J Neurol (2008) 255:246–254
DOI 10.1007/s00415-008-0708-y

ORIGINAL COMMUNICATION

Peter Hagell
Anna Lena Törnqvist
Jeremy Hobart

Testing the SF-36 in Parkinson's disease

Implications for reporting rating scale data

Received: 27 April 2007
Received in revised form: 28 June 2007
Accepted: 19 July 2007
Published online: 22 January 2008

P. Hagell, PhD (✉)
Dept. of Health Sciences
Lund University
P.O. Box 157
22100 Lund, Sweden
Tel.: +46-462221937
Fax: +46-462221934
E-Mail: Peter.Hagell@med.lu.se

P. Hagell, PhD
The Vårdal Institute
the Swedish Institute for Health Science
Lund University
Lund, Sweden

P. Hagell, PhD
Dept. of Neurology
University Hospital
Lund, Sweden

A. L. Törnqvist, PhD
Dept. of Neurosurgery
University Hospital
22185 Lund, Sweden

A. L. Törnqvist, PhD
Dept. of Clinical Sciences, Neurosurgery
Lund University
Lund, Sweden

J. Hobart, PhD
Dept. of Clinical Neuroscience
Peninsula Medical School
Room N16 ITTC Building
Tamar Science Park
Davy Road, Plymouth
Devon PL6 8BX, UK

J. Hobart, PhD
Neurological Outcome Measures Unit
Institute of Neurology
Queen Square
London, UK

Abstract Rating scales are increasingly the primary outcome measures in clinical trials. However, clinically meaningful interpretation of such outcomes requires that the scales used satisfy basic requirements (scaling assumptions) within the data. These are rarely tested. The SF-36 is the most widely used patient-reported rating scale. Its scaling assumptions have been challenged in neurological disorders but remain untested in Parkinson's disease (PD). We therefore tested these by analyzing SF-36 data from 202 PD patients (54 % men; mean age 70) to determine if it was legitimate to report scores for the eight SF-36 scales and its two summary measures of physical and mental health, and if those scores were reliable and valid. Results supported generation of the eight SF-36 scale scores and their reliabilities were generally good (≥0.74 in all but one instance). However, we found limitations that question the meaningfulness of four scales and other limitations that restrict the ability of four scales to detect change in clinical trials (floor/ceiling effects, 19.6–46.2 %). The two SF-36 summary measures were not found to be valid indicators of physical and mental health. This study demonstrates important limitations of the SF-36 and provides the first evidence-based guidelines for its use in PD. The limitations of the SF-36 demonstrated here may explain some unexpected findings in previous studies. However, the main implication is a general one for the clinical research community regarding requirements for reporting rating scale endpoints. Specifically, investigators should routinely provide scale evaluations based on data from *within* major clinical trials.

Key words clinical trials · outcome research · quality of life · Parkinson's disease

JON 2708

Introduction

Neurological clinical trials have traditionally measured outcomes using clinician reported rating scales. However, patients' and clinicians' perceptions of the impact of disease and therapy may differ, and be equally valid [16]. Accordingly, there is a growing emphasis on the importance of patient-reported outcome measures, particularly in chronic conditions such as neurological disorders [5, 16, 39].

Valid interpretation of rating scale data requires that certain criteria are met. For example, combining item responses into a total score assumes that this is a legitimate process [27]. This becomes particularly relevant when rating scales developed for one population are used in new patient groups, and is compounded by the fact that rating scale performance is sample dependent [27, 33]. As such, scales may not work as assumed and intended in specific populations.

The Medical Outcomes Study 36-item Short-Form Health Survey (SF-36) [37] is the most widely used generic (non-disease specific) patient-reported outcome measure. It is the main rating scale used for comparison of outcomes across diseases and populations, and is recommended for use in health policy evaluations, general population surveys, clinical research and practice [25, 37]. However, the validity of some SF-36 scores has been challenged in several neurological conditions [1, 4, 6, 17, 18, 21], whereas it has been supported in people with, e.g., osteo- and rheumatoid arthritis [22]. This underscores the need for documented supportive evidence of rating scale characteristics to enable valid inferences and claims regarding outcomes from clinical trials [12, 33].

In Parkinson's disease (PD), the SF-36 has been used in clinical trials, in studies comparing PD with the general population, and as a reference tool for new scales. It is also the recommended generic patient-reported outcome measure in neurosurgical PD trials [8]. Despite its important role, information on the performance of the SF-36 in PD is incomplete [7, 31]. In particular, no study has examined the basic requirements (scaling assumptions) for generating SF-36 scores. Such evaluations concern whether scores are valid representations of the variables they set out to measure and are essential for appropriate use of scales [12, 33].

Here we illustrate the fundamental importance of the basic, but rarely tested, assumptions that underpin the use of rating scale scores by examining the SF-36 in PD. Results provide evidence-based guidelines for using the SF-36 in clinical PD research. More importantly, the findings have general implications regarding reporting of any rating scale derived outcomes.

Methods

Patients and data collection

Data were collected by a postal survey. A total of 451 people with neurologist diagnosed PD seen at a South Swedish university hospital during one year were considered for inclusion. Of these, seven had passed away. People in terminal care (n = 23) were excluded together with people participating in other recent or ongoing questionnaire studies (n = 164). The latter group was excluded in order to avoid unnecessary respondent burden and compromised data quality resulting from questionnaire fatigue. The remaining 257 people were sent a questionnaire booklet containing the SF-36 [34, 37], Nottingham Health Profile [20], 39-item PD Questionnaire [30], a life satisfaction questionnaire [13], demographic-, PD- and survey related questions. Two weeks later (Time 2) patients received a second copy of the questionnaire including a question asking if their health had changed since Time 1. Reminders were sent out one week after each mailing. The study was performed in accordance with the ethical standards laid down in the Declaration of Helsinki and was approved by the local research ethics committee.

The SF-36

This generic rating scale has 36 items intended to reflect aspects of health from the perspective of the patient [37]. The SF-36 assumes that 35 of its items can be grouped into eight scales (Table 1). It is further assumed that the eight scales can be combined to form two summary measures of physical and mental health (the physical and mental component summary scores, PCS and MCS) [36, 38]. These summary measures have been suggested to have advantages over the eight scales for clinical trials by reducing the risk of chance findings and improving the potential to detect clinically significant change [36].

Analyses

Detailed descriptions of the analyses are provided elsewhere [1, 17, 18, 27, 35]. Briefly, for each of the eight SF-36 scales we first examined the percent missing data (data quality) as this indicates whether a scale is acceptable to a sample. Then we assessed the legitimacy of adding up items to generate scale scores without items being given different weights by examining item means and standard deviations (should be similar within each scale) and item-total correlations (should be > 0.3) [23, 27, 35]. If that process was supported, we then examined whether the grouping of items into eight scales (as suggested by Table 1) was empirically supported. That is, we examined whether the corrected item-total correlations exceed 0.4 and if, for each item, these correlations were significantly stronger than the item's correlation with any of the other scales (referred to as scaling success) [27, 35]. The amount of floor and ceiling effects (i.e., the proportion of people obtaining minimum and maximum scores, respectively) for each scale, and the reliability (internal consistency and test-retest reliability) were also examined. Floor/ceiling effects should not exceed 15 % [26] and scale reliabilities should be ≥ 0.7 and preferably ≥ 0.8 for group comparison studies [29]. Item-level reliability is considered acceptable when > 0.5 [9]. Finally, we examined if the eight scales measured different aspects of health by comparing the scale-to-scale correlations with each scale's internal consistency. Scales are considered measuring distinct constructs when their internal consistencies are larger than their inter-correlations [35].

For the two SF-36 summary measures (PCS and MCS), we determined whether they satisfied the criteria required for them to be computed using the published algorithm [36, 38]. A fundamental requirement for this algorithm to produce valid and interpretable PCS and MCS scores is that exploratory factor analysis (a data reduction tech-

Table 1 Measurement model of the SF-36[a]

Items		Scales	Summary measures
No.	Content (abridged)		
3a	Vigorous activities	Physical functioning (PF)	Physical health (PCS)
3b	Moderate activities		
3c	Lifting or carrying groceries		
3d	Climbing several flights of stairs		
3e	Climbing one flight of stairs		
3f	Bending, kneeling, stooping		
3g	Walking more than a mile		
3h	Walking several blocks		
3i	Walking one block		
3j	Bathing or dressing		
4a	Cut down time spent on work	Role-physical (RP)	
4b	Accomplished less than would like		
4c	Limited in the kind of work		
4d	Difficulty performing the work		
7	Pain – magnitude	Bodily pain (BP)	
8	Pain – interference with work		
1	Overall rating of general health	General health (GH)[b]	
11a	Get sick easier than others		
11b	As healthy as anyone I know		
11c	Expect health to get worse		
11d	My health is excellent		
9a	Feel full of pep	Vitality (VT)[b]	Mental health (MCS)
9e	Have a lot of energy		
9g	Feel worn out		
9i	Feel tired		
6	Extent of social limitations	Social functioning (SF)[b]	
10	Time of social limitations		
5a	Cut down time spent on work	Role-emotional (RE)	
5b	Accomplished less than would like		
5c	Didn't do work as carefully as usual		
9b	Been a nervous person	Mental health (MH)	
9c	Felt down in the dumps		
9d	Felt calm and peaceful		
9f	Felt downhearted and blue		
9h	Been a happy person		
2	Health now compared to a year ago	–[c]	–[c]

[a] As described by Ware et al. [36, 38].
[b] Scales considered to correlate with both summary measures, but strongest with that indicated in the table.
[c] Item 2 is not used in the scoring of SF-36 scales or summary measures.

nique) of the eight SF-36 scale scores results in two components [36]. Specifically, four SF-36 scales (PF, RP, BP, GH) should form one component (PCS), and the remaining four scales (MH, RE, SF, VT) should form the other (MCS) [36]. These two components should explain at least 75 % of the reliable variance in the eight SF-36 scales [36, 38].

In addition, we used confirmatory factor analysis to assess how well observed data fitted the hypothesized scales-to-summary measure structure. This technique is generally recommended over exploratory factor analysis when there is an *a priori* hypothesis regarding dimensionality, since it allows for testing whether empirical data fit an assumed structure [11]. In the case of the SF-36 summary measures, the *a priori* hypothesis is that the eight scales relate to two underlying constructs representing physical and mental health as outlined above (Table 1) [36, 38].

Analyses were performed using SPSS 12 (SPSS Inc., Chicago, IL) and AMOS 5 (SmallWaters Corp., Chicago, IL) for Windows.

Results

Patient characteristics and response rates are summarized in Table 2. The 202 patients included in the main analyses represented all five Hoehn & Yahr stages [19]; about 68 % experienced motor fluctuations and 49 % experienced dyskinesias. All but seven patients received levodopa with or without adjunct drugs, 18 had undergone neurosurgical interventions for their PD, three were on anti-PD drugs other than levodopa, and four were not yet on any medical therapy.

Table 2 Patient characteristics

	Time 1	Time 2[a]
Respondents (response rate)	209 (81)[b]	173 (67)[b]
Questionnaires self completed[c]	202 (97)[b]	168 (97)[b]
Unchanged self-reported health[d]	–	137 (79)[b]
Gender (men/women)	108 (53.5)/94 (46.5)[b]	74 (54)/63 (46)[b]
Age (years)	69.8 (10.0)[e]	70 (8.6)[e]
PD duration (years)	8.7 (6.6)[e]	8.7 (6.4)[e]
"Off"-phase Hoehn & Yahr stage of PD[g]	III (II-IV)[f]	III (II-IV)[f]
Subjective disease severity[h]	2 (2–2)[f]	2 (1–2)[f]
Motor fluctuations[i]	137 (67.8)[b]	89 (65)[b]
Dyskinesias[i]	99 (49)[b]	66 (48.2)[b]
Retired	143 (70.8)[b]	96 (70.1)[b]
Married or cohabitant	144 (71.2)[b]	101 (73.7)[b]
Living in own home	179 (88.6)[b]	120 (87.6)[b]

[a] Two weeks after time 1.
[b] n (%).
[c] Patients reporting that they had answered the questionnaires themselves. Those indicating that they had not answered the survey themselves were excluded.
[d] Self-reported change in health status since time 1 according to a 5-grade scale (much better – better – unchanged – worse – much worse).
[e] Mean (standard deviation).
[f] Median (q1–q3).
[g] From clinic visits within about 9 months of the postal survey. Higher values indicate more severe PD (range, I–V; I = mild unilateral disease, V = Confined to bed or wheelchair unless aided) [19].
[h] Self-reported as mild (= 1), moderate (= 2), or severe (= 3).
[i] Self-reported as present or absent.
PD Parkinson's disease

The eight SF-36 scales

The SF-36 appeared acceptable, as data quality was high with a mean of 3 % missing item responses (range, 0.5–5.9 %; data available on request) and scale scores could be computed (i.e., < 50 % missing item responses per scale) for 96.5 % (GH and RE) to 98.5 % (BP) of the sample. For all eight scales, there was general support for the legitimacy of generating scale scores by summing items without standardization or weighting. Roughly similar item mean scores and standard deviations within most scales (Table 3) indicate that they contribute about equally to their total scores and, therefore, that their variances do not need to be standardized before summation. Furthermore, all corrected item-to-own scale correlations were ≥0.4 (Table 3), which supports summation without applying item weights and indicates that the items in each scale measure a common construct [35].

We found evidence to challenge whether the proposed eight scales (Table 1) represent the best grouping of items. Although scaling successes (i.e., items correlating significantly stronger with their hypothesized own scale than with other scales) were 100 % for two scales (BP and RE) and nearly complete for another two (PF and RP), they were notably compromised (< 80 %) [32] for three scales (GH, VT and SF; Table 3). Scaling failure was detected in the SF (stronger item correlation with VT and MH than with SF) and MH (stronger item correlation with VT) scales. Because our sample size may have underestimated scaling success rates, and thus been overly harsh on the SF-36, we recomputed the scaling success rates with the sample size reset to 300 rather than the actual sample size [35]. This change in sample size decreases the standard error around the observed correlations and thus makes it easier for the item to score a scaling success. These analyses did not affect the overall findings (data available on request).

The distribution of scores for all eight SF-36 scales spanned the entire range (0–100). However, four scales (RP, RE, BP and SF) had notable (> 15 %) [26] floor or ceiling effects (Table 3).

Reliability was generally good (Table 3). Test-retest reliabilities of five items (numbers 4a, 4b, 5a, 5b, and 11b) did not meet the 0.5 item-level criterion [9]. Fifteen out of 16 scale reliability coefficients exceeded the recommended minimum of 0.70, and 12 exceeded the preferred value of 0.80 [29]. However, the reliabilities of the VT and MH scales fell within the 95 % confidence interval of the correlation between them (Table 4), suggesting some measurement overlap [35].

The two SF-36 summary measures

Factor analysis did not result in the eight scales grouping into two components. Instead, all scales grouped together as a single component, implying that they repre-

Table 3 Descriptive and psychometric statistics for SF-36 scales[a]

SF-36 scale (no. of items/ response categories)	Scale mean (SD) scores[b]	Ranges of item mean (SD) scores	Floor/ceiling effect (%)[c]	Scale reliability[d]		Item test-retest reliability[g]	Item-total correlation[h]	Scaling success[i]/ failure[j] (%)
				Internal consistency[e]	Test-retest[f]			
PF (10/3)	51.75 (29.79)	1.39–2.49 (0.63–0.82)	5.1/5.6	0.94	0.87	0.58–0.80	0.58–0.82	95.7/0
RP (4/2)	35.02 (39.98)	1.27–1.51 (0.45–0.50)	46.2/20.8	0.87	0.74	0.43–0.71	0.64–0.77	92.9/0
BP (2/5–6)	58.69 (27.45)	3.83–4.04 (1.41–1.44)	2.0/19.6	0.92	0.86	0.80–0.84	0.85–0.85	100/0
GH (5/5)	47.25 (21.34)	2.47–4.14 (1.07–1.29)	10/10	0.79	0.81	0.45–0.65	0.40–0.71	51.4/0
VT (4/6)	51.03 (24.76)	2.96–4.62 (1.45–1.60)	0.5/1.5	0.84	0.84	0.58–0.71	0.56–0.73	75/0
SF (2/5)	70.96 (26.23)	2.0–3.67 (1.09–1.23)	0.5/30.8	0.78	0.82	0.71–0.73	0.64–0.64	64.3/14.3
RE (3/2)	50.43 (44.95)	1.42–1.55 (0.50–0.50)	38.5/39.5	0.88	0.62	0.49–0.56	0.75–0.80	100/0
MH (5/6)	68.01 (20.74)	3.48–5.14 (1.17–1.52)	0.5/6.6	0.82	0.84	0.53–0.71	0.57–0.67	82.9/2.9

[a] All data but test-retest reliability are from time 1.
[b] Possible score range, 0–100 (100 = better health).
[c] Percentage scoring 0 (floor) and 100 (ceiling). Should not exceed 15 % [26].
[d] Should be ≥ 0.7 and preferably ≥ 0.8 for group comparison studies [29].
[e] Cronbach's coefficient alpha.
[f] One-way random intra-class correlation from scores of patients completing both administrations (2 weeks interval) and reporting unchanged health status at time 2 (n = 137).
[g] Range of quadratic weighted Kappa values for item scores from patients completing both administrations (2 weeks interval) and reporting unchanged health status at time 2 (n = 137). Should be > 0.5 [9].
[h] Range of item-to-own scale correlations corrected for overlap (i.e., the correlation between each item and the total score computed from the remaining items in that scale). Should be ≥ 0.4 to imply measurement of a common underlying construct, and ≥ 0.3 to allow unweighted summation [27, 35].
[i] Percentage of occasions when item-to-own scale correlations (corrected for overlap) exceed item-to-other scale correlations by > 2 standard errors ($2 \times 1/\sqrt{n}$), i.e., the approximate limit of the 95 % CI [27, 35] should be at least 80 % [32].
[j] Percentage of occasions when items correlated stronger with other subscales than with their hypothesized subscale.
PF physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health

Table 4 Inter-correlations among SF-36 scales[a]

SF-36 Scale	SF-36 Scale							
	PF	RP	BP	GH	VT	SF	RE	MH
PF	(0.94)[b]							
RP	0.55	(0.87)						
BP	0.37	0.38	(0.92)					
GH	0.48	0.55	0.48	(0.79)				
VT	0.61	0.62	0.54	0.64	(0.84)			
SF	0.52	0.56	0.46	0.54	0.62	(0.78)		
RE	0.46	0.67	0.37	0.46	0.49	0.44	(0.88)	
MH	0.43	0.46	0.46	0.57	0.73*	0.59	0.44	(0.82)

[a] Inter-correlations among scales should be substantially less than their respective alpha coefficients to support measurement of distinct constructs [35].
[b] Internal consistency reliability (coefficient alpha) in parentheses.
* Internal consistency reliabilities within 2 standard errors (±0.14) of the correlation between the VT and MH scales.
PF physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health

sent a single health dimension (Table 5). When data were re-factor analyzed specifying that the scales be forced into two components (i.e., a two-dimensional solution), the magnitude and pattern of scale-to-component correlations were notably different (Fig. 1A) from those observed in general populations (Fig. 1B, C). This suggests that in PD the PCS and MCS do not represent meaningful summary measures of physical and mental health.

Assessment of the fit of the observed data to the hypothesized scales-to-summary measures relationships (Table 1) using confirmatory factor analysis showed poor fit (Fig. 2). This further argues against the hypothesized measurement model in PD.

Discussion

The aim of this study was to illustrate the importance of comprehensive scale evaluation and to provide preliminary guidelines for using the SF-36 in PD. However, the results have general implications for any study using rating scales by calling for quality standards when reporting rating scale data from clinical trials.

Recommendations for selecting patient-reported outcome measures in clinical PD trials have been arbitrary rather than evidence based [8], and their use in clinical research is rarely accompanied by evidence of their measurement validity. With these facts in mind, we comprehensively tested the basic assumptions underpinning the scoring of the eight SF-36 scales and its two summary measures in PD. These tests are advocated by the developers of the SF-36 and considered pivotal for valid use of the questionnaire [27, 32, 35, 36, 38]. Results showed good data quality, general support for summa-

Table 5 Factor analysis of SF-36 scales[a]

SF-36 scale	Eigenvalue > 1 criterion[b]	Two component criterion[c]		
		PCS[d]	MCS[d]	h^2/r_{tt}[e]
PF	0.72	0.71	0.34	0.66
RP	0.80	0.82	0.34	0.91
BP	0.67	0.17	0.74	0.64
GH	0.77	0.34	0.73	0.82
VT	0.87	0.45	0.76	0.92
SF	0.79	0.46	0.65	0.81
RE	0.72	0.81	0.24	0.82
MH	0.78	0.29	0.79	0.86
Eigenvalue	4.72	4.72	0.76	
Variance	58.96	58.95	9.47	
Total variance[f]		68.43		

[a] Principal component analysis with orthogonal (varimax) rotation of extracted factors (Kaiser-Meyer-Olkin measure of sampling adequacy: 0.89; Bartlett's test of sphericity: χ^2, 775.35, $P < 0.0001$).
[b] Eigenvalue > 1 as criterion for factor extraction.
[c] Two components were pre-specified for extraction.
[d] The PCS should correlate strongly (> 0.70) with the PF, RP and BP scales, and weakly (< 0.30) with the MH, RE and SF scales, and vice versa for the mental MCS measure. The GH, VT and SF scales should correlate moderately with both PCS and MCS, with GH correlating higher with the PCS, and VT and SF correlating higher with the MCS [36, 38].
[e] Total reliable variance in each SF-36 scale explained by the two principal components (h^2 = sum of squared factor loadings for each scale, r_{tt} = coefficient alpha for each scale).
[f] Percent of total reliable variance in all SF-36 scales explained by the extracted factors.
PF physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health; *PCS* physical component summary measure; *MCS* mental component summary measure

tion of items without standardization or weighting, good reliability for the majority of the SF-36 scales, compromised scaling success for three scales, and notable floor or ceiling effects in four scales. Importantly, results did not support using the SF-36 PCS and MCS summary measures.

Our observations do not preclude use of the SF-36 in PD. However, they demonstrate some important limitations of the scale in this population that raise important questions about the use of the SF-36 as an outcome measure in PD. Until other studies prove otherwise, they also provide the following guidelines to assist clinicians in choosing SF-36 derived outcomes and to facilitate valid inferences based on the SF-36 in clinical PD research. First, while the process of generating SF-36 scale scores appears legitimate, clinicians should be aware that scores of half of the scales (GH, VT, SF, MH) should be interpreted cautiously as analyses of scaling success rates and scale-to-scale correlations relative to internal consistencies suggest that there is some ambiguity regarding their meaning. Clinical interpretation of these scores is therefore obstructed as it is uncertain what they represent. Second, the RP, BP, SF, and RE scales are not likely to be good choices as outcome measures in clinical trials since large floor and ceiling effects most probably will underestimate actual changes and differences between patients [1]. Large floor/ceiling effects indicate that the levels of functioning among those with minimum/maximum scores are not reflected by the available scale scores. Consequently, changes or differences outside the range covered by the scale will be undetected, the amount of change or difference required for the scale to respond is unknown, and only changes or differences in one direction can be detected.

For the SF-36 summary measures, our factor analytic results indicate that the PCS and MCS cannot be meaningfully interpreted as summary measures of physical and mental health in PD. This conclusion is based on the fact that the scoring algorithm for the summary measures assumes that the pattern of scale-to-component

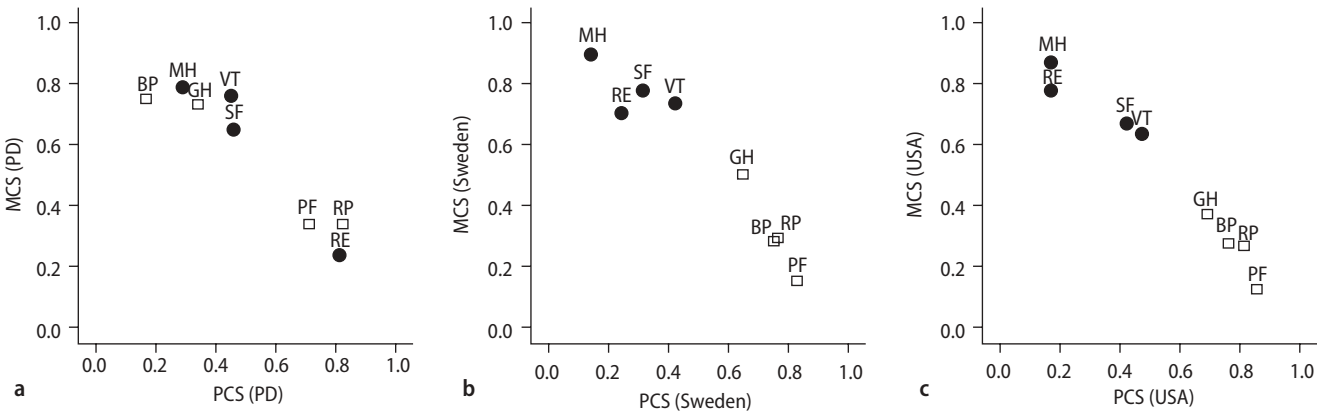
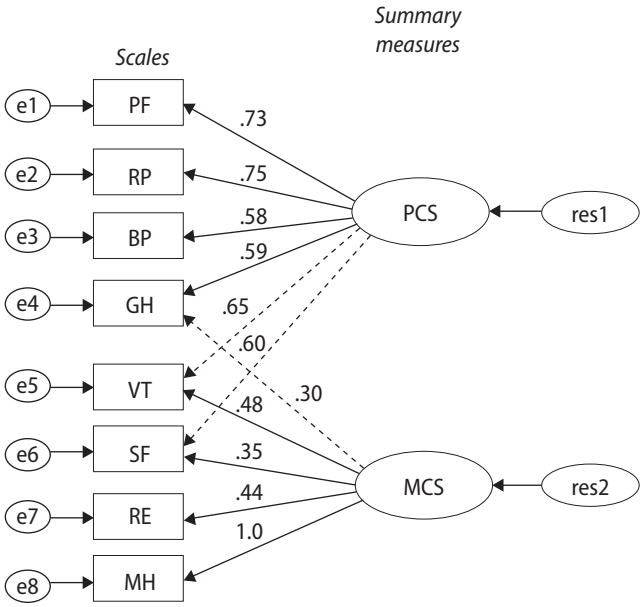


Fig. 1 SF-36 scale-to-component correlations derived from exploratory principle component factor analyses with orthogonal (varimax) rotation in (**A**) PD (n = 202; this study), (**B**) Swedish general population (n = 8930) [34], and (**C**) the United States general population (n = 2474) [36]. For the PD sample (**A**), two components were pre-specified for extraction. Open squares indicate scales expected to correlate most strongly with the PCS and filled circles indicate scales expected to correlate most strongly with the MCS (*PD* Parkinson's disease; *MCS* Mental Component Summary; *PCS* Physical Component Summary; *PF* physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health)



Model Fit

Fit statistic	Model Fit	Criteria
Chi-square	155.89 (P<0.0001)	P > 0.05
GFI	0.87	≥0.9
AGFI	0.72	≥0.9
CFI	0.82	≥0.9
RMSEA (90% CI)	0.21 (0.18–0.24)	<0.05

Fig. 2 Hypothesized relationships between SF-36 scales and summary measures assessed for fit with data from patients with Parkinson's disease (n = 202) by means of confirmatory factor analysis. Arrows indicate hypothesized primary relationships, and dashed arrows indicate hypothesized substantial secondary relationships [36]. Coefficients above each arrow are estimated standardized regression weights. Squares and circles represent observed and latent variables, respectively. The box summarizes model fit and accompanying criteria for acceptable fit [3] (*PF* physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health; *PCS* Physical Component Summary; *MCS* Mental Component Summary; *e* error term; *res* residual covariance; *GFI* goodness-of-fit index; *AGFI* adjusted goodness-of-fit index; *CFI* comparative fit index; *RMSEA* root mean square error of approximation; *CI* confidence interval)

correlations is consistent with the expected one [36]. If not, the algorithm produces scores that do not represent what they are believed to represent. Moreover, it is not known what they measure. Therefore, until other studies prove otherwise, results from this study suggest that the PCS and MCS should be avoided in PD. Instead, the PF and MH scales are recommended if SF-36 derived indices of physical and mental health are desired. PF fulfils all assessed criteria and ambiguities regarding the MH scale relate to an overlap with VT, which does not compromise its validity relative to PF. These guidelines are based on the currently best available evidence; we encourage others to examine their data to support their findings and to further clarify the role of the SF-36 in PD.

The contributions made by the SF-36 to understand illness and therapy from the patients' perspective should not be underestimated. However, we believe that when scales are used to make judgements about the effectiveness of therapy for chronically ill people there can be no compromise in scientific rigour. Problems with measurement validity cannot be compensated for by trial design and may explain some unexpected results in previous neurological trials. For example, in a double-blind clinical PD trial of adjunct entacapone, Fénelon et al. [10] found non-significant improvements and deteriorations in PCS and MCS scores, respectively. However, post-hoc analyses revealed significant improvements in the MCS-associated MH scale only. Furthermore, while the only study examining the measurement properties of the SF-36 in amyotrophic lateral sclerosis failed to support the validity of its physical and mental component summary scores [21], the PCS and MCS have been used to assess the effectiveness of, for example, non-invasive ventilation in this disorder [2]. Results showed improvements among people with good, but not among those with poor bulbar function. In contrast, other patient-reported scales showed significant (or nearly significant) improvements in both groups [2].

The most important implication of this study is a general one regarding the use of rating scales in clinical research. It goes without saying that any rating scale used in clinical trials, whether patient or clinician reported, should have documented reliability and validity [24]. However, in light of emerging recommendations from leading regulatory bodies [12], the results reported here and elsewhere call for the clinical research community to step up standards and begin taking score reliability and validity into account when reporting and interpreting study results, rather than presuming that rating scale assumptions are met in particular study samples [14, 15]. We therefore propose that any rating scale endpoints used in major clinical trials routinely should be evaluated using analyses such as those reported here. These analyses should be undertaken in the data generated by *the reported study*. Preferably, this information should be made available, e.g., as appendices or supplementary online data. Such practice would allow for transparent and valid interpretation of study results and facilitate accumulation of evidence for more firmly based future evidence-based guidelines regarding rating scale use. It is now time to complement the CONSORT guidelines [28] with formal guidance regarding using and reporting rating scale derived outcome measures in major clinical trials.

Acknowledgements The authors wish to thank the patients for their cooperation and Jan Reimer for assistance with data collection.

The study was supported by the Swedish Research Council, the Skane County Council Research and Development Foundation, Rådet för hälso- och sjukvårdsforskning (HSF), and the Department of Nursing at Lund University.

References

1. Baron R, Elashaal A, Germon T, Hobart J (2006) Measuring outcomes in cervical spine surgery: think twice before using the SF-36. Spine 31:2575–2584
2. Bourke SC, Tomlinson M, Williams TL, et al. (2006) Effects of non-invasive ventilation on survival and quality of life in patients with amyotrophic lateral sclerosis: a randomised controlled trial. Lancet Neurol 5:140–147
3. Byrne BM (2001) Structural equation modeling with AMOS. Basic concepts, applications, and programming. Lawrence Erlbaum Associates, Inc., Mahwah
4. Cano SJ, Thompson AJ, Bhatia K, et al. (2007) Evidence-based guidelines for using the Short Form 36 in cervical dystonia. Mov Disord 22:122–126
5. Clancy CM, Eisenberg JM (1998) Outcomes research: measuring the end results of health care. Science 282:245–246
6. Dallmeijer AJ, Dekker J, Knol DL, et al. (2006) Dimensional structure of the SF-36 in neurological patients. J Clin Epidemiol 59:541–543
7. Damiano AM, McGrath MM, Willian MK, et al. (2000) Evaluation of a measurement strategy for Parkinson's disease: assessing patient health-related quality of life. Qual Life Res 9:87–100
8. Defer GL, Widner H, Marié RM, et al. (1999) Core Assessment Program for Surgical Interventional Therapies in Parkinson's Disease (CAPSIT-PD). Mov Disord 14:572–584
9. Duruöz MT, Poiraudeau S, Fermanian J, Menkes CJ, Amor B, Dougados M, Revel M (1996) Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol 23:1167–1172
10. Fénelon G, Giménez-Roldán S, Montastruc JL, et al. (2003) Efficacy and tolerability of entacapone in patients with Parkinson's disease treated with levodopa plus a dopamine agonist and experiencing wearing-off motor fluctuations. A randomized, double-blind, multicentre study. J Neural Transm 110:239–251
11. Floyd FJ, Widaman KF (1995) Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess 7:286–299
12. Food and Drug Administration (2006) Draft guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Federal Register 71:5862–5863 (Available at: http://www, fda, gov/cber/gdlns/prolbl, htm)
13. Fugl-Meyer AR, Melin R, Fugl-Meyer KS (2002) Life satisfaction in 18- to 64-year-old Swedes: in relation to gender, age, partner and immigrant status. J Rehabil Med 34:239–246
14. Hagell P (2007) Self reported health in people with Parkinson's disease left untreated at diagnosis. J Neurol Neurosurg Psychiatry 78:442
15. Hobart J (2005) How severe is chronic pain in multiple sclerosis? Nat Clin Pract Neurol 1:80–81
16. Hobart JC, Freeman JA, Lamping DL (1996) Physician and patient-oriented outcomes in progressive neurological disease: which to measure? Curr Opin Neurol 9:441–444
17. Hobart J, Freeman J, Lamping D, et al. (2001) The SF-36 in multiple sclerosis: why basic assumptions must be tested. J Neurol Neurosurg Psychiatry 71: 363–370
18. Hobart JC, Williams LS, Moran K, Thompson AJ (2002) Quality of life measurement after stroke: uses and abuses of the SF-36. Stroke 33: 1348–1356
19. Hoehn MM, Yahr M (1967) Parkinsonism: Onset, progression and mortality. Neurology 17:427–442
20. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E (1980) A quantitative approach to perceived health status: A validation study. J Epidemiol Community Health 34:281–286
21. Jenkinson C, Hobart J, Chandola T, et al. (2002) Use of the short form health survey (SF-36) in patients with amyotrophic lateral sclerosis: tests of data quality, score reliability, response rates and scaling assumptions. J Neurol 249:178–183
22. Kosinski M, Keller SD, Hatoum TH, et al. (1999) The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: Tests of data quality, scaling assumptions and score reliability. Med Care 37(5 Suppl): MS10–MS22
23. Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 140:1–55
24. Marshall M, Lockwood A, Bradley C, et al. (2000) Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. Br J Psychiatry 176: 249–252
25. McDowell I (2006) Measuring health: a guide to rating scales and questionnaires, 3rd ed. Oxford University Press, Inc., New York
26. McHorney CA, Tarlov AR (1995) Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 4: 293–307
27. McHorney CA, Ware JE Jr, Lu JFR, Sherbourne CD (1994) The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. Med Care 32:40–66
28. Moher D, Schulz KF, Altman DG, the CONSORT Group (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357:1191–1194
29. Nunnally JC, Bernstein IH (1994) Psychometric theory, 3rd ed. McGraw-Hill, New York
30. Peto V, Jenkinson C, Fitzpatrick R, Greenhall R (1995) The development and validation of a short measure of functioning and well being for individuals with Parkinson's disease. Qual Life Res 4:241–248
31. Rubenstein LM, Voelker MD, Chrischilles EA, et al. (1998) The usefulness of the Functional Status Questionnaire and Medical Outcomes Study Short Form in Parkinson's disease research. Qual Life Res 7:279–290
32. Saris-Baglama RN, Dewey CJ, Chisholm GB, et al. (2004) SF health outcomes™ scoring software user's guide. QualityMetric, Inc., Lincoln
33. Scientific Advisory Committee of the Medical Outcomes Trust (2002) Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 11:193–205
34. Sullivan M, Karlsson J, Ware JE (1994) Hälsoenkät SF-36. Svensk manual och tolkningsguide (SF-36 Health Survey. Swedish manual and interpretation guide). Sahlgrenska University Hospital, Göteborg, Sweden

35. Ware JE Jr, Harris WJ, Gandek B, et al. (1997) MAP-R for Windows: multi-trait/multi-item analysis program – revised user's guide. Health Assessment Lab, Boston, MA
36. Ware JE Jr, Kosinski MA, Keller SD (1994) SF-36 physical and mental health summary scales: a user's manual. Boston: New England Medical Center, The Health Institute
37. Ware JE Jr, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I Conceptual framework and item selection. Med Care 30:473–483
38. Ware JE Jr, Snow KK, Kosinski M, Gandek B (1993) SF-36 Health Survey manual and interpretation guide. Nimrod Press, Boston, MA
39. Wheatley K, Stowe RL, Clarke CE, et al. (2002) Evaluating drug treatments for Parkinson's disease: how good are the trials? BMJ 324:1508–1511