

Content and Criterion Validity of the Preliminary Core Dataset for Clinical Trials in Fibromyalgia Syndrome

ERNEST H. CHOY, LESLEY M. ARNOLD, DANIEL J. CLAUW, LESLIE J. CROFFORD, JENNIFER M. GLASS, LEE S. SIMON, SUSAN A. MARTIN, C. VIBEKE STRAND, DAVID A. WILLIAMS, and PHILIP J. MEASE

ABSTRACT. *Objective.* Increasing research interest and emerging new therapies for treatment of fibromyalgia (FM) have led to a need to develop a consensus on a core set of outcome measures that should be assessed and reported in all clinical trials, to facilitate interpretation of the data and understanding of the disease. This aligns with the key objective of the Outcome Measures in Rheumatology (OMERACT) initiative to improve outcome measurement through a data driven, interactive consensus process.

Methods. Through patient focus groups and Delphi processes, working groups at previous OMERACT meetings identified potential domains to be included in the core data set. A systematic review has shown that instruments measuring these domains are available and are at least moderately sensitive to change. Most instruments have been validated in multiple languages. This pooled analysis study aims to develop the core data set by analyzing data from 10 randomized controlled trials (RCT) in FM.

Results. Results from this study provide support for the inclusion of the following in the core data set: pain, tenderness, fatigue, sleep, patient global assessment, and multidimensional function/health related quality of life. Construct validity was demonstrated with outcome instruments showing convergent and divergent validity. Content and criterion validity were confirmed by multivariate analysis showing R square values between 0.4 and 0.6. Low R square value is associated with studies in which one or more domains were not assessed.

Conclusion. The core data set was supported by high consensus among attendees at OMERACT 9. Establishing an international standard for RCT in FM should facilitate future metaanalyses and indirect comparisons. (J Rheumatol 2009;36:2330–4; doi:10.3899/jrheum.090368)

Key Indexing Terms:

FIBROMYALGIA
CLINICAL TRIALS

OMERACT

OUTCOME MEASURES
CORE DATA SET

Fibromyalgia (FM) is a common condition afflicting 2% of the population¹. It is characterized by chronic widespread pain with increased sensitivity to pressure elicited pain. The American College of Rheumatology (ACR) classification criteria in 1990 stipulated the presence of chronic widespread pain for at least 3 months and the presence of at least

11 out of 18 tender points². Direct and indirect medical costs associated with FM are high³ although using diagnosis positively can reduce healthcare utilization⁴. Aside from pain, FM is associated with many symptoms including fatigue, depression, anxiety, and poor sleep quality. Many clinical trials have been conducted in FM; however vari-

From the Sir Alfred Baring Garrod Clinical Trials Unit, Academic Department Rheumatology, King's College London, London, UK; Department of Psychiatry, University of Cincinnati College of Medicine, Cincinnati, OH; University of Michigan, Ann Arbor, MI, Division of Rheumatology and Women's Health, University of Kentucky, Lexington, KY; Research Center for Group Dynamics, Department of Psychiatry, Division of Substance Abuse, University of Michigan; Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA; RTI-Health Solutions, Ann Arbor, MI; Division of Immunology/Rheumatology, Stanford University, Portola Valley, CA; Internal Medicine, Psychiatry and Psychology, University of Michigan, Ann Arbor, MI; Seattle Rheumatology Associates, Chief, Division of Rheumatology Research, Swedish Medical Center and University of Washington, Seattle, WA, USA. The Academic Department of Rheumatology is supported by an Integrated Clinical and Academic Centre grant from the Arthritis Research Campaign, UK. Dr. Williams' participation was supported in part by Grant Number U01AR55069 from NIAMS/NIH. Dr. Arnold's participation was supported in part by Grant Number R01AR053207 from NIAMS/NIH. E.H. Choy, MD, Director, Sir Alfred Baring Garrod Clinical Trials Unit, Academic Department Rheumatology, King's College London;

L.M. Arnold, MD, Department of Psychiatry, University of Cincinnati College of Medicine; D.J. Clauw, MD, Professor of Medicine and Psychiatry, University of Michigan; L.J. Crofford, Gloria W. Singletary Professor of Internal Medicine, Chief, Division of Rheumatology and Women's Health, University of Kentucky; J.M. Glass, Research Assistant Professor, Research Center for Group Dynamics, Department of Psychiatry, Division of Substance Abuse, University of Michigan; L.S. Simon, MD, Associate Clinical Professor of Medicine, Harvard Medical School, Beth Israel Deaconess Medical Center; S.A. Martin, RTI-Health Solutions; C.V. Strand, MD, Clinical Professor of Medicine, Adj., Division of Immunology/Rheumatology, Stanford University; D.A. Williams, PhD, Professor Anesthesiology, Internal Medicine, Psychiatry and Psychology, University of Michigan; P.J. Mease, MD, Seattle Rheumatology Associates, Chief, Division of Rheumatology Research, Swedish Medical Center, Clinical Professor of Rheumatology, University of Washington. Address correspondence to Dr. E. Choy, Sir Alfred Baring Garrod Clinical Trials Unit, Academic Department Rheumatology, Weston Education Centre, King's College London, Cutcombe Road, London, SE5 9RJ UK. E-mail: ernest.choy@kcl.ac.uk

ances in outcome measurement methodology have made statistical comparison and pooling of results difficult.

The Outcome Measures in Rheumatology (OMERACT) initiative⁵ has helped to resolve the problem of outcomes measurement variability in rheumatic diseases such as rheumatoid and psoriatic arthritis, by establishing core data sets that should be collected and reported in randomized controlled trials (RCT). OMERACT offers guidance in selecting core data set domains. Applying the OMERACT filters (i.e., truth, discrimination and feasibility), an iterative process can unfold that continually refines the field's ability to access relevant aspects of disease/syndrome measurement (domains) with precision.

Previous works based on patient focus group and Delphi exercises have established a list of potential core data set domains for trials in FM. Remarkable consensus regarding the relevant domains for FM is supported by a Delphi exercise among clinician/researchers, patient focus groups⁶, a Delphi exercise conducted in patients with FM⁷, and through voting at OMERACT 7 and OMERACT 8⁸. Each of these studies provided empirical support for the selection of outcome domains that should be considered for inclusion in the core data set⁹. From these works, the relevant domains for FM appear to be (1) pain, (2) patient global, (3) fatigue, (4) health-related quality of life, (5) multidimensional function, (6) sleep, (7) depression, (8) physical function, (9) tenderness, (10) dyscognition (cognitive dysfunction), and (11) anxiety. The Delphi processes and patient focus groups helped to support the face validity (e.g., truth) of these potential domains. The feasibility and discriminatory power of specific instruments used to assess these domains were the topic of a separate systematic review of RCT in FM¹⁰. This latter review found that available instruments assessing these domains were at least moderately responsive to change with effect size of at least 0.4 and were feasible for use in trials of FM (with the exception of dyscognition). Most outcome measures in RCT of FM, however, have been adopted from other diseases (e.g., Beck Depression Questionnaire¹¹ used in evaluation of depression). Support for the valid use of these "adopted" questionnaires in some cases requires additional support. The objective of the current study was to examine some of the psychometric properties of existing outcomes measures being used in trials of FM. This information will help to evaluate the valid use of these "adopted" measures in the context of FM and will further help to establish a core set of domains for investigation in FM RCT.

METHODS

Data and analysis. The co-chairmen (PM and EC) on behalf of the steering committee approached 4 pharmaceutical companies, Forest Laboratories, Jazz Pharmaceuticals, Eli Lilly, and Pfizer, for de-identified access to each of their large RCT in FM for the purpose of evaluating the measurement characteristics of the instruments used by each for domain assessment. Data from 10 RCT of 4 compounds being investigated for the treatment of FM

were included: milnacipran, duloxetine, pregabalin, and sodium oxybate. Milnacipran and duloxetine are serotonin and norepinephrine reuptake inhibitors while pregabalin is an α 2- δ calcium channel agonist. Duloxetine, milnacipran, and pregabalin are licensed in the USA for management of FM. Duloxetine is also approved for treatment of depression and the pain of diabetic peripheral neuropathy. Pregabalin is also approved for the treatment of the pain of peripheral neuropathy and as an adjunct for the treatment of seizure disorder. Sodium oxybate is the sodium salt of gamma hydroxybutyrate. It is a central nervous system depressant and a sleep modifier. It is licensed for treatment of cataplexy and excessive daytime sleepiness in narcolepsy. Given that FM is a polysymptomatic condition, we included trials of different medications reasoning that medications acting on different pathways may have dissimilar impact on individual domains.

Data from RCT of the same medication have been pooled for analysis. For commercial sensitivity, medications are coded as A, B, C, and D. Change values were calculated for each outcome measure at baseline and after treatment at the primary endpoint of each trial.

Mapping outcomes measures to domains. All outcomes measures used in the clinical trials were mapped onto one or more of the following domains: pain, patient global, fatigue, health related quality of life (HRQOL), multi-dimensional function, sleep, depression, physical function, tenderness, dyscognition, and anxiety. For outcome measures that have multiple domains such as Medical Outcomes Survey Short Form 36 (SF-36)¹², the individual domains and summary component scores were mapped and included in the analyses separately.

Support for construct validity. Construct validity refers to the cumulative evidence supporting whether a given scale or instrument actually assesses the topic it purports to measure. Given almost all the instruments used in RCT of FM were developed and validated in other medical conditions, it can not be assumed that these "adopted" instruments actually measure FM signs and symptoms with the same measurement characteristics as those for the conditions for which they were originally designed. For example, a scale claiming to measure fatigue developed and validated in the context of sports medicine may not be measuring the same type of fatigue affecting individuals with FM. Thus despite the common name "fatigue," evidence would be needed to support a claim that the same fatigue construct was being assessed by this instrument in both populations.

Support for construct validity for measurement in FM, i.e., whether the instrument is really measuring what it is supposed to measure, has not been established, with the exception of SF-36 (personal communication). An example is the Medical Outcome Study Sleep Questionnaire¹³, which is a validated questionnaire developed to assess sleep in patients with sleep disorders. It has been used in a number of RCT in FM, but its validity and performance in FM have not been examined, creating a situation that requires that appropriateness of continued use of this instrument for the sleep domain in FM studies be evaluated.

The construct validity of the instruments is assessed by examining the convergent and divergent relationships of similar and dissimilar instruments. Instruments measuring similar constructs would be expected to have the strongest relationships (either positive or negative depending on the direction of the scale) and unrelated constructs would be expected to demonstrate weaker relationships. For this study, correlation matrices containing all the outcome measures used with a given compound were constructed. Thus 4 matrices were constructed in total. Either Pearson or Spearman correlation coefficients were used depending on the statistical distribution of the instrument. The mean correlation coefficient of outcome measures mapping to the same domain (intradomain correlation coefficient), was used as an indicator of convergent validity. The mean correlation coefficient of outcome measures of different domains (interdomain correlation coefficient), was used as an indicator of divergent validity.

Support for content validity of domains for the core dataset. Content validity refers to the extent to which a single or group of measures is able to capture the relevant facets of a given condition. For this study the content cov-

erage of the consensually derived domains was examined by multivariate analysis. Patient global impression of change (PGIC) was used as a surrogate of overall improvement and the dependent variable in multivariate regression analyses. The overall R square values from multivariate regression analyses were used to identify the adequacy of the domains and associated measurement instruments to evaluate overall improvement in these RCT of FM. For each regression equation, the instrument with the highest univariate correlation with PGIC from each of the domains was included as independent variable.

RESULTS

The domains and the outcome measures used to index the domains from the 10 RCT are listed in Table 1. Instruments such as the SF-36 or Fibromyalgia Impact Questionnaire (FIQ)¹⁴, which were mapped to HRQOL and multidimensional function domains, were almost identical. In trials of one medication, EuroQol was also used. Given the large overlap, HRQOL and multidimensional function were merged into one domain: multidimensional function. Not all the domains were measured in all RCT. The number of domains and instruments used in the trials of the 4 different medications are given in Table 2. While some domains were assessed in all trials (e.g., pain, fatigue), other domains were less consistently assessed (e.g., stiffness, tenderness), and some domains appeared in the evaluation of only one compound (e.g., dyscognition).

Construct convergence and divergence. Mean intradomain correlation coefficients were greater than mean interdomain correlation coefficients for pain, tenderness, fatigue and depression; therefore, instruments assessing these domains demonstrated convergent and divergent validity (Table 3). For multidimensional function and sleep, the difference between mean intra- and interdomain correlation coefficients was small. For multidimensional function, this was expected given the breadth of this construct. For sleep, lack of separation could be due to the failure of treatments to

Table 2. Number of domains and instruments used in clinical trials of 4 different medications.

Medication	A	B	C	D
Pain	> 1	> 1	> 1	> 1
Patient global	1	1	1	1
Fatigue	> 1	> 1	> 1	> 1
Multidimensional function	> 1	> 1	> 1	1
Sleep	1	> 1	1	> 1
Depression	> 1	> 1	> 1	> 1
Physical function	> 1	> 1	> 1	> 1
Tenderness	1	> 1	1	0
Dyscognition	1	0	0	0
Anxiety	1	> 1	> 1	> 1
Stiffness	1	1	1	0

improve sleep; or each of the instruments may have failed to assess facets of sleep that are of importance to individuals with FM. For example, the MOS sleep scale assesses snoring (correlation coefficient $r = 0.02$) and waking up with shortness of breath (correlation coefficient $r = 0.18$), which may be relevant for some sleep disorders but less relevant in FM. Thus sleep (despite clinical anecdotes and consensus as being of relevance to FM) did not correlate highly with PGIC and was rather insensitive to change. In some studies, a patient global rating of sleep quality based on a Likert scale was also used. It also showed a moderate correlation with PGIC (correlation coefficient: $r = 0.4$) as did the MOS sleep disturbance scale (PGIC: $R = 0.4$). These data suggest that subscales may be preferred to the overall indices on some instruments “adopted” from other medical conditions.

Originally, measures of tenderness were mapped onto the pain domain. The instruments used included tender point count (TPC) and dolorimetry. However, the correlation coefficient between tenderness and self-reported pain scale was at best moderate (≤ 0.4) while correlation between TPC

Table 1. List of outcome measures used in clinical trials.

Pain	VAS pain by paper, VAS pain by electronic diary, BPI total score, SF-36 bodily pain, FIQ pain item, BPI worst pain item, BPI least pain item, BPI average pain item, BPI pain right now item
Fatigue	Multidimensional fatigue inventory, Multidimensional assessment of fatigue, SF-36 vitality, FIQ tiredness
Depression	Beck Depression Inventory, Hamilton Rating Scale for Depression total score, HAMD Maier subscale, SF-36 mental component score, SF-36 mental health, HADS depression, FIQ depression
Sleep	MOS Sleep Scale indices I and II and all subscales, Jenkin’s sleep scale, Global sleep quality, FIQ restedness, HAMD sleep subscale, BPI interference with sleep item
Physical function	FIQ physical function, SF-36 physical function, SF-36 role physical, Sheehan disability total score, BPI interference with walking item, BPI interference with normal work item
Quality of life and multidimensional function	SF-36 mental and physical component scores, FIQ total score, EuroQol 5D
Patient global impression of change	Likert scale
Tenderness	Tender point count, Pressure pain threshold measured by dolorimeter
Dyscognition	Multiple abilities self-report questionnaire
Anxiety	FIQ anxiety, HADS anxiety, HAMD anxiety/somatization subscale
Stiffness	FIQ stiffness

VAS: visual analog scale; BPI: Brief Pain Inventory, SF-36: Short Form-36, FIQ: Fibromyalgia Impact Questionnaire, HAMD: Hamilton Rating Scale for Depression, HADS: Hospital Anxiety and Depression Scale.

Table 3. Convergent and divergent relationships.

	A		B		C		D	
	Intra-domain	Inter-domain	Intra-domain	Inter-domain	Intra-domain	Inter-domain	Intra-domain	Inter-domain
Pain	0.82	0.32	0.53	0.26	0.66	0.53	0.62	0.37
Tenderness	Single	0.32	0.59	0.21	Single	0.58	ND	ND
Fatigue	0.64	0.5	0.37	0.27	0.62	0.52	0.66	0.43
Sleep	Single	0.49	0.25	0.23	Single	0.52	0.3	0.26
Depression	0.63	0.42	0.37	0.22	0.64	0.36	0.58	0.35
Physical function	0.65	0.45	0.31	0.27	0.66	0.44	0.69	0.35
Multidimensional function	0.57	0.61	0.3	0.27	0.55	0.48	Single	0.5
Dyscognition	Single	0.23	ND	ND	ND	ND	ND	ND
Anxiety	Single	0.4	0.32	0.24	0.6	0.4	0.58	0.38
Stiffness	Single	0.57	Single	0.19	Single	0.57	ND	ND

Single: Single instrument; ND: not done.

and dolorimetry correlation was high: $R = 0.59$). This suggested tenderness and spontaneous self-report of pain may not be measuring the same construct in FM and should be treated separately.

In summary, instruments used in these RCT to measure patient self-reported pain, fatigue, depression, physical function, and multidimensional function supported the construct validity of these instruments for use in clinical trials of FM. For sleep, the subscale but not overall index was supported. For tenderness, support can only be demonstrated in the trials of one medication in which tenderness was assessed by more than one method. For stiffness, dyscognition and anxiety, convergent and divergent validities could not be determined as these domains were measured by only one instrument in these trials.

Content validity of the domains for the core dataset. Univariate analysis showed that correlations between instruments of different domains with PGIC were moderate to high (Table 4). For depression, the mean correlation coefficient with PGIC was less than 0.5. However, in all of these clinical trials, patients with severe co-morbid depression were excluded. In addition, patients with moderate depression were also excluded in trials of 3 of these compounds.

Table 4. Mean correlation coefficient between instruments of each domain with patient global impression of change.

Medication	A	B	C	D
Pain	0.64	0.42	0.64	0.54
Fatigue	0.65	0.4	0.68	0.5
Multidimensional function	0.8	0.57	0.8	0.59
Sleep	0.64	0.46	0.64	0.4
Depression	0.39	0.26	0.63	0.34
Physical function	0.56	0.37	0.56	0.37
Tenderness	0.4	0.34	0.4	ND
Dyscognition	0.3	ND	ND	ND
Anxiety	0.48	0.3	> 1	0.3
Stiffness	0.78	0.46	0.74	ND

ND: not done.

Consequently, baseline depression scores were low, reducing the effect size of these change scores.

Multivariate analyses showed moderate to high (0.4–0.67) values of R square, which was related to the number of domains assessed. In studies in which some of the potential domains were not assessed, such as tenderness, the R square value was also lower, suggesting that missing key domains will affect the overall coverage of content relevant to the condition of FM.

Regression analyses retained pain, fatigue, physical function, multidimensional function, and depression in all RCT of all 4 compounds. Tenderness was retained in all the trials of the 3 compounds in which it was assessed, further supporting inclusion of tenderness as a separate domain in the core data set. Sleep was retained in 2 of 3 possible clinical trial groups. Stiffness was retained in 2 of 4 groups and dyscognition was not retained in these regression analyses.

DISCUSSION

Data from our study and previous consensus exercises support inclusion of pain, fatigue, physical function, and multidimensional function as domains in a core data set for clinical trials in FM. Although “adopted” from other medical conditions, the instruments measuring these domains largely demonstrate characteristics supporting face, construct, content, and criterion validity in FM. A previous study has also shown that these instruments are at least moderately sensitive to changes¹⁰.

Depression is a common symptom in FM and rated as important by both patients and clinicians. This analysis showed that the correlation between depression and PGIC is only moderate. The main reason is likely that the exclusion of patients with moderate and severe depression in most clinical trials results in a low baseline depression score. Therefore, it is unlikely that any instrument would demonstrate a large effect size. Given this exclusion criterion is common in FM RCT, it seems unnecessary to stipulate its inclusion in the core data set. Nonetheless, the assessment of depression in FM is likely to be helpful in many clinical trials.

“Unrefreshed” sleep is common and thought to be pathogenically important in FM. Moldofsky, *et al* showed that symptoms similar to FM could be induced by disturbing the quality of sleep in healthy normal volunteers¹⁵. Both patients and clinicians agree regarding its importance. However, to date clinical trials have used instruments not developed in patients with FM and which may not assess the type of sleep problems specific to those patients. The practice of using total indices as the sole outcome endpoints for sleep may not be ideal. Our data suggested that using the sleep disturbance subscale of the MOS sleep scale would have improved the convergent and divergent characteristics of this measure. Since sleep was retained in regression analyses in all but one group, there is a strong argument for including some element of sleep in the core data set.

Results of the current study suggest tenderness should be included as a separate domain from patient self-reported pain. Pathophysiologically, this would be logical in that it may involve different pathways. Further, it mirrors the need to assess both patient reported pain and tender joint count in rheumatoid and psoriatic arthritis. Although tender point count and dolorimetry have deficiencies such as significant interobserver variation and may not be the perfect tool, they are feasible, and current analyses showed that they contribute significantly to the content validity when added to the core data set. Hence the conclusion to include tenderness in the core data set.

For anxiety, stiffness, and dyscognition, currently, there are insufficient data from available clinical trials to support their inclusion into the core data set. Researchers interested in these outcomes should include them in assessment, but it is not justifiable to stipulate their assessment in all clinical trials of FM.

Results of our study were presented at the OMERACT 9 FM module and were the basis, along with review of the previous clinician/researcher and patient Delphi exercises, outcome measures, and disease state discussion, for development of consensus on a core domain construct for fibromyalgia¹⁶.

ACKNOWLEDGMENT

We thank the support from Nooshine Dayani, Qu Peng, and Robert Palmer from Forest Laboratories Inc; Chinglin Lai, Yanping Zheng, and Diane Guinta from Jazz Pharmaceuticals Inc.; Daniel Kajdasz and Amy Chappell from Eli Lilly & Co.; and Gergana Zlateva and Emir Birol from Pfizer Inc.

REFERENCES

1. 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.
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. Boonen A, van den Heuvel R, van Tubergen A, et al. Large differences in cost of illness and wellbeing between patients with fibromyalgia, chronic low back pain, or ankylosing spondylitis. *Ann Rheum Dis* 2005;64:396-402.
4. Hughes G, Martinez C, Myon E. The impact of a diagnosis of fibromyalgia on health care resource use by primary care patients in the UK: An observational study based on clinical practice. *Arthritis Rheum* 2006;54:177-83.
5. Tugwell P, Boers M, Brooks P, Simons L, Strand V, Idzerda L. OMERACT: An international initiative to improve outcome measurement in rheumatology. *Trials* 2007;8:38-43.
6. Arnold LM, Crofford LJ, Mease PJ, et al. Patient perspectives on the impact of fibromyalgia. *Patient Edu Couns* 2008;73:114-20.
7. Mease PJ, Arnold LM, Crofford LJ, et al. Identifying the clinical domains of fibromyalgia: Contributions from clinical and patients delphi exercises. *Arthritis Care Res* 2008;59:952-60.
8. Mease P, Arnold LM, Bennett R, et al. Fibromyalgia syndrome. *J Rheumatol* 2007;34:1415-25.
9. Mease PJ, Clauw DJ, Arnold LM, et al. Fibromyalgia syndrome. *J Rheumatol* 2005;32:2270-7.
10. Carville SF, Choy EH. Systematic review of discriminating power of outcome measures used in clinical trials of fibromyalgia. *J Rheumatol* 2008;35:2094-105.
11. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry* 1961;4:561-71.
12. Stewart AL, Hays RD, Ware JE Jr. The MOS short-form general health survey. Reliability and validity in a patient population. *Med Care* 1988;26:724-35.
13. Stewart AL, Ware JE, Brook RH, Davies AR. Conceptualization and measurement of health for adults in the Health Insurance Study. Vol II. Physical health in terms of functioning. Santa Monica: RAND Corporation; 1978:236-359.
14. Burckhardt CS, Clark SR, Bennett RM. The fibromyalgia impact questionnaire: development and validation. *J Rheumatol* 1991;18:728-33.
15. Moldofsky H, Scarisbrick P, England R, Smythe HA. Musculoskeletal symptoms and non-REM sleep disturbance in patients with “fibrositis syndrome” and healthy subjects. *Psychosom Med* 1975;37:341-5.
16. Mease P, Arnold LM, Choy EH, et al. Fibromyalgia syndrome. *J Rheumatol* 2009;36:2318-29.