

July 20, 2001

Dockets Management Branch (HFA-305)
Food and Drug Administration
5630 Fishers Lane
Room 1061
Rockville, Maryland 20852

2059 01 JUL 20 P3:39



RE: Docket No. 01N-0197
Request for Assistance: Clinical Development Programs for Drugs, Biological Products, and Devices for the Treatment of Ankylosing Spondylitis (AS) and Related Disorders

Merck & Co., Inc., is a leading worldwide, human health product company. Merck's corporate strategy -- to discover new medicines through breakthrough research -- encourages us to spend more than \$2 billion annually on worldwide Research and Development (R & D). Through a combination of the best science and state-of-the-art medicine, Merck's R & D pipeline has produced many of the important pharmaceutical products on the market today.

Merck Research Laboratories (MRL), Merck's research division, is one of the leading U.S. biomedical research organizations. MRL tests many compounds as potential drug candidates through comprehensive, state-of-the-art R & D programs. Merck supports regulatory oversight of product development that is based on sound scientific principles and good medical judgment.

In the course of bringing Merck product candidates through developmental testing and clinical trials, Merck scientists regularly address issues affected by this proposed Guidance. We have experience in developing drugs for the treatment of ankylosing spondylitis (AS) and related disorders and therefore, are well qualified to comment on, and interested in, the development of a Guidance in this therapeutic area. The issuance of an updated Guidance for clinical study of AS would be helpful. A large body of data has been collected on AS in the last 10 years, particularly in regard to outcome measures.

We have organized our submission to this Docket by introducing the topic of AS and providing responses to each of the five key areas mentioned in the *Federal Register* notice (Volume 66, No. 98, Page 27984, May 21, 2001). We attach referenced journal articles for your convenience.

Introduction

Ankylosing Spondylitis (AS) is a progressive inflammatory disorder characterized by severe disabling arthritis of the axial skeleton. AS belongs to a broader group of inflammatory arthritides known as the spondylarthropathies. The spondylarthropathies, which include AS, psoriatic arthritis, enteric arthropathy (inflammatory bowel disease [IBD]-associated arthritis), and reactive arthritis (Reiter's syndrome), share a common theme, involvement of the axial skeleton with or without peripheral arthritis and/or enthesopathy. Of these, AS is the "purest" of these disorders, primarily involving the

01N-0197

C2

axial skeleton with additional involvement of peripheral arthritis in a small subset of patients. Patients may have extraarticular manifestations such as iritis or aortitis, but in most patients, spinal disease (spondylitis and/or sacroiliitis) is the principle manifestation of disease.

AS usually presents with arthritis of the sacroiliac (SI) joints with progressive involvement of the lumbar, thoracic, and eventually, cervical spine over time. Arthritis of the axial skeleton with resultant back (spinal) pain is the principle manifestation of disease. Other common manifestations include night pain, stiffness (particularly in the morning), decreased mobility, limited ability to carryout activities of daily living, and fatigue.

Among the spondylarthropathies, AS has been one of the best studied, and substantial progress has been made in terms of understanding both the clinical manifestations of this disease and outcome measures with which to follow them. This knowledge should form the basis for any regulatory Guidance developed by the FDA and international authorities.

Key Areas of Discussion

1. Scope: Should the guidance discuss AS alone, or a broader spondyloarthropathy rubric?

The Guidance should discuss AS alone. The manifestations of disease vary significantly between AS and most of the other spondylarthropathies and, thus, the criteria used to evaluate these diseases should be different. Compared to other spondylarthropathies, patients with AS represent a fairly homogeneous population in which to assess effect of treatment on outcome measures. A sufficiently large AS population (with a prevalence of approximately 0.1-0.2%) exists to merit a separate Guidance. Separate Guidances should be developed for distinct indications within the broader spondylarthropathy rubric.

AS, compared to the other spondylarthropathies, is primarily a disease of the spine. It has a slow, progressive, but fairly predictable course resulting in ankylosis and decreased mobility of the axial skeleton over time. While by definition essentially 100% of patients with AS have sacroiliitis and/or other spinal manifestations, only a minority (<50%) of patients with other spondylarthropathies have substantial spinal disease. When present, the spinal disease in the other spondylarthropathies often differs from that seen in AS. In AS, spondylitis (as detected by radiographic changes) usually involves contiguous regions of the lumbar, thoracic, and/or cervical spine. In psoriatic arthritis, spondylitis is often characterized by skip lesions, or non-contiguous involvement, of different regions of the spine. In most other spondylarthropathies, spondylitis is usually not seen and, when present, tends to be less severe than that observed for AS. The sacroiliitis seen in these conditions can also vary. AS is usually characterized by sacroiliitis with a symmetric pattern and in psoriatic arthritis and Reiter's syndrome, sacroiliitis, when present, is usually asymmetric.

AS also differs from other spondylarthropathies in the extent and pattern of peripheral arthritis. Only about 10-20% of patients with AS have peripheral arthritis, and peripheral arthritis is rarely the primary manifestation of disease. In contrast, the majority of patients with other spondylarthropathies have peripheral arthritis, and, in most cases, the peripheral arthritis is more prominent than the axial skeletal manifestations. Compared to AS, the pattern of peripheral arthritis in these other spondylarthropathies can also be quite varied. In AS, the typical pattern is large joint oligoarticular arthritis. In other spondylarthropathies, the arthritis may affect small joints or large joints, and have either a symmetric or asymmetric pattern depending upon the spondylarthritic variant and the individual patient.

Lastly, the other spondylarthropathies are more likely to have extra-articular manifestations as a primary manifestation of disease than AS. Many patients with psoriatic arthritis will present with psoriasis as their primary disease and have joint symptoms as a secondary manifestation. Likewise, many patients with enteric arthritis will have bowel symptoms as their primary manifestation of disease. In contrast, the primary symptoms of AS are localized to the axial skeleton and the majority are secondary to back disease. Extra-skeletal manifestations, while they do occur in AS, are rarely the primary manifestations of disease.

Scope: What about the clinical subgroups and pediatric expressions of the disorder(s)?

Pediatric variants should not be included in the Guidance.

AS is primarily a disease first experienced by young adults with a typical onset in late adolescence or early adulthood. In general, pediatric onset (age less than <15) is far less common than the typical adult disease, and, because it occurs infrequently, the clinical manifestations of pediatric AS have not been well-studied. The gathering of clinical data, particularly in regard to the validation of response measures in a pediatric subgroup, would be of great value, but given the paucity of clinical data in pediatric AS, the development of meaningful guidelines for improvement criteria is not possible at this time. The eventual development of such a Guidance would be of great value; the medical community should be encouraged to study these patients such that a database can be established from which a Guidance may be developed.

2. Claims: What type of claims structure is optimal to encompass the types of clinical benefit a therapeutic product might have on patients with AS? What type of evidence would be needed to support each proposed claim?

AS, like rheumatoid arthritis (RA), is a systemic, inflammatory, autoimmune disease leading to pain, discomfort, and functional disabilities. Over time, inflammatory changes associated with disease lead to joint erosion, reactive bone growth, and ankylosis (joint fusion). Similar to RA, it is expected that treatments for AS could be categorized as either symptom- or disease-modifying. At present, the most common treatment for AS is with nonsteroidal anti-inflammatory drugs (NSAIDs). NSAIDs have the proven benefit of symptom-modifying anti-rheumatic drugs (SMARDs) in improving many of the

clinical symptoms of AS. In contrast, no therapeutic agent to date has clearly demonstrated the ability to modify the long-term functional consequences of this disease, particularly in regard to the axial skeletal structural changes associated with long-term progression. Thus, the benefit of disease-modifying anti-rheumatic drugs (DMARDs) in the treatment of AS has yet to be proven.

Any new Guidance on AS should, at a minimum, distinguish between SMARDs and DMARDs with a specific requirement that the latter claim require an inhibition of the radiographic changes that characterize the long-term structural changes associated with this disease. A parallel may be drawn to rheumatoid arthritis in which SMARDs reduce the "signs and symptoms," and DMARDs reduce both the radiographic changes and the long-term functional consequences of disease.

Criteria to support an improvement (or reduction) in signs and symptoms should be based on endpoints that correlate well with disease activity, are believed by patients and/or clinical experts to be clinically important, and have the discriminant capacity to identify patients on clinically meaningful therapy. Guidelines for the study of symptom-modifying drugs should include a wide variety of clinical domains believed to be important in this disease (such as pain, physical function, inflammation, patient and physician global assessments, and mobility), but only well-validated endpoints with a high discriminant capacity should be required to support a claim. In NSAID trials (the only therapeutic agents approved for the treatment of AS in the United States), endpoints fulfilling these criteria include: patient and physician global assessments of disease activity, assessments of spinal pain, assessments of night pain, and assessments of physical function. Both the Bath AS Functional Index and the Dougados AS Functional Index have been validated and demonstrate a high discriminant capacity in this regard (Calin, A., Nakache, J-P., Gueguen, A., Zeidler, H., Mielants, H., Dougados, M. *Journal of Rheumatology*, 26, 975-979, 1999)(see Attachment 2). Of these, spinal pain is the predominant and most common disabling symptom and thus, is probably the best single measure of effect for a symptom-modifying agent.

A distinguishing characteristic for a DMARD claim should be that, in addition to reducing the signs and symptoms of disease, a DMARD prevents the structural damage characteristic of AS. The Guidance for Industry, *Clinical Development Programs for Drugs, Devices, and Biological Products for the Treatment of Rheumatoid Arthritis (RA)*, provides this distinction by having separate but non-exclusive claims for reduction of signs and symptoms and prevention of structural damage. We believe it unlikely that any one agent is likely to completely eliminate established radiographic changes associated with AS. Therefore, we believe it more reasonable to have inhibition (rather than prevention [cessation]) of further structural change as a claim, and that a criteria be devised in order to help determine the level of inhibition required to be clinically meaningful.

Various radiographic criteria have been devised to evaluate skeletal changes associated with AS. These criteria include the New York criteria to assess the SI joints; the Bath AS Radiology Index (BASRI) and the Stoke AS Spine Score (SASSS) to assess radiographic changes of AS in the spine; and the Larson and BASRI-hips to assess radiographic

changes of AS in the hips (Spoorenberg, A., de Vlam, K., van der Heijde, D., de Klerk, E., Dougados, M., Mielants, H., van der Tempel, H., Boers, M., van der Linden, S. *Journal of Rheumatology*, 26, 997-1002, 1999)(see Attachment 3). While some of these methods, particularly the SASSS and BASRI, have demonstrated good intraobserver reliability, none of these methods in present form was able to identify changes in AS patients during a one year period. Therefore, it is likely that these assessment tools will either need to be improved, or longer (multi-year) observation periods will be required to demonstrate a lack of radiographic progression with clinical intervention.

3. Measures of Disease Activity: Are currently available instruments for measuring disease activity adequate or are new measures required?

Measurements of disease activity have been best studied in NSAID trials, and a wealth of knowledge exists as to the validity and discriminant capacity of clinical endpoints in this setting. Useful measures currently exist to collect information on many of the domains associated with AS including: pain, global assessment, physical function, inflammation, and mobility. These response measures are discussed above in our response to Question 2.

Specific tools utilized to study these endpoints include both Likert and VAS scales for pain (back pain or low back pain) and global assessments of disease activity; both the BASFI and the DFI for physical function; measurements of both night pain (VAS or Likert) and morning stiffness (both duration and intensity) for inflammation; and measurements such as the Schober's test (or modified Schober's test), the occiput-to-wall test, and measurements of chest expansion for mobility. Each of these has been examined for its discriminant capacity (Calin, A., Nakache, J-P., Gueguen, A., Zeidler, H., Mielants, H., Dougados, M. *Journal of Rheumatology*, 26, 975-979, 1999)(see Attachment 2); in terms of the ability to distinguish from active therapy from placebo, the patient global assessment of disease activity provides the most power, followed by lumbar pain, night pain, physician global assessment, and the two functional indices. Measurements of mobility performed poorly and likely reflect the irreversible skeletal changes that are responsible for the decreased range of motion. Morning stiffness also performed poorly.

Other outcome measures not performing well include laboratory assessments of inflammation. The best studied of these include the erythrocyte sedimentation rate (ESR) and C-reactive Protein (CRP), useful markers of inflammation in other diseases, such as RA. Unfortunately, neither of these laboratory values correlated well with disease activity in AS. In fact, they appear to correlate more with the presence or absence of peripheral arthritis, rather than with disease activity (Spoorenberg, A., van der Heijde, D., de Klerk, E., Dougados, M., de Vlam, K., Mielants, H., van der Tempel, H., van der Linden, S. *Journal of Rheumatology*, 26, 980-984, 1999)(see Attachment 4). It may be that other potential biomarkers of disease activity, such as serum cytokine levels, may correlate better with disease activity; this is an attractive area for further exploration.

Clearly, well-validated endpoints for AS already exist. They cover a wide range of clinical domains. Despite this substantial knowledge on outcome measures in AS, there

is still room for improvement. In particular, there are some aspects of disease that have not been well-studied and could serve as additional endpoints in trials if the proper well-validated tools existed, such as fatigue and health-related quality of life. Detection of radiographic changes, for which criteria to assess changes already exist, might benefit from additional assessment measures (either observational or analytical), particularly a more sensitive assessment to measure change over time. Finally, there are other areas, such as development of composite response criteria (e.g. the ASAS criteria), where an initial version of an outcome measurement has been developed, but it has yet to be fully validated.

Measures of Disease Activity: Which disease activity should be measured in clinical trials in AS, and on what basis: (1) A consensus approach, which aims for agreement (clinicians, patients, and others) based on a blend of an observer-driven approach and performance characteristics; (2) a decision based on the comparative statistical characteristics of each measurement using concepts such as random measurement error; or (3) a fully data-driven approach where each measurement is tested in a standard venue to assess its predictive capacity.

Acceptance of a new clinical endpoint should require development, pilot testing, and validation. During development, a consensus approach should be used for item selection (physician and/or patient-based evaluation of the relevance of the endpoint to the disease process and whether, as applied, the endpoint gathers clinically significant information on the disease process). Draft instruments should be pilot-tested with AS patients and refined for ease of administration and completion. Validation should include testing in clinical settings using scenarios similar to the studies in which they will eventually be used. In order to be accepted, endpoints should correlate with other well-validated measures of disease activity and should have relatively high discriminant capacity (i.e. they should be able to distinguish those patients on active treatment from those patients on placebo).

The evidence-based validation processes are clearly most valuable before an endpoint can be accepted. For example, the ASAS response criteria, which has recently been devised for assessing treatment responses for AS (Anderson, J.L., Baron, G., van der Heijde, D., *et al. Arthritis and Rheumatism*, 43, S102, 2000)(see Attachment 1), would benefit from just such an analysis. This response criteria was derived based upon specificity and sensitivity analysis of previous NSAID and COX-2 inhibitor trials in AS and has the potential to become a valuable addition to the already validated endpoints, but it has not yet been prospectively validated. Prospective analysis in a clinical trial could provide just such a validation process. The response criteria could then be evaluated for correlation with other clinically important measures and to distinguish patients on different therapies (i.e. investigational drug versus placebo).

The response criteria would then need to be assessed for their clinical meaningfulness. A consensus approach is useful in this regard; do the clinical variables examined and the percent improvement required represent a clinically meaningful improvement in the opinion of clinicians and patients? The later process may occur in the setting of a formal consensus group such as the OMERACT organization, at an FDA advisory meeting, or in

a clinical trial setting in which outcome measures are discussed with clinical consultants, investigators, and regulatory agencies. This validation process is invaluable and insures that clinically meaningful results can be obtained.

4. Overall Trial Design: Are longitudinal comparison of means optimal? Because longer trials inevitably have substantial dropouts, would a survival analysis be more appropriate?

Both longitudinal comparison of means and survival analysis are of value in assessing the utility of a new drug for the treatment of AS. However, longitudinal comparison of means is generally more valuable as it allows a direct analyses of specific endpoints relevant to disease activity. Survival analysis (time-to-discontinuation) provides important supportive information, but as there may be many reasons for discontinuations from a trial, some of which may not be relevant to either the efficacy or safety of a new drug. Therefore, we recommend survival analysis as a secondary, not a primary, endpoint.

5. Intrinsic Trial Design: Which measures should be included in the primary analysis of the clinical trial to assess whether the therapeutic product is associated with a clinical benefit? Do all measures need equal-weight in the primary analysis? Can they be unequally weighted? Is the use of composites justified? Are outcomes of secondary endpoints essential for determining the success of the trial?

The primary analysis should be based on outcome variables that fairly represent disease activity, correlate to symptoms that are important to patients, have been well-validated in the past, and have a high discriminant capacity. The choice of which outcome measures to include in the primary analysis depends on the claim(s) being sought.

For a reduction in the signs and symptoms of AS, clearly the most important outcome variable is back (spinal) pain. AS is an inflammatory joint disease primarily affecting the axial skeleton. The predominate and most common disabling symptom is back pain, thought to be a direct result of the inflammatory process. Thus, the best judge of efficacy for a symptom-modifying agent is the significant reduction in spinal pain. Other helpful supporting information to establish the efficacy of a symptom-modifying drug include global assessments of disease activity and assessments of physical function.

All three of these endpoints have been examined in previous clinical trials, have a high discriminative capacity, and are well-validated for the evaluation of SMARD (NSAID) treatment responses in AS. Clearly, spinal pain is the primary symptom of disease and should be weighted appropriately, while the other two endpoints provide important supportive information. Assessment of spinal pain should be a primary endpoint; at a minimum, global assessments of disease activity and physical function should be considered key secondary endpoints.

The recently developed ASAS improvement criteria takes the approach that an improvement in 3 out of 4 response measures, including spinal pain, patient global assessment, the BASFI, and morning stiffness, should occur in order to consider a patient

as a clinical responder. This system weighs all four of these endpoints as important but allows for the possibility that not all patients will demonstrate improvement in each endpoint. Based on the studies from which the ASAS improvement criteria was derived, this system works well but has yet to be prospectively validated. Interestingly, one of the measures chosen for the ASAS improvement criteria, morning stiffness, has been shown in previous studies to be of low discriminant capacity (Calin, A., Nakache, J-P., Gueguen, A., Zeidler, H., Mielants, H., Dougados, M. *Journal of Rheumatology*, 26, 975-979, 1999)(see Attachment 2). Thus, it would be worthwhile to test the ASAS criteria in the presence and absence of this response measure (i.e. does the inclusion of morning stiffness add to the sensitivity and specificity or does it just broaden the analysis to include an additional domain?).

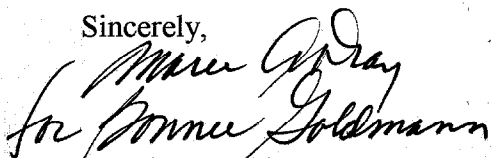
For trials seeking to claim a reduction in skeletal changes, the primary analysis should be based on the results of radiographic criteria (the best objective evidence that skeletal disease has not progressed). Other endpoints to be considered include measures of mobility such as the Schober's (or modified Schober's), occiput-to-wall test, and chest expansion test. Because decreases in mobility generally occur slowly (over a period of years), it may be difficult to show a significant reduction in loss of mobility over a shorter period. For this reason, mobility is likely to be of low discriminant capacity (even with a highly therapeutic drug), and it is probably unrealistic to require a significant reduction compared to placebo in a DMARD trial.

Conclusion

Guidelines are most helpful where they are relevant to the patient population. Spondylarthropathies are too diverse to cover in a single guideline. Therefore, we recommend that a Guidance focuses on the study population of interest (AS), and that endpoints concentrate on those manifestations of disease which are most typical of the disorder and correlate with disease activity. Well-validated measures (with the discriminant capacity to distinguish between placebo and active treatment groups), include patient and physician global assessments of disease activity, assessments of spinal pain, assessments of night pain, and assessments of physical function. Each of these measures has been used in multiple clinical trials and has proven value. The primary analysis of AS should concentrate on outcome variables that are clinically relevant. These include pain for a SMARD and inhibition of structural changes for a DMARD. We encourage the development of composite endpoints but believe they should be well-tested and prospectively validated before inclusion in a clinical Guidance.

We appreciate the opportunity to comment on this topic and welcome the opportunity to meet with you to discuss these issues.

Sincerely,



Bonnie J. Goldmann, M.D.

Vice President, Regulatory Affairs-Domestic

ABSTRACT SUPPLEMENT

PHIL  DELPHIA
2000

American College of Rheumatology
64th Annual Scientific Meeting

Association of Rheumatology Health Professionals
35th Annual Scientific Meeting

October 29 – November 2, 2000
Philadelphia, Pennsylvania
Pre-meeting courses begin October 28



Copyright © 2000 by the AMERICAN COLLEGE OF RHEUMATOLOGY, Atlanta, Georgia
The printing of the 2000 ACR Annual Scientific Meeting Abstract Supplement was supported
by an unrestricted educational grant from Wyeth-Ayerst Laboratories.



AMERICAN COLLEGE OF RHEUMATOLOGY
64th Annual Scientific Meeting



ASSOCIATION OF RHEUMATOLOGY HEALTH PROFESSIONALS
35th Annual Scientific Meeting

October 29 – November 2, 2000
Pre-meeting courses begin October 28

Philadelphia Convention Center
Philadelphia, Pennsylvania

ACR/ARHP PROGRAM OBJECTIVES

- To provide an in-depth presentation of the recent advances in the diagnosis, management and treatment of rheumatic diseases
- To provide a forum for exchange of new research data by scientists/investigators working in the area of rheumatic diseases
- To provide an opportunity to interact with experts in small group sessions
- To provide an opportunity to update knowledge concerning available pharmaceutical products and medical and assistive devices for use in the management of rheumatic diseases

The American College of Rheumatology is an independent professional, medical and scientific society. The ACR does not guarantee, warrant or endorse any commercial products or services.

This program is being sponsored by the American College of Rheumatology for educational purposes only. The material presented is not intended to represent the only or the best methods appropriate for the medical situations discussed, but rather is intended to present the opinions of the authors or presenters that may be helpful to other practitioners. Attendees participating in this medical education program sponsored by the American College of Rheumatology do so with full knowledge that they waive any claim they may have against the ACR for reliance on any information presented during these educational activities.

The 2000 ACR/ARHP Annual Scientific Meeting programs have been independently planned by the ACR Committee on Continuing Medical Education, ACR Annual Meeting Planning Committee, ARHP Program Committee, ARHP Clinical Focus Task Force and ARHP Rheumatology Practice Course Task Force.

Please be aware that the information and materials displayed and/or presented at all sessions of this meeting are the property of the American College of Rheumatology (and/or the presenter) and cannot be photographed, copied, photocopied, transformed to electronic format, reproduced or distributed without the written permission of the American College of Rheumatology (and/or the presenter).

Use of the ACR name in any fashion by any commercial entity for any purpose is expressly prohibited without the express written permission of the American College of Rheumatology.

CYTOKINE PROFILE IN PSORIASIS ARTHRITIS (PSA): PREDOMINANCE OF TH1-DERIVED PRO-INFLAMMATORY CYTOKINES. Raquel S Cuchacovich, Shankar Japa, Hernan Añis, Maria Cuellar, Cecilia Trejo, Abdul Aziz, Rodica S van Solingen, Luis R Espinoza New Orleans, LA and Santiago, Chile

The etiopathogenesis of PSA is not well understood, but accumulated evidence indicates that T cells and their products may play a major role. Conflicting data, however, has been published regarding the relative contribution of each Th1/Th2 subset.

Objective: To determine patterns of cytokine mRNA expression in peripheral blood mononuclear cells (PBMC), synovial tissue (ST) and skin from patients with PSA, as well as to determine their differential expression according to disease activity.

Material and Methods: Patients with PSA were recruited from our rheumatology clinics; all fulfilled Moll and Wright's criteria for PSA. Synovial tissue (ST), skin (S) and PBMC were obtained the same day, and frozen at -70 °C, until were tested. Spontaneous cytokine mRNA expression of interferon- γ (IFN- γ), Tumor necrosis factor- β (TNF- β) and IL-4 of ST, S and PBMC were determined using a semiquantitative RT-PCR. Quantification of cytokines after RT-PCR was carried out by comparing the intensity of each band to Glyceraldehyde-3-phosphate-dehydrogenase mRNA expression.

Results: Twenty patients with PSA (9 females and 11 males), mean age 38 \pm SD 11.7 years were studied. Active psoriatic skin and/or joint involvement was deemed to be present in 16 (80%) patients, and absent in 4 (20%). Duration of disease both skin and joint, did not differ significantly between both groups. IFN- γ mRNA and TNF- β mRNA were detected in 10/16 active (63%) and in 3/4 inactive (75%) PSA; and the levels of expression were not significant different between the two groups ($p > 0.05$). The expression, however, of IL-4 mRNA was not detected in the active group, but 1/4 patients with inactive PSA expressed high levels of IL-4 (26.5ug). The expression levels of IFN- γ , TNF- β and IL-4 mRNA did not differ significantly among PSA -PBMC, ST and S. Negligible levels expression, however, are seen in normal tissue.

Conclusion: Disregulation of the Th1/Th2 cytokine profile, with a predominance of Th1/Th2 pattern is present in active PSA.

Disclosure:

ONE YEAR OUTCOME OF PATIENTS WITH SEVERE PSORIASIS ARTHRITIS TREATED WITH INFILIXIMAB. Claudia Dechane, Christian Anconi, Joerg Wendler, Alexandra U Ogilvie, Matthias Lueck, Hans-Martin Lorenz, Joachim R Kalden, Bernhard Manger Erlangen, Germany

The anti-TNF-alpha antibody infliximab proved to be highly effective in treatment of rheumatoid arthritis (RA). For psoriatic arthritis it is known, that TNF-alpha is elevated in the synovial fluid and skin lesions. Therefore we wondered whether an anti-TNF-therapy could be similarly successful in the treatment of psoriatic arthritis.

We treated 10 patients with severe psoriatic arthritis with infliximab. All patients had a polyarticular disease with clinical and serological activity. 7 patients had concomitant MTX therapy, 1 sulfasalazine, and 2 had no DMARD. Patients received 5 mg/kg infliximab on week 0, 2, and 6. At week 10 all patients showed a dramatic response to infliximab treatment with reduction of signs and symptoms and serological activity (Arthritis Rheum 1999; 42 p 371). Thereafter infliximab treatment was adapted to the individual needs of the patients. Patients were followed for up to one year by consulting the ACR-criteria (for RA).

RESULTS: 1 patient with ACR 70 response at week 10 stopped infliximab therapy for personal reasons and his arthritis stayed in remission for 5 months until he got mild activity again. Of the remaining 9 patients, 5 patients of which with ACR 70 response and 1 with ACR 50 response at week 10 were further treated with infliximab, partly with a lower dose of 3.4 mg/kg and an infusion interval of 7-8 weeks. The infliximab therapy was discontinued in 3 cases after 5, 7, and 8 months because of remission and in 1 case after 8 months because of an infusion reaction and newly detected pregnancy. At follow up at year one after start of infliximab therapy all of these 5 patients still had an ACR 70 response. The pregnancy is uncomplicated so far. The other 4 patients received ongoing infusions with infliximab at a reduced dose of 3.4 mg/kg and enlarged infusion intervals of 8-8 weeks. 3 of these patients with an ACR 70 response at week 10 had an ACR 50 response at the one year evaluation. One patient with an ACR 50 response at week 10 developed an arthritis flare after 9 months after start of infliximab. By increasing the dose of infliximab and shortening the infusion intervals to 4 weeks a moderate response of ACR 50 could be achieved again. 6 of the 9 patients had no tender or swollen joints after one year.

CONCLUSIONS: These data show that infliximab was effective over one year. Therefore infliximab seems to be effective in the treatment of severe psoriatic arthritis as well.

Disclosure:

CROHN'S DISEASE WITH SPONDYLOARTHROPATHY: EFFECT OF TUMOR NECROSIS FACTOR α BLOCKADE WITH INFILIXIMAB ON THE ARTICULAR SYMPTOMS. F Van den Bosch, F Knuthof, M De Vos, P De Keyser, E M Veys, H Mielants Belgium

Introduction: The sustained beneficial effects of TNF α blockade in Crohn's disease (CD) have been confirmed in several randomised controlled trials. However the effects on articular manifestations were not assessed in these studies, although there exists an intriguing relationship between CD and spondyloarthropathy (SpA); articular symptoms compatible with SpA are observed in 33% of CD, and 25-27% of the SpA patients have microscopic gut inflammation which can evolve to full-blown CD. Therefore we evaluated the effect of TNF α blockade on articular symptoms in patients treated for CD.

Methods: Four patients were treated with infliximab (5 mg/kg) in a compassionate use program for refractory CD. All patients had therapy-resistant CD with or without fistulae. Moreover, all 4 patients had associated SpA, according to ESSG-criteria. Inflammatory parameters, gastrointestinal symptoms as well as axial and peripheral articular symptoms were assessed.

Results: In all 4 patients inflammatory parameters dropped and gastrointestinal remission was obtained. Two of the patients had peripheral synovitis that disappeared completely 2 weeks after treatment. They remained in articular remission for 3 months, and relapse of the synovitis was again successfully controlled by re-treatment with infliximab. In the third patient, the swollen joint count dropped from 12 to 6 after one dose of infliximab, and complete articular remission was obtained after a second dosing at week 2; this remission was sustained for up to 6 months. The fourth patient had HLA-B27 positive ankylosing spondylitis with severe inflammatory back pain. The axial night pain disappeared after treatment and the patient remained in remission during follow-up, which has now reached 6 months.

Conclusion: TNF α blockade is followed by a fast and significant improvement of articular as well as axial inflammation in SpA associated with CD. These observations warrant further investigation of the therapeutic potential of TNF α blockade in SpA.

Disclosure:

T CELL COUNTS AND TNF α SECRETION CAPACITY IN PATIENTS WITH ANKYLOSING SPONDYLITIS (AS) AFTER ANTI-TNF THERAPY. Jan Brandt, Hardy Maetzl, Andreas Thiel, Joachim Sieper, Jurgen Braun Berlin, Germany

Background: TNF α seems to be an important effector and regulatory cytokine. Anti-TNF α therapy has been successfully applied to patients with AS and other chronic inflammatory rheumatic and bowel diseases. The immunologic consequences of such treatment have not been completely elucidated. We have reported that the TNF α secretion capacity is rather reduced in AS patients (Rudwaleit et al. Ann Rheum Dis, in press).

Aims: To investigate changes of T cell counts and TNF α secretion capacity after treatment with anti-TNF α (ca2, infliximab) in AS patients.

Patients: All patients fulfilled the 1984 New York criteria for AS, the disease duration was < 10 years. Patients' characteristics have been described in more detail in a recent publication (A&R, June 2000).

Methods: Blood counts were performed before, and 1 and 2 weeks after therapy. PBMNCs of 6 AS patients were investigated before, 1 and 2 weeks after single infusion of 5mg/kg ca2. All patients had experienced significant clinical improvement. Cells were stimulated with PMA/Ionomycin for 6h, Brefeldin A was added and cells were fixed. FACS-analysis was performed by staining for surface CD3 and intracellular TNF α .

Results: The median total lymphocyte count (/ml) and the subset of CD3 positive T cells (%) decreased slightly from 1.4 and 2.2 at baseline to 0.8 and 2.0 at week 2 after ca2 treatment. The (median) percentage of CD3+ T/TNF α -producers increased from 5.6% at baseline to 9.9% at week 2. This was similar for IFN γ .

Conclusion: These results indicate a slightly decreased T cell count one week after infliximab and an increase of TNF α and IFN γ secretion capacity after therapy with anti-TNF α -antibodies in AS patients. These data suggest a systemic suppression of cytokine secretion which can be released by effective anti-cytokine therapy.

Disclosure:

ASAS PRELIMINARY DEFINITION OF SHORT-TERM IMPROVEMENT IN ANKYLOSING SPONDYLITIS. Jennifer J Anderson, Gabriel Baron, Desiric Van der Heijde, David T Felson, Maxime Dougados Boston, MA; The Netherlands and Paris, France

The Assessment in Ankylosing Spondylitis working group (ASAS) has established 5 domains for ankylosing spondylitis (AS) treatment: physical function (PF), pain (PA), spinal mobility (SM), patient global (PG), and inflammation (IN). We now develop symptomatic improvement criteria using data from 3 short-term placebo (PL) controlled clinical trials of NSAIDs (RX) in AS, with outcome measures in all 5 domains.

Trials (56 weeks) included 923 patients, 650 RX, 293 PL. Measures in 4 domains (PF, PA, PG, IN) were responsive (standardized response mean > 0.5) but not the SM measures, which have therefore been omitted from these criteria. We developed and tested candidate improvement criteria in a random 2/3 subset from the trials and used the remaining 1/3 for validation. Candidate definitions were indexes of multiple measures, performance of individual measures, and combinations of improvements in each of multiple domains. Based on clinical judgment, we required both % change and a minimum net change e.g., 20% of starting value and 10 on a 0-100 scale. Worsening per domain was defined as worsening by 20% of starting value and by 10 points. Partial remission (for comparison purposes) was defined as an end of trial value < 20/100 in each of the 4 domains. We selected the definition that best discriminated RX from PL by chi square (χ^2), and had PL response rate $\leq 23\%$.

Among 20 candidate criteria, change $\geq 20\%$ and ≥ 10 on each of 3 domains and absence of worsening on the 4th discriminated best in the development subset (51% RX, 25% PL, $\chi^2 = 36.4$, $p < 0.01$). Results were confirmed in the validation subset. Almost all patients in partial remission had also improved. Among all 923 patients improvement rates were 49% RX, 24% PL, with partial remission in 11%RX, 3% PL, and both in 10%RX, 3% PL subjects. Though further validation in data from trials currently underway is needed, we conclude that we have developed a clinically valid, easy to use measure of short-term improvement in AS that discriminates well.

Disclosure: Work reported in this abstract was supported in part by Searle France and Merck, with collaboration from Boehringer Ingelheim

THE EFFECT OF SUBREUM (OM 8980) ON THE PERIPHERAL MANIFESTATIONS IN SPONDYLOARTHROPATHY PATIENTS. H Mielants, F Van den Bosch, S Goemaere, K Goethals, K de Vlam, B Vannieuville, E M Veys Belgium

Introduction: Subreum (OM 8980) is a fractionated hypolytised extract from several forms of L. Colli containing 24mg of immunoreactive peptide fragments, including heat shock proteins (hsp70 and hsp60). Subreum has been proven by several double-blind controlled trials to be effective in rheumatoid arthritis with an excellent benefit-risk profile. Hypothesis was that the efficacy could be due to induction of oral tolerance and to switching the Th1 response towards Th2.

Objective: To study the clinical effect of subreum on the peripheral arthritis of patients with different forms of spondyloarthropathy (SpA).

Methods and Material: In a double-blind placebo-controlled monocenter study, Subreum was given in a dosage of 24mg/day for 48 weeks versus placebo. Included patients (pts) had to fulfil ESSG criteria for at least 3 months and had to present active disease with inflammatory pain and at least one swollen joint. 59 pts CGM, 28F, mean age 40y were evaluated in intention to treat analysis; 23 pts with ankylosing spondylitis (AS), 18 pts with psoriatic arthritis (PsA), 3 pts with reactive arthritis (ReA was not evaluated as subgroup) and 13 pts with undifferentiated spondyloarthropathy (USPA).

Results: Not in the total group, not in the subgroups of AS and PsA any significant difference in the different clinical and biological parameters was found between Subreum and placebo. In the subgroup of USPA, however, a statistical significant improvement was found for all clinical parameters in the Subreum group versus placebo; VAScore for pain ($p < 0.01$), pain at night ($p < 0.02$), articular pain score ($p < 0.05$), number of swollen joints ($p < 0.01$), patient's assessment ($p < 0.05$), physician's assessment ($p < 0.05$). Six out of the 8 pts treated with the active compound came into clinical remission (no synovitis, no tendinitis) versus 1 placebo patient (out of 7) after 24 to 36 weeks of treatment.

Conclusion: Subreum is effective in the subgroup of USPA. Since 20% of these patients during disease evolution will develop AS, Subreum administration could influence this evolution.

Disclosure:

Outcome Variables in Ankylosing Spondylitis: Evaluation of Their Relevance and Discriminant Capacity

ANDREI CALIN, JEAN-PIERRE NAKACHE, ALICE GUEGUEN, HENNING ZEIDLER, HERMAN MIELANTS, and MAXIME DOUGADOS

ABSTRACT. The clinical status of ankylosing spondylitis (AS) can be defined by several domains (e.g., pain, function, metrology, laboratory) and subcomponents within each domain (e.g., pain using visual analog scale, Schober's within metrology). Our aim was (1) To define groups of highly correlated variables in order to determine the most relevant; and (2) to evaluate the capacity of different clinical and biological variables that best discriminate between placebo and active nonsteroidal drugs in AS. Patients with active AS (n = 423) were followed prospectively over 6 weeks while receiving placebo (n = 121) or active nonsteroidal antiinflammatory drugs (n = 352). Eighteen variables were studied, including global assessment, pain, stiffness, functional indices, metrology, disease activity index, and laboratory markers. Statistics included (1) Evaluation of the relevance of the different domains by multivariate analysis (CART tree-structure classification; variable clustering); and (2) evaluation of the discriminant capacity by univariate analysis [i.e., differences in the standardized response mean (SRM) (mean change/SD) between placebo and active drug. A value ≥ 0.60 was considered relevant]. Four clusters were identified (patient's subjective perception, inflammatory symptoms, metrology, laboratory data) with multiple correlation R^2 revealing the most relevant variables to be the Bath Ankylosing Spondylitis Functional Index (BASFI; 0.75), night pain (0.62), Schober's test (0.58), and platelet count (0.55), respectively, within each cluster. In terms of discriminant power (SRM) the patient perceived global status (0.84), lumbar pain (0.73), night pain (0.71), physician global assessment (0.66), and BASFI (0.65) were most relevant in the univariate analysis. Among the 4 most relevant domains are subjective perception, inflammatory symptoms, metrology, and laboratory. Multivariate analysis of the data reveals that the spinal pain and the patient global assessment are the variables that best discriminate between placebo and active nonsteroidal drug in short term studies. (J Rheumatol 1999;26:975-9)

Key Indexing Terms:

ANKYLOSING SPONDYLITIS

OUTCOME

NSAID

Defining outcome in nonlethal chronic disease is notoriously difficult. In rheumatological practice, relevant endpoints such as loss of renal function in systemic lupus erythematosus may exist, while in others (e.g., rheumatoid disease) laboratory variables may be surrogate markers for underlying activity. In ankylosing spondylitis (AS), the situation is complicated because process markers are frequently absent (e.g., erythrocyte sedimentation rate) and few endpoints are clearly defined. Example of the latter include total hip

replacement or fracture of the porotic/immobile spine. Naturally, for the majority of patients, such phenomena do not occur, and the search for appropriate and validated outcome measures continues. No consensus exists as to how many variables are required to provide a full picture and it is unclear whether we should focus on individual outcomes, or whether indices combining a variety of variables would be more appropriate. With the advent of "evidence based medicine" the need for clearly defined and validated outcomes is ever more relevant.

Recently, independent initiatives have resulted in a series of OMERACT meetings to discuss a standardized approach to the study of rheumatoid disease, osteoarthritis, osteoporosis, and systemic lupus (Maastricht 1992,¹⁻³ Ottawa 1994, and Cairns 1996⁴). The situation for AS is less mature, in part because of the diffuse clinical spectrum of the disorder in part due to the lack of laboratory variables. The need for a core set of endpoint measures in this condition is self-evident. Once a consensus exists, research will be enhanced, and data from studies in different parts of the world would be more readily comparable.

From the Royal National Hospital for Rheumatic Diseases, Bath, UK; Hôpital National de Saint Maurice, Saint Maurice, France; Division of Rheumatology, Hannover School of Medicine, Hannover, Germany; University of Ghent, Ghent, Belgium; and Hôpital Cochin, Paris, France.

Supported by the Col. W.W. Pilkington Charitable Trust, John Coates Charitable Trust, and Boehringer Ingelheim Ltd.

A. Calin, MD, FRCP, Consultant Rheumatologist, Royal National Hospital for Rheumatic Diseases; J-P. Nakache, PhD; A. Gueguen, PhD, Department of Biostatistics, Hôpital National de Saint Maurice; H. Zeidler, Professor of Medicine and Rheumatology, Hannover School of Medicine; H. Mielants, Professor of Rheumatology, University of Ghent; M. Dougados, Professor of Rheumatology, Hôpital Cochin.

Address reprint requests to Dr. A. Calin, Royal National Hospital for Rheumatic Diseases, Upper Borough Walls, Bath, UK BA1 1RL.

To address these issues the "Assessments in Ankylosing Spondylitis Working Group" was formed in 1995, and following a literature search, nominal group discussions, plenary reviews, and data assessment, a preliminary core set of efficacy and endpoints was suggested⁵. However, in practical terms one must define precisely what domains should be assessed and which measures within each domain are relevant.

The present investigation explores some of these issues, studying 473 patients with AS who took part in a 6-week, placebo controlled randomized study of nonsteroidal antiinflammatory drug (NSAID) treatment using self-administered instruments and other variables. We define the discriminant capacity of tests to distinguish between placebo and fast acting drug, and develop a core set of outcome measurements, on the basis of actual data rather than by using a delphi approach⁵.

PATIENTS AND METHODS

Patient population. Outpatients with active disease fulfilling the modified New York criteria for AS were recruited. Disease activity was defined by patients requiring daily intake of NSAID during the preceding month.

Study design. The investigation was a longitudinal study of 6 weeks' duration, during which a complete examination was performed at baseline and 1, 3, and 6 weeks thereafter. Patients received either placebo or NSAID. The double blind, placebo controlled study was of 6 weeks' duration⁶.

Assessment criteria. Clinical evaluation was carried out at baseline and after 1, 3, and 6 weeks' therapy by the same investigator for each patient. The 15 assessments are summarized in Table 1. Two functional indices were used: The functional index of Dougados (ASFI)⁷ and the Bath Ankylosing Spondylitis Functional Index (BASFI)⁸ which focuses on 10 questions pertaining to function, measured on a visual analog scale (VAS). In addition, disease activity was assessed with the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)⁹ — a self-administered instrument with 6 questions relating to individual domains of fatigue, spinal pain, joint pain, and symptoms, together with perception of pain relating to the entheses (i.e., tender bony sites around the body) and 2 aspects of morning stiffness (quantity and quality).

Statistical analysis. Two distinct approaches were taken. First, our aim was to define the most relevant domains (i.e., group of interrelated clinical attributes) and the most important individual characteristics within these

Table 1. Variables measured at each visit.

1. Morning stiffness
2. Night pain
3. Patient perceived global status
4. Pain (visual analog scale)
5. Spinal pain
6. Ankylosing Spondylitis Functional Index (Dougados)
7. Bath Ankylosing Spondylitis Disease Activity Index
8. Bath Ankylosing Spondylitis Functional Index
9. Chest expansion
10. Finger-to-floor distance
11. Occiput-to-wall distance
12. Schober test
13. Hemoglobin
14. CRP
15. Platelet count

groups. Second, we evaluated the discriminant capacity of the different variables (i.e., the ability to recognize differences between active drug and placebo, in terms of simplicity, validity, sensitivity to change, and reliability).

A method of variable clustering, namely, the VARCLUS procedure of SAS software¹⁰, has been used to define groups of highly correlated variables to determine the most relevant domains to be evaluated. Three different clustering procedures have been performed on the whole population: (1) on the absolute baseline values of the variables collected at entry, (2) on the absolute last-available values of the variables, and (3) on the changes in the variables during the study. The clusters of variables so obtained appear to be relevant to explain the variability of the clinical representation of the patients. The multiple correlation R^2 between each variable included in a cluster and its own cluster allowed us to determine the most representative variable of the cluster; the other variables of the cluster are thereby considered redundant.

To explore the discriminant capacity of the variables, both univariate and multivariate analyses were conducted. The univariate analysis was based on the differences in the SRM between placebo and active drug. The SRM within each group (placebo vs active drug) was defined by the ratio of the mean changes of the variables during the study over the standard deviation of these changes (mean change/SD). A value of > 0.60 is usually considered a clinically relevant discriminant power¹¹.

In case of withdrawal of the patient during the study, we used the LOCF technique (last observation carried forward) to evaluate these patients.

The multivariate analysis used a CART tree-structured classification^{12,13}, with the treatment group (placebo vs NSAID) as the dependent variable and the changes during the study of all the variables plus the overall assessment of the patient at the end of the study as independent variables. Starting from the root node containing the whole sample, CART looks for the best split to separate the 2 treatment groups. Once the best split is found, CART repeats the search process for each child node, continuing, respectively, until further splitting is impossible. The large binary tree so obtained is then pruned and validated, leading to a smaller tree. In the last tree, relatively few variables appear explicitly in the splitting criterion, but CART keeps track of surrogate splits in the tree-growing process, providing a measure of the importance of each variable in the construction of the large tree.

RESULTS

Patients and study design. From the 603 screened patients, 473 were included in the study. During the 6 weeks of the study 110 patients withdrew either for lack of efficacy and/or toxicity or for other reasons. The main characteristics of the patients by treatment group are summarized in Table 2. There was no statistically significant difference between the 2 groups apart from an age difference (40.1 vs 43.5 years). The absolute values of the evaluated variables at

Table 2. Characteristics of patients.

	Total (n = 473)	Placebo (n = 121)	NSAID (n = 352)
Male (%)	78	72	80
Age (yrs)	42.6	40.1	43.5
Disease duration (yrs)	12.2	11.9	12.3
Peripheral joint disease (%)	28	30	27
Uveitis (%)	27	26	28
Family history (%)	27	26	28
HLA-B27 (%)	86	90	85

0 and their changes during the study per treatment group are summarized in Table 3.

Relevant outcome domains. Four clusters were identified with the multivariate analysis and the data are summarized in Table 4. A series of clusters of highly correlated variables were revealed at entry, last visit, and for changes during the study. For example, for changes during the study, the multiple correlation, R^2 , between a specific variable and its own cluster revealed values ranging from 0.30 to 0.75 [i.e., C-reactive protein (CRP) and BASFI, respectively]. In each case, the higher the value, the more relevant the endpoint. At the last visit, for example, within metrology, the "best" measurement was that of the Schober test (0.58), while chest expansion (0.48) was the "worst." Similarly, at entry, BASFI (0.75), BASDAI (0.68), and ASFI (0.67) were the optimum variables. It is also clear that 4 domains appear to be relevant, pertaining to: (1) patient subjective perception BASFI, BASDAI, ASFI, Global Status, pain VAS, spinal pain — in that order); (2) inflammatory symptoms of night pain and morning stiffness; (3) metrology; and (4) laboratory tests.

Discriminant capacity. Univariate analysis. For the discriminant power of each variable (differences in the SMR between the placebo and active drug), a value > 0.60 was

considered clinically relevant. The results in terms of the hierarchy of power for the individual assessments are shown in Figure 1. Thus, in terms of differentiating placebo from active drug, the global VAS derived from the patient provides the most power (0.84), followed by lumbar pain (0.73), night pain (0.71), physician opinion (0.66), and BASFI (0.65). At the other end of the spectrum, metrology and laboratory tests were of little value.

Multivariate analysis. All the variables were taken into account simultaneously to build a class probability tree using the CART procedure to discriminate between placebo and NSAID. This analysis, performed on the changes of all the variables as independent variables, led to the class probability tree presented in Figure 2.

The proportion of the 2 treatment groups in the subsample are provided within each node of this tree and the relevant variables are indicated on the branches. The predominant treatment group and its estimated probability are recorded within each terminal node (squared nodes). The difference between the observed proportions and the estimated probabilities in each terminal node are due to the fact that the treatment groups are equally treated regardless of the observed sample proportions.

This tree-structured classification reveals that spinal pain

Table 3. Changes during the study in the evaluated variables by treatment groups (placebo = 121, NSAID = 352 patients).

Variables	Placebo Changes	Treatment Group NSAID Changes	p
Global assessment			
Patient (VAS)	-7 ± 30	-29 ± 27	0.0001
Doctor (VAS)	-7 ± 27	-24 ± 26	0.0001
Pain			
VAS			
Spinal pain			
Cervical	-0.0 ± 0.9	-0.5 ± 1.0	0.0001
Dorsal	-0.1 ± 1.0	-0.6 ± 1.1	0.0001
Lumbar	-0.1 ± 1.2	-0.9 ± 1.1	0.0001
Inflammation			
Night pain	-0.0 ± 0.9	-0.7 ± 0.9	0.0001
Morning stiffness	-7 ± 29	-22 ± 31	0.0001
CRP	4.2 ± 13.6	0.4 ± 13.7	0.0075
Hemoglobin	0.0 ± 0.5	0.1 ± 0.7	0.5417
Platelets	2 ± 36	4 ± 39	0.1589
Functional disability			
ASFI	+0.5 ± 6.3	-3.0 ± 5.9	0.0001
BASFI	+0.4 ± 20.4	-12.7 ± 20.1	0.0001
Range of motion			
Schober test	+0.1 ± 1.4	+0.3 ± 1.2	0.0775
Finger-to-floor distance	+0.5 ± 11.0	+3.5 ± 9.7	0.0006
Occiput-to-wall	+0.1 ± 1.9	-0.5 ± 3.4	0.0161
Chest expansion	-0.2 ± 1.5	+0.4 ± 1.7	0.0002
BASDAI (total score)	-22 ± 74	-76 ± 105	0.0002
Analgesic consumption			
No. of pills/day	1.9 ± 2.1	0.9 ± 1.5	0.0001

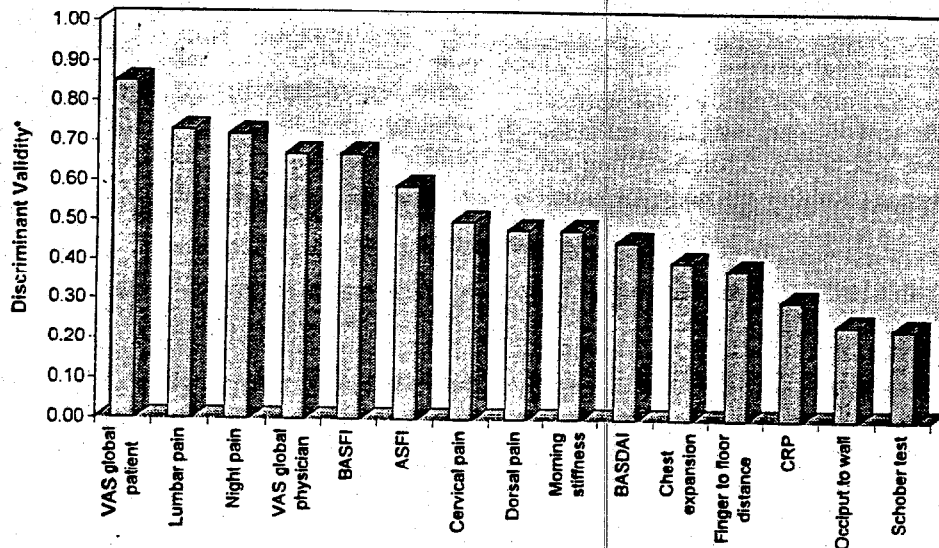


Figure 1. Discriminant validity of variables in AS between placebo and fast acting drug. *Discriminant power: differences in the standardized response mean (mean change/SD) between the placebo and the active drug.

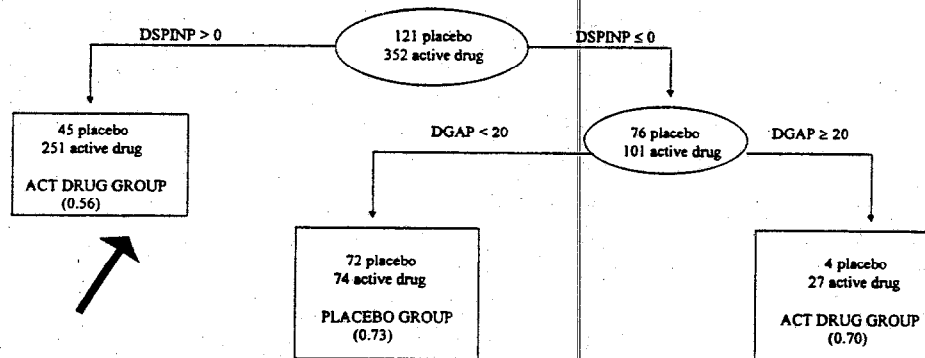


Figure 2. Nonparametric discrimination (CART) between treatment groups (active drug vs placebo). DSPINP: difference between spinal pain (sum of cervical, thoracic, and lumbar) at baseline and last visit. DGAP: difference between global assessment of pain at baseline and last visit.

and patient's global assessment are the variables that best discriminate between placebo and active drug. Moreover, as shown in Table 4, the 4 best variables in relative importance were patient global assessment, spinal pain, BASDAI, and BASFI.

DISCUSSION

Since 1980 there has been increased interest in outcomes research in the rheumatic diseases^{14,15}, culminating in a series of OMERACT meetings¹⁻⁴. Parallel to this, attempts by individual groups were made to address the various components of disease status in AS¹⁶, with, for example, a selection of a core set of variables proposed by a Dutch group¹⁷, together with a series of additional disease-specific measures to assess the disease from the UK and France^{7-9,18-20}.

Finally, preliminary core sets for endpoints in AS were recommended following a delphi/consensus approach —

Table 4. Using multivariate analysis (variable clustering) 4 clusters were identified. In each cluster the variables with the highest R² are the most relevant.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
BASFI 0.75	Night pain 0.62	Schober test 0.58	Platelets 0.55
BASDAI 0.68	Morning stiffness 0.62	Occiput-wall distance 0.47	Hb 0.36
ASFI 0.67		Finger to floor 0.47	CRP 0.30
Global (pt) 0.57		Chest expansion 0.45	
Pain (VAS) 0.55			
Spinal pain 0.33			

agreement reached after discussion with academics, clinicians, and patients with AS⁵.

We have reevaluated the various instruments available for the assessment of AS by following a large cohort (473 patients, 121 receiving placebo and 352 NSAID) over a 6 week investigation. We investigated which of the many variables best discriminate between active agent and placebo, and which core set of outcome measures best defines this phenomenon.

Specifically, we observed that simple inexpensive validated questions supplied by the patient provide data that discriminate better between active and dummy agents than results from metrology or laboratory tests. Specifically, with the a priori decision that a value of > 0.60 is clinically relevant, the analysis of the differences in the SMR between placebo and the active drug gave results ranging from a high of 0.84 for the patient VAS to 0.65 for BASFI. Intermediate values of 0.73, 0.71, and 0.66 resulted from VAS relating to lumbar pain, night pain, and physician impression of global status, respectively.

Additional results in descending order related to ASFI, cervical pain, dorsal pain, morning stiffness, and BASDAI (0.59 to 0.42, respectively), while chest expansion (0.37), finger-to-floor distance (0.31), analgesic consumption (0.30), CRP (0.28), occiput-to-wall distance (0.20), and Schober test (0.18) all performed rather poorly by comparison.

Parallel to this analysis, an evaluation of variable clustering revealed different groups that were internally highly correlated with one another at entry, at the last visit, and pertaining to change during the study: for example, groupings relating to laboratory tests, metrology, inflammation (sleep disturbance and morning stiffness) and a mixture of disease activity and functional attributes grouped together; another example, in descending order, BASFI with an R² value of 0.75, BASDAI of 0.68, followed by ASFI (0.67), VAS (patient global assessment) 0.57, pain scale 0.55, and spinal pain 0.33 on the one hand, with night pain and morning stiffness both scoring 0.62 on the other. In terms of laboratory variables, platelet count, hemoglobin, and CRP with an R² value of 0.55, 0.36, and 0.30 clustered together, while for metrology, the Schober test of 0.58, occiput-to-wall distance 0.47, finger-to-floor distance 0.47, and chest expansion 0.45 were also in a single group.

One limitation of our study is that not all possible domains have been studied in this investigation. For example, we cannot comment on extraarticular features or radiological change. Moreover, NSAID may have a limited effect on AS. In addition, we have no data on the longterm efficacy of NSAID or putative disease modifying drugs. However, it remains possible that NSAID and intensive physiotherapy may alter the structural changes seen in poorly managed AS.

In conclusion, these results were obtained in a short term

NSAID/placebo controlled study. Confirmation will be needed from longer term investigations using different treatment modalities.

REFERENCES

1. Conference on Outcome Measures in Rheumatoid Arthritis Clinical Trials, Maastricht, The Netherlands, April 29–May 3, 1992. *J Rheumatol* 1993;20:525–91.
2. Tugwell P, Boers M, Brooks P, et al. OMERACT II Conference proceedings. *J Rheumatol* 1995;22:980–99; 1185–207; 1399–433.
3. Tugwell P, Boers M, for the OMERACT Committee. Developing consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials. *J Rheumatol* 1993;20:555–6.
4. Brooks P, Boers M, Tugwell P. OMERACT III. The "ACT" Revisited. *J Rheumatol* 1997;24:764–5.
5. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan A, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. *J Rheumatol* 1997;24:2225–9.
6. Dougados M, Gueguen A, Nakache J-P, et al. Ankylosing spondylitis: what is the optimal duration of a clinical study? A one year vs a six week NSAID trial. *Br J Rheumatol* 1999; (in press).
7. Dougados M, Gueguen A, Nakache JP, et al. Evaluation of a function index for patients with ankylosing spondylitis. *J Rheumatol* 1990;17:1254–5.
8. Calin A, Garrett SL, Whitelock HC, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index (BASFI). *J Rheumatol* 1994;21:2281–5.
9. Garrett SL, Jenkinson TR, Whitelock HC, Kennedy LG, 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.
10. SAS Institute Inc. SAS/STAT User's Guide, version 6, vol 2. Cary, North Carolina: SAS Institute; 1991.
11. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopaedic evaluation. *Med Care* 1990;28:632–42.
12. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont: Wadsworth; 1984.
13. Steinberg D, Colla P. CART for Windows. San Diego, CA: Salford Systems; 1994.
14. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcomes in arthritis. *Arthritis Rheum* 1980;23:137–45.
15. Kirwan J. A theoretical framework for process, outcome and prognosis in rheumatoid arthritis. *J Rheumatol* 1992;19:333–6.
16. Bakker C, Boers M, van der Linden S. Measures to assess ankylosing spondylitis: taxonomy, review and recommendations. *J Rheumatol* 1993;20:1724–30.
17. Creemers MCW, van't Hof MA, Franssen MJAM, van de Putte LBA, Gribnau FWJ, van Riel PLCM. Disease activity in ankylosing spondylitis: selection of a core set of variables and a first step in the development of a disease activity score. *Br J Rheumatol* 1996;35:867–73.
18. Mander M, Simpson JM, McLellan A, et al. Studies with an enthesitis index as a method of clinical assessment in ankylosing spondylitis. *Ann Rheum Dis* 1987;46:197–202.
19. Jones SD, Steiner A, Garrett SL, Calin A. The Bath Ankylosing Spondylitis Global Score. *Br J Rheumatol* 1996;35:66–71.
20. Jenkinson T, Mallorie PA, Whitelock HC, Kennedy LG, Garrett SL, Calin A. Defining spinal mobility in ankylosing spondylitis: the Bath Ankylosing Spondylitis Metrology Index. *J Rheumatol* 1994;21:1694–8.

Radiological Scoring Methods in Ankylosing Spondylitis: Reliability and Sensitivity to Change Over One Year

ANNEKE SPOORENBERG, KURT de VLAM, DÉSIRÉE van der HEIJDE, ERIK de KLERK, MAXIME DOUGADOS, HERMAN MIELANTS, HILLE van der TEMPEL, MAARTEN BOERS, and SJEFF van der LINDEN

ABSTRACT. Our aim was to compare reliability and sensitivity to change of different radiological scoring methods in ankylosing spondylitis (AS). Two trained observers scored 30 AS radiographs twice with an interval of 4 weeks. The same two observers scored 187 AS radiographs in pairs, at baseline and after one year followup, to measure change and agreement on change. The sacroiliac (SI) joints were scored in 5 grades by the New York method and the SASSS (Stoke Ankylosing Spondylitis Spine Score). Hips were graded 0–5 (according to Larsen). Cervical and lumbar spine were graded (0–4, Bath Ankylosing Spondylitis Radiological Index, BASRI), and scored in detail (0–72, SASSS). SASSS of the cervical and lumbar spine scored on the anterior sites of the vertebrae proved most reliable, with both intra and interobserver intraclass correlation coefficients (ICC) between 0.87 and 0.97. BASRI was only moderately reliable, with Cohen's kappa ranging between 0.50 and 0.82 for intra, and 0.38–0.64 for interobserver reliability. Similarly, SI joint scores (New York, SASSS) showed intraobserver kappa between 0.56 and 0.84, and interobserver reliability with kappa between 0.37 and 0.47. Larsen hip scores proved unreliable: moderate intraobserver kappa of 0.47–0.58 and low interobserver kappa of 0.29. After retraining, interobserver kappa did not improve (0.45 and 0.17). In retrospect, a one year period was too short to measure sensitivity to change. Observers agreed that no change occurred in up to 89% of cases. A measurable change of deterioration or improvement occurred rarely. We conclude that in AS, only the SASSS method for the spine and the BASRI reached good reliability. Other methods for spine, SI joints, and hips were moderately reliable at best. There was moderate to good agreement on no change between the observers. No method showed change over a period of one year in a considerable number of patients. (J Rheumatol 1999;26:997–1002)

Key Indexing Terms:

RADIOLOGY	BATH ANKYLOSING SPONDYLITIS RADIOLOGY INDEX
STOKE ANKYLOSING SPONDYLITIS SPINE SCORE	OUTCOME
ANKYLOSING SPONDYLITIS	SPONDYLOARTHROPATHY

Structural damage is considered an important outcome in rheumatic disease, and ankylosing spondylitis (AS) is no exception. The Assessments in Ankylosing Spondylitis (ASAS) Working Group chose radiographs of the spine, sacroiliac (SI) joints, and hips as important endpoints in the

core set in trials concerning disease controlling antirheumatic therapy (DC-ART) in AS¹. Radiology has also proven to be an important endpoint in the core set of DC-ART trials in rheumatoid arthritis. In AS, however, the evaluation of radiological changes is very difficult; radiological sacroiliitis can easily be missed, syndesmophytes must be differentiated from osteophytes, and changes such as squaring may be an early or late change in AS². The SI changes are most frequently scored using the 5 grade New York criteria. The Stoke Ankylosing Spondylitis Spine Score (SASSS) also contains a 5 grade score for SI joints almost similar to the New York method. There are mainly 2 different scoring methods for AS changes in the spine. The Bath Ankylosing Spondylitis Radiology Index (BASRI) is a global scoring method graded from 0 to 4, while SASSS is a more detailed scoring method with a total scoring range from 0 to 72³. At the time we started this project there was no specific published scoring method for the hips in AS. Although the reliability and the sensitivity to change of the different scoring methods are assessed by the developers, little is known

From the University Hospital Maastricht, Maastricht, The Netherlands; University Hospital Gent, Belgium; Hôpital Cochin, Paris, France; Maasland Hospital, Sittard; and VU University Hospital, Amsterdam, The Netherlands.

A. Spoorenberg, MD, University Hospital Maastricht; K. de Vlam, MD, Rheumatologist, University Hospital Gent; D.M.F.M. van der Heijde, MD, PhD, Associate Professor in Rheumatology; E. de Klerk, MD, MSc, University Hospital Maastricht; M. Dougados, MD, PhD, Professor in Rheumatology, Hôpital Cochin; H. Mielants, MD, PhD, Professor in Rheumatology, University Hospital Gent; H. van der Tempel, MD, Rheumatologist, Maasland Hospital; M. Boers, MSc, MD, PhD, Professor in Clinical Epidemiology, VU University Hospital; S. van der Linden, MD, PhD, Professor in Rheumatology, University Hospital Maastricht.

*Address reprint requests to Dr. A. Spoorenberg, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ, Maastricht, The Netherlands;
E-Mail: jspo@sint.azm.nl*

about the reliability of these scoring methods in the hands of other investigators. Our objective was to compare the reliability and sensitivity to change over one year of the available radiological scoring methods developed for AS using the same set of radiographs for all methods.

MATERIALS AND METHODS

Study I

Inter and intraobserver reliability. Two observers (AS and KV) scored 30 sets of radiographs of 30 outpatients with AS who satisfied the modified New York criteria⁴. The 2 observers had 2 training sessions with 2 experienced rheumatologists to gain experience with the scoring methods. All abnormalities present on the radiographs were discussed in detail and a score according to the various scoring methods was assigned. After each training session the observers scored a set of radiographs independently and discussed the results with each other and with supervisors. Three training sessions were needed to minimize discrepancies between the 2 observers. To measure reliability of the scoring methods the University Hospital Maastricht, the University Hospital Gent, and the Hôpital Cochin in Paris each provided 10 sets of radiographs. These radiographs were a random sample of the radiographs taken at baseline of study II. The following radiographs were taken: the posterior-anterior view of the pelvis to score the SI joints and the hips, the anterior-posterior (AP) view and the lateral view of the lumbar and cervical spine to score the spine. For all scoring methods interobserver reliability was calculated. To assess intraobserver reliability the same 30 sets of radiographs were scored a second time after 4 weeks. When the results of one of the scoring methods appeared to be insufficient, this method was trained a 4th time, followed by a re-scoring of this method on 30 other sets of baseline radiographs. We randomized the scoring methods and the order in which the radiographs were scored (hips, SI joints, lumbar spine, cervical spine).

Scoring methods, SI joints. The SI joints were scored according to the New York method and the SASSS^{5,6}. Both methods score the lower half of the SI joints. The New York scoring method of the SI joint is graded: 0 = no disease; 1 = suspicious change (no specific abnormality); 2 = minimal sacroiliitis (loss of definition at the edge of the SI joints, there is some sclerosis and perhaps minimal erosions, there may be some joint space narrowing); 3 = moderate sacroiliitis (definite sclerosis on both sides, blurring and indistinct margins and erosive changes with loss of joint space); 4 = complete fusion or ankylosis of the joint (without some residual sclerosis). The SASSS scoring method of the SI joints is graded: 0 = normal; 1 = blurring of joint margin; 2 = 1+ periarticular sclerosis or pseudo-widening; 3 = 2+ erosions or partial bony bridging; 4 = complete ankylosis. The radiographs were scored missing when the radiograph was not available or when the quality of the radiograph made it impossible to score.

Scoring method, hips. The hips were scored according to Larsen grade: 0 = normal; 1 = slight abnormality; 2 = definite early abnormality; 3 = medium destructive abnormality; 4 = severe destructive abnormality; 5 = mutilating abnormality with gross bone deformation. To score the hips we also used Larsen's reference radiographs⁷. Originally Larsen developed his hip scoring method for patients with rheumatoid arthritis. Again, the radiographs were scored missing when they were not available or when the quality of the radiograph made it impossible to score.

Scoring methods, spine. The BASRI developed for the AP and lateral view of the lumbar spine and the lateral view of the cervical spine was also applied on the AP view of the cervical spine. Two versions of the BASRI have been developed. The first was published in 1995 and used in studies I and II⁸. The new BASRI appears in these proceedings and was used in Study II only. The old BASRI is a global score, graded: 0 = normal; 1 = suspicious changes (only squaring); 2 = indicative of obvious squaring of vertebrae with erosions or sclerosis; 3 = more widespread changes with obvious syndesmophyte formation; 4 = ankylosis⁸. The radiographs were scored

missing when they were not available or when the quality of the radiograph made it impossible to score the BASRI.

The SASSS was scored from the lower border of 12th thoracic vertebra up to and including the upper border of the sacrum on the lumbar lateral view. This scoring method was assessed on both the anterior and posterior site of the vertebrae with a score ranging from 0 to 36 for the total anterior and posterior scoring sites, so the total score ranges from 0 to 72⁶. The "modified" SASSS according to Creemers, *et al* was scored from the lower border of the 2nd cervical vertebra until the upper border of the 1st thoracic vertebra and the lower border of the 12th thoracic vertebra until the upper border of the sacrum on the cervical and lumbar lateral view⁹. This scoring method was only assessed on the anterior site of the vertebrae, with a scoring range from 0 to 36 for the cervical spine and 0 to 36 for the lumbar spine. Therefore the total score of the modified version also ranges from 0 to 72. The score is the same for both the original and modified SASSS: 0 = normal; 1 = erosion, sclerosis, squaring; 2 = obvious syndesmophyte formation; 3 = total bony bridging. We also attempted to score the SASSS on the posterior site of the lateral view of the cervical spine, but this proved to be technically impossible. Missing scoring sites for the SASSS method were handled as follows: when up to 3 scoring sites for each view (cervical anterior, cervical posterior, lumbar anterior, and lumbar posterior) could not be scored, the mean of the other scoring sites was applied; when more than 3 scoring sites could not be scored the whole SASSS method for that particular view was scored missing.

Study II

Interobserver reliability and sensitivity to change. We studied 2 sets of AS radiographs taken with an interval of one year of 187 consecutive outpatients with AS who satisfied modified New York criteria; participants were 124 patients from the University Hospital Maastricht and the Maasland Hospital Sittard, The Netherlands, 50 from the Hôpital Cochin, Paris, and 13 patients from University Hospital Gent⁴, all secondary and tertiary referral centers.

The sets of radiographs were scored paired without knowledge of the chronology of the radiographs in a random order by the same 2 observers. We used the same scoring methods as in the previous section describing Study I. We also used the recently developed new BASRI¹⁰. The main difference from the old BASRI is that the "new" method takes into account the number of vertebrae with a particular change. The new BASRI is graded 0 = no change; 1 = no definite change; 2 = erosions, squaring, sclerosis, or syndesmophytes on ≤ 2 vertebrae; 3 = syndesmophytes on ≥ 3 vertebrae \pm fusion involving ≤ 2 vertebrae; 4 = fusion involving ≥ 3 vertebrae. With the new BASRI the AP and lateral view of the lumbar spine are combined as one score. The BASRI-spine (BASRI-s) is the composite score of the new BASRI scored on the lumbar spine, cervical spine, and SI joints (New York method), with a total score ranging from 2 to 12¹⁰.

Because of poor intra and interobserver reliability found in Study I, the old BASRI applied to the cervical AP view and the SASSS applied to the posterior site of the lateral view of the cervical and lumbar spine were not scored in this second study.

Statistics. For simplicity, joint pairs (hips and SI joints) were regarded as independent units, i.e., their possible correlation was ignored. Inter and intraobserver reliability (agreement) of the different scoring methods were analyzed for categorical data by the unweighted kappa (κ) statistic and for data on a quasi-interval scale by the intraclass correlation coefficient (ICC) with observer as fixed facet. To visualize the observer agreement we plotted the data using the Bland and Altman method¹¹. Sensitivity to change of the scoring methods was assessed by interobserver reliability of change scores. For grading scales a change of one grade was defined as the minimum relevant difference. The interobserver agreement of change of at least one grade were quantified by the unweighted kappa. For data on a quasi-interval scale (SASSS spine) a smallest detectable difference (SDD) was estimated in the situation of 2 fixed observers yielding a mean change score, as described for RA in the Imaging Methods of the OMERACT IV proceedings¹².

RESULTS

Study I

Inter and intraobserver reliability of single radiographs.

Scores of the SI joints and the hips showed moderate intraobserver reliability, with kappa ranging from 0.47 to 0.84 (Table 1). The interobserver agreement was worse. For the 2 scoring methods of the SI joints the interobserver agreement was 0.47 and 0.37. The interobserver agreement for hip score was poorest, with a kappa of 0.29. After another training session and re-scoring, kappa improved to 0.45 (Table 1).

The old BASRI scored on both views of the lumbar and cervical spine also showed moderate intraobserver reliability: kappa 0.50–0.82. The interobserver reliability for the BASRI anterior view of the lumbar spine and the BASRI lateral view of the cervical spine was moderate (kappa 0.60 and 0.50). About half of the radiographs could not be scored

for the BASRI of the AP view of the cervical spine for technical reasons. Therefore, no reliability was calculated. The initial interobserver agreement of the BASRI scored on the lateral view of the lumbar spine was poor: kappa 0.38. After retraining and re-scoring the kappa improved to 0.64 (Table 1).

The SASSS scored on both sides of the lateral view of the lumbar and cervical spine showed good inter and intraobserver reliability, with ICC ranging from 0.84 to 0.95. Although the ICC for the SASSS scored on the lateral view of the lumbar spine was relatively high (0.84), the mean values of Observers 1 and 2 differed significantly ($p = 0.01$, re-scoring $p = 0.002$). The SASSS could not be applied on the posterior scoring sites of the lateral view of the cervical spine because, for technical reasons, about half of the radiographs could not be scored by one of the observers (Table 1).

Table 1. Study I: Inter and intraobserver reliability of single radiographs ($n = 30$). Values are mean \pm SD.

Method	Scoring Range	Intraobserver Reliability		Interobserver Reliability
		Observer 1	Observer 2	
SI joint total,				
New York	0–4	3.1 \pm 0.60, $\kappa = 0.84$	3.2 \pm 0.62, $\kappa = 0.65$	0.1 \pm 0.58, $\kappa = 0.37$
SASSS	0–4	3.3 \pm 0.69, $\kappa = 0.69$	3.2 \pm 0.73, $\kappa = 0.56$	0.2 \pm 0.65, $\kappa = 0.47$
Hips total		0.5 \pm 0.74, $\kappa = 0.58$	1.1 \pm 1.09, $\kappa = 0.47$	0.6 \pm 0.85, $\kappa = 0.29$
Larsen	0–4	0.7 \pm 0.79	0.8 \pm 0.89	0.1 \pm 0.62, $\kappa = 0.45$
Larsen (after retraining)	0–4			
BASRI lumbar spine		2.1 \pm 1.80, $\kappa = 0.79$	1.8 \pm 1.67, $\kappa = 0.64$	0.1 \pm 0.74, $\kappa = 0.60$
AP	0–4			
Lateral	0–4	2.1 \pm 1.53, $\kappa = 0.82$	1.6 \pm 1.36, $\kappa = 0.68$	0.3 \pm 1.17, $\kappa = 0.38$
Lateral (after retraining)	0–4	0.7 \pm 0.79	0.8 \pm 0.89	0.1 \pm 0.88, $\kappa = 0.64$
BASRI cervical spine		1.2 \pm 1.58, $n = 17$	1.5 \pm 1.10, $n = 14$	0.1 \pm 2.03, $n = 14$
AP	0–4			
Lateral	0–4	2.8 \pm 1.15, $\kappa = 0.71$	2.3 \pm 1.32, $\kappa = 0.50$	0.6 \pm 1.01, $\kappa = 0.50$
SASSS lumbar spine		9.7 \pm 7.0, ICC = 0.97	6.5 \pm 9.4, ICC = 0.98	7.4 \pm 9.3, ICC = 0.93
Lateral anterior	0–36	6.6 \pm 9.6, ICC = 0.91	3.6 \pm 7.8, ICC = 0.95	5.0 \pm 8.3, ICC = 0.84
Lateral posterior	0–36	2.5 \pm 6.3	6.7 \pm 10.3	3.4 \pm 7.0, ICC = 0.85
Lateral posterior (after retraining)	0–36			
SASSS cervical spine		11.9 \pm 9.9, ICC = 0.93	11.4 \pm 9.2, ICC = 0.91	11.7 \pm 9.7, ICC = 0.87
Lateral anterior	0–36	12.5 \pm 9.9, ICC = 0.95, $n = 25$	10.7 \pm 12.0, $n = 14$	11.6 \pm 10.8, $n = 14$
Lateral posterior	0–36			

Calculation of baseline descriptive (mean, standard deviation = SD): Intraobserver: SI joints, hips, BASRI \rightarrow mean of the 2 observations = (score of observation 1 + score of observation 2) divided by 2, for each observer. Interobserver: SI joints, BASRI \rightarrow difference between Observer 1 and Observer 2 at baseline = (score Observer 1 – score Observer 2 at baseline (perfect agreement mean = 0 and SD = 0)). SASSS \rightarrow mean of Observer 1 and Observer 2 at baseline = (score Observer 1 + score of Observer 2) at baseline divided by 2.

Study II

Interobserver reliability at baseline. We included 187 patients in this study; 127 were male, 60 female; this constitutes twice the normal male:female ratio in AS populations. The median age was 43 years (range 18–78). There is a striking difference between median duration of complaints (17.9 yrs, range 0.3–54) and the median duration of disease (9.4 yrs, range 0.1–41), confirming that patients with AS have complaints long before the diagnosis is made.

On the baseline radiographs, SASSS lumbar and cervical spine scores showed excellent interobserver reliability (ICC > 0.90; Table 2). The new BASRI-s was also good, with an ICC of 0.84 (Table 3). In contrast, the Larsen hip score again was poor, with a kappa of 0.17. For the other scoring methods, moderate interobserver reliability was confirmed (kappa around 0.60; Table 3).

Sensitivity to change. Over all, we found little change in damage over the course of one year. Except for the BASRI-s, both observers agreed in up to 89% of cases that no change had occurred (Table 2). Nevertheless, kappa were ± 0 due to an inherent problem of the kappa measure in situations with high levels of expected agreement. The last column of Tables 2 and 3 shows the distribution of (dis)agreement based on the minimum relevant difference of one grade or the SDD. The SDD is the smallest change that can be detected apart from measurement error. If a patient showed deterioration or improvement above the SDD, the

change was judged as real. In Tables 2 and 3 this is split for the percentage of patients that changed according to only one or according to 2 observers.

Figure 1 shows the Bland and Altman plot of the SASSS scored on the anterior site of the lateral view of the lumbar spine. There is a maximum difference of 13 points between the 2 observers on a scoring range from 0 to 36; the 95% confidence interval of the difference between the 2 observers is ± 2 times the standard deviation (3.4). It further shows that Observer 1 scores consistently higher versus Observer 2. The Bland and Altman plot of the SASSS cervical anterior gives a similar figure.

DISCUSSION

This study confirms that scoring radiographs in AS is difficult. With extensive training, it was possible to attain good to moderate intraobserver reliability. However, only 2 methods showed good or excellent interobserver reliability: the SASSS and the BASRI-s. The old BASRI and the SI joint scores showed moderate interobserver reliability. Scores for the hips (Larsen) and for the AP views of the cervical spine were unacceptable. The use of the Larsen reference radiographs did not improve interobserver reliability⁷. The reason for this may be the pathophysiologic difference in damage caused by rheumatoid arthritis or AS. Hip involvement in AS often shows as bony formations, which cannot be scored using the Larsen method.

Table 2. Study II: Summary statistics of the SASSS method for the spine.

Scoring Method (range)	Average of the 2 Observers ¹				Difference Between the 2 Observers ²				Reliability of Baseline Scores ³ and Change Over One Year ⁴
	Baseline		1 Year		Baseline		1 Year		
	Mean (SD)	Median (min-max.)	Mean (SD)	Median (min-max.)	Mean (SD) ³	Median (min-max.)	Mean (SD) ³	Median (min-max.)	
SASSS lumbar anterior (0–36)	8.9 (9.3)	6.0 (0,36)	9.1 (9.4)	6.0 (0, 36)	1.4 (3.4)	1.0 (-7,14)	1.5 (3.6)	1.0 (-8,13)	ICC 0.93 SDD 4.9, P0 76.9 P+ 0.5%, P(+) 7.6% P- 2.2%, P(-) 3.3%
SASSS cervical anterior (0–36)	9.2 (7.0)	7.0 (1,36)	9.9 (7.4)	8.0 (1, 34)	0.6 (3.2)	0.0 (-6,12)	0.5 (3.0)	0.0 (-10,9)	ICC 0.92 SDD 4.4, P0 88.6 P+ 0%, P(+) 3.4% P- 0.5%, P(-) 8.0%

¹Average score of the 2 observers = (score Observer 1 – score Observer 2) divided by 2.

²Difference between the 2 observers = score Observer 1 – score Observer 2.

³ICC: intraclass correlation coefficient.

⁴Level of reliability of a change of at least the SDD (see Methods). SDD: smallest detectable difference with 95% reliability.

P0: % of patients who did not change according to both observers.

P-: % of patients who deteriorated using the SDD, according to both observers.

P(-): % of patients who deteriorated using the SDD, according to only one observer.

P(+): % of patients who improved using the SDD, according to only one observer.

P+: % of patients who improved using the SDD, according to both observers.

Table 3. Study II: Interobserver reliability on baseline scores and change over 1 year of the scoring methods for the SI joints, hips, and BASRI. Values are mean \pm SD.

Method	Scoring Range	Interobserver Reliability of Baseline Scores ¹	Interobserver Reliability ² of Changed Scores ³
SI joint total			
New York (left and right)	0-4	n = 365, $\kappa = 0.65$ 0.04 \pm 0.42	0.03 \pm 0.43 PO 78.4 1 grade change: P+ 0.3%, P(+) 9.3%, P(-) 12.0%, P- 0%
SASSS total: (left and right)	0-4	n = 364, $\kappa = 0.66$ 0.07 \pm 0.44	0.02 \pm 0.49, PO 74.1 1 grade change: P+ 0.5%, P(+) 10.9%, P(-) 13.7%, P- 0.8%
Hips total			
Larsen (left and right)	0-5	n = 358, $\kappa = 0.71$ 0.40 \pm 0.80	0.04 \pm 0.53, PO 71.8 1 grade change: P+ 0.6%, P(+) 10.9%, P(-) 15.9%, P- 0.8%
Spine: BASRI lumbar			
AP	0-4	n = 181, $\kappa = 0.64$ 0.15 \pm 1.07	0.07 \pm 0.55, PO 83.0 1 grade change: P+ 0%, P(+) 4.4%, P(-) 12.1%, P- 0.5%
Lateral	0-4	n = 185, $\kappa = 0.48$ 0.21 \pm 1.09	0.04 \pm 0.46, PO 67.3 1 grade change: P+ 0%, P(+) 12.9%, P(-) 21.1%, P- 0%
BASRI cervical			
Lateral	0-4	n = 185, $\kappa = 0.53$ 0.20 \pm 0.95	0.03 \pm 0.50, PO 75.8 1 grade change: P+ 0%, P(+) 11.8%, P(-) 16.7%, P- 1.6%
New BASRI-s			
Lumbar, Cervical, Lateral, SI New York	2-12	n = 170, ICC = 0.84 0.25 \pm 1.35	0.05 \pm 0.83, PO 33.8 1 grade change: P+ 2.9%, P(+) 29.3%, P(-) 30.5%, P- 3.5%

¹Mean and SD are calculated from the mean difference of the 2 observers at baseline = (score Observer 1 - score Observer 2) divided by 2 (perfect reliability: mean = 0 and SD = 0).

²Mean and SD are calculated from the progression difference score over 1 year between the 2 observers = (score 1 year - score baseline) of Observer 1 - (score 1 year - score baseline) of Observer 2 (perfect reliability: mean = 0, SD = 0).

³: level of reliability of at least grade 1 change (see Methods).

PO: % of patients who did not change according to both observers.

P-: % of patients who deteriorated according to both observers.

P(-): % of patients who deteriorated according to only one observer.

P(+): % of patients who improved according to only one observer.

P+: % of patients who improved according to both observers.

In retrospect, the one year period was too short to measure sensitivity to change. With observers agreeing that no change occurred in up to 89% of cases, we may conclude that relevant change occurred only rarely. Measures that relate observed to expected agreement (such as kappa and ICC) are of no value in this situation because of high levels of expected agreement. The Bland and Altman method can only be applied reliably in scores with large ranges. That the methods are unable to detect a change over one year may be due to either lack of sensitivity of scoring methods or slow progression of structural damage in AS.

For the BASRI scoring method we reached about the same intra and interobserver agreement as the developers of this method. They also found no change of the BASRI score

over one year, but there was some change after 2 years¹⁰. The developers of the SASSS method found moderate intraobserver reliability and, surprisingly, a higher interobserver reliability using kappa statistics⁶. They found significant change in the SASSS after one year, a mean change of 4.1, with a scoring range from 0 to 72. The order in which the radiographs were scored was known in that study, in contrast with our study. This can markedly influence the results, as has been shown for RA¹³⁻¹⁵.

In conclusion, we recommend that Larsen's method not be used to assess hip damage in patients with AS. A newly developed method (BASRI-hips) is reported elsewhere in these proceedings¹⁰. Of the other methods, the SASSS shows the best reliability. Sensitivity to change of any of

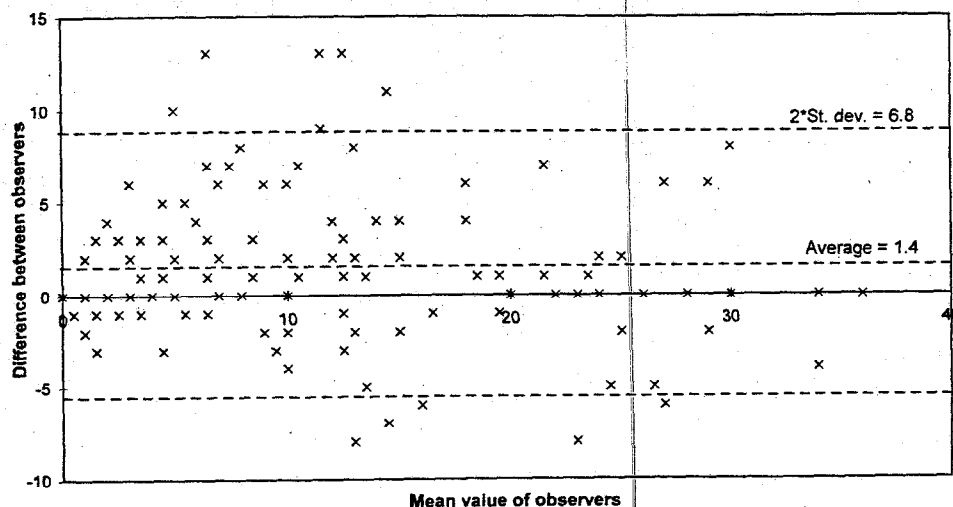


Figure 1. Bland and Altman plot: mean versus difference of 2 observers; SASSS score, lateral view of the lumbar spine (anterior site).

these methods will need to be reassessed in a data set where relevant change has occurred in a substantial number of cases.

REFERENCES

1. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S, on behalf of the Assessment in Ankylosing Spondylitis Working Group. Preliminary core sets for endpoints in ankylosing spondylitis. *J Rheumatol* 1997;24:2225-9.
2. Calin A. Ankylosing spondylitis. Seronegative spondylarthropathies. *Clin Rheum Dis* 1985;11:41-61.
3. van der Heijde D, Spoorenberg A. Plain radiographs as an outcome measure in ankylosing spondylitis. *J Rheumatol* 1999;26:ssss.
4. 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.
5. Dale K. Radiographic gradings of sacroiliitis in Bechterew's syndrome and allied disorders. *Scand J Rheumatol* 1979;32 Suppl 32:92-7.
6. Taylor HG, Wardle T, Beswick EJ, Dawes P. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. *Br J Rheumatol* 1991;30:330-5.
7. Larsen A, Dale K, Morten E. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference film. *Acta Radiol* 1977;18:481-91.
8. Kennedy LG, Jenkinson TR, Mallorie PA, Whitelock HC, Garrett SL, Calin A. Ankylosing spondylitis: the correlation between a new metrology score and radiology. *Br J Rheumatol* 1995;34:767-70.
9. Creemers MCW, Franssen MJAM, van't Hof MA, Gribnau FWJ, van de Putte LBA, van Riel PLCM. A radiographic scoring system and identification of variables measuring structural damage in ankylosing spondylitis [thesis]. 1994; University of Nijmegen, The Netherlands.
10. Calin A, Mackay K, Santos H, Brophy S. A new dimension to outcome. Application of the Bath Ankylosing Spondylitis Radiology Index. *J Rheumatol* 1999;26:988-92.
11. Bland JM, Altman DG. Comparing methods of measurement: why plotting differences against standard method is misleading. *Lancet* 1995;346:1085-7.
12. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. *J Rheumatol* 1999;26:731-9.
13. van der Heijde D, Boonen A, van der Linden S, Boers M. Reading radiographs in sequence, in pairs, or random in rheumatoid arthritis: influence on sensitivity to change [abstract]. *Arthritis Rheum* 1997;40 Suppl:S287.
14. Ferrara R, Priolo F, Cammisa M, et al. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the Grisar study. *Ann Rheum Dis* 1997;56:608-12.
15. Salaffi F, Carotti M. Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: comparison of 3 different reading procedures. *J Rheumatol* 1997;24:2055-6.

Relative Value of Erythrocyte Sedimentation Rate and C-Reactive Protein in Assessment of Disease Activity in Ankylosing Spondylitis

ANNEKE SPOORENBERG, DÉsirÉE van der HEIJDE, ERIK de KLERK, MAXIME DOUGADOS, KURT de VLAM, HERMAN MIELANTS, HILLE van der TEMPEL, and SJEF van der LINDEN

ABSTRACT. Our aim was to determine whether C-reactive protein (CRP) or erythrocyte sedimentation rate (ESR) is more appropriate in measuring disease activity in ankylosing spondylitis (AS). We studied 191 consecutive outpatients with AS in The Netherlands, France, and Belgium. Patients were attending secondary and tertiary referral centers. The external criterion for disease activity was: physician and patient assessment of disease activity on a visual analog scale (VAS) and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI). In each measure we defined 3 levels of disease activity: no activity, ambiguous activity, and definite disease activity. The patients with AS (modified New York criteria) were divided into 2 groups: those with spinal involvement only ($n = 149$) and those who also had peripheral arthritis and/or inflammatory bowel disease (IBD) ($n = 42$). For each criterion of disease activity, the patients with no activity and with definite activity were included in receiver operator curves and used to determine cutoff values with the highest sensitivity and specificity. We also calculated Spearman correlations. The median CRP and ESR were 16 mg/l and 13 mm/h, respectively, in the spinal group and 25 mg/l and 21 mm/h, respectively, in the peripheral/IBD group. In both groups the Spearman correlation coefficients between CRP and ESR were around 0.50. There was moderate to poor correlation between CRP, ESR, and the 3 disease activity variables (0.06–0.48). Sensitivity for both ESR and CRP was 100% for physician assessment and between 44 and 78% for patient assessment of disease activity and the BASDAI, while specificity was between 44 and 84% for all disease activity measures. The positive predictive values of CRP and ESR in our setting were low (0.15–0.69). We conclude that neither CRP nor ESR is superior to assess disease activity. (*J Rheumatol* 1999;26:980–4)

Key Indexing Terms:

ANKYLOSING SPONDYLITIS
ERYTHROCYTE SEDIMENTATION RATE

OUTCOME

DISEASE ACTIVITY
C-REACTIVE PROTEIN

Both erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) are frequently used to evaluate patients with ankylosing spondylitis (AS). Assessment of an acute phase reactant is also a recommended core set endpoint for disease controlling antirheumatic therapy (DC-ART) and clinical record keeping in AS by the international Assessment in Ankylosing Spondylitis Working Group¹. ESR is usually measured with the Westergren method. For CRP there is no formal consensus, but the nephelometric and turbidimetric methods are the most widely used.

From the University Hospital Maastricht, Maastricht, The Netherlands; University Hospital Gent, Gent, Belgium; Hôpital Cochin, Paris, France; Maasland Hospital Sittard, The Netherlands.

A. Spoorenberg, MD; D.M.F.M. van der Heijde, MD, PhD, Associate Professor in Rheumatology; E. de Klerk, MD, MSc, University Hospital Maastricht; M. Dougados, MD, PhD, Professor in Rheumatology, Hôpital Cochin; K. de Vlam, MD, Rheumatologist; H. Mielants, MD, PhD, Professor in Rheumatology, University Hospital Ghent; H. van der Tempel, MD, Rheumatologist, Maasland Hospital Sittard; S. van der Linden, MD, PhD, Professor in Rheumatology, University Hospital Maastricht.

Address reprint requests to Dr. A. Spoorenberg, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ, Maastricht, The Netherlands; E-mail: jspo@sint.azm.nl

Cross sectional data on the comparison of ESR and CRP show that they are highly correlated. The mean values of these 2 acute phase reactants are considerably lower than in rheumatoid arthritis (RA)². Data on the correlation of ESR and CRP in the assessment of disease activity in AS show ambiguous results. One important reason could be that there is no gold standard for disease activity in AS. Longitudinal evaluations of ESR and CRP are primarily focused on DC-ART clinical trials². In these proceedings Ruof and Stucki concluded, based on their literature review, that insufficient data are available to favor either ESR or CRP². The aim of our cross sectional study was to determine whether ESR or CRP is more appropriate in measuring disease activity in AS.

MATERIALS AND METHODS

We included 191 consecutive AS outpatients of the University Hospital Maastricht, The Netherlands, the Maasland Ziekenhuis Sittard, The Netherlands, the University Hospital Gent, Belgium, and the Hôpital Cochin, Paris, France, all secondary and tertiary referral centers. All patients fulfilled the modified New York criteria for AS³. Ours was a longitudinal, observational study with followup visits according to a fixed protocol. In this article only baseline data are reported. As differences with

respect to ESR and CRP in patients with AS with only spinal involvement and those with active peripheral arthritis and/or inflammatory bowel disease (IBD) may exist, we divided the patients into these 2 groups. Active peripheral arthritis was defined as synovitis of at least one large joint (wrist, elbow, shoulder, hip, knee, ankle) or 3 or more small joints (hand and feet joints, sternoclavicular joints). Because there is no gold standard for disease activity in AS we used 3 substitute clinical variables. Our first choice was physician assessment of disease activity on a 10 cm horizontal visual analog scale (VAS) anchored "no disease activity" at 0 cm and "very severe activity" at 10 cm. The 2 other clinical disease activity variables were patient assessment of disease activity on a VAS anchored "no disease activity" at 0 cm and "very severe activity" at 10 cm and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)⁴. This index contains 6 VAS format questions on fatigue, pain of the spine, pain and/or swelling of the peripheral joints, and duration and severity of morning stiffness of the spine. The total score ranges from 0 to 10. To define whether patients showed unambiguous disease activity or not, we subdivided the continuous scale. We defined 3 levels of disease activity for all 3 disease activity variables. A score of ≤ 4 meant no active disease, a score between 4 and 6 disease activity was ambiguous, and a score of ≥ 6 meant that there was definite disease activity. In our analysis we used only the 2 most contrasting groups, where disease activity was defined as definite or no disease activity.

ESR was assessed using the Westergren method (mm/h; normal range male 0-7, female 0-12) and CRP by the turbidimetric method (mg/l; normal range 2-9). The lowest detection limit for CRP was 2 and patients with undetectable levels were assigned 0.

Statistical analyses. To define cutoff values for ESR and CRP to measure disease activity with the best combination of sensitivity and specificity values we calculated receiver operator curves (ROC) for both acute phase reactants versus all 3 clinical disease activity variables both in the spinal and in the peripheral/IBD group. For these analyses only those patients with no disease activity and unambiguous disease activity were used as defined above. This was done in a similar way as described by Wolfe for RA⁵. A ROC is a curve of sensitivity (or true positivity) on the vertical axis and $1 - \text{specificity}$ (or false positivity) on the horizontal axis. The best cutoff values for ESR and CRP will be the point on the curve with the highest sensitivity and specificity, i.e., the highest point in the left upper corner of the graph. Further, the greater the area under the curve, the greater the diagnostic accuracy of the laboratory test. We calculated Spearman correlations. Positive predictive values were also calculated for the defined cutoff points for both ESR and CRP versus the 3 clinical disease activity variables in the spinal only and peripheral/IBD groups. The percentages of patients incorrectly classified according to the cutoff values of the acute phase reactants related to clinical disease activity were also calculated.

RESULTS

There were 149 patients with AS with spinal involvement only, and 42 with active peripheral arthritis and/or IBD. In both groups, the male/female ratio was 2:1. The median duration of complaints and disease was higher in the peripheral group. There was also a broad range in disease duration (Table 1). For the 3 clinical disease activity variables the median disease activity was lowest for the assessment of activity by the physician: 1.5 for the spinal group, and 2.5 for the peripheral arthritis/IBD group. The median ESR was significantly higher in the peripheral arthritis/IBD group: 21 mm/h versus 13 mm/h. Although CRP was also higher in the peripheral/IBD group, the difference was not statistically significant (Table 1). Also, more patients in the spinal groups showed ESR and CRP values in the normal range: 55

Table 1. Study variables, median (range).

	Spinal Group	Peripheral/IBD Group
No.	149	42
Demographics		
Sex (M:F)	2:1	2:1
Age, yrs	40.4 (19-77)	47.6 (22-78)
Duration of complaints, yrs	16.9 (0.3-52.4)	24 (2-53.9)
Disease duration, yrs	9.4 (0.3-40.7)	10.8 (2.5-53.9)
Clinical disease activity variables		
DA physician (0-10)	1.5 (0-9.6)	2.5 (1-8.7)
≤ 4	123 (83%)	24 (57%)
4.1-5.9	14 (9%)	8 (19%)
≥ 6	5 (3%)	5 (12%)
Missing	7 (5%)	5 (12%)
DA patient (0-10)	3.9 (0-10)	4.1 (1-10)
≤ 4	80 (54%)	20 (48%)
4.1-5.9	37 (25%)	8 (19%)
≥ 6	31 (21%)	13 (31%)
Missing	1 (0%)	1 (2%)
BASDAI (0-10)	3.6 (0-9.4)	4.3 (3-9.4)
≤ 4	87 (58%)	19 (45%)
4.1-5.9	42 (28%)	11 (26%)
≥ 6	16 (11%)	10 (24%)
Missing	4 (3%)	2 (5%)
Laboratory disease activity variables		
ESR, mm/h	13* (1-118)	21* (3-80)
CRP, mg/l	16 (0-125)	25 (0-139)

* $p < 0.02$, unpaired t test. DA: disease activity.

and 62%, respectively, compared to the peripheral/IBD group (38 and 39%, respectively). This was significantly different between the spinal and peripheral/IBD groups for CRP, but not for ESR (chi-squared, $p = 0.01$ and $p = 0.06$, respectively). In the spinal group 18% of patients had elevated ESR and normal CRP and 12% elevated CRP and normal ESR. These figures were 16% for both pairs in the peripheral/IBD group.

The Spearman correlations between ESR and CRP were similar in both groups, 0.50 and 0.48, respectively (Table 2). The correlations of the 2 acute phase reactants and the 3 clinical disease activity variables were considerably lower

Table 2. Spearman correlation coefficients.

	Spinal Group		Peripheral/IBD Group	
	ESR	CRP	ESR	CRP
ESR	—	0.50	—	0.48
CRP	0.50	—	0.48	—
DA physician	0.34	0.29	0.48	0.39
DA patient	0.31	0.26	0.31	0.21
BASDAI	0.19	0.23	0.06	0.06

in both groups, range 0.48 for ESR versus physician assessment of disease activity in the peripheral arthritis/IBD group to 0.06 for both ESR and CRP versus the BASDAI, also in the peripheral arthritis/IBD group (Table 2).

Figure 1 shows the ROC of ESR versus CRP against physician assessment of disease activity in the spinal group, with very low cutoff values for ESR and CRP, 15 mm/h and 14 mg/l, respectively. The figure also shows a very large area under the curve (AUC), especially for ESR, suggesting higher diagnostic accuracy of the test; however, the group considered to have active disease by the physician comprised 5 patients only. In all other groups there were more than 16 patients. The other ROC — ESR versus CRP against patient assessment of disease activity and BASDAI in the spinal group (Figures 2 and 3) — are more or less identical, with low cutoff values and no obvious difference between ESR and CRP. However, AUC are substantially smaller compared with Figure 1, suggesting lower diagnostic accuracy. The ROC for the peripheral arthritis/IBD group are not shown, but they express essentially the same trend as shown in the curves of the spinal group.

Table 3 shows the cutoff values, sensitivity, specificity, positive predictive value, and percentage of misclassified patients for the 3 clinical disease activity variables in both groups. Classifications of all 3 clinical disease activity measures in patients with AS, with either no disease activity or with definite disease activity, were compared with the classification according to ESR or CRP, based on the cutoff values of ROC. In general, the test characteristics sensitivity and specificity are reasonable, but the positive predictive

values — a relevant characteristic in clinical practice — are uniformly low, with large percentages of misclassified patients.

DISCUSSION

In the spinal group the majority of the patients have normal values of ESR and CRP, whereas the majority of patients in the peripheral/IBD group have elevated values of ESR and CRP. Also, the level of the ESR and CRP values is higher in the peripheral/IBD group compared to the spinal group. Thirty percent of the patients in both disease subgroups show either elevated ESR with a normal CRP or vice versa. The large majority of these cases show values just above normal in the acute phase reactant with a value outside the normal range. Especially in the spinal group many patients with AS have normal or slightly elevated values of ESR and CRP, in contrast to patients with rheumatoid arthritis⁶. These findings are comparable with the results in most other AS studies². One reason could be that disease activity especially in spinal AS is not well reflected in acute phase reactants such as ESR and CRP. The difference in the judgment of disease activity between the physician on one hand and the patients and BASDAI (also patient based) on the other hand, as reflected in quite different mean values on the same scale (0–10), is quite striking. Physicians classified only 5 patients as having disease activity ≥ 6 , contrasting with 31 and 16 according to the patient judgment and BASDAI, respectively. Also the correlations between the disease activity defined by the physician and ESR and CRP are considerably higher than those between disease activity defined by

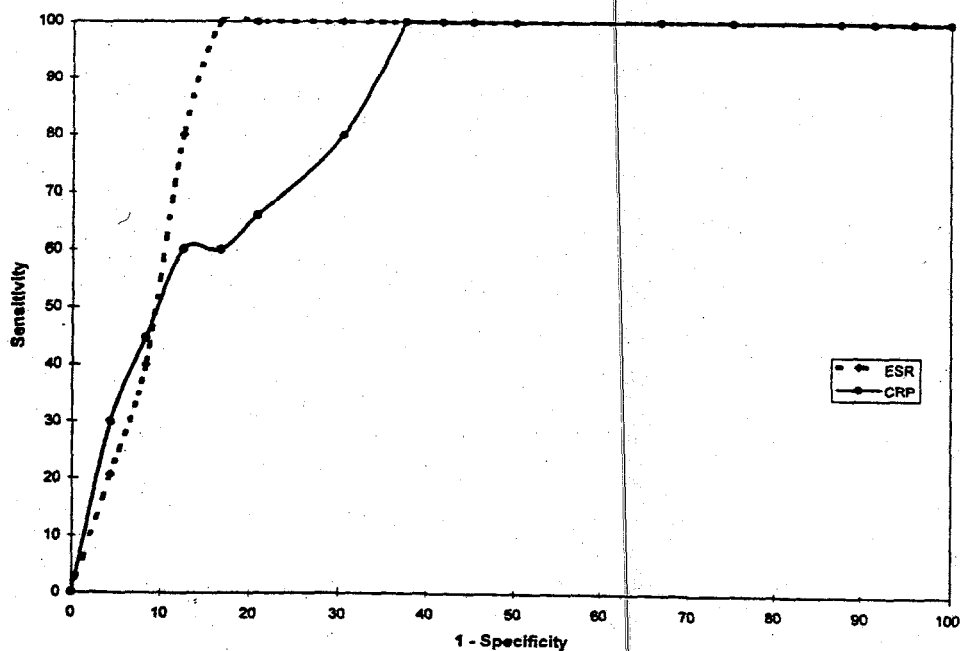


Figure 1. Receiver operator curve: ESR and CRP against physician assessment of disease activity.

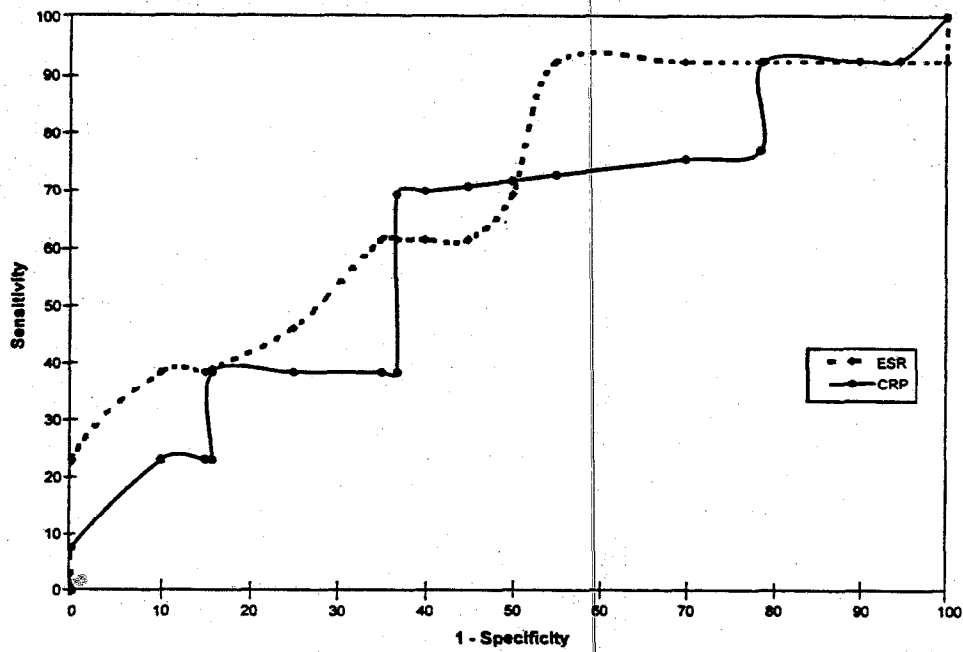


Figure 2. Receiver operator curve: ESR and CRP against patient assessment of disease activity.

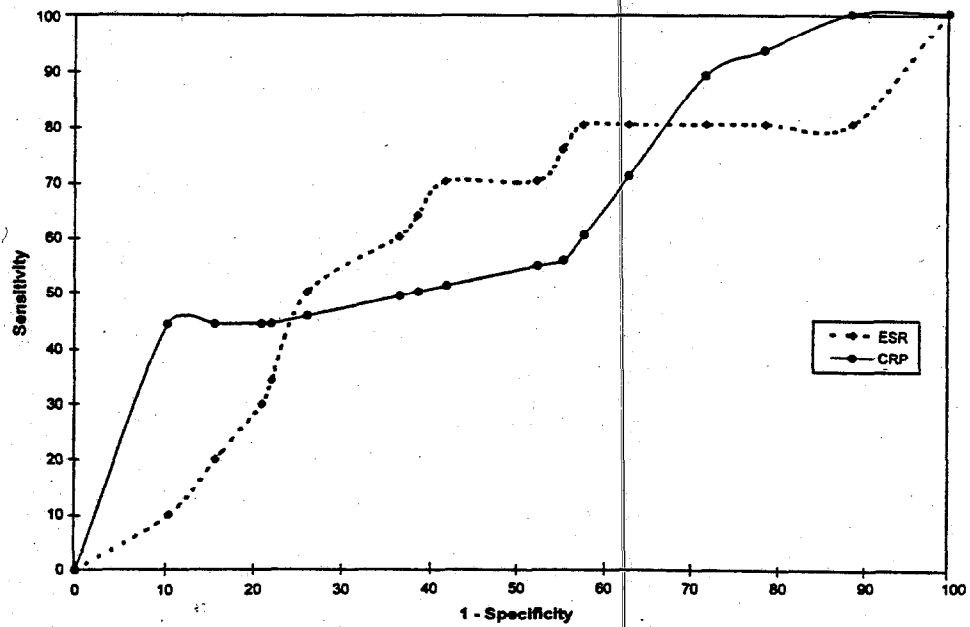


Figure 3. Receiver operator curve: ESR and CRP against Bath Ankylosing Spondylitis Disease Activity Index (BASDAI).

Table 3. Results of cutoff values from the ROC.

	Cutoff Value (%)	Specificity (%)	Sensitivity (%)	Positive Predictive Value (%)	Misclassified Patients (%)
Spinal group					
DA physician	ESR 15	77	100	15	22
	CRP 14	84	100	21	23
DA patient	ESR 15	83	55	56	24
	CRP 10	79	60	51	27
BASDAI	ESR 6	52	63	19	47
	CRP 12	81	44	30	21
Peripheral/IBD group					
DA physician	ESR 25	83	100	55	14
	CRP 15	70	100	58	25
DA patient	ESR 14	60	62	50	39
	CRP 10	63	69	44	34
BASDAI	ESR 17	58	70	46	34
	CRP 10	44	78	69	45

DA: disease activity.

the patient and BASDAI and ESR and CRP. In the peripheral/IBD group the correlations between ESR and CRP and BASDAI are virtually absent. It should be stressed that, when judging disease activity, the physician was not aware of ESR or CRP values, because blood for these assessments was taken after the visit to the physician.

The cutoff values based on the ROC are only slightly higher for ESR for the peripheral/IBD group than for the spinal group for the classification according to the physician and the BASDAI. A problem in all these sort of studies is that a gold standard for disease activity is lacking. Most studies on the comparison of ESR and CRP in AS use different definitions of disease activity². The results depend heavily on the definition used, as illustrated by this study, in which 3 definitions for disease activity were applied. Also, the disease spectrum in the sample (i.e., patients with spinal disease only and patients with extraspinal involvement) can greatly influence results.

This cross sectional study confirmed that there is no clear advantage to using either ESR or CRP in the assessment of AS. Longitudinal data are needed to evaluate whether ESR or CRP reflects fluctuation in disease activity better, and whether one of the 2 is correlated better to structural damage. There is a need for validated measures of disease activity in AS.

REFERENCES

1. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S, on behalf of the Assessment in Ankylosing Spondylitis Working Group. Preliminary core sets for endpoints in ankylosing spondylitis. *J Rheumatol* 1997;24:2225-9.
2. Ruof J, Stucki G. Validity aspects of erythrocyte sedimentation rate and C-reactive protein in ankylosing spondylitis. A literature review. *J Rheumatol* 1999;26:966-70.
3. 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.
4. 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.
5. Wolfe F. Comparative usefulness of C-reactive protein and erythrocyte sedimentation rate in patients with rheumatoid arthritis. *J Rheumatol* 1997;24:1477-85.