

---

# Assessing sleep in fibromyalgia: investigation of an alternative scoring method for the Jenkins Sleep Scale based on data from randomised controlled studies

---

B.K. Crawford<sup>1</sup>, E.C. Piau<sup>1</sup>, C. Lai<sup>2</sup>, P. Sarzi-Puttini<sup>3</sup>

---

<sup>1</sup>Mapi Values, Tokyo, Japan; <sup>2</sup>Jazz Pharmaceuticals, Inc.; <sup>3</sup>Luigi Sacco University Hospital, Milan, Italy.

Bruce K. Crawford, MA, MPH  
Elisabeth C. Piau, Pharm D, MA  
Chinglin Lai, PhD  
Piercarlo Sarzi-Puttini, MD

Please address correspondence  
and reprint requests to:

Bruce Crawford,  
General Manager, Asia,  
Mapi Values, Arco Tower 12F,  
1-8-1 Shimomeguro, Meguro,  
Tokyo 153-0064, Japan  
E-mail: bruce.crawford@mapivalues.com

Received on December 3, 2010; accepted  
in revised form on December 9, 2010.

Clin Exp Rheumatol 2010; 28 (Suppl. 63):  
S100-S109.

© Copyright CLINICAL AND  
EXPERIMENTAL RHEUMATOLOGY 2010.

**Key words:** fibromyalgia, sleep,  
outcome assessment (health care),  
psychometrics

*Competing interests:* This research was supported by Jazz Pharmaceuticals who participated in the study design, data analysis and interpretation, and preparation and review of the manuscript. Drs Crawford and Piau were employees of Mapi Values at the time the research was conducted and the manuscript was developed. Mapi Values is a consultancy whose activities on the project were funded by Jazz Pharmaceuticals. Dr Lai is an employee of Jazz Pharmaceuticals, Inc., Dr Sarzi-Puttini is an investigator in Study 06-009 sponsored by Jazz Pharmaceuticals, Inc.

## ABSTRACT

**Objectives.** To investigate the validity of a rescored version of the Jenkins Sleep Scale (JSS) to assess the extent of possible bias of a 4-week recall period in assessing sleep in patients with fibromyalgia.

**Methods.** A rescoring algorithm of the JSS was developed. The psychometric properties of the rescored JSS were examined using blinded, observed data from a Phase 2 trial (n=195) in subjects with fibromyalgia. In addition, data from two Phase 3, randomised, controlled trials (n=1,121) in subjects with fibromyalgia were used to further validate the rescored JSS by conducting correlation analyses with other assessments expected to correlate with sleep. These included fatigue and tiredness items from the Fibromyalgia Impact Questionnaire (FIQ), the Functional Outcomes of Sleep Questionnaire (FOSQ), and the Short Form-36 (SF-36™) Vitality scale.

**Results.** Construct validity of the rescored JSS was found to be acceptable, with an internal consistency reliability of  $\alpha=0.70$ . Test-retest reliability on stable subjects, defined using the FIQ total score, was also acceptable (ICC=0.70). Moderate to high correlations (Pearson  $r>0.66$ ) were found with two FIQ items, addressing fatigue and non-restorative sleep, and the SF-36™ Vitality scale; correlations with the original JSS were similar. Both JSS versions were found to be responsive ( $p<0.0001$ ), and the rescored version accounted for 90% of the variance captured in the original version.

**Conclusion.** These results showed the rescored JSS performed similarly to the original scale, suggesting the original scale's 4-week recall period did not introduce substantial bias in capturing the experience of fibromyalgia-related sleep disturbances.

## Introduction

Fibromyalgia is a complex disease with a hallmark symptom of widespread pain (1, 2). Fibromyalgia is a chronic musculoskeletal disorder with an estimated 2–3% prevalence rate in the United States and 2.9% in Europe (3); it is the second most common disorder observed by rheumatologists after osteoarthritis. Fibromyalgia is estimated to be approximately seven times more common in women than men, and its prevalence increases with age, with the highest prevalence among those 60–79 years of age (4).

Most patients with fibromyalgia report experiencing fatigue and sleep disturbances including non-restorative sleep and difficulties falling asleep (5). Thus sleep assessments are critical to capturing a patient's full experience with fibromyalgia and its treatment.

The Jenkins Sleep Scale (JSS) is a self-completed instrument assessing sleep symptoms (6). Improvement in quality of sleep, as measured by the JSS, has been shown to correlate with improved fibromyalgia pain symptoms (7). The four-item JSS captures how often during the last month a subject experienced sleep problems. At the time of the development of the questionnaire (1988), it was deemed relevant to capture clinically relevant variations in sleeping patterns through a frequency report of sleeping difficulties (e.g. insomnia) within the past month. A 4-week recall period was commonly used in generic instruments such as the Medical Outcomes Study Short Form-36 (SF-36™), (8, 9) and other sleep instruments such as the Leeds Sleep Evaluation Questionnaire (10) and the Medical Outcome Study (MOS) Sleep Questionnaire (11) as it is deemed relevant to provide a comprehensive assessment of patients' sleep disturbance in clinical practice setting because of

varying pain severity and day-to-day variation in sleep quality.

In a context of documenting treatment benefit, a patient-completed scale with a long assessment period might be subject to recall bias. To investigate the impact of potential recall bias on the assessment of sleep disturbances in fibromyalgia patients, a new, intuitive scoring algorithm of the JSS was developed and its rescored scale's psychometric performance (construct validity, reliability, and ability to detect a change) were examined in patients with fibromyalgia.

### Materials and methods

An intuitive rescored algorithm was developed prior to performing any data analysis to avoid bias. The psychometric properties (reliability, construct validity, and ability to detect a change) of the rescored JSS were examined using blinded, observed data from a randomised, double-blind, placebo-controlled, parallel-group, Phase 2 trial (n=195) (12).

Once the above psychometric properties of the rescored JSS were examined and found to be acceptable, data from two Phase 3 randomised, double-blind, placebo-controlled, parallel-group studies in fibromyalgia patients (n=548, Trial 1; n=573, Trial 2) were used to further evaluate the responsiveness of the rescored JSS with respect to pain, tiredness, and fatigue. When validating questionnaires, sample sizes are typically determined based on planned analyses rather than anticipated treatment effect size and power (13, 14). The sample size for the Phase 2 and 3 studies used in this research was determined based on the primary efficacy endpoints used in the respective studies (a composite of pain severity, functioning, and patient global assessment of change for the Phase 2 study and the proportion of subjects who had at least 30% reduction in pain severity from baseline to study endpoint for Phase 3 studies) and was much larger than the sample size of typical studies for validating questionnaires. The analyses in this research were conducted on the intent-to-treat (ITT) population, which included all randomised subjects in the studies used.

The three studies were conducted according to international standards for clinical research including documentation of patients' consent to participate in the trials.

### Instruments

The JSS was used to collect information about how the study medication affected subjects' sleep in the past month. This questionnaire comprises four items: *During the past month did you have trouble falling asleep?*; *During the past month did you wake up several times per night?*; *During the past month did you have trouble staying asleep, including waking up far too early?*; *During the past month did you wake up after your normal amount of sleep feeling tired and worn out?* Each item was rated on a 6-point Likert scale based on the frequency of the problems, where 0 indicated no sleep problems and 5 indicated frequent sleep problems. The JSS is scored as the sum of the four items, resulting in a score from 0–20 with higher scores reflecting greater sleep problems.

It was hypothesized that recalling the exact number of nights subjects experienced sleep problems within the previous month might be prone to recall bias, reflected by an increased likelihood of misclassification (*e.g.* responding 9 nights rather than 7 due to recall issues). To investigate the impact of the possible recall bias, the JSS was rescored based on an intuitive algorithm so that response options could be combined and capture the impact of the sleep problems in terms of a period of time (*e.g.* less than half the time) rather than exact number of days as described below:

- 0: not at all → 0: not at all
- 1: 1–3 nights → 1: less than ½ the time
- 2: 4–7 nights → 1: less than ½ the time
- 3: 8–14 nights → 1: less than ½ the time
- 4: 15–21 nights → 2: greater than ½ the time
- 5: 22–31 nights → 2: greater than ½ the time

The final score was the sum of the individual scores of all 4 items, ranging

from 0–8 with higher scores reflecting greater sleep problems. The final score was missing if any of the questions had a missing score.

In addition to the JSS, the following measures were collected in the Phase 2 and Phase 3 studies: Fatigue Visual Analogue Scale (VAS), Pain VAS, Fibromyalgia Impact Questionnaire FIQ (15), Functional Outcomes of Sleep Questionnaire (FOSQ) (16), SF-36™ v2, Patient Global Impression of Change (PGIC), and Clinical Global Impression of Change (CGIC).

The Fatigue VAS and the Pain VAS are global assessments of subjects' current severity levels of fatigue and pain, respectively, using a horizontal line anchored at "0 = No fatigue/pain" and "100 = Worst imaginable fatigue/pain."

The FIQ is composed of 20 items assessing 10 subscales. It measures the subject's fibromyalgia status, progress, and outcomes based on the experience in the previous 7 days. The total FIQ score (17), derived from the sum of the ten subscales, has been shown to reflect the total impact of fibromyalgia with scoring ranging from 0 to 100 (higher scores indicating a greater impact). Item 16 of the FIQ ("*How tired have you been?*") and item 17 ("*How have you felt when you get up in the morning?*"), both scored from 0–10, were of particular interest as they measure aspects related to sleep quality and were used to examine construct validity of the rescored JSS.

The FOSQ was used to determine the impact of daytime sleepiness and tiredness on daily activities. The questionnaire assesses general productivity, social outcome, activity level, vigilance, and intimate relationship and sexual activities, which are summed into a total score ranging from 5 to 20, with lower scores indicating a greater impact of sleepiness and tiredness on activities of daily living.

The SF-36™ v2, with a 4-week recall period, was used to collect information about how the study medication affected the subject's health-related quality of life (HRQL). The 4-item Vitality subscale ranges from 0 to 100, with higher scores indicating a higher level of energy.

The Epworth Sleepiness Scale (ESS), assessed only in the Phase 2 study, was used to collect subjective ratings of daytime sleepiness. The total score, the sum of all 8 items assessing the likelihood of dozing off or falling asleep during activities, ranges from 0 to 24, with higher scores indicating a higher level of daytime sleepiness.

Subjects (PGIc) and clinicians (CGIc) rated their impression of change in fibromyalgia severity and in the subject's overall condition since baseline (*i.e.* prior to first dose of study medication) using a 7-point scale anchored on 1 = "very much better/improved" and 7 = "very much worse."

### Analyses

The analysis of the psychometric properties began with the data from the Phase 2 study. Demographic characteristics of the study population were analysed at baseline (Visit 4, randomisation) using descriptive statistics. Quality of completion of the JSS was evaluated by examining the number of missing responses for each item over time. Item distributional characteristics of the rescored JSS, including floor and ceiling effects, were examined at baseline (Visit 4). Floor and ceiling effects were interpreted in relation to the severity of the condition experienced by the study population. If floor or ceiling effects were too pronounced and/or mean scores were very low or high (respectively) at baseline, it could interfere with the ability of the scale to detect changes over time.

The rescored JSS was tested for internal consistency reliability at baseline (Visit 4) to assess how well the items fit together in the same scale. This criterion was considered to be met if the Cronbach's alpha coefficient was  $\geq 0.70$  (18, 14). The impact of item removal on internal consistency reliability was examined. Cronbach's alpha was calculated with each item removed from the total scale to assess the impact. If the removal of an item caused the alpha to increase substantially, then that item might not be fitting well in the scale.

Test-retest reliability consists of measuring the degree to which an instrument yields similar scores at differ-

ent time points in stable subjects. The test-retest reliability was computed for the total score of the rescored JSS at Visit 5 (Week 2) and Visit 6 (Week 4). Test-retest reliability estimates were calculated using the Shrout and Fleiss intraclass correlation (ICC). A correlation  $\geq 0.70$  or greater was evidence of acceptable test-retest reliability.

It is important to evaluate a stable population in test-retest reliability as the goal is to evaluate the amount of measurement error in the instrument. Five criteria were selected to identify stable subjects for the evaluations of test-retest reliability:

- Responses from Pain VAS:  
Stable subjects were defined as those whose mean Pain VAS in the 2 weeks preceding Visit 5 and the 2 weeks preceding Visit 6 were similar (within  $\pm 5$  points).
- Responses from Fatigue VAS:  
Stable subjects were defined as those whose mean Fatigue VAS in the 2 weeks preceding Visit 5 and the 2 weeks preceding Visit 6 were similar (within  $\pm 5$  points).
- Responses from FIQ:  
Subjects whose FIQ total score changes from Visit 5 to Visit 6 were less than the minimal important change (range from -14% to +14%) were classified as stable (19).
- Responses from ESS:  
Stable subjects were defined as those whose absolute ESS total score change from Visit 5 to Visit 6 was  $\leq 2$ . This was based on a previous study, in patients with allergic rhinitis, where the placebo group had an increase of 1.16 points (20).
- Responses from FIQ item 17 ("How have you felt when you get up in the morning?"):  
As done with the FIQ total score, subjects whose FIQ item 17 changes from Visit 5 to Visit 6 were less than the minimal important change (range from -14% to +14%) were classified as stable (19).

Construct validity, including the scale's concurrent validity with other patient-completed measures, was measured by examining the correlations of the rescored JSS with other patient-reported

outcome (PRO) scales. The Pain VAS, Fatigue VAS, FOSQ, SF-36™, FIQ, and ESS were identified a priori for checking the construct validity. A logical pattern of correlations was expected to exist between the rescored JSS and these measures. Domains with similar content or symptoms evolving with the same pattern as patients' sleep problems were expected to correlate more highly than domains with less similar content. Pearson's Correlation Coefficient was used unless the data failed normality tests, then correlations were calculated using Spearman's Correlation Coefficient.

Responsiveness refers to the ability of a measure to reflect underlying change (21, 22). The rescored JSS was expected to be responsive to changes in sleep and was evaluated by changes from baseline to Week 8, the study duration of the Phase 2 study.

In the first set of responsiveness analyses, two groups of subjects, responders vs. nonresponders, were identified based on the following criteria:

- An average reduction of  $\geq 30\%$  in pain VAS from baseline to Visit 7 (Week 8);
- A reduction of  $\geq 30\%$  in FIQ total score from baseline to Visit 7 (Week 8); and
- A response of "Very much better" or "Much better" on the PGIc.

In the second set of responsiveness analyses, improved *versus* unimproved subjects were defined by the PGIc. The improved group included those subjects who reported "A little better", "Much better", or "Very much better" on their response to the PGIc at Week 8. The unimproved group was defined as those who reported "No change" or a worsening.

In the third set of responsiveness analyses, improved *versus* unimproved subjects were defined by the CGIc. The improved group included those subjects whose condition had been reported by their clinician as "Minimally improved", "Much improved", or "Very much improved" on the CGIc at Week 8. The unimproved group was defined by "No change" or a worsening on the CGIc.



Paired *t*-tests evaluated the within-group difference between two time points. Ideally, the nonresponders (unimproved) group was expected to show no significant change (mean change score near 0 or slight worsening), and the responder/improved group was expected to show significant improvement (negative mean change score). The level of change in the scale was also quantified by standardised effect sizes (SES). Based on Cohen's recommendations, the following values represented the magnitudes of responsiveness: small (0.20), moderate (0.50), and large (0.80) (23). Effect sizes were obtained by taking the difference in mean score between baseline and follow-up (Week 8) divided by the standard deviation (SD) of scores at baseline. Student's *t*-test was used to evaluate the between-group differences.

Additionally the responsiveness analyses were conducted by correlating change scores on the rescored JSS with changes on the FOSQ Total score, SF-36™ Vitality scores, and the ESS score. The rescored JSS was expected to be responsive to changes in the above scales.

The minimally important difference (MID), which is often used as a guideline for interpreting changes assessed by a questionnaire, was also calculated. The MID can be evaluated using either distributional or anchor-based methods, which have been shown to produce generally similar results (24).

The primary anchor-based MID for the rescored JSS was analysed using patients' responses to the PGIC as an anchor. The estimate was the mean change in the rescored JSS associated with the category "a little better". It should be noted that this provided a broad approximation as the PGIC asked about fibromyalgia changes and not about sleep specifically.

The distribution based evaluation of MID used the standard error of measurement (SEM), which is calculated as standard deviation  $\times \sqrt{(1-\text{reliability})}$  (14).

Separate regression models predicting Fatigue VAS change from baseline to Week 8 were conducted on the rescored and original JSS total score. The

**Table I.** Definition of responders used in the known-groups analysis (Trials 1 and 2).

Comparison scales	Responder definition
Pain VAS	$\geq 30\%$ reduction from baseline at Visit 11
FIQ total score and subscale scores	$\geq 30\%$ reduction from baseline at Visit 11
FIQ items 16 and 17	$\geq 30\%$ reduction from baseline at Visit 11
FOSQ score and subscale scores	change score $\geq$ MID (defined by SEM)
SF-36™ Vitality Scale	$\geq 10$ points improvement (MID)(28-36) from baseline at Visit 11
PGIC	a response of "Very much better" or "Much better"
CGIC	a response of "Very much improved" or "Much improved"

R-squared values from the two models were compared. If the difference in R-squared was small, then it was concluded that the new scoring accounted for a substantial amount of the original variance and could be used in place of the original scoring without losing important information.

The analyses listed above were conducted to assess the psychometric properties of the rescored JSS using data from the Phase 2 study. Once the psychometric properties were found to be acceptable, additional triangulation analyses using data from the Phase 3 studies were conducted to further support the construct validity and responsiveness to change.

#### Triangulation analyses

The performance of the rescored JSS was examined in relation to changes in other scales independently in the two Phase 3 studies. The results were contrasted between studies and between the original and rescored JSS.

The correlation coefficients between the rescored JSS and other scales that were expected to have strong correlations were calculated to show that the rescored JSS tracks closely with simpler and more concise measures of the same/similar concept.

The rescored JSS was also evaluated for its ability to discriminate among known responding groups of subjects. The known-groups methodology assesses the extent to which scores were linked to subjects' predefined health states. Evidence of known-groups validity was shown when a meaningful pattern of mean differences was observed across the defined subgroups. The Student's *t*-test was used in the known-group analysis to assess if there were significant differences in

the mean scores between the defined subgroups. In the situation where the normality assumption was violated, the non-parametric test was used in place of the *t*-test.

Groups used for these analyses were defined based on the responder definitions described in Table I. The assumption was that the subjects who were responders would tend to have less severe symptoms compared to nonresponders at endpoint.

A responder for the rescored JSS was defined as a subject with a change score from baseline to endpoint that was greater than or equal to the scale's MID. To test level of agreement between the responder definitions, Cohen's kappa statistics was used by organising the scores into a contingency table. Cohen's kappa has a range from 0.00-1.00, with larger values indicating better reliability or agreement. The following thresholds for agreement from Altman (25) were used for interpretation purposes: poor ( $<0.20$ ), fair (0.20-0.40), moderate (0.40-0.60), good (0.60-0.80), and very good ( $>0.80$ ).

## Results

### Validation of the psychometric properties using data from the Phase 2 study

A total of 195 subjects from the Phase 2 trial were used for the validation analysis of the rescored JSS. Subjects in the validation analysis were predominantly white (92.3%) and female (94.4%). The mean age of the subjects was 46.5 ( $\pm 11.35$ ) years, ranging from 20 to 83 years (Table II).

The percent of missing responses was low (2.05%) for all items and the total score of the rescored JSS. There were no floor effects on items or the total score at baseline. However, 64% to

**Table II.** Clinical trials: demographic and medical information.

	Phase 2 Validation, n=195	Trial 1, n=548	Trial 2, n=573
Age mean (SD) [range]	46.5 (11.35) [20–83]	47.0 (11.26) [18–79]	46.6 (10.72) [19–80]
Gender: Female (%)	184 (94.4)	500 (91.2)	513 (89.5)
Race/ethnicity			
Am. Indian or Alaska Native		6 (1.1)	3 (0.5)
Asian	1 (0.5)	7 (1.3)	4 (0.7)
Black/African American	11 (5.6)	33 (6.0)	39 (6.8)
White/Caucasian	180 (92.3)	498 (90.9)	524 (91.4)
Other	3 (1.5)	4 (0.7)	3 (0.5)
Time since first fibromyalgia symptoms (yrs)			
Mean (SD) [range]	10.6 (8.40) [0–41]	9.7 (8.49) [0–47]	9.7 (8.75) [0–51]
Median	9.0	7.0	8.0
Time since first fibromyalgia diagnosis (yrs)			
Mean (SD) [range]	6.0 (6.02) [0–41]	5.9 (6.75) [0–39]	4.9 (5.61) [0–48]
Median	5.0	4.0	3.0

**Table III.** Summary statistics for rescored and Original Jenkins Scale items and domain at baseline.

<i>Rescored Jenkins Scale</i>	n	Mean	SD	Median	Min	Max	Missing n (%) <sup>*</sup>	Floor Effect n (%) <sup>**</sup>	Ceiling Effect n (%) <sup>**</sup>
Trouble falling asleep	191	1.62	0.53	2.00	0.00	2.00	4 ( 2.05)	4 ( 2.09)	123 (64.40)
Waking up several times per night	191	1.80	0.40	2.00	1.00	2.00	4 ( 2.05)	0 ( 0.00)	152 (79.58)
Trouble staying asleep	191	1.75	0.43	2.00	1.00	2.00	4 ( 2.05)	0 ( 0.00)	143 (74.87)
Waking up feeling tired and worn out	191	1.92	0.27	2.00	1.00	2.00	4 ( 2.05)	0 ( 0.00)	176 (92.15)
Jenkins Scale Total Score	191	7.09	1.20	8.00	3.00	8.00	4 ( 2.05)	0 ( 0.00)	104 (54.45)
<i>Original Jenkins Scale</i>									
Trouble falling asleep	191	3.64	1.42	4.00	0.00	5.00	4 ( 2.05)	4 ( 2.09)	70 (36.65)
Waking up several times per night	191	4.24	1.05	5.00	1.00	5.00	4 ( 2.05)	0 ( 0.00)	106 (55.50)
Trouble staying asleep	191	4.14	1.16	5.00	1.00	5.00	4 ( 2.05)	0 ( 0.00)	105 (54.97)
Waking up feeling tired and worn out	191	4.62	0.73	5.00	1.00	5.00	4 ( 2.05)	0 ( 0.00)	140 (73.30)
Jenkins Scale Total Score	191	16.65	3.41	18.00	5.00	20.00	4 ( 2.05)	0 ( 0.00)	52 (27.23)

<sup>\*</sup>% missing is based on the total number of subjects.

<sup>\*\*</sup>% is based on number of completed responses for each domain.

92% of subjects chose the highest response option (*i.e.*, 22–31 nights [per month]) on the individual items, and 54% had the highest total score indicating very high ceiling effects before and after rescoring (Table III).

The rescored JSS was found to be internally consistent ( $\alpha=0.70$ ), meeting the threshold for acceptability on Cronbach's alpha reliability coefficient ( $\geq 0.70$ ). This implies that the individual JSS items were consistent with each other and reflected a single underlying construct. Similar results, although slightly higher, were found with the original JSS score ( $\alpha=0.78$ ) using data from the same Phase 2 study.

The test-retest reliability of the rescored JSS, using an intra-class correlation coefficient (ICC), met or approached the acceptable threshold ( $\geq 0.70$ ) for stability defined by the FIQ total score

**Table IV.** Concurrent validity of Jenkins Sleep Scale and other PRO scales at Visit 7 (Week 8) in Phase 2 study.

	Rescored Jenkins $r^1$	Original Jenkins $r^1$
With the Pain VAS	0.54	0.56
With the Fatigue VAS	0.57	0.59
With the FOSQ	-0.57	-0.60
With the SF-36 Vitality	-0.66	-0.70
With the ESS	0.43	0.47
With the FIQ Item 16: "How tired have you been?"	0.68	0.70
With the FIQ Item 17: "How have you felt when you get up in the morning?"	0.72	0.76

Note: PRO: Patient-reported outcome.

<sup>1</sup>The correlations were calculated using Pearson's Correlation Coefficient.

(ICC=0.70) and by the ESS total score (ICC=0.69). The ICC results for stable subjects based on Fatigue VAS and Pain VAS were low, 0.66 and 0.61, respectively. For the original JSS, the results were generally similar but slightly higher: the ICCs were 0.75 for the sta-

ble patients identified by the FIQ total score and the ESS, and were 0.70 and 0.66 for stable patients based on Fatigue VAS and Pain VAS, respectively. Very high ICC results were not expected since there was no true measure of stability as related to sleep (initiation,

**Table V.** Responsiveness analysis of Jenkins Sleep Scale total score from baseline to Visit 7 (Week 8) in Phase 2 study.

Assessed by	Responsiveness											
	Responders					Nonresponders					Between Group SES	t-test p-value
<i>Rescored Jenkins</i>	n	Mean Change	SD	p-value	Within Group SES	n	Mean Change	SD	p-value	Within Group SES		
Pain VAS + FIQ Total Score + PGIC <sup>1</sup>	38	-3.84	1.62	<.0001	-3.19	115	-1.60	1.85	<.0001	-1.33	-1.21	<.0001
PGIC <sup>2</sup>	53	-3.57	1.65	<.0001	-2.96	100	-1.41	1.82	<.0001	-1.17	-1.18	<.0001
CGIC <sup>3</sup>	48	-3.46	1.68	<.0001	-2.87	105	-1.56	1.91	<.0001	-1.30	-0.99	<.0001
<i>Original Jenkins</i>												
Pain VAS + FIQ Total Score + PGIC <sup>1</sup>	38	-11.5	4.03	<.0001	-3.38	115	-4.61	5.14	<.0001	-1.35	-1.35	<.0001
PGIC <sup>2</sup>	53	-10.5	4.56	<.0001	-3.09	100	-4.10	4.99	<.0001	-1.20	-1.29	<.0001
CGIC <sup>3</sup>	48	-10.2	4.66	<.0001	-3.00	105	-4.54	5.27	<.0001	-1.33	-1.08	<.0001

Note: Higher Jenkins Sleep Scale scores reflect greater sleep problems.

<sup>1</sup>The responders are defined as those who have reported an average reduction of  $\geq 30\%$  in pain VAS, a reduction of  $\geq 30\%$  in FIQ total score, and a response of "Very much better" or "Much better" on the PGIC.

<sup>2</sup>The responders group includes those subjects who reported a "Very much better/much better/a little better" condition on their response to the PGIC at Visit 7.

<sup>3</sup>The responders group includes those subjects whose condition has been reported as "Very much improved/much improved /minimally improved" on their response to the CGIC at Visit 7.

**Table VI.** Correlations between the Original and Rescored Jenkins Sleep Scale Change from baseline and the Comparison Scales Change from baseline - Trial 1/2.

	Trial 1		Trial 2	
	Rescored Jenkins	Original Jenkins	Rescored Jenkins	Original Jenkins
Pain VAS	0.45	0.46	0.39	0.39
FIQ Physical Impairment	0.32	0.34	0.36	0.35
FIQ Did not Feel Good	0.43	0.48	0.37	0.38
FIQ Work Missed	0.24	0.27	0.23	0.22
FIQ Difficulty with Work	0.43	0.45	0.37	0.39
FIQ Pain	0.48	0.49	0.42	0.44
FIQ item 16 ("How tired have you been?")	0.51	0.55	0.50	0.52
FIQ item 17 ("How have you felt when you get up in the morning?")	0.55	0.60	0.53	0.57
FIQ Stiffness	0.48	0.51	0.47	0.52
FIQ Anxiety	0.20	0.22	0.19	0.21
FIQ Depression	0.21	0.19	0.18	0.20
FIQ Total Score	0.53	0.56	0.50	0.52
FOSQ General Productivity	-0.40	-0.43	-0.41	-0.44
FOSQ Social Outcome	-0.28	-0.31	-0.32	-0.31
FOSQ Activity Level	-0.43	-0.47	-0.46	-0.46
FOSQ Vigilance	-0.35	-0.40	-0.40	-0.40
FOSQ Intimate Relationships	-0.24	-0.29	-0.34	-0.35
FOSQ Total Score	-0.41	-0.45	-0.46	-0.46
SF-36 Vitality Scale	-0.52	-0.58	-0.56	-0.58

maintenance, early arising, feeling re-freshed).

The exploratory analysis of the test-retest reliability of the rescored JSS using FIQ item 17 ("How have you felt when you get up in the morning?") to define stable subjects met the acceptable threshold (ICC=0.72) as did the original Jenkins Sleep Scale (ICC=0.77).

Concurrent validity results indicated an overall pattern of expected results for

all of the prespecified measures. The rescored JSS was moderately to highly positively correlated with FIQ item 16 ("How tired have you been?";  $r=0.68$ ), FIQ item 17 ("How have you felt when you get up in the morning?";  $r=0.72$ ), the Fatigue VAS ( $r=0.57$ ), the Pain VAS ( $r=0.54$ ), and the ESS ( $r=0.43$ ). The rescored JSS was also moderately negatively correlated with the SF-36™. Vitality scale ( $r=-0.66$ ) and with the FOSQ total score ( $r=-0.57$ ) where

lower scores equal worse symptoms or functioning. These correlations were similar, although slightly lower, than those seen with the original JSS, which was also positively correlated with FIQ item 16 ( $r=0.70$ ), FIQ item 17 ( $r=0.76$ ), the Fatigue VAS ( $r=0.59$ ), the Pain VAS ( $r=0.56$ ), and the ESS ( $r=0.47$ ), and negatively correlated with the SF-36™. Vitality scale ( $r=-0.70$ ), and the FOSQ total score ( $r=-0.60$ ) (Table IV). The JSS, using either the rescored or

the original version, was found to be responsive when groups of responders versus nonresponders were defined using the Pain VAS, FIQ and PGIC. The “responder” group significantly improved from baseline on the JSS with large effects (SES=-3.19 for the rescored; SES=-3.38 for the original; Table V). The “nonresponder” groups also significantly improved from baseline and had large effects (SES=-1.33 for the rescored; SES=-1.35 for the original). Although all groups showed large effects, responders demonstrated more than a two-fold improvement over nonresponders, which was statistically significant ( $p<0.0001$ ). The rescored and original versions of the JSS were also found to be responsive when groups of responders *versus* nonresponders were defined using the PGIC or the CGIC. The responder group was statistically significantly different from the nonresponder group ( $p<0.0001$ ). Larger effects for the responder group were found compared to the nonresponder group (Table V). In the regression models predicting Fatigue VAS change from baseline to Week 8, the R-squared for the rescored JSS change was 0.27, while the R-squared for the original JSS change was 0.30. The rescored JSS accounted for 90% of the variance captured by the original JSS, demonstrating the difference between the rescored and original JSS was small.

#### Triangulation analyses using data from the Phase 3 studies

In the two Phase 3 populations ( $n=1,121$ ), most subjects were female (90%) and White/Caucasian (91%). The average age of the subjects was 47 years (range: 18–80). The median time since onset of first fibromyalgia symptoms was 7.0 years (range: 0–51 years), and the median time since first diagnosis of fibromyalgia was 3.0 years (range: 0–48 years) (Table II).

As in the Phase 2 study, the rescored JSS in the Phase 3 trials correlated moderately with related measures (Table 6): FIQ total score ( $r=0.53$  in Trial 1,  $r=0.50$  in Trial 2), FIQ item 16 (“How tired have you been?”) ( $r=0.51$  in Trial 1,  $r=0.50$  in Trial 2), FIQ item

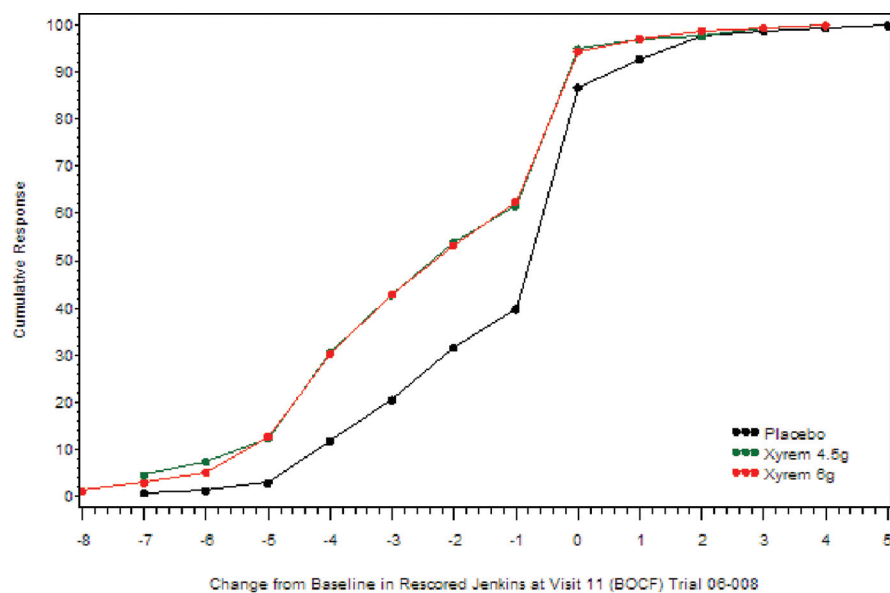


Fig. 1. Change from baseline in Rescored Jenkins Sleep Scale – Trial 1.

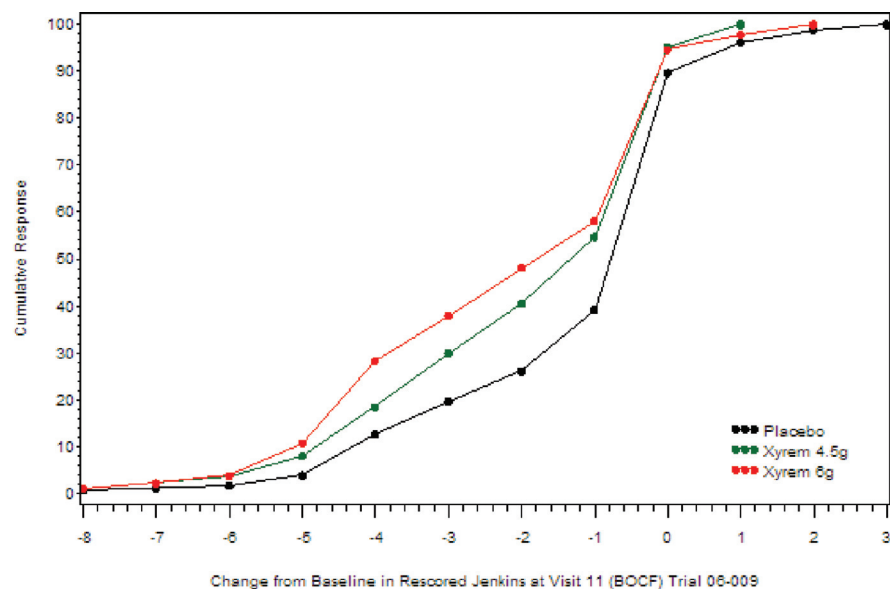


Fig. 2. Change from baseline in Rescored Jenkins Sleep Scale – Trial 2.

17 (“How have you felt when you get up in the morning?”) ( $r=0.55$  in Trial 1,  $r=0.53$  in Trial 2), and SF-36™ Vitality Scale ( $r=-0.52$  in Trial 1,  $r=-0.56$  in Trial 2). The rescored JSS correlated poorly to moderately with other measures: FIQ subscales except FIQ items 16 and 17 ( $r$  range from 0.20 to 0.48 in Trial 1,  $r$  range from 0.18 to 0.47 in Trial 2), FOSQ total score and subscales ( $r$  range from -0.24 to -0.43 in Trial 1,  $r$  range from -0.32 to -0.46 in Trial 2), and Pain VAS ( $r=0.45$  in Trial 1,  $r=0.39$  in Trial 2). The original JSS replicated these data as well.

In the known-groups analysis using responder definitions for other scales to evaluate differences, the rescored JSS was found to have good separation of mean changes between responders (mean change from -2.14 to -3.36) and nonresponders (mean change from -0.69 to -2.0) in both trials. All responder groups had significantly ( $p<0.01$ ) higher improvements compared to nonresponder groups. The mean changes in the responder groups were higher than the MID defined by PGIC (2 points). Similar results can be seen with the original JSS (Figs. 1–4).



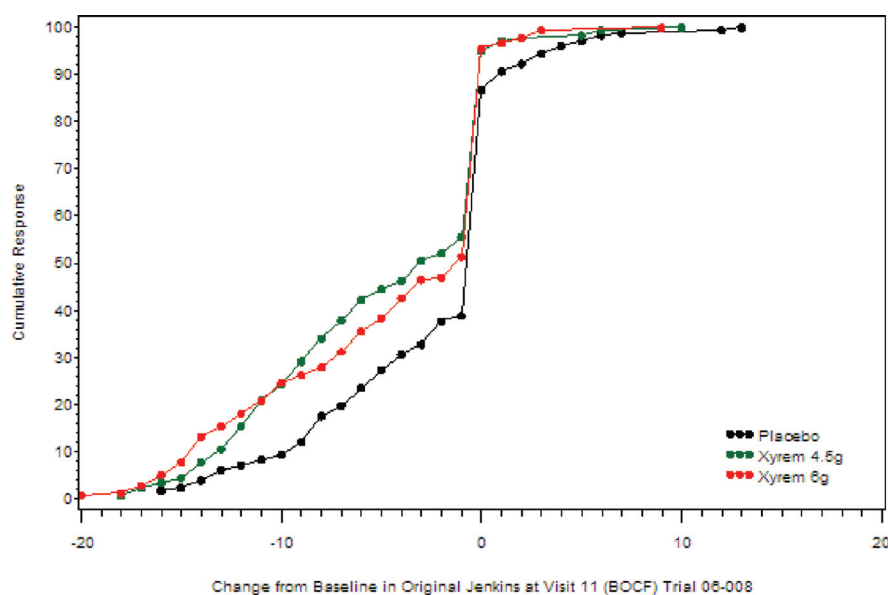


Fig. 3. Change from baseline in Original Jenkins Sleep Scale – Trial 1.

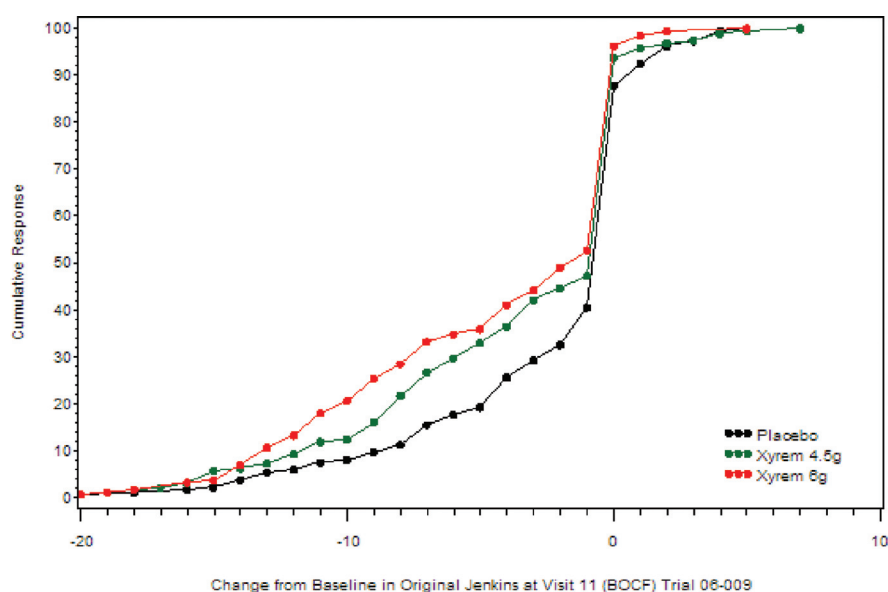


Fig. 4. Change from baseline in Original Jenkins Sleep Scale – Trial 2.

The ability of the rescored JSS to classify responders similarly to other definitions demonstrated fair to moderate agreement with FIQ total score ( $k=0.45$  in Trial 1,  $k=0.32$  in Trial 2), FIQ item 17 (*“How have you felt when you get up in the morning?”*) ( $k=0.38$  in Trial 1,  $k=0.41$  in Trial 2), PGIC ( $k=0.43$  in Trial 1,  $k=0.37$  in Trial 2), and CGIC ( $k=0.41$  in Trial 1,  $k=0.32$  in Trial 2). The agreement between the rescored JSS and other comparison scales (*i.e.* Pain VAS, FOSQ total score and subscale scores, SF-36™ Vitality scale) was poor to fair across both trials.

Similarly, the original JSS demonstrated moderate agreement with FIQ item 17 (*“How have you felt when you get up in the morning?”*) ( $k=0.43$  in Trial 1,  $k=0.44$  in Trial 2); fair to moderate agreement with PGIC ( $k=0.41$  in Trial 1,  $k=0.38$  in Trial 2) and CGIC ( $k=0.40$  in Trial 1,  $k=0.37$  in Trial 2); and fair agreement with the Pain VAS, FIQ item 16 (*“How tired have you been?”*), FOSQ total score and subscale scores, and SF-36™ Vitality scale.

## Discussion

This study sought to maintain the struc-

ture of the JSS by rescoring it in a manner unlikely to be fraught by recall issues and was not a study prospectively designed to study recall bias. By reducing the categorisations to less than half the time and more than half the time, the likelihood of misclassification was expected to be reduced. While this expectation is considered reasonable, a limitation of the study was that there was no daily sleep diary available to examine the true recall bias for these two scales.

It was important to assess the psychometric properties of the rescored JSS to ensure that it maintained the same general properties as the original. The construct validity of the rescored JSS was confirmed through an acceptable internal consistency reliability (Cronbach's  $\alpha=0.70$ ). Moderate correlations with scales of related content, including the fibromyalgia-specific FIQ item 16 (*“How tired have you been?”*), FIQ item 17 (*“How have you felt when you get up in the morning?”*) measuring non-restorative sleep, and the sleep-specific FOSQ supported the concurrent validity of the rescored scale.

High ceiling effects on the individual items and total score of the JSS were found, indicating that at baseline, the Phase 2 study population experienced sleep problems on most nights over the previous 4 weeks. In theory, a high ceiling effect may limit the ability of a score to show any deterioration in subjects' sleep that may occur during a trial. However, since subjects entered the Phase 2 study with chronic, moderate/severe fibromyalgia and since there were few effective treatments for fibromyalgia at the time of the study, it was not expected that further deterioration of their sleep would have occurred. Furthermore, the results found on the JSS typify the self reports of fibromyalgia patients, who generally report severe problems with sleep, fatigue, pain, and low levels of physical function (26, 27). To document the test-retest reliability, responsiveness to change, and known-group validity of the rescored JSS, groups of subjects whose sleep quality improved, worsened, or remained unchanged at endpoint should be defined. This is usually achieved through the



use of a global assessment of change in the concept being investigated (in this case, sleep quality). Since such a measure was not included in the Phase 2 study, the responder definition was based on change in scales capturing related concepts or fibromyalgia-related symptoms that were expected to have commonalities or similar evolution over time as sleep quality, such as fatigue and HRQL. The PGIC and CGIC as global assessments of change in the condition were also considered the best proxy measures available to capture change in sleep quality. Three sets of responder groups based on scores from (1) the composite of the FIQ, the pain VAS, and the PGIC, (2) the PGIC, and (3) the CGIC were defined for the responsiveness and known-group validity analyses. Significant group differences were found between the responder and nonresponder groups across all definitions of responsiveness. Yet, improvements were also found in the nonresponder group. This might be due to the criteria used for defining the groups: the use of non-sleep concepts (e.g. improvement of fatigue, improvement of condition) in the nonresponder definitions may have led to the classification of subjects with decreased fatigue or HRQL, yet with increased sleep quality, as nonresponders based on other criteria.

Reliability of the rescored JSS total score showed correlations that met or approached an acceptable level (ICC ~0.70) for subjects whose quality of sleep could be considered unchanged in a 2-week period as identified through the ESS and FIQ total scores. These results might be explained by the fact that groups were defined based on their changes (improvement or worsening) or no change in score on the ESS or on the FIQ, which evaluate daytime sleepiness and impact of fibromyalgia on HRQL, respectively. The ICC threshold was met (ICC=0.72) when using FIQ item 17 ("How have you felt when you get up in the morning?") in subjects whose sleep patterns could be considered stable. Results might have differed if a global assessment of change specific for sleep quality had been used instead.

In summary, the construct validity of the original and rescored JSS in addition to the scale's reliability and responsiveness were documented in patients with fibromyalgia. In addition, the JSS demonstrated moderate correlation with scales measuring conceptually related symptoms, including tiredness and fatigue, or symptoms evolving in a similar pattern, such as pain. These results were expected as the JSS assesses the frequency of the sleep problems, while the other scales measure the intensity/severity of the concepts. It is important to note that the reduction in the number of response options for the JSS as a result of the revised scoring algorithm reduces the scale's ability to detect differences. Nonetheless, when JSS data from the two Phase 3 studies were analysed, the rescored and original JSS both arrived at the same conclusion demonstrating the efficacy of the active treatment groups ( $p < 0.01$  for both studies).

These analyses support the use of the JSS in research to document treatment benefit with regard to sleep problems in patients with fibromyalgia. In addition, a new, intuitive scoring algorithm of the JSS was developed to assess the impact of potential recall bias. Based on the consistency of evidence supporting its validity, including its ability to demonstrate statistically significant improvements across two Phase 3 trials despite a reduction in the ability to detect change, the rescored JSS demonstrated adequate responsiveness to detect treatment benefit. The rescored JSS performed similarly to the original JSS, suggesting the original scale's 4-week recall period did not introduce substantial bias in capturing the experience of fibromyalgia-related sleep disturbances.

### Acknowledgements

The authors would like to thank Jonathan Stokes, Yueping Zhu, Andrew Yaworsky, Kathleen Rosa, Diane Guinta, and Kathy Ryan for their assistance in the study and manuscript preparation, and Mark Eller and Annette Madrid for their helpful review.

### Contributions of Authors

- Design and conduct of research

study (qualitative): BC, EP, CL

- Analysis and interpretation of the data (qualitative and quantitative): BC, EP, CL
- Preparation, review, and approval of the manuscript: BC, PSP, EP, CL

Bruce Crawford had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

### References

1. GOLDENBERG DL, BURCKHARDT C, CROFORD L: Management of fibromyalgia syndrome. *JAMA* 2004; 292: 2388-95.
2. WOLFE F, SMYTHE HA, YUNUS MB *et al.*: The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. *Arthritis Rheum* 1990; 33:160-72.
3. BRANCO JC, BANNWARTH B, FAILDE I *et al.*: Prevalence of fibromyalgia: a survey in five European countries. *Semin Arthritis Rheum* 2010; 39: 448-53.
4. WOLFE F, ROSS K, ANDERSON J, RUSSELL IJ, HEBERT L: The prevalence and characteristics of fibromyalgia in the general population. *Arthritis Rheum* 1995; 38: 19-28.
5. RUTLEDGE DN, MOUTTAPA M, WOOD PB: Symptom clusters in fibromyalgia: potential utility in patient assessment and treatment evaluation. *Nurs Res* 2009; 58: 359-67.
6. JENKINS CD, STANTON BA, NIEMCRYK SJ, ROSE RM: A scale for the estimation of sleep problems in clinical research. *J Clin Epidemiol* 1988; 41: 313-21.
7. DAVIES KA, MACFARLANE GJ, NICHOLL BI *et al.*: Restorative sleep predicts the resolution of chronic widespread pain: results from the EPFUND study. *Rheumatology (Oxford)* 2008; 47: 1809-13.
8. MCHORNEY CA, WARE JE, JR., RACZEK AE: The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993; 31: 247-63.
9. WARE JE, JR., SHERBOURNE CD: The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30: 473-83.
10. PARROTT AC, HINDMARCH I: Factor analysis of a sleep evaluation questionnaire. *Psychol Med* 1978; 8: 325-9.
11. CAPPELLERI JC, BUSHMAKIN AG, MCDERMOTT AM, SADOSKY AB, PETRIE CD, MARTIN S: Psychometric properties of a single-item scale to assess sleep quality among individuals with fibromyalgia. *Health Qual Life Outcomes* 2009; 7: 54.
12. RUSSELL IJ, PERKINS AT, MICHALEK JE: Sodium oxybate relieves pain and improves function in fibromyalgia syndrome: a randomized, double-blind, placebo-controlled, multicenter clinical trial. *Arthritis Rheum* 2009; 60: 299-309.
13. GORUSCH R: Factor Analysis. 2nd. ed.

- Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.
14. NUNNALLY J: Psychometric Theory. 2nd. ed. New York: McGraw Hill, 1978.
  15. BURCKHARDT CS, CLARK SR, BENNETT RM: The fibromyalgia impact questionnaire: development and validation. *J Rheumatol* 1991; 18: 728-33.
  16. WEAVER TE, LAIZNER AM, EVANS LK *et al.*: An instrument to measure functional status outcomes for disorders of excessive sleepiness. *Sleep* 1997; 20: 835-43.
  17. BENNETT R: The Fibromyalgia Impact Questionnaire (FIQ): a review of its development, current version, operating characteristics and uses. *Clin Exp Rheumatol* 2005; 23(5 Suppl. 39): S154-S162.
  18. CRONBACH LJ: Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 1951; 16: 297-334.
  19. BENNETT RM, BUSHMAKIN AG, CAPPELLERI JC, ZLATEVA G, SADOSKY AB: Minimal clinically important difference in the fibromyalgia impact questionnaire. *J Rheumatol* 2009; 36: 1304-11.
  20. HUGHES K, GLASS C, RIPCINSKI M *et al.*: Efficacy of the topical nasal steroid budesonide on improving sleep and daytime somnolence in patients with perennial allergic rhinitis. *Allergy* 2003; 58: 380-5.
  21. GUYATT GH, DEYO RA, CHARLSON M, LEVINE MN, MITCHELL A: Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol* 1989; 42: 403-8.
  22. HAYS RD, HADORN D: Responsiveness to change: an aspect of validity, not a separate dimension. *Qual Life Res* 1992; 1: 73-5.
  23. KAZIS LE, ANDERSON JJ, MEENAN RF: Effect sizes for interpreting changes in health status. *Med Care* 1989; 27 (Suppl.): S178-S189.
  24. NORMAN GR, SLOAN JA, WYRWICH KW: Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003; 41: 582-92.
  25. ALTMAN DG: Practical Statistics for Medical Research. London: Chapman & Hall CRC, 1991.
  26. MEASE P, ARNOLD LM, CHOY EH *et al.*: Fibromyalgia syndrome module at OMER-ACT 9: domain construct. *J Rheumatol* 2009; 36: 2318-29.
  27. SALAFFI F, SARZI-PUTTINI P, CIAPETTI A, ATZENI F: Assessment instruments for patients with fibromyalgia: properties, applications and interpretation. *Clin Exp Rheumatol* 2009; 27 (Suppl. 56): S92-105.
  28. ANGST F, AESCHLIMANN A, STUCKI G: Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum* 2001; 45: 384-91.
  29. GUYATT GH, OSOBA D, WU AW, WYRWICH KW, NORMAN GR: Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002; 77: 371-83.
  30. KOSINSKI M, ZHAO SZ, DEDHIYA S, OSTERHAUS JT, WARE JE, JR: Determining 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.
  31. KUJAWSKI SC, KOSINSKI M, MARTIN R *et al.*: Determining meaningful improvement in SF-36 scale scores for treatment studies of early, active RA. *Arthritis Rheum* 2000; 43.
  32. THUMBOO J, FONG KY, NG TP *et al.*: Validation of the MOS SF-36 for quality of life assessment of patients with systemic lupus erythematosus in Singapore. *J Rheumatol* 1999; 26: 97-102.
  33. WELLS GA, TUGWELL P, KRAAG GR, BAKER PR, GROH J, REDELMIEIER DA: Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. *J Rheumatol* 1993; 20: 557-60.
  34. WYRWICH KW, NIENABER NA, TIERNEY WM, WOLINSKY FD: Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care* 1999; 37: 469-78.
  35. WYRWICH KW, TIERNEY WM, WOLINSKY FD: Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol* 1999; 52: 861-73.
  36. ZHAO SZ, McMILLEN JJ, MARKENSON JA *et al.*: Evaluation of the functional status aspects of health-related quality of life of patients with osteoarthritis treated with celecoxib. *Pharmacotherapy* 1999; 19: 1269-78.