EXTENDED REPORT

Comparison of statistically derived ASAS improvement criteria for ankylosing spondylitis with clinically relevant improvement according to an expert panel

A van Tubergen, D van der Heijde, J Anderson, R Landewé, M Dougados, J Braun, N Bellamy, G Udrea, Sj van der Linden, for the ASAS Working Group

Ann Rheum Dis 2003;**62**:215–221

Objective: To investigate whether the recently developed (statistically derived) "ASsessment in Ankylosing Spondylitis Working Group" improvement criteria (ASAS-IC) for ankylosing spondylitis (AS) reflect clinically relevant improvement according to the opinion of an expert panel.

Methods: The ASAS-IC consist of four domains: physical function, spinal pain, patient global assessment, and inflammation. Scores on these four domains of 55 patients with AS, who had participated in a non-steroidal anti-inflammatory drug efficacy trial, were presented to an international expert panel (consisting of patients with AS and members of the ASAS Working Group) in a three round Delphi exercise. The number of (non-)responders according to the ASAS-IC was compared with the final consensus of the experts. The most important domains in the opinion of the experts were identified, and also selected with discriminant analysis. A number of provisional criteria sets that best represented the consensus of the experts were defined. Using other datasets, these clinically derived criteria sets as well as the statistically derived ASAS-IC were then tested for discriminative properties and for agreement with the end of trial efficacy by patient and doctor.

Results: Forty experts completed the three Delphi rounds. The experts considered twice as many patients to be responders than the ASAS-IC (42 *v* 21). Overall agreement between experts and ASAS-IC was 62%. Spinal pain was considered the most important domain by most experts and was also selected as such by discriminant analysis. Provisional criteria sets with an agreement of ⩾80% compared with the consensus of the experts showed high placebo response rates (27–42%), in contrast with the ASAS-IC with a predefined placebo response rate of 25%. All criteria sets and the ASAS-IC discriminated well between active and placebo treatment (χ^2=36–45; p<0.001). Compared with the end of trial efficacy assessment, the provisional criteria sets showed an agreement of 71–82%, sensitivity of 67–83%, and specificity of 81–88%. The ASAS-IC showed an agreement of 70%, sensitivity of 62%, and specificity of 89%.

Conclusion: The ASAS-IC are strict in defining response, are highly specific, and consequently show lower sensitivity than the clinically derived criteria sets. However, those patients who are considered as responders by applying the ASAS-IC are acknowledged as such by the expert panel as well as by patients' and doctors' judgments, and are therefore likely to be true responders.

See end of article for authors' affiliations

Correspondence to:
Professor D van der Heijde, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands;
dhe@sint.azm.nl

Accepted 3 July 2002

In the near future many new treatments will be studied in patients with ankylosing spondylitis (AS). Besides demonstrating the mean efficacy of a treatment at a group level, it is also important to ascertain how many (and which) patients actually improve, because improvement of an entire group may be based on (small) improvements of many patients, as well as on a marked improvement of only a few patients with the majority showing no change.[1] Advantages of defining an individual response are that such changes are easier to understand; it may reduce the number of tests because usually several domains are combined; it facilitates comparison across different trials; it may help doctors to decide whether a patient has responded adequately to treatment; and it may also be used to assess factors that predict response.[1–3] To establish an individual response, criteria for defining such improvement' are essential.

In 1998, during the Outcome Measures in Rheumatoid Arthritis Clinical Trial (OMERACT) Conference, the members of the ASsessment in Ankylosing Spondylitis (ASAS) Working Group selected "core sets" of outcome measures to be used in different kinds of trials in AS, in order to create uniformity and allow comparison among these studies.[4] The ASAS Working Group is an international group of rheumatologists, epidemiologists, patients with AS, and representatives of the pharmaceutical industry from more than 20 countries, who share their expertise in the field of AS. More recently, a small group of members from the ASAS have developed a preliminary definition of short term improvement in AS, based on the core set of clinical trials with non-steroidal anti-inflammatory drugs (NSAIDs).[5] These ASAS improvement criteria (ASAS-IC) consist of four outcome domains: physical function (measured with the Bath AS Functional Index (BASFI)),[6] spinal pain (measured on a 100 mm visual analogue scale (VAS)), patient global assessment (100 mm VAS), and inflammation (mean of the last two questions from the Bath AS Disease Activity Index (BASDAI),[7] concerning the intensity and duration of morning stiffness). Scores on each domain range from 0 (best) to 100 (worst). The ASAS-IC define improvement as ⩾20% and ⩾10 VAS points improvement on the 0–100 scale in at least three of the four domains, with on the 4th domain no worsening of ⩾20% and ⩾10 VAS points.

Abbreviations: AS, ankylosing spondylitis; ASAS-IC, ASsessment in Ankylosing Spondylitis improvement criteria; NSAIDs, non-steroidal anti-inflammatory drugs; VAS, visual analogue scale

Table 1 Number of responders and non-responders according to the ASsessment in Ankylosing Spondylitis improvement criteria (ASAS-IC) for ankylosing spondylitis versus the consensus of the experts

	ASAS-IC		
Consensus experts	Responder	Non-responder	Total
Responder	21	21	42
Non-responder	0	13	13
Total	21	34	55

The composition and validation of the ASAS-IC were performed by using a statistical approach. It is, however, important to know to what degree these statistically derived improvement criteria reflect the opinions of clinicians and patients with AS.[8] Our study aimed at comparing the ASAS-IC with clinically relevant improvement according to an expert panel in the field of AS by means of a three round, consensus building, Delphi exercise.

METHODS

Cases

For this study the profiles of real patients with AS, who had participated in a randomised placebo controlled NSAID efficacy trial with a flare design, were used.[9] All patients fulfilled the modified New York criteria for AS.[10] In total 473 patients entered the study, of whom 363 patients completed the six week part of the trial. For each of the patients the absolute and relative change scores between the last measurement and baseline in the four domains of the ASAS-IC (physical function, spinal pain, patient global assessment, and inflammation) were calculated, independently of treatment allocation or completion of the trial. Cases appropriate for the present study were selected manually by comparing the outcomes with the ASAS-IC. In total 55 cases were selected: 21 patients fulfilled the ASAS-IC, 23 patients approximated the ASAS-IC, but showed either too little improvement or too much worsening in the absolute or relative change score in one or two domains, and 11 patients were clearly non-responders showing either obvious worsening (>30%) in one or two domains while the others improved, or showed no/minor change in any domain.

Expert panel and Delphi exercise

The opinions of experts in the field of AS were examined by a three round, consensus building, Delphi exercise.[11] The expert panel consisted of members of the ASAS Working Group extended with a group of patients with AS who are actively involved in AS research. All 57 candidate participants with an email address or fax number known to us, and who were not involved in the development of the ASAS-IC, were invited to participate in this study.

In the first Delphi round, all 57 candidates were sent the profiles of the 55 cases, randomly presented in an Excel data spreadsheet. For each of the cases the values in each domain at baseline and at the end of the trial, as well as the absolute and relative change scores were provided. The experts were requested to indicate for each case whether they personally considered this patient as a responder (yes or no). After completion of the file, they emailed or faxed back their answers. In a second round, the same cases were sent to each participant, showing his/her initial judgment for each of the cases, including the anonymous, aggregated responses of the group. The experts were then again asked to supply their judgments for the cases, but with the possibility of modifying the initial decisions based on the group's responses. A final, third round was similar to that of the second round, at which the experts were provided with their individual decisions and the group's responses of the second round. In this final round all experts were also asked by questionnaire to describe how they came to a decision for judgments of the cases: whether they looked at both absolute and relative improvement or only one of them; how much improvement they considered at least necessary to certify a patient as a responder, and in how many domains; whether they considered some domains more important than others, and which domains were the least important; and how much worsening they allowed at most, and in how many domains.

For each round the experts were offered two weeks to judge the cases. If necessary, a reminder was sent. The exercise was performed before publication of the ASAS-IC.

Statistical analysis

The analyses were performed from both the viewpoint of the ASAS-IC and the final consensus of the experts. Firstly, the number of patients who improved or did not improve according to the ASAS-IC was compared with the consensus of the experts in a 2×2 table. The percentage of agreement was calculated as the total number of corresponding responder and non-responder cases divided by the total number of cases.

Secondly, the experts' rationale on how they came to a decision for each of the cases was investigated, including the opinions on which domains had prevailed in the final decision. We also applied discriminant analysis to determine statistically which domains were considered to contribute most for each participant to both the absolute and the relative change scores. Dependent variable was the individual judgment of each participant for all 55 cases. Independent variables were either the absolute or the relative change scores in the four domains of the ASAS-IC. This exercise was done for all participants separately.

Thirdly, because the consensus of the experts clearly differed from the ASAS-IC, a number of provisional criteria sets were defined in order to find criteria sets that best represented the consensus of the experts. These criteria sets and the ASAS-IC were analysed for their properties to discriminate between active and placebo treatment in 2×2 tables with χ^2 tests. For this purpose the same data subset as for validation of the ASAS-IC was used.[5] In this validation set a random selection of three NSAID efficacy trials was taken (197 placebo, 408 actively treated).

Fourthly, the ASAS-IC and the consensus of the experts were compared with the end of trial efficacy assessment by the doctor and patient of the 55 selected cases, thereby considering the efficacy assessment as the "gold standard". Also, the ASAS-IC and the provisional criteria sets that best represented the consensus of the experts were compared with the end of trial efficacy assessment of all patients (n=473) of the NSAID efficacy trial from which the 55 cases were initially selected.[9]

RESULTS

In total 57 ASAS members and patients with AS were invited to participate in this three round Delphi exercise. Thirteen did not respond, and two considered that they did not have (enough) expertise to complete the exercises. Forty two

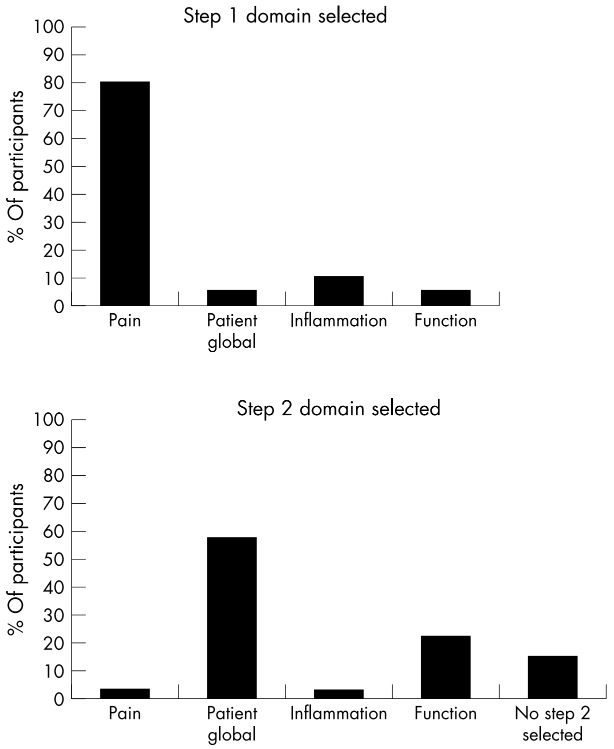

Figure 1 The percentage of participants in each domain selected as the first and second step in the discriminant analysis on the absolute change scores.

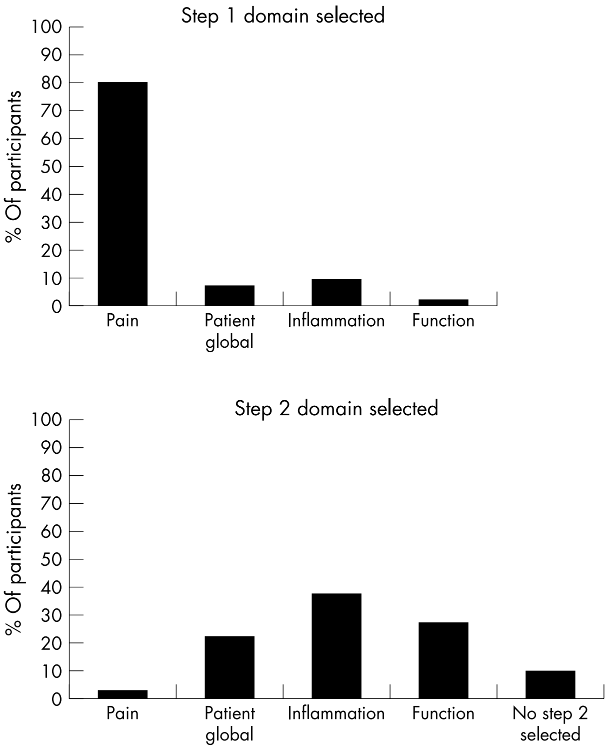

Figure 2 The percentage of participants in each domain selected as the first and second step in the discriminant analysis on the relative change scores.

participants returned the questionnaire in the first round, 41 in the second, and 40 (of whom four were patients with AS) in the last round. The reasons for the two non-responders during this exercise are unknown.

Agreement among the experts increased with each round for each of the cases. A median number of 4 (range 0–24) modifications were made in the second round by the experts, and a median number of 1 (range 0–7) in the third round. Agreement among the experts of ⩾70% was finally reached in all 55 cases, ⩾80% in 52/55, and ⩾90% in 44/55 cases. In this paper, the consensus of ⩾70% (including all cases) has been used for further analyses.

Agreement between ASAS improvement criteria and consensus of the expert panel

As already stated, of the 55 cases, by definition 21 (38%) fulfilled the ASAS-IC, and were therefore considered as responders (table 1). Thirty four (62%) cases did not fulfil the ASAS-IC, and were therefore considered as non-responders. The experts judged 42 (76%) cases to be responders and 13 (24%) to be non-responders. Sixteen of the 23 cases that approximated the ASAS-IC, and five of the 11 clearly non-responder cases (of which all showed large differences among the domains, with worsening up to 83% and 15 VAS points in one variable) were judged to be responders by the experts. The experts agreed on all 21 (100%) responder cases as defined by the ASAS-IC, and on 13 (38%) ASAS-IC non-responder cases. Improvement according to the ASAS-IC was acknowledged by the experts in all cases, but the proportion of patients considered to be responders by the experts was twice as high. Overall agreement between the ASAS-IC and the consensus of the experts was 62%.

Judgment of the cases by the experts and weight of domains

In the last Delphi round the experts were asked to describe how they came to a decision for each of the cases. The reasons for judging patients to be responders differed considerably among the experts: relative improvement of 10, 15, 20, 25% or

Table 2 Prevailing domain in discriminating between improvement and no improvement according to each participant versus discriminant analysis for each participant (n=38)

	Prevailing domain according to participants					
Discriminant analysis	Pain	Inflammation	Global	Function	All domains equal	Total
Pain	17	1	9	2	2	31
Inflammation	1	2	0	1	0	4
Global	1	0	1	0	0	2
Function	0	0	0	1	0	1
Total	19	3	10	4	2	38

Table 3 Improvement criteria sets showing ⩾80% agreement with the consensus of the experts and tested for their discriminative performances between placebo and actively treated patients

		Validation subset			
Improvement definition	% Agreement compared with consensus of the experts (n=55)	% Responders in placebo treated group (n=197)	% Responders in active treated group (n=408)	Difference	χ^2 Value
Relative improvement					
30% in 2 of 4	82	33	60	27	40
20% in 3 of 4	84	27	55	28	41
20% in 2 of 4	91	42	69	27	42
10% in 3 of 4	91	36	62	26	38
Combinations in relative improvement					
40% in 2 of 4 or 20% in 3 of 4	89	30	58	28	44
30% in 2 of 4 or 20% in 3 of 4	93	34	62	28	42
Relative and absolute improvement					
30% and 10 VAS in 2 of 4	80	33	60	27	39
20% and 10 VAS in 2 of 4	93	40	67	27	42
Combinations in relative and absolute improvement					
40% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4	87	29	58	29	44
30% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4	91	33	61	28	43
Relative improvement including pain					
30% in 2 of 4, including pain	82	29	56	27	38
20% in 3 of 4, including pain	80	26	53	27	41
20% in 2 of 4, including pain	93	37	64	27	40
Combinations in relative improvement including pain					
40% in 2 of 4 or 20% in 3 of 4, all including pain	87	27	56	29	43
30% in 2 of 4 or 20% in 3 of 4, all including pain	91	30	59	29	43
Relative and absolute improvement including pain					
30% and 10 VAS in 2 of 4, all including pain	80	29	56	27	39
20% and 10 VAS in 2 of 4, all including pain	93	35	63	28	42
Combinations in relative and absolute improvement including pain					
40% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4, all including pain	85	27	55	28	43
30% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4, all including pain	89	30	59	29	45
ASAS Improvement Criteria (ASAS-IC)					
20% and 10 VAS in 3 of 4, 4th no worsening of ⩾20% and 10 VAS	62	25	51	26	36

VAS, visual analogue scale; ASAS-IC, Assessment in Ankylosing Spondylitis improvement criteria.
All χ^2 values showed p values of <0.001

even more in ⩾2 domains, with or without absolute improvement of 10 or 20 VAS points in ⩾2 domains, and some experts required spinal pain to be part of the improving domains. Ninety per cent of the experts asserted that they looked at both absolute and relative improvement for their final judgment. Worsening was not included in most judgments. Spinal pain was considered as the most important domain by 50% of the experts, patient global assessment by 26%, physical function by 11%, inflammation by 8%, and 5% considered all domains to be equal. Inflammation was considered to be the least important domain by 43%, physical function by 30%, patient global assessment by 16%, spinal pain by 5%, and 5% considered all domains to be equal.

With discriminant analysis on both the absolute and the relative change scores the most important domains were selected for each of the 40 participants who had participated in the last round (figs 1 and 2, respectively). Spinal pain was selected as step 1 of the discriminant analysis for both absolute and relative change scores by 80% of the experts; thus this becomes the most important domain for making a decision on improvement. Patient global assessment was selected as the most important second step for the absolute change scores (fig 1), and inflammation for the relative change scores (fig 2). For the absolute change scores only one discriminant was selected in six participants, and for the relative change scores only one discriminant was selected in three participants.

Table 2 compares the opinions of the participants (n=38) regarding the prevailing domain with the results of the discriminant analysis. The results on the first step of the discriminant analysis for both the absolute and relative change scores were identical, except for one participant (who did not provide his personal judgment), and are therefore combined in table 2. It can be seen that in 31 participants, spinal pain was selected by discriminant analysis as the domain with the highest influence on decision making, whereas only 17 participants had stated that spinal pain was the most important domain. The remaining 14 participants assumed that their judgments were mainly based on other domains (in particular patient global assessment).

Provisional improvement criteria sets versus consensus of the expert panel

Because the consensus of the experts clearly differed from the ASAS-IC, a number of provisional criteria sets were defined in order to find criteria sets that best represented the consensus of the experts. Candidate criteria sets identical to those created for development of the ASAS-IC,[5] but also a number of provisional criteria sets adapted to the opinion of the experts, were tested for their agreement with the consensus of the experts. Because most experts did not consider worsening important in the judgments of the cases, a maximum in worsening was not included in most provisional criteria sets. These criteria sets always included relative improvement in at least two domains, with or without absolute improvement. Because spinal pain was selected by both the experts, as well as by discriminant analysis, to be the most important domain, a minimum improvement in this domain was regarded as a prerequisite in many provisional criteria sets.

In total 36 criteria sets were tested for their agreement with the consensus of the experts. Table 3 shows only those criteria sets with an agreement of ⩾80% compared with the consensus of the experts, based on all 55 cases. The highest agreement found was 93%. The criteria sets were thereafter applied to the validation subset of the ASAS-IC in order to analyse their properties to discriminate between active and

Table 4 Comparison of improvement criteria sets that best represented the consensus of the experts and ASAS-IC with the end of trial efficacy assessment by the doctor and the patient

Improvement definition	% Agreement compared with efficacy assessment (n=473)*	Sensitivity	Specificity
Relative improvement			
30% in 2 of 4	77	73	84
20% in 3 of 4	77	67	88
20% in 2 of 4	82	83	81
10% in 3 of 4	80	79	84
Combinations in relative improvement			
40% in 2 of 4 or 20% in 3 of 4	76	72	85
30% in 2 of 4 or 20% in 3 of 4	77	75	83
Relative and absolute improvement			
30% and 10 VAS in 2 of 4	76	72	85
20% and 10 VAS in 2 of 4	81	80	83
Combinations in relative and absolute improvement			
40% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4	75	71	86
30% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4	77	74	84
Relative improvement including pain			
30% in 2 of 4, including pain	76	72	86
20% in 3 of 4, including pain	73	66	88
20% in 2 of 4, including pain	81	80	84
Combinations in relative improvement including pain			
40% in 2 of 4 or 20% in 3 of 4, all including pain	76	71	88
30% in 2 of 4 or 20% in 3 of 4, all including pain	77	73	85
Relative and absolute improvement including pain			
30% and 10 VAS in 2 of 4, all including pain	75	71	86
20% and 10 VAS in 2 of 4, all including pain	80	78	84
Combinations in relative and absolute improvement including pain			
40% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4, all including pain	75	70	88
30% in 2 of 4 or 20% in 3 of 4 and 10 VAS in 2 of 4, all including pain	76	72	85
ASAS Improvement Criteria (ASAS-IC)			
20% and 10 VAS in 3 of 4, 4th no worsening of ≥20% and 10 VAS	70	62	89

VAS, visual analogue scale; ASAS-IC, Assessment in Ankylosing Spondylitis improvement criteria.
*Ten patients were omitted from this analysis, because of a disagreement between the patient and the doctor end of trial efficacy assessment. Another 27 patients were omitted from this analysis because of missing data on domains of the ASAS-IC.
Sensitivity (proportion of responders correctly classified) and specificity (proportion of non-responders correctly classified) were calculated considering the end of trial efficacy assessment as the "gold standard".

placebo treatment.[5] The percentage responders in the placebo treated group was large for many of the criteria sets (up to 42%). Ten of the criteria sets showed a placebo response rate of ≤30%. The ASAS-IC had by definition a placebo response rate of 25%.[5] All criteria sets and the ASAS-IC discriminated well between the placebo and active treated groups (χ^2=36–45; all p<0.001), and showed similar contrasts.

End of trial efficacy assessment

As a next stage, the ASAS-IC and the consensus of the experts were compared with the end of trial efficacy assessment by the doctor and patient for each of the 55 selected cases. Because disagreement between doctor and patient on the efficacy assessment was noted in only one of the 55 cases, the results were combined, and this case was omitted from the analyses. In total 44 patients were considered by themselves and by their doctors as responders, and 10 as non-responders. Agreement between the efficacy assessment and the consensus of the experts was 72%, sensitivity (proportion of responders correctly classified) was 80%, and specificity (proportion of non-responders correctly classified) was 40%. Agreement between the efficacy assessment and the ASAS-IC was 41%, sensitivity 36%, and specificity 60%.

Finally, the ASAS-IC and the provisional criteria sets that best represented the consensus of the experts were compared with the end of trial efficacy assessment of all patients (n=473) of the NSAID efficacy trial from which the 55 cases were initially selected (table 4).[9] In 10 of these 473 patients a disagreement between the doctor and the patient on the efficacy was observed. These patients were omitted from the analyses. Another 27 patients were also omitted from the analyses owing to missing data in one or more domains of the ASAS-IC. In total, 307 patients were considered by themselves and by their doctors to be responders (47 (43%) from the placebo group and 260 (80%) from the active group), and 129 patients to be non-responders (63 (57%) from the placebo group and 66 (20%) from the active group). Agreement between the efficacy assessment and the criteria sets ranged from 71% to 82%, sensitivity from 67% to 83%, and specificity from 81% to 88%. Agreement between the efficacy assessment and the ASAS-IC was 70%, sensitivity 62%, and specificity 89%.

DISCUSSION

In this study the ASAS-IC for AS recently developed by a statistical approach were compared with the opinion of an expert panel in the field of AS by means of a three round Delphi exercise, to assess the clinical relevance and applicability of the ASAS-IC. Advantages of using a Delphi technique to obtain the opinions of experts over, for instance, a consensus meeting are that participants are offered the opportunity to modify their initial judgments, based on the opinion of the group in following rounds, and that all opinions are collected anonymously.[11] Most modifications by the experts were made in the second round. In the final round only one (median) modification was made compared with the previous round, suggesting that three rounds are sufficient for achieving consensus.

According to the ASAS-IC, 21/55 patients were considered to be responders, whereas the experts judged 42 patients to be responders. Clearly, the experts personally used criteria for their judgments that were less strict than the ASAS-IC, but opinions among the experts varied considerably. Most experts considered improvement in only two domains sufficient to acknowledge a patient as a responder, and worsening was not included in most judgments. Spinal pain was regarded as the most discriminative domain. This was also reflected in the discriminant analysis, in which pain was selected as the first step by 80% of the experts. According to the experts' own opinion, pain was less often considered as the most important domain (in 50%). A number of participants defined other domains (particularly patient global assessment) as most discriminative, but apparently acted differently. This discrepancy between actual decision making and stated opinions on the importance of variables has been described previously by Kirwan *et al*.[12]

Because the consensus of the experts clearly differed from the ASAS-IC, a number of provisional criteria sets were defined in order to find criteria sets that best represented the consensus of the experts. Out of a total of 36 criteria sets, 19 sets showed an accuracy rate of ⩾80% with the consensus of the experts, and were therefore tested for their discriminative properties by using the same validation subset as was used for development of the ASAS-IC. All provisional criteria sets, as well as the ASAS-IC, discriminated well between active and placebo treatment. The placebo response rate was, however, relatively high (up to 42%) for the criteria clinically derived sets. For the development of the ASAS-IC, Anderson *et al* made the condition that the placebo response rate should not exceed 25%.[5] No placebo response rates below 25% were found for the provisional criteria sets, and only 10 criteria sets showed a placebo response rate of ⩽30%. The predefined maximum placebo response rate explains why the ASAS-IC behaved relatively worse than some of the provisional criteria sets for agreement and contrast between intervention and placebo. The ASAS-IC and the provisional criteria sets showed moderate to good agreement with the end of (NSAID) trial efficacy assessment by the patient and doctor. Five criteria sets showed an agreement of ⩾80%, and corresponding high sensitivity and specificity. The ASAS-IC showed lower agreement (70%), a relatively low sensitivity (62%), but the highest specificity (89%).

The ASAS-IC appeared to be strict in defining a patient as a responder, but those patients who were responders according to the ASAS-IC were also acknowledged as such by the expert panel. In addition, compared with the efficacy assessment by patient and doctor of the NSAID trial, the specificity of the ASAS-IC was high, implying that the number of patients incorrectly classified as responders by the ASAS-IC is very low (table 4). Thus, those patients who are classified as responders by the ASAS-IC are most likely to be true responders, because they are responders in the opinion of the experts, the patients themselves, and their treating doctors, independently of each other. This makes the ASAS-IC particularly valid for clinical trials, in which true responders are important to detect. Similar reasoning shows that applying the ASAS-IC in clinical practice is much more complicated; the ASAS-IC are not sensitive enough to pick up changes that are considered to be important in the opinion of the experts. In this respect the ASAS-IC can be compared with classification criteria (characterised by high specificity), and the consensus of the experts with diagnostic criteria (characterised by high sensitivity). In our opinion, the ASAS-IC should, however, not be substituted by any of the provisional criteria sets that best represented the consensus of the experts, because these criteria sets showed too high placebo response rates, which is an undesired effect in the assessment of potentially active treatment. We would, therefore, recommend application of the ASAS-IC particularly in clinical trials, in which the false positive rates should be kept low, and apply the ASAS-IC, with reservations, in daily practice, in which sensitivity is of greater importance.

The ASAS-IC have been validated in NSAID efficacy trials with "flare" as a selection criterion. Studies with flare based designs may, however, be associated with large placebo response rates.[13] It is therefore arguable whether the ASAS-IC would behave similarly in NSAID efficacy trials without this criterion. In addition, it is yet unknown how the ASAS-IC behave in other clinical trials assessing for instance the efficacy of disease modifying drugs, anti-tumour necrosis factor α treatment, or physical therapy. Also, the opinion of the experts on the degree of improvement and prevailing domains may vary for these different trials. It should be kept in mind that the opinion of experts on what they consider to be improvement may largely be determined by their clinical perception; if more effective treatment becomes available, then the opinion of the experts on what they call improvement may also change. In other words, experts' opinion, and thus criteria sets derived from this, is based on the current state of the art. In contrast, the ASAS-IC are statistically derived and consequently more "timeless" (that is, less influenced by current opinions). It may be expected that with the introduction of more effective treatment the experts will move more in the direction of the ASAS-IC. Future studies will be needed to validate the current ASAS-IC or to develop new improvement criteria for other kinds of trial.

In conclusion, the ASAS-IC are strict in defining patients as responders, but those patients classified as responders are acknowledged as such by the expert panel, by patients themselves, and by their treating doctors, independently of each other. Therefore, the opinion of the patients and the doctors is well reflected by the ASAS-IC.

ACKNOWLEDGEMENTS

We thank all the participating ASAS Working Group members and the patients with AS for their important contributions to this project, and Gabriel Baron for providing us with the validation dataset of the ASAS-IC.

A van Tubergen received the Novartis Rheumatology Grant 2001 from the Dutch Society for Rheumatology to perform this study.

Authors' affiliations

A van Tubergen, D van der Heijde, R Landewé, Sj van der Linden, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, Maastricht, The Netherlands
D van der Heijde, Limburg University Centre, Diepenbeek, Belgium
J Anderson, Department of Medicine, Clinical Epidemiology Research and Training Unit, Boston University Medical Center, Boston, USA
M Dougados, Department of Rheumatology, Hospital Cochin, Paris, France
J Braun, Department of Rheumatology, Rheumazentrum Ruhrgebiet, Herne, Germany
N Bellamy, Department of Medicine, Center of National Research on Disability and Rehabilitation Medicine CONROD, Brisbane, Australia
G Udrea, Department of Rheumatology, Dr Ioan Cantacuzino Hospital, Bucharest, Romania

REFERENCES

1 **van Riel PL**, van de Putte LB. DC-ART: what proportion of response constitutes a positive response? J Rheumatol Suppl 1994;41:54–5, discussion 56.
2 **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.
3 **Dougados M**, Leclaire P, van der Heijde D, Bloch DA, Bellamy N, Altman RD. Response criteria for clinical trials on osteoarthritis of the knee and hip: a report of the Osteoarthritis Research Society International Standing Committee for Clinical Trials response criteria initiative. Osteoarthritis Cartilage 2000;8:395–403.
4 **van der Heijde D**, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. J Rheumatol 1999;26:951–4.
5 **Anderson JJ**, Baron G, van der Heijde D, Felson DT, Dougados M. Ankylosing spondylitis assessment group preliminary definition of short-term improvement in ankylosing spondylitis. Arthritis Rheum 2001;44:1876–86.
6 **Calin A**, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, *et al.* A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994;21:2281–5.
7 **Garrett S**, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994;21:2286–91.

8 **Ward MM**. Response criteria and criteria for clinically important improvement: separate and equal? Arthritis Rheum 2001;44:1728–9.

9 **Dougados M**, Gueguen A, Nakache JP, Velicitat P, Veys EM, Zeidler H, *et al*. Ankylosing spondylitis: what is the optimum duration of a clinical study? A one year versus a 6 weeks non-steroidal anti-inflammatory drug trial. Rheumatology (Oxford) 1999;38:235–44.

10 **van der Linden S**, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361–8.

11 **Jones J**, Hunter D. Consensus methods for medical and health services research. BMJ 1995;311:376–80.

12 **Kirwan JR**, Chaput de Saintonge DM, Joyce CR, Currey HL. Clinical judgment in rheumatoid arthritis. II. Judging 'current disease activity' in clinical practice. Ann Rheum Dis 1983;42:648–51.

13 **Scott Lennox JA**, McLaughlin Miley C, Lennox RD, Bohlig AM, Cutler BL, Yan C, *et al*. Stratification of flare intensity identifies placebo responders in a treatment efficacy trial of patients with osteoarthritis. Arthritis Rheum 2001;44:1599–607.