

## Ankylosing Spondylitis Assessment Group Preliminary Definition of Short-Term Improvement in Ankylosing Spondylitis

Jennifer J. Anderson,<sup>1</sup> Gabriel Baron,<sup>2</sup> Desiree van der Heijde,<sup>3</sup> David T. Felson,<sup>4</sup> and Maxime Dougados<sup>5</sup>

**Objective.** To develop criteria for symptomatic improvement in patients with ankylosing spondylitis (AS), using outcome domain data from placebo-controlled clinical trials of nonsteroidal antiinflammatory drugs (NSAIDs).

**Methods.** Patient data from 5 short-term, randomized, controlled trials were used to assess equivalence, reliability, and responsiveness of multiple items in the 5 outcome domains for AS treatment: physical function, pain, spinal mobility, patient global assessment, and inflammation. At least one measure per domain was responsive (standardized response mean of >0.5), except for the spinal mobility domain, which was omitted from the criteria. We developed and tested candidate improvement criteria in a random two-thirds subset from the 3 largest trials and used the remaining one-third for validation. These 3 largest trials included 923 patients (631 receiving NSAIDs, 292 in placebo groups). We selected the multiple domain definition that best distinguished NSAID treatment from placebo by chi-square test and that had a placebo response rate of ≤25%.

**Results.** Candidate definitions were changes in single domains and in multiple measure indices, as well as combinations of improvements in multiple domains.

Supported in part by a grant from Searle France and Merck, in collaboration with Boehringer Ingelheim.

<sup>1</sup>Jennifer J. Anderson, PhD: Veterans Administration Medical Center, Bedford, Massachusetts, and Boston University, Boston, Massachusetts; <sup>2</sup>Gabriel Baron, MS: Université de Vannes, Vannes, France; <sup>3</sup>Desiree van der Heijde, MD: University Hospital Maastricht, Maastricht, The Netherlands; <sup>4</sup>David T. Felson, MD, MPH: Boston University, Boston, Massachusetts; <sup>5</sup>Maxime Dougados, MD: Université René Descartes, Hôpital Cochin, Paris, France.

Address correspondence and reprint requests to Jennifer J. Anderson, PhD, Boston University Arthritis Center, A203, 715 Albany Street, Boston, MA 02118.

Submitted for publication August 3, 2000; accepted in revised form March 14, 2001.

**Worsening in a domain was defined as a change for the worse of ≥20% and a net change for the worse of ≥10 units on a scale of 0–100. Partial remission (for comparison purposes) was defined as an end-of-trial value of <20/100 in each of the 4 domains. Among 20 candidate criteria, change of ≥20% and ≥10 units in each of 3 domains and absence of worsening in the fourth discriminated best in the development subset (51% of patients improved with NSAIDs, 25% with placebo;  $\chi^2 = 36.4$ ,  $P < 0.001$ ). Results were confirmed in the validation subset. Almost all patients satisfying the definition of partial disease remission at the end of the trial had also improved by this criterion. Among all 923 patients, improvement rates using this criterion were 49% for NSAID-treated patients and 24% for placebo-treated patients.**

**Conclusion.** Although further validation using data from new trials is still needed, we conclude that we have developed a clinically valid, easy-to-use measure of short-term improvement in AS.

A core set of 5 domains that are of importance in assessing ankylosing spondylitis (AS) symptomatic outcome has already been established by the Assessments in Ankylosing Spondylitis (ASAS) Working Group (1). The domains are physical function, pain, spinal mobility, spinal stiffness/inflammation, and the patient's global assessment. It is desirable to develop standard criteria for improvement in AS using these domains. The resulting definition of improvement should provide a single primary end point with good power for use in clinical trials, as well as an improvement definition that will be useful to both patient and clinician.

To reach this goal, a well-defined process is required in which relevant types of improvement are defined, both conceptually and in terms of the amount of

**Table 1.** Basic features of studies available for developing criteria for improvement in ankylosing spondylitis (AS)\*

Trial/study (ref.)	Design	Active treatments (mg/day)	Entry criterion	Duration, weeks	Time at which measures obtained, weeks	Time at which data available, weeks	Study group n for ASAS Working Group analysis	
							Active drug	Placebo
A/Dougados et al, 1988 (4)	Parallel RCT	Piroxicam (20)	Active disease	2	0, 1, 2	0, 1, 2	35	36
B/Dougados et al, 1989 (5)	Crossover	Ximoprofen (30)	Active disease	1†	0, 1	0, 1	18	18
C/Dougados et al, 1994 (6)	Parallel RCT	Ximoprofen (5, 10, 20, 30)	2-day washout	2	0, 1, 2	0, 1, 2	189	95
D/Dougados et al, 1999 (7)	Parallel RCT	Piroxicam (20), meloxicam (15, 22.5)	2–15-day washout	6	0, 1, 3, 6	0, 6	352	121
E/Dougados et al, 2001 (8)	Parallel RCT	Ketoprofen (200)	2–15-day washout	6	0, 1, 3, 6	0, 1, 3, 6	90	76

\* ASAS = Assessments in AS; RCT = randomized controlled trial.

† First period.

change that constitutes real improvement. Once this is done, candidate definitions of improvement can be proposed and tested for validity, discrimination, and feasibility using data from appropriate clinical trials, so that a definition that satisfies these 3 conditions can be selected.

A variety of information is now available to assist in achieving this goal. Recommendations for choices of measures to represent each domain have been made (2), and the reliability of different ways of measuring AS-specific outcomes within each of the 5 domains has been assessed (3). Although early trials of nonsteroidal anti-inflammatory drugs (NSAIDs) in AS did not always include a patient global assessment (4,5), there are now 3 recent, large, randomized, short-term clinical trials of NSAIDs versus placebo in which each of the 5 domains is represented by at least one measure, and for which the patient-level data are also available for detailed analysis (6–8). Thus, it is now possible to develop criteria for symptomatic improvement in AS.

Improvement/response criteria have already been developed in several areas within rheumatology, specifically in rheumatoid arthritis (RA) (9,10), juvenile arthritis (11), and osteoarthritis (OA) (12). The process and methods used in each case have been slightly different. Each has features to recommend it, and the AS-specific process draws from these experiences. We present details of the process and the results obtained in the case of AS, with a focus on short-term effects on symptoms of axial disease, and make some comparisons with criteria developed for other rheumatic diseases.

## PATIENTS AND METHODS

We completed 3 steps in developing criteria for symptomatic improvement in AS. First, we defined relevant levels of change/improvement per item within each of the 5 AS core domains and chose the most reliable and sensitive items from each domain. This work drew on information in 5 randomized, controlled trials (4–8). These were all the trials we could identify, from which patient-level data were available to us, that included measures representing at least 4 of the 5 AS core domains. Second, we prepared a conceptual list of ways of defining improvement in terms of some or all of these 5 core domains. The list was developed from a combination of clinical judgment, analogy with previous work in OA and RA, and results of published clinical studies of AS (6–8,13). Third, we determined the sensitivity and specificity of the selected candidate definitions of AS symptomatic improvement using a two-thirds random sample of data from the 3 large trials with information on patient global assessment (6–8). Preferred definitions have both a low response rate in the placebo group (i.e., good specificity) as well as a large difference between response rates for the active drug- and placebo-treated patients. We used the remaining one-third of the observations for validation before making a final choice of AS symptomatic response criteria. We also examined consistency between improvement rates and partial remission rates. These methods are now described in more detail.

**Choice of items to represent domains.** We worked with patient-level data from 5 trials of NSAIDs in AS. The trials, results of which were published between 1988 and 2001 (4–8), were all short-term (1–6 weeks) trials of  $\geq 1$  NSAIDs versus placebo in patients with axial manifestations of AS. Table 1 indicates the basic features of the trials (labeled A–E), and Table 2 shows the AS core domain measures available from each trial. Table 3 provides some characteristics of the patients for whom data were available for this study. Altogether, there were almost twice as many patients treated with NSAIDs as

**Table 2.** Outcome measures available by core domain in the 5 NSAID trials\*

	Trial				
	A	B	C	D	E
Physical function					
BASFI (0–100 scale)				x	x
DFI (0–40 scale)	x	x	x	x	
Pain					
100-mm VAS (past week or past 48 hours)	Past week	Past week	Past week	Past 48 hours	Past 48 hours
Spinal mobility (cm)					
Schober test	x	x	x	x	x
Fingers-to-floor distance	x	x	x	x	x
Chest expansion	x	x		x	x
Occiput-to-wall distance				x	
Patient global assessment					
Overall (5 grades)			x		
100-mm VAS				x	x
Inflammation					
Morning stiffness (minutes)	x	x	x	x	x
Nocturnal awakenings (no.)	x	x	x		
Sleep impairment (4-point scale)				x	x
Night pain VAS (past 48 hours)					x
BASDAI morning stiffness duration VAS				x	x
BASDAI morning stiffness intensity VAS				x	x
C-reactive protein (mg/dl)				x	x

\* NSAID = nonsteroidal antiinflammatory drug; BASFI = Bath Ankylosing Spondylitis Functional Index; DFI = Dougados Functional Index; VAS = visual analog scale; BASDAI = Bath Ankylosing Spondylitis Disease Activity Index.

were treated with placebo (684 versus 346) in the 5 trials, and 74.6% of all patients were men. The mean age of all patients was 40.7 years. The mean disease duration was 11.0 years, and 82.7% of patients were HLA-B27 positive.

We examined the distributional characteristics of the items contained within each of the 5 domains that were available in these data sets. We obtained the mean  $\pm$  SD and range for

each available measure both at baseline and at the end of the trial, and for the change in the measure during the trial. We calculated the effect size (ES), a standardized measure of the amount of change, as well as the standardized response mean (SRM) for each variable. The SRM divisor is a change SD rather than a baseline SD, and so measures the relative ease of detecting change or responsiveness of a measure.

**Table 3.** Characteristics of patients with ankylosing spondylitis in 5 trials (A–E) available for responder criteria development

Trial, treatment group	Age in years, mean $\pm$ SD	% male	HLA-B27 positive, no. (%) / no. missing values	Disease duration in years, mean $\pm$ SD	Patients dropping out, no. (%)
<b>A</b>					
Placebo (n = 36)	38.0 $\pm$ 11.8	80.6	22 (73.3)/6	8.4 $\pm$ 6.5	6 (16.7)
Active drug (n = 35)	34.4 $\pm$ 8.5	82.9	27 (90.0)/5	9.1 $\pm$ 6.5	1 (2.9)
<b>B</b>					
Placebo (n = 18)	39.1 $\pm$ 13.2	100	12 (80.0)/3	9.5 $\pm$ 8.1	0 (0)
Active drug (n = 18)	39.5 $\pm$ 12.7	83.3	15 (93.8)/2	12.4 $\pm$ 8.6	0 (0)
<b>C</b>					
Placebo (n = 95)	39.5 $\pm$ 10.6	68.4	71 (74.7)/0	9.7 $\pm$ 7.7	19 (20)
Active drug (n = 189)	40.0 $\pm$ 11.1	68.3	145 (76.7)/0	9.1 $\pm$ 8.7*	18 (9.5)
<b>D</b>					
Placebo (n = 121)	40.1 $\pm$ 11.6	71.9	77 (89.5)/35	11.9 $\pm$ 9.0	40 (33.1)
Active drug (n = 352)	43.5 $\pm$ 12.1	80.4	219 (85.2)/95	12.3 $\pm$ 9.8	47 (13.4)
<b>E</b>					
Placebo (n = 76)	40.2 $\pm$ 10.5	71.1	58 (84.1)/7	11.3 $\pm$ 9.1	32 (42.1)
Active Drug (n = 90)	38.1 $\pm$ 11.3	66.7	71 (88.8)/10	11.7 $\pm$ 9.7	23 (25.6)
<b>Total</b>					
Placebo (n = 346)	39.7 $\pm$ 11.2	73.1	240 (81.4)/51	10.7 $\pm$ 8.4	97 (28.0)
Active drug (n = 684)	41.2 $\pm$ 11.9	75.4	477 (83.4)/112	11.2 $\pm$ 9.4*	89 (13.0)

\* One value was missing.

We explored reliability of the variables, obtaining smallest detectable difference (SDD) estimates by the Bland and Altman technique (14) from repeated measures in two of the trials (A and C) and comparing these with other SDD estimates available for AS outcome measures (3). The SDD for a measure was defined as  $1.96 \times \text{SD}$  of the difference in that measure between two closely spaced visits occurring during a period in which the measure was expected to be stable. The SDD is a guide to the minimum amount of observed change in a single measure that can be considered to represent a real change in that measure as distinct from noise.

We also used correlations to examine associations between variables available within domains, to the extent permitted by multiplicities and overlap among the data sets, as a further aid in identifying the appropriate domain representatives to use.

These steps led to the omission of one domain from further consideration in defining short-term improvement criteria in AS. Responsiveness and reliability measures were used to refine clinical judgment regarding how large a change in observed variables would correspond to real improvement or real worsening.

A final task before setting up candidate improvement criteria for testing was to develop standardized forms of the variables for use in pooled indices. Standardization consisted of dividing the outcome variable value by a standardizing value, in this case the pooled baseline SD from both treatment and placebo groups.

**Conceptual list of ways of defining improvement.** As a basic principle, we determined that improvement should reflect a change from baseline values, which could include either an absolute amount of change or change as a percentage of baseline levels, or both, but which should not require reduction below a specific level of disease activity. We defined a state of partial remission as a separate entity to correspond to an end result with a low level of disease activity. We considered several different types of definition of improvement, as described below.

*Single outcomes.* Historically, substantial improvements in single outcomes have been used to define response. For example, as part of the presentation of results for two large trials in AS (6,7), a 50% response in pain (both trials) and also, separately, in physical function and in the patient's overall assessment (one trial only) had been considered natural definitions of response. We included single-item responder definitions in our set of candidate criteria. Criteria requiring only a percentage improvement and others requiring the combination of a percentage improvement with a minimum amount of improvement are represented.

*Multiple domains.* Results from a 1990 report of an open-label study of 985 patients (13) indicate that over a treatment period of 6 weeks, improvement in each of the 4 domains of pain, function, mobility, and stiffness contributed independently to improvement in the fifth domain (patient's global assessment). The results indicate that there is substantial, but not complete, overlap among these outcomes. Thus, all 5 domains should be relevant in defining improvement. Boolean-type improvement criteria can be expected to have a strong ability to discriminate active drug from placebo responses, because combinations would maintain sensitivity while possibly increasing specificity. Thus, we included a

variety of such combinations in the list of candidate improvement criteria including, for example, 1) improvement in all domains, and 2) improvement in at least 3 domains, without worsening in any.

*Indices.* Changes in the Disease Activity Score (DAS), which is a linear combination of functions of core set items and thus a particular type of index, are a primary component of the European League Against Rheumatism (EULAR) response criteria for RA (10). Work done in developing the American College of Rheumatology (ACR) improvement criteria (9) established that changes in standardized indices (other than the DAS) that were based on all or some of the RA core set items were quite sensitive, but not particularly specific. Results could differ for AS, so we included some pooled indices of this type among the candidates for evaluation. We used the baseline SD of each index component in standardization.

**Testing and validation of candidate definitions of improvement.** We used a random two-thirds of the clinical trial data from the 3 large trials to determine, via the chi-square test in  $2 \times 2$  tables, the sensitivity and specificity of each of the candidate definitions with respect to identification of actively treated patients. We treated cases with missing data conservatively. If outcome variable data were missing at baseline, we assumed a value of 0 on a scale of 0–100 ( $\leq 3$  cases in 4 variables). We used the last observation carried forward technique for imputation of outcomes missing at the end of a trial.

On the basis of high chi-square test values and placebo response rate of  $\leq 25\%$ , we selected a range of good performers for validation using the remaining one-third of the data. For a further validation, we also examined the overlap between response rates and partial remission rates for each definition of response.

## RESULTS

We used data on all available measures to select a measure to represent each domain, then tested the discrimination of candidate criteria formed using these measures.

**Choice of domain measure from multiple measures available.** Within the domain of physical function, scores from either the Bath Ankylosing Spondylitis Functional Index (BASFI) or the Dougados Functional Index (DFI) were available in each trial, with both obtained on all patients in study D. These two measures have been shown to have very similar properties (15). We used the overlap in study D to estimate a quadratic relationship between BASFI and DFI values so as to estimate the BASFI score if only the DFI score was available. The transformation,  $\text{BASFI} = 3.835 \times \text{DFI} - 0.03197 \times \text{DFI}^2$ , with  $R^2 = 0.934$ , produces close correspondence at both low and high ends of the scale (0 and 100) while capturing observed differences between the two measures in the rate of increase of the score with increasing problems with physical function.

For the pain domain, all 5 trials included a pain

**Table 4.** Effect sizes (ES) and standardized response means (SRM) in development subset of trials A–C\*

Variable (scale)	n <sub>p</sub>	n <sub>t</sub>	Δplacebo, mean ± SD	Δtreatment, mean ± SD	SD common baseline	ES†	SD common change	SRM†
DFI (0–40)	102	154	−0.8 ± 5.9	−4.4 ± 5.9	6.2	−0.58	6.1	−0.59
VAS pain (0–100)	102	154	−11.3 ± 24.9	−27.8 ± 23.9	17.4	−0.95	25.6	−0.64
Chest expansion in trials A and B (0–100)	35	38	−3.9 ± 13.8	−1.4 ± 7.9	22.7	0.11	11.4	0.23
Schober test (0–100)	100	151	−1.4 ± 10.4	−5.4 ± 12.3	12.4	−0.32	11.8	−0.34
Fingers-to-floor distance (0–100)	102	154	−2.1 ± 18.5	−9.2 ± 19.8	27.4	−0.26	19.6	−0.36
Morning stiffness (0–120)	102	154	−4.3 ± 33.5	−23.1 ± 39.0	37.7	−0.50	38.0	−0.50
Nocturnal awakenings (no.)	102	154	−0.5 ± 1.8	−1.1 ± 2.1	1.9	−0.30	2.0	−0.28
Global change in trial C (0–4)	64	119	−0.2 ± 0.9	−0.8 ± 1.1	0.7	−0.89	1.1	−0.55

\* ES = (Δtreatment − Δplacebo)/SD common baseline; SRM = (Δtreatment − Δplacebo)/SD common change; n<sub>p</sub> = no. of patients in placebo group; n<sub>t</sub> = no. of patients in treatment group (see Table 2 for other definitions).

† Some ES and SRM calculations do not match exactly because of rounding in the presentation of changes and SDs.

visual analog scale (VAS), although the definition and wording varied (earlier trials [A–C] referred to the past week, while the two most recent trials [D and E] referred to the past 48 hours). A variety of metrology measures were available to represent spinal mobility, although only the Schober test and the fingers-to-floor distance were available in all 5 trials. Trials D and E included a VAS for patient global assessment, while trial C included only a 5-grade measure (grades 0–4), and trials A and B lacked this assessment. Inflammation was represented in each trial by morning stiffness in minutes and in trials A–C by night awakenings as well, while trials D and E included a sleep impairment scale, the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) VAS questions on duration and intensity of morning stiffness, and C-reactive protein (CRP) levels. There was also a night pain VAS available (only in trial E).

Since some variables of interest were only available in trials A–C and others were only available in trials

D and E, we examined individual outcome variable performance in two separate data sets. The first was a random two-thirds subset of the combination of trials A–C (Table 4), while the second was a similar subset of trials D and E (Table 5). Table 4 shows that the physical function measure (DFI), pain VAS, and patient global assessment measure (the last available only in trial C) all exhibited substantial ESs (of >0.50) and SRMs (also of >0.50) in the development subset of trials A–C (in these tables, a negative ES and SRM correspond to greater improvement in the active drug treatment group than in the control group).

The 3 spinal mobility measures did not perform as well, with both ES and SRM considerably lower than 0.5. The direction of the effect for chest expansion (available only in the smaller trials of A and B) was counter to the prevailing direction for other outcome variables. Each of these spinal mobility variables was transformed linearly from its original scale in centime-

**Table 5.** ES and SRM in development subset of trials D and E\*

Variable (scale)	n <sub>p</sub>	n <sub>t</sub>	Δplacebo, mean ± SD	Δtreatment, mean ± SD	SD common baseline	ES†	SD common change	SRM†
BASFI (0–100)	133	289	1.74 ± 21.2	−11.7 ± 20.8	22.4	−0.60	21.8	−0.61
VAS pain (0–100)	133	287	−14.7 ± 26.8	−29.7 ± 26.3	16.0	−0.94	27.3	−0.54
Chest expansion (0–100)	132	288	0.7 ± 12.1	−3.2 ± 13.7	21.8	−0.18	13.6	−0.29
Schober test (0–100)	132	289	−0.9 ± 12.2	−3.7 ± 10.9	14.7	−0.19	11.4	−0.24
Fingers-to-floor distance (0–100)	132	289	1.4 ± 18.8	−6.0 ± 16.9	30.2	−0.25	17.9	−0.42
Global VAS (0–100)	133	288	−7.1 ± 26.1	−23.2 ± 27.8	20.0	−0.81	28.2	−0.57
Morning stiffness (0–120)	132	289	−4.0 ± 39.4	−21.4 ± 36.1	38.3	−0.45	38.0	−0.46
Night pain VAS in trial E (0–100)	53	55	−1.4 ± 29.6	−13.8 ± 33.5	29.1	−0.43	32.1	−0.39
Sleep impairment (1–4)	131	289	−0.1 ± 0.8	−0.7 ± 0.9	0.7	−0.80	0.9	−0.64
BASDAI morning stiffness VAS scores (mean)‡	133	289	−4.3 ± 27.9	−19.3 ± 26.7	24.7	−0.61	27.9	−0.54
C-reactive protein (mg/dl)	127	271	0.76 ± 12.16	−0.2 ± 13.29	15.8	−0.06	32.1	−0.03

\* See Tables 2 and 4 for definitions and other explanations.

† Some ES and SRM calculations do not match exactly because of rounding in the presentation of changes and SDs.

‡ Mean of two morning stiffness-related BASDAI VAS scores, one for duration of morning stiffness (0 = none, 100 = ≥120 minutes) and the other for intensity of morning stiffness (0 = none, 100 = very severe).

ters to a scale of 0–100, in which 0 corresponded to the best value ( $\geq 10$  cm for chest expansion or for the Schober test, 0 cm for the fingers-to-floor distance) and 100 corresponded to the worst value (0 cm for chest expansion or for the Schober test,  $\geq 50$  cm for the fingers-to-floor distance). Of the two variables available in the inflammation domain, morning stiffness (recoded to have a maximum of 120 minutes) performed adequately, with ES and SRM at  $-0.50$ , while the number of nocturnal awakenings had greater relative variability and thus rather low responsiveness as measured by ES and SRM.

The results in Table 5 from the development subset of trials D and E were similar, with the BASFI, the pain VAS, and the patient global assessment VAS each showing good responsiveness. In trial D, where the DFI and BASFI were both measured, the BASFI had slightly greater change and responsiveness (ES =  $-0.60$ , SRM =  $-0.65$ ) than the DFI (ES =  $-0.53$ , SRM =  $-0.58$ ; data not shown).

The results for the various spinal mobility measures were uniformly poor, however, and the results for the inflammation variables were mixed. It is clear that CRP level was not responsive in the short term (SRM =  $-0.03$ ), and the night pain VAS (available only in trial E) also was not particularly responsive (SRM =  $-0.39$ ). The sleep impairment scale was the most responsive (SRM =  $-0.64$ ), while the mean of the two morning stiffness-related BASDAI VAS scores (one for duration of morning stiffness [0 = none, 100 =  $\geq 120$  minutes] and the other for intensity of morning stiffness [0 = none, 100 = very severe]) had an SRM of  $-0.54$ , and morning stiffness duration alone was slightly less responsive (SRM =  $-0.46$ ). Pearson correlations of changes in each of night pain, morning stiffness, the BASDAI mean, and the sleep impairment scale (all in trial E) showed that the correlation between change in sleep impairment and change in night pain was high ( $r = 0.71$ ), as was the correlation between change in morning stiffness and change in BASDAI mean ( $r = 0.66$ ), while all other correlations were lower ( $r = 0.33$ – $0.50$ ). This led us to choose the BASDAI mean as the preferred measure of inflammation, with duration of morning stiffness to be used if BASDAI data were not available. Morning stiffness was chosen over the sleep impairment variable because it was the more responsive of the variables available in trial C and was very similar to the BASDAI-based measure.

We omitted spinal mobility from further consideration for short-term improvement criteria because of the lack of responsiveness of any available representa-

tive of this domain. We also restricted our further development and testing of criteria to include data from only the 3 later and larger trials (C–E), since only these trials included measures representing each of the other 4 domains, with physical function represented by BASFI (converted from DFI in trial C), pain represented by a VAS in all 3 trials, the patient global assessment represented by a VAS in trials D and E and by a scale of 0–4 in trial C, and inflammation represented by the BASDAI mean in trials D and E and by morning stiffness alone in trial C.

We estimated the smallest detectable difference (SDD) for variables in trials A and C. Patients in trial A had been asked whether or not they thought they had experienced improvement since the last visit. We used the repeated observations for those who indicated no improvement between week 1 and week 2 among both treated and control patients ( $n = 32$ ) to determine an SDD (on a scale of 0–100) by the Bland and Altman method (14) for each of DFI (SDD = 12.0), the pain VAS (SDD = 19.9), and duration of morning stiffness (SDD = 23.3). Also, in trial C, we used the week 1 and week 2 observations for the placebo-treated patients ( $n = 95$ ) to obtain separate SDD estimates, obtaining the higher values (expressed on a scale of 0–100) of 19.1 for DFI, 32.8 for the pain VAS, and 38.3 for morning stiffness, as well as 29.5 for the (0–4-scaled) patient global assessment measure (data not shown). These may be compared with the values obtained in a recent multinational reliability study (3) of 17.3 for DFI, 21.3 for BASFI, 33.0 for a pain VAS, 30.4 for morning stiffness, and 38.4 for a patient global VAS, each expressed on the same scale of 0–100. These 3 sources gave SDD estimates ranging from 12 to 21 for physical function, 20 to 33 for a pain VAS, 23 to 38 for morning stiffness, and 30 to 38 for a patient global assessment. This suggests that, for individual variables, the minimum change that should be considered detectable would be of the order of 20–30 units on a scale of 0–100.

Table 6 shows baseline mean  $\pm$  SD values for the 4 measures chosen to represent the domains in candidate criteria, as well as their pooled SDs and the mean  $\pm$  SD values of changes during the trials, expressed in the standard units. It is apparent from Table 6 that the typical patient enters one of the trials with levels of the variables near (in the case of physical function) or exceeding (for the other 3 domains) the midpoint of the possible range for the domain. The pooled SDs are comparable with SDDs for physical function and inflammation, but just below the range of SDDs obtained in the other domains. The differences between NSAID-

**Table 6.** Baseline values and standardized changes of variables used in forming improvement criteria\*

Domain	Domain measure used in criteria	Baseline value (0–100 scale), mean $\pm$ SD		Pooled SD used	Change during trial in standardized form, mean $\pm$ SD	
		Placebo group (n = 197)	Active drug group (n = 408)		Placebo group (n = 197)	Active drug group (n = 408)
Physical function	BASFI	47.1 $\pm$ 22.0	48.8 $\pm$ 20.2	20.8	0.0 $\pm$ 1.0	0.6 $\pm$ 1.0
Pain	Pain VAS	70.4 $\pm$ 15.7	69.0 $\pm$ 17.0	16.6	0.8 $\pm$ 1.6	1.8 $\pm$ 1.6
Patient global assessment	Global assessment	66.1 $\pm$ 19.9	64.1 $\pm$ 19.0	19.8	0.3 $\pm$ 1.3	1.1 $\pm$ 1.4
Inflammation	BASDAI morning stiffness VAS	55.1 $\pm$ 28.2	53.9 $\pm$ 27.1	27.4	0.1 $\pm$ 1.0	0.7 $\pm$ 1.1

\* See Tables 2 and 5 for definitions and explanations.

treated and placebo group changes observed during the trial were larger for pain and for global assessment (1.0 and 0.8 standard units, respectively, compared with 0.6 units in both inflammation and physical function). Pain and patient global assessment also had standardized change SDs somewhat greater than 1.0, indicating (as was also seen in Tables 4 and 5) that the SRM was less than the ES (i.e., that these larger changes were rela-

tively noisy). These results do suggest, however, that the 4 domains can be handled very similarly in developing improvement criteria.

#### Testing of candidate definitions of improvement.

We tested 20 different candidate criteria (see Table 7), including 12 individual domain-based criteria (50% improvement in each separate domain [criteria 1–4], the greater of  $\geq 20\%$  change and net change of  $\geq 10$  units in

**Table 7.** Development subset assessment of candidate improvement definitions\*

Criterion	Improvement definition	% improving in placebo-treated group (n = 197)	% improving in active drug-treated group (n = 408)	Difference	$\chi^2$	No. improved of 62 patients with partial disease remission
1	$\geq 50\%$ change in PF	13	34	21	30.7	57
2	$\geq 50\%$ change in PA	23	46	23	29.5	61
3	$\geq 50\%$ change in PG	19	42	23	32.1	61
4	$\geq 50\%$ change in IN	19	43	24	33.7	53
5	$\geq 20\%/10$ units change in PF	25	48	23	29.0	54
6	$\geq 20\%/10$ units change in PA	43	69	26	38.2	61
7	$\geq 20\%/10$ units change in PG	33	59	26	35.4	60
8	$\geq 20\%/10$ units change in IN	32	55	23	29.2	57
9	$\geq 30\%/20$ units change in PF	13	30	17	21.9	41
10	$\geq 30\%/20$ units change in PA	35	60	25	35.4	60
11	$\geq 30\%/20$ units change in PG	26	52	26	38.1	59
12	$\geq 30\%/20$ units change in IN	23	42	19	20.9	48
13	$\geq 30\%/20$ units change in PA, PG, and IN, and $\geq 20\%/10$ units change in PF	11	28	17	22.0	44
14	$\geq 20\%/10$ units change in all 4 domains	15	34	19	24.3	50
15	$\geq 20\%/10$ units change in any 3 of 4 domains, no worsening in fourth domain	25	51	26	36.4	59
16	$\geq 20\%/10$ units change in any 2 of 4 domains, no worsening in other 2 domains	35	64	29	45.7	61
17	$\geq 0.5$ units change in index†	51	78	27	44.4	62
18	$\geq 1.0$ units change in index†	48	75	27	45.0	62
19	$\geq 1.5$ units change in index†	41	72	31	53.8	62
20	$\geq 2.0$ units change in index†	37	67	30	50.8	62

\* Domains included the following: physical function (PF), pain (PA), patient global assessment (PG), and inflammation (IN).

† The final 4 criteria tested were changes that were multiples of a pooled index formed by averaging the standardized forms of the variables representing each domain. See text for details.

**Table 8.** Improvement definition performance showing development and validation sets combined\*

Criterion	Improvement definition	% improving in placebo-treated group (n = 292)	% improving in active drug-treated group (n = 631)	Difference	$\chi^2$	No. improved of 79 patients with partial disease remission
1	$\geq 50\%$ change in PF	13	32	19	37.8	72
2	$\geq 50\%$ change in PA	21	47	26	60.5	78
3	$\geq 50\%$ change in PG	16	41	25	55.7	78
4	$\geq 50\%$ change in IN	18	43	25	51.7	70
13	$\geq 30\%/20$ units change in PA, PG, and IN, and $\geq 20\%/10$ units change in PF	10	28	18	38.9	57
14	$\geq 20\%/10$ units change in all 4 domains	13	34	21	43.4	64
15	$\geq 20\%/10$ units change in any 3 of 4 domains, no worsening in fourth domain	24	49	25	54.2	75
16	$\geq 20\%/10$ units change in any 2 of 4 domains, no worsening in other 2 domains	33	63	30	72.3	78
19	$\geq 1.5$ units change in index	39	70	31	79.2	78
20	$\geq 2.0$ units change in index	35	66	31	78.7	78

\* See Table 7 for definitions and explanations.

each separate domain [criteria 5–8], and the greater of  $\geq 30\%$  change and net change of  $\geq 20$  units in each separate domain [criteria 9–12]). The 8 multiple-domain criteria tested included 4 Boolean combinations, as follows: for criterion 13,  $\geq 30\%$  change/net improvement of  $\geq 20$  units in each of inflammation, pain, and patient global assessment, coupled with  $\geq 20\%$  change/net improvement of  $\geq 10$  units in physical function; for criterion 14,  $\geq 20\%$  change/net improvement of  $\geq 10$  units in each domain; for criterion 15,  $\geq 20\%$  change/net improvement of  $\geq 10$  units in each of 3 domains, with the fourth domain not worsening; and for criterion 16,  $\geq 20\%$  change/net improvement of  $\geq 10$  units in each of 2 domains, with the other 2 domains not worsening. Worsening in a particular domain was defined symmetrically (based on clinical judgment and the simplicity principle) to be a change for the worse of  $\geq 20\%$  and a net change for the worse of  $\geq 10$  units on a scale of 0–100.

The final 4 criteria tested were changes that were multiples of a pooled index formed by averaging the standardized forms of the variables representing each domain. Thus, from Table 6 data, the index = average of physical function/20.8, pain/16.6, patient global assessment/19.8, and inflammation/27.4, where each variable is on a scale of 0–100. We tested criteria 17–20, corresponding to  $\geq 0.5$ ,  $\geq 1.0$ ,  $\geq 1.5$ , and  $\geq 2.0$  units of change, respectively, in the 4-domain index.

The assessment results are also shown in Table 7, where it is apparent that each candidate definition discriminated strongly between active drug- and placebo-treated patients ( $\chi^2 = 10.83$  was required for significance at the 0.001 level). The largest chi-square

value obtained corresponded to 1.5 units of change in the index, but this was associated with a 41% improvement rate in the placebo group. The criteria satisfying the additional condition that the placebo improvement rate should be  $\leq 25\%$  included each of the individual-domain 50% improvement definitions, the two other individual domain criteria for physical function, one for inflammation, criterion 12, and 3 of the 4 Boolean combinations, including criterion 15 ( $\geq 20\%$  change/net improvement of  $\geq 10$  units in each of 3 domains, with the fourth domain not worsening), which had the highest chi-square value (36.4) among those criteria that satisfied the additional condition.

We defined partial remission as a value of  $< 20/100$  in each of the 4 domains at the end of the trial. Among the 605 patients in the development subset, 62 (10%) had partial remission of their disease, and the last column of Table 7 shows the number among these 62 patients who also improved by each tested definition. Criterion 15 was among those that were most consistent with partial remission, with  $\geq 59$  patients having improved ( $> 95\%$  of the 62 patients with partial remission of their disease).

Based on their performance in the development subset, we selected 10 candidate criteria (the 4 individual-item 50% improvement criteria, the 4 Boolean combination criteria, and the criteria corresponding to  $\geq 1.5$ - and  $\geq 2.0$ -unit changes in the pooled index) for validation of performance in the remaining one-third of the trial data, and obtained very similar results. Table 8 shows their overall performance, in the full data set of all 923 observations, which confirmed that criterion 15 had the best discrimination as measured by the chi-square

**Table 9.** Assessments in Ankylosing Spondylitis Working Group criteria for response\*

---

Improvement of $\geq 20\%$ and absolute improvement of $\geq 10$ units (on a scale of 0–100) in $\geq 3$ of the following 4 domains:
Patient global assessment
Pain
Function
Inflammation
Absence of deterioration in the potential remaining domain, where deterioration is defined as a change for the worse of $\geq 20\%$ and net worsening of $\geq 10$ units (on a scale of 0–100)

---

\* Patient global assessment is represented by the VAS global assessment score (0–100 scale). Pain is represented by the VAS pain score (0–100 scale). Function is represented by the BASFI score (0–100 scale). Inflammation is represented either by (first choice) the mean of the two morning stiffness–related BASDAI VAS scores, or by (second choice) morning stiffness duration with a maximum of 120 minutes (0–100 scale). See Tables 2 and 5 for definitions and explanations.

statistic among multiple component criteria in which the placebo response rate did not exceed 25%.

## DISCUSSION

We have developed a preliminary definition of short-term improvement in AS that incorporates 4 outcome domains and has clinical validity and good performance in initial tests. We used clinical trial data to refine choices of items in the domains, then developed a variety of candidate definitions of improvement and tested and compared the performance of these candidate criteria. Our choice is improvement by  $\geq 20\%$  and net improvement of  $\geq 10$  units on a scale of 0–100 in each of 3 domains, with no worsening in the fourth, where the domains are physical function, pain, patient global assessment, and inflammation. Worsening in a specific domain has the mirror definition of deterioration, both by  $\geq 20\%$  and by  $\geq 10$  units on a scale of 0–100. The definition of improvement, summarized in Table 9, is straightforward and balanced with respect to the domains of interest in which short-term improvement can be expected to occur.

Occasionally, reports of individual trials in AS have defined improvement as a 50% change in a single domain, such as pain (6,7) or physical function or patient global assessment (7). These single-domain measures performed well in our tests, and since improvement in these 3 domains and in the domain of inflammation tends to be consistent, one could argue that it would be sufficient to consider a single domain when defining improvement, in which case either the pain or the patient global assessment domain would suffice. But a definition of improvement that takes multiple domains

into account has greater content validity and so is inherently preferable, since it is always possible that cases with inconsistent results could occur, and these would not be examples of real improvement.

Another point in favor of multiple domain–based measures of improvement, at least in rheumatic diseases, is the fact that many of the relevant outcome measures, considered individually, have relatively poor reliability, as evidenced by SDDs ranging between 20% and 40% of the full range of each variable. Therefore, a somewhat smaller amount of change, say, of the order of 0.5 SDD, in a single variable, for a single patient, is too small to be detectable as real. However, when consistent changes of this order of magnitude appear to occur in several domains simultaneously, one can have greater confidence in the idea that the change is real. Standardized indices also summarize information from multiple domains, and they tend to be highly sensitive, but not very specific, when used as a sole criterion. This, together with the difficulty in expressing the index or its change in clinically accessible terms, makes these particular representatives of multiple domain–based measures less desirable in practice.

In addition to defining improvement, we have created a definition of partial remission in AS that requires the patient to have a low level of disease activity (i.e.,  $< 20$  units on a scale of 0–100 in each of the 4 domains represented in the definition of improvement [see Table 10]). We used this definition to check on the reasonableness of our responder results. We have chosen to separate improvement conceptually from partial remission, since although there is typically considerable overlap in a clinical trial setting between those who improve and those who reach a low level of disease activity, this will not always be the case. In a study of severely ill patients, there could be a considerable

**Table 10.** Assessments in Ankylosing Spondylitis Working Group definition of partial remission\*

---

Value of $< 20$ (on a scale of 0–100) in each of the following 4 domains:
Patient global assessment
Pain
Function
Inflammation

---

\* Patient global assessment is represented by the VAS global assessment score (0–100 scale). Pain is represented by the VAS pain score (0–100 scale). Function is represented by the BASFI score (0–100 scale). Inflammation is represented either by (first choice) the mean of the two morning stiffness–related BASDAI VAS scores, or by (second choice) morning stiffness duration with a maximum of 120 minutes (0–100 scale). See Tables 2 and 5 for definitions and explanations.

number who improve without achieving partial remission of their disease. Conversely, in a study of patients who have already achieved some improvement, many may already have partial disease remission and be unable to improve further by our definition. This is largely because the lack of reliability of the outcome measures precludes an observed drop of <10 units on a scale of 0–100 in any one outcome from being described as real change. A partial remission definition may also be useful for across-trial comparisons and for characterizing and comparing patient populations with respect to the proportions in partial disease remission.

The AS short-term improvement criteria that we have developed may be compared with previously established criteria for improvement in RA (9,10), juvenile arthritis (11), and OA (12). In RA, the ACR improvement criteria require improvement by  $\geq 20\%$  both in tender and swollen joint counts and by  $\geq 20\%$  in  $\geq 3$  of the other 5 core set items (pain, physical function, patient global assessment, physician assessment, and erythrocyte sedimentation rate), while the EULAR response criteria for RA (10) factor in both change and absolute level of disease. The EULAR response definition requires that there be improvement in the DAS (by at least twice the estimated measurement error of the DAS) coupled with attainment of a low level of disease activity. In contrast, OA improvement criteria require a large amount (and percentage) of improvement in the single domain of pain, or smaller amounts and percentages of improvement in 2 of 3 domains (pain, functional impairment, and patient global assessment). In juvenile arthritis (11), the definition requires that 3 of 6 core set measures improve by 30%, and that no more than 1 measure worsen by  $>30\%$ .

Comparison of the improvement criteria in AS with these several previously developed sets of criteria indicates that a variety of approaches have been used, each of which combines clinical input with data-based testing. Improvement may be described in terms of percentage improvements, amounts of improvement, avoidance of worsening, and even level attained, but because of the overlap in practice among these various entities, the end results are quite similar in function and performance (for example, see ref. 16).

A potential concern with the proposed AS improvement definition is that the poor performance of the spinal mobility measures available in the short-term trials led us to drop them from further consideration in our development of improvement criteria. Spinal mobility is of obvious importance in AS, and constitutes a domain that is specific to AS with axial involvement, as

distinct from the other domains of physical function, pain, patient global assessment, and inflammation that are experienced in almost every rheumatic disorder and/or inflammatory process. It is therefore disappointing that spinal mobility did not have a role in these improvement criteria.

Several factors contributed to this. One is the relatively short-term nature of the trials studied and the fact that the NSAID treatment in the trials may have affected symptoms only. Longer trials, trials of disease-modifying antirheumatic drugs or new therapeutic agents, or trials of physical therapy could have different results. However, it is also clear that the measures available to us may not have been the most relevant (in that each included components other than spinal mobility alone) or the most responsive. Chest expansion in particular (somewhat akin to grip strength in RA) is a measure that may not be amenable to improvement once the patient has reached a certain stage of the disease. This is an area where further work is needed, both in developing more responsive measures that truly represent spinal mobility and in using them in trials, and then in devising expanded improvement criteria that can incorporate these new domains and thus reflect both the disease and the treatment aims more adequately.

In the other 4 domains as well, it is quite possible that measures will be identified that have performance superior to that of our choices to represent these domains. The Health Assessment Questionnaire (17), for example, has been found to be more responsive than the DFI (18). We chose the BASFI over the DFI in developing these criteria because of its slightly greater responsiveness in the data available to us. As evidence favoring other measures to represent specific domains becomes available, it can be used to revise our recommendations for specific choices of measures (Table 9).

In conclusion, we have been able to develop and test a measure of improvement in AS that has both clinical and empirical validity and that should be very useful as a summary outcome measure in clinical trials of symptom-modifying therapy. It summarizes change in the domains of physical function, pain, patient global assessment, and inflammation. Its simple definition,  $\geq 20\%$  improvement and net improvement of  $\geq 10$  units in each of 3 domains, with no worsening in the fourth, will make it easy to remember and use. Within each domain, this change can be readily visualized by thinking of it as 10 units of improvement if the patient starts the trial below the halfway mark (50 on the scale of 0–100) for a particular measure, and as improvement by 20% from baseline if the patient starts the trial at or above the

halfway mark (i.e., in a more severe disease state). This simplicity, together with the fact that the definition requires a minimum amount of change per domain while avoiding the explicit requirement of partial remission, suggests that it will be particularly useful in the reporting of clinical trials and may also help clinicians in determining whether individual patients with AS have experienced an overall improvement in their symptoms.

### ACKNOWLEDGMENTS

The authors thank the drug companies who provided the databases that permitted the conduct of this study, and also thank the members of the standing committee of the ASAS Working Group for their input in this project and for their endorsement of it.

### REFERENCES

1. Van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. *J Rheumatol* 1997;24:2225-9.
2. Van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden SM, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy and clinical record keeping in AS: progress report of the ASAS Working Group. *J Rheumatol* 1999;26:951-4.
3. Benbouazza K, Auleley G-R, Collantes E, Hajjaj-Hassouni N, van der Heijde D, Dougados M. Evaluation of the smallest detectable difference (SDD) in symptomatic outcome variables in ankylosing spondylitis [abstract]. *Ann Rheum Dis* 2000;59 Suppl 1:56.
4. Dougados M, Gueguen A, Nakache J-P, Nguyen M, Mery C, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988;15:302-7.
5. Dougados M, Caporal R, Doury R, Thiese A, Pattin S, Laffez B, et al. A double blind crossover placebo controlled trial of ximoprofen in AS. *J Rheumatol* 1989;16:1167-9.
6. Dougados M, Nguyen M, Caporal R, Legeais J, Bouxin-Sauzet A, Pellegrini-Guegnault B, et al. Ximoprofen in ankylosing spondylitis: a double blind placebo controlled dose ranging study. *Scand J Rheumatol* 1994;23:243-8.
7. Dougados M, Gueguen A, Nakache J-P, Velicitat P, Veys EM, Zeidler H, et al. Ankylosing spondylitis: what is the optimum duration of a clinical study? One year versus a 6 weeks non-steroidal anti-inflammatory drug trial. *Rheumatology (Oxford)* 1999;38:235-44.
8. Dougados M, Béhier J-M, Jolchine I, Calin A, van der Heijde D, Olivieri I, et al. Efficacy of celecoxib, a cyclooxygenase 2-specific inhibitor, in the treatment of ankylosing spondylitis: a six-week controlled study with comparison against placebo and against a conventional nonsteroidal antiinflammatory drug. *Arthritis Rheum* 2001;44:180-5.
9. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995;38:727-35.
10. Van Gestel AM, Prevoo MLL, van 't Hof MA, van Rijswijk MH, van de Putte LBA, van Riel PLCM. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis: comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism criteria. *Arthritis Rheum* 1996;39:34-40.
11. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. *Arthritis Rheum* 1997;40:1202-9.
12. Dougados M, Le Claire P, van der Heijde D, Bloch D, Bellamy N, Altman R. Preliminary definition of clinical response for osteoarthritis trials in knee and hip: a preliminary report of the Osteoarthritis Research Society International Standing Committee for Clinical Trials response criteria initiative [abstract]. *Arthritis Rheum* 2000;43 Suppl 9:S220.
13. Dougados M, Gueguen A, Nakache J-P, Nguyen M, Amor B. Evaluation of a functional index for patients with ankylosing spondylitis. *J Rheumatol* 1990;17:1254-5.
14. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-10.
15. Spoorenberg A, van der Heijde DMFM, de Klerk E, Dougados M, de Vlam K, Mielants H, et al. A comparative study of the usefulness of the BASFI and the DFI in the assessment of AS. *J Rheumatol* 1999;26:961-5.
16. Van Gestel AM, Anderson JJ, van Riel PLCM, Boers M, Haagsma CJ, Rich B, et al. ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. *J Rheumatol* 1999;26:705-11.
17. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;23:137-45.
18. Ward MM, Kuzis S. Validity and sensitivity to change of spondylitis-specific measures of functional disability. *J Rheumatol* 1999;26:121-7.