Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis

J A Freeman, J C Hobart, D W Langdon, A J Thompson

Abstract

***Objectives*—Understanding the properties of an outcome measure is essential in choosing the appropriate instrument and interpreting the information it generates. The MOS 36 item short form health survey questionnaire (SF-36) is widely acknowledged as the gold standard generic measure of health status; few studies however have evaluated its use for clinical trials in multiple sclerosis. Its clinical appropriateness, internal consistency reliability, validity, and responsiveness was investigated across a broad range of patients with multiple sclerosis.**

***Methods*—A prospective study in which 150 adults with clinically definite multiple sclerosis completed a battery of questionnaires evaluating generic health status, disability, handicap, and emotional well-being. Of these, 44 patients undergoing inpatient rehabilitation completed the questionnaires before and after intervention to evaluate responsiveness.**

***Results*—Score distributions demonstrated significant floor and ceiling effects in four of the eight dimensions which were particularly marked when patient selection was restricted to a narrow band of disease severity (as is the case in most clinical trials). Internal consistency exceeded the standard for group comparisons for all dimensions. Convergent and discriminant construct validity was supported by the direction, magnitude, and pattern of correlations with other health measures. In comparison with instruments measuring associated constructs, the responsiveness of the SF-36 was poor in evaluating change in moderate to severely disabled patients participating in a programme of inpatient rehabilitation.**

***Conclusions*—The SF-36 has some limitations as an outcome measure in multiple sclerosis. The results highlight the need for all instruments to be examined in the specific sample population under question and for the specific research question being investigated. In multiple sclerosis clinical trials, the SF-36 should be supplemented with other relevant measures.**

(*J Neurol Neurosurg Psychiatry* 2000;**68**:150–156)

Keywords: multiple sclerosis; SF-36; quality of life

Institute of Neurology, Department of Clinical Neurology, Queen Square, London WC1 N3BG, UK
J A Freeman
J C Hobart
D W Langdon
A J Thompson

Correspondence to:
Dr JA Freeman, Institute of Neurology, Department of Clinical Neurology, Queen Square, London WC1 N3BG, UK
email FreemanJR@compuserve.com

Received 2 June 1999 and in revised form 4 October 1999
Accepted 21 October 1999

Numerous clinical trials have been undertaken in the past decade to determine the effectiveness of a range of interventions in multiple sclerosis. Traditionally these trials have evaluated outcome on the basis of clinical end points (for example, relapse rate) and physiological parameters (for example, lesion load on MRI[1]). In recent years there has been a gradual broadening of the outcomes measured to include aspects of health status.[2 3] Alongside this advance, an increasing number of new measures of health status have been developed.[4–6] Unfortunately only preliminary information is available about many of these measures, particularly on their use in clinical trials. As a consequence researchers have found that they are faced with greater choice but limited information on which to base their selection.[7]

It is widely agreed that the choice of outcome measure(s) is crucial to the successful design of a clinical trial.[8] An informed decision is reliant on knowledge of the scientific (reliability, validity, and responsiveness) and clinical properties (feasibility, appropriateness to the study sample, respondent burden) of available measures.[9] Understanding the purpose of the study is also a key consideration as different questions may require different measures. Will the instrument be used to describe specific characteristics of the population? Will it make comparisons with other samples? Will it evaluate the effectiveness of an intervention? Such information is essential in choosing the most appropriate instrument and interpreting the results generated in a meaningful way.

Two different approaches to the measurement of health status are the generic and the disease specific models. The generic model seeks to assess basic health values thought to be relevant to health status regardless of disease, treatment, or age group.[10] By contrast, the disease specific model is not concerned with establishing universal standards but aims instead to reflect factors relevant to the person with a specific disease.

The SF-36 is generally considered as the gold standard generic measure of health status.[4] It is available in several languages and has been adopted and disseminated worldwide. A standard United Kingdom version has been developed[11] and norms determined for the healthy population both in the United States[12] and United Kingdom.[13] Although proved to be reliable and valid in a range of patient groups[12] relatively few studies have investigated its use in multiple sclerosis. Most have examined its value in describing the impact of multiple sclerosis on quality of life; often comparing their findings to other patient groups and the general

population.[14–17] Few have used the SF-36 as an outcome measure in clinical trials of multiple sclerosis.[18 19]

In 1996 we published the results of a cross sectional study which piloted the use of the SF-36 in patients with multiple sclerosis in a rehabilitation unit.[15] Our results showed that the SF-36 demonstrated marked floor effects in some dimensions in this group of moderately to severely disabled patients. We suggested that this was likely to limit its potential responsiveness in evaluating any changes that may occur as a result of interventions. We concluded that a systematic evaluation of the SF-36 in a broader range of patients with multiple sclerosis was necessary. This study investigated the appropriateness, reliability, and validity of the SF-36 in a broad range of patients with multiple sclerosis from the newly diagnosed to those in the advanced stages of the disease. It determined its responsiveness in a subgroup of patients undergoing a programme of inpatient rehabilitation.

Methods

SAMPLING

One hundred and fifty patients with a diagnosis of clinically definite multiple sclerosis[20] participated in this ethically approved prospective study. Consecutive patients were recruited from three different sources within a healthcare setting: a weekly outpatient assessment clinic, an inpatient neurorehabilitation unit, and those admitted under a single consultant (AJT) to acute hospital wards. Patients were excluded if they were cognitively impaired such that they were unable to reliably complete the questionnaires; had other diseases such as rheumatoid arthritis which may have influenced their health status; or were non-English speaking.

STRATIFICATION PROCESS

Data collected solely from one particular setting is often biased. For example, a larger percentage of severely disabled patients are more likely within the acute hospital setting than among those attending a follow up outpatient appointment. To ensure a more even spread of disability within our sample we undertook a stratification procedure to ensure that it was comprised of equal numbers of patients across the entire range of disease severity. This process involved a neurological registrar assessing all patients with Kurtzke's functional systems scale and expanded disability status scale (EDSS),[21] and then categorising them into one of three groups[22]: mild (EDSS 0–4.5), moderate (EDSS 5.0–6.5), or severe (EDSS 7.0–9.5). Consecutive patients were recruited until there were 50 patients in each category.

TEST BATTERY AND METHODS OF ADMINISTRATION

Demographic details were collected by interview and diagnostic details derived from the medical records. All patients were rated for level of disease severity as described above. Level of disability was scored by interview using the functional independence measure (FIM)[23] administered in accordance with published guidelines. Patients also completed a battery of self reported questionnaires measuring a range of health constructs including generic health status, handicap, emotional wellbeing, and a 0–10 point global rating scale of overall quality of life. Whenever possible this was undertaken independently but when necessary (for example, with visual disturbance, difficulty writing) physical assistance was provided by the researcher. No assistance was given in interpreting the questionnaires.

One hundred and six of the patients were assessed at a single time point. The other 44 subjects, who were all rehabilitation inpatients, were assessed at two time points (admission and discharge) to evaluate responsiveness.

THE HEALTH STATUS MEASURES

The anglicised version of the SF-36[13] was used. This 36-item generic health status questionnaire includes eight multi-item measures of functioning and wellbeing: physical function (PF-10 items), role limitations due to physical (RLP-four items) or emotional (RLM-three items) health problems, social function (SF-two items), emotional wellbeing (MH-five items), bodily pain (BP-two items), energy and fatigue (EV-four items), and general health perceptions (HP-five items). All items are coded, summed, and transformed onto a scale of 0–100 (0=worst health, 100=optimal health).[12] In addition, scores on these eight dimensions can be reduced to two summary scores, a physical (PCS) and a mental component (MCS), by means of principal components analyses.[24]

A global 0–10 point scale was used to rate overall quality of life (QoL).[25]

Instruments measuring related health constructs

Information was gathered from the following instruments to enable comparison with some of the SF-36 dimensions.

(1) Functional performance was assessed, by patient interview, using the FIM motor domain. This 13 item, seven level scale measures aspects of daily function in four subscales: self care, sphincter control, transfers, and locomotion. The total score range is 13–91 with higher scores indicating greater levels of independence.

(2) Handicap was assessed using the London handicap scale (LHS).[26] This six item, six level scale assesses the disadvantage experienced by the individual patients in the dimensions of mobility, physical independence, occupation, social integration, orientation, and economic self sufficiency. The total score range is 0–100 with higher scores indicating the least level of disadvantage.

(3) Emotional status was assessed using the 28 item general health questionnaire (GHQ).[27] This version has four subscales that measure disturbances in the areas of somatic complaints, anxiety, social dysfunction, and depression. The total score range is 0–28, with higher scores indicating greater levels of emotional disturbance.

Each of these instruments has been used in various multiple sclerosis populations and has been shown to be valid, reliable, and responsive within the rehabilitation[3 19] and hospital setting.[28]

THE REHABILITATION PROGRAMME

The inpatient rehabilitation programme consisted of a structured, goal oriented, multidisciplinary programme specifically aimed at considering the individual needs of the patient.[3] This typically included efforts to improve functional independence, mobility, bladder and bowel function, and communication. Advice and education regarding work and leisure pursuits, tone management, fatigue management, and strategies to compensate for memory dysfunction were also regular components of this programme.

DATA ANALYSES

Statistical analysis was performed using SPSS.[29] Descriptive statistics were used to describe demographic and disease characteristics of the sample.

Appropriateness

Appropriateness has been used to define whether the range of the construct measured within the study sample is similar to the range covered by the measurement instrument.[30] In essence this reflects how relevant the instrument is to the population being examined. This was assessed by examining the scale score distributions (range, mean, SD, floor (minimum), and ceiling (maximum) scores) of the eight dimensions and the two summary scales of the SF-36, as well as for each of the other measures.

Reliability

One aspect of reliability, internal consistency, was calculated by Cronbach's α statistic.[31] Alpha coefficients exceeding 0.7 are considered adequate for group comparison.[12]

Construct validity

Construct validity is the process used to establish the validity of a measurement instrument when no criterion or universe of content is accepted as entirely adequate to define the attribute being measured.[31] It is determined by examining the extent to which empirical data support hypotheses concerning the construct the instrument is purported to measure. We examined the data for evidence of:

(1) Convergent validity—by determining the relation between dimensions on the SF-36 and instruments measuring similar constructs. Pearsons product-moment correlations were examined for: SF-36 emotional wellbeing dimensions with the GHQ; SF-36 physical dimensions with the FIM and the EDSS; and SF-36 social and role dimensions with the LHS. To provide evidence of convergent validity we would expect, for example, to see substantial correlations between the SF-36 physical dimensions, the EDSS, and the FIM; and likewise between the SF-36 emotional dimensions and the GHQ.

(2) Discriminant construct validity—by determining the relation between dimensions on the SF-36 and instruments measuring different constructs. Pearson product-moment correlations were examined between the physical and emotional wellbeing dimensions of the SF-36; the FIM and the SF-36 emotional wellbeing dimensions; and the GHQ and the SF-36 physical dimensions. To provide evidence of discriminant validity we would expect, for example, to see weak correlations between the SF-36 emotional dimensions and the FIM; and between the SF-36 mental and physical summary scales.

(3) Group differences construct validity—by examining the differences in SF-36 scores between different groups. We investigated the ability of the mental and physical summary scales to distinguish between different levels of disease severity in multiple sclerosis by using a one way analysis of variance (ANOVA) with post-hoc comparison, adjusting for multiple comparisons using Bonferroni's test, with $\alpha=0.05$. To provide evidence of group differences construct validity we would expect, for example, that patients categorised into the severe group would report lower scores on both of the summary scales than patients in the mild group.

(4) Hypothesis testing—by examining whether the results produced are consistent with theoretical expectation. The following hypotheses were tested using independent *t* tests, with $\alpha=0.05$: (*a*) patients requiring carer assistance will report lower scores in the SF-36 physical function dimensions than those who are independent in their daily care; (*b*) patients with relapsing-remitting multiple sclerosis will report higher scores in the physical summary scale of the SF-36 than those with secondary progressive multiple sclerosis; (*c*) patients scoring $\geqslant$5.0 points on the GHQ (indicating emotional distress as defined by Dalos *et al*[27]) will report lower scores on the SF-36 mental summary scale than those scoring <5.0 points.

Responsiveness

Responsiveness is the ability of the instrument to measure clinically important change over time.[9] This was examined in a subgroup of 44 patients admitted for a short period of inpatient rehabilitation. This intervention has

Table 1 Demographic and diagnostic characteristics

	Total sample n=149	*Mild EDSS 0–4.5 n=47*	*Moderate EDSS 5.0–6.5 n=51*	*Severe EDSS 7.0–9.5 n=51*
% Female	68	60	67	76
Age (y) (mean (SD) (range))	44.6 (10.8) (24–78)	41.4 (10.2) (24–68)	45.4 (10.7) (24–78)	47.1 (10.6) (24–73)
Disease pattern:				
Secondary progressive (%)	50	2	71	76
Primary progressive (%)	11	13	8	12
Relapsing-remitting (%)	33	66	21	12
Benign (%)	6	19	0	0
Years since first symptoms (mean (SD) (range))	14.6 (8.9) (0.6–43)	12.3 (9.3) (2–41)	15.7 (8.3) (0.6–40)	16.0 (8.5) (1.4–43)
Years since diagnosis (mean (SD) (range))	10.2 (7.6) (0.1–38)	7.5 (7.8) (0.1–30)	10.8 (7.2) (0.2–38)	12.1 (7.2) (1.2–33)
Place of assessment:				
Outpatient clinic (%)	57	98	49	29
Inpatient ward (%)	7	2	4	14
Rehabilitation unit (%)	36	0	47	57

Table 2 SF-36 scores for three different multiple sclerosis (MS) populations

SF-36 dimensions	*MS this study n=149 mean (SD)*	*MS America n=171 mean (SD)*	*MS Canada n=97 mean (SD)*
Physical function	27.9 (28.1)	36.2 (32.3)	32.5 (27.4)
Emotional wellbeing	66.7 (19.8)	65.3 (19.9)	68.2 (20.8)
Role limitations physical	29.1 (38.9)	33.4 (39.3)	29.1 (35.8)
Role limitations emotional	59.1 (43.2)	61.3 (42.3)	55.7 (43.0)
Social function	49.7 (26.6)	58.3 (27.2)	60.9 (27.8)
Bodily pain	62.8 (29.6)	71.8 (26.8)	67.3 (27.7)
Energy and vitality	43.6 (22.0)	39.0 (21.6)	37.8 (21.7)
Health perceptions	50.0 (22.8)	43.0 (28.7)	52.0 (22.4)

American multiple sclerosis population compiled from Vickrey *et al* 1997; Canadian multiple sclerosis population compiled from Brunet *et al* 1996.

been previously evaluated and was shown to be effective in improving aspects of health status in people with multiple sclerosis in both the short[3] and long term.[19] In each of these outcomes patients change scores between admission and discharge were determined, and effect sizes calculated (where effect size=mean change/SD of the initial distribution of scores).[32] The criterion proposed by Cohen[33] was used to interpret the effect size, where 0.2 is small, 0.5 is moderate, and 0.8 or greater is large. Paired *t* tests were used to determine the statistical significance of these change scores.

Results

SAMPLE CHARACTERISTICS

Of the 150 patients entered into the study, one did not complete the battery of questionnaires and was excluded from the analyses. Of the remaining 149 people there were no missing data for any items. Table 1 presents the demographic and diagnostic characteristics of the study sample, of which 70% were married, 33% were employed, and 49% required assistance with their daily care. Table 2 shows the mean SF-36 scores for our sample alongside those of two other multiple sclerosis populations.

Table 3 Baseline SF-36 score distributions

SF-36 (scale range 0–100)	*Total sample (n=149)*	*Mild (n=47)*	*Moderate (n=51)*	*Severe (n=51)*
Dimensions				
Physical function:				
Sample range	0–100	5–100	0–65	0–20
Mean (SD)	27.9 (28.1)	58.7 (24.6)	22.9 (16.1)	4.4 (6.0)
Floor/ceiling effect (%)	21/2	0/6	10/0	52/0
Emotional wellbeing:				
Sample range	0–100	24–100	4–100	0–92
Mean (SD)	66.7 (19.8)	66.0 (18.3)	70.4 (22.4)	63.0 (18.0)
Floor/ceiling effect (%)	1/3	0/2	0/6	2/0
Role limitations physical:				
Sample range	0–100	0–100	0–100	0–100
Mean (SD)	29.1 (38.9)	50 (41.7)	28.5(37.1)	10.5 (27.2)
Floor/ceiling effect (%)	53/18	25/34	50/16	84/6
Role limitations emotional:				
Sample range	0–100	0–100	0–100	0–100
Mean (SD)	59.1 (43.2)	58.9 (41.8)	62.1 (42.2)	55.3 (45.9)
Floor/ceiling effect (%)	27/47	23/45	24/49	36/46
Social function:				
Sample range	0–88.9	0–88.9	0–88.9	0–88.9
Mean (SD)	49.7 (26.6)	58.9 (24.3)	51.8 (22.4)	38.9 (29.4)
Floor/ceiling effect (%)	8/0	4/0	2/0	18/0
Bodily pain:				
Sample range	0–100	12–100	10–100	0–100
Mean (SD)	62.8 (29.6)	65.5 (25.5)	64.5 (30.7)	58.3 (32.3)
Floor/ceiling effect (%)	2/28	0/25	0/33	6/26
Vitality and energy:				
Sample range	0–100	0–100	10–95	0–90
Mean (SD)	43.6 (22.0)	45.1 (22.2)	44.3 (21.9)	41.3 (22.4)
Floor/ceiling effect (%)	1/1	2/2	0/0	2/0
General health perceptions:				
Sample range	5–100	20–90	5–87	5–100
Mean (SD)	50.0 (22.8)	56.7 (18.9)	45.9 (20.8)	48.1 (27.1)
Floor/ceiling effect (%)	0/1	0/0	0/0	0/2
SF-36 PCS:				
Sample range	9.2–56.9	24.3 –56.9	9.2–47.3	9.7–40.2
Mean (SD)	31.8 (10.6)	40.9 (9.5)	29.9 (8.7)	25.3 (6.8)
Floor/ceiling effect (%)	0/0	0/0	0/0	0/0
SF-36 MCS:				
Sample range	15.9–73.0	16.9–62.6	22.3–73.0	15.9–72.4
Mean (SD)	47.1 (12.2)	44.2 (11.8)	49.4 (12.6)	47.3 (11.6)
Floor/ceiling effect (%)	0/0	0/0	0/0	0/0

PCS=Physical component summary scale; MCS=emotional component summary scale; Mild=EDSS 0–4.5; Moderate=EDSS 5.0–6.5; Severe=EDSS 7.0–9.5.

APPROPRIATENESS

Table 3 presents the score distributions for the SF-36 dimensions and summary scales. In the total sample, scores in all dimensions span virtually the entire range; however, floor effects in three dimensions (physical function, physical and emotional role limitations) and ceiling effects in two dimensions (emotional role limitations and pain) exceed the recommended criteria of 20%.[34] When patients are subdivided into groups according to disease severity the distribution of scores within each subgroup, in some cases, alters markedly. For example : (*a*) the physical function scores span only the bottom 20% of the range for the severe group; (*b*) the means fall substantially outside the midpoint of the scale for physical function and physical role limitations in the moderate and severe groups (mean PF moderate=22.9, severe=4.4; mean RLP moderate=28.5, severe=10.5); (*c*) floor effects increase markedly for physical function and role limitations (both emotional and physical), particularly in the severe group. The lowest possible score is reported by 84% of severe patients for physical role limitations and 36% of patients for emotional role limitations.

Table 3 demonstrates that the differences in score distributions between the total sample and the subgroups were less marked in the summary scales. Of importance, no ceiling or floor effects were present.

Table 4 presents the score distributions for instruments measuring related health constructs. In the total sample scores on the FIM, LHS, and the global rating scale of QoL span virtually the entire scale range; the mean scores were near the midpoint; and the floor and ceiling effects were minimal. This indicated that the scales were appropriate for the total study sample. When patients were subgrouped according to EDSS score the appropriateness of these instruments, while not ideal, remained satisfactory. Although the scores were restricted to a smaller range of the available scale, the floor and ceiling effects remained well below the recommended criteria of 20%.[34] By contrast, in both the total sample and each of the subgroups, the mean scores on the GHQ fell below the midpoint of the scale and the ceiling effects were above the recommended upper limit.

RELIABILITY

Internal consistency reliability for each of the eight dimensions and the component summary scales of the SF-36 was high with α coefficients ranging between 0.77 to 0.94.

Table 4 Baseline score distributions in instruments measuring a range of health related constructs

Measurement instrument (scale range)	*Total sample (n=149)*	*Mild (n=47)*	*Moderate (n=51)*	*Severe (n=51)*
EDSS (0–10):				
Sample range	1.0–9.0	1.0–4.5	5.0–6.5	7.0–9.0
Median	6.5	3.0	6.5	8.0
Mean (SD)	5.7 (2.1)	2.9 (0.8)	6.1	7.8 (6.5)
Floor/ceiling effect (%)	0/0	0/0	0/0	0/0
FIM motor (13–91):				
Sample range	13–91	74–91	43–90	13–81
Mean (SD)	69.7 (20.6)	87.5 (3.5)	75.5 (8.7)	47.7 (18.9)
Floor/ceiling effect (%)	1/5	0/17	0/0	2/0
LHS (0–100):				
Sample range	35–100	48–100	43–91.9	35–93.7
Mean (SD)	69.5 (13.3)	79.2 (13.2)	68.2 (10.1)	61.8 (10.7)
Floor/ceiling effect (%)	0/0	0/8.5	0/0	0/0
GHQ (0–28):				
Sample range	0–28	0–24	0–23	0–28
Mean (SD)	6.2 (6.5)	5.6 (6.8)	4.9 (4.9)	8.0 (7.2)
Floor/ceiling effect (%)	1/25	0/32	0/23.5	2/19.6
0–10 point global rating scale of QOL:				
Sample range	0–100	23.3–100	5–85	0–81.6
Mean (SD)	55.4 (19.6)	64.0 (19.9)	55.9 (15.6)	46.6 (19.6)
Floor/ceiling effect (%)	1/1	0/2	0/0	4/0

Mild=EDSS 0–4.5; Moderate=EDSS 5.0–6.5; Severe=EDSS 7.0–9.5.

CONVERGENT AND DISCRIMINANT CONSTRUCT VALIDITY

Intercorrelations between the SF-36 dimensions

Table 5 reports intercorrelations between SF-36 dimensions. Importantly, none of the correlations were strong (r=0.09–0.61) demonstrating that each dimension was measuring a related but distinct construct. As predicted, related dimensions were more strongly associated than less related dimensions. For example, physical function showed a stronger correlation with physical role limitations (r=0.57) than with emotional wellbeing (r=0.09) or emotional role limitations (r=0.14). Similarly emotional wellbeing showed a stronger correlation with emotional role limitations (r=0.54) than pain (r=0.29). Interestingly, emotional wellbeing showed a much stronger correlation with energy and vitality (r=0.61) than did physical function (r=0.18).

Correlations between SF-36 dimensions and instruments measuring related health constructs

As predicted, associations between SF-36 dimensions and instruments measuring related health constructs were strongest between those measuring similar concepts. For example, physical function correlated strongly with the FIM (r=0.68) and the EDSS (r=−0.82); and emotional wellbeing correlated substantially with the GHQ (r=−0.59). By contrast, associations were weak between instruments measuring unrelated constructs. For example, emotional role limitations was only weakly associated with the FIM (r=0.04), and pain was only weakly associated with the EDSS (r=−0.07). It is notable that the social function dimension correlated more strongly with scales measuring emotional constructs (for example, GHQ r=−0.56) than physical constructs (for example, EDSS r=−0.29; FIM r=0.34).

Table 5 Associations between SF- 36 dimensions (Pearson's product-moment correlations)

	PF	*MH*	*RLP*	*RLM*	*SF*	*BP*	*EV*
MH	0.09						
RLP	0.57	0.23					
RLM	0.14	0.54	0.38				
SF	0.35	0.45	0.41	0.29			
BP	0.22	0.29	0.29	0.22	0.29		
EV	0.18	0.61	0.34	0.46	0.50	0.27	
HP	0.26	0.37	0.34	0.26	0.41	0.18	0.50

PF=Physical function; RLP=physical role limitations; RLM=emotional role limitations; SF= social function; MH=emotional wellbeing; BP= bodily pain; EV=energy and fatigue; HP= general health perceptions.

Group difference construct validity—As expected, statistically significant differences between the patient subgroups occurred in three SF-36 dimensions (social function, physical function, and physical role limitations; p<0.05–0.0001). This finding demonstrates the ability of these dimensions to discriminate between different levels of disease severity. Significant differences were also demonstrated between all subgroups for the physical summary scale (p<0.001), but only between the mild and the moderate group in the mental summary scale (p<0.03).

Hypothesis testing—As predicted: (*a*) patients requiring carer assistance reported lower scores in the physical role limitations dimension than those who are independent (p<0.0001, mean scores=13.5 and 43.7 respectively); (*b*) patients with relapsing-remitting multiple sclerosis reported higher scores in the physical summary scale than those with secondary progressive multiple sclerosis (p<0.0001, mean scores=35.7 and 27.4 respectively); (*c*) patients scoring ⩾5.0 points on the GHQ reported lower scores on the mental summary scale than those scoring <5.0 points (p<0.0001, mean scores=40.4 and 52.2 respectively).

RESPONSIVENESS

Forty four patients participated in inpatient rehabilitation for an average of 20 days ((SD 6) range 13–39). Effect sizes for the SF-36 dimensions ranged from negligible to small (effect sizes 0.01–0.30). The dimensions demonstrating the largest effect size were the emotional role limitations (effect size 0.27) and pain (effect size 0.30). Of the eight dimensions, only pain (p=0.006) and physical function (p=0.01) demonstrated a statistically significant change in scores between admission and discharge. By contrast effect sizes on the FIM, LHS, and GHQ were all moderate in magnitude (effect size 0.56, 0.58, and 0.51 respectively) and statistically significant differences were demonstrated between scores on admission and discharge for each of these measures (p<0.002).

Discussion

This study provides information about a widely used generic measure of health status—the SF-36. The SF-36 was constructed to compare functional health and wellbeing across patient and general populations, and to evaluate and compare the benefits of alternative treatments.[24] The focus of this study was to examine its performance as an outcome measure in multiple sclerosis.

The generalisability of our results is supported by the fact that the demographic and diagnostic characteristics of our sample population are typical of those described in the literature.[35] Furthermore, as demonstrated in

table 2, the distribution of SF-36 scores is very similar to the results of previous multiple sclerosis studies[16 17 25] suggesting that our sample is representative of the general multiple sclerosis population.

The test-retest reliability of the SF-36 was not investigated in this study but the results of others, both in the United Kingdom[11] and the United States,[12] report excellent results at 2 weeks. In agreement with some other studies, our results demonstrate that the internal consistency for all dimensions of the SF-36 exceeds the 0.7 standard for group comparisons.[12] Similarly, convergent and discriminant construct validity are supported by the direction, magnitude, and pattern of correlations with other health measures. Further evidence for construct validity has been provided by support for the clinical hypotheses tested. These data support the internal consistency reliability and validity of the SF-36 as a measure of health status in multiple sclerosis. Consequently it would seem reasonable to choose the SF-36 as an outcome measure in clinical trials evaluating the effectiveness of interventions in multiple sclerosis.

When the data are examined in more detail, however, some limitations of this measure become apparent. For instance the large floor and ceiling effects in four of the eight dimensions indicate that the range of health status measured is unlikely to represent the range experienced by this population, and demonstrate limitations in the ability of the SF-36 to discriminate between individual patients in these dimensions. It is notable that the floor and ceiling effects do not simply apply to patients at the extremes of the disease severity range; the moderate group also exhibit significant floor effects in three dimensions. The data also show a polarisation of responses in the role limitations dimensions. This is perhaps not surprising when the dichotomous format of the questionnaire items is considered. For an example, refer to fig 1, which contains the emotional role limitations question.

These concerns as to the appropriateness of the SF-36 in multiple sclerosis are heightened when the population is subdivided into groups according to disease severity. This is very important as the selection criteria of most clinical trials will inevitably narrow the range of disease severity of the study sample, sometimes markedly (for example EDSS 1.0<3.5[36]; EDSS 3.0– 6.5[2]; EDSS<6.5[37]). These results highlight the importance of examining the appropriateness of an instrument for the specific population under investigation. Even though an instrument may prove to be appropriate for one group of patients this may not necessarily be the case for a different group, even within the same medical condition.

No floor or ceiling effects occur in the SF-36 mental and physical summary scales, suggesting that these scales may be more appropriate than the individual dimensions for discriminating between individual patients at a single point in time. Additionally they have the advantage of reducing the number of statistical comparisons required in the analysis of results, thereby reducing the role of chance in testing experimental hypotheses.[24] A disadvantage, however, is that it is impossible to interpret precisely where any changes have occurred; a common feature of all multidimensional instruments.

During the past 4 weeks have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? (circle one number on each line):

	Yes	*No*
(*a*) Cut down on the amount of time you spent on work or other activities	1	2
(*b*) Accomplished less than you would like	1	2
(*c*) Didn't do work or other activities as carefully as usual	1	2

Figure 1 Sample item from the SF-36 emotional role limitations dimension

The responsiveness of an instrument is of key importance in outcome studies. If the instrument is unable to detect change in health, an intervention that improves health status may show no apparent difference between treated and untreated patients. Unfortunately, this property is often overlooked and information about the responsiveness of the SF-36 in multiple sclerosis trials is scarce. The negligible to small effect sizes demonstrated by the SF-36 show the responsiveness of the SF-36 to be poor in evaluating the effectiveness of inpatient rehabilitation in people with moderate to severe disability. Although some may suggest that this is because little or no change has occurred, the moderate effect size results of the FIM (measuring physical function), the GHQ (measuring emotional health), and the LHS (measuring handicap) show that change has indeed occurred, at least in these select areas. Furthermore statistically significant changes were demonstrated between change scores on each of these three measures, but in only two of the dimensions of the SF-36. The poor responsiveness of the SF-36 may, in part, be explained by the fact that it measures broad issues of both function and wellbeing, which, taken together may not give a clear effect. By contrast, the FIM, GHQ, and LHS measure more specific health constructs. We would also suggest that the clustering of scores at either end(s) of the scale, found in half of the SF-36 dimensions, means that the range of the scale is too limited to enable small but possibly clinically significant changes to be recorded; thereby limiting responsiveness. It is stressed, however, that the responsiveness data in this study is restricted to patients with moderate to severe disability undergoing rehabilitation and that the SF-36 has therefore not been assessed in a population representative of the patients included in most multiple sclerosis trials. This is a limitation of this study.

Different approaches to consider some of the limitations of generic measures have been used in recent years. For example, the development of disease specific measures for multiple sclerosis has been undertaken either by adapting current measures (for example, the functional assessment measure[38] or the multiple sclerosis QoL-54[4]); by gathering together a wide range of symptom specific measures (for

example, the QoL inventory[39]); or by identifying key areas and then weighting them according to how important the patient thinks these areas are to their lifestyle (for example, the disability and impact profile[40]). All of these measures are in the early stages of evaluation.

Conclusions

Understanding the properties of an outcome measure is essential when choosing the most appropriate instrument for a study and interpreting the information it generates. The SF-36 is widely acknowledged as the gold standard generic measure of health status. It is being increasingly used as an outcome measure in a range of clinical trials to determine the effectiveness of interventions. The results of this study highlight some limitations of the SF-36 for this purpose. The marked floor and ceiling effects demonstrated in half of the dimensions, and across the range of disease severity, indicate a limited ability to discriminate between patients with multiple sclerosis at a single point in time. The poor responsiveness of the dimension scores suggest that it is limited in detecting change over time in people with moderate to severe disability. These results highlight the need for "generic" measures to be tested for specific populations and for specific purposes. We suggest that trials evaluating health status in multiple sclerosis should supplement the use of the SF-36 with other relevant and scientifically sound instruments to maximise the validity of health measurement.

1 Paty DW, McFarland H. Magnetic resonance techniques to monitor the long term evolution of multiple sclerosis pathology and to monitor definitive clinical trials. *J Neurol Neurosurg Psychiatry* 1998;**64**(suppl 1):S47–51.
2 European Study Group on Interferon 1-b in Secondary Progressive Multiple Sclerosis. Placebo-controlled multicentre randomised trial of interferon 1-b in treatment of secondary progressive multiple sclerosis. *Lancet* 1998;**352**: 1491–7.
3 Freeman JA, Langdon DW, Hobart JC, *et al*. The impact of inpatient rehabilitation on progressive multiple sclerosis. *Ann Neurol* 1997;**42**:236–44.
4 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.
5 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.
6 Sharrack B, Hughes, Soudain S, *et al*. The psychometric properties of clinical rating scales used in multiple sclerosis. *Brain* 1999;**122**:141–9.
7 Rothwell PM. Quality of life in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 1998;**65**:433.
8 Whitaker JN, McFarland HF, Rudge P, *et al*. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. *Multiple Sclerosis* 1995;**1**:37–47.
9 Hobart JC, Lamping DL, Thompson AJ. Evaluating neurological outcome measures: the bare essentials. *J Neurol Neurosurg Psychiatry* 1996;**60**:127–30.
10 Patrick DL, Deyo RA. Generic and disease specific measures in assessing health status and quality of life. *Medical Care* 1989;**27**(suppl 3):S17–32.
11 Brazier JE, Harper R, Jones NMB, *et al*. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992;**305**:160–4.
12 Ware JE, Snow KK, Kosinski M, *et al*. *SF-36 Health survey: manual and interpretation guide*. Boston, Massachusetts: The Health Institute, New England Medical Centre; 1993.
13 Jenkinson C, Coulter A, Wright L. Short form 36 (SF-36) health survey questionnaire: normative data for adults of working age. *B MJ* 1993;**306**:1437–44.
14 Hermann BP, Vickrey B, Hays RD, *et al*. A comparison of health-related quality of life in patients with epilepsy, diabetes and multiple sclerosis. *Epilepsy Res* 1996;**25**:113–18.
15 Freeman JA, Langdon DW, Hobart JC, *et al*. Health-related quality of life in people with multiple sclerosis undergoing inpatient rehabilitation. *Journal of Neurological Rehabilitation* 1996;**10**:185–94.
16 Brunet DG, Hopman WH, Singer MA, *et al*. Measurement of health-related quality of life in multiple sclerosis patients. *Can J Neurol Sci* 1996;**23**:99–103.
17 The Canadian Burden of Illness Study Group. Burden of illness of multiple sclerosis: part 2: quality of life. *Can J Neurol Sci* 1998;**25**:31–8.
18 Di Fabio RP, Choi T, Soderberg J, *et al*. Health related quality of life for patients with progressive multiple sclerosis: influence of rehabilitation. *Phys Ther* 1997;77:1704–16.
19 Freeman JA, Langdon DW, Hobart JC, *et al*. Inpatient rehabilitation in multiple sclerosis. Do the benefits carry over into the community ? *Neurology* 1999;**52**:50–6.
20 Poser CM, Paty DW, Scheinberg L, *et al*. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol* 1983;**13**:227–31.
21 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983;**33**:1444–52.
22 British Society of Rehabilitation Medicine. *Multiple sclerosis* (Working Party Report) London: Royal College of Physicians, 1993.
23 Granger CV, Cotter AC, Hamilton BB, *et al*. Functional assessment scales: a study of persons with multiple sclerosis. *Arch Phys Med Rehabil* 1990;**71**:870–5.
24 Ware JE, Kosinski M, Keller SD. *SF-36 Physical and mental health summary scales: a user's manual, 2nd ed*. Boston, Massachusetts: The Health Institute, New England Medical Centre, 1994.
25 Vickrey BG, Hays RD, Genovese BJ, *et al*. Comparison of a generic to disease-targeted health-related quality of life measures for multiple sclerosis. *J Clin Epidemiol* 1997;**50**: 557–69.
26 Harwood RH, Ebrahim S. *Manual of the London handicap scale*. University of Nottingham: Department of Health Care of the Elderly; 1995.
27 Dalos NP, Rabins PV, Brooks BP, *et al*. Disease activity and emotional state in multiple sclerosis. *Ann Neurol* 1983;**13**: 573–7.
28 Lykouras L, Adrachta D, Kalfakis N, *et al*. GHQ as an aid to detect mental disorders in neurological inpatients. *Acta Psychiatr Scand* 1996;**9**:3212–6.
29 SPSS. *SPSS for Windows Version 7.5*. Seattle: SPSS, 1996.
30 Van der Putten JMF, Hobart JC, Freeman JA, *et al*. Measuring change in disability following inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the functional independence measure. *J Neurol Neurosurg Psychiatry* 1999;**66**:480–4.
31 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
32 Kazis L, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**:S178–88.
33 Cohen J. *Statistical power analysis for the behavioural sciences*. London: Academic Press, 1977.
34 Holmes W, Shea J. Performance of a new HIV/AIDS-targeted quality of life (HAT-QOL) instrument in asymptomatic seropositive individuals. *Qual Life Res* 1997;**6**:561–71.
35 Weinshenker BG. Natural history of multiple sclerosis. *Ann Neurol* 1994;**36**:S6–11.
36 Jacobs LD, Cookfair DL, Rudick RA, *et al*. Intramuscular interferon β-Ia for disease progression in relapsing multiple sclerosis. *Ann Neurol* 1996;**39**:285–94.
37 IFNB Multiple Sclerosis Study Group and the University of British Columbia MS/MRI Analysis Group. Interferon β-1b in the treatment of MS: final outcome of the randomised controlled trial. *Neurology* 1995;**45**:1277–85.
38 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.
39 La Rocca N, Ritvo PG, Miller DM, *et al*. Quality of life assessment in multiple sclerosis clinical trials. In: Goodkin DE, Rudick RA, eds. *Multiple sclerosis advances in clinical trial design, treatment and future perspectives*. London: Springer Verlag, 1996:145–61.
40 Lankhorst G, Jelles F, Smits, *et al*. Quality of life in multiple sclerosis: the disability and impact profile (DIP). *J Neurol* 1996;**243**:469–74.