Determination of the minimal clinically important difference for seven fatigue measures in rheumatoid arthritis

Jacques Pouchot, MD[1,6], **Raheem B. Kherani, MD**[1], **Rollin Brant, PhD**[2], **Diane Lacaille, MD, MHSc**[1], **Allen J. Lehman, MA**[1,3], **Stephanie Ensworth, MD**[1], **Jacek Kopec, MD, PhD**[1,4], **John M. Esdaile, MD, MPH**[1], and **Matthew H. Liang, MD, MPH**[1,5]

1 *Arthritis Research Centre of Canada, Division of Rheumatology, University of British Columbia, Vancouver, BC, Canada*

2 *Centre for Community Child Health Research, Child and Family Research Institute, University of British Columbia, Vancouver, BC, Canada*

3 *Department of Educational Studies, University of British Columbia, Vancouver, BC, Canada*

4 *Department of Health Care and Epidemiology, University of British Columbia, Vancouver, BC, Canada*

5 *Division of Rheumatology, Immunology, and Allergy, Brigham & Women's Hospital and Harvard Medical School, Boston, MA, USA*

6 *Department of Internal Medicine, Hôpital Européen Georges Pompidou, Université René Descartes – Paris V, Paris, France*

Abstract

Objective—To estimate the minimal clinically important difference (MCID) of seven measures of fatigue in rheumatoid arthritis.

Study Design and Setting—A cross-sectional study design based on inter-individual comparisons was used. Six to eight subjects participated in a single meeting and completed seven fatigue questionnaires (nine sessions were organized and 61 subjects participated). After completion of the questionnaires, the subjects had five one-on-one 10-minute conversations with different people in the group to discuss their fatigue. After each conversation, each patient compared their fatigue to their conversational partner's on a global rating. Ratings were compared to the scores of the fatigue measures to estimate the MCID. Both non-parametric and linear regression analyses were used.

Results—Non-parametric estimates for the MCID relative to "little more fatigue" tended to be smaller than those for "little less fatigue". The global MCIDs estimated by linear regression were: FSS 20.2, VT 14.8, MAF 18.7, MFI 16.6, FACIT–F 15.9, CFS 9.9, RS 19.7, for normalized scores (0 to 100). The standardized MCIDs for the seven measures were roughly similar (0.67 to 0.76).

Conclusion—These estimates of MCID will help to interpret changes observed in a fatigue score and will be critical in estimating sample size requirements.

Keywords

minimal clinically important difference; sample size requirement; fatigue; rheumatoid arthritis; health status; interpretation

Rheumatoid arthritis (RA) is a chronic inflammatory disease that causes joint pain and destruction, and disability. The majority of persons with RA complain of fatigue [1–4], and describe it as different from their normal tiredness in that it is overwhelming and uncontrollable [5]. Patients identify physical, cognitive and emotional components of fatigue and complain that their symptom is commonly ignored by physicians [5]. In one study, 57% of patients reported that fatigue was the most important aspect of their disease [3] and in another, 42% of RA patients had clinically important levels of fatigue [4]. Fatigue in RA is associated with sleep disturbance, functional disability, pain, depressive symptoms and adverse psychosocial consequences [2,4,6]. Despite its high prevalence and its profound negative impact on quality of life [5,7,8], fatigue remains rarely evaluated in clinical studies.

Fatigue is a non-specific subjective symptom. In the absence of an objective measurement, fatigue can only be assessed by asking the subject. The measurement properties of available instruments need to be evaluated if they are to be used clinically or in clinical trials. Among the psychometric properties, longitudinal validity (responsiveness, sensitivity to change) is one of the most important. Closely related is the minimal clinically important difference (MCID), which is defined as "the smallest difference in score in the domain of interest (fatigue) which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" [9]. The MCID is essential to interpret the magnitude of longitudinal changes or differences when comparing two treatments or different group of patients. Knowledge of MCID is also essential for meaningful sample size calculations in clinical trials.

The aim of our study was to estimate the MCID of seven validated self-administered measures of fatigue in persons with RA. The fatigue instruments identified from a literature review as suitable for use in RA and studied were: the Fatigue Severity Scale (FSS) [10,11], the Vitality scale of the MOS-SF36 (VT) [12], the Multidimensional Assessment of Fatigue (MAF) [2], the Multidimensional Fatigue Inventory (MFI) [13], the Functional Assessment of Chronic Illness Therapy–Fatigue (FACIT–F) [14], the Chalder Fatigue Scale (CFS) [15], and a global numerical rating of fatigue with a 10-point scale (RS).

Patients and Methods

Patients

The study was conducted at the Mary Pack Arthritis Centre, Vancouver, Canada. All participants signed an informed consent. Eligible persons had RA as defined by the American College of Rheumatology (ACR) [16]. Participants had to be outpatients. Patients unable to complete self-administered questionnaires, or unable to read or converse in English, were excluded.

Study design

Six to eight subjects participated in each session. There were nine sessions with a total of 61 subjects, a sample size based on previous studies using a similar methodology [9,17–20]. After a detailed description of the study, each participant completed the seven fatigue instruments, and the Health Assessment Questionnaire (HAQ) [21] to assess their physical function. In

addition, demographic data, and self-assessment of disease activity and pain assessed by 10-point numerical rating scales were obtained. This component took 30 to 40 minutes to complete.

After completing the questionnaires, participants had five consecutive one-on-one conversations using the "inter-patient" cross-sectional method described by Redelmeier and Lorig [17]. Before the meeting, the principal investigator (JP) who did not know the participants grouped them into conversational pairs. The conversations were private and in separate offices, and lasted about 10 minutes. Prior to each conversation, we encouraged the participants to discuss issues they felt were important with respect to their fatigue. To assist this discussion and standardize the focus on fatigue, we provided participants with three questions about the components, the severity and the burden of fatigue. At the end of each conversation, both participants of the pair were asked to separately and confidentially rate their fatigue level relative to that of their conversational partner on a single global rating scale. The single rating item asked: "Thinking of the past week, compared to this person, I have:" and the seven response categories of the Likert scale used were: "Much more fatigue", "Somewhat more fatigue", "A little bit more fatigue", "About the same fatigue", "A little bit less fatigue", "Somewhat less fatigue", "Much less fatigue". Rating forms were collected after each round, following which the next set of pre-specified pairs of participants met and the process, including the instructions, was repeated.

Questionnaires

The literature review identified seven validated self-administered fatigue instruments that appeared to be the most suitable for RA patients (Table 1). Their selection was based on the content of the instrument, documented psychometric validity, availability in English, previous use in inflammatory rheumatic diseases, the ability to be self-administered, and the number of items. Not all the selected questionnaires have been used in RA. The recall period was standardized to one week for all seven fatigue measures (even if the time frame differed in the original version of the questionnaire).

Fatigue Severity Scale (FSS) [10,11]—The FSS measures the impact of fatigue on activities of daily living. This is a 9-item questionnaire with a 7-point rating scale response format ranging from 1 ("completely disagree") to 7 ("completely agree"). It has been widely used and validated in patients with systemic lupus erythematosus (SLE) [10,11,22,23], but has not been studied in RA. The nine items are combined in a global fatigue score computed as the average of the individual item responses. The scores can range from 1 (no fatigue) to 7 (maximum fatigue).

Vitality scale of the MOS-SF36 (VT) [12]—The VT subscale of the MOS-SF36 explores both fatigue and a related concept, energy level. Item responses are rated on a 6-point Likert scale from "all the time" to "none of the time". The score can vary from 0 (the worst score) to 100 (the best score). The VT subscale has been used in many chronic rheumatic conditions, including RA [7,24].

Multidimensional Assessment of Fatigue (MAF) [2]—This questionnaire contains 16 items and covers four dimensions of fatigue: severity, distress, degree of interference in activities of daily living, and timing. Items are rated using a 10-point numerical scale (14 items) or multiple-choice (4 choices) responses (2 items). A global fatigue index (GFI) can be computed using 15 out of the 16 items and ranges from 1 (no fatigue) to 50 (severe fatigue). The MAF has been used mainly in RA [2,3,25,26].

Multidimensional Fatigue Inventory (MFI) [13]—The MFI contains 20 statements organized into five dimensions of fatigue with four statements each (general fatigue, physical

fatigue, reduced activity, reduced motivation, mental fatigue). The response-scale has five choices from agreement "yes, that is true" to disagreement "no, that is not true". A global fatigue score combining the five results ranges from 20 to 100, with higher scores indicating higher levels of fatigue. The psychometric properties of the MFI have been well documented; it has been frequently used in oncology [12] and rheumatic conditions, including primary Sjögren's syndrome (SS) [27,28], SLE [28], ankylosing spondylitis (AS) [29], and RA [7, 27].

Functional Assessment of Chronic Illness Therapy–Fatigue scale (FACIT–F) [14]—This is a widely used instrument for cancer-related fatigue. It has 13 items and a five-point Likert type rating scale (0 = "not at all" and 4 = "very much"), and explores the severity of fatigue in a uni-dimensional basis. A total score is the sum of the individual items and ranges between 0 (maximum fatigue) and 52 (no fatigue). It has been used in primary SS [30] and in RA (results published only in abstract form).

Chalder Fatigue Scale (CFS) [15]—This self-administered questionnaire was developed for use for both normal and clinical populations [31]. It consists of 11 items, covering physical and mental aspects of fatigue. The responses are on 4-point Likert scales. A total fatigue score is obtained by adding the score of all 11 items and ranges from 0 (no fatigue) to 33 (maximum fatigue). It has been used in primary SS [32], SLE [23], but not in RA.

Global assessment of fatigue—We administered a 10-point numerical Rating Scale (RS) where 0 represented "no fatigue at all" and 10 represented "fatigue as bad as it could be". Global assessment of fatigue has been widely employed in chronic inflammatory rheumatic conditions including primary SS [32], SLE [22], and RA [4,5,7].

Health Assessment Questionnaire (HAQ) [21]—This widely used validated scale assesses physical function and disability due to RA. Scores range from 0 to 3 with higher numbers representing poorer physical functioning.

Data collection and analysis

Recommended scoring methods for each questionnaire were applied to obtain raw scores. To facilitate comparisons, the raw scores were then rescaled to 0–100 point scales of increasing fatigue. Following the approach described by Redelmeier and Lorig [17], we matched each self-reported comparison rating to the differences in fatigue scores of the associated subjects on the various questionnaires. Mean differences were calculated for each questionnaire after grouping according to comparison ratings. Since one would expect that the mean difference associated with the "about the same fatigue" category should be zero, mean differences for the other categorized were "standardized" by subtracting this value as an adjustment for self-report bias. Comparative tests and confidence intervals were derived using large sample normal theory for linear contrasts.

In addition to this essentially descriptive analysis, we applied linear regression based on the model proposed by Brant et al. [18]. The model relates the multiple pairwise ratings reported by patients to underlying perceptions of fatigue relative to the mean level in the overall population. Random effects repeated-measures analysis was applied to the pairwise ratings to provide imputed fatigue indicator scores on a interval scale with equi-spaced anchors (e.g. 2 = "somewhat more fatigue", −1 = "a little bit less fatigue"). Simple linear regression was then applied to relate the imputed global fatigue values to questionnaire scores, providing slope estimates to characterize MCIDs.

SAS statistical software (SAS Institute, Cary, NC) [33] and R [34] were used for data management and statistical analysis.

Results

Nine sessions were organized. The demographic features of the 61 participants are presented in Table 2. The mean age was 60 years, with the majority being women. Subjects had an average disease duration of more than 20 years, with self-reported HAQ, disease activity, and pain scales in the mild to moderate range. As a group the fatigue severity was quite high as reflected by the mean raw and normalized scores for fatigue (Table 2).

The range of reported fatigue among the participants was sufficient to permit the determination not only of the MCIDs but also the size of the other differences between patients (Tables 3 and 4). Distribution of the fatigue contrasts covered the full spectrum from "much more fatigue" all the way to "much less fatigue" (Table 2). The distribution of the contrasts' discrepancies demonstrated relatively few major discrepancies (Table 2).

Non-parametric analysis

The mean differences of patients' scores for the seven fatigue measurement tools according to the seven possible contrasts are presented in the Figure. The mean values for the "About the same fatigue" contrast ranges from 0.1 for the MFI to 3.7 for the MAF for normalized score (Table 3). The consistently positive values correspond to optimistic self-reference bias, though confidence intervals indicate that this is significant only for the MAF (Table 3). As expected, the greater the "distance" between the participants (at the maximum: much more or much less) the greater the difference between subjects' scores for all seven instruments (Table 4 and Figure). However, some differences were larger for the "little bit less fatigue" than for "somewhat less fatigue". Such unexpected reversals were possibly arising from misclassifications in the contrasts. These distances also seem reasonably equivalent between the contiguous categories of fatigue contrast for all seven fatigue measures (Figure). The normalized positive and negative MCID estimates are presented for the usual definition ("About the same fatigue" to "A little bit less fatigue" or "A little bit more fatigue") (Table 3). For all fatigue instruments but the FACIT-F, the MCIDs were larger for the positive MCIDs compared to the negative MCIDs, though the differences were not statistically significant (Table 3).

Model based analysis

The random effects model for contrasts accounted for 61% of the total variance. There was no evidence of self-reference bias or any interview-order effect. Separate graphs of the instruments versus the estimated scores were all consistent with linearity (Figure). The linear MCID estimates and corresponding confidence intervals are given in Table 3. They are consistently larger than the non-parametric estimates, which can be explained by the statistical theory of measurement error and misclassification as follows. The non-parametric estimates are differences in the means of groups which are derived from a classification process which is not entirely reliable. As a result, mean differences are "attenuated" towards zero in relation to those that would be obtained if the classification process was infallible. While it is also true that the imputed fatigue values used in the model-based approach are subject to statistical error, the mathematical rules used to compute them incorporate adjustments to adjust for attenuation bias. Additionally, the model based estimates combine information from all groups to provide a single estimate (whose validity then rests on the assumption of linearity) with increased accuracy as indicated by narrow confidence intervals.

Discussion

The high prevalence of fatigue and its chronicity in RA patients is a major source of disability and diminished quality of life [5,7,8]. It is described as a sensation of weakness, lack of energy or tiredness, or even a sustained exhaustion associated with a decreased capacity for physical and mental work [5,35]. Measuring fatigue can be used to assess and monitor fatigue in RA, and to determine more effective treatment.

Even though the participants were ambulatory, they reported high levels of fatigue on all seven instruments. Similar fatigue scores have been reported in RA patients with the VT scale of the MOS-SF36, the MAF and MFI [2,3,7,24–27]. In the study by Barendregt et al. [27] that used the MFI there was no difference in fatigue scores between patients with primary Sjögren's syndrome and patients with RA, and both groups reported significantly more fatigue than healthy controls. No comparative data are available in RA patients with the FSS, the CFS and the FACIT–F.

A challenge of using fatigue measures is to appreciate the significance of a given change in a scale. In that respect, sensitivity to change –longitudinal validity– is one of the most important psychometric properties of health outcome measures and is the ability of the instrument to accurately detect change when it has occurred [36,37]. The sensitivity is assessed by summary statistics such as the effect size (ES) or the standardized response mean (SRM) [38]. However, most data on sensitivity do not provide information on the ability of an instrument to detect important changes or differences from the patient's perspective – "responsiveness", the MCID [9,17–20,36–40]. For example, the usual improvement following a total hip replacement measured by a health status measure is certainly much larger than one that would be detected following a less dramatic intervention, such as rehabilitation, and a much smaller improvement (or worsening) in the health status measure may still represent an important change from a patient's perspective.

Knowledge of the MCID is essential for an instrument to be used in routine clinical practice and as an outcome in clinical trials. Statistically significant differences, as presented in most studies, are dependent on sample size and do not address clinical significance from the patient's perspective. Being able to translate changes or differences in scores into clinically meaningful terms is crucial to the interpretation of the results. We stress, however, that both anchor-based cross-sectional [17] and longitudinal [9] designs provide group-based MCID estimates [41]. They are obtained by relating mean values for clinical assessments for different subjective comparison ratings or different global ratings to measurements of the underlying concept (i.e. fatigue) taken as the independent variable. In this group-based context, instruments with a large MCID are preferred, as they require smaller sample sizes for clinical trials. The most commonly used definition for the MCID is in fact individual-based, referring to the smallest difference in measured health status (i.e. fatigue instrument) that signifies an important (or a detectable) difference in a patient's actual status. To derive the individual-based MCID, one should use models for predicting values of the measured concept (i.e. fatigue) from the measured health state values and in these, the smaller estimate the individual-based MCID correspond to the more sensitive instrument.

This seemingly paradoxical inverse relation between group-based and individual-based MCID's is easiest to describe for regression based estimates, where the individual-based MCID is mathematically equal to the group-based MCID times the ratio of the variance of both estimates (the variance being larger for the individual-based estimate reflecting the additional uncertainty in individual observations on patients). To summarize, the more sensitive an instrument, the larger the group-based MCID and the smaller the individual-based MCID. From

our results, the CFS appears to be the weakest instrument with the lowest standardized (group-based MCID) of 0.54; the other instruments are similar (0.67–0.76).

With experience, clinicians have developed an intuitive sense of MCID for many measures – especially the physiologic measures they use routinely. This may not hold for health status questionnaires where the meaning of change is less apparent intuitively, particularly because the measures are seldom used in clinical practice.

We found that for the seven tested fatigue instruments the estimates of the negative MCIDs (from "About the same fatigue" to "A little bit more fatigue") range from 4.4 for FSS to 13 for the FACIT-F and from 10.1 for the FACIT-F to 17.4 for the MAF for the positive MCIDs. Using the linear regression model the global estimation of the MCIDs range from 9.9 for the CFS to 20.2 for the global assessment. Our results showed a consistent linear relationship between the 7-point change scale and the MCIDs estimates (Figure) that is evidence in favor of the validity of the linear regression model that we used [18]. The standardized MCID ranges from 0.54 for the CFS to 0.76 for the MFI. The squared ratios of standardized MCIDs indicate relative sample size requirements. For example, a study incorporating the CFS as primary outcome would require twice the sample size as one utilizing the MFI. However, it must be emphasized that the concept of fatigue captured by these instruments is operationalized with different dimensions such as the severity/intensity, timing/frequency, duration, or impact (physical, mental or emotional, social) (Table 1).

The MCID of the VT scale of the MOS-SF36 has been estimated in another study in RA to 11.1 point (a 29% improvement compared to the baseline score that was 38.6) [42]. This is comparable to our MCID estimate of 11.9. This is the only previously published result of a MCID for the fatigue instruments that we tested.

Our study also gives an idea of the moderate and important differences defined from the patients' perspective (between 2 or 3 contiguous contrasts' categories, respectively). As one might anticipate, the mean change per question in questionnaire scores associated with a global rating of "unchanged" approximates zero, and the larger the difference as assessed by global ratings, the larger the change in the fatigue measurement tools.

The various methods proposed to estimate the MCID have been reviewed [39]. The explicit methods use anchor-based approaches based on cross-sectional or longitudinal designs to compare health status measures that have clinical relevance [39]. The cross-sectional technique of MCID estimation used in our study developed by Redelmeier and Lorig [17] has been used to estimate the MCID of the HAQ in RA patients [20] and in other non rheumatic chronic conditions [18,43]. An important limitation of this design is that subjective comparison ratings may not be a true change and should not be used to assess within patient changes [39]. These subjective comparison ratings may represent real differences in patients' perception of clinically important differences on a given instrument at a given point in time, but may not estimate the degree of clinical change over time considered by an individual to be meaningful.

The longitudinal or transitional technique of Jaeschke et al. [9] has been applied in rheumatic and non-rheumatic chronic conditions [19,39,42,44–46]. The estimation of MCID is based on the intra-personal variation of the outcome score between the onset and the end of a therapeutic intervention that is related to a single transition question about the change that occurred. The MCID for improvement is then defined as the difference between the mean effect of the intervention assessed by the instrument of those who rated themselves as slightly better and those who rated themselves as about the same (thus allowing one to eliminate the systematic bias previously discussed). The MCID for worsening is defined in a similar way using the slightly worse as anchor. The longitudinal design has several limitations. The presence of an intervention is likely to influence the result of the MCID compared to the cross-sectional study,

as the more the individual expects to improve due to the intervention, the more the individual will rate him/herself as having at least a small improvement. It may be more difficult to obtain estimates of MCIDs for worsening in the longitudinal design (at least for higher level of worsening in the transition item). Also, the longitudinal study design implies a retrospective judgment as the individuals have to recall their initial health state which may be difficult for long-term trials [39,44]. The cross-sectional design is easier to implement, as it does not need any follow-up or intervention. Nevertheless, the MCIDs estimates obtained with the intrapersonal and interpersonal judgment are similar [43].

The importance of the anchors used and the number of response categories of global rating scales or transition question has been rarely addressed. Response options vary between 5 [44] and 15-point scales [19]. The more common scales have seven points, as in our study. The anchors' wording is important as it is the arbitrary definition of the MCID; this has not been formally tested. Subjects seemed unable to distinguish more than seven different points in a scale [19,47].

The MCID as estimated in the literature applies to the detection of a minimally "detectable" or "perceptible" change/difference without consideration of whether the change/difference was "important" or "significant" from the patient's perspective [40,46]. It seems reasonable to assume that the true minimal clinically important difference would be at least the same or more probably greater than the one that is "only" perceptible. Certainly, patients' characteristics such as coping strategies, expectations and experiences with the health care system may influence the estimates.

Clinically meaningful changes in health status measures can also be assessed from distribution-based approaches with the computation of various indices to assess the responsiveness, including the ES or the SRM [39]. Arbitrarily, for example, an ES or a SRM of 0.20 has been proposed to represent "small" changes, representing a minimal clinically important difference [39]. The distribution-based methods are less appealing and less intuitive than the methods that explicitly examine the relationship between the health status measure and an independent transitional or comparative global item to explain the meaning of a change/difference. However, several studies show that both methods provide equivalent estimates of MCID [39, 48]. Although it was disputed [49,50], in a provocative paper, Norman et al. [47] attempted to show that, the minimally important difference of one half a standard deviation (0.5 SD) could be a universal standard for health status instruments.

Fatigue is a major issue for people suffering from RA and should be included in the outcome criteria of clinical trials and in the routine assessment of patients during clinical care. Our study provides clinicians and researchers with quantitative information regarding fatigue instruments that could be used in RA. It will improve communication about this symptom with patients and will assist in planning future therapeutic trials.

References

1. Belza BL, Henke CJ, Yelin EH, Epstein WV, Gilliss CL. Correlates of fatigue in older adults with rheumatoid arthritis. Nurs Res 1993;42:93–9. [PubMed: 8455994]

2. Belza BL. Comparison of self-reported fatigue in rheumatoid arthritis and controls. J Rheumatol 1995;22:639–43. [PubMed: 7791155]

3. Belza Tack B. Self-reported fatigue in rheumatoid arthritis. A pilot study. Arthritis Care Res 1990;3:154–7. [PubMed: 2285754]

4. Wolfe F, Hawley DJ, Wilson K. The prevalence and meaning of fatigue in rheumatic disease. J Rheumatol 1996;23:1407–17. [PubMed: 8856621]

5. Hewlett S, Cockshott Z, Byron M, Kitchen K, Tipler S, Pope D, et al. Patients' perceptions of fatigue in rheumatoid arthritis: overwhelming, uncontrollable, ignored. Arthritis Rheum 2005;53:697–702. [PubMed: 16208668]

6. Huyser BA, Parker JC, Thoreson R, Smarr KL, Johnson JC, Hoffman R. Predictors of subjective fatigue among individuals with rheumatoid arthritis. Arthritis Rheum 1998;41:2230–7. [PubMed: 9870880]

7. Rupp I, Boshuizen HC, Jacobi CE, Dinant HJ, van den Bos GAM. Impact of fatigue on health-related quality of life in rheumatoid arthritis. Arthritis Care Res 2004;51:578–85.

8. Suurmeijer TP, Waltz M, Moum T, Guillemin F, van Sonderen FL, Briançon S, et al. Quality of life profiles in the first years of rheumatoid arthritis: results from the EURIDISS longitudinal study. Arthritis Care Res 2001;45:111–121.

9. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15. [PubMed: 2691207]

10. Schwartz JE, Jandorf L, Krupp LB. The measurement of fatigue: a new instrument. J Psychosom Res 1993;37:753–62. [PubMed: 8229906]

11. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989;46:1121–3. [PubMed: 2803071]

12. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I Conceptual framework and item selection. Med Care 1992;30:473–83. [PubMed: 1593914]

13. Smets EMA, Garssen B, Bonke B, De Haes JCJM. The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39:315–25. [PubMed: 7636775]

14. Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. J Pain Symptom Manage 1997;13:63–74. [PubMed: 9095563]

15. Chalder T, Berelowitz G, Pawlikowska T, Watts L, Wessely S, Wright D, et al. Development of a fatigue scale. J Psychosom Res 1993;37:147–53. [PubMed: 8463991]

16. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24. [PubMed: 3358796]

17. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993;153:1337–42. [PubMed: 8507124]

18. Brant R, Sutherland L, Hilsden R. Examining the minimum important difference. Stat Med 1999;18:2593–603. [PubMed: 10495458]

19. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81–7. [PubMed: 8283197]

20. Wells GA, Tugwell P, Kraag GR, Baker PRA, Groh BJ, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557–60. [PubMed: 8478873]

21. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales. J Rheumatol 1982;9:789–93. [PubMed: 7175852]

22. Krupp LB, LaRocca NG, Muir J, Steinberg AD. A study of fatigue in systemic lupus erythematosus. J Rheumatol 1990;17:1450–2. [PubMed: 2273484]

23. Tench CM, McCurdie I, White PD, D'Cruz DP. The prevalence and associations of fatigue in systemic lupus erythematosus. Rheumatology 2000;39:1249–54. [PubMed: 11085805]

24. Talamo J, Frater A, Gallivan S, Young A. Use of the short form 36 (SF36) for health status measurement in rheumatoid arthritis. Br J Rheumatol 1997;36:463–9. [PubMed: 9159541]

25. Jump RL, Fifield J, Tennen H, Reisine S, Giuliano AJ. History of affective disorder and the experience of fatigue in rheumatoid arthritis. Arthritis Care Res 2004;51:239–45.

26. Neuberger GB, Press AN, Lindsley HB, Hinton R, Cagle PE, Carlson K, et al. Effects of exercise on fatigue, aerobic fitness, and disease activity measures in persons with rheumatoid arthritis. Res Nurs Health 1997;20:195–204. [PubMed: 9179174]

27. Barendregt PJ, Visser MRM, Smets EMA, Tulen JHM, van den Meiracker AH, Boomsma F, et al. Fatigue in primary Sjögren's syndrome. Ann Rheum Dis 1998;57:291–5. [PubMed: 9741313]

28. Godaert GLR, Hartkamp A, Geenen R, Garssen A, Kruize AA, Bijlsma JWJ, Derksen RHWM. Fatigue in daily life in patients with primary Sjögren's syndrome and systemic lupus erythematosus. Ann NY Acad Sci 2002;966:320–6. [PubMed: 12114289]

29. van Tubergen A, Coenen J, Landewé R, Spoorenberg A, Chorus A, Boonen A, van der Linden S, van der Heijde D. Assessment of fatigue in patients with ankylosing spondylitis: a psychometric analysis. Arthritis Care Res 2002;47:8–16.

30. Walker J, Gordon T, Lester S, Downie-Doyle S, McEvoy D, Pile K, et al. Increased severity of lower urinary tract and daytime somnolence in primary Sjögren's syndrome. J Rheumatol 2003;30:2406–12. [PubMed: 14677185]

31. Pawlikowska T, Chalder T, Hirsch SR, Wallace P, Wright DJM, Wessely SC. Population based study of fatigue and psychological distress. BMJ 1994;308:763–6. [PubMed: 7908238]

32. Lwin CTT, Bishay M, Platts RG, Booth DA, Bowman SJ. The assessment of fatigue in primary Sjögren's syndrome. Scand J Rheumatol 2003;32:33–7. [PubMed: 12635943]

33. SAS/STAT user's guide. Version 8. Cary, N.C: SAS Institute; 2000. software

34. Ihaka R, Gentleman R. R: a language for data analysis and graphic. Journal of Computational & Graphical Statistics 1996;5:99–314.

35. Berrios GE. Feelings of fatigue and psychopathology: a conceptual history. Compr Psychiatry 1990;31:140–51. [PubMed: 2178863]

36. Fortin P, Stucki G, Katz JN. Measuring relevant change: an emerging challenge in rheumatologic clinical trials. Arthritis Rheum 1995;38:1027–30. [PubMed: 7639797]

37. Liang MH. Longitudinal construct validity. Establishment of clinical meaning in patient evaluative instruments. Med Care 2000;38(Suppl II):II-84–II-90. [PubMed: 10982093]

38. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12:349–62. [PubMed: 12797708]

39. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407. [PubMed: 12812812]

40. Wright JG. The minimal important difference: who's to say what is important? J Clin Epidemiol 1996;49:1221–2. [PubMed: 8892487]

41. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. J Rheumatol 2001;28:406–12. [PubMed: 11246688]

42. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE. Determining the minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 2000;43:1478–87. [PubMed: 10902749]

43. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol 1996;49:1215–9. [PubMed: 8892486]

44. Angst F, Aeschlimann A, Michel BA, Stucki G. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. J Rheumatol 2002;29:131–8. [PubMed: 11824949]

45. Schwartz AL, Meek PM, Nail LM, Fargo J, Lundquist M, Donofrio M, et al. Measurement of fatigue determining minimally clinically important clinical differences. J Clin Epidemiol 2002;55:239–44. [PubMed: 11864794]

46. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369–78. [PubMed: 7490600]

47. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life. The remarkable universality of half a standard deviation. J Clin Epidemiol 2003;41:582–92.

48. Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med Care 2001;39:1039–47. [PubMed: 11567167]

49. Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. Med Care 2003;41:593–6. [PubMed: 12719682]

50. Wright JG. Interpreting health-related quality of life scores. The simple rule of seven may not be so simple. Med Care 2003;41:597–8. [PubMed: 12719683]
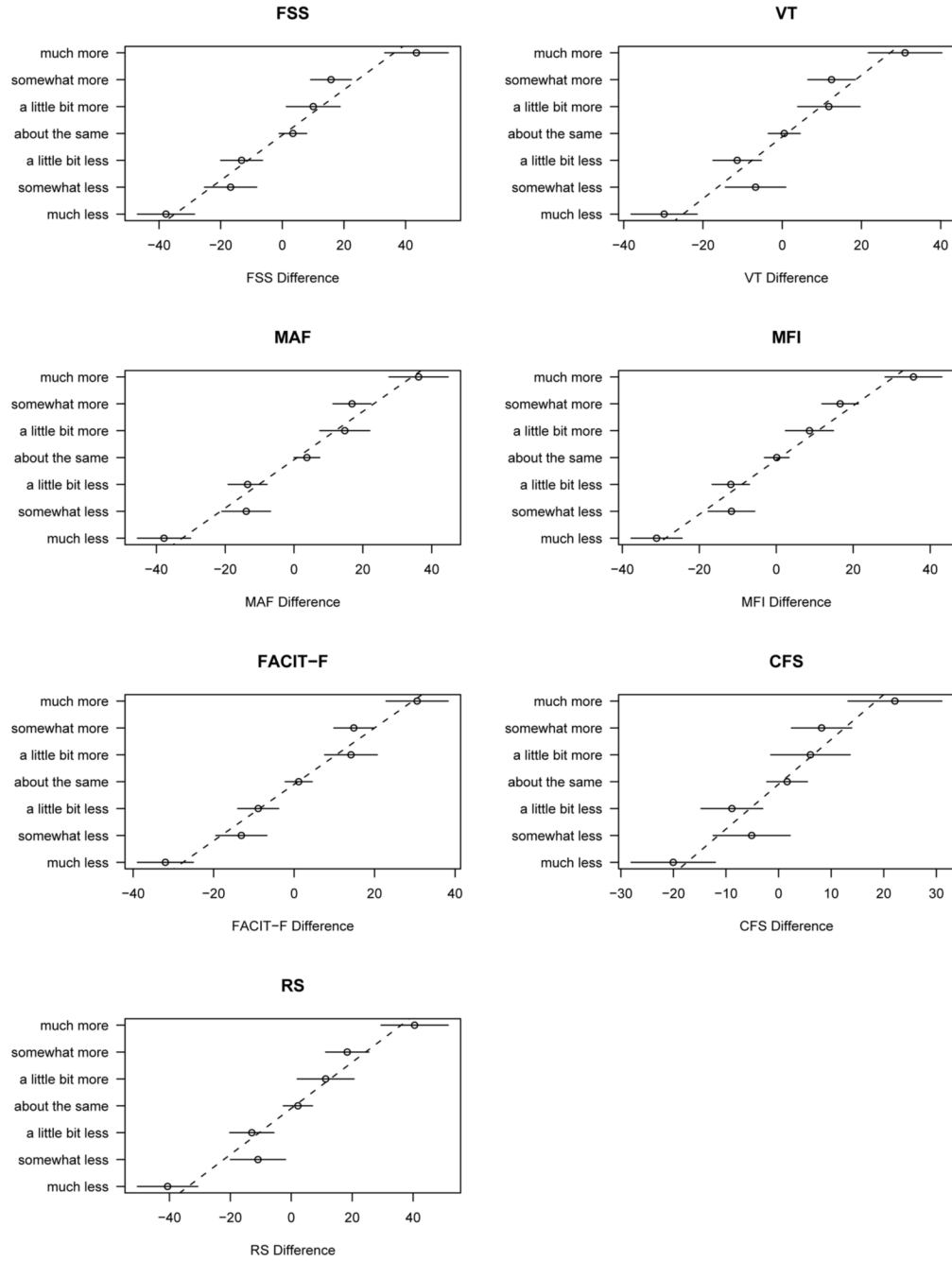
Figure. Mean differences of fatigue score for seven measurement tools in relation to pairwise contrasts for 61 patients with rheumatoid arthritis

Legend. The results are presented using normalized fatigue scores from 0 to 100 with mean differences of fatigue score and 95% confidence interval for all seven contrast categories, and regression lines. The plotted lines represent the regression of differences against integer scores (i.e., inverse regression).

Abbreviations. FSS: Fatigue Severity Scale, VT: Vitality scale of the MOS-SF36, MAF: Multidimensional Assessment of Fatigue, MFI: Multidimensional Fatigue Inventory, FACIT–F: Functional Assessment of Chronic Illness Therapy–Fatigue scale, CFS: Chalder Fatigue Scale, RS: global assessment of fatigue using a 10-point numerical Rating Scale.

Table 1

Characteristics of the seven self-administered fatigue instruments used in the study.

Fatigue instrument, abbreviated name [ref]	No Items	Response format	Score (range)[+]	Explored domains	Already used in RA	Documented psychometric properties[++]
FSS [10,11]	9	7-point RS	1–7	Severity, physical, mental, and social impact	−	Scaling, reliability, construct validity, responsiveness, IRT
VT [12]	4	6-point LS	0–100	Timing (fatigue to energy)	+	Extensively documented
MAF [2]	15	10-point RS	1–50	Severity, physical, and social impact	+	Reliability, construct validity
MFI [13]	20	5-point RS	20–100[*]	General fatigue, physical fatigue, reduced activity, reduced motivation, mental fatigue	+	Reliability, construct validity
FACIT–F [14]	13	5-point LS	0–52	Severity, role and social impact	−	Reliability, construct validity, IRT
CFS [15]	11	4-point LS	0–33	Physical and mental fatigue	−[**]	Reliability
Global RS	1	10-point RS	0–10	Severity	+	—

Legend. + for all but VT and FACIT–F, higher scores indicate higher level (severity or impact) of fatigue.

[*] a score for each of the 5 dimensions could be computed and ranges from 4 to 20.

[**] with slightly different anchors. ++ psychometric properties may have been documented in other diseases than RA.

Abbreviations. FSS: Fatigue Severity Scale, VT: Vitality scale of the MOS-SF36, MAF: Multidimensional Assessment of Fatigue, MFI: Multidimensional Fatigue Inventory, FACIT–F: Functional Assessment of Chronic Illness Therapy–Fatigue scale, CFS: Chalder Fatigue Scale, RS: Rating Scale. RA: rheumatoid arthritis. LS: Likert scale. IRT: item response theory.

Table 2

Clinical characteristics and fatigue scores for the seven fatigue measurement scales in the 61 participants, and distribution of fatigue contrasts.

Clinical characteristics	Mean (SD) or N (%)		
Age, yrs	62.1 (14.8)		
Women	52 (85%)		
Disease duration, yrs	20.2 (14.4)		
HAQ (0–3)	1.2 (0.8)		
Disease activity (0–10)	4.3 (2.3)		
Pain level (0–10)	4.3 (2.4)		
Fatigue scores	Mean (SD)	Median (range)	Normalized Mean (SD)
FSS	4.7 (1.6)	4.9 (1.3 – 6.9)	61.5 (26.2)
VT	45.9 (21.0)	45 (0 – 100)	54.1 (21.0)
MAF	27.9 (10.3)	29.5 (1 – 46.7)	49.1 (23.7)
MFI	60.1 (16.6)	62 (22 – 91)	50.1 (20.7)
FACIT–F	29.4 (10.6)	28 (8 – 52)	43.5 (20.3)
CFS	16.0 (5.8)	15 (1 – 28)	48.3 (17.6)
RS	5.1 (2.7)	5 (0 – 10)	51.1 (26.7)
Fatigue contrasts	N (%)	**Contrasts discrepancies**	N (%)
Much less fatigue	49 (15.9)	Mirror	54 (35.1)
Somewhat less fatigue	32 (10.4)	Minor	48 (31.2)
A little bit less fatigue	44 (14.3)	Moderate	29 (18.8)
About the same fatigue	69 (22.4)	Major	23 (14.9)
A little bit more fatigue	31 (10.1)		
Somewhat more fatigue	43 (14.0)		
Much more fatigue	40 (13.0)		

Legend. For all but VT and FACIT–F, higher raw scores indicate higher level (severity or impact) of fatigue. Normalized scores range from 0 to 100 with higher scores indicating higher fatigue levels. Disease activity and pain were self-rated on 10-point numerical rating scales. A contrast was defined as the subjective comparison rating obtained at the end of a one-on-one conversation, between both participants of the pair. Each one-on-one conversation provided 2 contrasts, and the 61 participants each involved in 5 to 6 one-on-one conversations provided 308 contrasts. A mirror contrast between two conversational partners was defined as one that should theoretically be expected ("About the same fatigue" and "About the same fatigue", "Much more fatigue" and "Much less fatigue", etc…). Minor, moderate and major discrepancies were defined for respectively 1, 2 and 3 or more unexpected category differences in the subjective rating scale (For example, a minor discrepancy was defined for "About the same fatigue" and "A little bit more fatigue", a moderate discrepancy as "About the same fatigue" and "Somewhat more fatigue", etc…).

Abbreviations. HAQ: Health Assessment Questionnaire. FSS: Fatigue Severity Scale, VT: Vitality scale of the MOS-SF36, MAF: Multidimensional Assessment of Fatigue, MFI: Multidimensional Fatigue Inventory, FACIT–F: Functional Assessment of Chronic Illness Therapy–Fatigue scale, CFS: Chalder Fatigue Scale, RS: 10-point Rating Scale.

Table 3

Non-parametric and regression-based estimates of minimal clinically important differences for seven fatigue instruments in rheumatoid arthritis.

Fatigue Instrument	Non-parametric MCID estimates				Regression based MCID estimates	
	About the Same *	About the same ↔ A little bit more	About the same ↔ A little bit less	P-value	Regression-based MCID **	Standardized MCID
FSS	3.4 (−1.0 – 7.9)	16.7 (8.1 – 25.2)	6.6 (−3.4 – 16.6)	0.14	20.2 (15.5 – 25.0)	0.74
VT	0.51 (−3.5 – 4.5)	11.9 (4.2 – 19.6)	11.3 (2.2 – 20.3)	0.92	14.8 (10.6 – 19.0)	0.67
MAF	3.7 (0.01 – 7.5)	17 (10.1 – 24.4)	11 (2.7 – 19.4)	0.27	18.7 (14.5 – 22.9)	0.75
MFI	0.13 (−3.1 – 3.3)	11.9 (5.8 – 18.1)	8.5 (1.3 – 15.7)	0.48	16.6 (13.0 – 20.2)	0.76
FACIT–F	1.1 (−2.2 – 4.5)	10 (3.6 – 16.5)	13 (5.5 – 20.5)	0.56	15.9 (12.3 – 19.6)	0.75
CFS	1.6 (−2.2 – 5.5)	10.5 (3.1 – 17.9)	4.4 (−4.25 – 13.1)	0.30	9.9 (5.9 – 13.8)	0.54
RS	2.2 (−2.6 – 7.0)	15.1 (6.0 – 24.3)	9.1 (−1.6 – 19.8)	0.41	19.7 (14.6 – 24.8)	0.70

Legend. Results are estimated values with 95% confidence intervals, computed for normalized scores (0–100). P-value is for test of equality of non-parametric MCID estimates for "A little bit more fatigue" and "A little bit less fatigue".

*
Mean differences scores for the seven fatigue measurement tools for the "About the same" contrast category.

**
Accounting for the non-reproducibility of the comparison ratings, and for possible artefacts such as self-reference bias and interview-order effects.

Abbreviations. MCID: minimal clinically important difference. FSS: Fatigue Severity Scale, VT: Vitality scale of the MOS-SF36, MAF: Multidimensional Assessment of Fatigue, MFI: Multidimensional Fatigue Inventory, FACIT–F: Functional Assessment of Chronic Illness Therapy–Fatigue scale, CFS: Chalder Fatigue Scale, RS: 10-point Rating Scale.

Table 4

Non-parametric estimates for seven fatigue instruments for the "Somewhat more (less)" and the "Much more (less)" fatigue comparison rating categories anchored to the "About the same fatigue" category.

Fatigue instrument	About the same ↔ Somewhat more	About the same ↔ Somewhat less	About the same ↔ Much more	About the same ↔ Much less
FSS	12.4 (3.7 – 21.1)	20.2 (9.2 – 31.3)	40.1 (28.5 – 51.7)	41.2 (30.4 – 52.0)
VT	11.9 (4.1 – 19.8)	7.2 (−2.7 – 17.2)	30.5 (20.1 – 40.9)	30.3 (20.6 – 40.0)
MAF	13.1 (5.9 – 20.4)	17.6 (8.4 – 26.9)	32.5 (22.8 – 42.1)	41.5 (32.5 – 50.5)
MFI	16.5 (10.2 – 22.7)	11.8 (3.8 – 19.7)	35.5 (27.2 – 43.8)	31.2 (23.5 – 38.9)
FACIT–F	13.7 (7.2 – 20.2)	14.2 (5.9 – 22.5)	29.4 (20.7 – 38.1)	33.1 (25.1 – 41.2)
CFS	6.6 (−0.99 – 14.1)	6.7 (−2.8 – 16.3)	20.5 (10.5 – 30.5)	21.7 (12.4 – 31.0)
RS	16.2 (6.9 – 25.5)	13.1 (1.3 – 24.9)	38.3 (25.9 – 50.7)	42.8 (31.3 – 54.3)

Legend. Results are means with 95% confidence intervals

Abbreviations. FSS: Fatigue Severity Scale, VT: Vitality scale of the MOS-SF36, MAF: Multidimensional Assessment of Fatigue, MFI: Multidimensional Fatigue Inventory, FACIT–F: Functional Assessment of Chronic Illness Therapy–Fatigue scale, CFS: Chalder Fatigue Scale, RS: 10-point Rating Scale.