Outcome and disease activity in ankylosing spondylitis : an international study

Citation for published version (APA):

Spoorenberg, J. P. L. (2003). *Outcome and disease activity in ankylosing spondylitis : an international study*. [Doctoral Thesis, Maastricht University]. Universiteit Maastricht.
<https://doi.org/10.26481/dis.20040206js>

Document status and date:

Published: 01/01/2003

DOI:

[10.26481/dis.20040206js](https://doi.org/10.26481/dis.20040206js)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

OUTCOME AND DISEASE ACTIVITY IN ANKYLOSING SPONDYLITIS

AN INTERNATIONAL STUDY

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van Prof. mr. G.P.M.F. Mols,
Rector Magnificus,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op vrijdag 6 februari 2004 om 16.00 uur

door

Johanna Petronella Lena Spoorenberg
Geboren op 18 februari 1966 te Leende

Promotores

Prof. dr. J.M.J.P. van der Linden

Prof. dr. D.M.F.M. van der Heijde

Beoordelingscommissie

Prof. dr. J. Metsemakers (voorzitter)

Prof. dr. M. Dougados (Groupe Hospitalier Cchin, Paris)

Prof. dr. B. Dijkmans (Vrije Universiteit Amsterdam)

Prof. dr. J.M.A. van Engelshoven

Dr. R. Landew

 Anneke Spoorenberg, Leeuwarden 2003

ISBN 90 9017 451 6

Printed by Hellinga b.v. Leeuwarden, The Netherlands

The publication of this thesis was financially supported by: MSD bv, Pfizer bv, Amgen bv, Abbott bv, Wyeth Pharmaceuticals, Aventis Pharma bv, Alliance for better bonehealth, Schering-Plough bv.

Voor mijn ouders
Christiaan
Emma & Sterre

CONTENTS

chapter 1	Introduction	7
chapter 2	A comparative study of the usefulness of the Bath Ankylosing Spondylitis Functional Index (BASFI) and the Dougados Functional Index (D-FI) in the assessment of ankylosing spondylitis	15
chapter 3	The relative value of Erythrocyte Sedimentation Rate (ESR) and C-reactive protein (CRP) in the assessment of disease activity in ankylosing spondylitis	25
chapter 4	The reliability of self-assessed joint counts in ankylosing spondylitis	35
chapter 5	Measuring disease activity in ankylosing spondylitis: from patient and physician perspective	47
chapter 6	The development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis	61
chapter 7	Radiological scoring methods in ankylosing spondylitis: reliability and change over one and two years	79
chapter 8	Summary and general discussion	95
chapter 9	Samenvatting en discussie	111
	Dankwoord	125
	Publications	129

Chapter I

INTRODUCTION

ANKYLOSING SPONDYLITIS (AS)

Ankylosing Spondylitis or Morbus Bechterew is a chronic inflammatory rheumatic disorder that primarily effects the axial skeletal.

Sacroiliitis is the hallmark of the disease and causes the most common and characteristic clinical symptom of AS: chronic inflammatory low back pain and stiffness. This back pain is often of insidious onset and difficult to localize but most often felt in the sacroiliac region or deep in the gluteal area. The pain can be quite severe, unilateral or bilateral and may appear intermittent or more persistent. This inflammatory back pain and stiffness tend to worsen after prolonged periods of inactivity and therefore pain may wake patients from sleep. Exercise may improve symptoms of inflammatory back pain and stiffness.

Spondylitis is the process of inflammation of the vertebral ligaments and facet joints and may lead to ossification of these structures. Pain, progressive loss of spinal mobility and functional capacity in AS is caused by both spondylitis and ossification. Spinal ankylosis appears most frequently in rather late stages of the disease and may not occur at all in patients with mild disease¹.

In approximately 25% of AS patients also the peripheral joints are affected in the disease process especially the shoulders and the hips ('root'-joints)². Patients may experience arthritis which causes pain, swelling, stiffness and loss of mobility of the affected joints. Enthesitis is also a frequent symptom. Typical localizations are at the achilles tendon, plantar fascia, pelvis and thorax.

Besides these manifestations patients may experience extra-articular features such as acute anterior uveitis, and inflammatory bowel disease. Clinical manifestations usually begin in late adolescence or early adulthood and onset after age 45 is rare.

The prevalence of AS in the European population has been reported in several studies to be approximately 0.1% but a recent study using magnetic resonance imaging techniques estimated a much higher prevalence of 0.86%³. As many males as females suffer from AS but in female patients the symptoms are often mild and the disease may not be detected². Therefore often three times as many males as females are included in most clinical AS studies.

AS is the most important disease of a specific group rheumatic disorders: the spondylarthropathies (SpA). These rheumatic disorders are classified together because of several common features such as inflammatory back pain (sacroiliitis, spondylitis), asymmetric peripheral arthritis predominantly of the lower limbs and the absence of rheumatoid factor. Other important extra-articular manifestations of SpA are acute anterior uveitis, inflammatory bowel disease, psoriasis, carditis, conjunctivitis, dactylitis, inflammatory bowel disease, balanitis and other genital inflammation. Furthermore there is a strong association of SpA with the class-I molecule HLA-B27⁴. The strongest association with the histocompatibility antigen HLA-B27 is found for AS which is demonstrated in about 90% of cases. The prevalence of HLA-B27 in the general European population is approximately 7 %. For a first degree HLA-B27 positive relative of a HLA-B27 positive AS patient the chance to develop AS is up to 30%².

OUTCOME VS DISEASE ACTIVITY IN AS

Measuring disease activity and disease progression (outcome) in AS is quite challenging. Progression of AS varies in rate and patterns, often independent of the degree in which observed symptoms such as pain and stiffness appear. There is a lot of variety in the clinical picture of AS therefore it is difficult to picture the whole spectrum of this disease in a comprehensive way. In contrast to the situation in rheumatoid arthritis, laboratory indicators of disease activity reflect neither clinical activity nor radiological progression and their use in AS is controversial⁵. Furthermore different instruments are used to follow the disease process in AS but often these instruments are not validated and therefore it is difficult to compare clinical studies in AS. Because of lack of information on validation of these instruments it is difficult to prefer one AS instrument over another.

In 1995 an international working group on Assessment in Ankylosing Spondylitis (ASAS) was formed. In 1997, 'core sets' for the following three settings were defined: disease controlling anti-rheumatic therapy (DC-ART), symptom modifying anti-rheumatic drugs (SM-ARD)/physical therapy and clinical record keeping⁶. The domains for all three core sets are physical function, pain, spinal mobility, spinal stiffness and patient global assessment. The core sets for clinical record keeping and DC-ART were extended with the domains acute phase reactants, peripheral joints and entheses. The core set of DC-ART includes also in addition spine and hip radiographs and fatigue. In a recent update of the core set, it was decided that the domain fatigue should be included in all three core sets. In 1999 this ASAS working group selected specific instruments for each core set according to the 'OMERACT (outcome measures in rheumatoid arthritis clinical trials) filter test' for relevance and feasibility^{7,8}. For each domain one or more instruments were selected. This selection procedure was undertaken to diminish the large number of assessments to create uniformity and comparability in AS clinical trials. There are several outstanding decisions on the optimal instruments for each domain. This should be based on data of reliability and sensitivity to change. To achieve this the remaining aspects of the 'OMERACT filter' should be used concerning truth (validity) and discrimination (reproducibility and responsiveness)⁸.

AIM AND DESIGN OF THE STUDY

As a follow-up of the work of the ASAS working group the present study focuses on different aspects of the assessment of disease activity and outcome in AS.

The objectives of this study were: (1) assess the validity of several widely used instruments in the follow up of AS concerning physical function and laboratory tests focusing on the acute phase reaction, (2) develop and assess the reliability of self assessed joint counts in AS, (3) investigate on which criteria AS patients and rheumatologists base their opinion on disease activity, (4) develop and validate a questionnaire designed to measure quality of life in AS and (5) assess reliability and change over time of existing AS radiological scoring methods.

To study these different objectives we started the Outcome in Ankylosing Spondylitis International Study (OASIS) in October 1996 conducted in Maastricht/Sittard (the Netherlands), Ghent (Belgium) and Paris (France). This multicenter open observational study in AS patients included a total of 217 consecutive outpatients who satisfied the modified New York criteria⁹. 137 of these AS patients were included in the Netherlands, 25 in Belgium and 55 in France. This OASIS study population concerns a cross sectional cohort of AS patients, followed longitudinally. During the first two years of the study an extended physical examination and laboratory assessments were done every six months. During this period in Belgium and the Netherlands the same physician and in France the same research nurse performed all physical examinations of each individual patient. Furthermore patients were asked to complete an extended package of self assessment questions concerning stiffness, pain, disease activity and health status such as disability and impairment biannually. Radiological assessments were based on yearly visits. Almost all manuscripts from this thesis are based on data derived from this OASIS study.

The first part of the thesis (chapter 2-5) highlights aspects of disease activity in AS.

In **Chapter 2** two widely applied specific physical function indexes used in the evaluation of AS are compared. The Dougados Functional Index (D-FI) was developed in France and the Bath Ankylosing Spondylitis Functional Index (BASFI) in the United Kingdom. These indexes were published in 1988 and 1994 respectively^{10,11}. The BASFI consists of 10 questions on a VAS, all questions deal with activities of daily living. The final score is the average of the scores of the 10 items. The D-FI consists of twenty 5-point Likert response items, assessing the ability to perform distinct daily activities. The total score (ranging from 0-40) is calculated as the sum of the item scores. Physical function in patients with AS is both related to disease activity and damage. In this study we compared the relation between these two functional indexes specific for AS with both disease activity and damage assessments specific for AS. The aim of the study was to investigate if one of these instruments performed better with respect to the various aspects of validity and should therefore be preferred.

Chapter 3 concerns the relative value of two laboratory assessments, erythrocyte sedimentation rate (ESR, mm/h) and C-reactive Protein (CRP, mg/l) in the assessment of disease activity in AS. The mean values of these two acute phase reactants are considerably lower in AS than in rheumatoid arthritis¹². As there might exist differences with respect to ESR and CRP in AS patients with only spinal involvement and those with active peripheral arthritis and/or inflammatory bowel disease, we divided the patients into these two groups for our analyses. Since there is no gold standard for disease activity in AS we studied the relation of CRP and ESR with three substitute clinical variables; physician assessment of disease activity and patient assessment of disease activity both on a Visual Analogue Scale (VAS range: 0 = inactive, 10 = extremely active) and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)¹³. This index contains six VAS format questions on fatigue, pain of the spine, pain and/or swelling of the peripheral joints, pain related to enthesitis, and duration and severity of morning stiffness of the spine. The total score ranges from 0 to 10. Another

aim was to study if elevated CRP or ESR are predictive for active disease defined by the three selected disease activity variables in AS.

Chapter 4 presents the reliability of patient self-assessment of swollen and painful joints marked on a mannequin. In AS a minor part of the population has peripheral arthritis ($\pm 25\%$). Traditionally, either a physician or a well-trained health care professional is involved in the clinical assessment of arthritis. If joint counts assessed by a physician could be replaced by self-assessed joint counts, this would lighten the task for rheumatologists in practice and researchers in particular. This would also make it easier to collect these data more frequently and even with postal questionnaires. The reliability of patient reported joint counts for swelling and pain are frequently studied in the assessment of disease activity in rheumatoid arthritis^{14,15,16,17,18,19,20}. However the reliability of patient self-reported joint counts in AS has never been reported.

Chapter 5 describes on which criteria AS patients and rheumatologists base their judgment on disease activity. In AS it is especially difficult to define disease activity because there is much variety in the clinical picture among different patients. Patients may only have axial involvement in all degrees of severity, but may also have extraspinal manifestations. The clinical diversity, both in severity and in localization, makes a high demand on the instruments that are supposed to measure disease activity. Furthermore AS patients and rheumatologists seem to have different understandings about active disease and at the moment no gold standard is available for measuring disease activity in AS²¹. Disease activity from the patient perspective as well as from the physician perspective was analysed by dichotomising both patient and physician global disease activity score on a VAS (VAS range: 0 not active and 10 extremely active) into 'high disease activity' and 'low disease activity'. Various AS instruments selected by the ASAS working group were assessed every six months for two years. Data reduction by principal components analysis (PCA) was performed to distinguish factors capturing correlated instruments. Discriminant analysis with the factor loadings was performed to discriminate between a low-and a high disease activity state for both patient and physician perspective of disease activity. Multiple regression analysis on the discriminant scores was performed to prioritise the instruments with respect to their contribution to each disease activity perspective. Our aim was to explore differences between the perspective of the patient and the perspective of the physician with respect to AS disease activity.

Chapter 6 and 7 focus on outcome measures in AS.

Chapter 6 describes the development of the Ankylosing Spondylitis Quality of Life questionnaire (ASQoL). There is a growing interest in the assessment of quality of life (QoL), particularly in chronic disabling conditions and it is becoming relatively common to measure QoL in studies designed to assess the impact of new pharmaceutical products or to compare different treatment regimes. AS instruments currently available focus predominantly on physical impairment and/or physical functioning. Such instruments include: the Bath Ankylosing Spondylitis Functional Index (BASFI)¹¹, the D-FI¹⁰, a version of the Stanford Health Assessment Questionnaire modified for the spondylarthropathies (HAQ-S)²² and a

modified version of the Arthritis Impact Measurement Scales II specific to AS (AS-AIMSII)²³. Generic health status instruments such as the Nottingham Health Profile, SF-36 and EuroQoL^{24,25,26} are available but no disease specific instrument exists for assessing quality of life (QoL) in AS patients. The ASQoL is a quality of life instrument specific to AS and was developed in parallel in the United Kingdom and the Netherlands and the content was derived from interviews with patients in each country. The development methodology used combines the theoretical strengths of the needs-based quality of life model²⁷ with the statistical and diagnostic power of the Rasch model²⁸. Our aim was to produce a valid and reliable AS-specific QoL measure that would be relevant and acceptable to respondents.

In **chapter 7** available radiological scoring methods in AS are studied. Radiological damage is considered an important outcome in AS⁷. The evaluation of radiological change proves to be very difficult. Changes of the sacroiliac joints are most frequently scored using the 5 grade New York criteria²⁹ or the nearly similar SI score of the Stoke Ankylosing Spondylitis Spine Score (SASSS)^{30,31}. To evaluate the lumbar and cervical spine in AS there are essentially two different scoring methods. The Bath Ankylosing Spondylitis Radiology Index (BASRI) is a global graded scoring method, quick and easy to perform^{32,33,34} and was also developed to score the hips³⁵. The several BASRI scores are also combined in composite scores^{33,34}. The SASSS for the spine is a more detailed scoring method assessing different features such as squaring, sclerosis and erosions at various locations of each vertebra^{31,36}. Our aim was to compare all these available AS radiological scoring methods for reliability and change over one and two years.

Finally, this thesis ends with a summary and general discussion provided in English and Dutch in **chapter 8** and **9** respectively.

REFERENCES

- 1 Khan MA. Ankylosing Spondylitis: Clinical Features. In: Klippel JH, Dieppe PA. *Rheumatology, Second Edition* 1998; 6: 16.1-16.10.
- 2 van der Linden Sj, van de Heijde D. Ankylosing Spondylitis. *Clinical features. Rheum Dis Clin North Am* 1998; 24: 663-76.
- 3 Braun J, Bollow M, Remlinger G, Eggens U, Rudwaleit M, Distler A, Sieper J. Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors. *Arthritis Rheum* 1998; 41: 58-67.
- 4 Wordsworth P. Genes in the spondylarthropathies. *Rheum Dis Clin North Am* 1998; 24: 845-63.
- 5 Taylor HG, Wardle T, Beswick EJ, Dawes PT. The relationship of clinical and laboratory measurements to radiological change in AS. *Br J Rheumatol* 1991; 30: 330-5.
- 6 van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden Sj. Preliminary core sets for endpoints in ankylosing spondylitis. *J Rheumatol* 1997; 24: 2225-9.
- 7 van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden Sj, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. *J Rheumatol* 1999; 26: 952-4.
- 8 Boers M, Brooks P, Strand CV, Tubwell P. The OMERACT filter for Outcome in Measures in Rheumatology [editorial]. *J Rheumatol* 1998; 25: 198-9.
- 9 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.
- 10 Dougados M, Gueguen A, Nakache JP, Nguyen M, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988; 15: 02-7.
- 11 Calin A, Garrett S, Whitelock H, Ken J, Mallorie P, Jenkinson T. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994; 21: 2281-5.
- 12 Wolfe F. Comparative usefulness of C-reactive protein and erythrocyte sedimentation rate in patients with rheumatoid arthritis. *J Rheumatol* 1997; 24: 1477-85.
- 13 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.
- 14 Stewart MW, Palmer DG, Knight RG. A self-report articular index measure of arthritic activity: investigations of reliability, validity and sensitivity. *J Rheumatol* 1990; 17: 1011-5.
- 15 Prevoo ML, Kuper IH, van't Hof MA, van Leeuwen MA, van de Putte LB, van Riel PL. Validity and reproducibility of self-administered joint counts. A prospective longitudinal follow up study in patients with rheumatoid arthritis. *J Rheumatol* 1996; 23: 841-5.
- 16 Hanly JG, Mosher D, Sutton E, Weerasinghe S, Theriault D. Self-assessment of disease activity by patients with rheumatoid arthritis. *J Rheumatol* 1996; 23: 1531-8.
- 17 Mason JH, Anderson JJ, Meenan RF, Haralson KM, Lewis Stevens D, Kaine JL. The rapid assessment of disease activity in rheumatology (RADAR) questionnaire. Validity and sensitivity to change of a patient self-report measure of joint count and clinical status. *Arthritis Rheum* 1992; 35: 156-62.
- 18 Abraham N, Blackmon D, Jackson JR, Bradley LA, Lorish CD, Alarcon GS. Use of self-administered joint counts in the evaluation of rheumatoid arthritis patients. *Arthritis Care Res* 1993; 6: 78-81.

- 19 Calvo FA, Calvo A, Berrocal A, Pevez C, Romero F, Vega E, et al. Self-administered joint counts in rheumatoid arthritis: Comparison with standard joint counts. *J Rheumatol* 1999; 26: 536-39.
- 20 Stucki G, Stucki S, Bruhlmann P, Maus S, Michel BA. Comparison of the validity and reliability of self-reported articular indices. *Br J Rheumatol* 1995; 34: 760-6.
- 21 Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van de Tempel H, van der Linden SJ. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. *J Rheumatol.* 1999; 26: 980-4.
- 22 Daltroy LH, Larson MG, Liang MH. A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. *J Rheumatol* 1990; 17: 946-950.
- 23 Guillemin F, Chailier B, Urlacher F, Vançon G, Pourcel J. Quality of life in ankylosing spondylitis: Validation of the ankylosing spondylitis Arthritis Impact Measurement Scales II, a modified Arthritis Impact Measurement Scales questionnaire. *Arthritis Care Res* 1999; 12: 157-162.
- 24 Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. *J Epidemiol Community Health* 1980; 34: 281-6.
- 25 Brazier JE, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *Br Med J* 1992; 305: 160-4.
- 26 The EuroQol Group. EuroQol-a new facility for the measurement of health related quality of life. *Health Policy* 1990; 16: 199-208.
- 27 Hunt SM, McKenna SP. The QLDS: A scale for the measurement of quality of life in depression. *Health Policy* 1992; 22: 307-19.
- 28 Rasch G. Probabilistic Models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1980.
- 29 Dale K. Radiographic gradings of sacroiliitis in Bechterews syndrome and allied disorders. *Scand Rheumatol* 1979; 32: 92-7.
- 30 Taylor HG, Wardle T, Beswick EJ, Dawes P. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. *Br J Rheum* 1991; 30: 330-5.
- 31 Dawes PT. Stoke Ankylosing Spondylitis Spine Score. *J Rheumatol* 1999; 26: 993-6.
- 32 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 Rheum* 1995; 34: 767-70.
- 33 MacKay K, Mack C, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI). A new validated approach to disease assessment. *Arthritis Rheum* 1998; 41: 2263-70.
- 34 Calin A, Makay k, Santos H, Brophy S. A new dimension to outcome. Application of the Bath Ankylosing Spondylitis Radiology Index. *J Rheumatol* 1999; 26: 988-92.
- 35 MacKay K, Brophy S, Mack C, Doran M, Calin A. The development and validation of a radiographic grading system for the hip in Ankylosing Spondylitis: the Bath Ankylosing Spondylitis Radiology Hip Index. *J Rheumatol* 2000; 27: 2866-72.
- 36 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.

Chapter 2

A COMPARATIVE STUDY OF THE USEFULNESS OF THE BATH ANKYLOSING SPONDYLITIS FUNCTIONAL INDEX AND THE DOUGADOS FUNCTIONAL INDEX IN THE ASSESSMENT OF ANKYLOSING SPONDYLITIS

Anneke Spoorenberg, Désirée van der Heijde, Erik de Klerk, Maxime Dougados, Kurt de Vlam, Herman Mielants, Hille van de Tempel, Sjeef van der Linden.

Journal of Rheumatology 1999; 26: 961-5.

ABSTRACT

Objective: To determine whether the Bath Ankylosing Spondylitis Functional Index (BASFI, score 0-10) or Dougados Functional Index (D-FI, score 0-40) is superior in measuring physical function in Ankylosing Spondylitis (AS).

Methods: We studied 191 consecutive outpatients with AS in the Netherlands, France and Belgium. The participating centers are secondary and tertiary referral centers. The external criterion for disease activity (DA) was: both patient and physician assessment of disease activity on a Visual Analogue Scale (VAS) and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI). The external criterion for damage were 2 radiological scores of the spine; BASRI-s (Bath Ankylosing Spondylitis Radiology Index-spine) and a modified SASSS (Stoke Ankylosing Spondylitis Spine Score).

Results: Median scores for BASFI and D-FI were 2.5 (range: 0-10) and 8.5 (range: 0-35) respectively. Spearman correlation coefficient between both indexes was 0.89. The average correlation with disease activity variables was 0.42 for BASFI and 0.41 for D-FI. For both BASFI and D-FI the correlation with BASRI-s was 0.42 and with SASSS 0.36. When distinguishing between patients with high and low disease activity, sensitivity for both indexes was between 76% and 94% while specificity was between 66% and 87% for all three DA measures. Average misclassification between BASFI, D-FI and DA was 23% and 27% respectively.

Conclusion: Both BASFI and D-FI correlate equally well with disease activity and damage.

INTRODUCTION

Two indexes are frequently used to evaluate physical function in ankylosing spondylitis (AS): the Dougados Functional Index (D-FI) and the Bath Ankylosing Spondylitis Functional Index (BASFI) published in 1988 and 1994 respectively^{1,2,3}. As physical function is both related to disease activity and damage, we compared the relation between these two functional indexes and both disease activity and damage. The international Assessment in Ankylosing Spondylitis (ASAS) working group has included physical function, assessed by D-FI or BASFI, in the core sets for endpoints for all settings in AS⁴. Ruof and Stucki reviewed the literature on the comparative usefulness of both indexes in this issue⁵. Based upon the literature data, no definite preference could be identified, although there seemed to be a slight preference for the BASFI. One problem in the evaluation of the D-FI was that data were only available with the answers to the items on a three point scale, not on a five point Likert scale. In this study we used a modification of the D-FI on the 5 point Likert scale. The goal of our cross-sectional study was to determine whether one of the two functional indexes is superior.

PATIENTS AND METHODS

We included 191 consecutive AS outpatients of the University Hospital Maastricht (the Netherlands), the Maasland Hospital Sittard (the Netherlands), the University Hospital Ghent (Belgium) and the Hôpital Cochin in Paris (France). These centers are secondary and tertiary referral centers.

All patients fulfilled the modified New York criteria for AS⁶. Since physical function reflects both disease activity and damage the indexes were compared with measures of clinical disease activity and also with measures of damage.

Because there is no gold standard for disease activity in AS we used three clinical variables as external criterion for disease activity; physician assessment of disease activity on a Visual Analogue Scale (VAS, anchored 'no disease activity' at 0 cm and 'very severe activity' at 10 cm), patient assessment of disease activity (VAS, anchored 'no disease activity' at 0 cm and 'very severe activity' at 10 cm) and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)⁷. This index contains six VAS format questions, related to: 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 of the BASDAI ranges from 0 to 10. Three levels of disease activity were chosen, using the same cutoff values for all three disease activity variables: a score of ≤ 4 meant no active disease, a score between 4 and 6 disease activity was labeled ambiguous, and a score of ≥ 6 indicated definite disease activity. In our analysis we used only the most contrasting groups where disease activity was defined as definite or no disease activity.

Further, we used two radiological damage measures as external criterion for damage. The Bath Ankylosing Spondylitis Radiology Index spine (BASRI-s) grades changes in the

sacroiliac (SI) joints, the lumbar spine and the cervical spine from 0-4, resulting in a scale ranging from 2 to 12⁸. The Stoke Ankylosing Spondylitis Spine Score (SASSS) is used to assess more detailed changes in the lumbar spine both at anterior and posterior sites (range 0 to 72)⁹. Creemers et al developed a modified version that scores the anterior sites of the lumbar and also of the cervical spine¹⁰. We used this modified SASSS as this has shown the best reliability in another study¹¹.

The BASFI consists of 10 questions on a VAS, all questions deal with activities of daily living. The final score is the average of the scores of the 10 items. The version of the D-FI we have applied consists of twenty 5-point Likert response items, assessing the ability to perform distinct daily activities. The total score (ranging from 0-40) is calculated as the sum of the item scores. The BASFI and D-FI were completed by the patients in randomized order (half of the patients completed BASFI first, other half completed D-FI first).

To deal efficiently with missing values of the D-FI, we assigned the average of the remaining items to the missing values if no more than 3 items were missing, concordant with the instructions of the D-FI authors. No corrections were applied to the BASFI.

Statistical analyses

Receiver Operator Curves (ROCs) were plotted for both functional indexes and the three clinical measures of disease activity. A ROC is a curve of sensitivity (or true positivity) on the vertical axis and 1-specificity (or false positivity) on the horizontal axis¹². The cutoff value of BASFI and D-FI 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 BASFI or D-FI.

The percentage of patients incorrectly classified according to the cutoff values of the functional indexes related to clinical disease activity were also computed. For correlations we used Spearman correlation coefficients.

RESULTS

There were 191 AS patients in our study, two times more males than females. *Table 1* shows striking differences between duration of complaints and disease duration (defined as years since diagnosis), indicating that AS patients have complaints long before the diagnosis is made.

It is of interest that the median physician based disease activity is only 1.5, while the scores for patient assessment of disease activity and BASDAI were somewhat higher: 3.6 and 3.9, respectively. Furthermore the BASRI-s appears to give a relatively higher score for radiological damage than the SASSS (*Table 1*).

A total of 7 BASFI questionnaires showed one or more missing answers versus 28 of the D-FI questionnaires. After correction for missing values 24 D-FI questionnaires could be calculated and analysed.

Table 1. Study variables (median, range, 191 patients)

	median (%)		range
Demographics:			
sex (M : F)	2 : 1		
Age (years)	43		18 - 78
Duration of complaints (years)	17.9		0.3 - 54.4
Disease duration	9.4		0.1 - 41
Clinical Disease Activity (DA) variables:			
DA physician (0 - 10)	1.7		0 - 9.6
≤4	147	(77%)	
4.1-5.9	22	(12%)	
≥6	10	(5%)	
missing	12	(6%)	
DA patient (0 - 10)	3.9		0 - 10.0
≤4	100	(52%)	
4.1-5.9	45	(24%)	
≥6	44	(23%)	
missing	2	(1 %)	
BASDAI (0 - 10)	3.7		0 - 9.4
≤4	106	(55%)	
4.1-5.9	53	(28%)	
≥6	26	(14%)	
missing	6	(3%)	
Damage variables:			
BASRI-s (2 - 12)	7.0		3 - 12
Modified SASSS (0- 72)	12.0		3 - 65
Functional Indexes:			
BASFI score (0 - 10)	2.5		0 - 10
D-FI score (0 - 40)	8.5		0 - 35

As expected both functional indexes were highly correlated (*Table 2*, see *page 20*). Correlation coefficients for both functional indexes and clinical disease activity and damage measures were comparable. Both showed a higher correlation with the BASDAI than with the other two clinical disease activity measures.

Figures 1 to 3 (see *page 21*) show the ROC of the two functional indexes and all three clinical disease activity variables. *Figure 3* shows the best curves with the highest sensitivity and specificity values for both functional indexes and the BASDAI with no real difference between the functional indexes. The plot lines of *Figures 1 and 2*, illustrating the two functional indexes and the patient or physician assessment of disease activity, are similar. In these curves

Table 2. Spearman correlation coefficients

	BASFI	D-FI
BASFI	****	0.89
D-FI	0.89	****
Clinical Disease Activity variables:		
DA physician	0.33	0.36
DA patient	0.33	0.32
BASDAI	0.59	0.57
Damage variables:		
BASRI-s	0.42	0.42
Modified SASSS	0.36	0.36

sensitivity and specificity values for the cutoff points are somewhat lower than for the BASDAI versus the D-FI and the BASFI (Table 3).

The cutoff values found with the ROC are provided in Table 3. The cutoff values for BASFI are consistently higher. Nevertheless, there are no real differences in specificity and sensitivity values for the cutoff points for both BASFI and D-FI. These values are uniformly rather high but in most cases there is still a large proportion (12-30%) of patients who are misclassified for disease activity by these two functional indexes. The proportion of misclassified patients is lowest (12%) for the cutoff value of the BASFI and clinical disease activity measured with the BASDAI.

Table 3. Results of cutoff values from the ROCs

	Cutoff values	specificity	sensitivity	misclassified patients
DA physician	BASFI 4.5	73%	80%	30%
	D-FI 10	70%	80%	30%
DA patient	BASFI 3.5	73%	76%	26%
	D-FI 8	66%	77%	31%
BASDAI	BASFI 4.5	87%	94%	12%
	D-FI 10	79%	93%	19%

Figure 1. Receiver Operator Curve of functional indexes: Bath Ankylosing Spondylitis Functional Index (BASFI) and Dougados Functional Index (DFI) versus Bath Ankylosing Spondylitis Disease Activity Index (BASDAI).

Figure 2. Receiver Operator Curve of functional indexes: Bath Ankylosing Spondylitis Functional Index (BASFI) and Dougados Functional Index (DFI) versus patient assessment of disease activity

Figure 3. Receiver Operator Curve of functional indexes: Bath Ankylosing Spondylitis Functional Index (BASFI) and Dougados Functional Index (DFI) versus physician assessment of disease activity.

DISCUSSION

Both BASFI and D-FI are validated and widely used instruments to assess physical function in patients with AS. The differences between the two have been described in detail, but no final choice could be made⁵. As physical function is a reflection of both disease activity and damage, we correlated the BASFI and the D-FI with both clinical disease activity measures and with radiographic damage measures. Because no gold standard is available to assess disease activity, we used three disease activity variables: disease activity assessed by the physician, disease activity assessed by the patient and using the BASDAI. A global grading system including SI joints, cervical and lumbar spine (BASRI-s), and a detailed scoring system of the lumbar and cervical spine (modified SASSS) were used to assess structural damage. Overall, both functional indexes correlate highly with each other and correlate about equally with the disease activity measures and the damage measures. The cutoff values to determine high versus low disease activity were for the BASFI considerably higher (around 35-45% of the full scale) compared to the D-FI (around 20% of the full scale). Using these cutoff values shows that a considerable percentage of patients will be misclassified as having low or high disease activity if solely based on their functional index scores. The whole scale was used for BASFI, whereas 35 out of a maximum of 40 was the highest score for the D-FI. More D-FI questionnaires (28) showed one or more question unanswered, compared to BASFI (7). Our study has important limitations. Sensitivity to change is an important aspect of a disease activity measure. This however, can not be analyzed with our cross-sectional data. Therefore, longitudinal data are needed. Such data will be collected during long term follow up of the patients in this ongoing observational study. Based upon the results of the present study, no definite choice can be made between BASFI and D-FI neither for the assessment of disease activity nor to approximate structural damage of the spine.

REFERENCES

- 1 Dougados M, Gueguen A, Nakache JP, Nguyen M, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988; 15: 302-7.
- 2 Dougados M, Gueguen A, Nakache JP, Nguyen M, Amor B. Evaluation of a functional index in ankylosing spondylitis. (letter) *J Rheumatol* 1990; 17: 1254-5.
- 3 Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, Jenkinson T. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994; 21: 2281-5.
- 4 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.
- 5 Ruof J, Stucki G. Comparison of the Dougados Functional Index (D-FI) and the Bath Ankylosing Spondylitis Functional Index (BASFI) - a literature review. *J Rheumatol* 1999; 26: 955-60.
- 6 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.
- 7 Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol* 1994; 21: 2286-91.
- 8 Calin A, Mackay K, Santos H, Brophy S. A new dimension to outcome: the Bath Ankylosing Spondylitis Radiology Index (BASRI). *J Rheumatol* 1999; 26: 988-92.
- 9 Taylor HG, Wardle T, Beswick EJ, Dawes PT. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. *Br J Rheumatol* 1991; 30: 330-5.
- 10 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
- 11 Spoorenberg A, de Vlam K, van der Heijde D et al. Radiological scoring methods in Ankylosing Spondylitis: reliability and sensitivity to change over one year. *J Rheumatol* 1999; 26: 997-1002.
- 12 Wolfe F. Comparative usefulness of C-Reactive Protein and Erythrocyte Sedimentation Rate in patients with rheumatoid arthritis. *J Rheumatol* 1997; 24: 1477-85.

Chapter 3

THE RELATIVE VALUE OF ERYTHROCYTE SEDIMENTATION RATE AND C-REACTIVE PROTEIN IN THE ASSESSMENT OF DISEASE ACTIVITY IN ANKYLOSING SPONDYLITIS

Anne Spoorenberg, Désirée van der Heijde, Erik de Klerk, Maxime Dougados, Kurt de Vlam, Herman Mielants, Hille van de Tempel, Sjef van der Linden.

Journal of Rheumatology 1999; 26: 980-4.

ABSTRACT

Objective: 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).

Methods: 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 analogue scale (VAS) and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI). In each measure we defined three levels of disease activity: no disease activity, ambiguous activity, and definite disease activity. The patients with AS (modified New York criteria) were divided into two 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.

Results: 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 three disease activity variables (0.06 - 0.48). Sensitivity for both ESR and CRP was 100% for physician assessment of disease activity 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).

Conclusion: We concluded that neither CRP nor ESR is superior to assess disease activity.

INTRODUCTION

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 (ASAS) 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.

Cross-sectional data on the comparison of ESR and CRP show that they are highly correlated. The mean values of these two 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 trials. 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 a 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 Ghent (Belgium) and the Hôpital Cochin in 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 AS patients with only spinal involvement and those with active peripheral arthritis and/or inflammatory bowel disease (IBD) may exist, we divided the patients into these two groups. Active peripheral arthritis was defined as synovitis of at least one large joint (wrist, elbow, shoulder, hip, knee, ankle) or three or more small joints (hand- and feet joints, sternoclavicular joints). Because there is no gold standard for disease activity in AS we used three substitute clinical variables. Our first choice was physician assessment of disease activity on a 10 cm horizontal Visual Analogue Scale (VAS) anchored 'no disease activity' at 0 cm and 'very severe activity' at 10 cm. The two 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 six 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 three levels of disease activity for all three 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 only used the two 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 three clinical disease activity variables in both 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-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 three 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 AS patients with spinal involvement only and 42 patients with active peripheral arthritis and/or IBD. In both groups the male/female ratio was 2:1. The median duration of complaints and disease duration was higher in the peripheral group. There was also a broad range of disease duration (*Table 1*). For the three 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*, see page 28). Also more patients in the spinal groups showed ESR and CRP values in the normal range: 55% and 62% respectively, compared to the peripheral/IBD group (38% and 39%, respectively). This was significantly different between the spinal and peripheral/IBD

Table 1: Study variables (median, range)

	Spinal group		peripheral	
Number	149		42	
Demographics:				
Sex (male: female)	2 : 1		2 : 1	
Age	40.4 yr. (19 - 77)		47.6 yr. (22 - 78)	
Duration of complaints	16.9 yr. (0.3 - 52.4)		24 yr. (2 - 53.9)	
Disease duration	9.4 yr. (0.3 - 40.7)		10.8 yr. (2.5 - 53.9)	
Clinical Disease Activity (DA) 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/hr	13* (1-118)		21* (3-80)	
CRP mg/l	16 (0-125)		25 (0-139)	

* p<0.02 unpaired t-test

groups for CRP but not for ESR (chi-squared, p=0.01 and p=0.06, respectively). In the spinal group 18% of the patients showed an elevated ESR and a normal CRP and 12% an elevated CRP and a 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, see page 30). The correlations of the two acute phase reactants

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

and the three clinical disease activity were considerably lower 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/hr 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 but 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

Table 3: Results of cutoff values from the ROCs

spinal group				positive predictive value	misclassified patients
	cutoff value	specificity	sensitivity		
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%

Figure 1. Receiver Operator Curve: ESR and CRP versus physician assessment of disease activity.

Figure 2. Receiver Operator Curve: ESR and CRP versus patient assessment of disease activity.

Figure 3. Receiver Operator Curve: ESR and CRP versus Bath Ankylosing Spondylitis Disease Activity Index (BASDAI).

activity and BASDAI in the spinal group (Figure 2 and 3, see page 31) - are more or less identical with low cutoff values and no obvious difference between ESR and CRP. However, AUC are substantial smaller when compared with Figure 1 (see page 31) 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 (see page 30) shows the cutoff values, sensitivity, specificity, positive predictive value and percentage of misclassified patients for the three clinical disease activity variables in both groups. Classifications of all three clinical disease activity measures in AS patients, 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 the 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 the 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 an 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 AS patients have normal or slightly elevated values of ESR and CRP which is in contrast to patients with rheumatoid arthritis⁵. These findings are comparable with the results found in most other AS studies². One reason could be that disease activity in especially 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 at one hand and the patient and BASDAI (also patient based) on the other hand as reflected in quite different mean values on the same (0-10) scale is quite striking. Physicians classified only 5 patients as having definite disease activity ($VAS \geq 6$), contrasting with 31 and 16 according to the patients' 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 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 the DA, the physician was not aware of the ESR or CRP value, because blood for these assessments was taken after the visit to the physician.

The cut-off 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 this type 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 three 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 influence the results greatly.

This cross-sectional study confirmed that there is no clear advantage to use either ESR or CRP in the assessment of AS. Longitudinal data are needed to evaluate whether ESR or CRP reflect fluctuation in disease activity better and whether one of the two 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 (ESR) and C-reactive protein (CRP) in ankylosing spondylitis a literature review. *J Rheumatol* 1998; this issue.
- 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* 1999; 26: 966-70.
- 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.

Chapter 4

THE RELIABILITY OF SELF-ASSESSED JOINT COUNTS IN ANKYLOSING SPONDYLITIS

Anneke Spoorenberg, Désirée van der Heijde, Maxime Dougados, Kurt de Vlam, Herman Mielants, Hille van de Tempel, Sijf van der Linden.

Annals of Rheumatic Diseases 2002; 61: 799-803.

ABSTRACT

Objective: To determine the reliability of self-reported joint counts to assess pain or swelling in ankylosing spondylitis (AS).

Methods: 217 outpatients with AS fulfilling the modified New York criteria were asked to mark painful and swollen joints on two mannequins presenting 44 and 40 joints respectively. A doctor or research nurse assessed the same joints for pain and swelling at the same day, after completion by the patient, without information on the results of the patient's assessment.

Results: 21% of the patients reported ≥ 1 swollen joint (mean number of swollen joints 0.5, range 0-8); the doctor found ≥ 1 swollen joint in 54 (25%) of the patients (mean number of swollen joints 0.8, range 0-31). The overall agreement on the number of swollen joints between patients and doctor was moderate (intraclass correlation coefficient (ICC) 0.53). Agreement on individual swollen joints was poor to moderate (kappa 0.1-0.64).

128 (60%) reported tender joints (mean number of painful joints 2.4, range 0-26). The doctors reported ≥ 1 tender joint in 50% of the patients (mean number of tender joints 2.2, range 0-34). The overall agreement was also moderate (ICC 0.71). The agreement on individual tender joints was again poor to moderate (kappa 0.19-0.42).

There is only high concordance between physicians and patients on the absence of swollen joints (82%). The concordance on the presence of monoarthritis, oligoarthritis or polyarthritis is low (17-22%).

Conclusion: Owing to these discrepancies in assessment of individual joints and total number of affected joints, joint counts in AS assessed by doctors cannot be replaced by joint counts reported by the patient. Patients are only able to judge if their joints are not swollen.

INTRODUCTION

In patients with ankylosing spondylitis (AS) a minor part of the population has peripheral arthritis ($\pm 20\%$). Traditionally, either a doctor or a well-trained health care professional is involved in the clinical assessment of arthritis. The reliability of joint counts for swelling and pain reported by the patients is often studied in the assessment of disease activity in rheumatoid arthritis (RA). To evaluate these joint counts different methods have been used. Some authors^{1,2,3} used mannequins to mark painful or swollen joints, Hanly et al used a questionnaire (modified version of the rapid assessment of disease activity in rheumatology (RADAR) questionnaire)^{4,5} and some authors evaluated both^{6,7,8,9,10,11}. The reported study results were ambiguous but there were no differences in the results related to the method used. Some authors reported good reliability and suggested that patients' self-reported joint counts can be used to measure disease activity in RA^{1,2,7,10,11}. Others found moderate to poor reliability and suggested that joint counts derived by the patients can be used but are not interchangeable with joint counts reported by doctors^{3,4,6,8}. As far as we know reliability of patient reported joint counts in AS has not been studied. Our aim was to determine the reliability of self reported swollen joint counts and tender joint counts marked on a mannequins by patients with AS.

PATIENTS AND METHODS

The study sample comprised consecutive outpatients with AS at the University Hospital Maastricht, The Netherlands, the Maasland Ziekenhuis Sittard, The Netherlands, the University Hospital Ghent, Belgium, and the Hôpital Cochin, Paris, France. These hospitals are secondary and tertiary referral centers. All patients fulfilled the modified New York criteria for AS¹² and are participating in a longitudinal, observational study with follow-up visits according to a fixed protocol. The patients were asked to mark the painful joints on a mannequin presenting 44 joints (*Figure 1*, see page 38) and the swollen joints on a mannequin presenting 40 joints (*Figure 1*). The mannequin diagram was designed after the method of Stewart et al². Shoulder and hip joints were not represented on the swollen joint mannequin because it is very difficult to see swelling of these joints, especially by untrained people. At the same day but after the patient assessment, two doctors and one research nurse, one person per participating center, assessed the joints for pain and swelling. These results were reported on similar mannequins without knowledge of the patients' assessments. Data were collected at yearly intervals. In this paper the results of the baseline and one year data are presented.

Statistics

Reliability was determined by the Intra Class Correlation Coefficient (ICC, type 3.1) and kappa statistic. Kappa was used for between rater agreement on categorical data such as the individual joints scores. ICC was used for overall agreement on linear data such as the

Figure 1. Joint figure

total number of tender and swollen joints. To visualise this overall agreement we plotted the data using the method of Bland and Altman¹³ and calculated the 95% limits of agreement. This method is also designed as an absolute measure of agreement between two instruments which are on the same scale of measurement. To visualize this, the difference between two observations is plotted against the mean of the pairs of observations. Furthermore Spearman's correlation coefficients were computed for data without a normal distribution. All analyses were done with SPSS 10.0 for Windows.

RESULTS

There were 217 outpatients in our study, with a male to female ratio of 2:1. Table I describes the demographic and clinical features of all patients. Sixty one (28%) patients had finished more than secondary school and 39 (18%) attended elementary school only. In 58 (27%) patients peripheral arthritis was diagnosed by the treating rheumatologist. Psoriasis was diagnosed in 10 (4.6%) of the patients and dactylitis in 20 (9.2%) during the whole course of the disease. The mean score of Bath Ankylosing Spondylitis Disease Activity Index, (BASDAI)¹⁴ and the Bath Ankylosing Spondylitis Functional Index (BASFI)¹⁵, indicate an

Table 1. Characteristics of the patients (n = 217) presented as mean (SD, min.-max.) or percentage

age (years)	43 (12, 18-77)
male	67%
disease duration since diagnosis(yrs)	9.2 (8.6, 0-42)
reported peripheral arthritis*	27%
reported psoriasis**	4.6%
reported dactylitis**	9.2%
formal education	>12 yrs 28%
	≤6 yrs 18%
BASRI †(patients with ≥ 3 syndesmophytes at lumbar and /or cervical spine)	43%
BASDAI †† (range 0-10)	3.5 (2.1, 0-9.7)
BASFI ††† (range 0-10)	3.4 (2.6, 0-10)
Mander enthesis index (range 0-90)	7.7 (11.1, 0-56)

* Diagnosed by treating rheumatologist, ever during the disease course

† Bath Ankylosing Spondylitis Radiology Index

†† Bath Ankylosing Spondylitis Disease Activity Index

††† Bath Ankylosing Spondylitis Functional Index

overall mild disease activity and mild functional impairment of this group patients with AS. The mean score of Bath Ankylosing Spondylitis Radiology Index (BASRI), of the lumbar and cervical spine¹⁶, was 1.9 (range 0-4) with 93 (43%) patients having at least 3 syndesmophytes at the lumbar and /or cervical spine indicating moderate to severe damage. The enthesis index according to Mander¹⁷ was also computed with a mean score of 7.7 (range 0-90) at baseline (Table 1)

For baseline and one year data, both patients and doctors reported more tender than swollen joints (Table 2, see page 40). The average tender joint count and swollen joint count were comparable between doctors and patients; however, there was a striking difference in the maximum number of swollen joints scored by the doctors (31) compared with the number scored by the patients (8). The overall between observer agreement (ICC) between the patients and doctors for the total number of tender joints was moderate (0.71 and 0.54 for baseline and one year respectively) and was slightly worse on the total number of swollen joints (0.53 and 0.51 for baseline and one year respectively). There was no difference in agreement between patients and the two doctors and patients and research nurse for both swollen joint counts (0.53 and 0.54 respectively) and tender joint counts (0.71 and 0.70 respectively). In general, the baseline and one year data were very similar. For the remaining analyses we present baseline data only. By contrast with RA, most patients

Table 2. Summary statistics of total number of tender and swollen joints (n=217)

	physician mean (median, percentiles 10 and 90)	patient mean (median, percentiles 10 and 90)	inter-observer agreement (ICC)
tender joints			
baseline	2.2 (1.0, 0.0-7.0)	2.4 (1.0, 0.0-6.0)	0.71
1-year follow-up	2.1 (0.0, 0.0-7.0)	3.1 (2.0, 0.0-8.7)	0.54
swollen joints			
baseline	0.8 (0.0, 0.0-2.0)	0.5 (0.0, 0.0-2.0)	0.53
1-year follow-up	0.5 (0.0, 0.0-1.0)	0.5 (0.0, 0.0-2.0)	0.51

with AS have just a few inflamed joints. Therefore we analyzed the data also in a different way: the number of patients with none, one, two or three, or more than three swollen joints (Table 3). The doctors found one or more swollen joints in 54 of 217 (25%) patients whereas 44 of 214 (21%) of the patients reported one or more swollen joint. These percentages were very similar. However, this is misleading as there was low concordance on one or more swollen joints between the patients and doctors (51%). So the patients who judged that they had one or more swollen joint were often different from those judged by the doctor to have swollen joints. When the doctors' assessment was used as the gold standard, our results indicate that AS patients can judge whether their joints are not swollen (specificity 93%) but have difficulty judging one or more swollen joint (sensitivity 61%).

Table 3. Number of painful and swollen joints, concordance and distribution of 'root'-joints¹

Affected joints	Painful				Swollen			
	physician	patient	perfect concordance	'root' joints	physician	patient	perfect concordance	'root' joints
0	107	86	53%		162	170	82%	
1	30	26	12%	80%	22	20	17%	50%
mono-arthritis								
2 or 3	30 patients	49 patients	23%	63%	20 patients	14 patients	17%	50%
oligo-arthritis	70 joints	115 joints			48 joints	32 joints		
>3	47 patients	53 patients	47%	27%	12 patients	10 patients	22%	0%
poly-arthritis	320 joints	372 joints			97 joints	54 joints		

¹ 'root' joints are shoulders, hips and sternal-clavicular joints. Only sternal-clavicular joint were scored at the swollen joint figure. The % of 'root' joints are given of the concordant pairs.

Table 4. Levels of agreement on individual joints (kappa)

	tender joints		swollen joints	
	left	right	left	right
sternoclavicular	0.38	0.23	0.53	0.36
shoulder	0.42	0.27	-	-
elbow	0.43	0.30	0.1	0.1
wrist	0.37	0.32	0.48	0.48
hand (mcp,pip)	0.73*	0.62*	0.05*	0.25*
hip	0.26	0.19	-	-
knee	0.42	0.34	0.52	0.50
ankle	0.40	0.27	0.48	0.49
foot (mtp)	0.72*	0.57*	0.64	0.23

* Intra-class Correlation Coefficient (because of linear data)

According to the doctors 50% of patients had one or more tender joint as opposed to 60% according to the patients. Again the concordance on assessing one or more tender joint was rather low (60%). Sensitivity of the patients' judgement on tender joints was rather high (82%) but the specificity was low (62%).

Table 3 shows the number of tender joints and swollen joints and concordance rate of patients and doctors if the assessed joints were split into four categories: no arthritis, monoarthritis, oligoarthritis, or polyarthritis. The only high concordance rate found was 82% (category, non-affected swollen joints) again suggesting patients could only judge whether their joints were not swollen. The other concordance rates were at best moderate but overall they were low.

The distribution of monoarthritis and oligoarthritis according to the doctors gave more or less the expected distribution in AS patients (Table 3). In tender joints we found 53% and 34% involvement of root joints (shoulders and hips) and mostly large joints were affected instead of small joints of hand and feet.

Analysis by kappa statistics showed moderate to poor and non-consistent agreement between doctor and patients on individual joint counts for either pain or swelling (Table 4). Because there were only very few affected small joints in the hand and feet we clustered these for statistical analysis using ICC instead of kappa statistics.

Figure 2 (see page 42) shows the Bland and Altman plot of the total number of tender joints. There was a maximum difference of 25 joints between the doctors and the patients on the scoring range of 0 to 40; the 95% limits of agreement of the difference is 6.2 (1.96*SD). It also shows that the doctors consistently scored somewhat lower as the patients (mean difference -0.4). The Bland and Altman plot of the total number of swollen joints showed similar results (Figure 3, see page 42). However, now the doctors scored consistently somewhat higher than the patients. There was one outlier, which showed a difference of 26

Figure 2. Bland and Altman: mean versus difference of patient and physician; total number of tender joints.

Figure 3. Bland and Altman: mean versus difference of patient and physician: total number of swollen joints

swollen joints between doctor and patient. Both Bland and Altman plots showed the influence of the number of affected joints: by increasing number of affected joints, the disagreement between patient and physician was slightly larger although this is based on few patients.

We also computed Spearmans' correlations of the total number of tender and swollen joints judged by doctors and the patients with the enthesitis index according to Mander et al at baseline and dactylitis. The highest correlation found was 0.66 of physician painful joints with the enthesitis index. The correlations of swollen joints assessed by doctors and tender and swollen joints assessed by the patient with the index of Mander et al were 0.38, 0.41 and 0.30 respectively. Correlations with dactylitis were low; 0.23 for both tender and swollen joints assessed by the doctors and 0.27 and 0.23 for tender joints assessed by the patients.

Discussion

Collecting accurate and reproducible information from patient in routine rheumatology practice, epidemiological surveys, and clinical trials often is labour intensive and time consuming. This is one of the reasons for an increased use of self administration forms such as questionnaires on function, disease activity, and quality of life to assess the course and outcome of the disease. If joint counts assessed by doctors could be replaced by joint counts assessed by the patients, this would again lighten the task for rheumatologists in practice and researchers in particular. It would also make it easier to collect these data more often and even with postal questionnaires. However, to be able to replace joints derived by doctors with joint counts derived by patients, the validity of the second needs to be assessed. So far, studies comparing results from patients with those from doctors have only been carried out on patients with RA. Five of these studies showed good reliability^{1,2,7,10,11} by contrast with four that showed only poor to moderate reliability^{3,4,6,8}. The authors from the second group of studies concluded joint counts derived by patients could be used, but were not interchangeable with those derived by doctors. Furthermore, reliability for swollen en tender joint counts were the same^{3,11} or slightly better for assessing tender joints^{6,9}.

As far as we know, this is the first study in AS to compare the patient's assessment of tender and swollen joint count with that of the doctor as a gold standard. A major difference between patient with RA and those with AS is the fact that in RA all patients' peripheral joints are affected, although to a different degree during the course of the disease. In AS patients only about 20 to 30% of the patients have involvement of peripheral joints. Root joints are involved in about 30% of patients and this is relatively more common in those with juvenile onset AS¹⁸. Moreover, if patients with AS have involvement of peripheral joints, this is often to a lesser extent than patients with RA and often a different pattern of joints is involved.

Our results show that, on a group level, there is a consistent difference between the number of tender and swollen joints assessed by the patients or by the doctors. The patients score

consistently more tender joints, and the doctors more swollen joints. An explanation for the first finding could be that it is difficult for patients to differentiate between a tender joint and pain caused by enthesitis, although according to our data the enthesitis index of Mander et al was only significantly correlated with total number of tender joints assessed by the doctor. This was possibly because both assessments use the same methodology and most peripheral entheses are located near the joint. The second finding could be caused by the fact that patients with AS are not educated as to what a swollen joint means. From the group level results it could be concluded that absolute scores assessed by patients can not replace those assessed by doctors. If we look on a patient level, the results are even worse. The 95% limits of agreement were ± 6.2 for tender joints indicating that there may be a difference of 12 joints between the patient and doctor assessment, whereas there is no real difference. For the swollen joints the 95% limits of agreement were ± 4.5 .

Although, actual joint counts assessed by a doctor cannot be replaced by self-assessed joint counts, self assessment could still be valuable if the patients could differentiate between absence of arthritis and presence of monoarthritis, oligoarthritis, or polyarthritis. Again the concordance rates were very low for all groups of tender joints, and for the various levels of swollen joints. The only good concordance rate was in the absence of swollen joints. Consequently, patients are able to tell if they do not have swollen joints. However, if they have swollen joints, they are unable to judge the extent of the swelling, even within rough categories of monoarthritis, oligoarthritis and polyarthritis. Perhaps further studies could investigate if training of the patients would make a difference.

A limitation of our study is that we did not assess test-retest reliability formally. However, the results obtained at baseline and after one year of follow-up showed very similar results, indicating good reliability.

Our study results show major discrepancy between the number of tender and swollen joints assessed by a doctor or the patient. Therefore, joint scores derived by doctors can not be replaced by self assessed joint scores in AS. The only reliable result is the judgment of the patient that no joints are swollen.

REFERENCES

1. Escalante A. What do self-administered joint counts tell us about patients with rheumatoid arthritis? *Arthritis Care Res* 1998; 11: 280-90.
2. Stewart MW, Palmer DG, Knight RG. A self-report articular index measure of arthritic activity: investigations of reliability, validity and sensitivity. *J Rheumatol* 1990; 17: 1011-5.
3. Prevoo ML, Kuper IH, van't Hof MA, van Leeuwen MA, van de Putte LB, van Riel PL. Validity and reproducibility of self-administered joint counts. A prospective longitudinal followup study in patients with rheumatoid arthritis. *J Rheumatol* 1996; 23: 841-5.
4. Hanly JG, Mosher D, Sutton E, Weerasinghe S, Theriault D. Self-assessment of disease activity by patients with rheumatoid arthritis. *J Rheumatol* 1996; 23: 1531-8.
5. Mason JH, Anderson JJ, Meenan RF, Haralson KM, Lewis Stevens D, Kaine JL. The rapid assessment of disease activity in rheumatology (RADAR) questionnaire. Validity and sensitivity to change of a patient self-report measure of joint count and clinical status. *Arthritis Rheum* 1992; 35: 156-62.
6. Alarcon GS, Tilley BC, Li SH, Fowler SE, Pillemer SR. Self-administered joint counts and standard joint counts in the assessment of rheumatoid arthritis. *J Rheumatol* 1999; 26: 1065-67.
7. Abraham N, Blackmon D, Jackson JR, Bradley LA, Lorish CD, Alarcon GS. Use of self-administered joint counts in the evaluation of rheumatoid arthritis patients. *Arthritis Care Res* 1993; 6: 78-81.
8. Calvo FA, Calvo A, Berrocal A, Pevez C, Romero F, Vega E, et al. Self-administered joint counts in rheumatoid arthritis: comparison with standard joint counts. *J Rheumatol* 1999; 26: 536-39.
9. Houssien DA, Stucki G, Scott DL. A patient-derived disease activity score can substitute for a physician-derived disease activity score in clinical research. *Rheumatology* 1999; 38: 48-52.
10. Stucki G, Stucki S, Bruhlmann P, Maus S, Michel BA. Comparison of the validity and reliability of self-reported articular indices. *Br J Rheumatol* 1995; 34: 760-6.
11. Wong AL, Wong WK, Harker J, Sterz M, Bulpitt K, Park G, et al. Patient self-report tender and swollen joint counts in early rheumatoid arthritis. *J Rheumatol* 1999; 26: 2551-61.
12. 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.
13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307-10.
14. 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.
15. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994; 21: 2281-5.
16. 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.
17. Mander M, Simpson JM, Mc Lellan A, Wahler D, Goodacre JA, Dick WC. Studies with an enthesitis index as a method for patients with ankylosing spondylitis. *Ann. Rheum. Dis* 1987; 46: 197-202.
18. Khan MA. Ankylosing Spondylitis: Clinical Features. In: Klippel JH, Dieppe PA, *Rheumatology*. Mosby. 1998; 6: 16.2-3.

Chapter 5

MEASURING DISEASE ACTIVITY IN ANKYLOSING SPONDYLITIS: FROM PATIENT AND PHYSICIAN PERSPECTIVE

Anneke Spoorenberg, Astrid van Tubergen, Robert Landewé, Maxime Dougados, Sjef van der Linden, Herman Mielants, Hille van de Tempel, Désirée van der Heijde.

Submitted for publication

ABSTRACT

Introduction: There is no 'gold standard' to assess disease activity in patients with ankylosing spondylitis (AS). It is known that patients and physicians have different opinions about disease activity.

Objective: To investigate on which criteria patients with AS, and physicians base their judgement on disease activity.

Patients and methods: A cohort of 203 AS outpatients fulfilling the modified New York criteria included in the ongoing long-term follow-up was analysed. The Assessments in Ankylosing Spondylitis working group (ASAS) has established different domains relevant for outcome in AS. Each domain includes a number of instruments for making assessments, and all these instruments are included in OASIS and were made every six months for two years. Disease activity from the patient perspective as well as from the physician perspective was analysed using the patient's- or the physician's global assessment of disease activity (Visual analogue scale (VAS): 0 (best)-10 (worst)) by dichotomising into 'high disease activity' (VAS $\geq$ 6.0) and 'low disease activity' (VAS $\leq$ 4.0). Data reduction by principal components analysis (PCA) was performed to distinguish factors capturing correlated instruments. Discriminant analysis with the factor loadings was performed to discriminate between a low disease and a high disease activity state from both the patient- and the physician perspective. Multiple regression analysis on the discriminant scores was performed to prioritise the instruments.

Results: PCA revealed four factors: spinal mobility, physician assessments, patient assessments, and laboratory assessments (Cronbach's alpha: 0.52-0.80; explained variance: 61%). Discriminant function analysis showed that the factor 'patient assessments' was most important (pooled correlation: 0.85) in discriminating between a low-and a high disease activity state as defined by the patient. The other 3 factors contributed marginally (pooled correlation: $<$ 0.30). In contrast, the factors 'physician's assessments' (pooled correlation: 0.62), 'spinal mobility' (pooled correlation: 0.52) and 'laboratory assessments' (pooled correlation: 0.48) contributed most to the physician perspective. The factor 'patients assessment' did not contribute at all (pooled correlation: 0.05). Multivariate analysis showed that the patient perspective of disease activity was best captured by the instruments 'pain spine', 'BASFI', 'pain joints' and 'BASDAI fatigue'. The physician's perspective was best captured by 'cervical rotation', 'swollen joint count', 'CRP' and 'intermalleolar distance'.

Conclusion: AS patients rate disease activity on the basis of complaints while physicians rate disease activity on the basis of instruments related to disease severity and inflammation.

INTRODUCTION

In most rheumatological disorders, disease activity and outcome cannot be measured by one single variable. In clinical practice the opinion about disease activity and outcome is based on different sources of information, such as patient complaints, clinical variables, laboratory variables (acute phase reactants) and imaging. All this information is compiled into an overall impression of disease activity.

Patients and physicians may think differently about how to define disease activity. In judging whether their disease is active or not, patients may rate their complaints higher than eg abnormal laboratory results, or rapidly progressive damage on X-rays, whereas physicians will tend to give weight to the latter observations, irrespective of patient complaints.

In ankylosing spondylitis (AS) it is especially difficult to define disease activity because there is much variety in the clinical picture among different patients. Patients may only have axial involvement in all degrees of severity, but they may also have extraspinal manifestations, such as enthesitis and joint inflammation or inflammation of the gastrointestinal tract. The clinical diversity, both in severity and in localization, makes a high demand on the instruments that are supposed to measure disease activity and outcome in AS. Since there is no gold standard available for measuring disease activity and outcome, many different instruments have been developed to assess a variety of signs and symptoms in AS. Some of these instruments emphasize the patient perspective of disease activity or disease outcome, whereas others represent the physician's point of view. Acute phase reactants such as C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR) are considered more objective. In AS their value in determining whether the disease is active or not is rather limited. Elevated CRP-levels and ESR are frequently absent in AS and do not correlate well with clinical activity and radiological progression^{1,2}.

In 1999, a core set to assess outcome in AS was selected by the 'Assessment in Ankylosing Spondylitis' (ASAS) working group. This AS outcome core set consists of different domains with specific instruments for each domain³. It includes instruments that are supposed to primarily measure the patient's perspective as well as the physician's perspective regarding disease activity and outcome.

The objective of this study was to explore differences between the perspective of the patient and the perspective of the physician with respect to disease activity. We hypothesized that patients with AS will rate the activity of their disease primarily on the basis of their complaints, whereas the treating physicians will use other parameters on which to base their judgment.

PATIENTS AND METHODS

For this study data derived from the OASIS cohort (Outcome in Ankylosing Spondylitis International Study) were used. The OASIS project is an international longitudinal

observational multicenter study performed at the rheumatology outpatient department of the University hospital in Maastricht (the Netherlands), the Maasland hospital in Sittard (the Netherlands), Hôpital Cochin in Paris (France) and the University hospital in Ghent (Belgium). In total, 203 AS outpatients who satisfied the modified New York criteria⁴ were included in this study. 73% of patients were male, a distribution usually found in AS populations. The mean age at baseline was 43 (SD:13) years. The mean duration of disease since diagnosis is 11 (SD: 8) years. 27% of the patients had a history of peripheral arthritis diagnosed by their treating rheumatologist. At each institution the same trained person (2 rheumatologists, 1 research nurse) assessed all patients every six months according to a pre-specified protocol. All patients were followed by their rheumatologists, independent of the evaluations of the researchers.

Assessments

The following assessments were made every 6 months for the first two years of the OASIS study: physician spinal pain assessment (0= no pain on firm palpation, percussion and on extreme motion of complete spine. No spasm, 1= slight pain on firm palpation, percussion or motion of complete spine and no more than slight limitation of motion, 2= moderate pain on moderate palpation, percussion or motion of complete spine and no more than slight limitation of motion, 3= moderate to severe pain on light palpation, percussion or slight motion of complete spine and moderate to severe limitation of motion, 4= extreme pain with inability to withstand even light palpation or percussion and essentially no mobility of spine) (FDA guidelines). The best of two tries was documented in case of: chest expansion (cm)⁵, finger to floor (cm)⁶, occiput to wall (cm)(FDA guidelines), tragus to wall (cm)⁷, modified Schober test (cm)⁸, cervical rotation (d°)⁹, lateral spinal flexion (cm)¹⁰ and intermalleolar distance (cm)¹¹. Tragus to wall, modified Schober test, cervical rotation, lateral spinal flexion and intermalleolar distance were combined to compute the Bath Ankylosing Spondylitis Metrology Index (BASMI)¹². Other physician derived assessments were: the articular index according to Dougados (range 0-30)¹³, enthesitis index according to Mander (0-90)¹⁴, Maastricht Ankylosing Spondylitis Enthesis Score (MASSES) (range 0-13)¹⁵, physician assessment of disease activity on a visual analogue scale of 10 cm (VAS, 0 = not active, 10 = extremely active), physician assessment of the number of tender joint (range 0-44) and swollen joints (0-40)⁸.

Patient derived assessments were: duration of morning stiffness of the spine (min), duration of morning stiffness of the peripheral joints (min), pain of the spine (VAS, 0 = no pain, 10 = unbearable pain), pain of the peripheral (VAS, 0 = no pain, 10 = unbearable pain), patient assessment of disease activity (VAS, 0 = not active, 10 = extremely active), fatigue (VAS, 0=not at all, 10=extremely), Bath Ankylosing Spondylitis patient Global Score (BASG) (VAS, 0 = no influence of AS on global wellbeing during the past week, 10 = global wellbeing completely influenced by AS during the past week)¹⁶, patient assessment of night pain (range 0 = no pain, 4 = extremely painful during the whole night), Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) (range 0-10)¹⁷, Bath Ankylosing Spondylitis Functional

Index (BASFI) (range 0-10)¹⁸ and Dougados Functional Index (D-FI) (range 0-40)¹³. The fatigue question of the BASDAI was used as a single variable for further analysis of the domain fatigue. The laboratory assessments were: erythrocyte sedimentation rate (ESR, mm/h)) and C-reactive protein (CRP, mg/l).

All measurements were obtained to follow the course of ankylosing spondylitis.

Analysis

First, all selected variables were assessed for their suitability for parametric statistical analysis. Variables with a skewness statistic ≥ 1 were logarithmically transformed in order to obtain a reasonably normal distribution.

If mutual correlation of the selected instruments was high (Pearson correlation coefficient > 0.6), it was assumed that these variables assess the same process (collinearity). To diminish collinearity we chose one of the two variables with high Pearson correlation for further analyses.

Data reduction

In order to structure the large number of instruments and to identify underlying constructs (factors), principal components analysis (PCA) (SPSS 10.0.7 factor analysis, rotation-varimax) was performed on two subsets of all patients: those with active disease from the patient perspective (1) and those with active disease from the physician perspective (2). Active disease from the patient and physician perspective was defined as a score of ≥ 6.0 (VAS) and low disease activity was defined as ≤ 4.0 (VAS) on the instruments 'patient's global assessment of disease activity' and 'physician's global assessment of disease activity' respectively (VAS range: 0 = not active to 10 = extremely active). Factor loadings were saved in order to use in further statistical analysis. Internal consistency of the resulting factors was calculated using Cronbach's alpha.

Discriminant function analysis and linear regression analysis

To investigate whether and how the constructs (factors) could discriminate between low and high disease activity from patient and physician point of view, discriminant function analyses using the factor loadings were performed for both perspectives. Pooled correlations were used to judge the relative contribution of each factor to the discriminant function. To test the robustness of the observations, the same analyses were performed for different time points of the study.

Subsequently, linear regression analysis with stepwise selection of the variables was performed, in order to rate the relative contribution of every single instrument to the variability in individual discriminant scores.

Table 1: Mean, standard deviation (SD), median, inter quantile range (IQR), minimum, maximum and skewness before and after transformation of baseline data from OASIS population (patients n=203).

	mean, SD	median, IQR	min.-max.	skewness before transformation	skewness after LN transformation
chest expansion (cm)	4.7, 2.2	4.4, 3.0 – 6.0	0.4 - 12.5	0.63	
finger to floor (cm)	14.6, 13.8	12.9, 1.0 - 22.7	0.0 - 56.5	0.74	
cervical rotation (d°)	63.9, 23.2	68.0, 50.3 - 80.8	3.3 - 107.0	-0.67	
lateral spinal flexion (cm)	10.9, 5.9	10.8, 6.8 - 15.2	0.0 - 26.1	0.17	
intermalleolar distance (cm)	104.5, 21.7	106.0, 93.0 - 118.0	15.2 - 150.0	-0.62	
occiput to wall (cm)	3.8, 5.6	0.0, 0.0 - 5.8	0.0 - 26.1	1.87	0.18
physician spinal pain assessment (0-4)	0.9, 0.9	1.0, 0.0 - 1.0	0 - 4.0	1.00	0.20
physician swollen joints (0-40)	0.8, 2.6	0.0, 0.0 - 1.0	0.0 - 31.0	8.29	2.1
physician painful joints(0-44)	3.3, 5.0	1.0, 0.0 - 5.0	0.0 - 41.0	3.04	0.56
physician assessment of disease activity (VAS 0-10)	2.1, 1.5	1.3, 0.5 - 3.0	0.0 - 9.6	1.41	-0.42
Articular Index Dougados (0-20)	2.7, 3.3	2.0, 0.0 - 4.0	0.0 - 20.0	2.08	0.22
pain spine patient (VAS 0-10)	3.5, 2.4	3.2, 1.7 - 5.2	0.0 - 9.5	0.36	
pain joints patient (VAS 0-10)	3.0, 2.6	2.6, 0.5 - 4.9	0.0 - 10.0	0.71	
patient night pain (0-4)	1.2, 0.8	1.0, 1.0 - 2.0	0 - 3.0	0.27	
BASFI (0-10)	3.4, 2.9	3.3, 1.0 - 5.2	0.0 - 10.0	0.51	
BASDAI fatigue (VAS 0-10)	4.5, 2.9	4.6, 1.8 - 7.0	0.0 - 10.0	0.03	
patient assessment of disease activity (VAS 0-10)	3.8, 2.8	3.5, 1.3 - 5.7	0.8 - 10.0	0.48	
ESR mm/h	14.3, 16.0	10.0, 4.0 - 18.0	0.0 - 118.0	2.96	-0.06
CRP mg/l	17.8, 24.9	7.0, 6.0 - 19.0	0.0 - 139.0	2.73	-0.18
BASDAI total (0-10)	3.5, 2.1	3.2, 1.6 - 5.3	0.6 - 9.7	0.39	

RESULTS

Selection of variables

The descriptive statistics of all assessments at baseline are presented in *Table 1*. Variables not normally distributed (skewness ≥ 1) were ln-transformed. A Pearson correlation matrix of all variables from the OASIS data set at baseline was constructed to trace collinearity ($r > 0.6$). We chose one of the two variables with high inter-variable correlation for further analysis. BASFI was selected instead of D-FI; occiput-to-wall distance was selected instead of tragus-to-wall distance; lateral spinal flexion was selected instead of the modified Schober test; pain spine was selected instead of BASG-week; physician assessment of painful joints was selected instead of MASES and Mander entheses index.

Finally, the following instruments were selected for further analysis: physician spinal pain assessment (ln-transformed), physician assessment of painful peripheral joints (ln-transformed), physician assessment of swollen peripheral joints (ln-transformed), chest expansion, finger-to-floor distance, occiput-to-wall distance (ln-transformed), lateral spinal flexion, cervical rotation, intermalleolar distance, physician's assessment of disease activity (ln-transformed), Dougados articular index, patient assessment of disease activity, patient night pain, patient pain joints, patient pain spine, BASFI, the BASDAI fatigue question, and ESR (ln-transformed) and CRP (ln-transformed).

Patients available for analysis

In all, 203 OASIS records with baseline data were available. In case of the patient perspective a total of 158 OASIS records fulfilled the criteria for low ($n = 112$) or for high ($n = 46$) disease activity. The 45 records with an intermediate level of disease activity were not used for further analyses. For the physician perspective of disease activity 145 OASIS records were available for further analyses; 128 with low disease activity and 17 with high disease activity. The 58 records with an intermediate level of disease activity according to the physician were neglected.

Factor analyses

Factor analysis (SPSS-factor analysis (PCA) with varimax rotation for the final solution) on the OASIS baseline data of the above-described variables was performed to structure the high number of variables. For both patient and physician perspectives of disease activity 4 factors were extracted with Eigen-values > 1 , and a cumulative percentage of explained variance of 61%. *Table 2* (see page 54) shows the distribution of the selected OASIS baseline variables in a four-factor model for each perspective of disease activity. Both perspectives gave the same factors that consisted of variables reflecting spinal mobility and function (1), variables reflecting assessments by the physician (2), variables reflecting assessments by the patient (3) and variables reflecting laboratory acute phase reactants (4). The order in which the factors appear slightly differed across the perspectives. The internal consistency of the factors was moderate to good (Cronbach's α ranging from 0.52 to 0.80) (*Table 2*).

Table 2: Four factor models with explained variance and internal consistency (Cronbach's α) of all selected OASIS baseline data reflecting the patient perspective and the physician perspective of disease activity

Perspective: patient assessment of disease activity				Perspective: physician assessment of disease activity			
factor 1 'mobility and function'	factor 2 'physician assessments'	factor 3 'patient assessments'	factor 4 'lab'	factor 1 'patient assessments'	factor 2 'mobility and function'	factor 3 'physician assessments'	factor 4 'lab'
chest expansion	spinal pain assessment	pain spine	ESR	pain spine	chest expansion	spinal pain assessment	ESR
fingers to floor	number of painful joints	pain joints	CRP	pain joints	fingers to floor	number of painful joints	CRP
cervical rotation	number of swollen joints	night pain		night pain	cervical rotation	number of swollen joints	
lateral spinal flexion	physician assessment of disease activity	BASFI		patient assessment of disease activity	lateral spinal flexion	articular index of Dougados	
intermalleolar distance	articular index of Dougados	BASDAI fatigue		BASDAI fatigue	intermalleolar distance		
occiput to wall				BASFI	occiput to wall		
α 0.52	α 0.80	α 0.73	α 0.71	α 0.80	α 0.52	α 0.75	α 0.71
explained variance: 61%				explained variance: 61%			

Discriminant analyses

The factor values of the four factors of both disease activity perspectives were used separately (as independent variables) in the discriminant function analyses. We performed discriminant function analyses on the baseline data, and subsequently on data of 6, 12, 18 and 24 months follow up. The factor values (dependent variables) were restricted to the groups of patients with defined high and low disease activity for both perspectives. *Table 3* shows the pooled- and canonical correlation coefficients for all 4 factors for each perspective of disease activity.

Factor 3 (patients assessments) and to a far less extent factor 2 (physicians assessments) contributed most to the discriminant score describing disease activity from the patients perspective, for all time periods analysed. The other 2 factors were not contributive at all. The pooled correlation for factor 3 remained high if tested at other time points. Factor 3

Table 3: Pooled correlation of discriminant functions of four factors according to the patient's perspective and the physician's perspective of disease activity in AS at baseline, 6, 12, 18, 24 months follow up.

Perspective and discriminating factor	baseline	6 months	12 months	18 months	24 months
Patient assessment of disease activity					
Factor 1 'mobility and function'	0.13	0.12	0.24	0.04	0.14
Factor 2 'physician'	0.27	0.43	0.42	0.34	0.43
Factor 3 'patient'	0.85	0.84	0.79	0.81	0.80
Factor 4 'lab'	0.23	0.04	-0.01	-0.14	0.06
Canonical correlation	0.63	0.49	0.49	0.59	0.65
Correctly classified (%)	79	69	72	77	78
Physician assessment of disease activity					
Factor 1 'patient'	0.05	0.41	0.07	-0.14	0.37
Factor 2 'mobility and function'	0.52	0.47	0.55	0.99	0.81
Factor 3 'physician'	0.62	0.72	0.74	-0.05	0.38
Factor 4 'lab'	0.48	0.21	0.41	-0.14	0.18
Canonical correlation	0.39	0.44	0.35	0.22	0.25
Correctly classified (%)	77	78	80	67	72

(physicians assessments) contributed most to the discriminant score describing disease activity from the physicians perspective, but factor 2 (spinal mobility and function) as well as factor 4 (laboratory acute phase reactants) also contributed significantly. Interestingly, factor 1 (patients assessments) was far less important. The contribution of factor 2 to the discriminant score was very consistent over time; the contribution of the factors 1, 3 and 4 was somewhat less consistent over time.

Regression analysis

Both discriminant functions (one for the patient perspective and one for the physician perspective) contain 4 factor values, which are each composed of several clinical variables. In order to get insight in which variables explain the discriminant scores best, we performed stepwise forward multiple linear regression analyses for each disease activity perspective. These regression analyses were performed on the baseline OASIS records, with the two discriminant functions (individual discriminant scores) as dependent variables, and all selected assessments as explanatory variables.

Table 4 shows the relative contribution of the variables as a result of this stepwise analysis. Disease activity from the patient perspective was best captured by the instrument 'pain

Table 4: Relative contribution of individual variables explaining disease activity from the perspective of the patient, as well as from the perspective of the physician.

variable	corresponding factor	model standardized coefficient Beta				
		1	2	3	4	5
perspective: patient's assessment of AS disease activity						
pain spine	3	.76	.55	.44	.37	.34
BASFI	3		.52	.42	.36	.31
pain joints	3			.30	.29	.29
BASDAI fatigue	3				.22	.22
Physician disease activity LN	2					.18
variance (R square)		.58	.81	.86	.90	.93
perspective: physician's assessment of AS disease activity						
cervical rotation	2	-.74	-.59	-.54	-.44	-.37
swollen joints LN	3		.46	.35	.37	.38
CRP LN	4			.33	.34	.30
intermalleolar distance	2				-.25	.24
finger to floor	2					.18
variance (R square)		.54	.73	.83	.88	.91

spine', followed by BASFI, the instrument 'pain joints', the BASDAI 'fatigue' question and the instrument 'physician global assessment of disease activity'. Disease activity from the physician's perspective was best captured by the instrument 'cervical rotation', followed by the instruments 'number of swollen joints', CRP, 'intermalleolar distance' and 'finger-to-floor distance'. The combination of these 5 instruments explained more than 90% of variance in discriminant scores for each perspective of disease activity. The standardized coefficients of the end solution reveal that 'pain spine', BASFI and 'pain joints' contribute approximately similarly to explaining disease activity from the patient perspective. The instruments 'cervical rotation' and 'swollen joint count' and to a somewhat lesser extent CRP, contribute almost similarly to explaining disease activity from the physician perspective.

DISCUSSION

The most important conclusion from this study is that patients with AS and physicians have different views on what active AS means.

Patients base their judgment primarily on the presence and severity of complaints related

to AS, which are registered by self-administered questionnaires. The impact of spinal mobility, the assessments made by the physician and the acute phase reactants can almost be neglected. Instruments measuring *pain* (pain spine, pain joints) contribute more to explaining disease activity from the patient perspective than instruments measuring *stiffness*, *fatigue* and *general well being*. An explanation may be that measuring pain already captures the latter. Apart from these pain instruments the BASFI, an index primarily designed to assess *function* also contributes to disease activity from the patient perspective. Apparently, patients base part of their estimation of disease activity on what they are able to (physically) perform. This is a phenomenon that we know from rheumatoid arthritis patients in whom function measured by the health assessment questionnaire (HAQ) is often strongly correlated with disease activity¹⁹.

A few more conclusions can be derived from the observations regarding the patient perspective. First, patients appear to properly distinguish disease *activity* (defined by them as complaints) from disease *severity* (spinal mobility). Second, measuring acute phase reactants does not capture disease activity from the patient perspective: laboratory assessments do not contribute to disease activity as perceived by the patients.

The physician judgment about disease activity rests on a combination of constructs: assessments made by the physician are particularly important, and patient-derived scores were far less contributive. Looking at the combination of instruments that best explains the physician's judgment, it is remarkable that 3 of the 5 variables include measures combining information on disease activity and severity (cervical rotation, intermalleolar distance and finger-to-floor distance) rather than pure activity (swollen joints and CRP). In a separate study designed to investigate on which criteria physicians judge whether a patient with AS should be treated with TNF-blocking drugs, we also found that physicians rated disease severity at least as important as disease activity (personal communication).

The disease activity markers that emerged from the multivariate analysis (CRP and swollen joint count) are known to be insensitive to trace disease activity from the patient perspective. CRP is poorly correlated with BASDAI, a patient-derived index of disease activity, and the swollen joint count is an insensitive marker for active disease, because only a subset of patients actually has swollen joints.

It is rather intriguing that we, as physicians, now commonly use patient-derived instruments (such as BASDAI) as a gold standard for measuring disease activity, both for including patients in clinical trials as well as establishing drug effects, whereas there is no appropriate evidence that patients and physicians perceive disease activity similarly. Another important consideration is the lack of evidence that either patient derived assessments or physician derived assessments of disease activity are somehow associated with long term outcome in AS (long-term function, structural damage, loss of participation etc). Since the options for drug therapy in AS are steadily increasing it becomes more and more important to define measures assessing a uniform construct of disease activity and long-term outcome to be used in clinical trials.

One may question the validity of this study with respect to extrapolation of the findings.

The OASIS cohort is an observational cohort with consecutive patients with AS from three different countries, from university hospitals as well as non-university general hospitals. Patients were included irrespective of age, gender, disease duration, severity or activity of their disease, and may therefore be considered an appropriate reflection of the average patient with AS. As shown in Table 1, the OASIS patients cover the entire range of activity- and severity variables scores, which may further add to this conclusion .

In order to be able to truly discriminate between active and inactive disease, we defined inactive disease at a value of ≤ 4.0 and active disease at a value of ≥ 6.0 (VAS from 0 to 10) for both patient and physician perspectives. By doing this, we may limit the interpretability of the findings in case of an 'indifferent' level of disease activity. The goal of this study, however, was to investigate how patients and physicians define inactive and active disease (which asks for a clear distinction), rather than classifying patients as active or inactive. However, the number of patients with high disease activity was rather low, especially in case of the physician perspective. The analyses at other time points revealed similar information as compared to the analyses performed at baseline data, which adds to the validity. These analyses are based on the same patient group and therefore do not contain independent information. Furthermore the patient perspective of disease activity is based on the opinion of a large number of patients, on the other hand the physician perspective is based on the opinion of three investigators only. Another remark is that one of the investigators assessing 'physician global disease activity' (in less than 25% of patients) is not a rheumatologist, but a very experienced research nurse trained to assess patients in many AS studies. Strictly spoken this is not a physician assessment and as a research nurse she lacks the experience of treating AS patients. The similarity in results over the entire two year follow up might serve as an external validation for the consistency in assessing disease activity by the involved assessors. Probably most important is that the treating physicians of the Dutch OASIS patients based their decisions on starting TNF-blocking agents more on spinal mobility and other disease severity measures than on disease activity measures (unpublished data), which is in consistence with the results of the physician perspective in this study.

In summary, in this study we showed that patients with AS perceive disease activity differently in comparison with physicians. Patients rate complaints and to a lesser extent function as important values of defining disease activity. Physicians rate variables reflecting inflammation and severity such as their own assessments and acute phase reactants as most important in assessing disease activity instead of patient perception.

REFERENCES

- 1 Spooenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van de Tempel H, van der Linden S. Relative value of erythrocyte sedimentation rate and C-reactive protein in ankylosing spondylitis. *J Rheumatol* 1999; 26: 980-4.
- 2 Ruof J, Stucki G. Validity aspects of Erythrocyte Sedimentation Rate (ESR) and C-reactive protein (CRP) in ankylosing spondylitis. a literature review. *J Rheumatol* 1999; 26: 766-70.
- 3 van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. Assessments in Ankylosing Spondylitis. *J Rheumatol* 1999; 26: 951-4.
- 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 Rheumatological Physical Examination. ISBN 0-8089-1824-9. Chapter 5: p.55.
- 6 Miller MH, Lee P, Smythe HA, Goldsmith CH. Measurements of spinal mobility in the sagittal plane: new skin contraction technique compared with established methods. *J Rheumatology* 1984; 11: 507-11.
- 7 Tomlinson M, Barefoot J, Dixar ASJ et al. Intensive in-patient physiotherapy courses improve movement and posture in ankylosing spondylitis. *Physiotherapy* 1986; 72: 238-40.
- 8 Spooenberg A, van der Heijde D, Dougados M, de Vlam K, Mielants H, van de Tempel H, van der Linden S. The reliability of self-assessed joint counts in ankylosing spondylitis (AS). *Ann Rheum Dis*. 2002; 61: 799-803.
- 9 Joint motion. Method of Measuring and Recording. Chicago; American Academy of Orthopaedic Surgeons 1965: 44-64.
- 10 Pile KD, Laurent MR, Salmond CE, Best MJ, Pyle EA, Moloney RO. Clinical assessment of ankylosing spondylitis: a study of observer variation in spinal measurements. *Br J Rheumatology* 1991; 30: 29-34.
- 11 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 Rheumatology* 1995; 34: 767-70.
- 12 Jenkinson TR, Mallorie PA, Whitelock H, Kennedy LG, Garrett SL, Calin A. Defining spinal mobility in ankylosing spondylitis. The Bath Ankylosing Spondylitis Metrology Index. *J Rheumatol*. 1994; 21: 1694-8.
- 13 Dougados M, Gueguen A, Nakache JP, Nguyen M, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988; 15: 302-7.
- 14 Mander M, Simpson JM, McLellan A, Walker D, Goodacre JA, Dick WC. Studies with an enthesitis index as a method of clinical assessment in ankylosing spondylitis. *Ann Rheum Dis*. 1987; 46: 197-202.
- 15 Heuft-Dorenbosch L, Spooenberg A, van Tubergen A, Landewé R, van der Tempel H, Mielants H, Dougados M, van der Heijde D. Assessment of enthesitis in ankylosing spondylitis. *Ann Rheum Dis* 2003; 62: 20-6.
- 16 Jones SD, Steiner A, Garrett SL, Calin A. The Bath Ankylosing Spondylitis Patient Global Score (BAS-G). *Br J Rheumatol* 1996; 35: 66-71.
- 17 Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease activity in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol* 1994; 21: 2286-91.
- 18 Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, Jenkinson T. A New approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994; 21: 2281-5.
- 19 Drossaers-Bakker KW, de Buck M, van Zeben D, Zwinderman AH, Breedveld FC, Hazes JM. Long-term course and outcome of functional capacity in rheumatoid arthritis: the effect of disease activity and radiologic damage over time. *Arthritis Rheum* 1999; 42: 1854-60.

Chapter 6

THE DEVELOPMENT OF THE **ASQoL**: A QUALITY OF LIFE INSTRUMENT SPECIFIC TO ANKYLOSING SPONDYLITIS

Lynda Doward, Anneke Spoorenberg, Sharon Cook, Diane Whalley, Philip Helliwell, Lesley Kay, Stephen McKenna, Alan Tennant, Désirée van der Heijde, M Anne Chamberlain.

Annals of the rheumatic diseases 2003; 62: 20-6.

ABSTRACT

Objective: Although disease-specific health status measures are available for ankylosing spondylitis (AS), no instrument exists for assessing quality of life (QoL) in the condition. The aim was to produce an AS-specific QoL measure that would be relevant and acceptable to respondents, valid, and reliable.

Methods: The ASQoL employs the needs-based model of QoL and was developed in parallel in the UK and the Netherlands (NL). Content was derived from interviews with patients in each country. Face and content validity were assessed via patient field-test interviews (UK and NL). A postal survey in the UK produced a more efficient version of the ASQoL, which was tested for scaling properties, reliability, internal consistency and validity in a further postal survey in each country.

Results: A 41-item questionnaire was derived from interview transcripts. Field-testing interviews confirmed acceptability. Rasch analysis of data from the first survey ($n=121$) produced a 26-item questionnaire. Rasch analysis of data from the second survey (UK: $n=164$; NL: $n=154$) showed some item misfit, but that items formed an hierarchical order and were stable over time. Problematic items were removed giving an 18-item scale. Both language versions had excellent internal consistency (alpha 0.89 to 0.91), test-retest reliability (Spearman 0.92 UK and 0.91 NL) and validity.

Conclusions: The ASQoL provides a valuable tool for assessing the impact of interventions for AS and for evaluating models of service delivery. It is well accepted by patients, taking approximately 4 minutes to complete, and has excellent scaling and psychometric properties.

Ankylosing spondylitis (AS) is a chronic inflammatory rheumatic condition affecting the sacroiliac joints, to a varying degree the spinal column and to a smaller extent the peripheral joints. Patients have pain, morning stiffness and disability which increases with duration of disease. A number of patients also experience extra-spinal and extra-articular manifestations as acute anterior uveitis and inflammatory bowel disease¹. Population studies report a prevalence of AS of between 0.5% and 1.6% and it is more commonly found in men than women^{2,3}. The pattern and rate of disease progression are variable but may be independent of disease duration⁴. Although major advances in the understanding of the disease pathogenesis have occurred in recent years, the optimal strategy for treatment is still unknown. Disease onset is generally in late adolescence or early adulthood and, consequently, the effects are present for a majority of the patient's life. Progression may continue through what should be economically active years⁵. Chamberlain⁶ reported that two-thirds of male patients experience difficulty at work, one third have social problems and up to two thirds report having difficulty with sexual activity. Reactive depression and frustration are noted together with impaired self-esteem and social skills⁶. Energy related problems are also widely reported⁶. All these features denote significant effects of the disease on lifestyle.

There is a growing interest in the assessment of quality of life (QoL), particularly in chronic disabling conditions. It is becoming relatively common to measure QoL in studies designed to assess the impact of new pharmaceutical products or to compare different treatment regimes. Although the concept has existed for many years, it is only within the last few decades that attempts have been made to operationalise QoL into a construct that can be measured in a meaningful way⁷.

Instruments currently available for use with AS patients focus predominantly on symptoms (impairment) or functioning (disability, or both, and are used to assess the presence or absence of disease and its consequences in these terms. Such instruments include: the Bath Ankylosing Spondylitis Functional Index (BASFI)⁸, the Leeds Disability Questionnaire (LDQ)⁹, the Ankylosing Spondylitis Assessment Questionnaire (ASAQ)¹⁰, the Dougados Functional Index (DFI)¹¹, a version of the Stanford Health Assessment Questionnaire modified for the spondylarthropathies (HAQ-S)¹², and, a modified version of the Arthritis Impact Measurement Scales II specific to AS (AS-AIMS II)¹³. Although such measures provide important information on the level of impairment and disability experienced by patients, they do not inform on the impact of the condition on QoL. The construct of QoL differs from impairment and disability in so far as it concerns the impact of disease from the patient's (rather than a clinical) perspective. By investigating how patient's lives are affected by impairment, disability and other influences it provides an outcome that is complementary to the traditionally assessed impacts of disease^{14,15}. Generic health status instruments such as the Nottingham Health Profile, SF-36 and EuroQoL also concentrate on impairment and disability rather than QoL. Furthermore, they have been shown to lack the responsiveness

necessary to detect real changes in health status associated with effective treatment^{8,16}. There is a clear need for a valid and reliable disease-specific instrument for assessing the impact of AS on QoL that is suitable for use in clinical practice. This paper describes the development of such a measure, the Ankylosing Spondylitis Quality of Life Questionnaire (ASQoL). The instrument was required to be suitable for monitoring patients, evaluating alternative treatment regimes, new pharmaceutical products and/or models of service delivery from the patient's perception. The development methodology employed is based on recent advances in the recognition and understanding of the conceptual and practical basis of measurement. The process combines the theoretical strengths of the needs-based quality of life model¹⁷ with the statistical and diagnostic power of the Rasch model¹⁸. The needs-based model of quality of life postulates that life gains its quality from the ability and capacity of the individual to satisfy his or her needs. QoL is high when these needs are fulfilled and low when few needs are satisfied. The model is well-established and has been applied successfully in the development of a large number of disease-specific QoL instruments, several of which have become established as the outcome instrument of choice for clinical trials and studies^{17,19,20,21,22,23,24,25,26}. The application of the Rasch model ensures that the fundamental scaling properties of the instrument (for example, unidimensionality and level of measurement) are assessed in addition to the traditional psychometric assessments of reliability and construct validity. Such basic measurement properties were considered at each stage of the development of the ASQoL.

PATIENTS AND METHODS

Figure 1 sets out the stages involved in the development of the ASQoL. The intention was to produce an instrument that would be equivalent for both the UK and the Netherlands (NL). Consequently, all stages were conducted simultaneously in both countries, with the exception of stage 4, which took place in the UK only. The purpose of stage 4 was to produce a more efficient instrument for final testing, by removing clearly problematic items.

Patient Samples

The study was approved by ethics committees in both countries and participants gave their written informed consent. All participating patients fulfilled the modified New York criteria for AS^{27,28}. Patients with significant co-morbidity such as psychiatric disorders, cancer or fibromyalgia were excluded. To ensure that a wide spectrum of clinical features were represented, each sample included patients with both axial and peripheral disease, a range of disease duration and patients with uveitis or inflammatory bowel disease, or both. Patients were recruited from three hospitals in the north of England and from three in the south of the Netherlands. In both countries, different patients participated at each stage of the study.

Figure 1: Stages in the production of the ASQoL

Stage 1: Interviews with patients

Deriving the content of a measure from individuals who are representative of the target population ensures that only relevant topics are included and that areas important to QoL are not omitted. For the ASQoL, the content of the questionnaire was derived from unstructured, qualitative interviews with relevant patients in both countries, conducted by experienced qualitative researchers. The interviews, took the form of informal, focused conversations. They were designed to explore the impact of AS on the patient, with emphasis on the persons's ability to fulfil his or her needs. For example, where interviewees indicated functional limitations associated with AS they were prompted to consider how such restrictions impacted on their lives – particularly, how they prevented the fulfilment of their needs. The interviews were audio recorded with the permission of the interviewee. Transcripts were produced from the tapes, which were then wiped clean. All traces of the interviewee's identity were omitted from the transcripts to maintain anonymity.

Stage 2: Selection of items and response format for the draft questionnaire

In both countries, the interview transcripts were subjected to independent content analysis to identify statements relating to need satisfaction. As far as possible, the actual words used by interviewees were selected for the questionnaire. Duplicate and idiosyncratic items were removed and the list was subjected to further scrutiny, with items retained if they were applicable to all potential respondents, reflected a single idea, were unambiguous and were short and simple. The item lists from each country were then compared at a meeting between the English and Dutch researchers. The purpose of this meeting was to decide on the content for the first draft of the questionnaire and to identify a response system that would be suitable for both languages.

A yes/no response system was selected for the draft measure as previous experience had indicated that this maximises language equivalence and ease of scoring and minimises respondent burden. In the development of the rheumatoid arthritis-specific instrument (the RAQoL) it was shown that a yes/no response format was more sensitive to change than a four-response Likert-type format²².

Stage 3: Field testing for face and content validity

The purpose of this exercise was to test the applicability, comprehensibility, relevance and comprehensiveness of the ASQoL with patients with AS. Participants completed the questionnaire in the presence of an interviewer. They were then asked to comment on its ease of completion and on the appropriateness of the instructions, items and response format. Items found to be problematic in either country were removed. Items were considered problematic if respondents found them ambiguous or difficult to understand. Results from this stage were used to compile a second draft version of the measure.

Stage 4: Postal Survey I (UK)

The new draft ASQoL was administered by post to patients in the UK. Analyses were

performed on the resulting data in order to identify items that failed to fit onto the underlying measurement construct and/or that worked differently by age (above or below the median), gender, AS diagnosis (axial only or axial with peripheral involvement), or disease duration (above or below the median). Such differential item functioning (DIF) would indicate that an item is valued differently by subgroups of the patient population. For example, in a disability measure it might be hypothesised that an item such as *I am unable to travel to my workplace* would be affirmed less often by respondents who have reached retirement age. Therefore, regardless of their level of disability, this item would appear to be less severe for younger respondents. DIF was identified through the application of the one parameter logistic item response theory model - the Rasch model¹⁸. In the context of a QoL scale, the Rasch model applies the premise that the likelihood of a person affirming a particular item depends on the level of QoL of the person and on the level of QoL represented by that item. The analysis provides estimates for the item and person parameters in log-odds units (logits). Such estimates are based on the assumption that the scale is indeed measuring a single underlying construct - that is, that the items form a unidimensional scale. The extent to which this assumption is justified is indicated by item fit statistics. For the present analysis, Rasch mean square (MNSQ) item fit statistics were identified through application of the computer program WINSTEPS²⁹. Two MNSQ statistics are given; an information-weighted fit statistic (INFIT) and an outlier-sensitive fit statistic (OUTFIT). OUTFIT is more sensitive to inconsistencies in the extreme responses, that is those made to items far removed from the individual's level of QoL. The INFIT statistic is weighted so that these outliers have less impact and is, thus, more sensitive to non-extreme responses. Taken together, these two MNSQ fit statistics inform on the extent to which the individual items map onto the underlying measurement construct, in this case, the QoL. Given the present sample sizes, MNSQ values between 0.7 and 1.3 were taken to reflect adequate fit to the model³⁰. As no Dutch data were included in Stage 4, only those items that were clearly problematic were removed. The third draft of the questionnaire was produced on the basis of these analyses and used in the subsequent postal survey in both countries.

Stage 5: Postal Survey 2 (UK and NL)

The purpose of the final postal survey was to assess the scaling properties, reliability, internal consistency and construct validity of the ASQoL in each country. Patients in both countries were sent a package consisting of the ASQoL, a demographic questionnaire, additional comparator measures and a reply paid envelope. Patients who completed and returned the first pack were sent a similar package timed to arrive two weeks later. The demographic questionnaire, which was consistent across countries, included questions on patient perceived disease activity and severity of illness. The Nottingham Health Profile (NHP)³¹ and the BASFI were used as comparator measures in both countries. In addition, the LDQ was used in the UK and the Dougados Functional Index (DFI)³² was selected in the Netherlands. The NHP is a measure of perceived distress and provides a profile of

scores in six sections; physical mobility, energy level, pain, emotional reactions, social isolation and sleep. It is scored out of a maximum of 100 for each of the sections, with a higher score indicating greater distress. The BASFI, the LDQ and the DFI each yield a single score. Scores on the BASFI can range from 0-100, on the LDQ, from 0-48 and on the DFI from 0-40. For each of these scales, a high score indicates greater disability. Each item on the ASQoL is given a score of '1' or '0'. A score of '1' is given where the item is affirmed, indicating adverse QoL. All item scores are summed to give a total score or index, with a high score indicating a worse QoL. Questionnaires with missing data were omitted from the analysis. The following properties of the two versions of the ASQoL were assessed: scaling properties, reliability, internal consistency and construct validity.

Scaling properties

Rasch analyses were conducted to confirm that items mapped onto the same underlying construct (unidimensionality), that they represented different amounts of the construct (hierarchical ordering) and that they worked in same way across different patient groups (differential item functioning). The level of measurement (that is, ordinal or interval level) provided by the measure was also examined.

Reliability

The reliability of the ASQoL was assessed using the test-retest method. This is an estimate of the instrument's reproducibility over time, assuming that no change in condition has taken place. For each country, ASQoL scores from each administration were correlated. Patients were excluded from these analyses if they reported significant changes to their perceived general health, severity of illness or perceived disease activity (that is, whether or not the patients considered their disease to be active at the time of completing the questionnaire) between administrations. Where an instrument is required for use in a clinical trial or for monitoring individual patients, a correlation coefficient of at least 0.85 is required³³. Due to the ordinal nature of the data, Spearman rank correlation coefficients were produced (intra-class correlation coefficients are also reported for information only).

Internal consistency

Internal consistency was assessed by Cronbach's alpha coefficients. This statistic indicates the degree of relatedness between items. A value of 0.70 or above was taken as reflecting adequate internal consistency³⁴.

Construct validity

ASQoL scores were related to the comparator instruments and to patient perceived general health and severity of illness and patient-perceived disease activity (that is, whether or not the patients considered their disease to be active at the time of completing the questionnaire). Patients describe disease activity in terms of whether they are having a 'good day' or a 'bad day'. This terminology is used throughout the results section. It was predicted

that there would be a moderate association between the ASQoL and the comparator measures indicating that they assess different but related constructs. It was also hypothesised that QoL would be worse for respondents experiencing a bad day (active disease), those reporting poorer general health or those describing their AS as severe.

RESULTS

Findings from the interviews (stage 1)

Thirty patients were interviewed in the UK and 25 in the Netherlands. Patient samples were comparable in each country. Approximately two thirds of those interviewed were male and a third reported having peripheral arthritis. The age of those interviewed ranged from 18 to 78 years, with disease duration ranging from 1.5 to 44 years. Interviews lasted for between 30 minutes and two hours with a median length of one hour and 10 minutes. All respondents chose to be interviewed in their own homes and all gave consent for the interview to be audio recorded.

Similar findings emerged from the Dutch and UK interviews. Respondents commented on the impact of pain and its effect on sleep, mood, motivation and ability to cope with the day ahead. One of the greatest fears expressed was that of losing independence. Many reported that they required some degree of assistance with everyday tasks such as dressing, washing and shopping (particularly for foodstuffs). In addition, many reported feeling that they were no longer in control of their own personal hygiene or grooming. A particular concern was about the future, particularly in relation to uncertainties surrounding disease progression. The AS had a major impact on interviewees' ability to meet their needs for stimulation and exploration, gender role fulfilment and feelings of worth. Major impacts were also reported on self image and self esteem, resulting from concerns over appearing slouched or slovenly. AS had a profound impact on relationships with family members and friends and social life was severely limited. For example, several interviewees commented that they chose places they could visit on the basis of how tolerable they found the seating. The condition was often cited as a major source of family tension and some interviewees reported taking out their frustration and anger on those closest to them.

Development of the draft questionnaire (stage 2)

Items for the questionnaire consisted of actual quotations from the transcripts in a majority of cases. However, it was necessary to change the actual words used by interviewees for some of the items. For example, some were shortened, had the word order altered, or were changed so that they were expressed in the first person and/or in the present tense. The item pool from each country was compared and items selected for the draft questionnaire that covered issues raised in both countries. Forty-one items were selected that best expressed the issues raised by the interviewees.

Table 1. Demographic and disease information (postal surveys). Results are shown as No (%)

	First postal survey (stage 4)	Second postal survey (stage 5)	
	UK	UK	The Netherlands
	n=121	n=210	n=154
Demographic details			
Males (%)	92 (76)	150 (71)	110 (71)
Females (%)	29 (24)	59 (28)	44 (28)
Age range (years)	21-77	19-82	20-79
Mean age (SD) (years)	47.6 (12.4)	46.1 (12.4)	47.6 (11.8)
Married or living as married (%)	85 (72)	144 (68)	126 (82)
Disease information			
Range of duration of illness (years)	1.5-50	1-62	3-51
Median (mean) duration of illness in years	15 (16.3)	18 (19.6)	19 (20.8)
No. reporting peripheral involvement (%)	85 (70)	174 (83)	112 (71)
No. reporting uveitis (%)	30 (25)	48 (23)	36 (23)
No. reporting IBD (%)	14 (12)	31 (15)	22 (14)
Perceived AS severity			
Mild (%)	18 (15)	22 (11)	36 (25)
Moderate (%)	50 (42)	76 (37)	69 (47)
Quite severe (%)	45 (38)	93 (45)	35 (24)
Very severe (%)	5 (4)	16 (8)	6 (4)
Perceived general health status			
Excellent / very good (%)	12 (10)	10 (5)	7 (4.6)
Good (%)	48 (40)	58 (28)	48 (31.6)
Fair (%)	43 (36)	100 (49)	83 (54.6)
Poor (%)	18 (15)	38 (18)	14 (9.2)
Perception of today			
Very good (%)	11 (9)	11 (5)	13 (8)
Good (%)	71 (61)	101 (50)	95 (62)
Bad (%)	34 (29)	83 (41)	42 (27)
Very bad (%)	1 (1)	8 (4)	4 (3)

Field-testing for face and content validity (stage 3)

In the UK 10 patients were interviewed in clinic and 5 in their home. In the Netherlands all 15 patients were interviewed in clinic. The ASQoL took between 2 and 16 minutes to complete (median 4 minutes in both the UK and NL). The measure was well accepted by interviewees in both countries, who generally found the items to be easily understood and relevant. Field testing of the questionnaire resulted in minor changes to the wording of two items and the removal of five more from both language versions. Items were removed because they were found to be problematic or were considered inappropriate by a number of respondents. For example, the item *I find it difficult to get moving in the morning* was among those deleted, as it was interpreted in different ways by UK respondents. The item *I often have to rest when doing jobs around the house* was removed due to gender bias. Although the item was intended to cover a range of household tasks such as cooking, cleaning, decorating or home maintenance, it was generally construed by patients in the UK to be solely related to housework. Many male respondents in the UK commented that they never undertook such tasks and, consequently, could not answer the question. Following these changes, a 36-item version of the ASQoL was produced for use in the first postal survey.

Testing the psychometric and scaling properties of the ASQoL

For both versions of the measure, a high score indicates worse QoL. For all tables in the following sections, *n* values deviating from the overall number are due to individual missing responses.

Results of the first postal survey (UK) (Stage 4)

Questionnaire packs were distributed to 180 people and returned by 121, a response rate of 67%. Table 1 shows the demographic details of the sample. Rasch analyses were performed on the data to identify items that were problematic in terms of misfit or DIF. While a number of items were found to misfit, DIF was minimal. As a result of these analyses, 10 items were removed from the measure, leaving a 26 item version of the ASQoL. This version was taken forward for further testing in each country.

Results of the second postal survey (Stage 5)

In the UK, 288 questionnaires were distributed at time 1 and 210 were returned, a response rate of 73%. Of these, 157 (75%) were returned at time 2. In the NL, 180 questionnaires were distributed at time 1 and 158 were returned, giving a response rate of 88%. Of these, 139 (88%) were returned at time 2. Four questionnaire sets from the Dutch sample were returned too late to be included in the analyses. Table 1 shows demographic details of the samples at time 1 in the UK and the NL. It can be seen from the table that the samples included in the postal surveys were similar demographically. Demographic characteristics of respondents at time 2 were also comparable. Table 1 also provides information on the respondents' perceived health status. The table shows that the UK respondents rated their health status worse than the Dutch participants. Respondents' scores on the comparator

instruments showed that, with the exception of social isolation, perceived distress (as shown by NHP section scores) is high for this patient sample and higher in the UK than in the NL. Rasch analyses were conducted on the data from each country. Eight items were removed as they were shown to misfit in one or both countries. The fit of the final 18 item ASQoL was good in both countries, with most MNSQ values within the required 0.7 to 1.3 range (table 2). Item stability over time was excellent in both countries, with Rasch item parameter estimates similar at times 1 and 2 (within 95% confidence intervals). Items were not equally spaced along the measurement continuum, indicating that the 18 item ASQoL produces raw scores at the ordinal level of measurement.

Scores on the 18-item ASQoL can range from 0-18. Median scores for the UK were 10.0 (Inter-quartile range (IQR) 5.0-14.0; mean 9.5, standard deviation (SD) 5.3) at time 1 and 9.0 (IQR 4.0-14.0; mean 8.8, SD 5.7) at time 2. For the NL, median scores were 6.0 (IQR 2.0-10.0; mean 6.7, SD 4.8) at time 1 and 6.0 (IQR 1.5-9.0; mean 6.2, SD 4.8) at time 2. Relatively few respondents scored at the extremes, although the basement effect was greater in the NL.

Table 2. Rasch item statistics for the 18-item ASQoL in the UK and the Netherlands

Item number	Mean square fit statistic (MNSQ)			
	UK	Netherlands		
	Infit	Outfit	Infit	Outfit
1	.90	.74	.91	.78
2	1.21	1.22	.93	.84
3	.95	.87	.97	.88
4	.79	.77	.84	.63
5	1.02	.70	1.25	1.23
6	.88	.75	.87	.71
7	1.36	1.87	1.00	1.06
8	.83	.65	.93	.68
9	.99	.96	.99	.91
10	1.13	.84	1.08	1.25
11	.86	.71	1.20	1.15
12	1.20	1.40	1.06	1.12
13	.99	.90	.89	.76
14	.88	.92	1.37	1.16
15	.98	.90	.90	.74
16	1.16	1.06	.98	1.08
17	1.11	1.02	.76	.76
18	.87	.73	1.19	1.16

Bold italics, MNSQ values above 1.3; *Italics*, MNSQ values below 0.7

Association with additional factors

ASQoL scores were not related to duration of illness or to the presence of uveitis. Patients with IBD scored higher on the measure (indicating worse QoL) than those without (UK $p<0.01$, NL $p<0.005$; Mann-Whitney U test).

Reliability and internal consistency of the ASQoL

The Spearman rank correlation coefficients for the test-retest reliability of the 18-item ASQoL was 0.92 in the UK ($n=129$) and 0.91 ($n=119$) in the NL, indicating that the measure has excellent reliability, producing low levels of random measurement error. Identical intraclass correlation coefficients were obtained (0.92 in the UK and 0.91 in the NL). Very few patients (two in the UK and one in the NL) reported any significant change in perceived general health, severity of illness or perceived disease activity. Therefore, removing such patients made little difference to the results obtained. The ASQoL also has good internal consistency in both countries (0.91 at time 1 and 0.92 at time 2 in the UK and 0.89 at time 1 and 0.90 at time 2 in the NL).

Table 3. Correlations* between scores on the 18-item ASQoL and those on the comparator measures

Comparator measure	UK Time 1	Netherlands Time 1
NHP sections		
Physical Mobility	.78	.79
Energy	.74	.73
Pain	.81	.79
Emotional reactions	.72	.73
Sleep	.54	.59
Social isolation	.53	.50
BASFI	.72	.75
LDQ	.70	-
DFI	-	.80

* Spearman rank correlation coefficients

Validity of the ASQoL

Evidence of construct validity was provided by examining the levels of association between the ASQoL and the comparator instruments. Moderate to high correlations were found between the ASQoL and all the comparator instruments (table 3). The pattern of association between the NHP section scores and the ASQoL was as expected, with the highest correlations being with the physical mobility, pain and energy level sections. The correlations with the emotional reactions section were also high. Further evidence of the validity of the ASQoL was gained through investigating the measure's ability to distinguish between

Table 4. ASQoL scores by specified groups

Grouping factor	UK			Netherlands		
	n	Median	IQR	n	Median	IQR
Disease Activity (good day / bad day)	p<.001*	p<.001*				
Very good / good	102	7.0	3.0-11.3	101	5.0	1.5-9.0
Very bad / bad	81	13.0	10.0-15.5	40	10.0	8.0-12.7
Perceived general health	p<.001**	p<.001**				
Excellent / good	61	3.0	1.0-7.0	52	2.0	0-4.0
Fair	91	11.0	8.0-13.0	76	9.0	6.0-10.7
Poor	34	15.0	14.0-17.0	12	13.5	12.0-14.7
Perceived AS severity	p<.001**	p<.001**				
Mild	19	3.0	1.0-5.0	34	1.5	0-4.2
Moderate	73	7.0	3.0-10.0	61	6.0	4.0-10.0
Quite / very severe	95	13.0	11.0-16.0	39	9.0	8.0-12.0

* Mann-Whitney U Test; ** Kruskal-Wallis one-way analysis of variance.

specified groups of patients (known groups validity). Table 4 shows that ASQoL scores differed significantly by whether the respondent was having a good or a bad day (disease activity); self perceived general health status, and self perceived AS severity.

Discussion

The efficient and cost effective management of any disease requires competing treatment regimes to be evaluated in terms of their ability both to control the disease and improve the QoL of patients. Existing instruments for use with subjects with AS focus on symptoms and functioning. Although these provide important information they do not provide information about the overall impact of the condition and its treatment on the patient's QoL. The ASQoL is based on a clear, conceptual model of QoL that has been successfully employed in the development of several other disease-specific QoL instruments^{17,19,20,21,22,23,24,25,26}. The development process was conducted in parallel in the UK and the NL. Consequently, it was possible to remove items that were problematic in one or other language version of the instrument at each stage of the testing procedure. This method of development is preferable to the standard one in which an instrument is produced in one country and then adapted for use in other languages. Such sequential development cannot overcome cultural and linguistic differences between countries. The content of the measure was derived from interviews with individuals diagnosed with AS in the UK and the NL. For each language version, the items are expressed (as far as

possible) in the original words of the patients. Consequently, respondents find the instrument acceptable, comprehensive and relevant to their condition. The ASQoL is quick and easy to complete (taking less than five minutes), making it suitable for use in clinical settings.

Application of item response theory in the form of the one parameter Rasch model showed the ASQoL to be unidimensional, had good item stability over time, and to have minimal DIF. The reliability of each language version of the measure has been shown to be excellent - the test-retest reliability coefficients obtained indicate that the ASQoL is suitable for use in routine clinical practice or for monitoring the progress of individual patients. Internal consistency was also adequate. It is essential to establish that a new instrument has construct validity, that is, that it is measuring the intended construct. Two prerequisites of this are; that the instrument is based on a model of the construct assessed and that it has good reliability³⁵. These requirements were met in both countries and hence, it is possible to infer that the ASQoL provides a valid assessment of the construct defined in the model. However, it is also necessary to determine construct validity formally through association with instruments measuring related constructs (convergent validity) and by comparing scores of patients at different stages of disease activity or with different disease severity (known groups validity). For the ASQoL, formal assessment was undertaken by correlating scores on the ASQoL with those on the NHP and the BASFI. ASQoL scores in the UK were also correlated with the LDQ and in the NL with the DFI. These comparator instruments measure a range of constructs; the NHP assesses perceived distress, while the BASFI, LDQ and DFI measure AS-specific disability. The relatively high levels of association between the ASQoL and these different constructs reflects the multifaceted nature of the impact of the disease on the patient. For example, pain, being a prominent feature of AS, would be expected to have a major influence on the QoL of the patient and, indeed, the correlation between these two measurements indicates approximately 66% shared variance. Similarly, QoL was moderately highly correlated with physical disability, energy and emotional reactions sections of the NHP. The results obtained show that the ASQoL and comparator instruments measure different though related constructs. Taken together, they provide a more complete picture of the impact of AS than any single measure can give alone.

The psychometric and scaling properties of the ASQoL suggest that researchers and clinicians can have confidence in the scores obtained by respondents on the measure. Further assessments of the instrument's validity will be possible as it is used in clinical studies. In addition, it is recommended that future studies are carried out to assess responsiveness, the instruments ability to detect meaningful changes in QoL.

The decision to adopt a dichotomous response system for the instrument was driven by practical issues related to language equivalence and ease of completion and scoring. There is often an assumption that such simplification is at the cost of some loss of sensitivity because it is presumed that multiple response items are able to provide more detailed information about the variable of interest. However, this assumption is not necessarily correct²². The ASQoL comprises 18 dichotomous items that have been shown through

Rasch analysis to form a single scale. Furthermore, the results from the assessment of known groups validity suggest that this scale is able to measure the QoL associated with a wide range of perceived disease severity and activity.

The ASQoL will serve as a valuable tool for assessing the impact of AS and its treatment on QoL in clinical settings and research studies. Such an instrument will allow accurate assessment of the effectiveness of interventions from the patient's perspective.

Acknowledgements

We would also like to thank Vicky Wilkinson at the University of Leeds and Gisela Mulder at the University Hospital Maastricht for their assistance in administering the postal surveys. This work in the UK was funded by the NHS Research & Development Programme.

References

- 1 Ryall NH, Helliwell PS. A critical review of ankylosing spondylitis. *Crit rev phys and rehabil* 1998; 10: 265-301.
- 2 Gran JT, Husby G. Ankylosing spondylitis: prevalence and demography. In: Klippel JH, Dieppe PA. *Rheumatology*. St Louis: Mosby. 1998; 6: 15.1-15.6.
- 3 Khan, MA. Ankylosing spondylitis: clinical features. In : Klippel JH, Dieppe PA. *Rheumatology*. St Louis : Mosby. 1994; 3: 25.1-25.10
- 4 Chamberlain MA. Socio-domestic and psychological factors in management. In Moll JMH, ed. *Ankylosing Spondylitis*. London. London: Churchill Livingstone 1980: 22735.
- 5 Lubrano E, Helliwell PS. Deterioration in anthropometric measures over six years in patients with Ankylosing spondylitis: An initial comparison with disease duration and reported exercise frequency. *Physiotherapy* March 1999; 85: 138-43.
- 6 Chamberlain MA. Socio-economic effects of ankylosing spondylitis in females : a comparison of 25 female with 25 male subjects. *Int Rehabil Med* 1983; 5: 149-53.
- 7 Doward LC, McKenna SP. Evolution of quality of life assessment. In Rajagopalan R, Sheretz EF, Anderson RT (Eds). *Care management of skin diseases: life quality and economic impact*. New York: Marcel Dekker, 1997: 9-33.
- 8 Calin A, Garrett S, Whitelock L, Kennedy G, O'Hea J, Mallorie P et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994; 21: 2281-5
- 9 Abbott CA, Helliwell PS, Chamberlain MA. Functional assessment in ankylosing spondylitis: evaluation of a new self-administered questionnaire and correlation with anthropometric variables. *Br J Rheumatol* 1994; 33: 1060-6.
- 10 Nemeth R, Smith F, Elwood, J, Calin A. ankylosing spondylitis - an approach to measurement of severity and outcome: Ankylosing Spondylitis Assessment Questionnaire (ASAQ) - a controlled study. *Br J Rheumatol* 1987; 26 (Abstract, Supp 1): 69-70.
- 11 Dougados M, Gueugen A, Nakache JP, Nguyen M, Mery C, Amor B. Evaluation of the functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988; 15: 302-7 .
- 12 Daltroy LH, Larson MG, Liang MH. A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. *J Rheumatol* 1990; 17: 946-50.
- 13 Guillemin F, Chailier B, Urlacher F, Vançon G, Pourel J. Quality of life in ankylosing spondylitis: Validation of the ankylosing spondylitis arthritis impact measurement scales II, a modified arthritis impact measurement scales questionnaire. *Arthritis Care Res* 1999; 12: 157-62.
- 14 Heinemann AW, Whiteneck GG. Relationships among impairment, disability, handicap and life satisfaction in persons with traumatic brain injury. *J Head Trauma Rehabil* 1995; 10: 54-63.
- 15 McKenna SP, Whalley D, Doward LC. Which outcomes are important in schizophrenia. *Int J Meth Psychiatric Res* 2000; 9 Suppl: S58-67.
- 16 McKenna SP. Quality of life assessment in the conduct of economic evaluations of medicines. *Br J Med Economics* 1995; 8: 33-8.
- 17 Hunt SM, McKenna SP. The QLDS: a scale for the measurement of quality of life in depression. *Health Policy* 1992a; 22: 307-19.

- 18 Rasch G. Probabilistic Models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1980.
- 19 McKenna SP, Hunt SM. A new measure of quality of life in depression: Testing the reliability and construct validity of the QLDS. *Health Policy* 1992; 22: 321-30.
- 20 Holmes SJ, McKenna SP, Doward LC, Shalet SM. Development of a questionnaire to assess the quality of life of adults with growth hormone deficiency. *Endocrin Metabol* 1995; 2: 63-9.
- 21 Wagner TH, Patrick DL, McKenna SP, Froese PS. Cross-cultural development of a quality of life measure for men with erection difficulties. *Qual Life Res* 1996; 5: 443-9.
- 22 Whalley D, McKenna SP, de Jong Z, van der Heijde D. Quality of life in rheumatoid arthritis. *Br J Rheumatol* 1997; 36: 884-8.
- 23 de Jong Z, van der Heijde D, McKenna SP, Whalley D. The reliability and construct validity of the RAQoL: a rheumatoid arthritis-specific quality of life instrument. *Br J Rheumatol* 1997; 36: 878-83.
- 24 Doward LC, McKenna SP, Kohlmann T, Niero M, Patrick D, Spencer B et al. The international development of the RGHQoL: a quality of life measure for recurrent genital herpes. *Qual Life Res* 1998; 7: 143-53.
- 25 McKenna SP, Doward LC, Mackenzie Davey K. The development and psychometric properties of the MSQOL: a migraine-specific quality-of-life instrument. *Clin Drug Invest* 1998; 15: 413-23.
- 26 McKenna SP, Doward LC, Alonso J, Kohlmann T, Niero M, Prieto L, et al. The QoL-AGHDA: an instrument for the assessment of quality of life in adults with growth hormone deficiency. *Qual Life Res* 1999; 8: 373-83.
- 27 Bennett PH, Burch TA. New York symposium on population studies in rheumatic diseases: new diagnostic criteria. *Bull Rheum Dis* 1967; 17: 453-58.
- 28 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.
- 29 Linacre JM, Wright BD. A user's guide to WINSTEPS. Chicago: Messa Press, 1997.
- 30 Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. *J Outcome Meas* 1998; 2: 66-78.
- 31 Hunt SM, McKenna SP, McEwen J, Williams J, Papp E. The Nottingham Health Profile: Subjective health status and medical consultations. *Soc Sci Med* 1981; 15: 221-9.
- 32 Creemers MC, Van 't Hof MA, Franssen MJ, Van de Putte LB, Gribnau FW, Van Riel PL. A Dutch version of the functional index for ankylosing spondylitis: development and validation in a long-term study. *Br J Rheumatol*, 1994 September; 33: 842-6.
- 33 Weiner E, Stewart B. Assessing individuals. Boston: Little Brown, 1984.
- 34 Nunally JC Jr. Psychometric theory. (2nd edn). New York, McGraw-Hill, 1978.
- 35 Streiner D, Norman G. Health Measurement Scales. Oxford: Oxford University Press, 1989.

Chapter 7

RADIOLOGICAL SCORING METHODS IN ANKYLOSING SPONDYLITIS: RELIABILITY AND CHANGE OVER ONE AND TWO YEARS

Anneke Spoorenberg, Kurt de Vlam, Sjef van der Linden, Maxime Dougados, Herman Mielants, Hille van de Tempel, Désirée van der Heijde.

Journal of Rheumatology [in press]

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 de Tempel, Maarten Boers, Sjef van der Linden.

Journal of Rheumatology 1999; 26: 997-1002.

ABSTRACT

Objective: To compare reliability and change over time of radiological scoring methods in AS.

Methods: Two trained observers scored 217 sets of radiographs from baseline, one and two years follow up. The sacroiliac (SI) joints were grade 0-4 by the New York method and SASSS (Stoke Ankylosing Spondylitis Spine Score). Hips, cervical and lumbar spine were graded 0-4 (Bath Ankylosing Spondylitis Radiological Index, BASRI). BASRI spinal scores, and New York SI are combined in BASRI-spine (2-12) and with the addition of BASRI-hips in BASRI-total (2-16). Cervical and lumbar spine were also scored in detail (SASSS, 0-36 each) and combined in SASSS-total or 'modified' SASSS (both range 0-72). To assess change a smallest detectable difference (SDD) was estimated for data on a quasi interval scale.

Results: The SI scoring methods showed intra- and interobserver kappas between 0.36 and 0.70. The BASRI-hip reached kappa's between 0.59-0.84. Combined SASSS scores were most reliable, with intra- and interobserver ICC's between 0.90 and 0.96. The ICC's of the combined BASRI scores were also very good, ranging from 0.85 to 0.95. For SI New York, SI SASSS and BASRI-hip 0.3-1.2% of patients deteriorated ≥ 1 grade. 7.5% deteriorated ≥ 1 grade (6.3% of maximum score) in BASRI-spine and BASRI-total and observers agreed in up to 48% of the cases that no change occurred. The SDD was lowest (7.5; 10% of maximum score) for 'modified' SASSS. Only 0.8% of patients deteriorated more than the SDD and observers agreed in up to 92% of the cases that no change occurred.

Conclusion: Radiological scoring methods for AS are moderately to excellently reliable. Under the selected scoring conditions (concealed time order, average of two observers, SDD based on interobserver data, unselected patient population) there was too little change over two years to be picked up reliably by the scoring methods.

Radiological damage is considered an important outcome in Ankylosing Spondylitis (AS)ⁱ. The evaluation of radiological change proves to be very difficult. There are several reasons for this. Radiological sacroiliitis can easily be missed because of the complex anatomy of the sacroiliac joints. The undulating articular surfaces make it hard to image these joints on conventional radiographs. Squaring, erosions and sclerosis appear in different stages of the diseaseⁱⁱ and syndesmophytes must be differentiated from osteophytes and disorders such as diffuse idiopathic skeletal hyperostosis (DISH). Usually AS is a slowly progressive disease and radiological change appears gradually: evaluation of radiographs with an interval of one year does not seem to be useful^{3,4}. However, a detailed scoring method showed some change after a period of one year⁵ and change after two years of follow up could also be detected by a graded scoring method³.

Changes of the sacroiliac joints are most frequently scored using the 5 grade New York criteria^{vi} or the nearly similar Stoke Ankylosing Spondylitis Spine Score (SASSS)^{5,7}. To evaluate the spine in AS there are essentially two different scoring methods. The Bath Ankylosing Spondylitis Radiology Index (BASRI) is a global graded scoring method, quick and easy to perform^{8,9,10}. The first version of this BASRI was described in 1995⁸ and a modified version was published in 1998. Several BASRI scores are also combined in composite scores^{9,10}. The SASSS for the spine is a more detailed scoring method assessing different features such as squaring, sclerosis and erosions at various locations of each vertebra^{6,7,11}. In an earlier study we compared the reliability and change over time over one year of these scoring methods. We concluded that both the SASSS method for the spine and BASRI reached good reliability. All other scoring showed moderate reliability at best. No method showed change over a period of one year in a considerable number of patients⁴. At the time of our first study no scoring method was available to evaluate the hip in AS. Therefore the Larsen scoring method designed to score the hips in rheumatoid arthritis was used. Recently a new graded scoring method was developed to evaluate the hip in AS, the BASRI-hip³. In this second study we used the BASRI-hip. The objective of this follow-up study was to compare all available AS radiological scoring methods for reliability and change over 2 years time.

MATERIALS AND METHODS

Patients. A total of 217 consecutive outpatients who satisfied the modified New York criteria¹² were included in our study. Our study population concerns a cross sectional cohort of AS patients, followed longitudinally. 69% of patients were male, a distribution usually found in AS populations. The median age at baseline was 42.2 year (range:18-78). There is a striking difference between median duration of complaints (17.0 year, range: 0.3-54) and duration of disease since diagnosis (9.4 year, range:0.1-41), indicating AS patients have complaints long before the diagnosis is made.

Scoring methods SI joints. The SI joints were scored according to the New York (NY) method (0-4) and the Stoke Ankylosing Spondylitis Spine Score (SASSS; 0-4). Both methods score the lower half of the SI-joints and are graded from 0 to 4^{5,6,7}. The main difference between these methods is grade 4 (complete ankylosis), where the New York method does not allow residual sclerosis. Both SI joints are scored separately, and thereafter the score is summed.

Scoring method hips. The hips were scored according to BASRI-hips, graded: 0= normal, 1= suspicious (possible focal joint space narrowing), 2= minimal (definite narrowing, leaving a circumferential joint space > 2mm), 3= moderate (narrowing with circumferential joint space ≤ 2mm or bone-on-bone apposition of < 2cm), 4= severe (bone deformity or bone-on-bone apposition ≥ 2cm or total hip replacement)¹⁰. Both hips are scored separately, and thereafter the score is averaged.

Scoring methods spine. The Bath Ankylosing Spondylitis Radiology Index (BASRI) was developed for the AP and lateral view of the lumbar spine and the lateral view of the cervical spine and is graded 0 to 4 for each view⁹. The BASRI-spine (BASRI-s) is the composite score of the BASRI scored on the lumbar spine, the cervical spine and the SI-joints (NY method), with scores ranging from 2 to 12⁹. The highest score of the two views of the lumbar spine is applied in this BASRI-s. In case of absence of one of the lumbar spine radiographs, the score of the available view was taken into account. The BASRI-total (BASRI-t) is a composite score of the BASRI-s and de BASRI-hips, with scores ranging from 2 to 16¹⁰. The SASSS is scored from the lower border of 12th thoracic vertebra down to and including the upper border of the sacrum. This scoring method is used on both the anterior and posterior site of the vertebrae with a score ranging from 0 to 36 for each site, so the total score will range from 0 to 72^{5,7}.

The 'modified' SASSS is 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 sacrum¹¹. This scoring method is only used on the anterior site of the vertebrae, with a score ranging from 0 to 36 for the cervical spine and 0-36 for the lumbar spine. Therefore the total score of the modified version will also range from 0 to 72. Missing scoring sites for the SASSS were handled as follows: when up to 3 scoring sites for each view 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 score for that particular view was scored missing.

In case of all 'spine' scoring methods syndesmophytes were differentiated from osteophytes using the following description; an osteophyte was defined as a bony deformity on the edge of the vertebra projecting more than 0.5 cm horizontally. Osteophytes were not included in the analyses. The scoring methods for the SI joint and spine are described in more detail in our first study⁴.

Inter- and intraobserver reliability. To obtain inter- and intra observer reliability of the scoring method for the hips (BASRI-hip) two experienced observers (AS and KV) scored

30 randomly selected baseline radiographs from the 217 consecutive outpatients with AS. The University Hospital Maastricht, the University Hospital Gent and the Hôpital Cochin in Paris each provided 10 blinded radiographs of anterior-posterior view of the pelvis to score the hips. The two observers had a training session to gain experience with the scoring method. All abnormalities present on the radiographs were discussed in detail. After this training session the observers scored a set of radiographs independently and discussed the results with each other and this session was followed by a consensus meeting with the two observers and two other experts in AS. The study on BASRI-hip reliability was started when few (≤ 5) discrepancies existed between the two observers. Two different sets of 30 radiographs were used for training and again a different set of 30 radiographs was used for assessment of reliability of BASRI-hip. In case of BASRI-hip interobserver reliability was calculated and in addition intra observer reliability based on the scores of the radiographs scored a second time after 2 weeks.

For all other scoring methods of the sacroiliac joints and the spine inter- and intraobserver reliability was assessed in our first study⁴. Baseline and 1 year radiographs were scored during our first study and again in this present study. Intraobserver reliability could be computed also using data from baseline and 1 year of our first study and this present study except for BASRI-total because the BASRI-hip scoring method which is part of BASRI-total was not available at the time of our first study.

This second intraobserver reliability was based on an interval of 2 years between the first and the second read. Interobserver reliability could also be calculated for baseline, 1 year and 2 year data of this study.

Change over time over two years. We included 217 consecutive outpatients who satisfied the modified New York criteria for AS¹²; 137 patients from the University Hospital Maastricht and Maasland Hospital Sittard (the Netherlands), 55 patients from the Hôpital Cochin, Paris (France) and 25 patients from the University Hospital Gent (Belgium). These hospitals are secondary and tertiary referral centers. Of these 217 patients we studied 3 sets of radiographs taken with an interval of one year. All three sets of radiographs were scored viewing the radiographs simultaneously (paired) without knowledge on the chronology of the radiographs in a random order by the same two experienced observers independently (AS and KV). The scoring methods were also scored in random order. As a result sacroiliac joints were scored separately from the hip joints. We used the scoring methods as described in the previous section.

Statistics. Inter- and intra- observer agreement of the different scoring methods was analyzed for categorical data by the linear weighted Kappa (k) statistic and for continuous data by the random effects, average measure Intraclass Correlation Coefficient (ICC, type 3.1) with observer as fixed facet¹³. Joint pairs (hips and SI- joints) were regarded as independent units, i.e., their possible correlation was ignored. To visualize the observer agreement we plotted the continuous data using the 'Bland & Altman' method¹⁴. Change over time of the scoring methods

Table 1: Inter- and Intra-observer reliability of all scoring methods with lower and upper border of the 95% confidence interval (baseline, 1 year and 2 year)

Method:	interobserver agreement			Intra-observer agreement the radiographs are scored with a 2 year interval			
	observer 1 and 2			observer 1		observer 2	
	T0	T12	T24	T0	T12	T0	T12
SI-joints N-York 0-4	k=0.68 0.61-0.72 N=400	k=0.69 0.62-0.74 N=388	k=0.66 0.62-0.72 N=365	k=0.63 0.56-0.69 N=348	k=0.55 0.47-0.63 N=342	k=0.40 0.32-0.48 N=348	k=0.40 0.32-0.48 N=339
SI-joints SASSS 0-4	k=0.68 0.63-0.72 N=400	k=0.69 0.64-0.74 N=388	k=0.70 0.64-0.74 N=365	k=0.67 0.63-0.74 N=352	k=0.62 0.56-0.64 N=344	k=0.42 0.34-0.49 N=348	k=0.36 0.28-0.43 N=339
BASRI-hip 0-4	k=0.60 0.54-0.67 N=30	k=0.59 0.52-0.66 N=30	k=0.60 0.52-0.67 N=30				
BASRI-s 2-12	ICC=0.92 0.90-0.94 N=192	ICC=0.94 0.93-0.96 N=190	ICC=0.93 0.91-0.95 N=176	ICC=0.89 0.84-0.92 N=154	ICC=0.90 0.86-0.92 N=163	ICC=0.85 0.79-0.89 N=144	ICC=0.86 0.80-0.89 N=161
BASRI-t 2-16	ICC=0.94 0.92-0.95 N=192	ICC=0.96 0.94-0.97 N=190	ICC=0.95 0.93-0.96 N=176				
SASSS modified 0-72	ICC=0.98 0.97-0.98 N=162	ICC=0.97 0.96-0.98 N=172	ICC=0.97 0.96-0.98 N=153	ICC=0.96 0.95-0.97 N=154	ICC=0.96 0.95-0.97 N=164	ICC=0.96 0.95-0.97 N=136	ICC=0.95 0.94-0.97 N=155
SASSS total 0-72	ICC=0.98 0.97-0.98 N=161	ICC=0.98 0.97-0.98 N=163	ICC=0.98 0.97-0.98 N=157	ICC=0.93 0.92-0.94 N=152	ICC=0.96 0.95-0.97 N=162	ICC=0.92 0.91-0.93 N=140	ICC=0.95 0.94-0.96 N=156

was assessed by a cut-off value based on interobserver reliability of change. For grading scales, such as all BASRI methods and the New York and SASSS method for the sacroiliac joints, a change of one grade was defined as the minimum detectable difference. For data on a semi-continuous scale, such as the SASSS methods for the spine, a smallest detectable difference (SDD) was estimated in the situation of 2 fixed observers yielding a mean change¹⁵. The SDD is the smallest change that can be detected apart from measurement error. In case of the BASRI scores and the SI scores, these are categorical scales, which do not allow to calculate a SDD.

RESULTS

Inter- and intra-observer reliability of BASRI-hip (single radiographs).

BASRI-h (0-4) scores of 30 baseline radiographs showed good to very good reliability with intraobserver kappa's of 0.73 and 0.84 and interobserver kappa 0.63 and 0.66.

Inter- and intra-observer reliability of all scoring methods (baseline, 1 and 2 years).

SI scoring methods (New-York and SASSS, 0-4) showed moderate to good intraobserver reliability with kappa ranging from 0.36 to 0.67. For both scoring methods interobserver reliability was good with kappa's between 0.66 and 0.70 (*Table I*).

The BASRI grading scores of the various parts of the spine (0-4) showed moderate to good reliability. For BASRI scores of the lumbar spine the intraobserver kappa with two year interval between the scoring sessions ranged from 0.61 to 0.65 and interobserver kappa's from 0.58-0.78. BASRI of the cervical spine showed intraobserver kappa's ranging from 0.41 to 0.56 and interobserver kappa's from 0.61 to 0.62. BASRI-h (0-4) scores again showed good interobserver reliability, kappa's ranging from 0.59 to 0.60 (*Table I*). Intraobserver reliability could not be calculated because at the time our first study was performed the scoring method for BASRI-h was not available.

The combined BASRI scores showed good to excellent reliability. For the BASRI-spine (2-12) the intraobserver ICC ranged from 0.85 to 0.90 and the interobserver ICC from 0.92 to 0.94. This was even slightly better for BASRI-total (0-16) with ICC's ranging from 0.94 to 0.96 for interobserver reliability (*Table I*). Intraobserver reliability could not be computed because the BASRI-hip scoring method which is part of BASRI-total was not available at the time of our first study.

The SASSS scores also showed excellent reliability. The SASSS scored on the anterior and posterior site of the lateral view of the lumbar spine (both 0-36) showed intraobserver ICC 0.94-0.95 and interobserver ICC ranging from 0.94-0.98.

The combined score, SASSS-total (0-72), showed intra- and interobserver ICC's of 0.92-0.96 and 0.98 respectively (*Table I*). The SASSS score applied on the anterior site of the lateral view of the cervical spine showed intra- and interobserver ICC's of 0.92-0.96 and 0.95-0.96 respectively. The combined score of the anterior sites of both the lateral view of the lumbar and cervical spine ('modified' SASSS, 0-72) showed good intra- and inter observer ICC's of 0.95-0.96 and 0.97-0.98 respectively (*Table I*).

Table 2: Concordance rate (%) observer 1 and 2

	SI SASSS	SI New York	BASRI-hip	BASRI-spine	BASRI-total	SASSS-total lwkt ant and post	'modified' SASSS lwkt ant and cwkt ant
	N=434 range 0-4	N=434 range 0-4	N=434 range 0-4	N=217 range 2-12	N=217 range 2-16	N=217 range 0-72	N=217 range 0-72
*perfect agreement	70%	74%	66%	32%	38%	35%	22%
< 2 grades difference				69%	72%		
< 6 points difference						78%	78%
1 year							
*perfect agreement	71%	76%	64%	35%	36%	33%	23%
< 2 grades difference				74%	75%		
< 6 points difference						80%	77%
2 years							
*perfect agreement	72%	76%	67%	31%	35%	27%	21%
< 2 grades difference				68%	68%		
< 6 points difference						77%	71%

* perfect agreement: < 1 grade/point difference between observers

Table 2 shows the concordance rates of the two observers at baseline, 1 year and 2 years follow-up for individual and combined scoring methods. For each moment perfect concordance rates are low for all scoring methods. The concordance rates for the combined SASSS methods are between 71% and 80%, accepting less than 6 points difference between the 2 observers on a scale from 0-72. Accepting less than 2 grades difference the concordance rates of BASRI-total (68-75%) are comparable with those of the combined SASSS methods accepting less than 6 points difference. One grade change in BASRI-total represents 6.3% of the maximum scoring range. One grade in the BASRI-total can be compared to 5 points in the combined SASSS methods which represents 4.9% of the maximum range.

To visualize observer agreement over the complete range of observed scores, figure 1 and 2 show Bland and Altman plots of baseline data and progression data of the modified SASSS. Progression data are based on the difference between baseline data and data of two year follow up. The Bland and Altman plot of baseline data of the modified SASSS shows a maximum difference of 26 points between the two observers on a scoring range from 0-72; the 95% limits of agreement of the difference between the two observers is ± 1.96 times the SD (4.4) (figure 1). Observer 1 scores systematically somewhat higher than observer 2. The Bland and Altman plots of 1 and 2 year data are very similar to this baseline plot (data not shown). Figure 2 concerning the progression data of the two observers over two years

Figure 1. Bland and Altman plot: Mean versus difference of 2 observers at baseline; SASSS-modified (SASSS lumbar anterior and cervical anterior)

o = 1 patient, every dash represents an extra patient

Figure 2. Bland and Altman plot: mean versus difference of the progression scores of the two observers over 2 years; SASSS-modified (SASSS lumbar anterior and cervical anterior).

o = 1 patient, every dash represents an extra patient

Table 3. Summary statistics on the group level at baseline, after 1 year and 2 years follow-up.

Scoring method (range)	average of the 2 observers ¹ mean, SD, median, min-max		
	baseline	1 year	2 years
SI SASSS (0-4)	3.1, 0.8 3.0, 0-4	3.1, 0.8 3.0, 0-4	3.1, 0.8 3.0, 0-4
SI New York (0-4)	3.1, 0.8 3.0, 0-4	3.1, 0.7 3.0, 0-4	3.1, 0.8 3.0, 0-4
BASRI-hip (0-4)	0.7, 1.0 0.0, 0-4	0.7, 1.0 0.0, 0-4	0.6, 1.0 0.0, 0-4
BASRI-spine (2-12)	6.9, 2.8 6.6, 0-12	6.8, 2.7 6.5, 0-12	6.9, 2.8 7.0, 0-12
BASRI-total (2-16)	7.6, 3.4 7.1, 0-16	7.5, 3.3 7.1, 0-16	7.5, 3.3 7.0, 0-16
SASSS- total: lumbar anterior and posterior (0-72)	11.5, 17.5 4.0, 0-72	12.2, 17.9 5.0, 0-72	12.3, 18.0 4.5, 0-72
SASSS-modified: lumbar and cervical anterior (0-72)	13.8, 16.4 6.7, 0-72	13.7, 16.3 6.7, 0-72	15.0, 16.8 7.6, 0-72

shows a maximum difference of 18 points between the observers with a 95% limits of agreement of 7.45. Because the scores of BASRI are on categorical scales it is not allowed to calculate SDD and Bland & Altman plots.

Change over time.

Overall, we found only little change over the course of 2 years. No difference was found in mean, median and SD of the entire group for SI New York, SI SASSS, BASRI-hip score, BASRI-s and BASRI-t (Table 3). In case of SASSS (spine) there was little difference in mean, median and SD on baseline, 1 year and 2 years (Table 3). These differences did not reach statistical significance. The distribution of (dis)agreement based on the minimum detectable

difference of 1 grade for data on graded scales or the SDD for data on a semi-continuous scale is shown in Table 4. (see page 90) If a patient deteriorated or improved more than the SDD or 1 grade, the change was judged as real. This is reported for the percentage of patients that changed according to only one or according to two observers. For the graded methods SI New York, SI SASSS and BASRI-hip 0.3-1.7% deteriorated $\geq$ 1 grade according to both observers (Table 4). Although there was some change in mean, median and SD in SASSS spine over this 2 year period only very few patients (0-1.1%) deteriorated more than the SDD in the combined SASSS scores (Table 4). Only BASRI-s and BASRI-t were able to detect change over this 2 year period in a considerable number of patients, 7.5% and 7.4% respectively. To avoid a possible ceiling effect we performed the same analysis excluding maximum scores. For the graded scales we excluded grade 4 and for the BASRI combined scores and SASSS scores we excluded all data above the 75th percentile from analysis. These analysis did not influence the results (data not shown).

DISCUSSION

Scoring radiographs in AS remains very difficult. For most radiological scoring methods developed for AS moderate to excellent intra- and interobserver reliability could be reached by two well trained observers. However, only the combined BASRI scoring methods (BASRI-s and BASRI-t) and especially the SASSS showed good to excellent reliability. Even with a scoring interval of two years the interobserver reliability remained very good. Because of this two year scoring interval intra-observer agreement was less high in comparison with the interobserver agreement. The reliability of the relatively new scoring method for the hips (BASRI-h) proved to be good. We found it to be more reliable than the Larsen scoring method for the hips used in our first study^{16,4}. Hip involvement in AS often shows as bony formations, which cannot be scored properly using the Larsen method. The BASRI-h seems to be more disease specific and the developers of the BASRI-h, also found good to even excellent intra- and interobserver agreement using unweighted kappa's³. In contrast to our first study and most other studies^{3,4,8,9,12} we used linear weighted kappa statistics instead of unweighted kappa statistics in this present study. In comparison the value of unweighted kappa is lower than of weighted kappa because large and small differences in assessments between observers are judged equally in unweighted kappa statistics. Furthermore, kappa indicates to what extent two observers are capable to perceive differences between radiographs. So kappa often turns out to be relatively low in case of a homogeneous group where every single radiograph receives more or less the same score. This could be an explanation for the relatively low intra- and interobserver agreement of the SI scoring methods because patients were included if they fulfilled the modified New York criteria. SI joints were at least scored grade 2 for both sites on a scale from 0 to 4. Measures that relate observed to expected agreement (such as kappa and ICC) are only of limited value in this situation because of high levels of expected agreement. This

Table 4: Sensitivity to change of AS scoring methods (values are mean $\pm$ standard deviation of the difference)

Method	Interobserver agreement ¹ of change ²		
	0-1 year	1-2 year	0-2 years
SI-joints New-York (left and right, 0-4)	0.04 $\pm$ 0.32 P0 85.6% $\geq$ 1 grade change: P- 0.3%, P+ 0.3%, P(-) 7.2%, P(+) 6.6%	0.02 $\pm$ 0.24 P0 88.5% $\geq$ 1 grade change: P- 0.3%, P+ 0.3%, P(-) 5.2%, P(+) 5.7%	0.03 $\pm$ 0.29 P0 88.8% $\geq$ 1 grade change: P- 1.2%, P+ 0.6%, P(-) 7.5%, P(+) 3.7%
SI-joints SASSS (left and right, 0-4)	0.07 $\pm$ 0.17 P0 89.7% $\geq$ 1 grade change: P- 0%, P+ 0%, P(-) 5.6%, P(+) 4.5%	0.05 $\pm$ 0.14 P0 91.7% $\geq$ 1 grade change: P- 0%, P+ 0%, P(-) 4.6%, P(+) 3.7%	0.01 $\pm$ 0.17 P0 90.5% $\geq$ 1 grade change: P- 0.3%, P+ 0%, P(-) 5.2%, P(+) 4.0%
BASRI-hips (left and right) (0-4)	0.02 $\pm$ 0.24 P0 84.6% $\geq$ 1 grade change: P- 0.3%, P+ 0.0%, P(-) 8.2%, P(+) 1.6%	0.02 $\pm$ 0.26 P0 84.6% $\geq$ 1 grade change: P- 0.9%, P+ 0.0%, P(-) 7.4%, P(+) 6.8%	0.0 $\pm$ 0.25 P0 85.3% $\geq$ 1 grade change: P- 0.3%, P+ 0.0%, P(-) 7.3%, P(+) 7.3%
BASRI-S lumbar- cervical spine, SI-New York (2-12)	0.03 $\pm$ 1.28 P0 48.6% $\geq$ 1 grade change: P- 5.6%, P+ 2.8%, P(-) 22.5%, P(+) 8.5%	0.14 $\pm$ 1.0 P0 51.9% $\geq$ 1 grade change: P- 1.8%, P+ 1.2%, P(-) 28.7%, P(+) 17.4%	0.15 $\pm$ 1.22 P0 48.6% $\geq$ 1 grade change: P- 7.5%, P+ 1.9%, P(-) 32.2%, P(+) 18.7%
BASRI-total lumbar- cervical spine, SI-New York, hip (2-16)	0.02 $\pm$ 0.59 P0 49.7% $\geq$ 1 grade change: P- 5.6%, P+ 2.8%, P(-) 28.2%, P(+) 1.5%	0.11 $\pm$ 0.86 P0 50.7% $\geq$ 1 grade change: P- 2.4%, P+ 0.6%, P(-) 30.1%, P(+) 19.3%	0.11 $\pm$ 0.73 P0 48.1% $\geq$ 1 grade change: P- 7.4%, P+ 1.8%, P(-) 32.7%, P(+) 19.8%
SASSS-total lwk post + ant (0-72)	sdd 8.2 P0 89.9% $\geq$ sdd change: P- 0.7%, P+ 0%, P(-) 8.6%, P(+) 0.7%	sdd 6.8 P0 89.8% $\geq$ sdd change: P- 0%, P+ 0.7%, P(-) 4.4%, P(+) 5.8%	sdd 9.8 P0 92.3% $\geq$ sdd change: P- 0%, P+ 0%, P(-) 6.2%, P(+) 1.5%
'modified' SASSS lwk ant + cwk (0-72)	sdd 6.8 P0 89.9% $\geq$ sdd change: P- 1.3%, P+ 0%, P(-) 6.1%, P(+) 2.7%	sdd 6.2 P0 89.4% $\geq$ sdd change: P- 2.1%, P+ 0.7%, P(-) 4.2%, P(+) 3.5%	sdd 7.5 P0 92.1% $\geq$ sdd change: P- 0.8%, P+ 0%, P(-) 4.7%, P(+) 2.4%

¹ Mean and SD are calculated from the difference in scores of the 2 observers over two years.
Example: mean((score 2 years observer 1 - score baseline observer 1) - (score two years observer 2 - score baseline observer 2)).

² Level of reliability of at least 1 grade or SDD change.

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

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

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

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

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

is confirmed by the relatively low median scores for the SASSS-spine methods (Table 3, see page 95). The low prevalence of radiological damage in SASSS inflates the ICC statistics with a tendency to overestimate the ICC.

We also decided to show perfect concordance rates as a measure of (complete) agreement between the two observers not depending on statistical techniques used such as kappa and ICC. For all scoring methods the perfect concordance rates for the 2 observers were rather low. The developers of the BASRI method found good to excellent perfect concordance rates for the hips between 78 and 95%^{3,10}. They found good concordance rates (73-81%) for BASRI applied on lumbar and cervical spine and they reached comparable concordance rates for the SI New York method (78-86%)^{9,10}. Concordance rates for the SASSS method were not reported by the developers. Visual presentation of a Bland and Altman method adds to understanding the data especially because it visualizes the distribution of the data and outliers over the entire range of observed data. Visual presentation of agreement using the Bland and Altman method can only be applied reliably in scores with large ranges such as the SASSS and not for the various BASRI scores.

In our present study only BASRI-s and BASRI-t were able to detect change in a considerable number of patients over a two year period. This change could not be identified by the other graded and detailed scoring methods. In case of BASRI-s and BASRI-t observers agreed in up to 52% that no change occurred. Unfortunately we may still conclude that relevant change occurred rarely because observers agreed in only 7.5% of cases that real change of at least 1 grade occurred. A reason for this could be that observer variation or error can not be distinguished from radiological progression. An important reason could be that we followed an unselected group of patients, without a particular request for disease activity. In a group of AS patients selected for high disease activity, the situation might be different with a better signal to noise ratio. The developers of the BASRI-h found significant change after 1 year using Wilcoxon signed rank test for nonparametric data ($n=60$)³. For BASRI-s they found significant change after 2 years ($n=31$) and after 1 year 30% of 20 cases showed change of at least 1 grade but this was not significant⁹. In 1999 they reported the magnitude of change for the BASRI-s was from 7.0 to 7.9 in 2 years and 42% of 31 patients showed change in BASRI-s score¹⁰. In these studies change over time was not specified for BASRI-t. These results are based on a small number of patients could be caused by a selection of severe cases. The developers of the SASSS methods found significant change over a group of 28 patients in 1 year using Mann-Whitney U test with a mean change of 4.1 points (range 0-72) in SASSS-total and a mean change of 1.02 grade in SASSS for the sacroiliac joints⁷. In this study the order in which the radiographs were scored was known in contrast with our study. This can markedly influence the results, as has been shown for rheumatoid arthritis (RA)^{17,18,19}. All these sensitivity to change studies report over a relatively small number of patients.

Comparing all radiological AS studies available at the moment we recommend to use the New York method for the SI-joints because it is most widely used and the reliability is similar to the SASSS score for SI joints. The BASRI-hip should be used because it is the only AS

disease specific method for the hips available and it has good intra- and interobserver reliability. To score the spine the choice is not unequivocal. The BASRI-s and BASRI-t can be preferred above the SASSS methods by its feasibility. According to face validity, the BASRI and modified SASSS score highest because both include the cervical spine in addition to the lumbar spine. In our study the BASRI was the only method which showed change in a considerable number of patients over a two year period. However, this might be misleading information as we set a change of 1 grade arbitrarily as a cutoff. Looking at the concordance rates within 1 grade difference, only in about 70% of the cases the observers agree. For the SASSS the comparable data for concordance within 6 points is somewhat higher (in 78% of the cases). However, the calculated SDD for the SASSS is higher (9.8 for SASSS and 7.5 for modified SASSS). So the cutoff used for SASSS is very strict and that for BASRI is much looser. This might be an important reason why we were unable to detect changes if we applied the SASSS. Further study is needed with sets of radiographs in which progression of damage is likely, e.g. sets with a 5 year interval or in a population with AS with a short disease duration because these patients tend to show more radiological change or selected for high disease activity. Additional studies where AS radiographs are scored in both random and chronological order are warranted to assess the difference in methodology as was done for RA^{17,18,19}. Given the conditions used in this study (paired reading without information on sequence, average score of two observers, cut-off based on SDD on interobserver data, unselected patient population) the scoring methods are unable to detect change over two year time reliably in a considerable number of patients.

REFERENCES

- 1 van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden SJ. 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 Calin A, Makay K, Santos H, Brophy S. A new dimension to outcome. Application of the Bath Ankylosing Spondylitis Radiology Index. *J Rheumatol* 1999; 26: 988-92.
- 4 Spoorenberg A, de Vlam K, van der Heijde D, de Klerk E, Dougados M, Mielants H, van de Tempel H, van der Linden SJ. Radiological scoring methods in ankylosing spondylitis: reliability and sensitivity to change over one year. *J Rheumatol* 1999; 26: 997-1002.
- 5 Taylor HG, Wardle T, Beswick EJ, Dawes P. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. *Br J Rheum* 1991; 30: 330-5.
- 6 Dale K. Radiographic gradings of sacroiliitis in Bechterew's syndrome and allied disorders. *Scand Rheumatol* 1979; 32 (suppl. 32): 92-7.
- 7 Dawes PT. Stoke Ankylosing Spondylitis Spine Score. *J Rheumatol* 1999; 26: 993-6.
- 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 Rheum* 1995; 34: 767-70.
- 9 MacKay K, Mack C, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI). A new validated approach to disease assessment. *Arthritis Rheum* 1998; 41: 2263-70.
- 10 MacKay K, Brophy S, Mack C, Doran M, Calin A. The development and validation of a radiographic grading system for the hip in ankylosing spondylitis: the Bath Ankylosing Spondylitis Radiology Hip Index. *J Rheumatol* 2000; 27: 2866-72.
- 11 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.
- 12 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.
- 13 Shrout P, Fleiss J. Intraclass correlation: using in assessing rater reliability. *Psych Bull* 1979; 86: 420-8.
- 14 Bland JM, Altman DG. Comparing methods of measurement: why plotting differences against standard method is misleading. *Lancet* 1995; 346: 1085-87.
- 15 M. Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. *J Rheumatol* 1999; 26: 731-9.
- 16 Larsen A, Dale K, Morten E. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference film. *Acta Radiologica Diagnosis* 1977; 18: 481-91.
- 17 van der Heijde D, Boonen A, van der Linden SJ, 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.
- 18 Ferrara S, Priolo F, Cammisà 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.
- 19 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.

Chapter 8

SUMMARY AND GENERAL DISCUSSION

Summary and general discussion

To create more uniformity in studies concerning aspects of outcome and disease activity in AS, the international working group on ASsessment in Ankylosing Spondylitis (ASAS) defined 'core sets' for the following three settings: disease controlling anti-rheumatic therapy (DC-ART), symptom modifying anti-rheumatic drugs (SM-ARD)/physical therapy and clinical record keeping^{1,2}. The domains for all three core sets are physical function, pain, spinal mobility, spinal stiffness, patient global assessment and fatigue. The core sets for clinical record keeping and DC-ART were extended with the domains acute phase reactants, peripheral joints and enthesitis and the core set of DC-ART includes also radiographs of spine and hip. As a follow up of the work of the ASAS working group this thesis focuses on different aspects of disease activity and outcome in AS. The first part of the thesis (chapter 2-5) highlights aspects of disease activity and chapters 6 and 7 focus on outcome measures in AS. Most of the results described and discussed in the chapters are derived from an international observational multicenter project: the Outcome in Ankylosing Spondylitis International Study (OASIS). A total of 217 consecutive outpatients with AS who satisfied the modified New York criteria³ were included in OASIS. This cross sectional cohort of AS patients is followed longitudinally and patients are derived from outpatient clinics of several university hospitals and general hospitals in three European countries: 137 patients from the university hospital Maastricht and the Maasland hospital Sittard (the Netherlands), 55 patients from the hospital Cchin, Paris (France) and 25 patients from the university hospital Ghent (Belgium). These hospitals are secondary and tertiary referral centers. Approximately two third of the OASIS patients are male, a distribution usually seen in AS populations. At baseline of the study the mean age of the patients was 43 years (SD: 13 years) and the mean disease duration since diagnosis was 11 years (SD: 8 years). 27% of the patients had peripheral arthritis diagnosed by their treating rheumatologist. In each country the same trained person (2 rheumatologists and 1 research nurse) assessed all patients every six months according to a pre-specified protocol for a period of two years. All patients were followed by their rheumatologist, independently of the evaluations of the researchers.

Comparison of two functional indexes in AS

Physical function is both related to disease activity and damage in AS. The ASAS working group has also included physical function, assessed by the Dougados Functional Index (D-FI)⁴ or the Bath Ankylosing Spondylitis Functional Index (BASFI)⁵, in the core sets for all settings¹. In **chapter 2**, we compared these two widely applied and validated functional indexes specific for AS. The main purpose of this cross-sectional study was to investigate the relation of BASFI and D-FI to aspects of disease activity and damage, which are both related to physical function. If one of the two indexes would perform better, this could be

selected as the preferred measure to assess physical function. The BASFI consists of 10 questions on a visual analogue scale (VAS), all questions deal with activities of daily living. The final score is the average of the scores of the 10 items. The D-FI consists of twenty 5-point Likert response items, assessing the ability to perform distinct daily activities. The total score (ranging from 0-40) is calculated as the sum of the item scores. Because there is no 'gold standard' for disease activity in AS available we used three external criteria for disease activity: both patient and physician assessment of disease activity on a VAS, (0-10) and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI, range 0-10)⁶. A global and a detailed radiological scoring system specific for AS: the Bath Ankylosing Spondylitis Radiology Index-spine (BASRI-s, range 0-12)^{7,8,9} and the modified Stoke Ankylosing Spondylitis Spine Score (SASSS, range 0-72)^{10,11} were chosen as the external criterion for damage. In our analyses we used the most contrasting groups of AS patients in which disease activity on the three disease activity measures was defined as high (score ≥ 6.0) or low disease activity (score ≤ 4.0). Furthermore, Receiver Operator Curves (ROC) were plotted for both functional indexes with the three measures of disease activity.

Results of this study showed relatively low disease activity scores and functional scores in our patient population. A total of 7 BASFI questionnaires showed one or more missing answers versus 28 D-FI questionnaires. As expected both functional indexes were highly correlated with each other (Spearman: 0.89). Correlation of BASFI and D-FI with the disease activity measures were all comparable with the highest correlation for the BASDAI (Spearman: 0.59, 0.57 resp.). ROC's of the two functional indexes and all three disease activity measures showed the best curve with the highest sensitivity and specificity values for both functional indexes and the BASDAI (BASFI: 94%, 87% resp. and D-FI: 93%, 79% resp.). The cutoff values to determine high versus low disease activity were considerable higher for the BASFI ($\geq 40\%$ of the full scale) compared to D-FI ($\geq 20\%$ of full scale). However, using these cutoff values showed that considerable percentages of patients were misclassified (12-30%) as having high or low disease activity if solely based on their functional scores. The proportion of misclassified patients was lowest for the BASFI cutoff values in combination with disease activity measured with the BASDAI (12%). So, disease activity assessed with BASDAI comes most closely to the disease activity aspect of function assessed with BASFI. A reason for this may be that both BASDAI and BASFI are developed by the same research group and completely patient reported. Physical function in AS is not solely based on disease activity therefore it could not be expected that all patients were correctly classified. Of the external criteria chosen for damage, BASRI-s appears to give a relatively higher median score for radiological damage than the modified SASSS method: 7.0 (range 2-12) and 12.0 (range 0-72) respectively. Correlation of both functional indexes with BASRI-s and SASSS were about similar, 0.42 and 0.36 respectively.

Although, in this cross-sectional study, BASFI seems to perform slightly better assessing the disease activity aspect and in feasibility (BASFI takes less time to complete and less missing values were found) otherwise no definite choice can be made between BASFI and D-FI based on these results.

The BASFI was developed several years after the D-FI and avoids some presumably redundant items and items assessing symptoms instead of function¹². Furthermore the BASFI included three items that improve content validity. For two of these items this was proven in the development of the HAQ-s¹³. In case of the third item, concerning physically demanding activities, it was shown in rheumatoid arthritis¹⁴ and osteoarthritis¹⁵ that only this item might discriminate for almost but not perfectly healthy patients who would not score on any other item. At the other hand the BASFI comprises a few items reflecting unusual tasks, and the DFI includes several items covering additional domains not included in the BASFI. A literature review was published including all available studies comparing the performance of BASFI and D-FI. The one study concerning a head to head comparison of the two indexes showed that both instruments were able to discriminate inpatients from outpatients, but only the BASFI could discriminate the effects of a 3 weeks intensive physical therapy period¹². In four physical therapy trials the D-FI did not discriminate between the treatment arms while the BASFI could discriminate between treatment arms in two other physical therapy trials¹². A reason for this could be that the distribution of the D-FI scores show a tendency towards normal scores in most studies where it was used. This may not allow further measurement of improvement in patients with only mild disability. To improve the sensitivity of the D-FI the authors proposed the 5 point Likert response scale instead of the 3 point Likert response scale,

The D-FI discriminated well between treatment arms in all but one SMARD trial and in one DC-ART trial between treatment arms¹². The BASFI did not discriminate between treatment arms in one of the two SMARD trials where it was used¹². The one SMARD trial concerning a head-to-head comparison of the two indexes showed that both instruments discriminated equally well between placebo and treatment group¹². There is no DC-ART study available which shows the results of a direct comparison of the two indexes. The only study that evaluates a conventional DC-ART (salazopyrine) the D-FI was used and in three other studies evaluating the effects of inhibition of tumor necrosis factor α only the BASFI was used^{12,16,17,18}. In all these studies the applied instrument discriminated between the treatment arms. Based on all these results there is a slight preference for using the DIFI in SMARD trials whereas the BASFI should be preferred in trials concerning physical therapy. Given the efficacy of biological drugs in the treatment of AS direct comparison of both instruments should be preformed in future DC-ART trials and should include calculation of the effect sizes or standardized response mean of the instruments in these settings,

Acute phase reactants in AS

Both laboratory blood test, Erythrocyte Sedimentation Rate (ESR, mm/hr) and C-Reactive Protein (CRP, mg/l), are frequently used to evaluate disease activity in patients with AS. Assessments of acute phase reactants is also a recommended core set endpoint for DC-ART and clinical record keeping by the ASAS Working Group^{1,2}. It is well known that the

mean values of these two acute phase reactants are considerably lower for patients with AS in comparison with patients suffering from rheumatoid arthritis¹⁹. **Chapter 3** describes a study which was conducted to determine whether ESR or CRP is more appropriate in measuring disease activity in AS. Because there might exist differences with respect to ESR and CRP in AS patients depending on the clinical disease presentation, we divided our study population into two groups: patients with only spinal involvement (n=149) and patients with active peripheral arthritis and/or inflammatory bowel disease (n=42). Since there is no 'gold standard' for disease activity in AS we studied the relationship of CRP and ESR with three substitute clinical disease activity variables following the same methodology and statistical approach as described in the previous study on the comparison of BASFI and DFI. A second aim of this cross-sectional study was to determine if elevated CRP or ESR reflect active disease defined by one of these three selected disease activity variables.

The results showed that in the spinal group the majority of patients have normal values for ESR and CRP whereas the majority of patients in the peripheral arthritis/IBD group have slightly elevated levels for both acute phase reactants. Only for ESR this difference between the two groups was statistically significant. Thirty percent of patients in both disease subgroups showed either an elevated ESR with normal CRP or vice versa and in most of these cases values just above normal were seen in case the acute phase reactant was increased. Overall, relatively low disease activity scores were seen in both study groups. Quite striking is the difference in judgement of disease activity between the physician at one hand and both the patient derived measures (BASDAI and patient assessment of disease activity) on the other hand. This is reflected in very different mean values at baseline of physician assessment of disease activity versus BASDAI and patient assessment of disease activity (for the spinal group: 1.5 versus 3.6/3.9 resp. and for the arthritis/IBD group: 2.5 versus 4.3/4.1 resp.). Also in the spinal group only 3% of patients were classified as having high disease activity according to the judgment of the physician. In contrast, the percentage of patients in the spinal group with high disease activity defined by BASDAI and patient was 11% and 21% resp.. Overall, the ROCs showed low cutoff values for ESR and CRP in both groups. Based on these cutoff values sensitivity and specificity were reasonable with the highest sensitivity for physician assessment of disease activity (100%). Unfortunately the corresponding positive predictive values, which are of importance in clinical practice, were uniformly low with large percentages of misclassified patients. This cross-sectional study showed that neither ESR nor CRP are good reflections of active disease as defined in our study. So, no preference can be made to use either of these acute phase reactants in the assessment of disease activity in AS based on these results. There might be a difference between ESR and CRP in relation to the progression of damage in AS, which needs to be further evaluated. Other studies concerning a head-to-head comparison of ESR and CRP are difficult to interpret since different definitions of active disease activity in AS were used. A literature review was published including all available studies comparing the performance of ESR and CRP²⁰. In five of these seven studies ESR

and CRP performed equally and two studies indicated that CRP is more closely related to disease activity²⁰. There are two studies indicating that elevated ESR and CRP are more likely in patients with peripheral manifestations of AS²⁰.

There is no SMARD trial available concerning a direct comparison of ESR and CRP. In the available SMARD trials no discriminant capacity of the acute phase reactants was found²⁰. One of these SMARD trials also reports the standardized response mean for CRP which was low²¹. Nine DC-ART trials (one methotrexate trial and eight sulfasalazine trials) showed no significant effect of the tested drugs and possibly therefore CRP or ESR did not discriminate between therapy and placebo²⁰. The two DC-ART trials allowing a direct comparison of the two acute phase reactants provided opposite results²⁰. In two other studies, evaluating the effects of inhibition of tumor necrosis factor α , both ESR and CRP discriminated between treatment arms^{16,17}. Unfortunately the exact effect sizes or standardized response mean for ESR and CRP were not given, although these were high for both (>3). Based on all these results no definite decision can be made to use either ESR or CRP in all three clinical settings defined for AS. Since the promising results of biological drugs in the treatment of AS direct comparison of both acute phase reactants is possible and also calculation of the effect size or standardized response mean is needed. Furthermore, the relation of CRP and ESR with the progression of damage in AS should be investigated in future.

Patient self-assessed joint counts in AS

In **chapter 4** a reliability study on patient self-assessed swollen and painful joints is presented. In AS a minor part of the population has peripheral arthritis ($\pm 25\%$). Traditionally, either a physician or a well-trained healthcare professional is involved in the clinical assessment of arthritis. If joint counts assessed by a physician could be replaced by patient self-assessed joint counts, this would be an advantage for rheumatologists in clinical practice and especially for researchers. The reliability of patient self-reported joint counts in AS has never been studied. In rheumatoid arthritis, reliability of self-assessed joint counts is studied extensively and the results found are both of good reliability^{22,23,24,25,26} and poor reliability^{27,28,29,30}. In our study, 217 AS patients were asked to mark their painful and swollen joints on a mannequin designed after the method of Stewart²³ presenting 44 and 40 joints respectively. At the same day, without knowledge of the patient assessment, three investigators (one person for each research center) assessed painful and swollen joints on similar mannequins. Our results showed that, on a group level, there is a consistent difference between the number of tender and swollen joints assessed by the patients and by the physicians with only moderate agreement (Intraclass Correlation Coefficient between 0.51 and 0.71) on the total number of joint counts and even poor to moderate agreement (kappa between 0.23 and 0.64) on individual joint counts. Patients scored consistently more tender joints and the physicians scored more swollen joints. Possible explanations for these findings are: (1) AS patients can not differentiate between a tender

joint or pain caused by enthesitis since the entheses are located near the joint and (2) AS patients are not trained to detect a swollen joint. The enthesitis index of Mander was only significantly correlated with the total number of swollen joints assessed by the physician. On a patient level the results were even worse shown by visualizing our data with the Bland and Altman method³¹. Self assessed joint counts could still be valuable if patients could differentiate between the absence of arthritis and the presence of mono-, oligo- or polyarthritis. However, for these differentiations, perfect concordance rates between patients and physicians were also very low (17%, 17% and 22% resp.). The only good concordance rate found was in case of absence of swollen joints (82%). Consequently, AS patient can tell if their joints are not swollen but in case of swollen joints they are unable to judge the extent of swelling even within the rough categories of mono-, oligo-, or polyarthritis. We did not formally assess test-retest reliability in this study but the results obtained at baseline and after one year follow up showed similar results. Based on the results, joint scores derived by physicians cannot be replaced by patient self-assessed joint counts in AS in general. Only information from patients that there are no swollen joints is sufficiently reliable to be useful.

Disease activity in AS

Since there is extended variety in the clinical picture among different AS patients it is very difficult to define disease activity in AS. Patients may experience axial involvement in all degrees of severity, but may also have extra spinal manifestations. This clinical diversity, both in severity and in localization, makes a high demand on instruments that are supposed to measure disease activity in AS. Furthermore AS patients and rheumatologists seem to have very different understandings about active disease³². **Chapter 5** describes on which criteria AS patients and rheumatologists base their judgment on disease activity. Our goal was to explore differences between the patient and the physician perspective of AS disease activity. For this study, data of the OASIS patient cohort were used. The patients in this cohort may be considered to appropriately reflect the spectrum of AS patients seen by rheumatologists, since the patients were included irrespective of gender, age, disease duration, disease severity or disease activity,

In this study disease activity from patient perspective as well as from physician perspective was analysed by dichotomising both patient and physician global disease activity score on a VAS (VAS range: 0 not active and 10 extremely active) into 'high disease activity' (VAS $\geq$ 6.0) and 'low disease activity' (VAS $\leq$ 4.0). Various AS instruments selected by the ASAS working group were assessed every six months for two years. Data reduction of these instruments by principal components analysis (PCA) was performed and distinguished four factors capturing correlated instruments, therefore assumed to measure the same underlying construct: spinal mobility, physician assessments, patient assessments and laboratory assessments (Cronbachs alpha between 0.52 and 0.81; explained variance 61%).

Discriminant function analysis with the factor loadings was performed to discriminate between the low- and high disease activity state for both patient and physician perspective. This analysis showed that the factor patient assessments was most important (pooled correlation: 0.84) in discriminating between low and high disease activity state as defined by the patient. The other three factors contributed marginally (pooled correlation: <0.30). In contrast, the three factors: physician assessments, spinal mobility and laboratory assessments contributed most in discriminating between the two defined levels of disease activity of the physician perspective (pooled correlation: 0.62, 0.48, 0.48 respectively). The factor patient assessments did not contribute at all (pooled correlation: 0.05). The discriminant function analysis of baseline data and the analyses of data from other time points revealed similar information. Multiple regression analysis on the discriminant scores was performed to prioritise the instruments with respect to their contribution to each disease activity perspective. In case of the patient perspective disease activity was best captured by the instruments: 'pain spine', 'BASFI', 'pain joints' and 'fatigue'. The physician perspective was best captured by the instruments: 'cervical rotation', 'swollen joint count', 'CRP' and 'intermalleolar distance'.

According to our results AS patients and their physicians indeed have very different views on what disease activity in AS means. AS patients seem to rate disease activity on the basis of complaints while physicians rate disease activity on the basis of instruments assessing inflammation and disease severity. There are a few more conclusions that can be derived from this study. The BASFI, an index primarily designed to assess function in AS also contributes to disease activity from the patient perspective. It also seems that AS patients base part of their estimation of disease activity on what they are able to physically perform. Overall, AS patients appear to properly distinguish disease activity (defined by them as complaints) from disease severity. Disease activity from patient perspective is not captured by acute phase reactants. The physician judgement of disease activity is based on a combination of constructs including measures that combine information on disease activity and severity. Remarkably, CRP was included as a variable while this information was not available to the investigator at the time the judgement of disease activity was made. At the moment fully patient derived instruments such as BASDAI are combined with physicians' assessment of disease activity and/or elevated CRP are used as a 'gold standard' for the assessment of disease activity in AS for including patients in clinical trials and for the start of anti-TNF therapy in clinical practice. Furthermore, there is still lack of evidence that either patient- or physician derived assessments of disease activity are associated with long-term outcome in AS. This important information is needed to be able to select the instruments that best reflect real disease activity leading to the final outcome.

Since the options for drug therapy in AS are increasing it becomes more important to define measures assessing a uniform construct of disease activity and outcome to be used in clinical trials.

In chronic disabling conditions there is a growing interest in the assessment of quality of life (QoL). Especially in studies designed to assess the impact of new pharmaceutical products or to compare different treatment regimes it is becoming relatively common to measure QoL. Disease specific instruments used to evaluate the course of AS focus predominantly on physical impairment and/or physical functioning. Generic health status instruments are available but currently no disease specific instrument exists for assessing quality of life (QoL) in AS patients.

Chapter 6 describes the development of the Ankylosing Spondylitis Quality of Life questionnaire (ASQoL). Our goal was to produce a valid and reliable AS-specific QoL measure that would be relevant and acceptable to respondents. The ASQoL is a quality of life instrument specific to AS and was developed in parallel in the United Kingdom and the Netherlands. All included AS patients fulfilled the modified New York criteria. The methodology used to develop the ASQoL combines the theoretical strengths of the needs-based quality of life model³³ with the statistical and diagnostic power of the Rasch model³⁴. The development of the questionnaire enclosed five stages. The first content of the questionnaire was derived from interviews with 30 patients in the UK and 25 patients in the Netherlands (stage 1). Stage 2 concerned the selection of items and response format which formed the first 41-item draft-questionnaire. To assess face and content validity 15 patient field-test interviews were done in both the UK and NL which lead to a 36-item questionnaire (stage 3). Stage 4 concerned a postal survey in the UK (n = 121) to produce a more efficient version of the ASQoL with Rasch analysis the number of questions was reduced to 26 items. Rasch analysis of data from a final postal survey (UK: n = 164; NL: n = 154) was done to assess scaling properties, reliability, internal consistency and construct validity in each country (stage 5). This analysis showed some item misfit, but showed that items formed a hierarchical order and were stable over time. The problematic items were removed resulting in the 18 items ASQoL. Both language versions of the ASQoL showed excellent internal consistency (Conbach's alpha: 0.89-0.91), test-retest reliability (intraclass correlation coefficient: UK: 0.92; NL: 0.91), and validity. The ASQoL may serve a valuable tool in both clinical settings and research for assessing the impact of AS and its treatment on quality of life from the patients perspective. Independently of this study good reliability, validity and responsiveness of the ASQoL was found in another study comparing disease specific patient assessed measures of health outcome in AS³⁵. Since the development of the ASQoL this instrument was also used in two trials. One study evaluates the effects of spa therapy and the two other study evaluates the effects of inhibition of tumor necrosis factor α and both studies show that ASQoL discriminated between treatment arms^{36,18}. The standardized response mean (SRM) and effect size (ES) were only calculated in case of the spa therapy trial with moderate responsiveness scores (SRM: 0.24 and ES: 0.22) reflecting the moderate treatment effect. In a study on the effect of etanercept there was a high responsiveness of the ASQoL and at least similar to that of the BASDAI³⁸.

Radiological scoring methods in AS

Radiological damage is considered as an important outcome in AS. The evaluation of radiological change proves to be very difficult in AS. Changes of the sacroiliac joints (SI) are most frequently scored using the 5 grade New York criteria (0-4)³ or the nearly similar SI score described by the Stoke group¹⁰. To evaluate the lumbar and cervical spine in AS there are essentially two different scoring methods. The Bath Ankylosing Spondylitis Radiology Index (BASRI) is a global graded scoring method, quick and easy to perform and developed to score the lateral and anterior-posterior view of the lumbar spine (both views combined (0-4), the lateral view of the cervical spine (0-4)⁶ and the hips (BASRI-hip, 0-4)³⁹. The mean score of the New York scoring method of the SI-joints and the several BASRI scores are also combined in two composite scores: BASRI-spine (2-12) and BASRI-total (2-16)^{8,9}. The SASSS for the spine is a more detailed scoring method assessing different features such as squaring, sclerosis and erosions at various locations of each vertebra^{10,11}. This method is scored on the lateral view of the lumbar spine on both the anterior and posterior sites of the vertebrae (0-72). The 'modified' SASSS is scored on the lateral view of the lumbar spine only at the anterior site of the vertebrae and on the lateral view of the cervical spine also at the anterior site of the vertebrae (0-72)⁴⁰.

In **chapter 7**, we compared reliability and changes over one and two years of all these available radiological scoring methods in AS. Two well trained observers scored sets of radiographs from the OASIS cohort at baseline, one and two years follow up. These sets were scored viewing the radiographs simultaneously (paired) without knowledge of the chronology and in random order. The sets of radiographs available for reliability analyses varied from 136 to 200 depending on the number of missing scores for the various scoring methods.

Our results showed good intra- and interobserver reliability for almost all radiological scoring methods. For categorical data, observer agreement was analyzed with linear weighted kappa statistics and in case of continuous data with the Intraclass Correlation Coefficient (ICC). The combined BASRI scoring methods (BASRI-s and BASRI-t) and especially the SASSS showed excellent reliability (ICC 0.85-0.98). Even with a scoring interval of two years the intraobserver reliability remained very good (ICC 0.85-0.96). The reliability of the relatively new scoring method for the hips (BASRI-h) proved to be good (kappa 0.59- 0.60). Of consideration is that kappa indicates to what extent two observers are capable to perceive differences between radiographs. So kappa often turns out to be relatively low in case of a homogeneous group where every single radiograph receives more or less the same score. This could be an explanation for relatively low intra-and interobserver agreement found for the SI scoring methods (kappa 0.36-0.70). Furthermore, measures that relate observed to expected agreement (such as kappa and ICC) are of limited value in this situation because of high levels of expected agreement. This is also confirmed by the relatively low median scores for the SASSS-spine scoring methods (median SASSS-total 17.5-18.0, median modified-SASSS 16.3-16.8, range 0-72). Furthermore,

the low prevalence of radiological damage in SASSS inflates the ICC statistics resulting in a tendency to overestimate the ICC.

Because of these considerations concerning ICC and kappa statistics we also calculated concordance rates. The results showed that the perfect concordance rates between the two observers were overall low (21-76%). Also the visual presentation of a Bland and Altman method³¹ adds to the understanding of continuous data (SASSS method) especially because it visualizes the distribution of the data and outliers over the entire range of observed data. These plots showed a maximum difference of 26 points (possible range 0-72) between both observers,

In our study only BASRI-s and BASRI-t were able to detect change based on a binomial cut-off in a small percentage of patients over a two year period (7.5% and 7.4% resp.). This change could not be identified by the other graded and detailed scoring methods. In case of BASRI-s and BASRI-t observers agreed in up to 52% that no change occurred. Unfortunately we may still conclude that relevant change occurred rarely because observers agreed in only 7.5% of cases that real change of at least 1 grade occurred. However, this might be misleading information as we set a change of 1 grade arbitrarily as a cut off. The calculated smallest detectable difference (SDD) for the SASSS is relatively larger⁴¹. So the cut off used for SASSS seems to be very strict in comparison to the cut off used for the BASRI. This might be a reason why we were unable to detect changes if we applied the SASSS. Furthermore it could be that observer variation or error cannot be distinguished from radiological progression in our study. Moreover, the use of a binomial cut-off induces considerable loss of information and consequently loss of power to detect differences. Another consideration may be that we followed an unselected group of patients, without a particular request for disease activity. In a group of AS patients selected for high disease activity, the situation might be different.

In this study the reliability of AS scoring methods seems to be moderate till good. Unfortunately the scoring methods were unable to detect change over two-year time reliably in a considerable number of patients under the given scoring conditions (paired reading without knowledge of chronology, results based on average score of two observers, cut-off based on SDD, unselected AS population).

Résumé and perspective

Since there was a great need to create more uniformity in different studies focusing on AS the international ASAS working group was formed and this working group defined domains for three AS core sets (DC-ART, SM-ARD/physical therapy, clinical record keeping). In the past 5 years, as a follow up of the work of the ASAS working group, several study groups worldwide, which focus on disease activity and outcome in AS have done a lot of work. Results presented in this thesis are derived from a large cohort of AS patients followed longitudinally (OASIS) and relate to both aspects of outcome and disease activity. Chapter

6 and 7 are focusing on outcome measures in AS. Chapter 6 describes the development of a valid disease specific quality of life instrument (ASQoL). Chapter 7 describes the comparison of available AS radiological scoring methods. These scoring methods prove to be reliable but none of the methods showed considerable change in two year time. In detecting structural change in AS the role of Magnetic Resonance Imaging (MRI) may be become more clear in near future because with MRI it is possible to assess both features of damage and disease activity of the spine and sacroiliac joints in AS^{18,42,43}. In case of conventional radiography in AS the possible influence of aspects such as the knowledge of the chronology of the radiographs on sensitivity to change are under investigation, Chapter 2-5 highlights aspects of disease activity. The results presented in these chapters are all confirming that measuring aspects of disease activity in AS remains very difficult. Although there has been a lot of effort in studying disease activity in AS there is still no uniform measure which reflects AS disease activity in all its aspects. The acute phase reactants such as ESR and CRP are elevated in a minority of AS patients and of most important consideration is that AS patients and their treating physicians seem to have very different understandings about disease activity. Therefore, until now fully patient derived instruments such as BASDAI in combination with physicians assessment of AS disease activity and/or elevated CRP are used as a 'gold standard' to include patients in clinical trials as well as establish anti TNF α therapy in clinical practice. The main reason for this is the persistent lack of a valid tool which combines all aspects of disease activity in AS. Recently more potent biological drugs such as anti TNF α have come available in the treatment of AS and the effects of these drugs are very promising^{16,17,18}. In future studies evaluating the effects of these potent drugs can be used to validate and compare ASAS selected instruments used in the follow-up of AS and hopefully these studies will finally lead to the development of a disease activity measure which reflects AS disease activity in all its aspects.

REFERENCES

- 1 van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden SJ. Preliminary core sets for endpoints in ankylosing spondylitis. *J Rheumatol* 1997; 24: 2225-9.
- 2 van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden SJ, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. *J Rheumatol* 1999; 26: 952-4.
- 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 Dougados M, Gueguen A, Nakache JP, Nguyen M, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988; 15: 02-7.
- 5 Calin A, Garrett S, Whitelock H, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994; 21: 2281-5.
- 6 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.
- 7 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 Rheum* 1995; 34: 767-70.
- 8 MacKay K, Mack C, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI). A new validated approach to disease assessment. *Arthritis Rheum* 1998; 41: 2263-70.
- 9 Calin A, Makay k, Santos H, Brophy S. A new dimension to outcome. Application of the Bath Ankylosing Spondylitis Radiology Index. *J Rheumatol* 1999; 26: 988-92.
- 10 Taylor HG, Wardle T, Beswick EJ, Dawes P. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. *Br J Rheum* 1991; 30: 330-5.
- 11 Dawes PT. Stoke Ankylosing Spondylitis Spine Score. *J Rheumatol* 1999; 26: 993-6.
- 12 Ruof J, Stucki G. Comparison of the Dougados Functional Index and the Bath Ankylosing Spondylitis Functional Index. a literature review. *J Rheumatol* 1999; 26: 955-60.
- 13 Daltroy LH, Larson MG, Liang MH. A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. *J Rheumatol* 1990; 17(7): 946-50.
- 14 Sangha O, Büchi S, Klaghofer R, Rau R, Stucki G. Development of a new short instrument to assess functional status in rheumatoid arthritis patients [abstract]. *Arthritis Rheum* 1997; 40 Suppl: S111.
- 15 Stucki G, Liang MH, Phillips C, Katz JN. The short-form-36 is preferable to the SIP as a generic health status measure in patients undergoing elective total hip arthroplasty. *Arthritis Care Res* 1995; 8: 174-81.
- 16 Gorman JD, Sack KE, Davis JC. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor. *N Engl J Med* 2002; 18: 1349-56.
- 17 Braun J, Brandt J, Listing J, Zink A, Alten R, Golder W, Gromnica-Ihle E, Kellner H, Krause A, Schneider M, Sörensen, Zeidler H, Thriene W, Sieper J. Treatment of active ankylosing spondylitis with infliximab: a randomised controlled multicentre trial. *Lancet* 2002; 359: 1187-93.
- 18 Marzo-Ortega H, McGonagle D, O'Connor P, Emery P. Efficacy of Etanercept in the treatment of enthesal pathology in resistant spondylarthropathy. *Arthritis Rheum* 2001; 44: 2112-17.

- 19 Wolfe F. Comparative usefulness of C-reactive protein and erythrocyte sedimentation rate in patients with rheumatoid arthritis. *J Rheumatol* 1997; 24: 1477-85.
- 20 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.
- 21 Calin A, Nakache JP, Gueguen A, Zeidler H, Mielant H, Dougados M. Outcome variables in ankylosing spondylitis: evaluation of their relevance and discriminant capacity. *J Rheumatol* 1999; 26: 975-9.
- 22 Escalante A. What do self-administered joint counts tell us about patients with rheumatoid arthritis? *Arthritis Care Res* 1998; 11: 280-90.
- 23 Stewart MW, Palmer DG, Knight RG. A self-report articular index measure of arthritic activity: investigations of reliability, validity and sensitivity. *J Rheumatol* 1990; 17: 1011-5.
- 24 Abraham N, Blackmon D, Jackson JR, Bradley LA, Lorish CD, Alarcon GS. Use of self-administered joint counts in the evaluation of rheumatoid arthritis patients. *Arthritis Care Res* 1993; 6: 78-81.
- 25 Stucki G, Stucki S, Bruhlmann P, Maus S, Michel BA. Comparison of the validity and reliability of self-reported articular indices. *Br J Rheumatol* 1995; 34: 760-6.
- 26 Wong AL, Wong WK, Harker J, Sterz M, Bulpitt K, Park G, et al. Patient self-report tender and swollen joint counts in early rheumatoid arthritis. *J Rheumatol* 1999; 26: 2551-61.
- 27 Prevoo ML, Kuper IH, van't Hof MA, van Leeuwen MA, van de Putte LB, van Riel PL. Validity and reproducibility of self-administered joint counts. A prospective longitudinal follow up study in patients with rheumatoid arthritis. *J Rheumatol* 1996; 23: 841-5.
- 28 Hanly JG, Mosher D, Sutton E, Weerasinghe S, Theriault D. Self-assessment of disease activity by patients with rheumatoid arthritis. *J Rheumatol* 1996; 23: 1531-8.
- 29 Alarcon GS, Tilley BC, Li SH, Fowler SE, Pillemer SR. Self-administered joint counts and standard joint counts in the assessment of rheumatoid arthritis. *J Rheumatol* 1999; 26: 1065-7.
- 30 Calvo FA, Calvo A, Berrocal A, Pevez C, Romero F, Vega E, et al. Self-administered joint counts in rheumatoid arthritis: Comparison with standard joint counts. *J Rheumatol* 1999; 26: 536-9.
- 31 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307-10.
- 32 Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van de Tempel H, van der Linden SJ. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. *J Rheumatol* 1999; 26: 980-4.
- 33 Hunt SM, McKenna SP. The QLDs: A scale for the measurement of quality of life in depression. *Health Policy* 1992; 22: 307-19.
- 34 Rasch G. Probabilistic Models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1980.
- 35 Haywood KL, Garratt AM, Jordan K, Dziedzic K, Dawes PT. Disease-specific, patient-assessed measures of health outcome in ankylosing spondylitis: reliability, validity and responsiveness. *Rheumatol* 2002; 41: 1295-1302.
- 36 van Tubergen A, Landewé R, van der Heijde D, Hidding A, Wolter N, Asscher M, Falkenbach A, Genth E, Goei Thè H, van der Linden SJ. Combined spa-exercise therapy is effective in patients with ankylosing spondylitis: a randomized controlled trial. *Arthritis Care Res* 2001; 45: 430-8.

- 37 van Tubergen A, Landewé R, Heuft-Dorenbosch L, Spoorenberg A, van der Heijde D, van der Tempel H, van der Linden S. Assessment of disability with the WHODAS II in patients with ankylosing spondylitis. *Ann Rheum Dis* 2003; 62: 140-5.
- 38 Marzo-Ortega H, McGonagle D, Emery P. Etanercept treatment in resistant spondyloarthritis: Imaging, duration of effect and efficacy on reintroduction. *Clin Exp Rheumatol* 2002; Suppl 28: S175-7.
- 39 MacKay K, Brophy S, Mack C, Doran M, Calin A. The development and validation of a radiographic grading system for the hip in Ankylosing Spondylitis: the Bath Ankylosing Spondylitis Radiology Hip Index. *J Rheumatol* 2000; 27: 2866-72.
- 40 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.
- 41 Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. *J Rheumatol* 1999; 26: 731-9.
- 42 Braun J, Sieper J, Bollow M. Imaging of sacroiliitis. *Clin Rheumatol* 2000; 19: 51-7.
- 43 Braun J, Baraliakos X, Golder W, Brandt J, Rudwaleit M, Listing J, Bollow M, Sieper J, van der Heijde D. Magnetic resonance imaging examinations of the spine in patients with ankylosing spondylitis, before and after successful therapy with infliximab: evaluation of a new scoring system. *Arthritis Rheum* 2003; 48: 1126-36

Chapter 9

SAMENVATTING EN DISCUSSIE

Samenvatting en discussie

Een internationale werkgroep (ASsessment in Ankylosing Spondylitis, ASAS) heeft sets met kernpunten gedefinieerd voor wetenschappelijk onderzoek bij de ziekte van Bechterew (ofwel spondylitis ankylopoëtica) met als doel meer uniformiteit te scheppen in wetenschappelijk onderzoek naar ziekte-activiteit en ziekte-uitkomst bij deze aandoening. Deze sets met kernpunten werden gedefinieerd voor drie onderzoeksettings: therapie die het beloop van de ziekte beïnvloedt (disease controlling anti-rheumatic therapy, DC-ART), medicatie en therapie die de symptomen van de ziekte beïnvloeden (symptom modifying anti-rheumatic drugs, SM-ARD/physical therapy) en reguliere behandeling (clinical record keeping). De volgende domeinen werden voor alle drie onderzoeksettings geselecteerd: lichamelijk functioneren, pijn, mobiliteit en stijfheid van de wervelkolom, globale indruk van de patiënt tav ziekte en moeheid. De domeinen voor 'clinical record keeping' en 'DC-ART' zijn uitgebreid met acute fase reacties, perifere gewrichten en peesaanhechtingen (enthesis), 'DC-ART' bevat ook de domeinen: röntgen onderzoek van wervelkolom en heupen.

In navolging van het werk van de ASAS werkgroep heeft dit proefschrift vooral betrekking op verschillende aspecten van ziekte-activiteit en ziekte-uitkomst bij de ziekte van Bechterew. Het eerste deel van het proefschrift (hoofdstuk 2-5) gaat vooral over ziekte-activiteit terwijl de hoofdstukken 6 en 7 zich met name richten op ziekte-uitkomst. Bijna alle resultaten beschreven in dit proefschrift zijn afkomstig van een internationaal observationeel 'multicenter' onderzoek: 'Outcome in Ankylosing Spondylitis International Study' (OASIS). In deze studie werden 217 opeenvolgende poliklinische patiënten met de ziekte van Bechterew geïncludeerd. Dit cross-sectioneel cohort van Bechterew patiënten voldeed aan de gemodificeerde 'New York' criteria en werd longitudinaal gevolgd in verschillende academische- en perifere ziekenhuizen in Europa. Honderdzevenendertig patiënten zijn afkomstig uit het academisch ziekenhuis Maastricht en het Maasland ziekenhuis in Sittard (Nederland), 55 patiënten uit het Hôpital Cochin in Parijs (Frankrijk) en 25 patiënten uit het academisch ziekenhuis Gent (Belgie). Al deze ziekenhuizen zijn secundaire en/of tertiaire referentie centra. Overeenkomstig met andere Bechterew populaties is ongeveer twee derde van de OASIS patiënten van het mannelijk geslacht. Aan het begin van de OASIS studie was de gemiddelde leeftijd van de patiënten 43 jaar (SD 13 jaar) en hadden de patiënten een gemiddelde ziekte duur van 11 jaar (SD 8 jaar). Bij 27% van de patiënten werd een perifere artritis vastgesteld door de behandelend reumatoloog. In elk van de drie participerende landen werden de OASIS patiënten ieder half jaar gedurende 2 jaar onderzocht door steeds dezelfde getrainde persoon (2 reumatologen en 1 onderzoeksverpleegkundige) volgens een vastgesteld protocol. Onafhankelijk van de bevindingen van deze onderzoekers werden de patiënten ook regulier gezien door de behandelend reumatoloog.

De vergelijking van twee indexen voor fysiek functioneren bij de ziekte van Bechterew

Fysiek functioneren is gerelateerd aan ziekte-activiteit en uiteindelijke schade aangericht door de ziekte van Bechterew. De ASAS werkgroep selecteerde fysiek functioneren gemeten met de 'Dougados Functionele Index' (DFI) en de 'Bath Ankyloserende Spondylitis Functionele Index' (BASFI) voor de domeinen van alle drie de onderzoeksettings. In **hoofdstuk 2** werden deze twee veel gebruikte en gevalideerde ziekte specifieke functionele indexen met elkaar vergeleken. Aangezien ziekte-activiteit en schade beide gerelateerd zijn aan fysiek functioneren was het belangrijkste doel van deze cross-sectionele studie om de relatie van BASFI en DFI met deze twee aspecten van de ziekte te onderzoeken. Wanneer de resultaten van deze studie zouden laten zien dat één van deze twee indexen beter zou presteren, dan zou die index uniform gebruikt kunnen worden voor het meten van fysiek functioneren.

De BASFI bestaat uit 10 vragen op een visuele analoge schaal (VAS) en alle vragen betreffen activiteiten uit het dagelijks leven. Het gemiddelde van de scores van de afzonderlijke 10 vragen vormt de uiteindelijke score. De DFI bestaat uit twintig 5 punts Likert respons vragen over het in staat zijn om verschillende dagelijkse activiteiten uit te voeren. De totale score (range van 0-40) wordt berekend door de som van de afzonderlijke vragen te berekenen. Aangezien er geen 'gouden standaard' bestaat voor het meten van ziekte-activiteit bij de ziekte van Bechterew hebben we voor ziekte-activiteit drie meetinstrumenten gekozen: ziekte-activiteit aangegeven op een visuele analoge schaal (VAS, 0-10 cm) door de arts en de patiënt en de Bath Ankylosing Spondylitis Disease Activity Index (BASDAI, range 0-10). Als meetinstrumenten voor schade werden een globale en een gedetailleerde röntgen scoringsmethode specifiek voor het scoren van de wervelkolom bij de ziekte van Bechterew gekozen: de Bath Ankylosing Spondylitis Radiology Index-spine (BASRI-s, range 0-12) en de gemodificeerde Stoke Ankylosing Spondylitis Spine Score (SASSS, range 0-72). Voor de analyses hebben we de twee meest contrasterende groepen patiënten gebruikt: Bechterew patiënten waarbij ziekte activiteit hoog (score ≥ 6.0) en laag (score ≤ 4.0) was volgens de drie ziekte-activiteit instrumenten. Verder werden er Receiver Operator Curves (ROC) gemaakt voor beide functionele indexen versus de drie ziekte-activiteit instrumenten.

In onze patiënten populatie werden relatief lage waarden voor de functionele indexen en de ziekte-activiteit instrumenten gevonden. Er waren 7 BASFI vragenlijsten versus 28 DFI vragenlijsten met één of meer niet ingevulde vragen. Zoals verwacht waren beide functionele indexen hoog met elkaar gecorreleerd (Spearman 0.89). De correlaties van de BASFI en DFI met de ziekte-activiteit instrumenten was vergelijkbaar voor beide indexen waarbij correlatie met de BASDAI het hoogst was (respectievelijk 0.59 en 0.57). ROC curven voor elk van de twee functionele indexen met de drie ziekte-activiteit instrumenten liet de beste curve met de hoogste sensitiviteit en specificiteit zien voor beide functionele indexen versus de BASDAI (BASFI: respectievelijk 94% en 87% en DFI: respectievelijk 93%

en 79%) De afkappunten voor het onderscheiden van lage en hoge ziekte-activiteit waren duidelijk hoger voor de BASFI ($\pm 40\%$ van de volledige schaal) in vergelijking met de DFI ($\pm 20\%$ van de volledige schaal). Wanneer we echter deze afkappunten gebruiken om onderscheid te maken tussen hoge en lage ziekte-activiteit gebaseerd op de functionele scores worden er behoorlijk hoge percentages patiënten fout geclassificeerd (12-30%). Het percentage fout geclassificeerde patiënten was het laagste voor de BASFI afkappunten in combinatie met ziekte-activiteit gemeten met de BASDAI (12%).

Ziekte-activiteit gemeten met de BASDAI komt dus het dichtst bij het ziekte-activiteit aspect van de BASFI. Een reden hiervoor zou kunnen zijn dat beide indexen geheel door de patiënt gerapporteerd worden en door dezelfde onderzoeksgroep ontwikkeld zijn. Aangezien fysiek functioneren bij de ziekte van Bechterew niet alleen bepaald wordt door ziekte-activiteit kan ook niet verwacht worden dat alle patiënten goed geclassificeerd werden. Ten aanzien van de instrumenten gekozen voor het bepalen van schade laat de BASFI-s een relatief hogere mediaan score voor radiologische schade zien dan de gemodificeerde SASSS: respectievelijk 7.0 (range 2-12) en 12.0 (range 0-72). Correlaties van beide functionele indexen met BASFI-s en gemodificeerde SASSS waren vergelijkbaar (respectievelijk 0.42 en 0.36).

In deze cross-sectionele studie lijkt de BASFI iets beter te presteren dan de DFI met betrekking tot aspecten van ziekte-activiteit en uitvoerbaarheid (BASFI kost minder tijd om in te vullen en er waren minder ontbrekende antwoorden). Op basis van de resultaten van dit onderzoek is er anderszins geen voorkeur ten aanzien van BASFI en DFI uit te spreken. De BASFI werd enkele jaren na de DFI ontwikkeld en vermijdt een aantal vermoedelijk overvullende vragen en vragen met betrekking op symptomen in plaats van fysiek functioneren. Verder bevat de BASFI drie vragen die de validiteit van de inhoud verbeteren. Voor twee van deze vragen is dit bewezen bij de ontwikkeling van de HAQ-s. Bij patiënten met reumatoïde artritis en artrose werd aangetoond dat alleen de derde van deze vragen, betreffende fysiek belastende activiteiten, kon discrimineren tussen bijna maar niet geheel gezonde patiënten die niet scoorden op één van de andere vragen.

Aan de andere kant bevat de BASFI vragen over ongewone taken en de DFI bevat diverse vragen met betrekking op verschillende aanvullende domeinen die niet in de BASFI geïncludeerd zijn. Er is een literatuur overzicht gepubliceerd van alle bestaande studies met betrekking tot de vergelijking van de BASFI en DFI. De enige studie met een rechtstreekse vergelijking van de twee functionele indexen liet zien dat beide indexen onderscheid konden maken tussen poliklinische en klinische patiënten maar alleen de BASFI kon de effecten van een intensieve fysiotherapie gedurende drie weken aantonen. In zes studies met betrekking tot fysiotherapie was de DFI vier maal niet in staat onderscheid te maken tussen de verschillende behandelingen terwijl in de twee andere studies de BASFI wel kon discrimineren tussen de verschillende behandelingen. Een reden hiervoor kan zijn dat in de studies waar de DFI gebruikt werd de scores van de DFI vragen dicht bij de normale waarden lagen. Het gevolg hiervan kan zijn dat het niet mogelijk is verder te verbeteren wanneer er sprake is van alleen milde invaliditeit. Om de sensitiviteit van de DFI te verbeteren stelden de auteurs voor de vragen om te zetten van een drie punt Likert

respons schaal naar een vijf punt Likert respons schaal.

De DFI discrimineerde goed tussen de verschillende behandelingen in op één na alle SMARD trials en in één DC-ART studie waarbij deze index werd gebruikt. De BASFI discrimineerde tussen de verschillende behandelingen in één van de twee SMARD trials waarbij de index werd gebruikt. De enige SMARD trial waarbij een rechtstreekse vergelijking van de twee indexen is gedaan liet zien dat beide indexen even goed discrimineerden tussen placebo en behandelgroep. Er is geen DC-ART studie beschikbaar die de BASFI en de DFI rechtstreeks vergelijkt. In de enige studie die een conventionele DC-ART (sulfasalazine) evalueerde werd de DFI gebruikt en in drie andere studies waar de effecten van anti tumor necrosis factor (TNF) a werden geëvalueerd werd de BASFI gebruikt. In al deze studies discrimineerde de gebruikte index goed tussen de verschillende behandelarmen.

Samenvattend lijkt er een lichte voorkeur te bestaan om de DFI te gebruiken in SMARD trials en de BASFI in trials met betrekking tot fysiotherapie. Gezien de goede effecten van biologische geneesmiddelen zoals anti-TNF a bij de behandeling van de ziekte van Bechterew is een rechtstreekse vergelijking van beide functionele indexen in toekomstige DC-ART trials goed mogelijk; daarbij zullen dan ook 'effect sizes' en/of 'standardised reponse mean' van beide indexen berekend moeten worden.

Bezinking (BSE) versus C-reactive protein (CRP) bij de ziekte van Bechterew

De acute fase bloed testen, BSE (mm/uur) en CRP (mg/l), worden beide regelmatig gebruikt ter evaluatie van de ziekte activiteit bij Bechterew patiënten. Het bepalen van acute fase reacties is door de ASAS werkgroep voorgesteld als een domein voor de onderzoeksettings DC-ART trial en reguliere behandeling. Het is bekend dat de gemiddelde waarde van deze twee acute fase reacties laag zijn bij patiënten met de ziekte van Bechterew in vergelijking met patiënten met reumatoïde artritis.

Hoofdstuk 3 van dit proefschrift beschrijft een onderzoek met als doel te bepalen of er onderscheidt gemaakt kan worden tussen BSE en CRP voor het meten van ziekte-activiteit bij de ziekte van Bechterew. Gezien er verschillen kunnen bestaan in het klinisch beeld bij Bechterew patiënten ten aanzien van BSE en CRP is onze patiënten populatie in twee groepen verdeeld: patiënten met alleen spinale betrokkenheid (n=149) en patiënten met ook actieve perifere artritis en/of inflammatoire darmziekte (n=42). Zoals eerder werd aangegeven bestaat er geen 'gouden standaard' voor het meten van ziekte-activiteit bij de ziekte van Bechterew. We bestudeerde daarom de relatie van BSE en CRP met drie vervangende klinische ziekte-activiteit instrumenten volgens dezelfde methodologie en statische procedure als beschreven in de vorige studie met betrekking tot de vergelijking van BASFI en DFI. Een tweede doel van deze studie was na te gaan of hoge waarden voor BSE en CRP de mate van ziekte-activiteit reflecteren gedefinieerd door de drie geselecteerde ziekte-activiteit instrumenten.

De resultaten lieten zien dat in de groep met alleen spinale betrokkenheid de meeste patiënten normale waarden voor BSE en CRP hadden. Daarentegen lieten de Bechterew patiënten met ook perifere artritis en/of inflammatoire darmziekte licht verhoogde waarden voor BSE en CRP zien. Allen voor de BSE was dit verschil tussen de twee subgroepen statistisch significant. Dertig procent van de patiënten in beide subgroepen hadden een verhoogde BSE en een normale CRP of vice versa en in de meeste van deze gevallen werden waarden van net boven de normale grens gevonden als de acute fase reactie verhoogd was. Over het algemeen werden er in beide subgroepen relatief lage scores voor de ziekte-activiteit instrumenten gevonden. Opvallend is het verschil in beoordeling van ziekte-activiteit door de arts (VAS) aan de ene kant en de twee op de patiënt gebaseerde instrumenten aan de andere kant (BASDAI en ziekte-activiteit aangegeven op een VAS door de patiënt). Aan het begin van de studie werd dit weergegeven in zeer verschillende gemiddelde waarden van ziekte-activiteit aangegeven door de arts versus BASDAI en ziekte-activiteit aangegeven door de patiënt (voor de groep met spinale betrokkenheid: respectievelijk 1.5 versus 3.6 en 3.9 en voor de groep met perifere artritis en/of inflammatoire darmziekte: respectievelijk 2.5 versus 4.3 en 4.1). In de groep met alleen spinale betrokkenheid had maar 3% van de patiënten hoge ziekte activiteit beoordeeld door de arts in tegenstelling tot ziekte activiteit beoordeeld door de patiënt en BASDAI (respectievelijk 21% en 11%). Over het algemeen lieten de ROC curven lage afkappunten zien in beide subgroepen voor zowel BSE als CRP. Gebaseerd op deze afkappunten waren sensitiviteit en specificiteit redelijk met de hoogste sensitiviteit (100%) voor ziekte-activiteit aangegeven op een VAS door de arts. Helaas waren de bijbehorende positief voorspellende waarden, die belangrijk zijn in de klinische praktijk, laag met hoge percentages fout geclassificeerde patiënten. Deze cross-sectionele studie laat zien dat zowel BSE als CRP niet overeenkomen met ziekte-activiteit zoals gedefinieerd in deze studie. Op basis van deze resultaten kan er dus geen duidelijke voorkeur worden gegeven aan één van deze twee acute fase reacties. In relatie met progressie van schade bij de ziekte van Bechterew zou er wel een duidelijk verschil kunnen bestaan tussen BSE en CRP en dit zal in de toekomst ook verder geëvalueerd moeten worden.

Overige studies die een rechtstreekse vergelijking van BSE en CRP laten zien zijn moeilijk te interpreteren doordat er steeds verschillende definities voor ziekte-activiteit worden gebruikt. Er is een literatuuroverzicht gepubliceerd van alle bestaande studies met betrekking tot de vergelijking van BSE en CRP bij de ziekte van Bechterew. In vijf van deze zeven studies werd er geen verschil gevonden tussen BSE en CRP en de resultaten van de twee andere studies geven aan dat CRP beter gerelateerd is aan ziekte activiteit dan BSE. Twee studies geven aan dat verhoogde BSE en CRP vaker gezien worden in Bechterew patiënten met perifere manifestaties van de ziekte. BSE en CRP werden niet rechtstreeks vergeleken in een SMARD trial. In de beschikbare SMARD trials werd geen onderscheidend vermogen gevonden van BSE en CRP tussen de verschillende behandelingen. Eén van deze SMARD trials rapporteerde een lage 'standardised response mean' voor CRP. Negen DC-ART trials (één methotrexaat trial en acht sulfasalazine trials) lieten geen significant effect

van de gebruikte medicatie zien en waarschijnlijk is hierdoor ook geen discriminerend vermogen van BSE en CRP zichtbaar tussen de geteste medicatie en placebo. De twee DC-ART trials met een rechtstreekse vergelijking van de twee acute fase reacties lieten tegengestelde resultaten zien. In twee studies naar de effecten van anti-TNF α konden zowel BSE als CRP erg goed onderscheid maken tussen de verschillende behandelingen. Alhoewel de 'effect sizes' en 'standardised response mean' voor BSE en CRP hoog waren in deze studies (>3) werden de exacte waarden hiervan helaas niet gerapporteerd. Op basis van deze resultaten kan voor geen van de drie onderzoeksettings een definitieve keus gemaakt worden tussen BSE en CRP. Gezien de veelbelovende resultaten van biologische geneesmiddelen zoals anti-TNF α bij de behandeling van de ziekte van Bechterew is er een rechtstreekse vergelijking van BSE en CRP mogelijk waarbij door berekening van 'effect sizes' en/of 'standardised response mean' van beide acute fase reacties een betere vergelijking mogelijk is. Tevens zal in de toekomst de relatie van BSE en CRP met de progressie van schade bij de ziekte van Bechterew onderzocht moeten worden.

Patiënt gerapporteerde gewricht scores bij de ziekte van Bechterew

In **hoofdstuk 4** wordt een studie gepresenteerd naar de betrouwbaarheid van het rapporteren van pijnlijke en gezwollen gewrichten door de patiënt zelf. Een klein gedeelte van de Bechterew patiënten heeft een perifere artritis ($\pm 25\%$). Normaal gesproken wordt artritis klinisch gediagnostiseerd door een arts of een goed getrainde verpleegkundige. Wanneer de gewrichtsscores aangetoond bij lichamelijk onderzoek door de arts vervangen kunnen worden door gewrichtsscores door de patiënt zelf zou dit een groot voordeel betekenen voor reumatologen en in het bijzonder klinisch onderzoekers. De betrouwbaarheid van door de patiënten zelf gerapporteerde gewrichtsscores bij de ziekte van Bechterew is nooit eerder onderzocht. Bij patiënten met reumatoïde artritis is de betrouwbaarheid van patiënt gerapporteerde gewrichtsscores uitgebreid onderzocht en de resultaten lieten zowel hoge als lage betrouwbaarheid zien. In onze studie werd aan 217 Bechterew patiënten gevraagd hun pijnlijke en gezwollen gewrichten aan te kruisen op een mannequin, ontwikkeld volgen de methode van Stewart waarop respectievelijk 44 en 40 gewrichten aangekruist kunnen worden. Op dezelfde dag maar zonder dat de resultaten van de patiënten bekend waren rapporteerden drie onderzoekers (één persoon voor elk onderzoekscentrum) de pijnlijke en gezwollen gewrichten van deze patiënten op vergelijkbare mannequins. De resultaten lieten zien dat er op groepsniveau een consistent verschil was tussen het aantal gerapporteerde pijnlijke en gezwollen gewrichten door patiënten en onderzoekers met een bijbehorende matige overeenstemming (Intraclass Correlatie Coëfficiënt tussen 0.51 en 0.71) van het totaal aantal pijnlijke en gezwollen gewrichten en zelfs een slechte tot matige overeenstemming (kappa tussen 0.23 en 0.64) van de individuele gewrichtsscores. Patiënten scoorden consistent meer pijnlijke gewrichten en de onderzoekers scoorden meer gezwollen gewrichten. Mogelijke verklaringen hiervoor

zijn : (1) Bechterew patiënten kunnen niet goed differentiëren tussen een pijnlijk gewricht en pijn veroorzaakt door enthesitis gezien de aanhechtingen vlak bij het gewricht gelokaliseerd zijn en (2) Bechterew patiënten zijn niet getraind om gezwollen gewrichten te detecteren. De enthesis index volgens Mander was alleen significant gecorreleerd met het totaal aantal gezwollen gewrichten gerapporteerd door de onderzoekers. Op patiëntniveau waren de resultaten zelfs nog slechter en dit is duidelijk zichtbaar gemaakt in Bland en Altman plots. Door patiënt gerapporteerde gewrichtsscores kunnen nog steeds van waarde zijn als patiënten kunnen differentiëren tussen de aanwezigheid van mono- oligo- en polyartritis. Ook voor deze onderverdeling was volledige overeenstemming tussen onderzoekers en patiënten erg laag (respectievelijk 17%, 17% en 22%). Alleen wanneer er geen sprake was van gezwollen gewrichten was de overeenstemming tussen onderzoekers en patiënten goed (82%). Bechterew patiënten kunnen dus wel oordelen over de afwezigheid van gezwollen gewrichten maar ze zijn niet in staat om om de aanwezigheid van zwelling aan te geven zelfs niet in de grove categorieën van mono- oligo- en polyartritis. In deze studie hebben we officieel geen test-retest betrouwbaarheid onderzocht maar de resultaten verkregen met de baseline en 1 jaars data waren vergelijkbaar. Op basis van deze resultaten kunnen gewrichtsscores gerapporteerd door onderzoekers/artsen dus niet vervangen worden door gewrichtsscores gerapporteerd door de Bechterew patiënten zelf. Alleen informatie van de patiënten ten aanzien van het afwezig zijn van gezwollen gewrichten is voldoende betrouwbaar om bruikbaar te kunnen zijn.

Ziekte-activiteit bij de ziekte van Bechterew

Aangezien er een grote variatie bestaat in het klinische beeld tussen verschillende Bechterew patiënten is het erg moeilijk om ziekte-activiteit te definiëren bij deze ziekte. Patiënten hebben axiale betrokkenheid in verschillende gradaties maar kunnen daarbij ook verschillende extra-spinale manifestaties van de ziekte hebben. Deze klinische diversiteit in zowel ernst als in lokalisatie zorgt ervoor dat instrumenten die gebruikt worden om ziekte-activiteit te meten aan hoge eisen moeten voldoen. Verder hebben Bechterew patiënten en hun behandelend reumatologen zeer uiteenlopende inzichten ten aanzien van ziekte-activiteit. **Hoofdstuk 5** beschrijft op basis van welke criteria Bechterew patiënten en reumatologen ziekte-activiteit beoordelen. Ons doel was om verschillen in inzicht ten aanzien van ziekte-activiteit tussen patiënten en reumatologen te verkennen. Voor deze studie werden data van het OASIS cohort gebruikt. Aangenomen mag worden dat de patiënten uit dit cohort het hele klinische spectrum van Bechterew patiënten omvat dat normaal gesproken door reumatologen gezien worden. De patiënten werden geïncludeerd onafhankelijk van geslacht, leeftijd, duur van de ziekte, ernst van de ziekte en mate van ziekte-activiteit.

In deze studie werd ziekte-activiteit bestudeerd vanuit het perspectief van zowel de patiënt als van de arts. Ziekte-activiteit door de arts en de patiënt aangegeven op een visuele

analoge schaal (VAS range: 0 = niet actief en 10 = zeer actief) werd onderverdeeld in 'hoge' ziekte-activiteit (VAS ≥ 6.0) en 'lage' ziekte-activiteit (VAS ≤ 4.0). Meetinstrumenten die door de ASAS werkgroep geselecteerd zijn voor gebruik bij de evaluatie van de ziekte van Bechterew werden iedere zes maanden toegepast over een periode van twee jaar. Datareductie van deze instrumenten werd verricht met behulp van factoranalyse. Dit resulteerde in vier factoren met onderling correlerende meetinstrumenten: 'metingen van de mobiliteit van de wervelkolom', 'metingen door de arts', 'metingen door de patiënt' en 'laboratoriumbepalingen' (Cronbachs alpha tussen 0.52 en 0.81; verklaarde variantie 61%). Er werd aangenomen dat de instrumenten binnen een factor hetzelfde onderliggende construct bepalen. Een discriminantanalyse met de factorwaarden werd verricht om onderscheid te kunnen maken tussen 'lage' en 'hoge' ziekte-activiteit voor zowel het perspectief van de patiënt als dat van de arts. Deze analyse liet zien dat de factor, 'metingen door de patiënt', het meest bijdragend was (gezamenlijke (pooled) correlatie 0.84) in het onderscheid maken tussen de twee niveaus van ziekte-activiteit gedefinieerd door de patiënt. De bijdrage van de andere drie factoren was maar minimaal (pooled correlatie < 0.30). Daarentegen droegen de drie factoren: 'metingen door de arts', 'metingen van de mobiliteit van de wervelkolom' en laboratoriumbepalingen' het meest bij in het discrimineren tussen de twee niveaus van ziekte-activiteit gedefinieerd door de arts (pooled correlatie: respectievelijk 0.62, 0.48 en 0.48). De factor, 'metingen door de patiënt', droeg in het geheel niet bij (pooled correlatie 0.05). De discriminantanalyse verricht met baseline data en de data van de andere momenten liet geen verschillen zien. De discriminantscores werden gebruikt voor multiële regressie analyse om de prioriteit van de instrumenten te bepalen ten aanzien van de bijdrage aan ieder perspectief van ziekte-activiteit. Het ziekte-activiteit perspectief van de patiënt werd het beste verklaard door de instrumenten: 'pijn van de wervelkolom', 'BASFI', 'gewrichtspijn', en 'moeheid'. Het perspectief van de arts kwam het beste tot uiting met de instrumenten: 'cervicale rotatie', 'aantal gezwollen gewrichten', CRP' en 'intermalleolaire afstand'. Onze resultaten bevestigen dat patiënten en artsen een heel ander zicht hebben op wat ziekte-activiteit bij de ziekte van Bechterew betekent. Bechterew patiënten beoordelen ziekte-activiteit op basis van klachten terwijl artsen ziekte-activiteit beoordelen op basis van instrumenten die ontsteking en ernst van de ziekte meten.

Uit deze studie volgen nog een aantal andere conclusies. De BASFI, een index primair ontworpen om fysiek functioneren bij Bechterew patiënten te meten, sluit ook aan bij het perspectief van ziekte-activiteit van de patiënt. Het lijkt er tevens op dat Bechterew patiënten hun inschatting van ziekte-activiteit gedeeltelijk maken op basis van wat ze lichamelijk kunnen doen. Over het algemeen kunnen Bechterew patiënten goed onderscheid maken tussen ziekte-activiteit (bepaald door klachten) en de ernst van de ziekte. Ziekte-activiteit vanuit het patiënten perspectief wordt niet bepaald door acute fase reacties. Het oordeel over ziekte-activiteit van de arts wordt gebaseerd op een combinatie van constructen met instrumenten die informatie over ziekte-activiteit en ernst van de ziekte combineren. Opvallend is dat CRP als variabele werd geïncludeerd aangezien de waarde hiervan niet bij de arts bekend was op het moment dat er een oordeel werd

gegeven over ziekte-activiteit. Momenteel worden volledig op het oordeel van de patiënt gebaseerde ziekte-activiteit instrumenten, zoals BASDAI, in combinatie met ziekte activiteit aangegeven door de arts en/of verhoogde CRP gebruikt als 'gouden standaard' voor het meten van ziekte-activiteit bij de ziekte van Bechterew voor het includeren van patiënten in klinische trials en voor de besluitvorming rondom het starten van anti TNF α therapie in de klinische praktijk. Er is echter nog steeds geen wetenschappelijk bewijs dat ziekte-activiteit instrumenten gebaseerd op het oordeel van de patiënt of de arts geassocieerd zijn met ziekte-uitkomst op de lange duur. Deze belangrijke informatie is nodig om instrumenten te kunnen selecteren die zowel de ziekte-activiteit als de uitkomst van de ziekte het beste weergeven. Aangezien de keuzes voor medicamenteuze behandeling van de ziekte van Bechterew groter worden is het ook belangrijker om instrumenten te definiëren die een eenduidig construct van ziekte-activiteit en ziekte-uitkomst meten zodat deze gebruikt kunnen worden in klinische trials.

Kwaliteit van leven bij de ziekte van Bechterew

Er is een groeiende belangstelling voor het meten van kwaliteit van leven (QoL) bij chronische invaliderende aandoeningen. Met name in studies ontwikkeld om de impact van nieuwe farmaceutische producten te meten of om verschillende behandelmethoden met elkaar te vergelijken wordt steeds vaker kwaliteit van leven gemeten. Ziektespecifieke instrumenten gebruikt om het beloop van de ziekte van Bechterew te evalueren richten zich vooral op fysieke beperkingen en/of fysiek functioneren. Er zijn wel generieke instrumenten die de algehele gezondheid meten maar er bestaat nog geen ziektespecifiek instrument voor het meten van kwaliteit van leven bij de ziekte van Bechterew.

Hoofdstuk 6 beschrijft de ontwikkeling van de Ankylosing Spondylitis Quality of Life vragenlijst (ASQoL). Ons doel was om een valide en betrouwbaar Bechterew specifiek kwaliteit van leven instrument te ontwikkelen dat relevant en acceptabel is voor respondenten. De ASQoL is gelijktijdig ontwikkeld in Engeland en Nederland. Alle geïncludeerde Bechterew patiënten voldeden aan de gemodificeerde New York criteria. De methodologie gebruikt om de ASQoL te ontwikkelen combineert de theoretische principes van het 'needs-based' kwaliteit van leven model met de statistische en diagnostische kracht van het Rasch-model.

De ontwikkeling van de vragenlijst bestaat uit vijf stadia. De inhoud van de eerste versie van de vragenlijst was afkomstig van interviews met 30 patiënten in Engeland en 25 patiënten in Nederland (stadium 1). Stadium 2 betreft het selecteren van items en antwoord type waarbij de eerste proef vragenlijst met 41 items werd gevormd. Er werden test interviews gedaan met deze proef vragenlijst bij 15 patiënten in Engeland en 15 patiënten in Nederland om informatie over validiteit van vorm en inhoud te verkrijgen. Hierbij ontstond een vragenlijst met 36 items (stadium 3). Stadium 4 was een onderzoek per post in Engeland (n=121) met als doel een efficiëntere versie van de ASQoL te ontwikkelen. Met behulp van Raschanalyse werd het aantal vragen verminderd tot 26 items. Raschanalyse van de data van

een laatste onderzoek per post werd in beide landen (Engeland: $n=164$; Nederland: $n=154$) gedaan om eigenschappen van de schaal, betrouwbaarheid, interne consistentie en construct validiteit te meten (stadium 5). Deze analyse liet enkele misplaatste items zien maar liet tevens zien dat de items een hiërarchische volgorde hadden en stabiel waren over de tijd. De problematische items werden verwijderd hetgeen resulteerde in de definitieve ASQoL met 18 items. Beide taal versies van de ASQoL hadden een uitstekende interne consistentie (Cronbach's alfa: 0.89-0.91), test-retest betrouwbaarheid (Intraclass Correlatie Coëfficiënt: Engeland: 0.92; Nederland: 0.91) en validiteit. De ASQoL kan een waardevol instrument zijn in zowel de klinische situatie als bij wetenschappelijk onderzoek wanneer de invloed van de ziekte van Bechterew en bijbehorende behandeling op kwaliteit van leven gemeten wordt. Onafhankelijk van deze studie werd een goede betrouwbaarheid, validiteit en respons van de ASQoL gevonden in een studie waarbij door de patiënt bepaalde ziekte specifieke instrumenten werden vergeleken. De ASQoL werd sinds de ontwikkeling gebruikt in twee trials. Eén studie evalueert het effect van kuurtherapie en de andere het effect van anti-TNF a therapie. Beide studies laten zien dat de ASQoL in staat is te discrimineren tussen de verschillende behandelingen. De 'standardised reponse mean' (SRM) en 'effect size' (ES) werden alleen berekend voor de kuurtherapie trial met matige respons scores (SRM: 0.24 en EF: 0.22) passend bij het matige behandeloeffect. Er was een onderling vergelijkbaar hoge respons van de ASQoL en de BASDAI in de studie naar het effect van etanercept.

Radiologische scoringsmethoden bij de ziekte van Bechterew

Radiologische schade is een belangrijke uitkomst maat bij de ziekte van Bechterew. Het is erg moeilijk om deze radiologische schade te evalueren. Veranderingen van de sacroiliacale gewrichten (SI) worden meestal gescoord door gebruik te maken van de New York criteria of de bijna identieke scoringsmethode ontwikkeld door de Stoke groep. Beide SI scoringsmethoden differentiëren 5 graden (range 0-4 voor ieder SI gewricht). Er zijn twee scoringsmethoden om de lumbale en cervicale wervelkolom te scoren. De 'Bath Ankylosing Spondylitis Radiology Index' (BASRI) is een globale, snelle en makkelijke scoringsmethode ontwikkeld voor het scoren van de voor-achterwaartse en de laterale opname van de lumbale wervelkolom (range beide opnames gecombineerd 0-4), de laterale opname van de cervicale wervelkolom (range 0-4) en de heupen (BASRI-h, range 0-4 voor iedere heup; daarna gemiddelde rechter en linker heup). De gemiddelde score van de New York scoringsmethode voor de SI gewrichten gecombineerd met de verschillende BASRI scoringsmethoden vormen twee samengestelde scoringsmethoden: BASRI-'spine' (range 2-12) en BASRI-'total' (range 2-16). De SASSS methode voor de wervelkolom is een meer gedetailleerde scoringsmethode die verschillende aspecten zoals 'squaring', sclerose en erosies op verschillende plaatsen van iedere wervel scoort. Deze methode wordt gescoord op de laterale opname van de lumbale wervelkolom aan de voor- en achterzijde van iedere wervel (SASSS, range 0-72). De 'gemodificeerde' SASSS wordt gescoord op de laterale

opname van zowel de lumbale als cervicale wervelkolom waarbij alleen de voorzijde van iedere wervel gescoord wordt (range 0-72). In **hoofdstuk 7**, werd de betrouwbaarheid en radiologische verandering na 1 en 2 jaar follow-up van al deze beschikbare radiologische scoringsmethoden vergeleken. Twee goed getrainde 'observers' scoorden sets met röntgenfoto's van het OASIS cohort gemaakt op baseline, 1 en 2 jaar follow-up. De röntgenfoto's van één set werden gelijktijdig (gepaard) gescoord zonder kennis van de chronologische volgorde. De volgorde waarin alle sets werd gescoord was willekeurig. Afhankelijk van het aantal missende scores voor iedere methode varieerde het aantal beschikbare sets voor betrouwbaarheidsanalyse tussen de 100 en 136.

Onze resultaten lieten goede 'intra- en interobserver' betrouwbaarheid zien voor bijna alle scoringsmethoden. 'Observer' betrouwbaarheid voor categoriale data werd geanalyseerd met behulp van lineair gewogen kappa statistiek. Continue data werd geanalyseerd met behulp van de Intraclass Correlatie Coëfficiënt (ICC). De gecombineerde BASRI scoringsmethoden (BASR-s) en BASRI-t) en in het bijzonder de SASSS lieten een zeer goede betrouwbaarheid zien (ICC 0.85-0.98). Zelfs met een scoringsinterval van twee jaar was de 'intra-observer' betrouwbaarheid goed (ICC 0.85-0.96). De betrouwbaarheid van de relatief nieuwe scoringsmethode voor de heupen (BASRI-h) bleek ook goed te zijn (kappa 0.59-0.60).

Kappa geeft aan in welke mate twee 'observers' verschillen tussen röntgenfoto's kunnen waarnemen. Kappa is dus laag bij een homogene groep data waarbij iedere röntgenfoto dus min of meer dezelfde score heeft gekregen. Dit zou een verklaring kunnen zijn voor de relatief lage 'intra- en interobserver' betrouwbaarheid gevonden voor de SI scoringsmethoden (kappa 0.36-0.70). In deze situatie zijn statistische methoden die de geobserveerde overeenstemming relateren aan de verwachte overeenstemming (zoals kappa en ICC) van beperkte waarde gezien de hoge mate van verwachte overeenstemming. Dit wordt nog eens bevestigd door de relatief lage mediaan scores voor de SASSS scoringsmethoden (mediaan SASSS-totaal 17.5-18.0, mediaan 'gemodificeerde' SASSS 16.3-16.8, range 0-72). Verder geldt dat vanwege de lage prevalentie van radiologisch schade bij de SASSS de ICC toeneemt met als gevolg overschatting van de ICC.

Vanwege deze overwegingen ten aanzien van kappa en ICC hebben we ook concordanties berekend. De volledige concordantie tussen de twee 'observers' was over het algemeen laag (21-76%). Visuele presentatie van de continue data (SASSS scoringsmethoden) door middel van een Bland and Altman plot geeft ook meer inzicht in de data zeker omdat deze plots de verdeling van de data met 'outliners' over de gehele range van geobserveerde data laat zien. Deze plots lieten een maximaal verschil van 26 punten (range 0-72) tussen de twee 'observers' zien.

BASRI-s en BASRI-t waren de enige scoringsmethoden die verandering lieten zien bij een klein aantal patiënten na twee jaar follow-up (respectievelijk 7.5% en 7.4%) gebaseerd op een binomiaal afkappunt. Deze verandering kon niet worden aangetoond met de andere globale of gedetailleerde scoringsmethoden. De 'observers' kwamen in maximaal 52% overeen dat er geen verandering in BASRI-s en BASRI-t score plaats vond. Toch moeten we

helaas concluderen dat er zelden sprake was van relevante verandering omdat de 'observers' maar in 7.5% van de patiënten overeenstemde dat een significante verandering van minimaal 1 graad had plaats gevonden. Dit laatste kan misleidend zijn aangezien de definitie van 1 graad verandering een arbitrair afkappunt is. Het berekende kleinste aantoonbare verschil (SDD) gedefinieerd als afkappunt voor de SASSS scoringsmethoden is relatief groter. Het afkappunt gebruikt voor de SASSS methoden is dus strenger gedefinieerd dan het afkappunt (1 graad verandering) gebruikt voor de BASRI methoden. Dit laatste zou weer een reden kunnen zijn waarom we geen verandering hebben gevonden in SASSS scores. Een mogelijke andere reden is dat 'observer variatie' of 'observer error' niet te onderscheiden is van radiologische progressie in onze studie. Het gebruik van binomiale afkappunten kan betekenen dat men informatie verliest en dat het daardoor niet meer mogelijk is verschillen aan te tonen. Een andere overweging is dat we een niet geselecteerde populatie Bechterew patiënten gebruikt hebben zonder een gegarandeerde mate van ziekte activiteit. In een groep Bechterew patiënten met een hoge ziekte activiteit zouden de resultaten wel eens anders kunnen zijn.

In onze studie is de betrouwbaarheid van de radiologische scoringsmethoden bij de ziekte van Bechterew goed. Helaas waren de scoringsmethoden onder de gegeven scorings condities (gepaard scoren met onbekende chronologische volgorde, resultaten gebaseerd op gemiddelde scores van twee 'observers', binomiale afkappunten en een niet geselecteerde patiënten populatie) niet in staat om na twee jaar follow-up bij een groot aantal patiënten radiologische progressie of verandering aan te tonen.

Resumé en perspectief

De ASAS werkgroep is opgericht vanwege gebrek aan uniformiteit op het gebied van ziekte-activiteit en ziekte-uitkomst in wetenschappelijke studies met betrekking tot de ziekte van Bechterew. Deze werkgroep heeft verschillende domeinen gedefinieerd met kernpunten voor drie verschillende onderzoeksettings bij de ziekte van Bechterew ('DC-ART', 'SM-ARD'/'physical therapy' en 'clinical record keeping'). In navolging op het werk van de ASAS werkgroep hebben wereldwijd diverse onderzoeksgroepen de laatste vijf jaar veel werk verricht door zich voornamelijk te richten op aspecten van ziekte-activiteit en ziekte-uitkomst bij de ziekte van Bechterew. De resultaten in dit proefschrift zijn voor het grootste deel afkomstig van de gegevens van een groot cohort Bechterew patiënten die longitudinaal gevolgd zijn (OASIS). Deze resultaten zijn gericht op aspecten van zowel ziekte-activiteit als ziekte-uitkomst. Hoofdstuk 6 en 7 gaan vooral over uitkomstmaten bij de ziekte van Bechterew waarbij hoofdstuk 6 de ontwikkeling en validatie van een ziekte specifieke kwaliteit van leven vragenlijst (ASQoL) beschrijft. Hoofdstuk 7 beschrijft de vergelijking van alle beschikbare radiologische scoringsmethoden bij de ziekte van Bechterew. De resultaten van deze studie laten zien dat de methoden betrouwbaar zijn maar dat geen van de methoden in staat is bij een groot aantal patiënten radiologische verandering aan te tonen.

over een periode van twee jaar. In de nabije toekomst wordt waarschijnlijk duidelijk welke rol 'Magnetic Resonance Imaging' (MRI) gaat krijgen bij het aantonen van structurele verandering bij de ziekte van Bechterew. Het is met behulp van de MRI mogelijk om aspecten van schade en van ziekte-activiteit aan te tonen op het niveau van zowel de SI gewrichten als van de wervelkolom. Het mogelijke effect van bijvoorbeeld een bekende chronologische volgorde van röntgenfoto's op het aantonen van radiologische verandering bij conventionele radiologie wordt nog onderzocht.

Hoofdstuk 2 tot en met 5 beschrijven aspecten van ziekte-activiteit en de resultaten gepresenteerd in deze hoofdstukken bevestigen dat het meten van ziekte-activiteit bij de ziekte van Bechterew erg moeilijk blijft. Hoewel wereldwijd veel werk is gedaan, is er nog steeds geen uniforme maat die alle aspecten van ziekte-activiteit reflecteert. Acute fase reacties gemeten met CRP en BSE zijn slechts bij een minderheid van de Bechterew patiënten verhoogd. Verder blijkt dat Bechterew patiënten en hun behandeld artsen een heel ander zicht op ziekte-activiteit hebben. Tot nu toe wordt om deze reden ziekte-activiteit vanuit het patiënten perspectief, zoals BASDAI, in combinatie met ziekte-activiteit vanuit het perspectief van de arts en/of verhoogde CRP gebruikt als 'gouden standaard' voor het includeren van patiënten in klinische trials en voor de besluitvorming rondom het starten van anti TNF α therapie in de klinische praktijk. Anti TNF α en andere nieuwe biologische geneesmiddelen zijn recent beschikbaar gekomen bij de behandeling van de ziekte van Bechterew en de effecten lijken veelbelovend. In de toekomst zullen studies die de effecten van deze potente geneesmiddelen evalueren gebruikt kunnen worden voor validatie en vergelijking van de door de ASAS geselecteerde meetinstrumenten. Hopelijk leiden de resultaten van deze studies uiteindelijk naar de ontwikkeling van een uniform en valide meetinstrument dat alle aspecten van ziekte-activiteit bij de ziekte van Bechterew reflecteert.

DANKWOORD

Dankwoord

Als eerste wil ik alle patiënten bedanken die zonder direct eigenbelang heel veel tijd in het OASIS project hebben gestoken: het invullen van de vele vragenlijsten en de uitgebreide onderzoeken waren geen sinecure. Tijdens deze veelvuldige contacten heb ik heel veel van jullie geleerd.

Bij de uitvoering van dit project zijn behalve de patiënten ook vele anderen betrokken geweest. De onderzoeksassistenten: Gisela, Lily en Anita in Maastricht en Maryse in Parijs, jullie hebben het werk, dat niet altijd even enerverend was, met veel inzet verricht. Heel erg veel dank hiervoor.

Reumatoloog Hille van der Tempel en de dames van de poliklinieken reumatologie in Sittard, Maasmechelen en Maastricht bedankt voor jullie gastvrijheid en bijdrage. The OASIS teams from Ghent (Belgium) and Paris (France) reumatologists professor Herman Mielants, Kurt de Vlam and professor Maxime Dougados many thanks for all your contributions. Dear Maxime Dougados you are known as an authority not only in the field of Ankylosing Spondylitis but in several other fields of research, thank you for your support. Beste Kurt, het samen scoren van 1736 en later nog eens 2604 röntgenfoto's zal ik niet licht vergeten. Al mijn directe collega's in Maastricht waar ik op enig moment mee heb samen gewerkt op het gebied van onderzoek en de klinische reumatologie: Liesbeth, Erik, Thea, Arco, Astrid, Simone, Guy, Astrid, Karen, Debbie, Annelies, Marijke, Peter, Christine en Piet wil ik bedanken voor de fijne samenwerking en hun gezelligheid.

Maarten Boers, jij hebt mij tijdens mijn co-schappen enthousiast gemaakt voor de reumatologie en het wetenschappelijk onderzoek. Ik heb met heel veel plezier aan ons 'spalken'project gewerkt. Jij hebt mij geleerd dat in wetenschappelijk onderzoek het proces minstens zo belangrijk is als het uiteindelijke resultaat.

Beste Erik, jou wil ik speciaal bedanken voor je onontbeerlijke steun bij het maken van de enorme OASIS database. Je hulpvaardigheid en je heldere uitleg over verschillende statistische technieken die betrekking hadden op mijn eerste analyses heb ik als zeer bijzonder ervaren.

Yolanda Soons, wat zouden ze zonder je moeten beginnen? Je was en bent de spil van de organisatie van de werkgroep reumatologie in Maastricht.

Beste Robert, na mij vertrek uit Maastricht ben jij betrokken geraakt bij OASIS en het Bechterew onderzoek. Onze samenwerking tijdens de ontwikkeling en het schrijven van het laatste artikel (hoofdstuk 5) van dit proefschrift heb ik, ondanks de afstand en de bijbehorende communicatie problemen, als zeer motiverend ervaren.

Mijn paranyf en lieve collega Liesbeth Heuft heeft samen met Astrid Wanders het OASIS project van mij overgenomen. Ik wens jullie veel plezier en wijsheid bij het uitvoeren, analyseren en het schrijven. De resultaten worden alleen al door de duur van het project steeds leuker en interessanter.

Mijn promotor Désirée van der Heijde wil ik graag bedanken voor haar begeleiding. Beste Désirée, ik heb bewondering voor je gedreven, doelmatige, zakelijke en perfectionistische

manier van wetenschap bedrijven. Jij bent de geestelijk moeder van OASIS en inmiddels kent OASIS verschillende oppasmoeders waarvan ik de eerste was. De strijd over tijdsbesteding aan klinisch werk en onderzoek heb ik van vrijwel het begin tot aan het einde van deze promotie moeten voeren. Vaker voelde ik me een koorddanser die ternauwernood het evenwicht kan bewaren en in het begin van het project ben ik wel eens van het koord gevallen, maar al doende leert men. Al met al is het een enorme verrijking.

Mijn opleider en tweede promotor Sjef van der Linden was tijdens het project altijd op de achtergrond aanwezig en indien nodig had hij een sturende rol. Beste Sjef, het was vaak moeilijk jouw gedachten en doelstellingen in te schatten, waarbij jouw introverte en mijn extroverte persoonlijkheid wel eens botsten maar ik ken weinig mensen in een soortgelijke functie die zo zuiver, ongevoelig voor status en macht, met hun positie om kunnen gaan. Bedankt voor een prima opleiding waarvan je de waarde pas kunt inschatten als je het schip verlaten hebt.

Mijn collega's uit het academisch ziekenhuis Groningen: Martin, Miek, Martha, Hendrika, Marcel, Bouke, Ingrid, Liesbeth en Thea, bedank ik voor hun begrip en respect ten aanzien mijn keuze voor een andere werkplek. Betere collega's had ik me niet kunnen wensen!

Beste Nella, George, Ed en Tim, het reumatologisch netwerk in Friesland staat als een huis, daar kunnen velen nog wat van leren. Jullie kleurrijke persoonlijkheden zullen er voor zorgen dat de samenwerking nooit saai wordt en steeds weer nieuwe impulsen bevat.

Lampkje, Nynke, Foka, Joke, Jeanine en Astrid bedankt voor jullie hand- en spandiensten.

Beste Elsa, bedankt voor je 'finishing touch' zichtbaar in de lay-out van dit boekje. De Friezen zijn zoals het land eruit ziet, recht door zee zonder veel franje. Ik vind het net als jij heel prima om in Friesland te wonen.

Lieve vrienden zonder de afleiding en steun van jullie vriendschap had ik het niet volgehouden.

Lieve pappa, mamma en Ellen jullie zijn mijn basis en dat zegt genoeg! Oneindig veel dank voor jullie liefde, grenzeloze steun en vertrouwen.

Liefste Christiaan, het moest maar eens af zijn! Ik ken niemand met zo veel geduld en vertrouwen. Beide hebben we concessies moeten doen ten aanzien van onze ambities en vrijetijdsbesteding maar intussen hebben ook wij aardig leren koorddansen. Ik ben heel trots op je keuze om naast je werk als oncologisch chirurg ook een deel van de tijd voor Emma en Sterre te zorgen met veel meer geduld dan ikzelf. Deze zorgtaak is binnen de chirurgie zeker nog geen voor de hand liggende en alom gerespecteerde keuze.

Liefste Emma voor jou heb ik het meeste bewondering. Verhuisd van Maastricht naar Leeuwarden, van meer dan fulltime opvang naar bijna fulltime opvang en dit allemaal voor de ambities van je ouders. Je doorstond dit alles alsof het de gewoonste zaak van de wereld was, je flexibiliteit lijkt grenzeloos. Door je innemende en vasthoudende persoonlijkheid en daarbij een leuk snoetje heb je al heel wat mensen voor je gewonnen.

Liefste Sterre, je bent er net maar je lach is nu al onvergetelijk. Ik hoop dat we samen nog veel plezier mogen hebben.

PUBLICATIONS

Publications

Spoorenberg A, Boers M, van der Linden S. Wrist splints in rheumatoid arthritis: what do we know about efficacy and compliance? *Arthritis Care Res* 1994; Jun; 7: 55-7. Review. No abstract available.

Spoorenberg A, Boers M, Van der Linden S. Wrist splints in rheumatoid arthritis: a question of belief? *Clin Rheumatol* 1994; Dec; 13: 559-63.

Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van der Tempel H, van der Linden S. A comparative study of the usefulness of the Bath Ankylosing Spondylitis Functional Index and the Dougados Functional Index in the assessment of ankylosing spondylitis. *J Rheumatol*. 1999 Apr; 26: 961-5.

Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van der Tempel H, van der Linden S. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. *J Rheumatol* 1999; Apr; 26: 980-4.

van der Heijde D, Spoorenberg A. Plain radiographs as an outcome measure in ankylosing spondylitis. *J Rheumatol* 1999; Apr; 26: 985-7. Review.

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. Radiological scoring methods in ankylosing spondylitis: reliability and sensitivity to change over one year. *J Rheumatol* 1999; Apr; 26: 997-1002.

van Tubergen A, Coenen J, Landewe R, Spoorenberg A, Chorus A, Boonen A, van der Linden S, van der Heijde D. Assessment of fatigue in patients with ankylosing spondylitis: a psychometric analysis. *Arthritis Rheum* 2002; Feb; 47: 8-16.

Boonen A, van der Heijde D, Landewe R, Spoorenberg A, Schouten H, Rutten-van Molken M, Guillemin F, Dougados M, Mielants H, de Vlam K, van der Tempel H, van der Linden S. Work status and productivity costs due to ankylosing spondylitis: comparison of three European countries. *Ann Rheum Dis* 2002; May; 61: 429-37.

Spoorenberg A, van der Heijde D, Dougados M, de Vlam K, Mielants H, van der Tempel H, van der Linden S. Reliability of self assessed joint counts in ankylosing spondylitis. *Ann Rheum Dis* 2002; Sep; 61: 799-803.

Jansen TL, Spoorenberg A, Houtman PM, van Roon EN. The effect of comedication with folic or folinic acid on the toxicity and efficacy of methotrexate in rheumatoid arthritis: a randomized, double blind, placebo controlled study of 48 weeks. *Ned Tijdschr Geneeskd* 2002; Nov 9; 146: 2168; author reply 2168. Dutch. No abstract available.

Auleley GR, Benbouazza K, Spoorenberg A, Collantes E, Hajjaj-Hassouni N, van der Heijde D, Dougados M. Evaluation of the smallest detectable difference in outcome or process variables in ankylosing spondylitis. *Arthritis Rheum* 2002; Dec 15; 47: 582-7.

Doward LC, Spoorenberg A, Cook SA, Whalley D, Helliwell PS, Kay LJ, McKenna SP, Tennant A, van der Heijde D, Chamberlain MA. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. *Ann Rheum Dis* 2003; Jan; 62: 20-6.

Heuft-Dorenbosch L, Spoorenberg A, van Tubergen A, Landewe R, van der Tempel H, Mielants H, Dougados M, van der Heijde D. Assessment of enthesitis in ankylosing spondylitis. *Ann Rheum Dis* 2003; Feb; 62: 127-32.

van Tubergen A, Landewe R, Heuft-Dorenbosch L, Spoorenberg A, van der Heijde D, van der Tempel H, van der Linden S. Assessment of disability with the World Health Organisation Disability Assessment schedule II in patients with ankylosing spondylitis. *Ann Rheum Dis* 2003; Feb; 62: 140-5.

Boonen A, van der Heijde D, Landewe R, Guillemin F, Rutten-van Molken M, Dougados M, Mielants H, de Vlam K, van der Tempel H, Boesen S, Spoorenberg A, Schouten H, van der Linden S. Direct costs of ankylosing spondylitis and its determinants: an analysis among three European countries. *Ann Rheum Dis* 2003; Aug; 62: 732-40.

Boonen A, van der Heijde D, Landewe R, Guillemin F, Spoorenberg A, Schouten H, Rutten-van Molken M, Dougados M, Mielants H, de Vlam K, van der Tempel H, van der Linden S. Costs of ankylosing spondylitis in three European countries: the patient's perspective. *Ann Rheum Dis* 2003; Aug; 62: 741-7.

Spoorenberg A, de Vlam K, van der Linden S, Dougados M, Mielants H, van der Tempel H, van der Heijde D. Radiological scoring methods in ankylosing spondylitis: reliability and change over one and two years. *J Rheumatol* [in press]

Curriculum Vitae

Anneke Spoorenberg werd geboren op 18 februari 1966 te Leende in Noord-Brabant. Haar HAVO- en haar VWO diploma behaalde zij respectievelijk in 1982 en in 1984 aan het van Maerlantlyceum te Eindhoven. Alvorens zij in 1986 startte met de studie geneeskunde aan de Universiteit van Maastricht heeft zij in 1985 de propaedeuse gezondheidswetenschappen aan deze universiteit behaald. In 1990 ontving zij haar doctoraal diploma geneeskunde en in 1992 haar arts diploma. Tijdens de 2 jaar voorafgaande aan het behalen van het arts examen heeft zij onderzoek gedaan naar het spalk gebruik bij patiënten met reumatoïde artritis bij de werkgroep reumatologie in het Academisch Ziekenhuis Maastricht (begeleider: prof. dr. M. Boers). Na het behalen van het arts examen is zij ruim 3 jaar werkzaam geweest als arts assistent interne geneeskunde in het toenmalig Sint Joseph ziekenhuis te Veldhoven (thans Máxima Medisch Centrum) in het kader van de vooropleiding tot reumatoloog (opleider: dr. P. Gerlag). In 1996 was zij werkzaam als arts assistent interne geneeskunde in het academisch Ziekenhuis Maastricht en is zij tevens met de Outcome in Ankylosing Spondylitis International Study (OASIS) gestart (begeleider: prof. dr. D. van der Heijde). Deze studie heeft onder andere geleid tot de onderzoeken waarvan de resultaten verwerkt zijn in dit proefschrift. Van 1997 tot 2000 was zij werkzaam als arts assistent reumatologie in het academisch ziekenhuis Maastricht (opleider: prof. dr. Sj. van der Linden) waarna in november 2000 de registratie tot reumatoloog volgde. Na het afronden van de opleiding tot reumatoloog was zij 1,5 jaar werkzaam als stafid in het Academisch Ziekenhuis te Groningen. Sinds april 2002 werkt zij als reumatoloog in de Friese reumatologen maatschap. In 1999 is zij getrouwd met Christiaan Hoff en samen hebben zij twee dochters: Emma (2000) en Sterre (2003).