Responsiveness of patient-reported outcome measures in multiple sclerosis relapses: the REMS study

Andrea Giordano, Eugenio Pucci, Paola Naldi, Laura Mendozzi, Clara Milanese, Federica Tronci, Maurizio Leone, Nerina Mascoli, Loredana La Mantia, Giorgio Giuliani, et al.

Full title: Responsiveness of patient-reported outcome measures in multiple sclerosis relapses: the REMS study

Corresponding Author:

Alessandra Solari, MD

Unit of Neuroepidemiology

Foundation IRCCS Neurological Institute C. Besta

Via Celoria,11

20133 Milan, Italy

E-mail: solari@istituto-besta.it

Tel. +39 02 23942391

Fax +39 02 70606233

Co-authors:

Andrea Giordano[1], Eugenio Pucci[2], Paola Naldi [3], Laura Mendozzi[4], Clara Milanese[5], Federica Tronci[4], Maurizio Leone [3], Nerina Mascoli[5], Loredana La Mantia[5], Giorgio Giuliani[2], and Alessandra Solari[1]

1. Unit of Neuroepidemiology, Foundation IRCCS Neurological Institute C. Besta, Milan, Italy
2. Neurology Department, Macerata Hospital, Macerata, Italy
3. Neurology Clinic, University Hospital "Maggiore della Carità", Novara, Italy
4. MS Unit, Foundation IRCCS Don C. Gnocchi Onlus, Milan, Italy
5. Multiple Sclerosis Unit, Foundation IRCCS Neurological Institute C. Besta, Milan, Italy

Key words: Quality of life; multiple sclerosis; randomised trials.

Word count: 2829

Declaration:

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence on a worldwide basis to the BMJ Publishing Group Ltd and its Licensees to permit this article (if accepted) to be published in the Journal of Neurology, Neurosurgery & Psychiatry editions and any other BMJPGL products to exploit all subsidiary rights, as set out in the BMJPGL licence.

ABSTRACT

Objectives. To assess the responsiveness of the three most-used patient-reported multiple sclerosis (MS)-specific questionnaires: the Functional Assessment of MS (FAMS), the MS Impact Scale (MSIS-29), and the 54-item MS Quality of Life (MSQOL-54).

Design. Prospective multicentre longitudinal study on 104 MS patients treated with i.v. steroids for clinical exacerbation.

Methods. Patient-reported data, EDSS score and clinical information were collected at admission and 8 weeks later. 'Internal' (distribution-based) responsiveness was assessed by standardized response means (SRM). 'External' (anchor-based) responsiveness was assessed by receiver operating characteristic (ROC) curves in relation to corresponding changes in a pre-specified reference measure (anchor). The pre-specified anchor was patients' self-reported recovery assessed on a five-point Likert scale.

Results. SRM was 0.39 for FAMS, 0.58 for MSIS-29 physical scale, 0.45 for MSIS-29 psychological scale, 0.71 for MSQOL-54 physical health composite, and 0.57 for MSQOL-54 mental health composite. Seventy three patients (70%) reported they had improved; physicians agreed substantially with patient assessments (kappa statistic 0.70, 95% CI 0.54-0.85). Areas under ROC curves differed significantly from 0.50 only for the MSIS-29 and MSQOL-54 scales, where areas ranged from 0.65 (95% CI 0.53-0.76) for the MSIS-29 psychological scale, to 0.70 (95% CI 0.58-0.81) for the MSQOL-54 mental health composite. Areas under ROC curves assessed using a physician-based anchor were similar to the patient-based areas.

Conclusions. The responsiveness of the MS-specific instruments was less than ideal. The MSIS-29 and the MSQOL-54 were significantly more responsive, using both distribution-based and anchor-based approaches, than FAMS, and should be preferred in longitudinal studies.

INTRODUCTION

Patient-reported assessment is increasingly incorporated into multiple sclerosis (MS) clinical research. Several health-related quality of life (HRQOL) instruments have been devised for use as outcome measures in MS, and are now widely accepted.[1] As these instruments become more widely used to evaluate the safety and efficacy of new treatments, it is imperative that issues concerning their appropriateness as evaluative tools be addressed. Specifically, cross-sectional validity, reliability and responsiveness must all be adequate.[2, 3] Although the assessment of cross-sectional validity and reliability are straightforward, evaluation of responsiveness (i.e. ability to accurately detect change when it has occurred), is more complex.[2] In fact relatively few studies have assessed responsiveness in MS-specific patient-reported instruments, [4-9] and head-to-head comparisons are rare.[6-8]

There is also a lack of consensus about the conceptualization of responsiveness. Several statistical approaches have been used, and none has emerged as standard.[10, 11] Husted et al. described two basic types of responsiveness: (a) 'internal' (or distribution-based) responsiveness defined as the ability of a measure to detect change over a pre-specified time, and often assessed using various effect size statistics; and (b) 'external' (or anchor-based) responsiveness which evaluates the extent to which changes in a measure over time relate to corresponding changes in a reference measure (anchor) of health status.[12] Responsiveness is usually evaluated using a single-group repeated measurement design, in which patients are assessed before and after a specific treatment that has known efficacy. To examine anchor-based responsiveness, a clinically meaningful change is defined and evaluated from the perspective of the patient, proxy, caregiver or clinician. The clinician is the most appropriate evaluator of impairment, while for measures of functioning and HRQOL the patient should be the prime assessor.[3, 10]

In the present study, Responsiveness in MS (REMS), we prospectively assessed the relative responsiveness of the three most-used MS-specific patient-reported instruments in patients receiving i.v. steroids for exacerbation. The instruments were the Functional Assessment of MS [FAMS], the 29-item MS Impact Scale [MSIS-29], and the 54-item MS Quality of Life [MSQOL-54]).[13-15] We used both distribution-based and anchor-based approaches. We also compared the patient-derived anchor of clinical improvement with the physicians' assessment of improvement.

MATERIALS AND METHODS

Design and participants

The REMS study is a prospective, open label multicentre study (pre-post design). The study population were consecutive patients treated with i.v. steroids for exacerbations at four Italian MS centres, satisfying all of the following criteria: (a) MS diagnosis according to the revised McDonald criteria; [16] (b) new exacerbation severe enough to necessitate i.v. steroids; (c) age between 18 and 55 years; (d) signed informed consent; (e) absence of severe cognitive compromise (defined as a score >4 in mental functional system);[17] (f) no exacerbations in the three months prior to enrolment. Before starting the study, the study protocol, patient information leaflet, informed consent form, and other relevant documents were approved by the Ethics Committee of each participating centre.

HRQOL instruments

The FAMS consists of a generic core HRQOL measure (the Functional Assessment of Cancer Therapy – general scale), supplemented with MS specific items.[14] The 59 items of the FAMS are divided into 6 subscales: mobility, symptoms, emotional well-being, general contentment, thinking/fatigue, and family/social well-being. Only items 1 through 44 are included in the total score

(0-4 score range for each); items 45 through 59 being excluded.[14] Thus, scores can range from 0 (worst HRQOL) to 176 (best HRQOL).

The MSIS-29 consists of 29 items divided into two scales: physical (20 items) and psychological (nine items).[15] For each item, participants rate their symptoms on a 5-point Likert scale, lower scores indicate independence and higher scores indicate greater compromise. The physical scale score can range from 20 (best functioning) to 100 (worst functioning), and psychological scale score can range from 5 (best) to 45 (worst).

The MSQOL-54 includes the 36-item Short Form (SF-36), a generic measure of self-reported health plus 18 MS-specific items.[13] The entire MSQOL-54 contains 52 items divided into 12 domains, plus 2 individual items (sexual satisfaction and change in health over previous year). As with the SF-36 alone, physical and mental health composite scores are calculated.[18, 19]

We used the published Italian versions of the FAMS and MSQOL-54,[20, 21] while an unpublished Italian version of the MSIS-29 was kindly provided for the present study by Professor Hobart. The recall periods are one week for the FAMS, two weeks for MSIS-29 and four weeks for MSQOL-54. However we used an 'acute' MSQOL-54 that, by analogy with the SF-36 version 2 acute form,[22] uses a one-week recall period instead of the four-week period of the standard instrument. To facilitate comparison, FAMS and MSIS-29 scores were transformed into a 0 (worst) -100 (best) scale.

Patient assessment of change

Global assessment of recovery from the exacerbation was assessed using a standardized health transition question (SHTQ) in which each patient's assessment of his/her recovery was rated on five-point Likert scale (1 much worse than two months ago; 2 somewhat worse; 3 about the same; 4 somewhat better; 5 much better). An overall assessment of exacerbation outcome was also given by the caring neurologist using a physician SHTQ rated on five-point Likert scale.

Baseline and follow-up visits

At baseline eligible MS patients who had signed the consent form were given the 3 questionnaires to complete. The order of administration was determined randomly. The neurologist then examined the patient (standard neurological examination, EDSS,[17] and recorded demographic and clinical data on the REMS case report form. Intravenous steroids were then started.

At the follow-up visit, scheduled 8 (±1) weeks after baseline, each patient first received the SHTQ form, followed by the three questionnaires to complete again, in the same order of administration as at baseline. Recent disease history (including duration and course of exacerbation, therapies received in the previous 8 weeks), standard neurological examination including EDSS, and overall assessment of exacerbation outcome (physician SHTQ) were then assessed by the caring neurologist and recorded on the REMS case report form.

STATISTICAL ANALYSES

Descriptive analyses were performed using the chi square test for proportions, and one-way ANOVA, the Wilcoxon rank-sum test or Kruskal-Wallis test for continuous data. Test-retest reliability was assessed with the kappa statistic for ordinal data.[23]

For each instrument we assessed missing items, internal-consistency of scales, and score distributions, including the percentages of patients obtaining floor and ceiling scores for individual scales or subscales.[24] Anchor-based responsiveness was determined from receiver operating characteristic (ROC) curves. The SHTQ was used as the reference measure for external responsiveness: patients scoring >3 were classified as having improved. Distribution-based responsiveness was evaluated by the standardized response mean (SRM), which is the ratio of the mean score change to the standard deviation of the score change.[10]

Although there are no absolute standards for the SRM, it has been suggested (as for other measures of effect size) that small, moderate and large responsiveness should correspond to SRMs of 0.20-0.49, 0.50-0.79, and 0.80 or more, respectively.[25, 26] With regard to anchor-based responsiveness, the area under the ROC curve (AUC) can range from 0.5 (no accuracy in distinguishing improved from unimproved patients) to 1.0 (perfect accuracy).[27]

Data were analyzed using Stata Statistical Software, release 9.0 (Stata, College Station, Texas). All tests were two-tailed and $p < 0.05$ was considered statistically significant.

RESULTS

Between January 2007 and August 2008 209 exacerbating MS patients were screened at the four participating centres, and 107 (51%) were enrolled. Seventy three excluded patients were women (72%), they had a mean age of 40.7 years (SD 11.8), and a median EDSS score of 3.0 (range 0 – 8).The main reasons for exclusion were: inability to return for follow-up visit (19 patients); age > 55 years (16 patients); exacerbation in the preceding three months (14 patients); refusal to participate (12 patients); decision not to administer i.v. steroids (10 patients); and marked cognitive impairment (9 patients).

Of the 107 enrolled patients, one 43 year-old woman did not receive i.v. steroids because of spontaneous recovery and two patients did not complete the study: a 37 year-old woman did not attend for follow up, and a 25 year-old woman did not complete the questionnaires at follow-up. Table 1 shows the characteristics of the 104 people with MS who completed the study, whose mean age was 39 years. One 60 year-old woman was included (protocol violation); 82% were women, average disease duration was 12 years (range 6 months to 30 years), and 37 patients were on disease-modifying drugs. Median EDSS score was 3.5 (range 1.0-7.5); most patients had multiple symptoms

during the exacerbation, and the most frequently affected Kurtzke functional systems were pyramidal (60%), sensory (45%), and brainstem (20%).

The clinimetric properties of the three questionnaires satisfied recognized standards in terms of missing items (ranging from 4% for MSIS-29 physical subscale to 9% for FAMS total scale), internal consistency (Cronbach's alpha <0.70 only for the MSQOL-54 social function scale), floor effect (three patients scoring minimum on the MSIS-29 physical scale; one scoring minimum on the MSIS-29 psychological scale), and ceiling effect (one patient scoring maximum on MSQOL-54 physical health composite scale). Changes from baseline to follow-up were significant for most scale and subscale scores. Exceptions were the mobility, emotional well-being, general contentment, thinking/fatigue, and family/social well-being subscales of FAMS; and the general health, overall quality of life, sexual function, sexual satisfaction, and cognitive function subscales of MSQOL-54 (Table 2).

To allow direct comparison of physical and psychological domains, and to compare our results with those from other studies [6, 7], we also calculated responsiveness indexes for the FAMS mobility and emotional well-being subscales (Table 2).

With regard to distribution-based responsiveness, the SRM of FAMS were total score 0.39, mobility subscale 0.32, and emotional well-being subscale 0.22 (Table 2). The SRMs of MSIS-29 were 0.58 (physical scale) and 0.45 (psychological scale). The SRMs of MSQOL-54 were 0.71 (physical health composite), and 0.57 (mental health composite).

Seventy-three patients (70%) reported they had improved over the eight-week period (SHTQ score >3); corresponding figures from treating neurologists (physician SHTQ score >3) were 83 (80%). Overall agreement between patients and neurologists was 86% (kappa 0.70 [95% CI 0.54-0.85; p<0.001]).

Fig 1 shows box plots of changes in scale and subscale scores in unimproved and improved patients. The median change in scale/subscale scores was close to zero in unimproved patients, and was

positive in improved patients, ranging from 3.6 (interquartile range, IR – 5. 6 to 17.9) for FAMS mobility to 10.9 (IR 1.2 to 20.6) for MSQO-54 mental health composite. There was considerable overlap of IRs (boxes in Fig 1) between unimproved and improved patients for all scale/subscale score changes. Box plots of changes in the EDSS score are shown in Fig 2. The median change in EDSS score in unimproved patients was zero (IR 0 to 0.5) and was -1.0 in improved patients (IR 0.5 to 1.5) with no overlap of boxes.

ROC curves with AUCs for physical domains and EDSS are shown in Fig 3; corresponding curves and AUCs for psychological domains are shown in Fig 4. For physical domains, patient-reported scale/subscale AUCs ranged from 0.52 (FAMS mobility subscale) to 0.68 (MSIS-29 physical scale); while the AUC of 0.79 (95% CI 0.69 - 0.90) for EDSS indicated excellent responsiveness. For the psychological domains, AUCs ranged from 0.57 (FAMS emotional wellbeing subscale) to 0.70 (MSQOL-54 physical health composite). Most AUCs differed significantly from 0.50, the exception being that for FAMS. AUCs determined from the physician-based anchor for improvement were in all cases consistent with patient based AUCs (Table 2).

DISCUSSION

Responsiveness is one of the least well studied clinimetric properties of patient-reported measures in MS.[1, 8] Our distribution-based approach showed that the MSIS-29 and the MSQOL-54 had good internal responsiveness, but that responsiveness was less satisfactory for the FAMS. The anchor-based approach supported the distribution-based findings, with AUCs above 0.5 for the MSIS-29 scales and MSQOL-54 composite scales, but not for the FAMS. Two papers that studied different MS populations in the UK (patients admitted for exacerbation, patients admitted for rehabilitation, and primary progressive patients)[6, 7] also found low internal responsiveness of the FAMS compared to

the MSIS-29. However, we did not find the high internal responsiveness for the MSIS-29 scales (SRM 0.90 for physical, and 0.72 for psychological) reported by Riazi et al.[6]

With regard to external responsiveness, we used a patient-determined anchor as base to assess clinically relevant improvement, as recommended for patient-reported instruments.[28, 29] Nevertheless this patient-determined anchor was highly consistent with the physician-based anchor (as shown by kappa statistic of 0.70); furthermore patient-based and neurologist-based ROC AUCs were closely similar (Table 2).

To our knowledge, only two studies have assessed the anchor-based responsiveness of patient-reported measures in MS.[8,9] Both evaluated changes (mainly worsening) over the long term, using the MSIS-29 physical scale,[9] and the Disability Impact Profile.[8] The former used a physician-based anchor, the latter used both physician and patient based anchors. AUCs for the MSIS-29 physical scale (in patients with EDSS in the range 0-5.0 − to which 85% of our sample belonged) were 0.73 (95% CI 0.63-0.82) in the Costelloe et al. study [9] and 0.68 (95% CI 0.57-0.78) in our study. The Disability Impact Profile is a less widely employed MS-specific HRQOL measure, thus direct comparison with our study is not possible.[8] It is noteworthy however that EDSS AUCs in both studies, obtained using a patient-reported criterion for change, were similar: 0.70 (95% CI 0.62– 0.79) in de Groot et al. [8] and 0.79 (95% CI 0.69–0.90) in our study.

A possible limitation of our study is that we assessed responsiveness to health status changes in the context of treatment for a newly appearing exacerbation, whereas responsiveness over a longer period would have been more pertinent for use of patient-reported instruments in clinical trials on disease modifying treatments. We decided to consider exacerbations because SHTQs are retrospective evaluations and subject to recall bias which may be considerable in a long-term study or (for a patient-based anchor) if the patient has cognitive compromise.[29] Parenthetically we note that recall bias could have accounted for the poor patient-physician agreement regarding clinical change

found by de Groot et al.[8] We decided on an eight week period – sufficiently long to detect exacerbation recovery, but short enough to avoid recall problems.[5] Since we were interested in a sub-acute event (MS exacerbation) we changed the MSQOL-54 referral period from four weeks to one week. [22] Note, however, that this modified version of the MSQOL-54 has not been validated.

Overall the two approaches we used to assess responsiveness produced results that were consistent with each other, confirming that head to head comparisons of instruments in a single setting provide responsiveness rankings that are relatively insensitive to the method used to assess responsiveness.[30]. Nevertheless responsiveness was less than excellent, particularly external responsiveness. Cognitive compromise could have reduced responsiveness: however patients with overt cognitive impairment were excluded from the study, and patient judgment of clinical improvement was in agreement with that of the neurologists.

Finally, responsiveness is negatively affected by variability of score changes and measurement error.[27] We made every effort to limit variability due to study conduct: investigators were instructed and trained on the study protocol, and data were recorded on case report forms which were collected, processed and analyzed by a central unit. Thus, the variability of patient-reported scale scores (Fig 1) appears mainly due to the heterogeneity of the study population and the multidimensionality of the instruments − which have a positive effect on the generalizability of our findings. Additional strengths are the prospective design and completeness of follow-up (97%).

To conclude, we have found that the overall responsiveness of the three most used MS-specific patient-reported instruments was less than ideal, but that two instruments – the MSIS-29 and MSQOL-54 – emerged as significantly more responsive and should be used for longitudinal studies on persons with MS. It is disconcerting that the EDSS proved to be more responsive than the MS-specific patient-reported instruments, particularly since we used a patient-based criterion of change. In fact the

responsiveness is the least-assessed property of patient-reported health status questionnaires, and assessment criteria are ill defined. [31] Further studies that prospectively determine the responsiveness of MS-specific patient-reported outcome measures are clearly necessary. It is also essential that the findings of such instruments – habitually used as secondary outcome measures in randomized trials – should be fully accessible to the scientific community. Finally, it is important to assess these instruments in diverse linguistic and cultural settings, since it is not clear that responsiveness transfers stably across cultures.[32]

ACKNOWLEDGMENTS: We are indebted to Giusi Ferrari for study management, Don Ward for help with the English, and Christoph Heesen for critical comments on the manuscript. Elisabetta Cartechini and Elisabetta Medicato helped with patient evaluation at the Macerata centre.

COMPETING INTEREST: None.

FUNDING: The FISM (Fondazione Italiana Sclerosi Multipla) funded the study (Grant No. 2005/R/19 to A. Solari) and supported A. Giordano with a training fellowship.

Table 1 Characteristics of 104 MS patients who completed the study

Characteristic			Number of patients (%)
Women			85 (81.7)
Age (years)*			38.7, 9.1 (18 - 60)
Disease course:	Relapsing-remitting		88 (84.6)
	Relapsing-progressive		16 (15.3)
Time since first symptom (years)*			11.9, 7.4 (0 - 30)
Time since diagnosis (years)*			8.9, 6.8 (0 - 25)
Functional system affected in current relapse:		Pyramidal	62 (59.6)
		Cerebellar	19 (18.2)
		Brainstem	21 (20.1)
		Sensory	47 (45.1)
		Visual	11 (10.5)
		Mental	1 (0.9)
		Bowel/bladder	13 (12.5)
EDSS score**			3.5 (1.0 - 7.5)
Disease-modifying drugs:	Beta Interferon		23 (22.1)
	Glatiramer acetate		4 (3.8)
	Natalizumab		1 (0.9)
	Mitoxantrone		3 (2.8)
	Azathioprine		6 (5.7)

* Mean, SD (minimum-maximum)

**Median (minimum-maximum)

Table 2 Scale (subscale) values and responsiveness indexes. SRM is the standardized response mean, AUC the area under the receiver operating curve, and CI is the confidence interval

Scale or subscale	Baseline	Follow-up	SRM	AUC (95% CI)	
	Mean (SD)			Patient-based anchor	Physician-based anchor
FAMS					
Mobility	55.6 (17.8)	60.6 (16.1)	0.32	0.52 (0.41-0.64)	0.51 (0.38-0.65)
Symptoms*	64.8 (22.0)	69.3 (21.6)			
Emotional well-being	69.1 (24.0)	72.4 (22.3)	0.22	0.57 (0.45 - 0.70)	0.46 (0.32 - 0.60)
General contentment	65.2 (19.9)	67.8 (19.2)			
Thinking/fatigue	55.9 (23.8)	59.8 (23.6)			
Family/social well-being	73.6 (17.7)	74.4 (19.6)			
Total*	63.7 (15.6)	67.0 (15.8)	0.39	0.61(0.49 - 0.73)	0.55(0.42 - 0.69)
MSIS-29					
Physical*	70.9 (21.8)	80.0 (19.2)	0.58	0.68 (0.57 - 0.78)**	0.60 (0.47 - 0.73)
Psychological*	66.0 (23.6)	72.8 (23.1)	0.45	0.65 (0.53 - 0.76)**	0.65 (0.52 - 0.78)**
MSQOL-54					
Physical function*	61.1 (29.3)	69.4 (27.4)			
Role limitation – physical*	25.09 (37)	57.7 (42.8)			
Pain*	63.9 (28.4)	75.4 (22.6)			
Mental health*	54.1 (22.5)	61.5 (24.4)			
Role limitation – emotional*	53.1 (43.1)	69.5 (41.2)			
Energy*	42.2 (21.1)	50.8 (23.0)			
Social function*	67.8 (19.5)	74.3 (19.7)			
General health	48.6 (17.3)	48.9 (17.1)			
Change in health*	33.7 (21.8)	44.1 (21.9)			

Overall quality of life	58.6 (16.8)	60.9 (17.8)			
Health distress*	58.7 (24.0)	71.0 (24.0)			
Sexual function	78.1 (29.4)	79.4 (26.4)			
Sexual satisfaction	69.2 (26.0)	69.7 (23.6)			
Cognitive function	70.4 (22.8)	73.4 (24.2)			
Physical health composite*	55.5 (18.5)	64.8 (18.7)	0.71	0.67 (0.56 - 0.79)**	0.62 (0.48 - 0.75)
Mental health composite*	57.9 (22.1)	66.4 (23.2)	0.57	0.70 (0.58 - 0.81)**	0.67 (0.54 - 0.80)**

* $p \leq 0.001$ for comparison between baseline and follow up scores (paired t-test).

** AUC differs significantly from 0.50.

REFERENCES

1. **Solari A.** Role of health-related quality of life measures in the routine care of people with multiple sclerosis. *Health Qual Life Outcomes* 2005;**3**:16.
2. **Guyatt GH,** Kirtschner B, Jaeschke R. Measuring health status. What are the necessary measurement properties? *J Clin Epidemiol* 1992;**45**:1341–5.
3. **Guyatt GH,** Osoba D, Wu AW, *et al.* Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;**77**:371–83.
4. **Pfennings LE,** van der Ploeg HM, Cohen L, *et al.* A comparison of responsiveness indices in multiple sclerosis patients. *Qual Life Res* 1999;**8**:481–9.
5. **Bethoux F,** Miller DM, Kinkel RP. Recovery following acute exacerbations of multiple sclerosis: from impairment to quality of life. *Mult Scler 2001*;**7**:137–42.
6. **Riazi A,** Hobart JC, Lamping DL, *et al.* 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.
7. **Hobart JC,** Riazi A, Lamping DL, *et al.* How responsive is the Multiple Sclerosis Impact Scale (MSIS-29). A comparison with some other self-report scales. *J Neurol Neurosurg Psychiatry* 2005;**76**:1539–43.
8. **de Groot V,** Beckerman H, Uitdehaag BM, *et al.* The usefulness of evaluative outcome measures in patients with multiple sclerosis. *Brain* 2006;**129**:2648–59.
9. **Costelloe L,** O'Rourke K, Kearney H, *et al.* The patient knows best: significant change in the physical component of the Multiple Sclerosis Impact Scale (MSIS-29 physical). *J Neurol Neurosurg Psychiatry* 2007;**78**:841–4.
10. **Liang MH.** Longitudinal construct validity. Establishment of clinical meaning in patient evaluative instruments. *Med Care* 2000;**9** (suppl 2):84–90.

11. **de Boer MR,** Moll AC, de Vet HC, *et al.* Psychometric properties of vision-related quality of life questionnaires: a systematic review. *Ophtal Physiol* 2004;**24**:257–73.
12. **Husted JA,** Cook RJ, Farewell VT, *et al.* Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol* 2000;**53**:459–68.
13. **Vickrey BG**, Hays RD, Harooni R, *et al.* A health-related quality of life measure for multiple sclerosis. *Qual Life Res* 1995;**4**:187–206.
14. **Cella DF**, Dineen K, Arnason B, *et al.* Validation of the Functional Assessment of Multiple Sclerosis quality of life instrument. *Neurology* 1996;**47**:129–39.
15. **Hobart JC**, Lamping DL, Fitzpatrick R, *et al.* The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. *Brain* 2001;**124**:962–73.
16. **Polman CH**, Reingold SC, Edan G, *et al.* Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Ann Neurol* 2005;**58**:840–6.
17. **Kurtzke JF.** Rating neurologic impairment in multiple sclerosis: an expanded disability status scale. *Neurology* 1983;**33**:1444–52.
18. **Ware JE.** *SF-36 Health Survey: manual and interpretation guide.* Boston: The Health Institute, New England Medical Centre; 1993.
19. **Apolone G,** Mosconi P. The Italian SF-36 Health Survey: translation, validation and norming from a clinical epidemiology perspective. *J Clin Epidemiol* 1998;**51**:1025–36.
20. **Provinciali L,** Ceravolo MG, Bartolini M, *et al.* A multidimensional assessment of multiple sclerosis: relationships between disability domains. *Acta Neurol Scand* 1999;**100**:156–62.
21. **Solari A,** Filippini G, Mendozzi L, *et al.* Validation of Italian multiple sclerosis quality of life 54 questionnaire. *J Neurol Neurosurg Psychiatry* 1999;**67**:158–62.
22. **Ware JE,** Kosinski M, Dewey JE. *How to score version two of the SF-36 Health Survey*. Lincoln, RI: QualityMetric Inc.;2000.

23. **Landis JR,** Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.

24. **Hays RD,** Hayashi T. Beyond internal consistency: Rationale and user's guide for Multitrait Analysis Program on the microcomputer. *Behav Res Methods Instrum Computers* 1990;**22**:167.

25. **Stucki G,** Liang MH, Fossel AH, *et al.* Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol* 1995;**48**:1349–78.

26. **Zou GY.** Quantifying responsiveness of quality of life measures without an external criterion. *Qual Life Res* 2005;**14**:1545–52.

27. **Deyo RA,** Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Controlled Clin Trials 1991*;**12**:142s–58s.

28. **Paltamaa J,** Sarasoja T, Leskinen E, *et al.* Measuring deterioration in International Classification of Functioning Domains of people with multiple sclerosis who are ambulatory. *Phys Ther* 2008;**88**:176–90.

29. **Revicki DA,** Gnanasakthy A, Weinfurt K. Documenting the rationale and psychometric characteristics of patient reported outcomes for labelling and promotional claims: the PRO Evidence Dossier. *Qual Life Res* 2007;**16**:717–23.

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

31. **Terwee CB,** Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;**60**:34–42.

32. **de Vet HC,** Terwee CB, Bouter LM. Current challenges in clinimetics. *J Clin Epidemiol* 2003;**56**:1137–41.

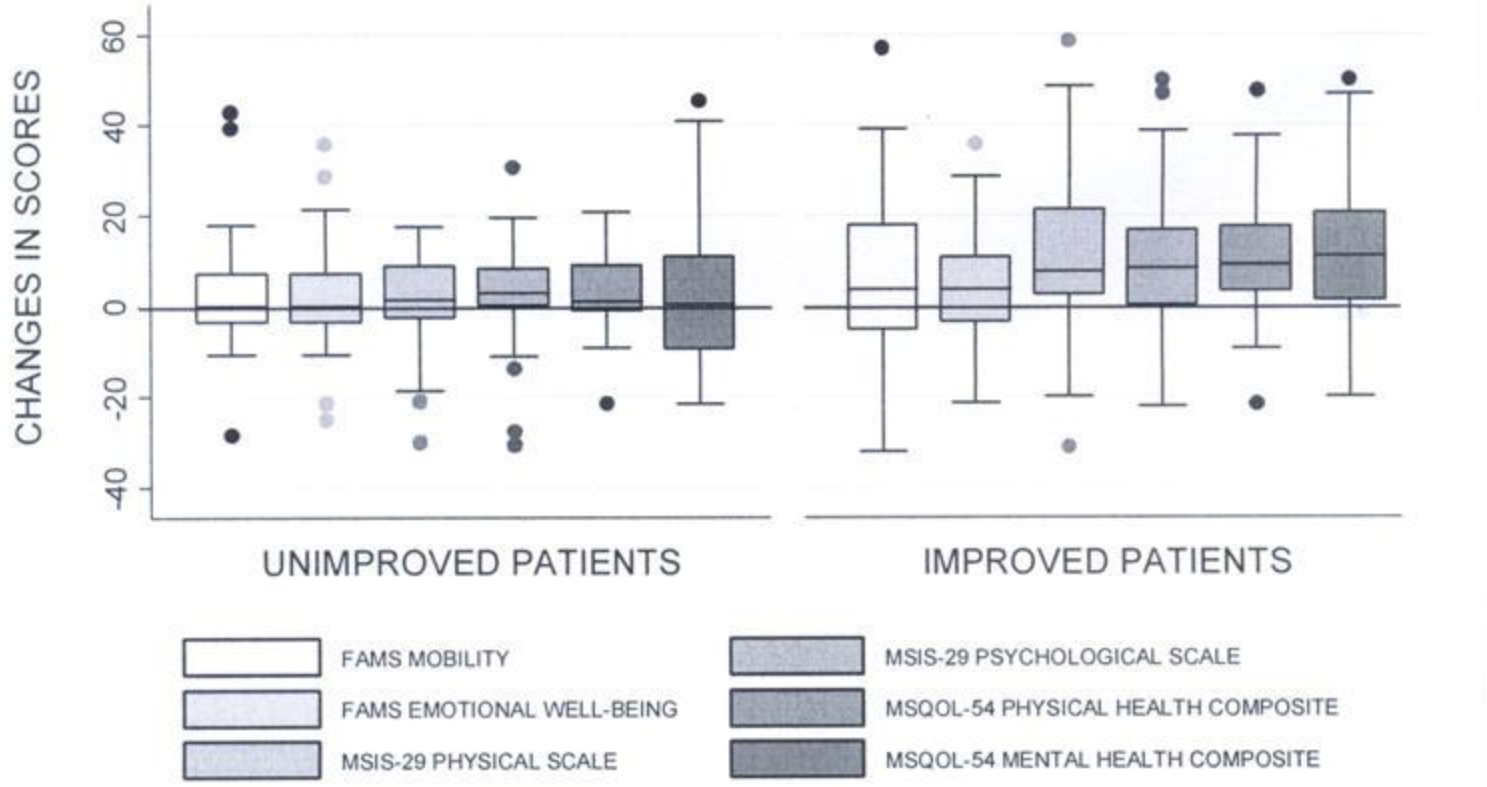

Figure 1 Box plots for changes in score of patient-reported measures for unimproved (n=31) and improved patients (n=73). Boxes represent the interquartile range, horizontal lines inside boxes represent medians, and tails represent the 5th – 25th and the 75th - 95th percentile range.

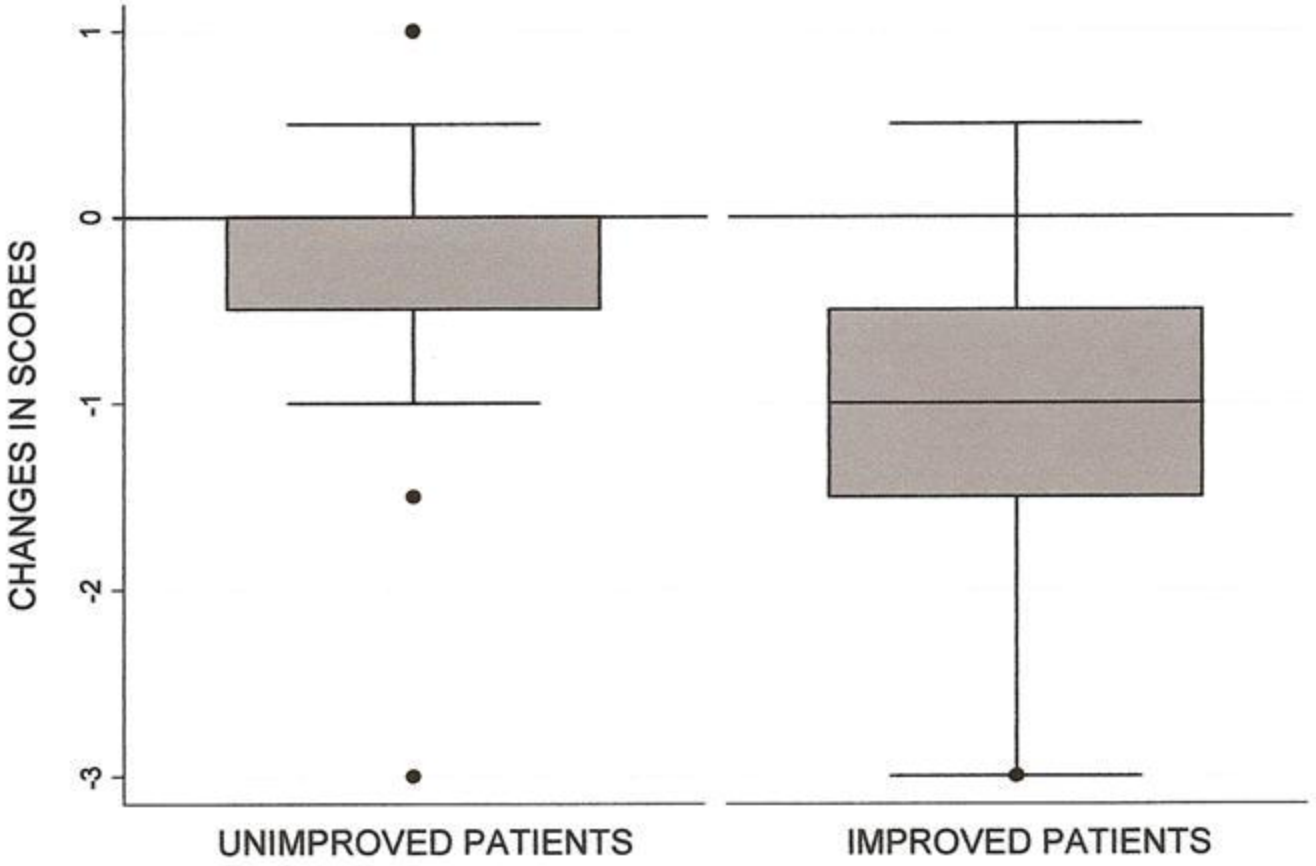

Figure 2 Box plots for changes in EDSS score from baseline for unimproved (n=31) and improved patients (n=73). Boxes represent the interquartile range, horizontal lines inside boxes represent medians, and tails represent the 5^{th} – 25^{th} and the 75^{th} – 95^{th} percentile range.

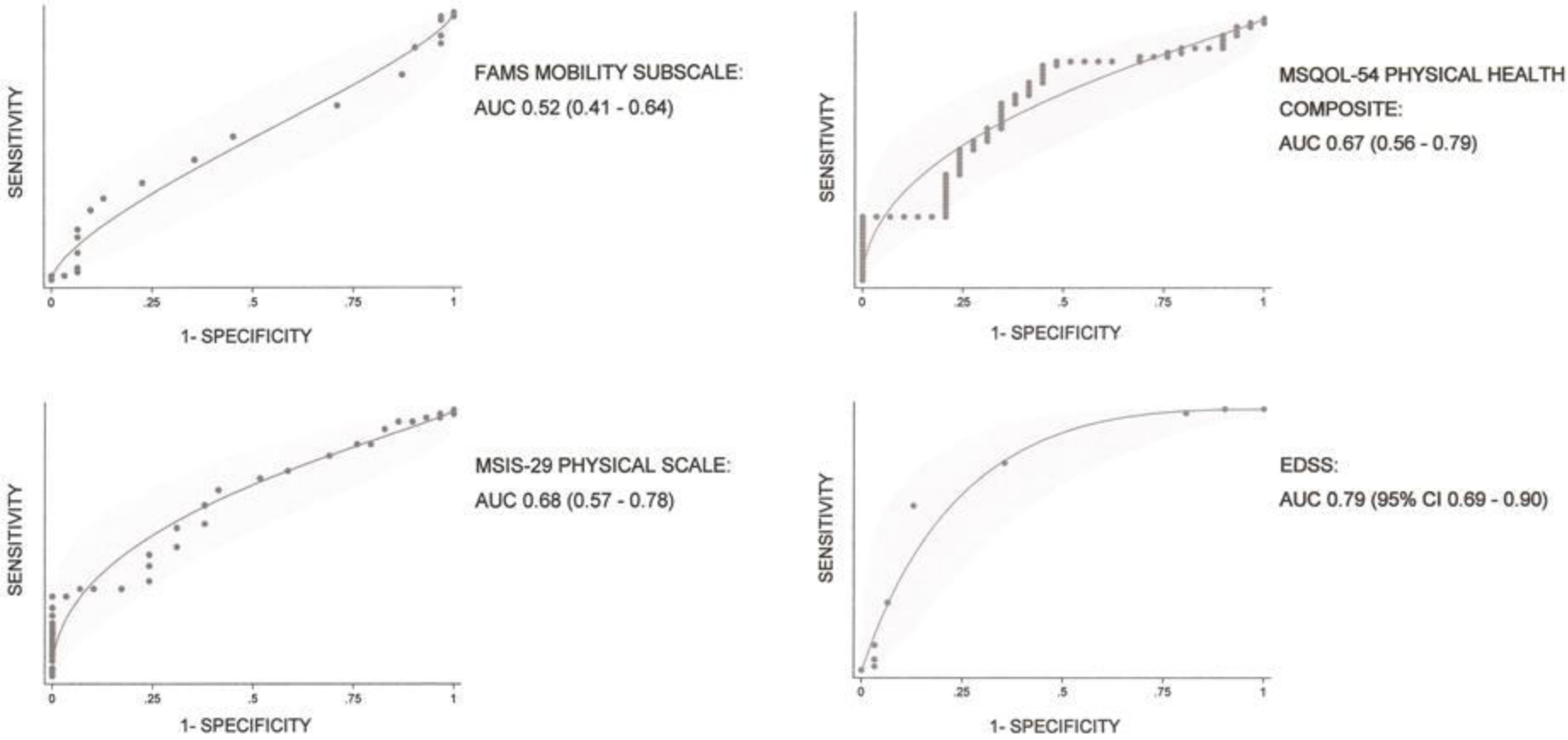

Figure 3 Plots of the receiver operating characteristics curves for changes in the physical scale/subscale scores of the three patient-reported instruments, and EDSS scores for improved patients (SHTQ score >3). AUC is the area under the curve (95% confidence interval).

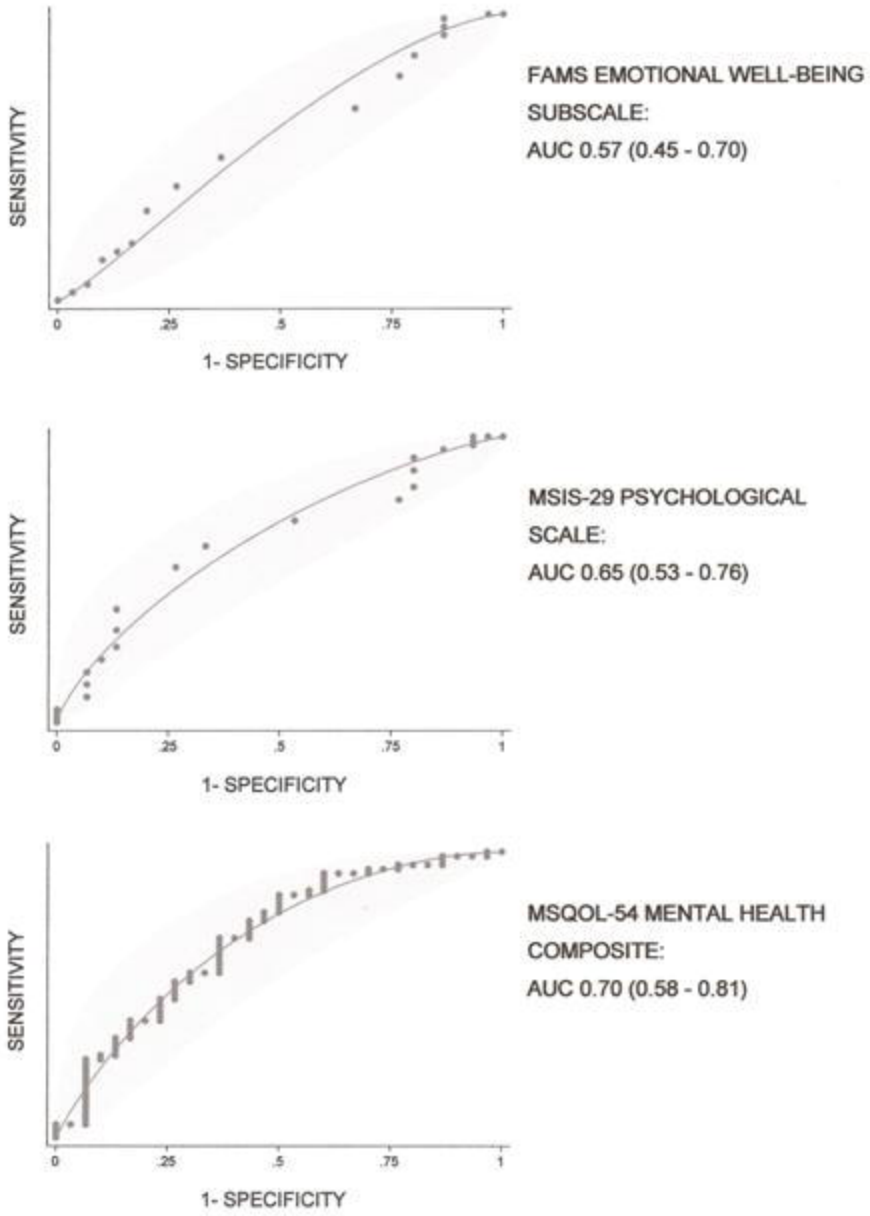

Figure 4 Plots of the receiver operating characteristics curves and 95% confidence intervals for changes in the psychological scale/subscale scores of the three patient-reported instruments for improved patients (SHTQ score > 3). AUC is the area under the curve (95% confidence interval).