible external criteria, we made insightful what the effect of the external criteria is.

A global rating requires that patients are able to mentally subtract a previous situation from the present situation (Liang, 1995; Stratford *et al.*, 1996). Criticism about the use of a GRS concerns the fact that this rating has often been found to show stronger associations with the present situation than with the previous situation (Guyatt *et al.*, 2002). In an attempt to overcome this problem, we coupled the previous situation to an important life-event for the patient. In this way, we tried to facilitate the mental subtraction, and hoped for more equal associations of the GRS with the previous and the present situation. We considered the time of diagnosis as an important life-event. Because in our study patients were not diagnosed until some time after their exacerbation and because the mean time between diagnosis and first measurement is relatively short (3.5 months), we decided that it was valid to use it as reference point. Our strategy was partly successful. The mean correlation coefficient between the GRS at 3 years and the outcome measures at baseline was 0.26 (range 0.15–0.43), at 6 months it was 0.30 (range 0.14–0.44), at 1 year it was 0.33 (range 0.14–0.49), at 2 years it was 0.37 (range 0.09–0.56), and at 3 years it was 0.40 (range 0.14–0.59).

Another point of discussion about the use of the GRS as external criterion is the choice of the cut-off point used for the calculation of the MIC. We decided to use the category 'slightly deteriorated' or 'slightly improved' as indicator of minimally important change. In our opinion, the next category ('much deteriorated' or 'much improved') is, at least semantically, not equivalent to minimally important change. Others have argued that using 'much deteriorated' or 'much improved' is more appropriate than 'slightly deteriorated' or 'slightly improved', 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-P_{sens}) with the MIC-C. For 17 outcome measures the MIC-P was closer to the MIC-C than the MIC-P_{sens}, indicating that there is a greater correspondence between the MIC-P and the MIC-C than between the MIC-P_{sens} 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 ($\text{SRC}_{\text{group}}$), 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 $\text{SRC}_{\text{individual}}$. 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 SaGAS 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 SaGAS 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

Concept and design: V.deG., H.B., B.M.J.U., H.C.W.deV., G.J.L., C.H.P., L.M.B.

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

Analysis and interpretation of the data: V.deG., H.B., B.M.J.U., H.C.W.deV., G.J.L., C.H.P., L.M.B.

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

Critical revision of the manuscript for important intellectual content: V.deG., H.B., B.M.J.U., H.C.W.deV., G.J.L., C.H.P., L.M.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

Antonucci G, Aprile T, Paolucci S. Rasch analysis of the Rivermead mobility index: a study using mobility measures of first-stroke inpatients. Arch Phys Med Rehabil 2002; 83: 1442–9.

Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001; 10: 571–8.

Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH, et al. Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. Med Care 1998; 36: 491–502.

Beurskens AJHM, De Vet HCW, Köke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996; 65: 71–6.

Brunet DG, Hopman WM, Singer MA, Edgar CM, MacKenzie TA. Measurement of health-related quality of life in multiple sclerosis patients. Can J Neurol Sci 1996; 23: 99–103.

Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the functional assessment of cancer therapy (FACT) anemia and fatigue scales. J Pain Symptom Manage 2002a; 24: 547–61.

Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002b; 11: 207–21.

Cohen JA, Cutter GR, Fischer JS, Goodman AD, Fedor RH, Jak AJ, et al. Use of the multiple sclerosis functional composite as an outcome measure in a phase 3 clinical trial. Arch Neurol 2001; 58: 961–7.

Cohen JA, Fischer JS, Bolibrush DM, Jak AJ, Kniker JE, Mertz LA, et al. Intrarater and interrater reliability of the multiple sclerosis functional composite outcome measure. Neurology 2000; 54: 802–6.

Cohen L, Pouwer F, Pfennings LE, Lankhorst GJ, Van der Ploeg HM, Polman CH, et al. Factor structure of the disability and impact profile in patients with multiple sclerosis. Qual Life Res 1999; 8: 141–50.

Collen FM, Wade DT, Robb GF, Bradshaw CM. The Rivermead Mobility Index: a further development of the Rivermead motor assessment. Int Disabil Stud 1991; 13: 50–4.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003; 56: 395–407.

Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999; 122: 871–82.

De Vet HCW, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care 2001; 17: 479–87.

Fischer JS, Rudick RA, Cutter GR, Reingold SC. The multiple sclerosis functional composite measure (MSFC): an integrated approach to multiple sclerosis clinical outcome assessment. National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force. Mult Scler 1999; 5: 244–50.

Forlander DA, Bohannon RW. Rivermead mobility index: a brief review of research to date. Clin Rehabil 1999; 13: 97–100.

Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis. J Neurol Neurosurg Psychiatry 2000; 68: 150–6.

Goodkin DE, Cookfair D, Wende K, Bourdette D, Pullicino P, Scherokman B, et al. Inter- and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke expanded disability status scale (EDSS). Multiple Sclerosis Collaborative Research Group. Neurology 1992; 42: 859–63.

Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: a study of persons with multiple sclerosis. Arch Phys Med Rehabil 1990; 71: 870–5.

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987; 40: 171–8.

Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002; 55: 900–8.

Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000; 123: 1027–40.

Hobart JC, Lamping DL, Fitzpatrick R, Riazi A, Thompson A. The Multiple sclerosis impact scale (MSIS-29): a new patient-based outcome measure. Brain 2001a; 124: 962–73.

Hobart JC, Lamping DL, Freeman JA, Langdon DW, McLellan DL, Greenwood RJ, et al. Evidence-based measurement: which disability scale for neurologic rehabilitation? Neurology 2001b; 57: 639–44.

Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome. Health Technol Assess 2004; 8: 1–60.

Hoogervorst EL, Kalkers NF, Van Winsen LML, Uitdehaag BMJ, Polman CH. Differential treatment effect on measures of neurologic exam, functional impairment and patient self-report in multiple sclerosis. Mult Scler 2001a; 7: 335–9.

Hoogervorst EL, Van Winsen LM, Eikelenboom MJ, Kalkers NF, Uitdehaag BM, Polman CH. Comparisons of patient self-report, neurologic examination, and functional impairment in multiple sclerosis. Neurology 2001b; 56: 934–7.

Hoogervorst EL, Zwemmer JN, Jelles B, Polman CH, Uitdehaag BMJ. Multiple sclerosis impact scale (MSIS-29): relation to established measures of impairment and disability. Mult Scler 2004; 10: 569–74.

Hsieh CL, Hsueh IP, Mao HF. Validity and responsiveness of the rivermead mobility index in stroke patients. Scand J Rehabil Med 2000; 32: 140–2.

Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000; 53: 459–68.

Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10: 407–15.

Jonsson A, Dock J, Ravnborg MH. Quality of life as a measure of rehabilitation outcome in patients with multiple sclerosis. Acta Neurologica Scandinavica 1996; 93: 229–35.

Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994; 47: 81–7.

Kalkers NF, De Groot V, Lazeron RH, Killestein J, Ader HJ, Barkhof F, et al. multiple sclerosis functional composite: relation to disease phenotype and disability strata. Neurology 2000; 54: 1233–9.

Kalkers NF, Bergers E, Castelijns JA, Van Walderveen MA, Bot JC, Ader HJ, et al. Optimizing the association between disability and biological markers in multiple sclerosis. Neurology 2001; 57: 1253–8.

Kaufman M, Moyer D, Norton J. The significant change for the timed 25-foot walk in the multiple sclerosis functional composite. Mult Scler 2000; 6: 286–90.

Kidd D, Howard RS, Losseff NA, Thompson AJ. The benefit of inpatient neurorehabilitation in multiple sclerosis. Clin Rehabil 1995; 9: 198–203.

Koziol JA, Lucero A, Sipe JC, Romine JS, Beutler E. Responsiveness of the Scripps neurologic rating scale during a multiple sclerosis clinical trial. Can J Neurol Sci 1999; 26: 283–9.

Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33: 1444–52.

Laman H, Lankhorst GJ. Subjective weighting of disability: an approach to quality of life assessment in rehabilitation. Disabil Rehabil 1994; 16: 198–204.

Lankhorst GJ, Jelles F, Smits RCF, Polman CH, Kuik DJ, Pfennings LE, et al. Quality of life in multiple sclerosis: the disability and impact profile (DIP). J Neurol 1996; 243: 469–74.

Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995; 22: 1191–2.

Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993; 2: 221–6.

Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res 1981; 4: 483–92.

Mancuso CA, Peterson MG. Different methods to assess quality of life from multiple follow-ups in a longitudinal asthma study. J Clin Epidemiol 2004; 57: 45–54.

Marolf MV, Vaney C, Konig N, Schenk T, Prosiegel M. Evaluation of disability in multiple sclerosis patients: a comparative study of the functional independence measure, the extended Barthel index and the expanded disability status scale. Clin Rehabil 1996; 10: 309–13.

McGuigan C, Hutchinson M. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. J Neurol Neurosurg Psychiatry 2004; 75: 266–9.

Miller DM, Rudick RA, Cutter G, Baier M, Fischer JS. Clinical significance of the multiple sclerosis functional composite: relationship to patient-reported quality of life. Arch Neurol 2000; 57: 1319–24.

Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997; 50: 869–79.

Noseworthy JH. Clinical scoring methods for multiple sclerosis. Ann Neurol 1994; 36: S80–5.

Noseworthy JH, Vandervoort MK, Wong CJ, Ebers GC. Interrater variability with the expanded disability status scale (EDSS) and functional systems (FS) in a multiple sclerosis clinical trial. The Canadian Cooperation Multiple Sclerosis Study Group. Neurology 1990; 40: 971–5.

Ostelo RW, De Vet HC. Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 2005; 19: 593–607.

Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. Arch Phys Med Rehabil 1996; 77: 1226–32.

Patzold T, Schwengelbeck M, Ossege LM, Malin JP, Sindern E. Changes of the multiple sclerosis functional composite and EDSS during and after treatment of relapses with methylprednisolone in patients with multiple sclerosis. Acta Neurol Scand 2002; 105: 164–8.

Pfennings LE, Van der Ploeg HM, Cohen L, et al. A health-related quality of life questionnaire for multiple sclerosis patients. Acta Neurol Scand 1999a; 100: 148–55.

Pfennings LE, Van der Ploeg HM, Cohen L, Polman CH. A comparison of responsiveness indices in multiple sclerosis patients. Qual Life Res 1999b; 8: 481–9.

Poser CM, Paty DW, Scheinberg L, et al. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983; 13: 227–31.

Riazi A, Hobart JC, Lamping DL, Fitzpatrick R, Thompson AJ. Evidence-based measurement in multiple sclerosis: the psychometric properties of the physical and psychological dimensions of three quality of life rating scales. Mult Scler 2003; 9: 411–9.

Rudick R, Antel J, Confavreux C, Cutter G, Ellison G, Fischer J, et al. Clinical outcomes assessment in multiple sclerosis. Ann Neurol 1996; 40: 469–79.

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.

Schunemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and the St. George's respiratory questionnaire in patients with chronic airflow obstruction. J Clin Epidemiol 2003; 56: 1170–6.

Schwid SR, Goodman AD, Apatoff BR, Coyle PK, Jacobs LD, Krupp LB, et al. Are quantitative functional measures more sensitive to worsening multiple sclerosis than traditional measures? Neurology 2000; 55: 1901–3.

Schwid SR, Goodman AD, McDermott MP, Bever CF, Cook SD. Quantitative functional measures in multiple sclerosis: what is a reliable change? Neurology 2002; 58: 1294–6.

Sharrack B, Hughes RA. The Guy's neurological disability scale (GNDS): a new disability measure for multiple sclerosis. Mult Scler 1999; 5: 223–33.

Solari A, Radice D, Manneschi L, Motti L, Montanari E. The multiple sclerosis functional composite: different practice effects in the three test components. J Neurol Sci 2005; 228: 71–4.

Stratford PW, Binkley FM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther 1996; 76: 1109–23.

Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995; 48: 1369–78.

Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003; 12: 349–62.

Uitdehaag BMJ, Ader HJ, Roosma TJ, De Groot V, Kalkers NF, Polman CH. Multiple sclerosis functional composite: impact of reference population and interpretation of changes. Mult Scler 2002; 8: 366–71.

Van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. The rehabilitation activities profile: a validation study of its use as a disability index with stroke patients. Arch Phys Med Rehabil 1995; 76: 501–7.

Van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the rehabilitation activities profile and the Barthel Index. J Clin Epidemiol 1996; 49: 39–44.

Van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the action research arm test and the Fugl-Meyer assessment scale in chronic stroke patients. J Rehabil Med 2001; 33: 110–3.

Van der Putten JJ, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the functional independence measure. J Neurol Neurosurg Psychiatry 1999; 66: 480–4.

Van der Windt DA, Van der Heijden GJ, De Winter AF, Koes BW, Devillé W, Bouter LM. The responsiveness of the shoulder disability questionnaire. Ann Rheum Dis 1998; 57: 82–7.

Vaney C, Vaney S, Wade DT. SaGAS, the short and graphic ability score: an alternative scoring method for the motor components of the multiple sclerosis functional composite. Mult Scler 2004; 10: 231–42.

Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health-related quality of life measure for multiple sclerosis. Qual Life Res 1995; 4: 187–206.

Whitaker JN, Mcfarland HF, Rudge P, Reingold SC. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. Mult Scler 1995; 1: 37–47.

Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999; 52: 861–73.

Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42: 121–30.