Health Outcomes In Ankylosing Spondylitis:
An Evaluation Of Patient-Based And Anthropometric Measures

Kirstie Louise Haywood

BSc (Honours) Physiotherapy Victoria University of Manchester

A thesis presented for the degree of Doctor of Philosophy

University of York

Department of Health Sciences and Clinical Evaluation

September, 2000

Abstract

The thesis makes three contributions to the evidence base for outcome measurement in Ankylosing Spondylitis (AS). First, a systematic review and evaluation of the entire range of patient-based and anthropometric measures of outcome applied in published studies of AS represents the first detailed and explicit synthesis of evidence relating to the development, measurement properties, acceptability and feasibility of outcome measures applied in current practice in AS. Secondly, the development of the first AS-specific individualised measure of disease-related quality of life, the Patient Generated Index for AS (PGI-AS). Thirdly, the first comparative evaluation of the PGI-AS and an evidence-based selection of disease-specific, anthropometric and generic measures of outcome in patients with AS. This study is the largest clinic-based and one of the largest multi-centre postal evaluations of outcome measures in AS within the United Kingdom. The study also describes the most rigorous process of instrument testing reported in AS.

The systematic review describes the wide diversity of outcome measures applied in the evaluation of AS. The first standardised and evidence-based package of instruments to fulfil domains considered important in the evaluation of AS is described. However, the evaluation of these instruments indicated that no instrument adequately fulfilled the required measurement properties and practical criteria considered necessary for use in individual evaluation, and no clear recommendation for the disease-specific evaluation of health-related quality of life, disease activity, functional disability or spinal mobility could be made. Several instruments can be recommended for use in group analysis and in clinical trials. Gaps in the availability of evaluative instruments for specific domains of health and the need for modifications and further research to evaluate measurement properties of new, modified and widely used instruments are described. Initial evidence suggests satisfactory measurement properties for the PGI-AS and further evaluation of the role of individualised measures in routine practice, research, clinical decision making and directing patient-centred management is required.

List of Contents

List of Tables

List of Figures

Acknowledgement

The patient data reported in this thesis was collected over a period of 18-months. 200 patients were involved in the various sections of the clinic-based study and an additional 353 patients provided information by returning postal questionnaires. This was made possible through the cooperation of the Staffordshire Rheumatology Centre which permitted access to their patient database for both the clinic based and postal studies. In addition, the participation of an additional five rheumatology centres across the UK supported the recruitment of patients for the postal survey. Many thanks must go to all Consultant Rheumatologists and Physiotherapists who supported this collaboration. Thanks must also go to the health-care professionals, clinic and secretarial staff from all participating centres for their support in patient recruitment and to ensuring the smooth running of the study.

Particular thanks go to Ms. Jackie Waterfield for her assistance with patient evaluation during the clinic survey; to Dr. Kelvin Jordan who provided valuable computing and statistical advise; and to Mrs. Mary Simpson for her assistance with the collation of survey data. The support from members of the Stoke-on-Trent branch of the National Ankylosing Spondylitis Society (NASS) throughout the study has also been greatly appreciated. Thanks must go to Mrs Irene Fenton and her colleagues at the North Staffordshire Medical Institute Library for their support in database searching and article retrieval.

The following agencies are acknowledged for their generosity and financial support: the Staffordshire Rheumatology Centre, the Hospital Savings Association and the Arthritis Research Council.

I would like to thank my three supervisors Drs. Andrew Garratt, Krysia Dziedzic and Peter Dawes for their continued support, comments and helpful criticism throughout the study. I would also like to thank Marc Haywood, my family and close friends for their continued strength and encouragement, without whose support this thesis could not have been written. To all of these people I am deeply indebted.

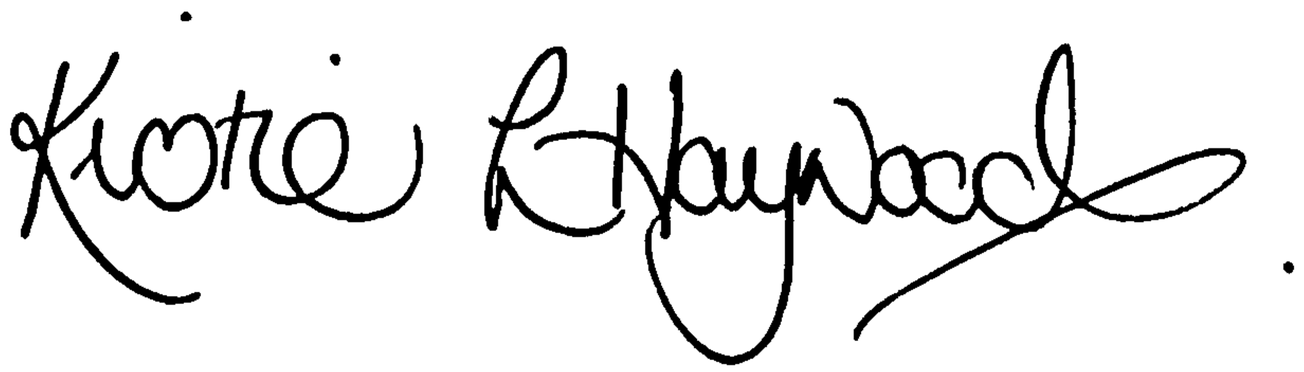
'And let your best be for your friend.
If he must know the ebb of your tide, let him know its flood also.'

(Gibran, 1980)

Authors Declaration

The opportunity for undertaking this thesis arose through being awarded a full-time studentship from the Department of Health Sciences and Clinical Evaluation, University of York. The focus of the thesis is the evaluation of health outcome measurement in Ankylosing Spondylitis and during a four year period I have worked full-time on this project with support from members of a multi-disciplinary research advisory group (RAG). In addition to myself the RAG included a specialist in health outcome measurement, a consultant rheumatologist and a chartered physiotherapist. I took lead responsibility for developing the protocol, day to day management of the project, patient recruitment, data collection, data management, statistical analysis and report writing.

This thesis has been composed by the candidate and has not been accepted in any previous application for a degree. All quotations have been distinguished by quotation marks and sources of information acknowledged.



Kirstie Louise Haywood

September, 2000

Chapter 1 The Measurement of Health Outcome and Ankylosing Spondylitis

1.1 Introduction

This chapter provides an overview of the measurement of health outcome in general and an introduction to Ankylosing Spondylitis, and places the work that follows in context. Section 1.2 describes the status of outcome measurement in general and the role of patient-based evaluation. Ankylosing Spondylitis (AS), its impact and management, and the present status of outcome measurement in AS are discussed in sections 1.3 and 1.4 respectively. Section 1.5 describes the foundation for the empirical work that follows.

1.2. The measurement of health outcome

Historically the outcome of health care was based on the biomedical model reflecting a disease-based view of outcome that considered ill-health to be an objective, measurable concept (Jones, 1992; Jenkinson, 1994). At an individual level this model focussed on the presence or absence of disease. The belief was that a quantitative relationship between organ impairment, ill-health and well-being existed and correction of disease at the organ level would positively influence wider issues of ill-health (Jones, 1992). History of the disease process and laboratory and radiographic based assessments were the mainstay of a clinician's evaluative repertoire. At a population level indicators of disease consisted of morbidity rates, disease incidence and prevalence, and most commonly mortality rates (McDowell and Newell, 1996).

Where initially medical intervention was accepted for the ability to prolong life, with advancing medical technologies acute, life-threatening illnesses no longer dominated the medical picture (McDowell and Newell, 1996). Long term, chronic illness associated with increased survivorship became the prime focus of health care, demanding increasing health care resources and a change in emphasis in management and evaluation (Fitzpatrick et al, 1998a). Where mortality is no-longer the main concern of outcome evaluation there is increased relevance of the World Health Organisations (WHO) broad definition of health as:

> 'physical, mental and social well-being, not merely the absence of disease and infirmity'
>
> (WHO, 1947)

When it was first proposed the definition was considered immeasurable (McDowell and Newell, 1996), but advances in measurement techniques have operationalised concepts proposed in the definition and improved acceptance over recent years. Within this definition the impact of disease may be considered in terms of the impairment, disability and handicap associated with ill-health (Carr, 1996). Impairment represents disease impact at the anatomical, physiological or psychological level and disability relates to the disadvantage experienced by an individual when performing an activity as a result of the impairment. Handicap considers the broader impact of ill-health as the role and social disadvantage experienced by an individual. Although traditional measures of impairment, for example, laboratory based assessments, are informative to clinicians the simple dichotomy of health and illness presented by the biomedical model was no longer considered an adequate representation of disease (van der Linden and van der Heijde, 1995; Fitzpatrick et al, 1998a). Health is a complex and abstract concept that requires indicators that focus attention on the quality of survival and wider issues of relevance to patients, health care professionals and health care providers (McDowell and Newell, 1996; Ware, 1998).

Associated with the acknowledged inadequacy of the biomedical assessment was increasing evidence that a patient's subjective perception of health, the psychosocial impact of disease and treatment were important to the evaluation of outcome and allocation of health care (Barlow et al, 1992; Jenkinson, 1994). This information was considered complimentary to traditional assessments by demonstrating the broad impact of health care and was supported by evidence that the patient is the best judge of disease impact (Albrecht, 1994). The importance of the patient in this evaluative role is highlighted by the following definition of medical outcome as:

> 'the extent to which a change in a patients behavioural functioning or well-being meets the patients needs or expectations' (Ware, 1997)

The measurement of a patient's subjective perception of health-related quality of life (HRQL) is now recommended as a core component in the assessment of health outcome (Albrecht, 1994; Fitzpatrick et al, 1998a), and treatment that improves only traditional biomedical features without benefiting HRQL may be considered to have only limited medical success.

Despite the lack of consensus definition of HRQL the construct develops on the WHO definition of health through the inclusion of health status and social well-being (Guyatt et al, 1993; Albrecht, 1994). Most authors agree that HRQL is a multi-dimensional concept, with patient-reported symptoms of physical and mental health major components (Fitzpatrick, 1993a; Ward, 1998; Ware, 1998). Questionnaires which attempt to place the patient at the centre of the evaluative process have been developed for a wide range of health problems and have invariably been referred to as measures of quality of life or HRQL, often without further definition of the concept addressed by the instrument. The evaluation of quality of life should, in theory, consider a very broad concept of life that does not focus simply on the impact of ill-health (Wolfe, 1995). The focus of HRQL evaluation is towards aspects of life that might in principle be influenced by health and health care (Patrick and Erickson 1993; Ware, 1997; Jenkinson et al, 1998a). However, despite the apparent congruency between the concepts of impairment, disability, handicap and HRQL this relationship is not hierarchical and evaluation should consider all elements to provide a complete representation of disease impact (Carr and Thompson, 1994; Carr, 1996).

The goals of management of chronic and often incurable conditions focus on symptom amelioration, the restoration and preservation of function and well-being and enhancement of a patients HRQL (Barlow and Barefoot, 1996). Multi-disciplinary health care teams are generally involved in the management process with patients playing a central role. Traditional biomedical assessments provide an incomplete picture of the wide impact of various therapeutic interventions on health (Streiner and Norman, 1995). Therefore, methods of evaluation that provide appropriate feedback on the relative success of interventions associated with the goals of management are required. Asking patients their own views about ill-health and health care retains their position in the multi-disciplinary team throughout the cyclical process of treatment planning, implementation and evaluation whilst providing a broad representation of health. Where relative and informative evidence is gained from evaluation this may influence the effectiveness and quality of care and empower patients to undertake or maintain an active role in management. Health care reforms have focussed attention towards the evaluation of medical outcomes, especially towards those reflecting consumer or patient subjective perceptions and preferences (Albrecht, 1994; Ware, 1998). Evidence based practice with a foundation based on a

combination of traditional objective measurement and patient-based measures of outcome may be more informed and responsive to the challenges and demands of ill-health for which health care aims to provide.

1.2.1 Taxonomy of measures of health-related quality of life

Two broad approaches to measuring patient perceptions of HRQL can be described: generic instruments that provide a broad summary of HRQL, and specific instruments that focus on issues of relevance to a specific disease or patient group. Generic instruments are not age, disease or treatment specific and contain multiple HRQL concepts of relevance to patients and the general population, supporting application in both populations (Guyatt et al, 1989a; Ware, 1997). Population-based normal values can be calculated, which supports data interpretation from disease-specific groups (Ware, 1997).

Two classes of generic instrument can be described: health profiles and utility measures. Scores on different domains of HRQL covered by a single health profile are presented separately to support data interpretation, therefore reflecting a clinical perspective (McDowell and Newell, 1996). Sometimes a single or summary score may be generated, but proponents argue that measurement is most meaningful within separate domains. The Short Form 36-item Health Survey Questionnaire (SF-36) is a widely used example of a generic health profile (Ware, 1997). The items cover eight domains of HRQL including physical and social functioning and mental health. Responses to each item are summed (0-100), where 0 is the worst possible HRQL, and 100 the best. Mental and physical component summary scales may also be generated. Population norms have been calculated in several countries (Ware, 1997).

The values and preferences for outcome generated by the patient (direct weighting) or the general population (indirect weighting) provide external weightings for utility measurement (Garratt et al, 2000). Although utility measures can cover several domains of HRQL, the weighting generates a single index that relates HRQL to death (0) or perfect health (1)(Guyatt et al, 1993). The EuroQol (EQ-5D)(EuroQol Group, 1990) is an example of a utility measure that incorporates indirect valuations of health states (Kind et al, 1998). A benefit of utility measures is the recommendation for use in cost-utility economic analysis, but a disadvantage is that the single score limits data interpretation (Guyatt et al, 1993).

Specific instruments may be specific to a particular disease (e.g. AS), to a patient population (e.g. child health), to a specific problem (e.g. pain, limited range of movement), or to a described function (e.g. functional ability) (Guyatt et al, 1993). For example, the Revised Leeds Disability Index (RLDQ) is an AS-specific measure of functional disability (Abbott et al, 1994). Responses to each item are summed (0-48), where 0 is the best possible functional ability.

The broad content of generic instruments supports identification of co-morbid features and treatment side-effects that may not be captured by specific instruments, but this may reduce instrument responsiveness to small and important disease-specific changes. Disease-specific instruments may have greater clinical appeal due to the specificity of content, and an associated increased responsiveness to specific change in condition (Guyatt et al, 1993; Garratt, 1997). Their combined use is therefore recommended in the evaluation of health outcome (Guyatt et al, 1993; McDowell and Newell, 1996).

1.2.2 Application of measures of health outcome

To be suitable for use in an evaluative role, instruments should be acceptable and feasible for the required application, and possess certain measurement properties: reliability, validity and responsiveness to change (Fitzpatrick et al, 1998a). These properties are addressed in detail in the ensuing chapters.

Patient-based measures of health outcome are intended to provide supplementary information to traditional biomedical assessments and several forum for their application have been described, including clinical research, routine clinical practice and health policy (Guyatt et al, 1993; Ware, 1997). By far the greatest evidence is available to support application in clinical research, where the results may, for example, support evaluation of comparative management strategies in a controlled trial. A measure of patient perceived physical fitness (Astrand Fitness Index)(Astrand and Rodahl, 1977) and the Health Assessment Questionnaire for the Spondyloarthropathies (HAQ-S)(Daltroy et al, 1990), a disease-specific measure of functional ability, suggested improved levels of physical fitness and function in AS patients randomly assigned to receive supervised group exercise therapy over a nine-

month period, versus those pursuing a home exercise programme only (Hidding et al, 1994a).

Potential benefits from the application of patient-based instruments in clinical practice include: improved patient-clinician communication (Jenkinson et al, 1996), the identification of functional and psychosocial problems that may have previously been missed (Greenhalgh and Meadows, 1999), and regular standardised patient monitoring where the information may assist in clinical decision making at the individual patient level (Ware, 1997). However, evidence suggests that there has been a limited adoption of these instruments in routine practice (Bellamy et al, 1998, 1999; Greenhalgh and Meadows, 1999). Attitudinal, practical and methodological barriers have been cited as reasons for their limited uptake (Bellamy et al, 1999). A review of the effectiveness of including patient-based instruments in routine clinical practice reported that there was limited evidence to support the proposed benefits (Greenhalgh and Meadows, 1999). In addition, there was little evidence to indicate that their use substantially influenced patient management or improved outcome.

At a managerial and policy level, measures of HRQL support the comparison of costs and benefits of competing health care programmes, where managers will seek to provide the best health care for the best price (Ware, 1997). Rationing of health care is the inevitable consequence of limited resources, and the use of reliable, valid and responsive measures of HRQL may provide beneficial information to support the distribution of resources for health care (Ware, 1997; Garratt, 1997). Guyatt et al (1993) suggest that generic instruments are of greatest interest to the policy maker or manager because they consider consumer needs and preferences, whilst facilitating comparison of HRQL and economic evaluation across populations and conditions. Alternatively, disease-specific measures of HRQL, due to the significance of item content, are of greatest relevance to patients and health professionals. However, the limitations of data from patient-based instruments and its interpretation when supporting important clinical decisions at both individual and policy level should be recognised (Bindman et al, 1990; Jenkinson, 1995).

1.3 Ankylosing Spondylitis

The focus of the empirical work that follows in subsequent chapters is the evaluation of approaches to measuring health outcome in Ankylosing Spondylitis (AS).

AS is a chronic systemic, often progressive, inflammatory disorder, primarily affecting the sacro-iliac joints of the pelvis, the axial skeleton and the thoracic cage (Russell, 1998). Peripheral joints, entheses, and extra-articular sites may also be affected (Dawes et al, 1988). The subsequent impact of AS on a patient's health-related quality of life encompasses broad multi-dimensional issues including social interactions, role and physical functioning, psychological well-being, impact of treatment, and the actual disease symptoms.

1.3.1 Epidemiology and aetiology

The true prevalence of AS is unknown but in virtually any racial group it is reported to reflect the prevalence of a genetic marker, the Human Leucocyte Antigen (HLA) B27 (Rigby, 1991; van der Linden and van der Heijde, 1995). However, HLA B27 lacks specificity for AS, and where in the healthy caucasian population HLA B27 may have a prevalence of between 7-12%, the prevalence of AS has been estimated as between 0.1-0.4% (Rigby, 1991; Johnsen et al, 1992), and as high as 1-2% of the caucasian population in certain circumstances (Pal, 1987; Johnsen et al, 1992). In caucasian patients with AS, 90-95% are HLA B27 positive (Albert and Scholz, 1987; van der Linden and van der Heijde, 1995). These results suggest that there may be 60-70,000 clinically diagnosed cases of AS in the United Kingdom (UK)(Barlow et al, 1993a). However, where results are based on hospital records, and possibly biased towards the more severe cases, the true prevalence of AS may be underestimated. It is suggested that 750,000 individuals may have AS, if sub-clinical or very mild forms of the disease are taken into account (Barlow et al, 1993a).

The aetiology of AS is unknown. It is hypothesised that environmental factors, for example infection, may act as a trigger in genetically predisposed individuals (Calin, 1985; Carbone et al, 1992). The role of genetics in causation is strengthened by the accumulated evidence of a familial link, and the inherited susceptibility marked by HLA B27 (Carbone et al, 1992; van der Linden and van der Heijde, 1996). However, the relative importance of these factors, and the mechanism by which AS and HLA B27 are related is unknown and remains under investigation.

AS was traditionally described as a disease of young men. However, recognition of the disease in females has improved and male to female ratios of between 2-5:1 have

been variously reported (Calin, 1985; Gran and Husby, 1998). The peak incidence of disease onset is between 25-34 years of age, and onset after 55 years of age is unusual (Carbone et al, 1992). However, evidence suggests that the disease does not 'burn out', and most patients remain symptomatic throughout most of their life, with pain a dominant feature (Thompson and Chalmers, 1993). Although AS is considered generally not to be life threatening, there are few studies of AS mortality, and no studies prior to the treatment of AS with radiotherapy in the 1940's and 1950's (Symmons, 1996). Patients with most severe disease often received radiotherapy during this period, and excess mortality related to leukaemia and malignancy has been reported (Ramos-Remus and Russell, 1992; Symmons, 1996), thereby confounding subsequent studies of mortality. However, deaths attributable to AS generally occur in patients with longer-standing disease, of more than 20 years duration, and a population based study reported an 88% relative survival rate when compared to a sex and age matched group (Carbone et al, 1992). Cardiovascular abnormalities or 'violent death' are often the cause of death (Symmons, 1996). For example, fracture of the immobile and osteoporotic vertebrae following a fall or even minor trauma. However, outcome is notoriously difficult to predict and disease progress varies widely between patients (Calin, 1985; Goodacre et al, 1991).

1.3.2 Pathophysiology

The primary pathological site of AS is the enthesis (Calin, 1985). That is, the insertion of ligament, tendon or joint capsule into bone. Entheses are found in synovial and cartilagenous joints and at extra-articular sites. For example, the insertion of intercostal ligaments at the sternocostal margins. Although not specific to AS, the ensuing enthesopathy is a hallmark of the disease (Haslock, 1993; van der Linden and van der Heijde, 1996). The major feature of the pathological process involves inflammation which affects the synovium, articular capsules, fibrocartilagenous joints and entheses (Freemont, 1987). Inflammation is followed by a healing process typified by calcification and bony ankylosis (Haslock, 1993). At the discovertebral junction calcification leads to the development of slender outgrowths from the vertebral margin, referred to as syndesmophytes or enthesophytes (Dziedzic, 1998; Haslock, 1993). With continuing inflammation and repair these may grow and eventually bridge the gap between vertebra, resulting in the bony ankylosis characteristic of AS. Fusion of syndesmophytes, and capsular and ligamentous

ossification leads to the characteristic, progressive spinal rigidity, and the 'bamboo spine' of advanced disease.

Spinal involvement varies between patients, as does the speed of disease progression. In some patients the disease may be relatively benign with pathology limited to the pelvis (Carette et al, 1983). Alternatively, the disease may follow a rapidly progressive course with involvement of the whole spine, thoracic cage, peripheral joints and extra-articular features (Calin, 1985). Reduced joint mobility, particularly limited spinal mobility and chest expansion, feature strongly in the diagnostic criteria for AS (Modified New York Criteria - van der Linden et al, 1984), and have influenced assessment in AS for many years. However, diagnosis of AS is difficult and consideration of all presenting features is important in clinical practice (Dziedzic, 1998). A delay in diagnosis of up to six years in men, and nine years in females has been reported (Calin et al, 1988), but this has reduced with improved recognition of the female presentation of the disease (Dalyan et al, 1999).

Multiple entheses may be involved in the disease process, but the tarsal region, including the insertion of the Achilles tendon and the plantar fascia, has been reported to account for between 26.7-43.5% of all enthesitis in adult onset AS (Burgos-Vargas, 1990). Patients may experience pain following palpation at the site of entheses actively involved in the disease process, and two clinical measures have been developed in an attempt to quantify enthesitis as a reflection of disease activity (Dawes et al, 1987; Mander et al, 1987). Peripheral joint involvement has been reported in between 20-30% of adults with AS, primarily involving the gleno-humeral or hip joints (Dalyan et al, 1999). The incidence of peripheral joint involvement in patients with disease onset before 20 years of age is almost double that seen in patients with a disease onset after this age (40% versus 22%), and involvement of the hip is reported to be most likely in the first ten years of symptoms (Carrette et al, 1983). Extra-articular features have been reported in 5-25% of patients and include uveitis, cardiovascular and gastrointestinal complications, respiratory embarrassment, and renal disease (Dziedzic, 1998).

1.3.3 Impact of AS on health-related quality of life

The following domains will be adopted to summarise evidence relating to the disease impact on a patients HRQL:

- symptoms
- physical function
- role function
- social interaction
- emotional well-being

Symptoms

Symptoms have been defined as a patient's subjective perception of an abnormal physical, emotional or cognitive state (Anderson and Burckhardt, 1999). Pain and stiffness are the most frequently mentioned symptoms of patients with AS (Dziedzic, 1998), closely followed by reports of fatigue and sleep disturbance (Ward, 1998).

The cause of pain in AS is multi-faceted and may be contributed to by the pathophysiological disease-process, the biomechanical impact, AS-related systemic illness, and depression. An early study of AS reported that pain and/or stiffness in the lumbar spine or buttocks were the presenting symptoms in 73.4% of patients, whereas 24% indicated that the initial symptoms involved peripheral joints (Dudley Hart, 1955). Evidence suggests that pain fluctuates throughout the course of the disease (Thompson and Chalmers, 1993; Ward, 1998). A longitudinal cohort evaluation of 151 male Army veterans with AS reported the most severe pain experience in the first 10 years after diagnosis (Carrette et al, 1983). At re-examination (average disease duration 38 years) 68% reported pain as a predominant feature of the disease, and in 30% of these it was considered moderate or severe. 30% of patients reported no pain at re-examination (Carette et al, 1983). However, these results may be influenced by the dominance of male patients, and the loss to follow-up (n=51). A hospital based postal survey of 1492 AS patients reported pain and disease activity to be equivalently high in patients with a disease duration of less than 10 years or of more than 30 years, when assessed by an AS-specific instrument (Bath Disease Activity Index)(Kennedy et al, 1993).

Stiffness is an important clinical feature in AS and is included in the diagnostic criteria. Worse on awakening (Jamieson et al, 1995), or after prolonged periods of immobility, it often lasts for more than two hours (Garrett et al, 1994) but generally eases with movement. Stiffness is a complex symptom and although differentiation

from limited mobility or pain is difficult, patients are often able to distinguish between symptoms (Dziedzic, 1998). Although a strong association between stiffness severity and duration has been reported (Garrett et al, 1994), the assessment of severity is more informative than duration in inflammatory conditions (Hazes et al, 1993). A high correlation between stiffness and pain (Garrett et al, 1994) and stiffness and change in global health following group exercise therapy has been reported (Hidding and van der Linden, 1995).

Fatigue has been recognised as an important complaint by up to 65% of patients with AS (Garrett et al, 1994; Ward, 1998), and up to 11% of patients report major difficulties with sleep or rest (Bakker et al, 1995; Jamieson et al, 1995). Fatigue has been reported to correlate highly with increased pain, stiffness and functional disability (Jones et al, 1996b), and discomfort in bed, a frequent complaint of patients with AS, often leads to disturbed sleep which may result in complaints of tiredness and fatigue. This pattern may also be associated with adverse mental health (Walker et al, 1993). Effective analgesic control may facilitate improved sleep, but evidence suggests that better sleep is associated with increased stiffness on awakening (Jamieson et al, 1995).

There is no gold standard measure of symptomology or disease activity in AS, and the unique disease profile seen in many patients may complicate assessment (Goodacre et al, 1991; Dalyan et al, 1999). Therefore, a combination of measurements including pain, stiffness, articular and enthesitis indices, laboratory-based assessment, analgesic consumption and the presence of extra-articular features are traditionally adopted parameters. Patient reported change in pain and stiffness have been recorded in routine practice and clinical research for many years. Often measurement involves the representation of severity on single item visual analogue (VAS) or likert-type scales and the multi-dimensional nature of pain or stiffness is infrequently addressed (Dziedzic, 1998). Fatigue and sleep disturbance are less frequently recorded. However, recent developments in patient-based evaluation have produced the Bath AS Disease Activity Index (BASDAI), an AS-specific measure of disease activity containing items addressing pain, stiffness and fatigue (Garrett et al, 1994), and the Body Chart, a global representation of bodily pain in AS (Dziedzic, 1997). These instruments are considered further in Chapters 2 and 4 respectively.

Physical function

Functional disability is one of the most important complaints in AS following pain and stiffness (Dougadas et al, 1988), and refers to limitation in activities of daily living, mobility and self-care. The impact of pain, stiffness and altered biomechanics on axial and peripheral joints, and the limited mobility of advancing AS, may limit physical functioning with a resulting loss of independence in certain activities. However, many patients may underestimate and under-report functional difficulties (Hidding et al, 1992). This may reflect patients adjusting to functional difficulties over the years of often relentless disease progression, and thus failing to report difficulties. Patients may adopt unusual movements or gadgets to assist in the performance of activities (Abbott et al, 1994), and no longer consider the activity impossible or difficult to perform. Alternatively, patients may learn to accept the reality of the disease, and adjust their functional expectation accordingly. A survey of 129 AS out-patients reported physical difficulties with routine daily activities as a consequence of AS (53%), the majority indicating greater problems with general mobility (47%), as opposed to self-care (6%)(Bakker et al, 1995). A further survey of members of an American AS self-help group rated difficulty with physical functioning highly, and four areas were described (Nemes, 1991): firstly, limited neck mobility restricting activities such as sleeping prone, driving, reaching and hugging. Secondly, sexual function was affected. This may relate to the pain and/or stiffness and immobility associated with AS. Alternatively, a spouse or partner may avoid sexual contact for fear of causing pain. Pain associated with rest was a third factor, and was closely associated with sleep disturbance, tiredness or fatigue. Finally, axial dysfunction and difficulty with activities requiring degrees of spinal mobility was reported. For example, bending to put on socks, or getting into/out of the bath. Following a review of available evidence, Ward (1998) suggests that sexual functioning may be a 'substantial problem' for patients, especially females, with more than 30% of patients experiencing moderate or severe pain. In a subsequent study, mild sexual difficulties were reported in 25% of patients (n= 44 out-patients), leading the investigators to include items relating to sexual function in a modified AS-specific disability index (Dalyan et al, 1999).

Several risk factors for increased functional disability have been described, and include younger age at onset, neck, hip and/or gleno-humeral joint involvement, increased disease activity and depression (Ward, 1998). A survey of relatively young

AS patients (mean age 36 years), with severe disease (n= 17) (Brown et al, 1987) rated their top four most important problems as stiffness (82%), inability to do everyday tasks (82%), sexual problems (71%) and pain (65%). 50% of these patients had undergone hip replacement surgery, and 35% were unable to work due to the extensive AS-associated deformities and related fatigue and demotivation. However, accumulated evidence suggests that most patients remain functionally independent, despite often chronic discomfort (Calin, 1985; Dalyan et al, 1999).

There is no gold standard for the evaluation of functional disability in AS, and many investigators adopt patient-based questionnaires originally developed for patients with Rheumatoid Arthritis (RA). For example, the Health Assessment Questionnaire (HAQ) (Fries, 1980). However, these instruments focus on peripheral joint impairment and difficulty with prehensile activities. Therefore, item content has little relevance to patients with a predominantly axial disease. The functional assessment of AS has lagged behind that of other rheumatic diseases, such as RA, but with the increasing realisation that rheumatic disease has an important impact on both functional and psychosocial issues, instruments to evaluate these domains have been developed, and are evaluated in Chapter 2.

The axial and thoracic dysfunction of AS is a dominant feature of the disease. Although reflecting a very limited aspect of disease impact, measurement of the limitation in spinal mobility and chest expansion have dominated the evaluation of outcome in AS for many years. Proponents suggest that rigorous and regular anthropometric assessment is essential to describe the clinical outcome in AS and to support clinical decision making (Lubrano and Helliwell, 1999). Serial measurement may also provide an insight into the natural history of disease progression reflecting either structural, irreversible change in axial status (Kennedy et al, 1995) or reversible change in mobility (Roberts et al, 1988), and identify sub-groups of patients in relation to disease severity (Dawes, 1999; van der Heijde and Spoorenberg, 1999). Numerous anthropometric measurement techniques can be described, and are evaluated in Chapter 2.

Role function

Role performance describes the ability of a patient to continue with daily life-style obligations, such as employment and household chores (Jenkinson et al, 1998a). The

early onset of AS, striking at the prime of life, suggests that the impact on paid employment could be considerable. Up to one third of patients experience at least one prolonged period of sick leave from work (Worsdworth and Mowat, 1986), this rate increasing for patients involved in manually demanding jobs (Guillemin et al, 1990). Many patients report changes in employment to less physically demanding roles, but accumulated evidence suggests that 60-85% of patients with a disease duration of 14 years or more remain in paid employment (Ward, 1998; Dalyan et al, 1999).

Social interaction

There has been little specific research on the impact of AS on social function and interaction (Ward, 1998). When the impact of AS on daily routine problems was considered (n= 129), few reported difficulties with social function (Bakker et al, 1995): 6% indicated that leisure activities were limited, 3% were limited in role activities, 2% with communication, and only 1% reported difficulties with social interaction. However, 41% of patients with severe disease (mean age 36 years; n= 17) described depression, loneliness and boredom as important features (Brown et al, 1987). Dalyan et al (1999) indicated that few patients reported marital strain, or marriage avoidance due to AS.

Emotional well-being

A patient with AS may experience involvement of the 'whole system' in the disease process and attention to the locomotor system alone may detract from possible extra-articular manifestations and symptoms (Dziedzic, 1998; Reynolds et al, 1999). Almost 25% of patients reported emotional well-being as an important area of routine daily life affected by AS (n= 129)(Bakker et al, 1995). 20% of patients related this impact to their emotional health, and a further 4% demonstrated concern over their physical appearance. One third of patients with AS, and a significantly higher proportion of females than males, may experience clinical depression (Barlow et al 1993b). A strong association with pain, particularly in females, and a weaker association with functional disability, was indicated. A strong association between depression and poor physical function, social inadequacy and low self-esteem has also been reported (Barlow et al, 1992; Hidding et al, 1994b; Ward, 1998). The AS pathological process leading to the adoption of an altered posture, may have a multi-faceted impact on the well-being of a patient, being associated with pain and reduced function, thereby strengthening the possible association with depression. It has been

suggested that the rarity of psychological problems in patients with AS is associated with the ability of the patient to accommodate to the gradual physical and psychological demands of the disease (Dalyan et al, 1999). Alternatively, this may reflect the infrequency with which psychosocial aspects of AS have been addressed in research and routine practice. Patients with detected depressive problems may have a better outcome than patients where symptoms are not identified, and evaluation of this important symptom in AS is recommended (Barlow and Barefoot, 1996).

1.3.4 Management

AS is incurable, progressive and unpredictable in its progress. Therefore, long-term management centering around the control of pain and improvement of function is indicated, with the responsibility for daily management lying primarily with the individual patient (Barlow and Barefoot, 1996). This is a significant undertaking and requires notable physical and psychological adjustment (Barlow et al, 1993a).

There are two main facets to management in AS: drug therapy and physiotherapy. Drug therapy may involve three categories of therapy: 1) disease-controlling anti-rheumatic therapy (DC-ART), such as sulphasalazine, which influence the disease process; 2) symptom-modifying anti-rheumatic drugs (SMARD), such as non-steroidal anti-inflammatory drugs (NSAID), which suppress inflammation without influencing the disease process; and finally, 3) analgesics and muscle relaxants for pain relief. Drug management is decided on an individual basis, and many patients chose not to take medication due to the potential side effects. However, the benefits of pain-free movement and reduced stiffness, with the ensuing ability to continue with normal activities of daily life afforded by selective medication may outweigh any possible side effects (Dziedzic, 1998).

Physiotherapy, including daily exercise therapy and education is recognised as an essential part of any management programme in AS (Viitenan and Suni, 1995). The classical image of a patient with advanced AS, is that of a 'question-mark' posture (Hyde, 1980). That is, flattening of the lumbar spine with associated hip flexion, increased thoracic kyphosis, and protraction of the upper cervical spine to facilitate a forward looking gaze. It is suggested that the change of posture is initially adopted as a pain relieving response to inflammation of the spinal zygoapophyseal joints (Dziedzic, 1998), but subsequent ankylosis and soft tissue shortening may result in a

fixed posture, with its resulting adverse impact on functional ability and emotional well-being. Management aims to maintain or improve mobility, posture and general fitness, and thus enhance the quality and expectations of a young life. Education plays an important role in this process and patients should be empowered to incorporate such a routine into their everyday life. Although physical exercise has been demonstrated to improve functional outcome, patient education may have a greater impact on psychological well-being, for example, feelings of depression and patient self-efficacy with disease management requirements (Barlow and Barefoot, 1996). However, these issues are rarely evaluated in AS.

Various surgical procedures may be indicated in patents with severe AS, and range from total hip joint replacement to spinal wedge osteotomies or stabilisation to reduce pain and improve posture and functional outcome (Dziedzic, 1998).

1.4 Measurement of outcome in AS

Mortality is not such an important outcome measure in a chronic, incurable disease such as AS, where accumulated evidence suggests that life expectancy is not significantly reduced, but the impact of disease on HRQL is great. Patients and health professionals require more relevant information about the impact of disease and disease management than is provided by the traditional biomedical assessments frequently encountered in AS. However, little attention has been paid to the evaluation of HRQL in AS. A review of outcome measures applied in published studies of AS showed that physician assessed measures of impairment or disability prevailed in 79% of studies, the majority recording spinal mobility, pain at sites of entheses or joints, and laboratory based assessment (Bakker et al, 1993b). Some studies included a physician generated global assessment of patient health on a single visual analogue scale (VAS). 67% of studies included patient-based evaluation, but in 65% of cases this included only the assessment of pain or stiffness on single item VAS or Likert-type scales.

Measures of health outcome which adequately fulfil the measurement and practical properties deemed necessary for evaluative purposes (Kirshner and Guyatt, 1985; Fitzpatrick et al, 1998a) provide the evidence upon which evidence-based clinical decision-making in routine practice, research, medical audit and health policy is based (Ruta et al, 1998b; Bowker, 1998). However, despite the wide acceptance of many

anthropometric measures, and to a lesser extent patient-based instruments, in routine practice and research, few measures of outcome have adequate evidence of their measurement properties, acceptability or feasibility to support their adoption in the evaluation of AS (Laurent et al, 1991; Bellamy et al, 1991a). There is no standardisation of measurement practice in AS, a feature common with many other chronic disorders, and it remains unclear how patient-based and anthropometric measures relate to each other and how best to incorporate these instruments in evaluation. These are important and unresolved issues. There has been no systematic review of the wide range of available instruments in AS or an explicit appraisal of instrument development and application to describe the best available instruments. Also, no empirical evaluation of the comparative role of the many available patient-based and anthropometric measures has been described.

Awareness of the need to standardise and reduce the number of instruments frequently applied in the evaluation of AS has resulted in the recent recommendations by the Assessment in Ankylosing Spondylitis group (ASAS) (van der Heijde et al, 1997; 1999a,b,c)(table 2.65). Recommendations were based on expert opinion and followed the identification of several domains considered important in AS evaluation and the subsequent fulfillment of these domains by AS-specific measures of outcome. Recommendations are heavily biased towards the measurement of impairment and are considered further in Chapter 2.

1.5 The current research

The apparent knowledge gap in how best to evaluate outcome in AS provides a strong case for further methodological research, and is addressed in the current research. The following chapters describe three stages in this process.

Firstly, the entire range of patient-based and anthropometric measures of outcome applied in published studies of AS were identified and assessed as a reflection of current evaluative practice. Evidence for the acceptability, feasibility and measurement properties was systematically reviewed and appraised to produce the first explicit evaluation of all instruments. Any gaps in evaluation were identified. Historically the selection of instruments for inclusion in research or routine practice has been based upon 'usual' practice, historical precedence (Jenkinson et al, 1994a), or on expert opinion (van der Heijde et al, 1997), and consequently has resulted in a lack

of standardisation. This review has described the first evidence-based selection of patient-based and anthropometric measures of health outcome to reflect the different domains of HRQL considered important in the evaluation of AS. Secondly, the first individualised measure of AS-related quality of life, the Patient Generated Index for AS (PGI-AS) was developed and tested for the first time.

Thirdly, the study describes the first empirical comparison of the selected instruments, the PGI-AS and two generic measures of HRQL in a large population of AS patients. The study describes one of the largest clinic-based and multi-centre postal surveys of outcome measures in AS in the United Kingdon (UK). The detailed assessment of data quality, scaling assumptions and measurement properties for such a broad selection of measures of outcome has not previously been undertaken in AS. The concurrent evaluation supports instrument comparison and final recommendations are based on accumulated evidence of instrument acceptability, feasibility and measurement properties (McHorney and Tarlov, 1995).

The first standardised and evidence-based package of patient-based and anthropometric measures of outcome for application in AS clinical practice and research that fulfills the domains considered important in the evaluation of AS will be described. The study will also support the reduction in multiple measures of impairment and disability traditionally adopted in AS evaluation, and provide guidance for the role of disease-specific and generic measures of HRQL in AS alongside the more traditional measures of impairment. The role of individualised patient assessment of disease-related quality of life in AS will be introduced to the evaluation of AS for the first time.

Chapter 2 Systematic review of patient-based and anthropometric measures of outcome in AS

2.1 Introduction

This chapter presents a systematic review of patient-based and anthropometric measures of outcome applied in published studies of AS between 1990 and May 2000. The status of outcome measurement in AS is described in section 2.2 and the role of systematic reviews is discussed in section 2.3. The methods for performing a systematic review and explicit evaluation of outcome measures are described in section 2.4. The results of the literature review, identification of articles and measures of outcome are described in section 2.5. Section 2.6 describes the data evaluation and selection of instruments. The chapter closes with a discussion.

2.2 AS and measures of outcome

There is no current consensus on the best approach to take in the evaluation of a disease with such a wide clinical spectrum as AS (van der Heijde et al, 1997). Diverse issues from observable clinical manifestations to the impact on HRQL may be considered, and a multitude of instruments are often applied in published studies and in clinical practice (Laurent et al, 1991; Jenkinson et al, 1994a). However, AS evaluation has largely focussed on clinical measures of impairment, disease process and the presence of subjective symptoms of disease activity such as pain and stiffness (Laurent et al, 1991; Bakker et al, 1993b). A recent survey of routine practice revealed that two anthropometric measures from the AS diagnostic criteria (lumbar anterior flexion and chest expansion) were usually or always included in AS longitudinal evaluation by more than 70% of clinicians, but less than 50% included patient-based instruments (Bellamy et al, 1998; 1999). A lack of familiarity with patient-based instruments, logistic restraints and a lack of emphasis on formalised measurement of outcome were offered as suggestions for the described measurement practice (Bellamy et al, 1998;1999).

There is a need for the standardisation of approaches used in the evaluation of AS to foster comparison of results across studies and to reduce the unnecessary burden to both patient and clinician of completing a large number of potentially inappropriate instruments. The result will be a package of outcome measures that are of relevance to both patient and clinician. In recognition of the difficulties in selecting appropriate evaluative instruments the Assessment in Ankylosing Spondylitis (ASAS) group was

formed in 1995 as a sub-committee to the larger Outcome Measurement in Rheumatology (OMERACT) initiative (van der Heijde et al, 1997; 1999a,b,c). ASAS is an 'invited' international working group consisting of clinical experts in AS, clinical epidemiologists, patient representatives and delegates from the pharmaceutical industry.

The ASAS group first described three different settings in which therapy could occur: firstly, disease-controlling anti-rheumatic therapy (DC-ART); secondly, symptom modifying anti-rheumatic drugs (SMARD) and physical therapy; thirdly, clinical record keeping and clinical practice (van der Heijde et al, 1997). Following a Medline (index medicus on-line: 1986-1995) database search, further supported by a hand-search of the bibliographies of selected articles, 110 instruments applied in AS during this period were identified. This total included both single component measures and composite indices (van der Heijde et al, 1997).

Members chose those instruments they felt should be included in a core set for each setting, spending points to indicate the relative importance of each instrument. Instruments were subsequently ranked and the domains described to produce a core set of domains for each setting (table 2.1).

DC-ART	SMARD / Physical therapy	Clinical record keeping
	Common to all settings	
	Physical function	
	Pain	
	Spinal mobility	
	Spinal stiffness	
	Patient global assessment	
Peripheral joints / entheses		*Peripheral joints / entheses*
Acute phase reactants		*Acute phase reactants*
Spine radiograph		
Hip radiographs		
Fatigue		

Table 2.1 Preliminary core sets of domains for the evaluation of patients with AS identified by the ASAS working group (van der Heijde et al, 1997).
Domains in italic print are not definitely included, but are on the ASAS research agenda.

OMERACT and ASAS have proposed a filter to support instrument selection that relies upon evidence of truth, discrimination and feasibility (Bellamy, 1999). Truth

considers validity and discrimination the reliability and responsiveness of an instrument as a collective attribute. Feasibility addresses instrument brevity, simplicity and scoring when applied in routine practice or research. Subsequent to domain identification instrument selection was initially based upon evidence of feasibility and relevance, as determined by expert opinion and group consensus only (van der Heijde and van der Linden, 1998; van der Heijde et al, 1999a,b,c). This process did not appraise evidence to support instrument development or measurement properties. ASAS acknowledges that further appraisal of instrument measurement properties is required and indicate that any necessary amendments to recommendations following this process will be made (van der Heijde et al, 1999a,b,c).

The ASAS instrument selection fulfilling the core set of proposed domains was published subsequent to the initial systematic review described in section 2.4. The selection will be discussed in relation to the instruments selected as a result of the systematic review in section 2.7.

2.3 Systematic reviews of scientific evidence

A systematic review has been defined as:

> 'the process of systematically locating, appraising and synthesising evidence from scientific studies in order to obtain a reliable overview.'
>
> (Centre for Reviews and Dissemination (CRD) Report 4,1996)

Available literature on the methodology for performing systematic reviews concentrates predominantly on the critical appraisal of randomised controlled trials (RCTs)(CRD Report 4, 1996). This study differs from standard reviews because the majority of the studies included in the evaluation are not RCTs, and the main focus of the study is to determine the quality of patient-based and anthropometric measures of outcome.

Several structured surveys and more practical approaches to evaluating the developmental and measurement properties of patient-based measures of outcome to support making recommendations for the adoption of certain instruments have been described. For example, Beurskens et al (1995) critically appraised the evidence in

support of four widely used functional measures of low back pain, and McDowell and Newell (1996) and Bowling (1996) have produced texts which describe and appraise various outcome measures and rating scales. Also, standards for instrument development, testing and appraisal have been described by several authors (Streiner and Norman, 1995; McDowell and Jenkinson, 1996; Fitzpatrick et al, 1998a).

The review of outcome measures described by ASAS, although extensive, was not systematic. The review design and inclusion criteria by which instruments were identified were not defined, the search was not exhaustive and article retrieval not systematic. Assessment and instrument selection was dictated by expert opinion and group consensus and not by an explicit appraisal of available evidence. Also, although experts offer extensive knowledge in relation to AS, 'content experts' may lack the objectivity desirable in critical appraisal (Oxman, 1995).

The first article to describe a systematic literature review of two specific patient-based outcome measures was published subsequent to the initial review (section 2.4) (Ruof and Stucki, 1999a). The investigators compared the properties and performance of the Dougadas Functional Index (DFI)(Dougadas et al, 1988) and the Bath Ankylosing Spondylitis Functional Index (BASFI)(Calin et al, 1994). Article retrieval followed a database search of Medline (index medicus online), scanning reference lists and contact with the instrument developers. Evidence was appraised in accordance with the OMERACT filter (Boers et al, 1999) modified to consider instrument development. Contact with the instrument developers was made to improve identification of all relevant published data. However, permission to use either instrument is not required and contact with the developers assumes an awareness of all studies applying each instrument. A more exhaustive search of the major electronic databases and hand-searching of relevant journals would improve the systematic nature of the search and improve study replication (Jadad et al, 1998).

When designing a systematic review structured, thorough and replicable methods of data collection are essential to ensure the identification and retrieval of all, or nearly all, relevant studies (Dickersin et al, 1995; Jadad et al, 1998). Data collection should clarify methods adopted to identify data, study inclusion criteria and a structured data extraction to support analysis. The adoption of explicit methodology improves the

validity of results, allows study replication and supports appreciation of why results and conclusions of similar reviews may differ (Mulrow, 1995).

The review described in this chapter has adapted the guidelines for performing a systematic review of RCTs proposed by the CRD in York (CRD Report 4, 1996) and available evidence to support the requirements of evaluative instruments to develop a systematic review and explicit evaluation outcome measures. Due to the large number of outcome measures included in AS evaluation and the limited study resources, the review has focussed on the identification of patient-based and anthropometric measures of outcome and available publications to support the development and testing of these instruments.

Patient-based measures reflect a relatively new approach to the evaluation of patient outcome reflected by the growth in availability of instruments in rheumatology and other specialities (McDowell and Newell, 1996; Bowling, 1996). Broadly defined, patient-based measures of outcome record a patients perspective about various domains of health, illness and the effects of health care (Fitzpatrick et al, 1998a; Greenhalgh and Meadows, 1999).

Anthropometric measurement has clinical relevance to both AS research and clinical practice (Lubrano et al, 1998). Despite a lack of standardisation and paucity of data supporting the measurement properties (Laurent et al, 1991; Bakker et al, 1993b), clinicians routinely include these measures in the longitudinal evaluation of patients (Lubrano et al, 1998; Bellamy et al, 1998, 1999).

Although playing an important role in diagnosis and as a long-term end-point in assessment, radiographic evaluation is not well established as an evaluative procedure in AS (Dawes, 1999; van der Heijde and Spoorenberg, 1999) and has been excluded from the review. Laboratory based measures are considered 'unhelpful' in the evaluation of AS (Calin, 1995a) and evidence of a relationship between laboratory based measures of disease activity and axial disease is weak (Ruof and Stucki, 1999b; Dawes, 1999). These measures have also been excluded.

A 'triple research question' (CRD Report 4, 1996), modified to suit to focus of the review, was proposed:

1) *Instruments:* What patient-based and anthropometric measures of outcome are used in the evaluation of patients with the adult expression of AS?
2) *Measurement properties:* What is the available published evidence supporting the development, testing and application of these instruments?
3) *Success of instrument:* Do these instruments successfully fulfil the necessary attributes required of an evaluative instrument to support recommendation for use in clinical practice and research? (Fitzpatrick et al, 1998a).

The primary aim of the review is two-fold: first, to provide the first comprehensive report of the entire range of patient-based and anthropometric measures of outcome applied in published studies of patients with AS between 1990 - 2000; and secondly, to synthesise the evidence base in support of the development and subsequent testing of identified instruments.

A secondary aim is to make recommendations in support of the first evidence-based selection of instruments for use in research and clinical practice which fulfil the domains considered important in the evaluation of AS. These instruments will be adopted in a comparative study (Chapter 4) where measurement and practical properties will be further evaluated before a final recommendation is made.

2.4 Methodology for a systematic review and evaluation of patient-based and anthropometric measures of outcome in AS

A systematic review and evaluation of patient-based and anthropometric measures of outcome requires selection criteria for the identification of articles and measures of outcome. Search strategies for a systematic and exhaustive literature search are described. Data extraction from selected articles and the explicit evaluation of instruments then follows a systematic format. A grading scheme to provide a quality assessment and quantiative summary of the evidence for instrument measurement properties is also described.

The initial literature search covered the years 1990 - April 1998. The start date of 1990 was chosen because very little work on patient-based measures of outcome had

taken place in AS before this time. Although anthropometric measures had been in use prior to 1990, the review was to be a reflection of current practice, therefore representing those approaches accepted into routine practice or research. Little evidence for the measurement properties of anthropometric measurement existed before this time (Laurent et al, 1991; Bellamy et al, 1991a). The search was subsequently extended to May 2000 to identify further published evidence and progress in the field of outcome measurement in AS. This will inform the discussion in subsequent chapters following instrument selection for the empirical evaluation.

2.4.1 Inclusion criteria

All articles and measures of outcome were required to satisfy certain criteria of relevance to the study question, patient population, type of outcome and language. Article inclusion criteria is shown in table 2.2.

Articles			
Inclusion		Exclusion	
i.	Published articles (1990-2000) focussing on evaluation of adult form of AS and containing identifiable patient-based or anthropometric measures of outcome	i.	Not specific to evaluation of adult form of AS
		ii.	Non-English language
		iii.	Development, testing or use of laboratory, radiographic or imaging techniques
ii.	Published articles referring to development / testing of patient-based or anthropometric measures of outcome applied in studies of AS between 1990-2000	iv.	Do not describe instruments in sufficient detail to allow identification
		v.	Non-published data
		vi.	Narrative reviews

Table 2.2 Article inclusion and exclusion criteria.

A measure of outcome was selected if it was patient-based or anthropometric and applied in the evaluation of adults with AS between 1990-2000, as shown in table 2.3.

Measures of Outcome			
Inclusion		Exclusion	
i.	Applied in evaluation of adult AS patients (1990 – May 2000): - Published (Anglicised) patient-based measures of outcome - Examination-based anthropometric measures (spinal and thoracic)	i.	Only used in relation to other conditions, or in childhood forms of AS
		ii.	Anthropometric peripheral joint assessment
		iii.	Laboratory, radiographic and imaging techniques
		iv.	Single item measures
		v.	Instruments not clearly identified in published text
		vi.	Non-Anglicised instruments
		vii.	Instruments for which only stage III information can be obtained [a]

Table 2.3 Patient-based and anthropometric measures of outcome inclusion and exclusion criteria. [a] stage III refers to article type in which instrument was identified (defined in section 2.4.2)

AS predominantly affects the axial skeleton (van der Heijde and van der Linden, 1998). Although peripheral joints, in particular the large axial joints, may be involved in the disease process, this is less common and with a tendency to be episodic and therefore difficult to assess (Kidd et al, 1988). Assessment of peripheral joints has therefore been excluded and anthropometric evaluation restricted to the assessment of spinal and thoracic mobility.

A preliminary review of selected articles identified a wide range of single item measures adopted for the evaluation of features ranging from pain or stiffness to global health. Given the large variation in the number of items identified and the frequent lack of methodological detail these instruments were excluded from the review.

Although limiting the extent and generalisability of the review, due to limited resources only English language articles and Anglicised instruments were included. Where instruments have been translated into other languages this has been referred to. Where articles describe the application of instruments that have not been translated into English these articles and instruments have been excluded from the review.

Where communication with experts identified the development of new instruments this has been acknowledged. However, resource constraints meant that only published instruments and published evidence in support of development and testing have been included.

2.4. 2 Review search strategies

The strategy adopted for obtaining articles and the identification of instruments is outlined in figure 2.1. Three stages can be described: 1) identifying articles; 2) identifying measures of outcome; 3) assessment of articles and instrument relevance.

1) Identifying articles

The strategy adopted in developing the literature search for articles used the following four steps: i) developing search terms; ii) electronic database searches; iii) hand searching key journals; and iv) scanning reference lists (Jadad et al, 1998).

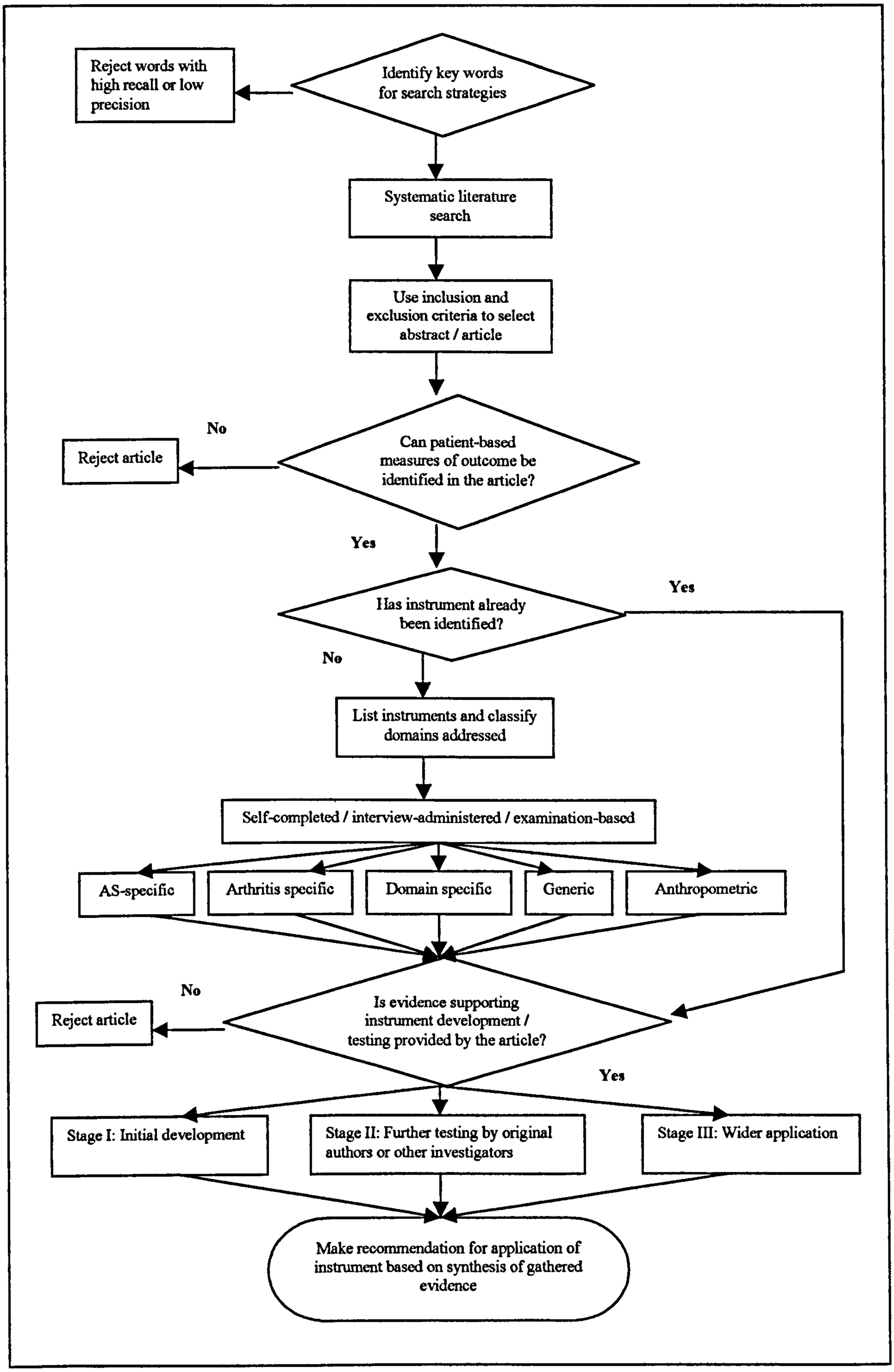


Figure 2.1 Strategy for obtaining relevant articles, identification and evaluation of measures of outcome

i. Developing search terms

Contacts with experts in rheumatology and outcome measurement

Consultation was made in order to gain a wide-ranging perspective of current issues in the measurement of outcome and AS and to call upon expert knowledge of relevant literature and ongoing research. Individuals were identified due to exceptional publications in AS and from identification in the British Society for Rheumatology Handbook (1995-1996)(table 2.4). Information regarding the use of and knowledge of the development and testing of patient-based measures of outcome, evaluation guidelines and recommendations for monitoring in AS was requested.

Health professionals	United Kingdom	Health professionals	International
Dr. Julie Barlow	Coventry University	Dr. Maarten Boers	Netherlands
Dr. Robin Butler	Shropshire	Professor Peter Brooks	Australia
Dr. Andre Calin	Bath	Dr. Maxime Dougadas	France
Dr. Martin Davis	Cornwall	Dr. Desiree van der Heijde	Netherlands (Chair of ASAS)
Dr. Chris Eastmond	Scotland	Dr. M.A. Khan	USA
Professor Ian Haslock	Cleveland	Professor S. Van der Linden	Netherlands
Dr. Philip Helliwell	Leeds		
Dr. Nigel Hurst	Edinburgh	*Associations*	
Dr. Dereck Jacoby	South Devon	American College of Rheumatology (ACR)	America
Dr. Andrew Keat	London		
Dr. Gabrielle Kingsley	London	Arthritis Rheumatism Council (ARC)	England
Dr. David Perry	London		
Professor Roger Sturrock	Scotland	National Ankylosing Spondylitis Society	England
Dr. Paul Wordsworth	Oxford		

Table 2.4 Contact with health professionals and rheumatology associations

Retrospective searching

Articles and text-books were identified that provide overviews of the development, evaluation and application of patient-based measures of outcome. Articles providing a more general overview of the application of patient-based and anthropometric measures of outcome in AS, rheumatology and other disorders were also identified. Searching these items provided references to relevant publications and provided an indication of the state of outcome measurement within AS and rheumatology.

These two steps provided a basis from which to develop search terms to be used in electronic database searches. Key words that regularly appeared in the title or text of publications already identified and associated with outcome measurement and / or AS were used to develop the search strategies.

ii. Electronic database searches

The selection of instruments for inclusion in the review was based on a search of the literature covering the years 1990-April 1998. The extension of the search to May 2000 made little difference to the availability of evidence in support of instrument selection. Therefore, the combined search results (1990-May 2000) are presented.

A comprehensive search of the literature used the most important electronic databases: Medline (index medicus on line), EMBASE (excerpta medica online), CINAHL (cumulative index of nursing and allied health online), PsycLIT (psychological abstracts online), AMED (allied and alternative medicine online), Cochrane Library (Cochrane database of systematic reviews (CDSR), Database of abstracts of reviews of effectiveness (DARE), Cochrane Controlled Trials Register (CCTR)), Centre for Reviews and Dissemination Reviews Database (CRD)) and ASSIA (applied social sciences online).

The comprehensiveness of an electronic database search depends on the search strategy adopted (Brettle et al, 1998). Test searches were run on each database to establish the recall and precision of various search terms. Terms that failed to retrieve any articles or retrieved a large number of inappropriate articles were omitted from the final search strategy. However, the heterogeneity of terms used in the field of outcome measurement has been commented on by other authors (Fitzpatrick et al, 1998a) and an extensive list of search terms was required to make the search as sensitive as possible. All searches specified 'Ankylosing Spondylitis', 'Spondylitis Ankylosing', or an alternative (Bechterew's or Marie-Strumpell Disease) as the main focus. All additional terms were required to be associated with these key terms. Individual terms varied for each database searched. The Medline (Silverplatter software) search strategy is shown in figure 2.2. This was modified to suit the specific requirements of the other databases.

All abstracts were searched and articles retrieved following the application of the inclusion criteria (Cook et al, 1997). It was not always possible to be certain of the relevance of the article based purely upon the search results. Articles of definite and possible relevance were retrieved in full and the inclusion criteria re-applied.

#1	(Spondylitis-Ankylosing).sh.
#2	Ankylosing Spondylitis.tw.
#3	#1 or #2
#4	Bechterews Disease
#5	#3 or #4
#6	Health status.tw.
#7	(Health-status).sh.
#8	(Health-status-indicators).sh.
#9	(outcome-and-process-assessment-(health care)).sh.
#10	(outcome-assessment-(health care)).sh.
#11	(process-assessment-(health care)).sh.
#12	outcome$ or proces$.tw.
#13	outcome measure$.tw.
#14	(treatment-outcome).sh.
#15	assess or assessing or assessment$.tw.
#16	(quality-of-life).sh.
#17	health-related quality of life or health related quality of life.tw.
#18	(severity-of-illness-index).sh.
#19	severity of disease.tw.
#20	disease activity.tw.
#21	index or indice$.tw.
#22	(self-assessment-(psychology)).sh.
#23	self-assess or self assess or self-assessment or self assessment.tw.
#24	end point$ or end-point$ or endpoint.tw.
#25	measure$ or measuring or measurement$.tw.
#26	(physical-examination).sh.
#27	(range-of-motion-articular).sh.
#28	metrology.tw.
#29	function$ or functional.tw.
#30	functional impairment or functionally impaired.tw.
#31	functional disabilit$ or functionally disabled.tw.
#32	functional abilit$ or functionally able.tw.
#33	functional activit$.tw.
#34	(activities-of-daily-living).sh.
#35	physical therapy.tw.
#36	health status.tw.
#37	mental status.tw.
#38	handicap.tw.
#39	(disability-evaluation).sh.
#40	disabilit$.tw.
#41	(evaluation-studies).sh.
#42	(clinical-trials).sh.
#43	(pain-measurement).sh.
#44	(questionnaires).sh.
#45	flexion or extension or rotation.tw.
#46	chest expansion.tw.
#47	physical mobility.tw.
#48	score$ or scoring.tw.
#49	monitor$ or monitoring.tw.
#50	(reproducibility-of-results).sh.
#51	reliable or reliability.tw.
#52	valid or validity or validate.tw.
#53	responsive or responsiveness.tw.
#54	6 or 7 o r8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51 or 52 or 53
#55	#5 and #54
#56	PY > 1989
#57	LA = English

Figure 2.2 Medline (index medicus online) database search strategy (1990-2000)

Notation: sh. - MeSh heading (medical subject heading); tw. - word in any of the text; $ - truncation symbol.

iii. Hand searching key journals

It is possible that articles are indexed inaccurately in the online sources or are overlooked in the search process (CRD Report 4,1996). Hand searching may also identify articles yet to be registered with online sources (Jadad et al, 1998). The following journals were considered most relevant to the topic and were hand

searched: Rheumatology (formerly the British Journal of Rheumatology), Journal of Rheumatology, Current Opinion in Rheumatology, Annals of Rheumatic Diseases, Arthritis and Rheumatism, Arthritis Care and Research, Seminars in Rheumatology, Clinical and Experimental Rheumatology, Spine, Quality of Life Research.

iv. Scanning reference lists

Scanning the reference lists of retrieved articles identified further articles for consideration. Specifically, this search supported the acquisition of articles published pre-1990 that may report the development and testing of instruments.

2. Identifying measures of outcome

Following application of the inclusion criteria all patient-based and anthropometric measures of outcome listed in the article were identified. If instruments fulfilling the inclusion criteria could not be identified the article was rejected. Instruments were then listed according to their completion format. That is, patient self-completed, interview-administered or examination-based / anthropometric. Finally, the disease specificity or domain addressed was indicated. Instruments may be broadly classified as either specific or generic (Guyatt et al, 1993). Further divisions within this simple dichotomy provide a more detailed analysis of the potential instrument properties. Within this review instruments were classified as: AS-specific, arthritis-specific, domain-specific, generic or anthropometric.

Where an inadequate description of the instrument, the methodology adopted or an inadequate reference to support the approach was found the instrument was excluded.

3. Assessment of relevance

The final stage in the search strategy was to assess the identified articles and measures of outcome for their relevance to the research question (figure 2.1).

Application of the review inclusion criteria ensured that patient-based and anthropometric measures of outcome applied in AS evaluation were identified. Evidence describing the development and / or subsequent testing of instruments was required from the published articles. Articles could be classified into three different stages of instrument development reflecting the original development, testing and subsequent application of an instrument (table 2.5)(figure 2.1).

Stage of development	Properties of article
I	Original development and testing
II	Further development and testing Performed by original authors or other investigators. Studies have a specific aim to evaluate the instruments measurement properties
III	Wider testing and use of the measure in AS-specific clinical practice, research, and audit, etc.

Table 2.5 Stage of instrument development described by article

A hierarchy of instrument development has been described to support the degree to which evidence from the review could be incorporated in the data evaluation (section 2.4.4). Studies from the developers of each instrument were described as Stage I articles and were sought for each instrument. These articles should describe instrument purpose and conceptual base, clarify the intended population and provide evidence in support of the development and initial testing. Stage II articles provide further evidence of instrument application and build upon evidence in support of the measurement properties. These investigations may be performed by the original developers or by subsequent investigators. In stage III, the wider application of the instrument in clinical research and routine practice is sought. These articles may provide further evidence in support of instrument reliability, validity and responsiveness. Evidence of the generalisability of the results in terms of AS, and the feasibility and acceptability of instruments beyond application by the original developers is sought. Selected articles were listed under each registered instrument and the stage of instrument development reflected by an article indicated. Where published evidence in support of stages I and II could not be identified instruments were excluded from the review.

2.4.3 Data extraction

Using criteria considered important in the evaluation of patient-based measures of outcome (Streiner and Norman, 1995; McDowell and Newell, 1996) a data extraction sheet was developed to retrieve data reflecting the development, testing and evaluation of instruments from selected articles (Appendix 1). Tabulated evidence in support of the identified instruments was created that follows the structure of the data extraction sheet.

Data extraction was performed by the lead investigator only (KLH), and all articles and data-extraction sheets were double-checked by the same investigator for consistency and accuracy of content. Appraisal by a second reviewer to check for human error and reduce any potential of bias in data extraction was not possible within the given resources.

2.4.4 Data evaluation

The data evaluation of all selected instruments was performed by the lead investigator (KLH) and is based on published evidence from retrieved articles (Cook et al, 1997). Using widely cited criteria adapted from several publications a structured and explicit qualitative framework was applied (Streiner and Norman, 1995; Beurskens et al,1995; McDowell and Newell, 1996). The framework allows for appraisal of each instrument in terms of the areas summarised in table 2.6.

Review criteria	Key questions
Title	Title provided by the original author and any subsequent revisions
Author	Lead author in original development
Year	Year of first publication and the year of any major revision
Purpose	What is the purpose of the instrument? What does it aim to evaluate?
Conceptual base	Do the authors provide a conceptual definition of what they aim to measure? What is the rational behind the design and development? What is the focus and relevance of the instrument? Is there justification for the items included? Does the rational relate to the 'purpose'?
Population	On which populations has the instrument been developed and tested?
General description, item development and scale structure	What method of item development is used? Are all participants involved in this process identified? What domain(s) are covered by the items? What method of item scaling is used and why? Time specificity? Performance or capacity based? Are developers active in developing / modifying the instrument?
Measurement properties	Published evidence of - reliability, validity and responsiveness; population investigated and study design (developers and subsequent investigators)
Acceptability	Is the instrument acceptable to patients?
Feasibility and application	Is the instrument practical and acceptable to all users? Has respondent/clinician burden been considered in administration and scoring? Is there a 'users manual'? Has the instrument been applied by investigators other than the original developers?
Commentary	Synthesis of evidence

Table 2.6 Summary of data evaluation.

Following the guidelines proposed by McDowell and Newell (1996), the appraisal of each instrument was based upon published information together with an accompanying users manual if available. As a minimum, published information relating to an instruments development should consider the purpose and conceptual base, development and subsequent testing of measurement properties, the standardisation of application and scoring procedures, and identify the definitive version (McDowell and Newell, 1996).

In addition to the retrieval of descriptive data and statistical results, a grading scheme to provide a quantitative summary of the quality of evidence supporting the reliability, validity and responsiveness of identified instruments has been developed based on previous work (McDowell and Newell, 1996)(table 2.7).

Thoroughness of testing		Results of testing	
0	No reported evidence	0	No numerical results reported
+	Basic information only	+	Weak evidence only
++	Several types of test, or several studies reporting evidence	++	Moderate levels of evidence
+++	All major forms of validity / reliability / responsiveness reported. Several good quality trials reporting evidence	+++	Strong evidence to support

Table 2.7 Grading scale summary of the reliability, validity and responsiveness of identified instruments – (adapted from McDowell and Newell, 1996)

Test-retest reliability describes the stability of scores over time and is most often assessed by the calculation of the correlation coefficient. Internal consistency reliability may be assessed by calculating Cronbach's alpha coefficient which evaluates the homogeneity of items in a multi-item instrument based on classical test construction theory. Validity examines if the instrument measures what it purports to measure, and both qualitative (face and content) and quantitative (construct) assessments of validity are described. Responsiveness describes the ability of the instrument to detect clinically important change over time. The concepts of reliability, validity and responsiveness are addressed in detail in Chapters 5, 6 and 7 respectively. The quality of evidence considers both the thoroughness and the results of testing the measurement properties considered important for evaluative instruments (Kirshner and Guyatt, 1985). The grading scheme summarises evidence in terms of

ordered categories and consists of a four-point scale: a score of '0' indicates 'no support' for the underlying criteria, whereas a score of '+++' indicates a strong level of published evidence (table 2.7).

2.5 Results of review

The results of the review will be considered in three stages: 1) identification of articles; 2) identification of measures of outcome; 3) the data evaluation (section 2.6).

2.5.1 Identification of articles

For each electronic database searched a total number of abstracts were identified, as shown in table 2.8.

Electronic database	No. of abstracts	No. of abstracts / articles reviewed	No. of articles included	No. of articles (additional to Medline)
Medline	499	129	80	80
EMBASE	589	123	83	21
CINAHL	135	20	16	3
AMED	45	24	21	0
ASSIA (until 1997)	2	2	2	0
Cochrane - CTR	210	65	32	0
Cochrane - DARE	3	1	0	0
Cochrane - SR	9	9	0	0
PsychLIT	15	7	3	0
Database total	-	-	-	104
Handsearching and citation searching	-	-	-	54
TOTAL	-	-	-	158

Table 2.8 Results of systematic literature review (1990-2000)

Following application of the inclusion criteria the number of articles selected for the review was 104. The hand and citation search produced an additional 54 articles, making a total of 158 articles.

2.5.2 Identification of measures of outcome

A total of 33 self-completed and one interview-administered patient-based measures of outcome (table 2.9), and 46 examination based / anthropometric measures (table 2.10) were identified.

The patient-completed instruments included six generic measures of HRQL, six AS-specific measures of functional disability (including the interview-administered

MACTAS/PET), two AS-specific measures of disease activity, and a further two AS-specific measures of global health or HRQL (BAS-G; AS-AIMS2)(table 2.9).

The Bath AS-Global Score (BAS-G) quantifies the impact of AS on well-being (Jones et al, 1996a). It consists of two 10cm horizontal VAS: one records the impact of AS over the last week and the second the impact over the previous six months. However, the results of the VAS are not combined and are not routinely applied together. The BAS-G should be considered a single item measure and was excluded from the data evaluation. Only the AS-AIMS2 (Guillemin et al, 1999) was published subsequent to the initial review and was not available for consideration for inclusion in the comparative study (Chapter 4).

Category	Original reference	Instrument
Generic		
Health profile	Bergner et al (1976)	Sickness Impact Profile.
	Hunt et al (1989)	Nottingham Health Profile.
	Ware (1997)	Short Form 36-item Health Survey.
Utility measure	Torrence (1976)	Standard Gamble.
	Torrence (1976)	Rating Scale.
	Bennett et al (1991)	McMaster Utility Measurement Questionnaire.
AS-specific		
Functional disability	Nemeth et al (1987)	Ankylosing Spondylitis Assessment Scale.
	Dougadas et al (1988)	Dougadas / Spondylitis Functional Index.
	Daltroy et al (1990)	Health Assessment Questionnaire - Spondyloarthropathics
	Abbott et al (1994)	Leeds (Revised Leeds) Disability Questionnaire.
	Calin et al (1994)	Bath AS Functional Index.
	Bakker et al (1995)	McMaster Toronto AS Patient Preference Disability Questionnaire / Patient Elicitation Technique *
Disease activity	Kennedy et al (1993)	Bath Disease Activity Index.
	Garrett et al (1994)	Bath Ankylosing Spondylitis Disease Activity Index.
AS HRQL / global health	Jones et al (1996a)	Bath Ankylosing Spondylitis Global Score.
	Guillemin et al (1999)	AS Arthritis Impact Measurement Scales 2.
Arthritis-specific		
Functional disability	Fries et al (1980)	Health Assessment Questionnaire.
	Meenan et al (1980)	Arthritis Impact Measurement Questionnaire.
	Helewa et al (1982)	Toronto Activities of Daily Living Questionnaire.
	Kirwan & Reeback (1983)	Modified - HAQ.
Self-efficacy	Nicassio et al (1985)	Arthritis Helplessness Index.
	Barlow et al (1996)	Generalised Self-Efficacy Scale.
Sexual functioning	Blake et al (1987)	Sexual Activity and Satisfaction Questionnaire.
Domain specific		
Perceived fitness	Astrand & Rodahl (1977)	Astrand Fitness Index.
	Borg (1978)	Borg Scale - subjective effort / physical performance
Social function	Carlson & Levy (1968)	Carlson Adjective Checklist – Social Personal Orientation
	Mari et al (1985)	Self-Report Questionnaire 20.
	Funch (1986)	Social Support Scale.
	De Witte (1991)	Self-Assessed Function Questionnaire.
Pain	Melzack R (1975)	McGill Pain Questionnaire.
	Melzack R (1975)	Pain Rating Index (Rank).
Depression	Radloff DP (1977)	Centre for Epidemiological Studies Depression Scale.
	Wallston et al (1978)	Multi-dimensional Health Locus of Control.
	de Jong-Gierveld & Kampohuis (1985)	Loneliness Scale.

Table 2.9 Self-completed patient-based measures of outcome. * requires interview completion

Category	Original reference	Instrument
AS-specific		
Disease-activity	Mander et al (1987)	Newcastle Enthesitis Index.
	Dawes et al (1987)	Stoke Enthesitis Index.
	Dougadas et al (1988)	Dougadas / Spondylitis Articular Index.
	Creemers et al (1996)	Ankylosing Spondylitis Disease Activity Scale.
Anthropometric	Jenkinson et al (1994a)	Bath Ankylosing Spondylitis Metrology Index.
Arthritis-specific		
Disease-activity	Steinbrocker et al (1949)	Steinbrocker Functional Criteria (for RA).
	Ritchie et al (1968)	Ritchie Articular Index.
Generic anthropometric		
Cervical rotation	Cheshire (1957)	Cervical rotation - 'Large protractor'.
	AAOS (1965)	Universal goniometer.
	O'Driscoll (1978)	Simple or Spirit inclinometer.
	Viitanen (1992)	Myrin inclinometer.
	Viitenan (1998)	Tape measure.
Cervical flexion / extension	AAOS (1965)	Universal goniometer.
	O'Driscoll (1978)	Simple inclinometer.
	Calcraft (1974)	Tape measure: occiput to C7; Chin to chest.
Cervical lateral flexion	AAOS (1965)	Universal goniometer.
	AAOS (1965)	Tape measure: tragus to acromioclavicular joint.
	O'Driscoll (1978)	Simple inclinometer.
Chest expansion	Hart et al (1963)	Nipple level.
	Moll & Wright (1972)	4th intercostal space; hands on head.
	Tomlinson (1986)	Xiphisternum; hands on head / by side.
Thoracolumbar flexion	AAOS (1965)	C7 - iliac crest.
	Calcraft (1974)	C7 - 10cm proximal to L5/S1 junction.
	Hyde (1980)	C7 - sacrococcygeal point.
	Armstrong (1984)	C7 - posterior superior iliac spines.
	Viitanen (1992)	C7 - S1.
	Averns (1996)	C7 - L5.
Thoracic rotation	Viitanen (1993)	Thoracolumbar rotation frame.
	Viitanen (1999)	Pavlaka method.
Fingertip to floor distance (anterior flexion)	Miller (1984)	Tape measure; patient stands on floor.
	Tomlinson (1986)	Vertical mounted ruler; patient stands on floor.
	Kippers & Parker (1987)	Ruler / tape; patient stands on raised stool.
	Stokes (1988)	'Portable spinal mobility scale' (PSMS).
Lumbar flexion	Von Schober (1937)	Schober '10cm' index.
	Macrae & Wright (1969)	Modified Schober Index (15cm) (lumbar flexion).
	Adrichem & van der Korst (1973)	Lumbar Flexion Index.
Lumbar extension	Dunham (1949)	Dunham Spondylometer (flexion / extension).
	Moll (1972b)	'Plumb-line' extension.
	Miller (1984)	Smythe technique (flexion / extension).
Lumbar lateral flexion	Moll (1972a)	Skin distraction technique (ipsi/contralateral).
	Domjan (1990)	Fingertip markings, lateral thigh (ipsilateral).
	Pile (1991)	Fingertip to floor - vertical mounted ruler (ipsilateral).
	Little (1986)	Fingertip to fibula (ipsilateral).
Spinal posture	Tomlison (1986)	Tragus to wall distance - 't'-square.
	Stokes (1988)	Occiput to wall distance - PSMS.
	ARA (1984)	Occiput to wall distance - tape measure.

Table 2.10 Anthropometric / examination-based instruments

Footnote: AAOS-American Academy of Orthopaedic Surgeons, ARA- American Rheumatology Association.

The examination based measures include four AS-specific measures of disease activity and one AS-specific battery of anthropometric measures (Bath Ankylosing Spondylitis Metrology Index, BASMI)(Jenkinson et al, 1994a). However, the Ankylosing Spondylitis Disease Activity Scale (AS-DAS)(Creemers et al, 1996) represents a battery of several AS-specific instruments including laboratory-based measures and has been excluded from the evaluation. A total of 39 generic anthropometric measures were identified and are listed under the evaluation of cervical, thoracic or lumbar mobility, or spinal posture (table 2.10).

2.6 Results of data evaluation

The data evaluation has been restricted to the AS-specific measures of: 1) functional disability, 2) disease activity and 3) all anthropometric measures. The domains were selected in light of the core domains identified by the ASAS working group (table 2.1).

2.6.1 AS-specific functional disability

This section presents the evaluation of the six AS-specific measures of functional disability (table 2.9). A general description and the scale structure is summarised in table 2.11, and table 2.12 details the item content.

Purpose and conceptual base

The Ankylosing Spondylitis Assessment Questionnaire (ASAQ)(Nemeth et al, 1987) is proposed as a simple self-administered measure of spinal mobility or disability in AS. The purpose is not defined further and the conceptual base is not described.

The Dougadas Functional Index (DFI)(Dougadas et al, 1988) is the first AS-specific measure of functional disability. It is an evaluative instrument defining functional ability as an appreciation of how a patient functions within their own environment. The initial choice of items was based on the expert opinion of three rheumatologists familiar with AS and not in relation to any stated theory of functional disability.

The Health Assessment Questionnaire for the Spondyloarthropathies (HAQ-S) (Daltroy et al, 1990) describes difficulties with functional activities of daily life as a result of an inflammatory spondyloarthropathy. The foundation for the instrument is the Health Assessment Questionnaire (HAQ)(Fries, 1980) which focuses on the

	ASAQ [1]	DFI [2]	HAQ-S [3]	LDQ (RLDQ) [4]	BASFI [5]	MACTAS/PET [6]
Number of items	2	20	25 (13) disability index 2 pain and discomfort index	19 (16)	10	0-15
Response options	8-point scale with adjectival anchors: 0 'no difficulty' to 8 'very severe difficulty' Score 0-11 0 = best health 11 = worst health	3-point adjectival scale 0= yes, with no difficulty 1= yes, but with difficulty 2= no Score 0-40 0 = best health 40 = worst health	Functional disability: 4-point adjectival scale 0= without difficulty 1= some difficulty 2= much difficulty 3= unable to do Pain and Discomfort: 2 x VAS (10cm) Score 0 – 3 0 = best health 3 = worst health	4-point adjectival scale 0= able to do 1= able to do with difficulty 2= use unusual movements or gadgets 3= unable to do RLDQ provides more guidance for response '2' than the LDQ Score 0-3 LDQ (0-48 RLDQ) 0= best health 3 (48) = worst health	VAS with endpoints 'easy' and 'impossible' Mean score of ten VAS scales Score 0-10 0 = best health 10 = worst health	Likert scales: numerical (0–7) and adjectival (4 descriptors) a. 3 groups (difficulty/ severity/ frequency). Score each problem (0-15) in each group b. Importance of each problem scored on Likert scale 0–7 'least' to 'most important' Score 0-49 0 = best health 49 = worst health
Subscales	Disability / spinal mobility: combine 2 gross 'spinal' movements –lumbar and cervical mobility (1 response) Pain Scale: omitted from final scale	Functional disability: 20 items	Functional disability: 25 items 8 components - 2 subscales added to 'activities' component of original HAQ a. 3 items-carrying, sitting and working b. 2 items - driving Pain and discomfort: 2 VAS items - severity of pain and stiffness	Functional disability grouped into 4 sub-scales: Mobility (4 / 4) Bending down (4 / 4) Reaching up and neck movements (6 / 4) Posture (5 / 4) LDQ =19items RLDQ=16items	Functional disability: 10 items 8 functional anatomy 2 coping with everyday life (items 9 and 10)	Patient identifies 0 to 15 AS-related problem areas Each problem 'weighted' by the patient
Time specificity	-	-	Past week	LDQ - RLDQ - past week	Past week	Past week
Administration	-	Interview 2 – 3 minutes [2] Self-complete 2 minutes[2]	HAQ [7]: self-complete 5-8 minutes, score 1-2 minutes HAQ-S not reported	Not reported (suggest 1-2 minutes)	Self-completed 1- 2 minutes [5]	Interview completed 10-15 minutes first completion 5-10 minutes subsequent [6]

Table 2.11 General description and scale structure of AS-specific measures of functional disability. Abbreviations defined in Glossary.

References: [1] Nemeth et al (1987), [2] Dougadas et al (1988), [3] Daltroy et al (1990), [4] Abbott et al (1994), [5] Calin et al (1994), [6] Bakker et al (1995), [7] Fries (1980), [8] Spoorenberg et al (1999a), [9] Kennedy et al (1993), [10] Bakker et al (1994), [11] Hidding & van der Linden (1995), [12] Hidding et al (1994b), [13] Hidding et al (1994a), [14] Jones et al (1996a), [15] Jones et al (1996b), [16] Jones et al (1996c), [17] Koh et al (1997a), [18] Ward & Kuzis (1999), [19] Santos et al (1998), [20] Taylor et al (1998), [21] Dougadas et al (1994), [22] Dougadas et al (1995), [23] Dougadas et al (1999a), [24] Calin et al (1999a), [25] Hidding et al (1993a), [26] Koh et al (1997b), [27] Band & Calin (1998), [28] Band et al (1997), [29] Maksymowych et al (1998), [30] Calin et al (1999b).

[a]DFI [2]	HAQ-S [3]	RLDQ [4]	BASFI [5]	MACTAS/PET [6]
Put on your shoes (1)	Dress yourself, including shoelaces / buttons? (1a)	Getting into and out of the bath (1a)	Putting on your socks or tights without help or aids (e.g. sock aid)(1)	(9 functional probes)
Pull on trousers (2)	Shampoo your hair (1b)	Getting into and out of the car (1b)	Bending forward from the waist to pick up a pen from the floor without an aid (2)	Mobility
Pull on a pullover (3)	Carry heavy packages such as grocery bags (SPAR 1a)	Getting up and out of bed in the morning (1c)	Reaching up to a high shelf without help or aids (e.g., helping hand)(3)	Self-care
Get into a bathtub (4)	Sit for long periods of time, such as at work (SPAR 1b)	Rolling over in bed (1d)	Getting up out of an armless dining room chair without using your hands or any other help (4)	Leisure activities
Remain standing 10 minutes (5)	Work at a flat topped table or desk (SPAR 1c)	Wiping yourself after using the toilet (2a)	Getting up off the floor without help from lying on your back (5)	Communication
Climb one flight of stairs (6)	Stand up from an armless straight chair (2a)	Putting on and taking off your socks (2b)	Standing unsupported for 10 minutes without discomfort (6)	Social interaction
Run (7)	Get in and out of bed (2b)	Putting on your shoes and tying your laces (2c)	Climbing 12-15 steps without using a handrail or walking aid (one foot on each step) (7)	Role activities
Sit down (8)	Look in the rear view mirror (SPAR 2a)	Cutting your toe nails (2d)	Looking over your shoulder without turning your body (8)	Emotional health
Get up from a chair (9)	Turn your head to drive in reverse (SPAR 2b)	Opening high windows (3a)	Do physically demanding activities (e.g. physiotherapy, exercises, gardening, or sports)(9)	Sleep and rest
Get into a car (10)	Cut your meat (3a)	Looking both ways before crossing the road (e g. do you have to move your feet?)(3b)	Doing the full day's activities whether it be at home or at work (10)	Appearance
Bend over to pick up an object(11)	Lift a full cup or glass to your mouth (3b)	Looking at what you are reaching on a high shelf (3c)		
Crouch (12)	Open a new milk carton (3c)	Drinking from a small glass or can (e.g. Do you have to bend your knees?) (3d)		
Lie down (13)	Walk outside on flat ground (4a)	Walk on your heels (4a)		
Turn in bed (14)	Climb up five steps (4b)	Coughing or sneezing (4b)		
Get out of bed (15)	Wash and dry your entire body (5a)	Sleep on your back (4c)		
Sleep on your back (16)	Take a tub bath (5b)	Sleep on your stomach (4d)		
Sleep on your stomach (17)	Get on and off the toilet (5c)			
Do your job or housework (18)	Reach and get down a 5 pound object (such as a bag of sugar) from just above your head (6a)			
Cough or Sneeze (19)	Bend down to pick up clothing from the floor (6b)			
Breathe deeply (20)	Open car doors (7a)			
	Open jars which have previously been opened (7b)			
	Turn faucets on and off (7c)			
	Run errands and shop (8a)			
	Do chores such as vacuuming or yardwork (8b)			
	Get into and out of the car (8c)			
	Severity of Pain (VAS) over the last week (9a)			
	Severity of Stiffness (VAS) over the last week (SPAR)			

Table 2.12 Item content of AS-specific measures of functional disability. Item number in parentheses. [a]References defined in table 2.11. Abbreviations defined in Glossary.

disability and discomfort associated with peripheral arthritis. The HAQ-S represents the impact of a disease characterised by axial involvement and oligoarthritis.

The Leeds Disability Questionnaire (LDQ)(Abbott et al, 1994) and the revised instrument, the Revised LDQ (RLDQ), evaluate AS-specific functional disability. Items were identified to fulfil the multi-dimensional nature of functional disability (Dougadas et al, 1986) and to describe functional groupings described by Badley et al (1984) (table 2.12).

The Bath Ankylosing Spondylitis Functional Index (BASFI)(Calin et al, 1994) is proposed as an evaluative measure to assist in the definition and monitoring of functional ability in AS. However, functional ability is not defined and item content is not related to any stated theory (table 2.12).

The MacMaster Toronto Ankylosing Spondylitis patient priority questionnaire / Patient Elicitation Technique (MACTAS/PET)(Bakker et al, 1995) provides a patient elicited AS-specific evaluation of functional disability and social dysfunction. Patients identify their own preference for improvement in patient-derived areas of functional handicap (Tugwell et al, 1987), an approach which is intended to assist in clinical decision-making. Following a broad definition of function capturing AS-specific difficulties with both physical and social function, nine functional disability groups act as prompts for item generation (table 2.12).

Study population

All developers indicate that the instruments are specific to AS evaluation, although the developers of the HAQ-S suggest application in patients with inflammatory spondyloarthropathy. However, patient populations involved in instrument development are often not clarified. The developers of the ASAQ provide no information about the patient population involved in instrument development. Only the developers of the HAQ-S described the sampling frame from which patients were selected and patient inclusion criteria, but method of patient selection was not indicated. No other developers clearly define the study sample frame, beyond the distinction of in-patient or out-patient status. All instruments, except for the MACTAS/PET, have been applied in both in-patient and out-patient populations but

the method of patient selection or inclusion criteria is largely unclear, with many reporting the participation of consecutive patients.

General description and scale structure

The ASAQ was proposed as a measure of pain and spinal immobility, but the pain item was deleted from the final instrument. Two items describing gross spinal movements of the lumbar and cervical spine were retained. The process of item development is not described. The response to both items is combined but there is little detail about the response scale. Score interpretation is not described.

The DFI was developed for interview-administration to be completed by a clinician following the verbal response of a patient. However, the majority of published studies have administered the DFI in a self-completed format. The DFI consists of 20 items relating to functional ability and activities of daily living (table 2.12). Following initial item selection, further development and initial testing was undertaken as part of a trial of AS out-patients with active disease (Dougadas et al, 1988). A principle component analysis (PCA) reduced the number of items from 29 to 20 and a strong level of concurrent validity was found between the original and the revised indices ($r = 0.95$). Each item is scored on a three-point ordinal scale of ability (table 2.11). Item scores are totalled (range 0-40); lower scores indicate better functional ability. The treatment of missing values is not described. Items are capacity based and readily understood by patients (Dougadas et al, 1988), although the time specificity is not indicated. Average completion time is 100 seconds (Calin et al, 1994).

In developing the HAQ-S five items were added to the disability index of the HAQ (total 25 items) and an additional VAS (stiffness severity) added to the pain and discomfort dimension. A postal survey of British AS patients was reviewed as a basis for item development. Items, referred to as sub-scales (SPAR 1 and SPAR 2), ask about ability over the previous week and are capacity based (tables 2.11 - 2.12). Items in the disability index are scored on a four-point ordinal scale of ability (table 2.11). However, the HAQ check-list for registering the use of aids or assistance, or the scoring procedure for the ordinal scale or VAS is not described. The HAQ totals the highest item score within each section (total range 0 – 24) and divides by the number of sections (8) to produce a score between 0-3 (Fries, 1980); lower scores

indicate better functional ability. The treatment of missing values is not described. Daltroy et al (1990) refers to a 'modified HAQ' without further description or reference. A 'Modified HAQ' (MHAQ) has been described (Pincus et al,1983) which contains only one item in each of the eight disability components and scoring includes all item scores.

The LDQ is a self-administered questionnaire describing four areas of AS-specific functional disability: mobility, bending down, reaching up and neck movements, and posture (table 2.12). Items derived from external sources related to the evaluation of functional disability in AS and rheumatology (Badley et al, 1984; Dougadas et al, 1986) were selected by the developers. A pre-pilot evaluation involved group discussion between selected AS in-patients (n= 12). The form of the discussion is not described. Following item selection patients graded items within each sub-section in increasing order of difficulty. The methods of patient selection and their characteristics are not provided. Following testing items were removed from the LDQ, but criteria for removal are not clearly specified. The revised instrument, RLDQ, contains 16 items (table 2.12). Each item is scored on a four-point ordinal scale similar to that used in the HAQ, but to improve discrimination between options the third option indicates the use of 'unusual movements or gadgets' (Abbott et al, 1994), an option that was further clarified in the RLDQ (table 2.11). The LDQ is scored in the same way as the HAQ: score range 0 - 3; lower scores indicate better functional ability. The treatment of missing values is not described. Scoring of the RLDQ has been revised to consider all items: score range 0 - 48; lower scores indicating better functional ability. If more than 2 items in any one section are missing then no final score is given (Helliwell P. – personal communication, 1999). Although capacity-based the time specificity of the LDQ is not described, but the RLDQ reflects a patients ability over the past week.

The BASFI consists of ten items that encompass functional anatomy (8 items) and the ability of a patient to cope with activities of everyday life (2 items)(table 2.12). Although a multidisciplinary development group including patients was indicated, further detail about members and their selection are not provided. The generation of items was not reported and the level of agreement between members before items were retained or rejected is not defined. The form of the discussion is not described. Each item is scored on 10cm horizontal VAS anchored by adjectival descriptors

'easy' and 'impossible', with no further distinguishing marks (table 2.11). The position marked on each VAS is recorded (0-10) and the mean value for the ten items gives the final BASFI score (0-10); lower scores indicating better functional ability. No guidance for missing values is provided. A normal score distribution covering 95% of the BASFI scale was reported in both in-patients and out-patients, and was accepted as support for instrument data quality. The BASFI is capacity based, patients scoring their ability with each activity over the last week. It requires a maximum completion time of 100 seconds.

The MACTAS/PET is a modification of the McMaster Toronto Rheumatoid Arthritis patient priority questionnaire (MACTAR)(Tugwell et al, 1987) and the Patient Elicitation Technique (PET)(Bell et al, 1990; Bakker et al, 1995). Administration with a trained interviewer follows a scripted format. Completion is in three stages. Patients identify up to 15 disease-related functional difficulties with normal activities of daily life over the last week, assisted by the nine functional probes generated following completion of a PET questionnaire in a large group of AS out-patients (n= 144)(Bakker et al, 1995). Patient identified problems are divided into three groups: mobility and role activity, social interaction and appearance. Each problem is ranked using seven-point Likert scaling for the relative difficulty / severity / frequency of the item, and an additional scale is used to assess the importance of each problem (0-7). To score the MACTAS/PET, for each problem the importance score (0–7) is multiplied against the relative difficulty / severity / frequency score (0–7). Results are totalled and divided by the number of problems identified (range 0-15) producing a possible range of 0-49; higher scores represent greater patient-perceived disability. Completion time is approximately 10-15 minutes on the first administration and 5-10 at subsequent completions. At follow-up patients are allowed to view the baseline questionnaire but are not allowed to alter the areas identified which are reassessed for difficulty/ severity/ frequency and importance, and a revised MACTAS/PET score calculated.

Reliability

Evidence of the reliability of all instruments is shown in table 2.13. Only a six week test-retest reliability assessment of the ASAQ has been performed and good levels of reliability calculated ($r > 0.78$).

Reliability[a]	ASAQ [1]	DFI [2]	HAQ-S [3]	LDQ (RLDQ) [4]	BASFI [5]	MACTAS/PET [6]
Test-retest	6-weeks retest. Self-completed (n= 52) AS out-patients [1] Blind completion (n=26) r=0.78 Open completion (n=26) r= 0.80	1-week retest. Interview-completion. (n=15) AS out-patients with axial disease [2] ICC = 0.86 1-day retest. Self-completed (n= 30) AS in-patients [5] DFI r= 0.96	-	5-day retest. Self-completed (n= 9) AS in-patients [4] ICC = 0.98 1-day retest - clinic and home Self-completed (n= 24) AS out-patients [4] ICC = 0.93	1-day retest. Self-completed (n= 30) AS in-patients [5] BASFI r= 0.89	-
Inter-observer	-	Interview-completion (n= 2 observers) (n= 15) AS out-patients with axial disease [2] ICC = 0.99	-	-	-	-
Intra-observer	-	-	-	-	-	-
Internal consistency	-	-	-	(n= 33) AS out-patients [4] Cronbachs alpha LDQ α= 0.93 Four sections - range Posture α= 0.62 Neck movements α= 0.90	(n= 2740) AS outpatients – retrospective analysis only [16] Cronbachs alpha α = 0.94	-

Table 2.13 Reliability of AS-specific measures of functional disability.

[a]References defined in table 2.11. Abbreviations defined in Glossary.

Test-retest reliability of the DFI has been assessed in both in-patient and out-patient populations and over one-week and one-day retest periods, and high levels of reliability have been calculated (ICC > 0.86). Condition stability in the original study was based upon the clinical opinion of the investigating physician (Dougadas et al, 1988). High inter-observer reliability has also been calculated (ICC 0.99). Internal consistency reliability of the DFI has not been evaluated.

Test-retest reliability of the LDQ has been assessed in both in-patient and out-patient populations and over one-day and five-day retest periods respectively. High levels of test-retest reliability have been calculated (ICC 0.93 - 0.98). A high level of internal consistency reliability was calculated for the total LDQ (α= 0.93) and for each instrument section. There is no published evidence of the reliability of the RLDQ.

High levels of one-day test-retest reliability of the BASFI in an in-patient population has been reported (r= 0.89). Internal consistency reliability has only been reported in a published letter reflecting a retrospective analysis of patients. A high level of alpha was calculated (α= 0.94).

There was no assessment of condition stability in the assessment of test-retest reliability for the ASAQ, LDQ and BASFI, and hence some patients may have changed thereby weakening the results. There is no published evidence of the test-retest reliability or the internal consistency reliability of the HAQ-S or the MACTAS/PET.

Validity

Evidence of the validity of all instruments is shown in table 2.14. Issues of face and content validity were not specifically addressed by any development authors. No investigators have established a priori hypothesised relationships between variables in the assessment of construct validity.

Evidence of the construct validity of the ASAQ is limited. The result of the comparison between the HAQ and the ASAQ was not reported and it was impossible to verify the nature of other outcome measures against which validity was assessed.

Validity	ASAQ [1]	DFI [2]	HAQ-S [3]	LDQ (RLDQ) [4]	BASFI [5]	MACTAS/PET [6]
Physical tests / signs	(n= 52) AS out-patients [1] 'Metrology' r= 0.60	(n= 80) AS outpatients with axial, active disease [2] FFD r= 0.35, Chest expansion r= -0.17 MSI r= -0.19 (n= 216) AS out-patients[18] Csp rotation r= -0.23, OWD r= 0.14 Chest expansion r= -0.27 Schober 10cm index r= -0.13 Smythes extension r= -0.09	(n= 44) AS outpatients [3] Csp rotation r= -0.57, Cexp r= 0.33, FFD r= 0.32, Smythes Test r= -0.33 (n= 216) AS out-patients[18] Csp rotation r= -0.50, OWD r= 0.33 Cexp r= -0.29 Schober 10cm index r= -0.36, Smythes extension r= -0.27	(n= 42) AS outpatients[4] ROM: strong correlation with 'Neck movements' and cervical spine ROM (ICC 0.63-0.73). Bending section and FFD (ICC 0.42)	-	(n=144) AS outpatients[6] Chest exp, Csp rot, Smythes technique (flex and ext). Correlation not reported. No statistically significant association reported
Pain	-	(n= 80) AS outpatients [2] Morning stiffness(duration) r= 0.54, Pain (VAS) r= 0.45, Sleep disturbance r= 0.22	-	-	-	(n=144) AS outpatients[6] (VAS) Pain r= 0.44, Stiffness r= 0.44
Disability assessments	(n= 52) AS out-patients [1] HAQ – result not reported	(n= 191) AS outpatients [8] BASFI r= 0.89 (n= 216) AS out-patients [18] HAQ r= 0.66, HAQ-S r= 0.64, AIMS2 r= 0.55	(n=44)AS outpatients [3] HAQr=0.98 (n=144)AS outpatients[12]SIP physical r= 0.59, psychosocial r=0.36, SAF physical r= -0.60, psychosocial r=-0.32 (n= 216) AS out-patients [18] HAQ r= 0.96, DFI r= 0.64, AIMS2 r= 0.80	-	(n= 191) AS outpatients [8] DFI r= 0.89	(n=144) AS outpatients[6] AIMS subscales - Anxiety r= 0.49, Pain r= 0.48, Depression r= 0.48, Health perception r= 0.41, SIP r= 0.43 Subjective health r= -0.41, DFI r= 0.34, HAQ-S r= 0.26
Other	(n= 52) AS out-patients [1] 'clinician-opinion' r= 0.61 'radiography' r= 0.56	(n= 80) AS outpatients with axial, active disease [2] Self-Physio r= 0.29, DAI r= 0.32 (n= 65) AS inpatients [9] Bath DAI r= 0.78 (n= 20) AS in-patients, 2 observers [5] 'external validity' DFI r= 0.90 (n=33)ASout-patients[10]Utility rating scale r=-0.49 Standard gamble r= -0.02 (n=144) AS out-patients [6] PET r= 0.34 (n= 191) AS out-patients [8] Disease activity: physician r = 0.36, patient r= 0.32, BASDAI r=0.57, Damage variables: BASRI r = 0.42, SASSS(modified) r= 0.36	(n= 44) AS outpatients [3] 4 groups of variables Highest correlation between – Personal variables and Flexion (initial)r= 0.69; Flexion (initial) and Function r= 0.61 (n= 33) AS out-patients [10] Utility rating scale r= -0.41, Standard gamble r= 0.04 (n=119) AS out-patients [6] PET r= 0.26, p< 0.05 VAS (HAQ-S) with PET r= 0.44, p< 0.05	(n= 17) AS in-patients, 2 observers - 'external validity' r= 0.96 [4] (n= 42) ASoutpatients[4] Patients stratified according to LDQ scores - Low (0-1.25) to High (1.26-3.00) Low LDQ = older, longer disease duration , greater ROM restriction & worse posture	(n= 20) AS in-patients, 2 observers [5] - 'external validity' BASFI r= 0.97 - 0.89 (n= 177) in-patients with (n= 215) postal respondents [14] BAS-G r= 0.54, BAS-G with items of BASFI range r= 0.30 – 0.59 (n= 295) AS postal survey [15]Major symptom of fatigue - 'significantly worse function' - higher mean BASFI score (5.87mean, 2.42 sd, p< 0.001) (n= 16) AS outpatients following cervical spine surgery and (n= 81) AS inpatients[17]Worse functional score in post surgical group (mean 7.07 vs 4.35). (n= 191) AS outpatients [8] Disease activity: physician r = 0.33, patient r= 0.33, BASDAI r= 0.59, Damage variables: BASRI r= 0.42, SASSS(modified)r=0.36 Self-report exercise, AS out-patients[19] Significant difference in score between non (n=915) and moderate (n= 1491) and between non and intensive (n- 393) exercisers (paired t-test) (n= 2965) AS out-patients [20] Disease duration r= 0.19	-

Table 2.14 Validity of AS-specific measures of functional disability. [8]References defined in table 2.11. Abbreviations defined in Glossary.

The evaluation of construct validity of the DFI is limited (table 2.14). Small to moderate correlation between the DFI and several poorly established clinical instruments and measures of disease activity and a high correlation with another AS-specific measure of functional ability (r= 0.89) has provided evidence for the validity of the DFI.

There is moderate evidence in support of the validity of the HAQ-S as a measure of functional disability (Daltroy et al, 1990; Hidding et al, 1994b). A strong level of concurrent validity with the HAQ was reported, and moderate to strong levels of correlation with an established generic measure (Sickness Impact Profile - SIP, Bergner et al, 1976) and an arthritis-specific measure of HRQL (Arthritis Impact Measurement Scale 2 - AIMS2, Meenan et al, 1992) were found (table 2.14). Small to moderate correlation with anthropometric measures were found.

The evaluation of construct validity for the LDQ has been limited to comparison with anthropometric measures and disease characteristics, and moderate to strong correlations were calculated (table 2.14). The concurrent validity of the RLDQ and the LDQ was not calculated and there is no evidence for the validity of the RLDQ.

The evaluation of construct validity of the BASFI is limited. Small to moderate correlation with several poorly established clinical variables and a high correlation with another AS-specific measure of functional ability (r= 0.89) was reported (table 2.14).

The evaluation of the construct validity of the MACTAS/PET is limited and relies on the results of one published study. However, a wide range of instruments were applied (table 2.14). Moderate correlation with generic (SIP) and arthritis-specific (AIMS2) measures of HRQL and small correlation with other AS-specific measures of functional ability (DFI, HAQ-S) were reported.

Responsiveness

Evidence of the responsiveness of all instruments is shown in tables 2.15 and 2.16.

There is no evidence of the responsiveness of the ASAQ or the RLDQ.

Study design	Used in trial with known efficacy
DFI [2]	RCT: known active compound (Piroxicam) and placebo (n= 80) AS outpatients. Stepwise logistic regression to determine the best variable with which to differentiate between groups. Clinicians 'overall judgement' best variable (p< 0.0001); followed by DFI (p< 0.001); DAI, Pain (VAS), and hand to ground distance (p< 0.01) [2]
	RCT: to determine efficacy of Ximoprofen (placebo controlled) 2-weeks duration (n= 285) AS outpatients. Statistically significant improvement in all variables – a 'better' improvement in the Ximoprofen groups irrespective of dosage p< 0.05 (Dunnett's t-test) for the DFI, DAI and pain (VAS) [21]
	Non-RCT: 3-week in-patient AS rehabilitation programme (n= 47). Compare questionnaires completed day '0' and '18'(n= 30). Non-significant change in DFI p= 0.19, 5.9% improvement [5]
	RCT: to evaluate efficacy and tolerability of sulphasalasine (SSZ)(placebo controlled) 6-months duration (n= 134) AS outpatients. DFI demonstrated a statistically significant change in favour of the treatment group at 6-months (p= 0.015)(ANOVA) [22]
	RCT: efficacy of addition of supervised group physiotherapy to unsupervised therapy (n= 144) AS out-patients. 9-month follow-up. Effect size calculated: low sensitivity of DFI E= 0.36 [6]
	RCT: supervised therapy and unsupervised therapy (continuation) (n= 68) AS out-patients. 9-month follow-up. Non-significant change in DFI and HAQ-S for both groups baseline to 9-months; significant change in the SIP for continuation group. Significant difference between groups for DAI, global health (VAS) and HAQ-S (favouring continuation)(t-test of change scores, p< 0.05). [12]
	RCT: efficacy and tolerability of NSAID therapy (placebo controlled)(n= 473) AS out-patients. Statistically significant change in favour of treatment groups seen in the DFI, patient global assessment (VAS) and Pain (VAS) at 6-weeks (p< 0.0167) and at 1-year (p< 0.05)(mean change compared by treatment groups using ANOVA) [23]
	RCT: to determine efficacy of NSAID (placebo controlled) 2-weeks duration (n= 473) AS out-patients. Statistically significant improvement in all variables for treatment group when compared to placebo group (DFI p= 0.0001) [24]
HAQ-S [3]	RCT: efficacy of addition of supervised therapy versus unsupervised therapy (n= 144) AS out-patients. 9-month follow-up. No statistically significant difference between groups for 'functioning' and all secondary outcomes. Baseline score of zero in 25% for HAQ-S and SIP- indices re-analysed excluding patients scoring zero. No significant additional effect of group therapy relative to unsupervised therapy. Suggest that HAQ-S and SIP may not be sensitive enough to detect change in AS. Statistically significant difference between groups for thoracolumbar mobility, fitness and patient assessed global effect (VAS)(comparison of groups for mean improvement – t-test of change scores) [25]
	RCT: efficacy of addition of supervised group therapy to unsupervised therapy (n= 144) AS out-patients. 9-month follow-up. Effect size for HAQ-S low E= 0.05 (Physical fitness E= 3.50) [6]
	RCT: supervised therapy and unsupervised therapy (continuation) (n= 68) AS out-patients. 9-month follow-up. Non-significant change in DFI and HAQ-S for both groups baseline to 9-months; significant change in the SIP for continuation group. Significant difference between groups for DAI, global health (VAS) and HAQ-S (favouring continuation)(t-test of change scores, p< 0.05). [13]
LDQ [4]	RCT: 3 physiotherapy regimes (n= 42) AS outpatients. Assess pre-randomisation, post-treatment (6-weeks), 6-months. Significant improvement in functional score between baseline and 6-weeks (t= 2.79, p< 0.01). Interpretation of 6-month results difficult due to poor response rate. 4 items in LDQ 'not sensitive to change', therefore omitted in the revised instrument (RLDQ) [4]
BASFI [5]	Non-RCT: 3-week in-patient AS rehabilitation programme (n= 47). Compare questionnaires completed day '0' and '18'(n= 30). Significant improvement in BASFI p= 0.004; 19.6% improvement [5]
	RCT: efficacy of amitriptyline in reduction of pain, fatigue and stiffness (placebo controlled)(n= 81) AS in-patients (rehabilitation programme). Baseline and 2-week completion . Significant change in BASFI score for both treatment groups (improvement range 23-29%, p< 0.001). Non-significant difference between treatment and placebo groups when assessed by the BASFI [26]
	Non-RCT: 2-week in-patient AS rehabilitation programme. 6-week follow-up (n= 100) (response rate 79%). Comparison between SF-36, BASFI, BASDAI and BAS-G. Effect size calculated – BASFI and physical functioning, role functioning, social function and vitality sections of the SF-36 similar small effect size (BASFI 0.19, SF-36 range 0.12 to 0.15) [27]
	Non-RCT: 2-week in-patient AS rehabilitation programme (n= 236). Compare questionnaires completed day '0' and '14'. Significant improvement in BASFI p< 0.001; 25% improvement [28]
	RCT: to determine efficacy of NSAID (placebo controlled) 2-weeks duration (n= 473) AS out-patients. Statistically significant improvement in all variables for treatment group when compared to placebo group (BASFI p= 0.0001) [24]
MACTAS / PET [6]	RCT: efficacy of addition of supervised group therapy to unsupervised therapy (n= 144) AS outpatients. 9-month follow-up. Effect size calculated: high level of sensitivity for the MACTAS / PET E= 0.60. Largest effect sizes for Borg Scale (physical fitness - E 3.50), Smyth Technique (spinal ROM - E 0.76) and patient assessed improvement (VAS - E 0.76) [6]

Table 2.15 Responsiveness results I (trials of known efficacy) of AS-specific measures of functional disability. [a]References defined in table 2.11. Abbreviations defined in Glossary.

Strategy applied	Correlation of scale change with changes in other measures
DFI [2]	RCT: known active compound (Piroxicam) and placebo (n= 80) AS outpatients. 2-weeks treatment. Correlation between change in DFI score and 'overall judgement' of treatment efficacy (clinician & patient) – rated nil / moderate / good / very good. Change in score of DFI (and DAI) highly correlated with 'clinician overall judgement' regardless of treatment ($p< 0.0001$) [2]
	Longitudinal analysis of NSAID (n= 985) AS outpatients. Assess baseline and 6-weeks. Stepwise logistic regression to determine 'clinical relevance' of DFI. Patient 'overall assessment' of treatment success or failure dependent variable. DFI and other measures independent variables. DFI and Pain (VAS) most significant variables to discriminate between patients considering treatment a success or failure – most appropriate variables for assessing efficacy of NSAID therapy in AS [28]
	RCT: efficacy of the addition of supervised group therapy to unsupervised therapy (n= 144) AS outpatients. 9-month follow-up. Comparison of change between measures (change = follow-up minus baseline result). Significant correlation between DFI and MACTAS / PET ($r= 0.34$, $p< 0.005$) [6]
	Non-RCT: Longitudinal evaluation group physiotherapy (n= 68) AS out-patients. 9-month follow-up. Correlation of change in patient-assessed global health (VAS) $r= 0.16$ [11]
	Longitudinal study of outcomes in AS (n= 155) out-patients, 6-month follow-up. External criteria of change = patient reported change in disease activity (4-categories). SRM calculated for patients reporting a 1 or 2-category change in disease activity. Higher levels of responsiveness calculated for 2-category reported change: DFI 0.40, HAQ-S 0.55, HAQ 0.51 [18]
	Longitudinal study of outcomes in AS (n= 153) out-patients, 2-year follow-up. Relationship between change in score over time and change in patient rating of pain and stiffness. Significant relationship between change in DFI and change in stiffness only ($p > 0.0001$) [18]
HAQ-S [3]	Non-RCT: Longitudinal evaluation group physiotherapy (n= 68) AS out-patients. 9-month follow-up. Correlation of change in patient-assessed global health (VAS) $r= 0.30$ [11]
	Longitudinal study of outcomes in AS (n= 155) out-patients, 6-month follow-up. External criteria of change = patient reported change in disease activity (4-categories). SRM calculated for patients reporting a 1 or 2-category change in disease activity. Higher levels of responsiveness calculated for 2-category reported change: HAQ-S 0.55, HAQ 0.51, DFI 0.40 [18]
	Longitudinal study of outcomes in AS (n= 153) out-patients, 2-year follow-up. Relationship between change in score over time and change in patient rating of pain and stiffness. Significant relationship between change in HAQ-S and change in pain and stiffness ($p > 0.0001$) [18]
BASFI [5]	Longitudinal 'open analysis' of pamidronate in treatment of AS (n= 8) outpatients (3-month intervention and 3 / 6 / 9-month follow-up). BASFI, BASDAI, BASMI and laboratory assessments. Comparison between baseline and follow-up - a non-significant ($p> 0.05$) reduction in BASFI scores, but a significant improvement in BASDAI, BASMI and ESR observed (paired t-test) [29]
MACTAS / PET [6]	
	RCT: efficacy of the addition of supervised group therapy to unsupervised therapy (n= 144) AS outpatients. 9-month follow-up. Comparison of change between measures (change = follow-up minus baseline result). Significant correlation between MACTAS / PET and two AIMS sub-scales (pain $r= 0.41$, $p< 0.005$, physical activity $r= 0.29$, $p< 0.05$), subjective health score ($r= -0.36$, $p< 0.005$), rating scale utilities ($r= -0.28$, $p< 0.05$), DFI ($r= 0.34$, $p< 0.005$) and cervical rotation ($r= -0.29$, $p< 0.05$) [6]

Table 2.16 Responsiveness results II (correlation of scale change with changes in other measures) of AS-specific measures of functional disability.

[a]References defined in table 2.11. Abbreviations defined in Glossary.

The strongest evidence of the responsiveness of the DFI has been reported in drug therapy trials which suggest statistically significant change in functional ability following active treatment. The observed change in the DFI was often closely associated with change in other clinical variables (tables 2.15-2.16). However, weak evidence of responsiveness following physiotherapy intervention has been reported.

Accumulated evidence suggests that the HAQ-S is not responsive to clinically important change in functional ability in AS (Hidding et al, 1993a; Bakker et al, 1995), although it may be able to detect differences in functional ability between groups of patients (Hidding et al, 1993a). Comparison of change in score of the HAQ-S with change in score of various other measures of health and functional status strengthen the evidence of poor responsiveness (table 2.16).

Limited evidence of LDQ responsiveness suggests that it is responsive to clinically important change in AS out-patients receiving physical therapy, a treatment of known efficacy, over the short term (6-weeks)(table 2.15).

Evidence suggests that the BASFI is capable of detecting statistically significant change in functional ability over the short-term following physical therapy, but not following evaluations of drug therapy. Comparison with an established generic measure of HRQL, the SF-36, following a short-term evaluation of in-patient physiotherapy failed to provide strong evidence of responsiveness for either instrument (table 2.15).

Several different evaluations on the same data-set provide good evidence in support of the responsiveness of the MACTAS/PET following physical therapy intervention in a large and heterogeneous group of patients. The instrument compared favourably to other instruments demonstrating evidence of responsiveness (tables 2.15-2.16).

Acceptability

All developers have reported good levels of acceptability by patients, although no developer indicates how the level of acceptability was derived or defined. Completion rates following in-patient or postal completion have not been reported.

Feasibility and application

The DFI and BASFI have become the most widely applied AS-specific measures of functional disability, although the majority of articles using the BASFI have been published by members of the development team (n= 12/14 articles). The DFI has been most widely accepted by investigators outside of the development team (n= 13/19 articles). The original DFI was developed in the French language, but procedures for translation into English have not been described. Both the DFI and BASFI have been translated into other languages (Dutch Functional Index - Creemers et al, 1994; Turkish Functional Index - Dalyan et al, 1999; French BASFI - Claudepierre et al, 1997; Swedish BASFI - Cronstedt et al, 1999).

The HAQ-S has been widely used by other investigators, but seven articles refer to different stages in the same randomised trial of physical therapy. There is only one published article describing the development and initial testing of the LDQ, which proposed adoption of the revised instrument, the RLDQ (Abbott et al, 1994).

Commentary

The purpose of the six instruments is similar - to evaluate AS-specific functional disability. Table 2.17 provides a summary of the data synthesis. All instruments lack sufficient information about intended purpose and conceptual base. The DFI and LDQ provide the most detailed attempt at a definition of purpose. The LDQ has a similar theoretical grounding to the DFI but a stronger conceptual base due to the association with an arthritis disability classification (Badley et al, 1984). However, the theoretical approach adopted is poorly explained.

There is a lack of information supporting the item development of all instruments. Although most instruments involve patients at some stage of item development there is limited information to indicate patient selection or characteristics, beyond the diagnosis of AS, and the nature of the patient role in item selection. The content of all instruments has been dominated by expert opinion but there is little information to support the experience of these experts and the criteria by which items were selected. Without a clear definition of instrument purpose and conceptual base an appreciation of item development is limited. In combination these factors also restrict an appraisal of content and construct validity.

	AS-specific functional disability [a]												
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
	Scale	Number of items	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
ASAQ [1]	ordinal	2 (3)	'Epidemiological, comparative & analytical studies'	Self-administered 1 minute	1	3	0	+	+	+	+	0	0
DFI [2]	ordinal	20	Not specified - clinical practice, research	Interview completion 2-3 minutes Self-complete 1-2 minutes	1	7	13	+/++	++	++	++	++	+/++
HAQ-S [3]	Disability Index - ordinal Pain & Discomfort - VAS	8 components – 25/13 items 2 components – 2x VAS	Clinical practice, research	Self-administered 5-8 minutes Time to score 1-2 minutes	1	1	7	0	0	++	++	+	0
LDQ [4]	ordinal	19	Clinical practice, research	Self-administered 1-2 minutes	1	0	0	+	++	+	+	+	+
RLDQ [4]	ordinal	16	Clinical practice, research	Self-administered 1-2 minutes	1	0	0	0	0	0	0	0	0
BASFI [5]	VAS	10	Not specified – clinical practice, research	Self-administered 1-2 minutes	1	6	9	+/++	+/++	++	++	+/++	+/++
MACTAS / PET [6]	Likert Scale (0-7)	0-15 patient generated	Clinical practice, research	Interview Initial 10 -15 minutes Follow-up5 -10 minutes	1	0	0	0	0	+/++	++	+/++	+/++

Table 2.17 AS-specific measures of functional disability - summary of data evaluation (after McDowell and Newell, 1996).

[a]References defined in table 2.11. Abbreviations defined in Glossary.
Superscript: [b]Stage of instrument development: I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability/Validity/Responsiveness: 0 = no evidence to +++ = strong evidence.

The degree of limitation in functional disability investigated by each instrument varies. The ASAQ is limited as a measure of functional disability. The DFI has good face validity, containing items that cover a wide range of functional activities (table 2.12). However, it does not address neck mobility. Subsequent investigators have described the importance of neck mobility in AS-specific functional disability (Daltroy et al, 1990; Nemes, 1991) and items to address neck mobility are recommended in a modification of the DFI.

The version of the HAQ adopted for the HAQ-S is not clear. Clarification is needed to support standardisation of application. There is evidence that the HAQ-S fails to focus sufficiently on spondyloarthropathy related disability and that HAQ items with a focus on peripheral joint arthritis are of limited relevance to AS patients. A ceiling effect in 25% of AS out-patients completing the HAQ-S suggested that their functional ability could not get any better (n= 144)(Hidding et al, 1993a). Low mean values and inadequate score coverage further supports the inadequacy of the HAQ-S as a measure of functional disability in AS (Hidding et al, 1993a; 1994b). A revision of item content is required.

The LDQ does not include more strenuous functional activities such as running or housework, activities included in both the DFI and BASFI. The lack of more arduous activities may be a reflection of the patient population included in instrument development. In-patients may reflect the more severe spectrum of disease (Kennedy et al, 1995) and patients with long-standing disease often report adaptation in functional activities and changed priorities. As a result, the more basic requirements of functional ability may appear to be paramount. The inclusion of more difficult functional activities should improve instrument content validity.

The BASFI covers a wide spectrum of functional disability sharing a similarity of item content with other instruments (table 2.12). Although it does not include items reflecting easier functional activities, score distribution suggests that it captures the extremes of functional disability when completed by patients representing a broad spectrum of disease (Calin et al, 1994).

The functional probes included in the MACTAS/PET and the results of validity testing suggest that it may represent a broad reflection of HRQL, as opposed to functional disability.

In light of criticisms of insensitivity the response scale of the DFI has recently been modified to a five-point adjectival scale (Spoorenberg et al, 1999a). The involvement of the original developers (Dougadas) should help to distinguish the official modified version from other 'modifications' referred to in published articles. However, the modification is inadequately described by Spoorenberg et al (1999a).

The HAQ-S and the LDQ adopt the same four-point ordinal response scale as the HAQ, despite critisism for its relative insensitivity to change (Liang et al, 1985; McDowell & Newell, 1996). Results from completion of the HAQ-S would suggest that it suffers similar problems of insensitivity. However, this may be related to both an inadequacy of item content and of the response scale. Although offering more response options than the original DFI, there is limited evidence of the responsiveness of the LDQ (RLDQ), and the five-point adjectival response scale of the modified DFI (Spoorenberg et al, 1999a) may improve the sensitivity of the DFI beyond that offered by the RLDQ.

The BASFI includes ten VAS. The acceptance of VAS by clinicians and patients is not clear, and reservations about the feasibility of, and understanding associated with, these scales have been expressed (Streiner & Norman, 1995; Fitzpatrick et al, 1998a). A recent survey suggests that Likert type response scales are more readily accepted than VAS by clinicians in routine evaluation (Bellamy et al, 1998; 1999).

Notwithstanding the wide range of retest intervals and the lack of consideration for disease stability shown by many authors, evidence of the test-retest reliability of the ASAQ, DFI, LDQ and BASFI seems satisfactory and supports application of each instrument in group evaluation (> 0.70). The reliability of the DFI and the LDQ supports application in individual evaluation (> 0.90). Improved evidence of test-retest reliability with appropriate retest periods and the external evaluation of condition stability is required, particularly for the ASAQ, HAQ-S, RLDQ, BASFI and MACTAS/PET. The assessment of internal consistency reliability is required for all

instruments, except for the MACTAS/PET for which the calculation is not appropriate.

There is limited evidence to support the validity of all instruments. All investigators fail to hypothesise expected relationships between various constructs a priori to the analysis and the results are therefore logically weak (McDowell and Jenkinson, 1996). Comparison of the DFI, LDQ/RLDQ and BASFI with other more established measures of functional ability or HRQL, either generic or disease-specific, has not been undertaken and are recommended with a priori hypothesised relationships stated to provide a wider appreciation of the constructs addressed.

All instruments, apart from the ASAQ and the RLDQ, have been applied in trials of known efficacy and several strategies for assessing responsiveness have been applied to the DFI, HAQ-S, BASFI and MACTAS/PET. There is strong evidence to support the responsiveness of the DFI following drug therapy intervention and increasing evidence for the BASFI following short-term rehabilitation programmes. There is limited evidence for the responsiveness of the LDQ and the MACTAS/PET following trials of physical therapy. Accumulated evidence suggests that the HAQ-S is not responsive to change following physical therapy.

All instruments are easy to understand although the lack of detail relating to the ASAQ makes interpretation difficult. All instruments, apart from the MACTAS/PET which requires interview-administration, are quick to administer (2-8 minutes) and are self-completed. The DFI, LDQ and BASFI have all been administered in both in-patient, out-patient and postal evaluations.

Conclusion

All instruments have been developed integrating clinical expertise with a theoretical appreciation of functional ability in AS. Based upon the limited data available following the initial literature search (April 1998), three of the six AS-specific measures of functional disability have acceptable evidence in support of their development and measurement properties and appear to be acceptable to patients in a self-completed format: DFI, LDQ/RLDQ and BASFI.

The ASAQ and HAQ-S are not recommended for application in AS. These inadequacies relate to the lack of clarity in instrument purpose, development and structure, irrelevance of certain items, inadequacy of response scale and lack of standardisation in instrument administration. However, the HAQ-S highlighted the relevance of neck mobility in the assessment of AS-specific functional disability.

The MACTAS/PET provides a wider view of the impact of AS than a simple reflection of functional disability. Although providing a novel approach to evaluation it has not been widely accepted into research or routine practice. Methods of dealing with patients unable to identify problem areas, and further evidence of the acceptability, reliability and responsiveness is necessary.

The final selection was made between the DFI, LDQ/RLDQ and BASFI. All instruments had positive and negative points with important gaps relating to development and performance (table 2.18), making selection of one instrument difficult.

All instruments	DFI	LDQ/RLDQ	BASFI
Data quality and scaling assumptions	Item content modification - items relating to neck mobility	Item content modification - more arduous functional activities	Test-retest reliability - longer retest period and external evaluation of change
Response scale comparison - Likert, VAS and numerical rating scales	Testing of revised response scale	Further testing of revised instrument - RLDQ	
Explicit rules for missing data	Time specificity	Further evaluation of responsiveness - drug trials and longer term follow-up	
Internal consistency reliability	Compare interview and self-completed formats		
Validity testing against established instruments - hypothetical constructs proposed *a priori*	Formal cross-cultural adaptation into English		

Table 2.18 Gaps in the evidence base of three AS-specific measures of functional disability.

The original DFI has more published evidence in support of its development and testing. The application and testing of the BASFI has not been widely described outside the development base. The LDQ/RLDQ remains at the very early stages of

development, and only one published article has been identified (Abbott et al, 1994), which recommends adoption of the revised instrument (RLDQ).

There is greater clarity of instrument purpose and conceptual base for the LDQ than seen for both the DFI and BASFI, and an attempt by developers to include items in response to a definition of functional disability. All instruments involved clinical experts in item development and the LDQ and BASFI included patients. All instruments have a similarity of item content, although the DFI lacks items addressing neck mobility, and the LDQ lacks the more arduous activities included in the DFI and BASFI.

There is greater evidence in support of both the test-retest and internal consistency reliability of the LDQ than for the DFI and BASFI, but evidence for validity is limited for all instruments. There is limited evidence in support of the responsiveness of all instruments, with the strongest evidence for the DFI following placebo-controlled drug trials.

In recommending both the DFI and the BASFI in the evaluation of AS functional ability (van der Heijde et al, 1999a,b,c; Ruof & Stucki, 1999a), the ASAS make no reference to the LDQ and no reference to the modified DFI. This is a serious omission by the ASAS group in making recommendations for a standardised set of instruments and the modified DFI should be specified. It also suggests that the extent of the instrument review was limited.

In light of the proposed revision to the DFI response scale (Spoorenberg et al, 1999a), the support for the DFI may increase. However, this modification was reported after instrument selection for the comparative study (Chapter 4). The conceptual base, development methodology and early evidence of measurement properties suggests that the RLDQ is worthy of further testing, and was selected for the proposed study.

2.6.2 AS-specific disease activity

This section describes the evaluation of the five AS-specific measures of disease activity (tables 2.9 and 2.10). A general description and the scale structure is summarised in table 2.19. The NEI, SEI and DAI require clinician completion and the Bath DAI and BASDAI are patient-completed.

	NEI [1]	SEI [2]	DAI [3]	Bath DAI [4]	BASDAI [5]
Response options	4-point adjectival and numerical scale to score 15 sites of entheses Score 0-45 0 = no disease activity 45 = maximum disease activity	4-point adjectival and numerical scale to score 22 sites of entheses Score 0-66 0 = no disease activity 66 = maximum disease activity	4-point adjectival and numerical scale to score 10 joint sites Score 0-30 0 = no disease activity 30= maximum disease activity	3x vertical, adjectival and numerical (1-4 and 1-6), Likert scales. 1x VAS with numerical and adjectival descriptors (1-10) Score 3-26 3 = no disease activity 26= maximum disease activity	6x VAS with endpoints 'none' and 'very severe' (10cm) Mean score of items 1-4 plus mean score of items 5-6, divided by 2. Score 0-10 0 = best health 10 = worst health
Subscales	Entheses: nuchal crests, manubrio-sternal joint, costochondral joints, humeral greater tuberosity, humeral epicondyles, anterior superior iliac crests, femoral greater trochanters, pelvic adductor origin, femoral lateral epicondyles, Achilles tendon insertion, plantar fascia insertion, cervical / thoracic / lumbar spinous processes, ischial tuberosities, posterior superior iliac spines.	Entheses: symphysis pubis, vertebral processes (C1/2, C7/T1, T12/L1, L5/S1), femoral greater trochanters, pelvic adductor origin, anterior superior iliac crests, ischial tuberosities, sternoclavicular joints, sternocostal joints, Achilles tendon insertion, plantar fascia insertion.	Joint sites: thorax (antero-lateral pressure), hip flexion, buttock (right / left pressure), cervical spine rotation, dorsolumbar spine rotation	Severity of pain, frequency of severe pain, patient assessed disease activity, AS-related well-being	Fatigue / tiredness, spinal pain, peripheral joint pain / swelling, localised tenderness / bodily discomfort, severity and duration of morning stiffness
Time specificity	Present time	Present time	Present time	Past month	Past week
Administration	Clinician completed Approximately 3-minutes.	Clinician completed Approximately 3-4 minutes.	Clinician completed Approximately 5-minutes	Self-completed Approximately 4-minutes	Self-completed range 30 seconds to 2-mintues Time to score?

Table 2.19 General description and scale structure of AS-specific measures of disease activity. Abbreviations defined in Glossary.

References: [1] Mander et al (1987), [2] Dawes et al (1987), [3] Dougadas et al (1988), [4] Kennedy et al (1993), [5] Garrett et al (1994), [6] Creemers et al (1996), [7] Hidding et al (1993a), [8] Hidding et al (1993b) [9] Bellamy et al (1991a), [10] Jones et al (1996c), [11] Calin et al (1999a), [12] Dougadas et al (1990), [13] Goodacre et al (1991), [14] Zukovskis et al (1991), [15] Bakker et al (1995), [16] Hidding & van der Linden (1995), [17] Jones et al (1996a), [18] Koh et al (1997a), [19] Santos et al (1998), [20] Taylor et al (1998), [21] Spoorenberg et al (1999b), [22] Dougadas et al (1994), [23] Band et al (1997), [24] Koh et al (1997b), [25] Maksymowych et al (1998), [26] Dawes (1999), [27] Bakker et al (1994b), [28] Hidding et al (1994a), [29] Calin et al (1999b).

Purpose and conceptual base

The Newcastle Enthesitis Index (NEI)(Mander et al, 1987) and the Stoke Enthesitis Index (SEI)(Dawes et al, 1987) were the first non-invasive clinical methods to evaluate disease activity in AS as a reflection of the extent and severity of enthesitis. The assessment of disease activity had previously relied upon laboratory-based assessment, or the evaluation of pain and stiffness.

The Dougadas Articular Index (DAI)(Dougadas et al, 1988) was the first published AS-specific instrument to score joint tenderness. The initial choice of joint sites was based on the expert opinion of three rheumatologists familiar with the pathological process of AS.

The Bath Disease Activity Index (Bath-DAI)(Kennedy et al, 1993) was the first self-administered AS-specific measure of disease activity focussing on a patient's perception of symptoms, disease progression and global well-being. The developers theorised that a patients perception of disease activity could be described by pain, stiffness and the need for persistent medication, providing a more relevant assessment of disease activity than laboratory-based assessment and radiographic analysis, but provide no further support for this theory.

The Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)(Garrett et al, 1994) was proposed as a simple self-administered measure in response to the need to provide a comprehensive measure of AS-specific disease activity (Garrett et al, 1994). Although the developers do not define disease activity, the need to distinguish between disease activity and severity is highlighted.

Study population

All developers indicate that the instruments are specific to the evaluation of patients with AS. However, there is a lack of clarity about the patient populations involved in the development of most instruments. No developers clearly describe the population sampling frame beyond the distinction of in-patient or out-patient status, and method of patient selection was generally unclear. Many investigators do not detail patient variables or disease characteristics. All instruments, except for the SEI have been applied in both in-patient and out-patient populations. Only the BASDAI has been applied in a postal survey (Jones et al, 1996b).

General description and scale structure

The NEI consists of 15 items reflecting sites of entheses involved in the pathological process of AS. Items (entheses) were generated by the clinical opinion of the development team. However, the number and experience of participants was not described and the extent of the literature search was not detailed. Patient-reported discomfort following palpation of chosen sites by a trained clinician is recorded. Following testing (n= 6 AS out-patients with active disease) non-responsive items were removed from the draft instrument; that is, no change in entheses activity over a one-week period. Each item is scored on a four-point ordinal scale of discomfort (table 2.19), and although patient input was not described, a standard palpation in an area devoid of enthesis (clavicle) is recommended before each evaluation to foster patient discrimination between pain and palpatory pressure. Item scores are totalled (range 0-45); a lower score indicating less disease activity. The treatment of missing values is not described. All items are performance based and completion time is approximately three minutes.

The SEI consists of 22 items reflecting 13 enthesis zones involved in the AS pathological process. The level of patient reported discomfort is recorded following palpation of chosen sites by a trained clinician. There is a similarity of item content to the NEI (table 2.20), but no detail about item development for the SEI.

Spine	Pelvis	Lower limb
Cervical, thoracic and lumbar spinous processes = 1 group (NEI)	Anterior superior iliac spines (bilateral)(NEI)	Femoral greater trochanters (bilateral)
(Separate) Vertebral processes at (SEI):	Anterior superior border of iliac crests (bilateral)(SEI)	Achilles tendon insertion (bilateral)
C1 / 2	Ischial tuberosities (bilateral)	Plantar fascia insertion (bilateral)
C7 / T1		
T12 / L1		
L5 / S1		

Table 2.20 Similarity of item content between the NEI and SEI

Dawes et al (1987) indicate that the index is scored in a similar way to the NEI, and describe a four-point ordinal scale of patient discomfort in response to firm palpation over the enthesis site. Item scores are totalled (range 0-66); a lower score indicating less disease activity. The treatment of missing values is not described. All items are

performance based, but completion time is not calculated. No attempt to standardise palpation or to improve patient discrimination is described.

Following the initial selection of items for the DAI by three rheumatologists, further development and initial testing was undertaken as part of a trial of AS out-patients with active disease (n= 88)(Dougadas et al, 1988). A principle component analysis (PCA) reduced the number of items to 10 (table 2.19) and a strong level of concurrent validity was found between the original and revised indices (r= 0.84). The homogeneity of items was further supported in a subsequent, larger study (Dougadas et al, 1990). Completion requires a trained clinician. A patient response is recorded following joint movement or firm digital pressure to a designated site. Each item is scored on a four-point ordinal scale of discomfort (range: 'no tenderness' (0) to 'patient said that it was painful, winced and withdrew the limb' (3)). Item scores are totalled (range 0–30); a lower score indicates less disease activity. The treatment of missing values is not described.

The Bath-DAI contains four items with a focus on pain, disease activity and well-being 'selected' from the AIMS, an arthritis-specific measure of HRQL (Meenan et al, 1980). The criteria for item selection or item development is not provided. Two items (pain severity, and frequency of severe pain) are scored on vertical six-point Likert scales ('none' / 'never' (1) to 'very severe' / 'always' (6) respectively). A third item, patient-assessed disease activity, is scored on a four-point scale ('not at all active' (1) to 'very active' (4)). The fourth item, AS-related well-being, is scored on a horizontal 10cm VAS with adjectival (very well, well, fair, poor and very poor) and numerical (0, 2.5, 5.0, 7.5 and 10) descriptors. Item scores are totalled (range 3-26); a lower score indicates a lower patient perceived disease activity. Each response is considered over the previous month. The treatment of missing values is not described.

The BASDAI is self-completed containing six items representative of AS disease activity (table 2.19). Item development was based upon the clinical experience of a multi-disciplinary group including patients, but the number of participants and the form of discussion is not described. Item generation and selection was not reported. Patient selection, disease characteristics and level of input is not clarified. The item relating to fatigue was included due to recent research from the development group.

Following testing, the quality of morning stiffness was added and item wording modified. Each item is scored on a 10cm horizontal VAS anchored by adjectival descriptors 'none' and 'very severe' (items 1-5). Item 6 (duration of morning stiffness) is anchored by a time scale ('0' to '2-hours') with marks every quarter of an hour, and further adjectival support at half-hour intervals. The two hour period was selected following a retrospective analysis of patient questionnaires (unpublished data)(Garrett et al, 1994). The VAS scoring procedure is not described. It is assumed that the position marked on each VAS is recorded (0-10). The mean of the two morning stiffness items (items 5 and 6) is calculated to ensure equal weighting of symptoms. All items are totalled (range 0-50) and the total score converted to a 0-10 scale; a lower score indicates less disease activity. No guidance for missing values is provided. The BASDAI is capacity based, considers symptoms over the previous week, and has an average completion time of 67 seconds (range 30 seconds to 2 minutes).

Reliability

Evidence of the reliability of all instruments is shown in table 2.21. Low levels of test-retest reliability (ICC 0.52) and significant inter-observer variability has been reported for the NEI, prompting the developers to suggest that the same observer should be employed during clinical trials. A higher level of reliability was reported by Creemers et al (1996) but the methodology is unclear.

Test-retest reliability of the DAI has been assessed following one-week and 48-hour retest periods in out-patient populations and moderate to high reliability reported (ICC 0.59 to 0.83). High inter-observer reliability has been calculated (ICC > 0.90).

High one-day test-retest reliability of the Bath-DAI and BASDAI has been reported in the same in-patient population supporting their application in individual assessment ($r > 0.93$). A high level of internal consistency reliability has been reported for the BASDAI in a published letter reflecting a retrospective analysis of patients (Jones et al, 1996c), and more recently following a drug therapy trial (Calin et al, 1999a)(α 0.84).

There is no published evidence for reliability of the SEI, and internal consistency reliability has not been evaluated for the NEI, SEI, DAI or the Bath-DAI.

[a] Reliability	NEI [1]	SEI [2]	DAI [3]	Bath DAI [4]	BASDAI [5]
Test-retest	-	-	2 trained observers, 1-week retest. (n=15) AS out-patients with axial disease [3] ICC 0.83	1-day retest. Self-completed (n= 46) AS in-patients [5] r= 0.96	1-day retest. Self-completed (n= 46) AS in-patients [5] r= 0.93
Inter-observer	3 trained observers, (n= 18) AS out-patients (ANOVA)[1] Significant variability between observers ($p< 0.01$) at initial assessment. No significant difference between observers in subsequent re-assessments at 2 and 3-weeks. No significant difference in assessment of pain and stiffness (VAS)	-	2 trained observers, (n=15) AS out-patients with axial disease [3] ICC 0.94 5 trained observers, (n= 7) AS out-patients [9] Pre-standardisation r= 0.98 Post-standardisation r= 0.90	-	-
Intra-observer	1 trained observer , 48-hour retest (n= 19) AS out-patients [7,8,28] ICC 0.54 ? retest, ? observers, (n= 59) AS out-patients [6] NEI r= 0.74	-	1 trained observer, 48-hour retest (n= 19) AS out-patients [7,28] ICC 0.59	-	-
Internal consistency	-	-			(n= 2740) AS outpatients - retrospective analysis only [10] Cronbachs alpha $\alpha = 0.84$ (n= 473) AS out-patients [11] Cronbachs alpha $\alpha = 0.84$

Table 2.21 Reliability of AS-specific measures of disease activity.

[a] References defined in table 2.19. Abbreviations defined in Glossary.

Validity

Evidence of the validity of all instruments is shown in table 2.22. Issues of face and content validity have only been specifically addressed for the Bath-DAI. No investigators have established a priori hypothesised relationships between variables in the assessment of construct validity.

Although limited, the assessment of construct validity of the NEI has been more extensive than that of the SEI. Both instruments have been assessed for their relationship with traditional clinical measures of disease activity. Similar moderate levels of correlation were reported with pain (VAS), but a stronger correlation between the SEI and stiffness (VAS). Generally, low and non-significant correlation between both indices and laboratory based measures of disease activity and anthropometric measures have been reported. A limited range of scores was observed when the NEI was completed in an in-patient population (approximately 50% of available range). This was in contrast to the 95% coverage of score range observed in the BASDAI and the Bath-DAI in the same patient population, prompting the investigators to suggest that the NEI had a limited ability to represent disease activity in patients with a wide spectrum of disease (Garrett et al, 1994).

The DAI was compared to traditional measures of outcome employed in clinical practice and research. Dougadas et al (1988) wished to determine if the new index measured 'something other than (or the same thing as)' these instruments, but a priori hypothesised relationships were not stated. Low to moderate correlations were calculated with all clinical instruments (table 2.22). Low correlations were also calculated with the DFI, a measure of AS-specific functional disability, and with change in patient-assessed global health following physical therapy ($r = 0.18$)(table 2.22).

The face and content validity of the Bath-DAI was critisised by the original development team for its focus on pain and well-being and the failure to address important issues of AS disease activity, such as fatigue, and Garrett et al (1994) subsequently recommended the BASDAI. Evidence of the construct validity of the Bath-DAI is limited but a strong correlation with the BASDAI and with the DFI (functional disability) has been found ($r > 0.75$)(table 2.22). The Bath-DAI was also able to discriminate between an in-patient and an out-patient population ($p < 0.001$).

[a] Validity	NEI [1]	SEI [2]	DAI [3]	Bath DAI [4]	BASDAI [5]
Physical tests / signs	(n= 22) AS out-patients [13] No significant correlation with spinal range of movement (no detail)	(n= 26) AS out-patients [2] No significant correlation with FFD, MSI, OCP and FVC (n= 40) AS out-patients [14] No significant relationship with measures of spinal deformity	(n= 15) AS out-patients with axial disease[3] FFD r= 0.49 MSI r= -0.38 Chest expansion r= -0.46	-	-
Pain	(n= 19) AS out-patients [1] Pain severity (VAS) r= 0.67 Morning stiffness(VAS) r= 0.46 ESR r= -0.21 (Pain with Stiffness r= 0.71)	(n= 26) AS out-patients [2] Pain severity (VAS) r= 0.70 Morning stiffness(VAS)r= 0.70 (n= 40) AS out-patients [14] Pain severity(VAS)p< 0.05, ESRp< 0.0005	(n= 15) AS out-patients with axial disease[3, 12] Pain (VAS) r= 0.35 Morning stiffness(VAS) r= 0.27 Nigh-time awakenings r= 0.21	-	(n= 46) AS in-patients [5] Correlation between separate items of the index: Morning stiffness - quality with quantity r= 0.79 Fatigue with joint pain r= 0.34
Disability assessments	-	-	(n= 15) AS out-patients with axial disease[3] DFI r= 0.32	(n= 65) AS out-patients [4] DFI r= 0.79 (p< 0.001)	(n= 191) AS out-patients, postal survey [21] BASFI r= 0.59, DFI r= 0.57
Other	(n= 25) AS in-patients [5] Convert scores to 0-10 range for comparison with BASDAI: NEI mean value 1.96 (range 0-5.33) (n= 22) AS out-patients [13] No significant correlation with gender, age, disease duration or Ritchie Articular Index (n= 59) AS out-patients [27] Utility rating scale r=-0.49, Standard gamble r= 0.09 (n= 33) AS out-patients [15] PET - no statistically significant relationship	(n= 26) AS out-patients [2] No significant correlation with laboratory base measures of disease activity (n= 70) AS out-patients [26] SASS XR evaluation lumbar spine. Non-significant correlation	(n= 59) AS out-patients [27] Utility rating scale r=-0.44 Standard gamble r= 0.03 (n= 33) AS out-patients [15] Patient Elicitation Technique (PET) r= 0.10	(n= 46) AS in-patients and (n= 108) AS out-patients [5] Convert scores to 0-10 range for comparison with BASDAI: Bath DAI mean value: in-patients 5.13, out-patients 3.97 (p= 0.001) (range 0-.9.5) (n= 46) AS in-patients [5] BASDAI r= 0.75 (p< 0.001)	(n= 46) AS in-patients and (n= 108) out-patients [5] Mean value: in-patients 5.06, out-patients 4.00 (p= 0.05) (range 0.5-10.0) (n= 46) AS in-patients [5] Bath-DAI r= 0.75 (n= 200) AS out-patients[17]BAS-G r= 0.73 (n= 16) AS out-patients after cervical spine surgery, (n= 81)AS in-patients[18] Greater score post-surgical group (mean 5.56 vs 4.54) Self-report exercise, AS out-patients[19] Significant difference between non (n=915) and moderate (n= 1491) exercisers. No difference between non and intensive (n-393) exercisers (paired t-test) (n= 2979 AS out-patient [20] AS duration r= -0.05 AS out-patients (n= 149) axial disease, (n= 42) with peripheral involvement [21] ESR range r= 0.06 to 0.19, CRP r= 0.06 to 0.23

Table 2.22 Validity of AS-specific measures of disease activity.

[a] References defined in table 2.19. Abbreviations defined in Glossary.

Evidence for the construct validity of the BASDAI is limited. Strong correlation with the Bath-DAI and the BAS-G, an AS-specific measure of global health, and moderate to strong correlation with AS-specific measures of functional activity have been reported. Direct correlation with more traditional measures of disease activity, for example, patient reported pain or stiffness, or laboratory based assessment has recently been assessed and very low correlation with Erythrocyte Sedimentation Rates (ESR)(r= 0.06 - 0.19) and C-Reactive Protein (CRP)(r= 0.06 - 0.23) have been reported in patients with active disease (Spoorenberg et al, 1999b). The investigators suggest that the result supports the inability of the laboratory based measures (acute phase reactants) to reflect disease activity in AS. However, a strong correlation of change in BASDAI score with change in ESR has been reported following a drug trial of known efficacy (Maksymowych et al, 1998)(table 2.24).

Responsiveness

Evidence of the responsiveness of all instruments is shown in tables 2.23 and 2.24. Evidence suggests that the NEI is not responsive to change following physical therapy intervention. However, following a drug trial of known efficacy evidence suggests that the NEI and the assessment of pain severity (VAS) are both capable of discriminating between patients receiving the active drug or placebo (Mander et al, 1987). There is no evidence for the responsiveness of the SEI.

The strongest evidence of the responsiveness of the DAI has been reported in drug therapy trials, with evidence to suggest that the DAI is capable of detecting statistically significant change in disease activity following active treatment. The observed change in the DAI was often closely associated with change in other clinical variables (tables 2.23-2.24). There is little evidence to support the responsiveness of the DAI following physical therapy intervention.

There is limited evidence of the responsiveness of the Bath-DAI following physiotherapy intervention, but greater score improvement was calculated for the Bath-DAI (22.8%) than for the BASDAI (16.4%). The BASDAI has been assessed following both physiotherapy and drug therapy trials of known efficacy (tables 2.23-2.24). Although the evaluations of physiotherapy had limited follow-up periods, significant improvement in BASDAI score for active treatment groups for both treatment modalities have been recorded.

Study design	[a] Used in trial with known efficacy
NEI [1]	RCT: known active drug (NSARDs) and no drug; cross-over design, 2-week duration (n= 14) AS out-patients. NEI and Pain severity (VAS) significantly lower in patients taking NSARDs (NEI 10.5 vs 13.7; Pain 26.8 vs 31.7) [1]
	RCT: efficacy of addition of supervised therapy versus unsupervised (n= 144) AS out-patients. 6-week follow-up. Statistically significant change in score ($p< 0.001$) - but investigators suggest that result should not be considered of clinical significance due to the poor test-retest result (range r= 0.32-0.52) [8]
	Non-RCT: 3-week intensive in-patient physiotherapy AS rehabilitation programme (n= 47). Comparison of questionnaires completed day '8' and '18'(n= 30). Non-significant improvement in NEI score p= 0.35; % improvement not reported [5]
	RCT: efficacy of the addition of supervised group physiotherapy to unsupervised therapy (n= 144) AS outpatients. 9-month follow-up. Effect size calculated: low level of sensitivity for the NEI E= 0.31. Largest effect sizes calculated for Borg Scale (physical fitness - E 3.50), Smythe Technique (spinal ROM - E 0.76) and patient assessed improvement (VAS - E 0.76) [15]
	RCT: supervised therapy and unsupervised therapy (continuation) (n= 68) AS out-patients. 9-month follow-up. Non-significant difference between groups for NEI (t-test of change scores, $p> 0.05$) [28]
SEI [2]	-
DAI [3]	RCT: known active compound (Piroxicam) and placebo (n= 80) AS outpatients. Stepwise logistic regression to determine the best variable with which to differentiate between groups. Clinicians 'overall judgement' best variable ($p< 0.0001$); followed by DFI ($p< 0.001$); DAI, Pain (VAS), and hand to ground distance ($p< 0.01$) [3]
	RCT: to determine efficacy of Ximoprofen (placebo controlled) 2-weeks duration (n= 285) AS outpatients. Statistically significant improvement in all variables – a 'better' improvement in the Ximoprofen groups irrespective of dosage $p< 0.05$ (Dunnett's t-test) for the DAI, DFI and pain (VAS) [22]
	RCT: supervised therapy and unsupervised therapy (continuation) (n= 68) AS out-patients. 9-month follow-up. Significant difference between groups for DAI, global health (VAS) and HAQ-S (favouring continuation)(t-test of change scores, $p< 0.05$). [28]
Bath DAI [4]	Non-RCT: 3-week intensive in-patient physiotherapy AS rehabilitation programme (n= 47). Comparison of questionnaires completed day '0' and '18'(n= 30). Significant improvement in Bath DAI p= 0.002; 22.8% improvement [5]
BASDAI [5]	Non-RCT: 3-week intensive in-patient physiotherapy AS rehabilitation programme (n= 47). Comparison of questionnaires completed day '0' and '18'(n= 30). Significant improvement in BASDAI p= 0.009; 16.4% improvement [5]
	Non-RCT: 2-week intensive in-patient physiotherapy AS rehabilitation programme (n= 236). Comparison of questionnaires completed day '0' and '14'. Significant improvement in BASDAI $p< 0.001$; 18% improvement [23]
	RCT: efficacy of amitriptyline in reduction of pain, fatigue and stiffness (placebo controlled)(n= 81) AS in-patients (rehabilitation programme). Baseline and 2-week completion. Statistically significantly greater improvement in BASDAI scores for active treatment compared to placebo (23% improvement vs 10%)($p< 0.001$) [24]
	Longitudinal 'open analysis' of pamidronate in treatment of AS (n= 8) out-patients (active disease)(3-month intervention and 3 / 6 / 9-month follow-up). BASFI, BASDAI, BASMI and laboratory assessments. Comparison between baseline and follow-up - a significant improvement in BASDAI (p= 0.03), BASMI and ESR scores observed at 6-months (paired t-test)[25]
	RCT: to determine efficacy of NSAID (placebo controlled) 2-weeks duration (n= 473) AS out-patients. Statistically significant improvement in all variables for treatment group when compared to placebo group (BASDAI p= 0.0002) [11]

Table 2.23 Responsiveness results I (trials of known efficacy) of AS-specific measures of disease activity.

[a] References defined in table 2.19. Abbreviations defined in Glossary.

Strategy applied	[a] Correlation of scale change with changes in other measures
NEI [1]	RCT: efficacy of addition of supervised therapy versus unsupervised (continuation of trial) (n= 68) AS out-patients. 9-month follow-up. . Low correlation between patient-assessed change in global health (VAS) and NEI r= 0.02 (p = 0.43) [16]
	Non-RCT: Longitudinal evaluation group physiotherapy (n= 68) AS out-patients. 9-month follow-up. Correlation of change in patient-assessed global health (VAS) r= 0.02 [16]
SEI [2]	-
DAI [3]	RCT: known active compound (Piroxicam) and placebo (n= 80) AS outpatients. 2-weeks treatment. Correlation between change in DAI score and 'overall judgement' of treatment efficacy (clinician & patient) – rated nil / moderate / good / very good. Change in score of DAI highly correlated with 'clinician overall judgement' regardless of treatment [3]
	RCT: supervised therapy and unsupervised therapy (continuation) (n= 68) AS out-patients. 9-month follow-up. Statistically significant difference between groups seen in DAI, DFI, patients global health (VAS) and HAQ-S (favouring continuation)(t-test of change scores, p< 0.05). Non-significant change in DAI between baseline and 9-months [25]
	RCT: efficacy of the addition of supervised group physiotherapy to unsupervised therapy (n= 144) AS outpatients. 9-month follow-up. Comparison of change between measures (change = follow-up minus baseline result). Significant correlation between DFI and MACTAS / PET (r= 0.34, p< 0.005) [5]
	RCT: efficacy of addition of supervised therapy versus unsupervised (continuation of trial) (n= 68) AS out-patients. 9-month follow-up. Low correlation between patient-assessed change in global health (VAS) and DAI (r= 0.18) [16]
	Non-RCT: Longitudinal evaluation group physiotherapy (n= 68) AS out-patients. 9-month follow-up. Correlation of change in patient-assessed global health (VAS) r= 0.18 [16]
Bath DAI [4]	-
BASDAI [5]	Longitudinal 'open analysis' of pamidronate in treatment of AS (n= 8) out-patients (active disease)(3-month intervention and 3 / 6 / 9-month follow-up). Significant association between change in BASDAI score between baseline and 9-months and change in ESR (r= 0.79, p= 0.032) [25]

Table 2.24 Responsiveness results II (correlation of scale change with changes in other measures) of AS-specific measures of disease activity.

[a] References defined in table 2.19. Abbreviations defined in Glossary.

Acceptability

All developers have reported good levels of acceptability of respective instruments by patients, although no developer defines acceptability or indicates how the level of acceptability was derived. Completion rates have not been reported.

Feasibility and application

The NEI and DAI have become the most widely applied clinician completed AS-specific measures of disease activity. However, few published articles have referred to the use of the NEI and DAI since 1998. The BASDAI is the only recommended patient-completed instrument and recent years have seen a rapid increase in published articles referring to its use, although the majority of articles have been published by members of the development team (n= 10/13 articles). The original DAI was developed in the French language but procedures for translation in to English have not been described. The BASDAI has been translated into French (Claudepierre et al, 1997) and Swedish (Waldner et al, 1999).

Commentary

Although all five instruments purport to measure AS-specific disease activity, the nature of disease activity and the approach adopted by each instrument differs. The NEI, SEI and DAI all require clinician administration and record a patient response to movement or palpation of specific sites. However, the NEI and SEI describe the pathological involvement of entheses and the DAI records the involvement of articular sites in the disease process. The Bath-DAI and BASDAI adopt a very different approach, requiring a patient-completed evaluation of symptoms associated with AS disease activity. Table 2.25 provides a summary of the data synthesis. The purpose and conceptual base of all instruments has practical appeal and clinical relevance. However, no instrument provides a clear definition of disease activity and inferences from results rely heavily on the definition used (Spoorenberg et al, 1999b).

The NEI is described as both a measure of disease activity and disease severity. These terms communicate different aspects of disease status and are not interchangeable (Symmons, 1995). Inflammation is a key feature of AS and the measurement of disease activity aims to quantify the inflammatory process at one point in time whilst reflecting the extent of associated reversibility. For example, the

	AS-specific disease activity [a]												
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
	Scale	Number of items	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
NEI [1]	ordinal	15	Not specified - clinical practice, research	Trained clinician – 3-minutes	1	1	9	++	+	++	+	++	+
SEI [2]	ordinal	22	Not specified - clinical practice, research	Trained clinician 3-4 minutes?	1	2	2	0	0	+	+	0	0
DAI [3]	ordinal	10	Not specified - clinical practice, research	Trained clinician - 3-5 minutes	1	1	8	++	++	++	++	++	+
Bath DAI [4]	3 x ordinal 1 x VAS	4	Not specified - clinical practice, research	Self-administered 1 minute	1	1	0	+	+	+	+	+	+
BASDAI [5]	VAS	6	Not specified - clinical practice, research	Self-administered 1 minute	1	5	6	++	++	+	+	++	++

Table 2.25 AS-specific measures of disease activity - summary of data evaluation (after McDowell and Newell, 1996)

[a] References defined in table 2.19. Abbreviations defined in Glossary.

Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

evaluation of pain, stiffness or fatigue. The measurement of disease severity has a more prognostic role and may describe the structural and irreversible change that occurs as a result of disease process (Symmons, 1995).

Although a pragmatic approach to instrument development was adopted by all authors, there is a lack of detail supporting item development. Without a clear definition of instrument purpose and conceptual base an appreciation of item development and instrument validity is limited. Knowledge of the pathoanatomy of AS was required to support item generation for the NEI, SEI and DAI, but item development for all instruments has been dominated by experts in clinical rheumatology, with little information to support their experience with AS or the criteria by which items were selected. Despite the requirement for a patient response to a clinicians palpation, there is no indication of patient involvement in developing the NEI, SEI or DAI. Only the developers of the BASDAI indicate that patients were involved in the development process but do not detail patient selection, characteristics and input.

The developers of the BASDAI (Garrett et al, 1994) critisised the NEI and the Bath-DAI for their failure to adequately represent the wide spectrum of AS disease activity. Although the BASDAI contains a range of items relating to pain (spinal and peripheral joints), fatigue and tiredness, bodily discomfort (representative of enthesitis) and morning stiffness (severity and duration), there remains a focus on pain. Although the items have good face validity, due to the failure to define disease activity, a lack of conceptual base and a paucity of detail about item development, an appraisal of content validity is limited. Appraisal of content validity is similarly limited for other study instruments. However, when comparing the NEI and SEI, greater detail about item development and selection for the NEI is provided, and although containing fewer items the entheses cover a wider range of sites, thus benefiting from greater face validity.

Notwithstanding the wide range of retest intervals and the lack of consideration for disease stability shown by many authors, evidence of the test-retest reliability of the DAI, Bath-DAI and BASDAI seems satisfactory and supports application in group evaluation (> 0.70). However, the DAI does not clarify if joint movements are active, active-assisted or passive. This is an important methodological issue when evaluating

the impact of pain on joint movement and requires standardisation to support consistent administration. The reliability of the Bath-DAI and BASDAI supports application in individual evaluation (> 0.90). Despite the attempt to standardise administration the test-retest and inter-observer reliability of the NEI is low, suggesting significant error in repeat administrations. There is no evidence for the reliability of the SEI. The BASDAI is the only instrument with evidence of both test-retest and internal consistency reliability. Greater evidence of test-retest reliability, with appropriate retest periods and the external evaluation of condition stability is required for all instruments. The assessment of internal consistency reliability is also required.

There is limited evidence to support the validity of all instruments. Results are weak due to the failure of all investigators to hypothesise expected relationships between constructs a priori (McDowell and Jenkinson, 1996). A 'gold standard' measure of disease activity in AS is lacking (Symmons, 1995; Spoorenberg et al, 1999b), and the quantification of disease activity has traditionally relied upon the measurement of subjective domains such as pain or stiffness, or on laboratory based assessment. The direct relationship between these traditional measures and all instruments, except for the Bath-DAI and the BASDAI, have been reported. Generally, moderate relationships with subjective measures, but small correlation with laboratory based measures were found, providing support for some authors to suggest that the ESR is a poor reflection of AS disease activity (Mander et al, 1987; Symmons, 1995). However, a strong correlation between change in BASDAI score and change in ESR following a drug therapy trial has been reported (Maksymowych et al, 1998).

A direct comparison between two AS-specific measures of disease activity has only been reported between the BASDAI and Bath-DAI (strong relationship). The NEI has been applied in the same study as the Bath-DAI and the BASDAI and scores were converted to 0-10 to help comparisons. The NEI was critisised for the low mean value (1.96) and for covering just over 50% of the available score range in comparison to wider range covered by the Bath-DAI and BASDAI (both 95%)(Garrett et al, 1994). However, the correlation relationship between instruments was not reported. Comparison of all instruments with other measures of disease activity and with more established generic measures of HRQL have not been

undertaken, and are recommended with a priori hypothesised relationships stated to provide a wider appreciation of the constructs addressed.

All instruments, apart from the SEI, have some evidence of responsiveness, but accumulated evidence suggests that the NEI is not responsive to change. There is good evidence for the responsiveness of the DAI following drug therapy and limited evidence for the responsiveness of the Bath-DAI following trials of physical therapy. There is increasing evidence for the responsiveness of the BASDAI following drug therapy and physical therapy. A greater percentage improvement in Bath-DAI score (22.8%) than calculated for the BASDAI (16.4%) was reported following an evaluation of physical therapy (Garrett et al, 1994). The authors suggest that items relating to pain are very responsive to intensive physiotherapy but not to change in disease activity as a whole. Consequently the responsiveness of the Bath-DAI reflects the greater focus on pain and patient's well-being (Garrett et al, 1994). However, three of the six items of the BASDAI also address pain and discomfort.

All developers have reported good levels of acceptability, but completion rates have not been reported. All instruments are easy to understand, although the lack of detail relating to the SEI makes interpretation difficult. The patient-completed instruments are quick to complete (1-2 minutes). The clinician-completed instruments require longer for administration, all taking a similar time to complete (3-5 minutes), but none require special equipment. All instruments have been administered in both in-patient and out-patient populations, and the BASDAI has also been administered in a postal survey (Jones et al, 1996b).

Conclusion

All instruments have been developed integrating clinical expertise with a theoretical appreciation of disease activity in AS. Based on the limited published evidence at the time of instrument selection there was acceptable evidence in support of the development and measurement properties of the DAI, Bath-DAI and the BASDAI. However, the developers of the Bath-DAI recommend adoption of the BASDAI (Garrett et al, 1994). The comparative study (Chapter 4) required a patient-completed instrument and so the BASDAI was the only available option.

There is inadequate published evidence to support the development and measurement properties of the NEI and SEI and they are not recommended as disease-specific measures of disease activity. These inadequacies relate to the lack of clarity in instrument development (SEI) and, where reported, the poor levels of reliability and responsiveness.

2.6.3 Anthropometric measures

This section describes the evaluation of the anthropometric measures listed in table 2.10. The measures will be considered in terms of the evaluation of cervical mobility, chest expansion, thoracolumbar mobility, fingertip to floor distance, lumbar mobility and spinal posture. One composite index, the Bath Ankylosing Spondylitis Metrology Index (BASMI; Jenkinson et al, 1994a) is also evaluated.

Purpose and conceptual base of anthropometric measurement

The identified anthropometric measures represent the available range of movement, or postural status, at a specific area of the axial skeleton. Limitation in spinal mobility is an objective feature of the pathological process of AS, and in 1968 recommendation was made for spinal movement in AS to be assessed in 'three planes of movement' (Bennett and Wood, 1968 cited by Moll et al, 1971).

Study population

There is a lack of clarity about the patient population involved in most studies. Many investigators do not clearly describe the population sampling frame from which patients were selected beyond the distinction of in-patient or out-patient status and the method of patient selection was generally unclear. Many investigators do not adequately describe patient variables and disease characteristics. Also, the training or experience of observers recording the measurements is often inadequately detailed, an important consideration when assessing the feasibility and reliability of measurement.

Cervical mobility

General description

Thirteen approaches for the assessment of cervical mobility have been identified: five rotation, four lateral flexion and four flexion / extension. A general description, and methodology is summarised in table 2.26. All approaches require special equipment

[a] Cervical mobility assessment			Methodology			
	Earliest application - AS	Identified need	Land markings and special equipment	Starting position	Application	Time
Rotation						
Large protractor [1]	Roberts et al (1988) [7]	Clinically important parameter [8]	Large protractor O-degree center point placed over episternal notch. 90-degree lines aligned with antra of the ears	Sitting	Maximal rotation. Total change = addition of left and right rotation	-
Universal goniometer [2]	Hidding et al (1993b) [9]	-	Univeral goniometer placed horizontally on crown of head. Pivot centered on occiput with a line through the nose	Sitting	Maximal rotation. Total change = addition of left and right rotation	-
Simple inclinometer [3]	O'Driscoll et al (1978)	-	Simple inclinometer placed on patients forehead	Supine	Maximal rotation. Total range right and left measured	-
Myrin inclinometer [4]	Viitenan et al (1992) [4]	-	Myrin inclinometer. Positioning not reported	Supine	Total rotation form max leftward to max rightward	-
Tape measure [5]	Viitenan et al (1998) [5]	Quick and easy methodology [5]	Distance between tip of chin and cornoideus process of claviculae measured in neutral with a tape measure	Sitting	Maximal rotation - decrease in distance between landmarks on both sides	-
Lateral flexion						
Universal goniometer [2]	Pile et al (1991) [8]	-	Pivot of universal goniometer centered on sternal notch. Vertical arm passes in line with nose. Second arm perpendicular to horizontal plane	Sitting	Max lateral flexion to both sides. Total or separate range calculated (degrees)	-
Tragus to ACJ [2]	Pile et al (1991) [8]	-	Distance between tragus of the ear and acromioclavicular joint (ACJ) measured in neutral with tape measure	Sitting	Max lateral flexion to both sides. Decrease in distance between landmarks, shoulders relaxed	-
Simple inclinometer [3]	O'Driscoll et al (1978) [6]	-	Simple inclinometer positioned on vertex of head	Sitting	Total range between maximal lateral flexion to both sides recorded (degree)	-
Myrin inclinometer [5]	Viitenan et al (1998) [5]	-	Simple inclinometer positioned on vertex of head	Sitting	Total maximal lateral bending of the head to both sides (degrees)	-
Flexion / extension						
Universal goniometer [2]	Pile et al (1991) [8]	-	Universal goniometer centered laterally over ACJ. Vertical arm aligned with tragus (zero). Horizontal arm perpendicular to this.	Sitting	Total or separate range of full flexion /extension measured relative to zero (degrees)	-
Occiput to C7 [6]	Calcraft (1974) [6]	-	Distance between occipital protuberance and 7th cervical process (C7) measured in neutral with a tape measure	Sitting	Total flexion and total extension measured separately (centimetres)	-
Chin to chest [6]	Calcraft (1974) [6]	-	Distance between tip of chin and chest (suprasternal notch) measured in neutral with a tape measure	Sitting	Total flexion and total extension measured separately (centimetres)	-
Simple inclinometer [3]	O'Driscoll et al (1978) [3]	-	Spirit inclinometer placed on crown of head	Sitting	Total range of flexion and extension measured (degrees)	-

Table 2.26 General description and structure of anthropometric assessments of cervical mobility. [a] References defined in table 2.28. Abbreviations defined in Glossary.

and goniometers, inclinometers and tape measures have been adopted by different authors for the evaluation of each range of movement.

Certain methodologies describe a composite movement. For example, combined left and right rotation is described for the protractor (Calcraft et al, 1974), the goniometer (American Academy of Orthopaedic Surgeons (AAOS, 1965) and the inclinometer (O'Driscoll et al, 1978; Viitanen et al, 1992). However, the separate movements of right and left rotation are described by the tape measure methodology (Viitenan et al, 1998)(table 2.26.) Difficulty in finding the neutral starting position is often cited in support for composite measurement, but the inability to determine if a problem exists with left or right movement is support for separate analysis of each movement.

Reliability

Evidence of the reliability of all measures of cervical mobility is shown in table 2.27. High levels of inter and intra-observer reliability have been reported for all approaches for the measurement of cervical rotation (inter-observer r> 0.84, ICC 0.89)(intra-observer r> 0.96, ICC> 0.89), although inter-observer reliability has not been calculated for measurement with a goniometer.

More moderate levels of inter-observer reliability was reported for the assessment of cervical lateral flexion using a goniometer or tape measure (r< 0.76). High levels of inter-observer reliability were reported by Viitenan et al (1998) for a selection of anthropometric measurements including cervical lateral flexion (tape measure: AAOS, 1965), but the specific result was not reported. Evidence for the reliability of the simple inclinometer in patients with AS is limited. Pile et al (1991) report that 90% of measurements of cervical lateral flexion repeated by one or more trained observer will vary by up to 10 degrees when measured with a goniometer, and by up to 2.5cm when measured with a tape measure, and that changes of less than this could be attributed to methodological variation and not due to true change in cervical mobility.

Low levels of inter-observer reliability were calculated for the assessment of cervical flexion and extension with a goniometer (AAOS, 1965; Pile et al, 1991), but several studies report moderate to high levels following assessment with a tape measure (Calcraft et al, 1974; Bellamy et al, 1991a; Viitenan et al, 1995a,b). There is no

[a] Cervical mobility	Inter-observer	Intra-observer
Rotation		
Large protractor [1]	5 trained observers, 2 repetitions (n= 10) AS out-patients, minimal AS activity [8] r= 0.84 – 0.90. (degrees) mean 5.3^0 – 5.9^0, median 5.0^0, 90^{th} percentile 11.4^0 – 12.5^0	1 trained observer, 2 observations over 1-hour (n= 20: 10 AS out-patients, 10 normal) [7] AS r= 0.96, normal r= 0.83. Significant warm-up effect cervical rotation and FFD.
		5 trained observers, 2 repetitions (n= 10) AS out-patients (minimal disease activity) [8] right mean 3.9^0, median 0.0, 90th percentile 10.0^0; left mean 7.0^0, median 5.0^0, 90th percentile 15.0^0
Universal goniometer [2]	-	1 trained observer, 48-hour retest (n= 19) AS out-patients [9] r= 0.96
Simple inclinometer [3]	3 trained observers, blinded, 1 observation (n= 20) AS in-patients [10] r= 0.99	3 trained observers, blinded, 24-hour retest (n= 20) AS in-patients [10] r = 0.98
		1 trained observer, 3-day retest (n= 5) AS in-patients (paired t-test) [20] Non-significant difference (1.0^0+/- 8.5^0)
Myrin inclinometer [4]	2 trained observers (n= 39) AS in-patients [11] ICC 0.96	2 trained observers, 2-hour retest (n= 38) AS in-patients [11] ICC 0.97
	2 trained observers (n= 39) AS in-patients [12] ICC 0.96	1 trained observer, 72-hour retest (n= 52) AS out-patients [5] ICC range 0.89 - 0.98
	2 trained observers (n= 52) AS out-patients [5] ICC range 0.89 - 0.98 (cervical mobility)	
Tape measure [5]	2 trained observers (n= 52) AS out-patients [5] ICC range 0.89 - 0.98 (cervical mobility)	1 trained observer, 72-hour retest (n= 52) AS out-patients [5] ICC range 0.89 - 0.98
Lateral flexion		
Universal goniometer [2]	5 trained observers, 2 repetitions (n= 10) AS out-patients (minimal disease activity) [8] r= 0.68 - 0.74. Probability that underlying assessor means differ (p< 0.05)	5 trained observers, 4 repetitions (n= 10) AS out-patients (minimal disease activity) [8] 90^{th} percentile 8.8 – 10.0 degrees
Tragus to ACJ [2]	5 trained observers, 2 repetitions (n= 10) AS out-patients (minimal disease activity) [8] r= 0.69 - 0.76. Probability that underlying assessor means differ (p< 0.05)	5 trained observers, 4 repetitions (n= 10) AS out-patients (minimal disease activity) [8] 90^{th} percentile 2.0 – 2.5 cm
	2 trained observers (n= 52) AS out-patients [5] ICC range 0.89 - 0.98 (cervical mobility)	
Simple inclinometer [3]	-	? observers (n= 59) AS out-patients, ? retest [29] r= 0.86 - 0.87
Myrin inclinometer [5]	2 trained observers (n= 52) AS out-patients [5] ICC range 0.89 - 0.98 (cervical mobility)	-
Flexion / extension		
Universal goniometer [2]	5 trained observers, 2 repetitions (n= 10) AS out-patients, minimal AS activity [8] Flexion r=0.21, Extension r= 0.59. Probability that underlying assessor means differ (p< 0.05)	5 trained observers, 4 repetitions per patient (n= 10) AS out-patients, minimal AS activity [8] flexion 90^{th} percentile 22.3 degree, extension 90^{th} percentile 12.0 degrees
Occiput to C7 [6]	6 trained observers, random order, (n= 7) male AS out-patients [13] Methodology standardisation: Flexion- pre r= 0.85- 0.99, post r= 0.94 - 0.98. Extension- pre r= 0.91-1.00, post- r= 0.91- 0.99	-
Chin to chest [6]	6 trained observers, random order, (n= 7) male AS out-patients [13] Methodology standardisation: Flexion- pre r= 0.85- 0.99, post r= 0.94 - 0.98. Extension- pre r= 0.91-1.00, post- r= 0.91- 0.99	2 trained observers, 2-hour interval (n= 38 AS in-patients) [11] ICC 0.95
	2 trained observers (n= 39) AS in-patients [11] ICC 0.72	
	2 trained observers (n= 39) AS in-patients [12] ICC 0.72	
Simple inclinometer [3]	-	? observers (n= 59) AS out-patients, ? retest [29] flex r= 0.77, extension r= 0.81

Table 2.27 Reliability of anthropometric assessments of cervical mobility. [a] References defined in table 2.28. Abbreviations defined in Glossary.

evidence of reliability for the measurement of flexion or extension with an inclinometer in patients with AS.

Validity

Limited evidence of the validity of all measures of cervical mobility is shown in table 2.28. There is no evidence in support of the validity of cervical lateral flexion measured with a goniometer or hydrogoniometer, or for flexion / extension measured with a goniometer or a tape measure (occiput to C7 distance). Most frequently, the relationship between measurement techniques and spinal radiographic change have been assessed, with small to moderate relationships reported. The strongest relationships were with cervical rotation (spirit inclinometer)(r= 0.59)(Kennedy et al, 1995) and cervical lateral flexion (tape measure)(r= 0.58)(Viitenan et al, 1998). The lowest correlation was between cervical rotation (Myrin inclinometer) and cervical spine age-adjusted radiographic change (Viitenan et al, 1995a). Small correlation between cervical lateral flexion measured with a tape measure (AAOS, 1965) or with a Myrin inclinometer and radiographic change in the thoracic and lumbar spine, and in the sacroiliac joints have also been reported (Viitenan et al, 1998).

Moderate correlation between cervical rotation (protractor) and the HAQ-S (functional disability) (Daltroy et al, 1990), and between rotation (simple inclinometer) and sub-sections of the LDQ (functional disability) (Abbott et al, 1994) have been reported. A small to moderate correlation between flexion / extension (simple inclinometer) and sub-sections of the LDQ have also been reported, the strongest relationship with the sub-section addressing neck mobility. Moderate to strong correlation between the protractor assessment of rotation and other anthropometric assessments have also been reported. No other investigators have reported the relationship between cervical mobility and other anthropometric assessments.

Responsiveness

Limited evidence of the responsiveness of all measures of cervical mobility is shown in table 2.29. Evidence suggests that cervical rotation measured with a goniometer or either form of inclinometer may be responsive to change following physical therapy. There is no evidence of responsiveness for rotation measured with a protractor or tape measure. Evidence suggests that cervical lateral flexion, flexion or extension are not

[a] Cervical mobility	Physical tests / signs	Disability assessment	Other
Rotation			
Large protractor [1]	(n= 46) AS out-patients [14] Smythe technique r= 0.77, FFD r= -0.69, Chest expansion r= 0.54	(n= 46) AS out-patients [14] HAQ-S (global) r= -0.57 (n= 216) AS out-patients[23] HAQ-S r= 0.50, DFI r= 0.32, HAQ r= 0.38, AIMS2 r= 0.38	(n= 46) AS out-patients [14] Age r= -0.45, AS duration r= -0.43
Universal goniometer [2]	-	(n= 42) AS in-patients [22] Turkish DFI r= 0.67	-
Simple inclinometer [3]	-	(n= 45)AS out-patients [16] Sections of LDQ range r= 0.36- 0.72	(n= 53) AS in-patients [17] Cervical XR r= 0.59, p< 0.001 (n= 11) AS in-patients [24] No relationship with subjective or objective sleep
Myrin inclinometer [4]	-	-	(n= 73) male AS in-patients[12] Age-adjusted AS-specific lumbar XR [(18)] r= -0.26
Tape measure [5]	-	-	(n= 52) AS out-patients [5] AS-specific XR [(12)] Cervical r= 0.42, thoracic r= 0.15, lumbar r= 0.15, sacroiliac joint r= 0.25
Lateral flexion			
Universal goniometer [2]	-	(n= 42) AS in-patients [22] Turkish DFI r= 0.64	-
Tragus to ACJ [2]	-	-	(n= 52) AS out-patients [5] AS-specific XR [(12)] Cervical r= 0.58, thoracic r= 0.27, lumbar r= 0.33, sacroiliac joint r= 0.43
Simple inclinometer [3]	-	-	(n= 11) AS in-patients [24] No relationship with subjective or objective sleep
Myrin inclinometer [5]	-		(n= 52) AS out-patients [5] AS-specific XR [(12)] Cervical r= 0.57, thoracic r= 0.18, lumbar r= 0.28, sacroiliac joint r= 0.36
Flexion / extension			
Universal goniometer [2]	-	-	-
Occiput to C7 [6]	-	-	-
Chin to chest [6]	-	-	(n= 73) male AS in-patients [5] AS-specific lumbar XR [(12)] of cervical flexion r= 0.37
Simple inclinometer [3]	-	(n= 45)AS out-patients [16] LDQ sections: flexion range r= 0.32 - 0.72; extension range r= 0.30 - 0.72	(n= 11) AS in-patients [24] Objective sleep r= 0.60. No relationship with subjective sleep

Table 2.28 Validity of anthropometric assessments of cervical mobility. Abbreviations defined in Glossary.

References: [1] Cheshire (1957), [2] American Assoication of Orthopaedic Surgeons (AAOS)(1965), [3] O'Driscoll et al (1978), [4] Viitenan et al (1992), [5] Viitenan et al (1998), [6] Calcraft et al (1974), [7] Roberts et al (1988), [8] Pile et al (1991), [9] Hidding et al (1993b), [10] Jenkinson et al (1994a), [11] Viitenan et al (1995b), [12] Viitenan et al (1995a), [13] Bellamy et al (1991a), [14] Daltroy et al (1990), [15] Hidding & van der Linden (1995), [16] Abbott et al (1994), [17] Kennedy et al (1995), [18] Dale & Vinje (1985). [19] Kirwan et al (1993), [20] Helliwell et al (1996), [21] Roberts et al (1989), [22] Dalyan et al (1999), [23] Ward & Kusiz (1999), [24] Jamieson et al (1995), [25] Corkhill et al (1990), [26] Hidding et al (1994a), [27] Hidding et al (1993a), [28] Bakker et al (1994b), [29] Creemers et al (1996).

[a] Cervical mobility	Used in trial with known efficacy ([b] correlation of change score with change in other variable)
Rotation	
Large protractor [1]	-
Universal goniometer [2]	RCT: Home exercise vs home plus group exercise (n= 144) AS out-patients. 9-month follow-up (treatment period). Non-significant difference between groups for cervical rotation [27,28]
	Non-RCT: 6-week before and after evaluation of supervised individual physiotherapy (n= 144) AS out-patients. Statistically significant change in cervical rotation ($p< 0.01$)[9]
	RCT: Supervised vs unsupervised therapy (n= 68) AS out-patients. 9-month follow-up. Non-significant difference between groups, and non-significant change for cervical rotation and all ROM[26]
	Non-RCT[b]: Longitudinal evaluation group physiotherapy (n= 67) AS out-patients. 9-month follow-up. Correlation with change in patient assessed global health (VAS) r= 0.05 [15]
	RCT: Sulphasalasine vs placebo (n= 62) AS out-patients. 48-week follow-up.No significant difference between groups (t-test). No significant difference over time for either group (paired t-test)[25]
Simple inclinometer [3]	Non-RCT: 1) comparison of measurements 3-weeks and immediately prior to rehabilitation. Statistically significant difference in cervical rotation ($p< 0.01$). 2) 3-week rehabilitation (n= 25) AS in-patients. Measurement day 0 and day-18. Statistically significant change in all ROM ($p< 0 .001$). 3) 3-month follow-up (n= 11). Statistically significant change in cervical rotation ($p< 0.001$) [3]
	Non-RCT: Retrospective analysis rehabilitation programme (n= 48) AS in-patients. 3-weeks- significant improvement (19.0^0)($p< 0.0001$). 5-years- significant deterioration (8.0^0)($p= 0.012$). Cervical rotation may be of benefit in the evaluation of short and long term change[21]
	RCT: 3 physiotherapy regimes (n= 44) AS out-patients. Assess pre-randomisation, post-treatment (6-weeks), 6-months. Significant differences between groups immediately after treatment only for cervical rotation ($F= 3.73$, $p= 0.03$), pain and stiffness (VAS)($F= 10.88$, $p= 0.001$). No significant difference between groups at 6-months for cervical rotation ($F= 1.95$, $p= 0.16$) [16, 20]
Myrin inclinometer [4]	Non-RCT: Retrospective analysis 3-week rehabilitation (n= 165)AS in-patients. Significant change ($p< 0.001$) Cervical rotation 22.6% change (range Schober index 12.4% to FFD 36.6%)[4]
	Non-RCT: 3-week rehabilitation (n= 79) male AS in-patients. Assess day-1 and 18. Effect size (ES) cervical rotation 0.26 (range Tlflexion 0.23 to TLrotation and FFD 0.73) [12]
Tape measure [5]	-
Lateral flexion	
Universal goniometer [2]	-
Tragus to ACJ [2]	-
Simple inclinometer [3]	Non-RCT: 1) comparison of measurements 3-weeks and immediately prior to rehabilitation (n= 9). No statistically significant difference in lateral flexion. 2) 3-week rehabilitation (n= 25) AS in-patients. Measurement day 0 and 18. Statistically significant change in all ROM ($p< 0 .001$). 3) 3-month follow-up (n= 11). Statistically significant change in lateral flexion ($p< 0.01$) [3]
	RCT: Sulphasalasine vs placebo (n= 89) AS out-patients, established disease. 3-year follow-up. Non-significant improvement all measurements. Cervical lateral flexion unchanged both groups [19]
Myrin inclinometer [5]	-
Flexion / extension	
Universal goniometer [2]	RCT: Sulphasalasine vs placebo (n= 62) AS out-patients. 48-week follow-up.No significant difference between groups (t-test). No significant difference over time for either group (paired t-test)[25]
Occiput to C7 [6]	-
Chin to chest [6]	Non-RCT: Retrospective analysis - 3-week rehabilitation programme (n= 505) AS in-patients. Statistically significant change in all measures ($p< 0.001$).% change chin to chest (flexion) 21.7% [4]
	Non-RCT: 3-week rehabilitation (n= 79) male AS in-patients. Measurements day-1 and day-18. Effect size (ES): chin to chest (flexion) ES= 0.31 [12]
Simple inclinometer [3]	Non-RCT: 1) comparison of measurements 3-weeks and immediately prior to rehabilitation (n= 9). Non-significant difference in flexion/extension ($p> 0.10$). 2) 3-week rehabilitation (n= 25) AS in-patients. Measurement day 0 and 18. Statistically significant change in ROM ($p< 0 .001$). 3) 3-month follow-up (n= 11). Non-significant change in flexion/extension ($p> 0.10$) [3]

Table 2.29 Responsiveness of anthropometric assessments of cervical mobility. [a] References defined in table 2.28. Abbreviations defined in Glossary.

responsive following physical therapy. There is limited data suggesting a lack of responsiveness for cervical mobility assessment following drug therapy.

Acceptability

Acceptability to patients has not been reported.

Feasibility and application

Cervical rotation is the most frequently measured range of cervical mobility assessed in patients with AS identified in the review. All items of equipment have the benefit of being applicable to the measurement of any range of cervical mobility, but only the tape measure and universal goniometer can be used for other anthropometric assessments, improving the acceptance in clinical practice. They are also the cheapest and most portable items.

Chest expansion

General description

Three approaches for the assessment of chest expansion have been clearly described in the literature, all using a tape measure but adopting different patient starting positions and body land-markings. All land marks are easily identified and a general description and methodology is summarised in table 2.30.

Reliability

Evidence of the reliability of chest expansion is shown in table 2.31. Low levels of inter-observer ($r < 0.53$), and high levels of intra-observer reliability ($r > 0.91$) have been reported for chest expansion measured at both the 4th intercostal space and at the xiphisternum. A high level of inter-observer reliability, improving following standardisation of technique, has been reported for measurement at nipple level (Bellamy et al, 1991a).

Validity

Evidence of the validity of chest expansion is shown in table 2.32. There is no evidence of the validity of measurement at nipple level (Hart et al, 1963). Small to moderate correlation between chest expansion (4th intercostal space) and other anthropometric measures have been reported by several investigators. Very small levels of correlation were reported with age, disease duration, laboratory based

Chest expansion			Methodology			
	Earliest application - AS	Identified need	Land markings and special equipment	Position	Application	Time
Nipple level[1]	Hart et al 1963[1]	-	Circumferential girth of torso at nipple level, measured with a tape measure	Standing, unclad to waist	Measurement taken at height of inspiration and expiration - difference calculated	-
4th intercostal space [2]	Moll & Wright 1972 [2]	Clinically important parameter	Circumferential girth of torso at level of 4th intercostal space, measured with a tape measure	Standing, unclad to waist, hands on head and arms flexed in frontal plane	Measurement taken at height of inspiration and expiration - difference calculated	-
Xiphisternum[3]	Tomlinson et al 1986 [3]	-	Circumferential girth of torso at level of xiphisternum, measured with a tape measure	Standing, unclad to waist	Measurement taken at height of inspiration and expiration - difference calculated	-

Table 2.30 General description and structure of anthropometric assessments of chest expansion. Abbreviations defined in Glossary.

References: [1] Hart et al (1963), [2] Moll & Wright (1972), [3] Tomlinson et al (1986), [4] Viitenan et al (1995a), [5] Viitenan et al (1995b), [6] Fisher et al (1990), [7] Bellamy et al (1991a), [8] Creemers et al (1996), [9] Pile et al (1991), [10] Daltroy et al (1990), [11] Hidding et al (1993b), [12] Moll & Wright (1973), [13] Franssen et al (1986), [14] Adrichem & van der Korst (1973), [15] Abbott et al (1994), [16] Dale & Vinje (1985), [17] Hidding & van der Linden (1995), [18] Averns et al (1996a), [19] Averns et al (1996b), [20] Viitenan et al (1992), [21] Taggart et al (1996), [22] Clegg et al (1996), [23] Roberts et al (1989), [24] Helliwell et al (1996), [25] Battle-Gualda et al (1996), [26] Dalyan et al (1999), [27] Hidding et al (1993a), [28] Roberts et al (1988), [29] Hidding et al (1994a), [30] Bakker et al (1994b).

[a] Chest expansion	Inter-observer	Intra-observer
Nipple level[1]	6 trained observers, random order, 2-day retest period, (n= 7) male AS out-patients [7] Methodology standardisation: pre- r= 0.86, post- r= 0.94	-
	? observers (n= 52) AS out-patients, ? type of reliability [8] r= 0.86	
4th intercostal space [2]	2 trained observers (n= 39) male AS in-patients [4] ICC 0.53 (range Cexp 0.53 to FFD 0.98)	1 trained observer (n= 33) AS out-patients [5] CV (%) 14.0
	2 trained observers (n= 39) male AS in-patients [5] ICC 0.53 (range Cexp 0.53 to FFD 0.98)	2 trained observers (n= 38) AS in-patients [6] ICC 0.93 (range 0.93 to OWD 0.99)
Xiphisternum[3]	5 trained observers, 2 repetitions per patient (n= 10) AS out-patients (minimal disease activity)[9] r= 0.15. Significant probability that underlying assessor means are different (p= 0.018)	1 trained observer (n= 44) AS outpatients [10] r= 0.92
		5 trained observers, 4 repetitions (n= 10) AS out-patients (minimal disease activity) [9] mean 0.6, median 0.5, 90th percentile 1.2 cm
		1 observer, 2-day retest (n= 19) AS out-patients [11] Cexp r= 0.91
		1 trained observer, 4 reps over 1-hour, reps 3 and 4 compared (n= 10) AS out-patients [28] r= 0.95
		1 trained observer, 3-day retest (n= 5) AS in-patients (t-test) [24] No significant difference

Table 2.31 Reliability of anthropometric assessment of chest expansion. [a] References defined in table 2.30. Abbreviations defined in Glossary.

[a] Chest expansion	Physical tests / signs	Pain measures	Disability assessment	Other
4th intercostal space [2]	(n= 85) AS out-patients [12] lumbar flexion r=0.06, LLF r=0.56, lumbar extension r=0.50 (n= 33) male AS out-patients [13] Vital capacity r=0.59, LFI [14] r=0.56 (n=33) AS out-patients [6] MSI r= 0.60, vital capacity r= 0.71, thoracic kyphosis r= 0.16	(n= 33) male AS out-patients [13] Thoracic pain (VAS) r=-0.21	-	(n= 85) AS out-patients [12] Age r= -0.28, AS duration r= -0.29 (n= 33) male AS out-patients[13] AS duration r=-0.38. AS activity, ESR and IgA r < 0.20. (n=30) AS out-patients [6] AS duration r= 0.14, ESR r= 0.13 (n=73) male AS in-patients[4] Age-adjusted change lumbar XR [16] r= 0.53
Xiphisternum[3]	-	-	(n= 44) AS outpatients[10] HAQ-S (global) r= 0.33 (n= 42) AS out-patients[15] LDQ sections range r= 0.34 to 0.54 (n= 42) AS out-patients[26] Turkish DFI r= 0.57	(n= 61) AS out- patients [18] SASSS XR lumbar spine r= -0.39 (n= 42) AS out- patients [19] Smoking and outcome. Non-significant difference in Chest expansion between smokers and non-smokers

Table 2.32 Validity of anthropometric assessment of chest expansion. [a] References defined in table 2.30. Abbreviations defined in Glossary.

[a] Chest expansion	Used in trial with known efficacy
4th intercostal space [2]	Non-RCT: Retrospective analysis - 3-week rehabilitation (n= 505) AS in-patients. Statistically significant change all measures (p< 0.001). Cexp 31.3% (range Schober index 12.4% - FFD 36.6%) [20]
	Non-RCT: 3-week rehabilitation (n= 79) male AS in-patients. Measurements day 1 and day-18. Effect size (ES): Cexp ES= 0.42. Most responsive TLrot and FFD (ES 0.73). Least TL flexion (ES 0.23) [4]
	Non-RCT: 3-week rehabilitation (n= 141) AS in-patients. End of course – significant improvement all measures (p< 0.001). No improvement in chest expansion at 15-months follow-up [5]
	RCT: 3 active drugs (n= 90) AS out-patients, active disease. 26-week follow-up. No significant change in chest expansion between treatments or over time [21]
	RCT: Sulphasalasine and placebo (n= 264) AS out-patients, active disease. 36-week follow-up. No statistically significant change for Cexp in treatment or control group, or between groups (p= 0.55) [22]
	RCT: 2 x NSAID (no placebo)(n= 310) AS out-patients. Baseline, 15/30/60/90 days follow-up. Cexp improvement (0.6mm to 0.8mm +/- 0.1mm (p< 0.01). Non-significant difference between groups[25]
Xiphisternum[3]	Non-RCT: 3-week rehabilitation (n= 180) consecutive AS in-patients. Measurements day 0 and day-18. Statistically significant improvement in all variables (p< 0.001). Chest expansion p< 0.0025 [3]
	Non-RCT: Retrospective analysis - 3-week physiotherapy, 5-year follow-up (n= 52) AS in-patients. Short term: significant improvement in Cexp (p< 0.0001). Long term: Non-significant deterioration in Cexp (p= 0.82). No significant difference between disease duration groups (p> 0.10).Cexp may be of benefit in the evaluation of short term change, but not for long term change[23]
	RCT: 3 physiotherapy regimes (n= 44) AS out-patients. Assess pre-randomisation, post-treatment (6-weeks), 6-months. No significant difference between groups at 6-weeks and 6-months for Cexp (F= 0.81, p= 0.46). Independent of treatment group and irrespective of disease severity, Cexp had improved in majority of patients at 6-months. Significance of change from baseline not determined[15]
	RCT: 3 physiotherapy regimes (n= 44) AS out-patients. Assess 2,4 and 6-months post-treatment. No significant difference in Cexp between groups at all follow-up. But majority of patients had improved in movement at 6-months. Significance of change from baseline not determined [24]
	Non-RCT:6-week before/after evaluation of supervised individual physiotherapy (n= 144) AS out-patients.Non-significant change chest expansion (p=0.06)Significant change cervical rotation (p<0.01)[11]
	RCT: Home exercise vs home plus group exercise (n= 144) AS out-patients. 9-month follow-up (treatment period). Non significant difference between groups in chest expansion or cervical rotation [27,30]
	RCT: Supervised vs unsupervised therapy (n= 68) AS out-patients. 9-month follow-up. Non-significant difference between groups, and non-significant change for chest expansion and all ROM[29]
	Non-RCT$_a$: Longitudinal evaluation group physiotherapy (n= 67) AS out-patients. 9-month follow-up. Correlation with change in patient assessed global health (VAS) r= 0.31 [17]

Table 2.33 Responsiveness of anthropometric assessment of chest expansion. [a] References defined in table 2.30. Abbreviations defined in Glossary.

measures of disease activity and radiographic change in the lumbar spine. The strongest correlation was reported with the measurement of vital capacity, a reflection of lung function (Franssen et al, 1986). Small to moderate levels of association between measurement at the xiphisternum level and the HAQ-S (Daltroy et al, 1990), sub-sections of the LDQ (Abbott et al, 1994), and correlation of change score with patient reported change in global health following physical therapy (Hidding and van der Linden, 1995) (table 2.33) have been reported.

Responsiveness

Evidence of the responsiveness of chest expansion is shown in table 2.33. There is no evidence of the responsiveness of measurement at nipple level (Hart et al, 1963). Evidence following three-week rehabilitation programmes suggests that chest expansion (4th intercostal space and xiphisternum) may be responsive to short term change, but that improvement is not retained at a longer term follow-up. However, evidence following a six-week study of physiotherapy suggests a non-significant improvement (Hidding et al, 1993b). Evidence following drug therapy trials suggests that chest expansion is not a responsive measurement.

Acceptability

Acceptability to patients has not been reported.

Feasibility and application

Chest expansion is widely measured in clinical trials and in clinical practice (Bellamy et al, 1998; 1999). However, further evidence in support of clear methodologies is limited due to inadequate descriptions and referencing by investigators.

Thoracolumbar flexion

General description

Six methods for the assessment of thoracolumbar flexion have been described. All use a tape measure, but identify different bony landmarks. A general description and methodology is summarised in table 2.34.

Reliability

Evidence of the reliability of thoracolumar flexion is shown in table 2.35. There is no evidence for two approaches (Hyde, 1980; Averns et al, 1996a), but high levels of

Thoracolumbar mobility			Methodology			
Earliest application - AS		Identified need	Land markings and special equipment	Starting position	Application	Time
Flexion						
C7 to iliac crest [1]	Pile et al (1991) [9]	-	7th cervical C7 spinous process (C7) and horizontal line between superior aspects of iliac crests [9]	Standing	Maximal spinal flexion. Distraction distance between landmarks measured with tape	-
C7 to 10cm proximal to LSJ [2]	Calcraft (1974) [2]	-	C7 and a point 10cm proximal to lumbosacral junction (LSJ) [2]	Standing	Maximal spinal flexion. Distraction distance between landmarks measured with tape. Difference = thoracic flexion[2]	-
C7 to SC junction [3]	Hyde (1980) [3]	-	C7 and sacrococcygeal (SC) junction (not defined) [3]	Standing	Maximal spinal flexion. Distraction distance between landmarks measured with tape. Repeat in full extension.	-
C7 to dimples of Venus [4]	Armstrong et al (1984) [4]	-	C7 and a line joining the dimples of Venus (posterior superior iliac spine - PSIS) [4]	Standing	Overall spinal movement. Distance between landmarks measured. Not clear if flexion, extension or both assessed	-
C7 to S1 [5]	Viitenan et al (1992) [5]	-	C7 and first sacral spinous process (S1) [5]	Standing	Maximal spinal flexion. Distraction distance between landmarks measured with tape.	-
C7 to L5 [6]	Averns et al (1996a) [6]	-	C7 and 5th lumbar spinous process (L5) [6, 10]	Standing	Maximal spinal flexion. Distraction distance between landmarks measured with tape.	-
Rotation						
Rotation frame [7]	Viitenan et al (1993) [7]	Lack of adequate methodology[7]	Frame - semicircular degree scale anterior to patient. 43cm needle indicator attached at xiphisternum (belt), horizontal and positioned at zero. Pelvis manually fixed.	Sitting on stool	Maximal trunk rotation both sides (left, then right).- thoracolumbar rotation (degrees) between xiphisternum and S1	-
Pavlaka method [8]	Viitenan et al (1999) [8]	Quick and easy methodology[8]	Tape measure distance between the xiphisternum and first sacral spinous process (S1)	Standing	Maximal trunk rotation, in inspiration. Contralateral increase in distance between skin marks on both sides	-

Table 2.34 General description and structure of anthropometric assessment of thoracolumbar mobility. Abbreviations defined in Glossary.

References: [1] American Academy of Orthopaedic Surgeons (AAOS)(1965), [2] Calcraft et al, (1974), [3] Hyde (1980), [4] Armstrong et al (1984), [5] Viitenan et al (1992), [6] Averns et al (1996a), [7] Viitenan (1993), [8] Viitenan et al (1999), [9] Pile et al (1991), [10] Averns et al (1996b), [11] Bellamy et al (1991a), [12] Viitenan et al (1995b), [13] Viitenan et al (1995a), [14] Viiitenan et al (1995c), [15] Dalyan et al (1999), [16] Taylor et al (1991a), [17] Dale & Vinje (1985), [18] Viitenan et al (1995d), [19] Battle-Gualda et al (1996), [20] Taylor et al (1991c), [21] Bland & Altman (1986)

[a] Thoracolumbar	Inter-observer	Intra-observer
Flexion		
C7 to iliac crest [1]	5 trained observers, 2 repetitions (n= 10) AS out-patients, minimum AS activity [9] r= 0.73 Non-significant difference between assessor means	5 trained observers, 4 repetitions, random order (n= 10) AS out-patients, minimum AS activity [9] mean 0.6cm, median 0.5cm, 90th percentile 1.5cm
C7 to 10cm proximal to LSJ [2]	6 trained observers, random order (n= 7) male AS out-patients [11] Methodology standardisation: pre r= 1.00, post r= 1.00	-
C7 to SC junction [3]	-	-
C7 to dimples of Venus [4]	6 trained observers, random order (n= 7) male AS out-patients [11] Methodology standardisation: pre r= 0.93, post r= 0.97	-
C7 to S1 [5]	2 trained observers, 2-hour interval (n= 39) AS in-patients [12] ICC 0.91	2 trained observers, 2-hour interval (n= 38) AS in-patients [12] ICC 0.95
C7 to L5 [6]	-	-
Rotation		
Rotation frame [7]	2 trained observers, 1-2 hour retest (n= 39) randomly selected AS in-patients [7] ICC 0.89 Schober Index ICC 0.88, FFD ICC 0.98	2 trained observers. 24-hour retest (n= 39) randomly selected AS in-patients [7] ICC 0.93 Schober Index ICC 0.95, FFD ICC 0.97
	2 trained observers, 2-hour retest (n= 39) randomly selected AS in-patients [12] ICC 0.89 (ICC range Cexp 0.53 to FFD 0.98)	2 trained observers, 2-hour retest (n= 39) randomly selected AS in-patients [12] ICC 0.93 (ICC range TLrot and Cexp 0.93 to OWD 0.99)
	2 trained observers, same day retest (n= 39) randomly selected male AS in-patients [13] ICC 0.89 (ICC range Cexp 0.53 to FFD 0.98)	2 trained observers, 2-hour retest (n= 39) randomly selected AS in-patients [14] Inter and intra-observer reported together, not specific to measurement (range r= 0.53 to r= 0.99).
	2 trained observers, 2-hour retest (n= 39) randomly selected AS in-patients [14] Inter and intra-observer reported together, not specific to measurement (range r= 0.53 to r= 0.99)	
Pavlaka method [8]	2 trained observers, random order, 48-hour retest, (n= 52) consecutive male AS in-patients [8] Limits of agreement reported [21] Suggest good reliability. Not discuss clinical implication of result	1 trained observer, 48-hour retest, (n= 52) consecutive male AS in-patients [8] Limits of agreement reported [21] Suggest good reliability. Not discuss clinical implication of result

Table 2.35 Reliability of anthropometric assessment of thoracolumbar mobility.

[a] References defined in table 2.34. Abbreviations defined in Glossary.

inter-observer reliability are reported for the remaining methods. The highest reliability is for the measurement C7 to 10cm proximal to the lumbo-sacral junction, reflecting thoracic flexion (Calcraft et al, 1974). High levels of intra-observer reliability have been reported for C7 to iliac crest (AAOS, 1965) and C7 to S1 (Viitenan et al, 1992).

Validity

Evidence of the validity of thoracolumar flexion is shown in table 2.36. Validity testing has only been reported for the measurement C7 to the sacrococcygeal junction (Hyde, 1980), and C7 to L5 (Averns et al, 1996a). Moderate to high correlation with radiographic evaluation of the spine and moderate correlation between C7 to sacrococcygeal junction and other anthropometric measures was reported (Taylor et al, 1991a).

Responsiveness

Limited evidence of responsiveness for thoracolumar flexion is shown in table 2.37. Accumulated evidence following physiotherapy suggests poor responsiveness over the active treatment period (three-weeks), and no improvement over the longer term follow-up (Viitenan et al, 1992, 1995a,b). Further evidence suggests that thoracolumbar flexion is not responsive following drug therapy and does not discriminate between patients receiving active or placebo treatment (Taylor et al, 1991c).

Acceptability

Acceptability to patients has not been reported.

Feasibility and application

Thoracolumbar flexion is often referred to in published articles. However, the methodology for five of the six techniques is very similar and standardisation of technique is recommended to allow data to be combined to further support evaluation.

Thoracic rotation

General description

Two approaches for the assessment of thoracic rotation have been described, both developed by the same author (table 2.34). The first requires a 43cm long indicator

[a] Thoracolumbar	Physical tests / signs	Disability assessment	Other
Flexion			
C7 to iliac crest [1]	-	(n= 42) AS out-patients [15] Turkish DFI r= 0.62	-
C7 to 10cm proximal to LSJ [2]	-	-	-
C7 to SC junction [3]	(n= 32) AS out-patients [16] Correlation with Cexp, FFD and OWD. Range r= 0.60 – 0.65	-	(n= 32) AS out-patients [16] SASSS XR lumbar spine, total spinal flexion, Cexp, FFD and OWD: range r= 0.37 – 0.47
C7 to dimples of Venus [4]	-	-	-
C7 to S1 [5]	-	-	-
C7 to L5 [6]	-	-	(n= 53) AS out-patients [6] SASSS XR evaluation lumbar spine r= -0.72 (n= 42) AS out-patients [10] Significant difference between smokers and non-smokers p< 0.001
Rotation			
Rotation frame [7]	(n= 135) AS in-patients (90 male)[7] Disease duration before age-adjustment r=-0.55. Age-standardised correlation r=-0.17	-	(n= 135) AS in-patients (90 male) [7] Descriptive data only - XR change lumbar spine [(17)] - strong negative correlation. AS-specific lumbar changes - strong negative correlation. XR evidence of sacroiliitis - no significant correlation. XR degenerative change lumbar spine only (n= 9) equivalent TLrot range to patients without XR change (n= 30) (n= 73) randomly selected male AS in-patients [13] Age-adjusted AS-specific XR change lumbar spine [(17)] r= -0.41 (range Cexp r= -0.15 to Schober Index r= -0.66) (n= 73) male AS in-patients [18] AS duration r= -0.47, age-adjusted r= -0.40 (range FFD r= 0.17 to TLrot r= -0.47; age-adjusted Cexp r= -0.10 to TLrot r= -0.40) (n= 151) AS in-patients (male 108) [14] XR lumbar spine r= -0.41(Cexp r= -0.27 to Schober index r= -0.71). XR SIJ r= -0.27 (FFD r= 0.20 to Schober index r= -0.48) (n= 52) consecutive male AS in-patients [8] AS-specific XR change in spine [(17)] range thoracic spine r= -0.13 to lumbar spine r= -0.29
Pavlaka method [8]	-	-	(n= 52) consecutive male AS in-patients [8] AS-specific XR change in spine[(17)] range thoracic spine r= -0.32 to cervical spine and SIJ r= -0.36).

Table 2.36 Validity of anthropometric assessment of thoracolumbar mobility.

[a] References defined in table 2.34. Abbreviations defined in Glossary.

[a] Thoracolumbar	Used in trial with known efficacy
Flexion	
C7 to iliac crest [1]	RCT: 2 x NSAID (no placebo)(n= 310) AS out-patients. Baseline, 15/30/60/90 days follow-up. TLflex 11-14% improvement (0.4cm+/- 0.2cm ($p< 0.01$). Non-significant difference between groups [19]
C7 to 10cm proximal to LSJ [2]	-
C7 to SC junction [3]	RCT: Sulphasalasine versus placebo (n= 40) AS out-patients. 12-mth follow-up. Non-significant difference between baseline and follow-up for either group. Non-significant difference between groups for TLflex. Pain (VAS) the only variable to discriminate between groups at 12-months [20]
C7 to dimples of Venus [4]	-
C7 to S1[5]	Non-RCT: Retrospective analysis of 3-week rehabilitation programme (n= 505) AS in-patients. Assessments pre and post treatment. Statistically significant improvement all variables ($p< 0.001$). TLF 15% change (range Schober index 12.4% to FFD 36.6%) [5]
	Non-RCT: 3-week rehabilitation (n= 141) AS in-patients. 15-month follow-up. Significant improvement short term all variables (baseline to end of course)($p< 0.001$). No significant change in thoracolumbar flexion over long term (15-months). Significant long-term improvement only for cervical rotation, FFD and Astrand Fitness Index ($p< 0.001$) [12]
	Non-RCT: 3-week rehabilitation (n= 79) male AS in-patients. Assessed day-1 and 18. Effect size (ES) range thoracolumbar flexion (C7-S1) 0.23 to thoracolumbar rotation 0.73 and FFD 0.71 [13]
C7 to L5 [6]	-
Rotation	
Rotation frame [7]	Non-RCT: 3-week rehabilitation (n= 79) male AS in-patients. Assessed day-1 and 18. Effect size (ES) Thoracolumbar rotation 0.73 (range thoracolumbar flexion (0.23) to thoracolumbar rotation)[13]
	Non-RCT: 3-week rehabilitation (n= 141) AS in-patients. Assessed baseline, 3-weeks, 15-months. Change from baseline calculated. Short term (end of programme) – significant improvement all measures ($p< 0.001$). Long term (15-months) - significant improvement only in cervical rotation, FFD and a fitness index [14]
Pavlaka method [8]	-

Table 2.37 Responsiveness of anthropometric assessments of thoracolumbar mobility.

[a] References defined in table 2.34. Abbreviations defined in Glossary.

needle to be strapped to the patients chest, who sits behind a large semi-circular degree scale (Viitenan, 1993). The second, the Pavlaka method (Viitenan et al, 1999) requires only a tape measure, and is the only new anthropometric measure to be identified following extension of the initial review period. A general description and methodology is summarised in table 2.34.

Reliability

Evidence of the reliability of thoracolumbar rotation is shown in table 2.35 and suggests good intra and inter-observer reliability for both methodologies. However, the correlation coefficient is not reported for the Pavlaka method, the developers reporting only the 95% limits of agreement (Bland and Altman, 1986). The limits of agreement describe a statistical range of error that may relate to clinical acceptability (Bruton et al, 2000), but the developers do not discuss the implications of the result.

Validity

Evidence of validity of thoracolumbar rotation is shown in table 2.36. Small to moderate correlation between thoracolumbar rotation and radiographic evaluation of the spine have been reported, the strongest relationship with AS-specific change in the lumbar spine. A stronger, but still moderate relationship between the Pavlaka method and AS-specific spinal change in the same patient group has been reported (Viitenan et al, 1999).

Responsiveness

Evidence of the responsiveness of thoracolumbar rotation is shown in table 2.37. Two studies suggest that the measurement is responsive over the short term following intensive physiotherapy, although the improvement was not maintained over the long term follow-up (Viitenan et al, 1995a,b).

Acceptability

Acceptability to patients has not been reported.

Feasibility and application

Use of the rotation frame in the assessment of AS has only been published by the instrument developers. The frame may not be readily accepted in clinical practice or

research. The same development team have recently proposed the Pavlaka method which only requires a tape measure and may prove to be more feasible.

Fingertip to floor distance

General description

Four approaches for measuring fingertip to floor distance (FFD) following anterior spinal flexion have been described. All require special equipment ranging from a tape measure (Miller et al, 1984) to a Portable Spinal Mobility Scale (PSMS)(Stokes et al, 1988). A general description and methodology is summarised in table 2.38.

Reliability

Evidence of the reliability of FFD is shown in table 2.39. High levels of inter and intra-observer reliability were reported for all approaches (inter-observer $r > 0.94$, ICC > 0.98; intra-observer $r > 0.98$, ICC > 0.97), except for the PSMS where a significant difference between observers was reported ($p = 0.04$)(Stokes et al, 1988). Inter-observer reliability of FFD following the technique described by Kippers and Parker (1987) has not been calculated in AS patients.

Validity

Evidence of the validity of FFD is shown in table 2.40. Moderate correlation between FFD (mounted ruler) and sub-sections of the LDQ (Abbott et al, 1994), and a small correlation between FFD (patient on a stool; Kippers and Parker, 1987) and the HAQ-S (Daltroy et al, 1990) have been reported. Small correlation between the measurement of FFD with a mounted ruler (Tomlinson et al, 1986) or tape measure (Miller et al, 1984) and radiographic change in the lumbar spine have been reported (Averns et al, 1996a; Viitenan et al, 1995a). A significant difference between the measurement of FFD using a tape measure or the PSMS (Stokes et al, 1988) has been reported.

Responsiveness

Evidence of the responsiveness of FFD is shown in table 2.41. Evidence is not available for the approach described by Kippers and Parker (1987). The measurement of FFD with a mounted ruler (Tomlinson et al, 1986) or a tape measure (Miller et al, 1984) have both been used in the longitudinal evaluation of physiotherapy in AS, with evidence to suggest a statistically significant change over both short and long term

Fingertip to floor				Methodology		
Earliest application - AS		Identified need	Land markings and special equipment	Starting position	Application	Time
Tape measure, standing on floor [1]	Miller et al (1984) [1]	Reliable and valid measure	Tip of the longest finger. Tape measure (cm)	Standing on floor	Maximal forward flexion, fingertips towards the floor. Tape measure extended vertically to tip of longest finger	13.3 +/-7.4 seconds range 2-30 seconds[1]
Vertically mounted ruler, standing on floor [2]	Tomlinson et al (1986) [2]	-	Tip of longest finger. Fabricated lucite rule (cm) mounted on a floor stand [6], or a vertical mounted ruler (cm) with a sliding marker [7]	Standing on floor	Patient bends forward, pushing sliding marker down vertically mounted ruler, whilst maintaining knee extension. Position of slide recorded (cm)	-
Ruler / tape measure, standing on stool [3]	Roberts et al (1988) [5]	Relationship between FFD and vertebral mobility	Tip of third digit of hand [3] 20cm stool with scale (cm) fixed anteriorly (median plane) [3], or a 23cm stool and a tape measure (cm) [5] Stance position marked on stool	Standing on stool	Patient bends forward, fingers reaching towards the floor. Distance between fingertip and floor is measured against ruler. If reach beyond toes, values recorded as negative.	-
Portable Spinal Mobility Scale [4]	Stokes et al (1988) [4]	-	Fingertips. Portable Spinal Mobility Scale (PSMS) and a styrofoam block (30x30x10cm)	Standing on block	Patient bends forward, fingers reaching towards the floor. Horizontal arm of PSMS raised to touch fingertips. If unable to reach the horizontal arm whilst on the block, block is removed and 10cm added to score.	-

Table 2.38 General description and structure of anthropometric assessment of fingertip to floor distance (anterior flexion). Abbreviations defined in Glossary.

References: [1] Miller et al (1984), [2] Tomlinson et al (1986), [3] Kippers & Parker (1987), [4] Stokes et al (1988), [5] Roberts et al (1988), [6] Roberts et al (1989), [7] Pile et al (1991), [8] Bellamy et al (1991a), [9] Viitenan et al (1995b), [10] Creemers et al (1996), [11] Viitenan et al (1995a), [12] Dale & Vinje (1985), [13] Abbott et al (1994), [14] Averns et al (1996a), [15] Averns et al (1996b), [16] Daltroy et al (1990), [17] Viitenan et al (1992), [18] Clegg et al (1996), [19] Kragg et al (1990), [20] Kragg et al (1994).

[8] Fingertip to floor	Inter-observer	Intra-observer
Tape measure, standing on floor [1]	6 trained observers, random order, 2-day retest (n= 7) male AS out-patients [8] Methodology standardisation: pre r= 1.00, post r= 1.00	? observers, ? retest (n=59) AS out-patients [10] r= 0.94
	2 trained observers (n= 39) AS in-patients [9] ICC 0.98 (range Cexp 0.53 to FFD)	
Vertically mounted ruler, standing on floor [2]	5 trained observers, 2 repetitions (n= 10) AS out-patients, minimal AS activity [7] r= 0.95 Non-significant difference between assessor means	5 trained observers, 4 repetitions (n= 10) AS out-patients, minimal AS activity [7] mean 2.8cm, median 1.0cm, 90th percentile 9.0cm
	6 trained observers, random order, 2-day retest (n= 7) male AS out-patients [8] Methodology standardisation: pre r= 0.97, post r= 1.00	
Ruler / tape measure, standing on stool [3]	-	1 trained observer, 4 reps over 1-hour, reps 3 and 4 compared (n= 10) AS out-patients [5] r= 0.98 Also report a significant warm-up effect for FFD and cervical rotation (p< 0.002)
Portable Spinal Mobility Scale [4]	5 trained observers, random order, (n= 5) AS out-patients [4] Observers differed significantly overall (p= 0.04)	5 trained observers, random order, (n= 5) AS out-patients [4] ICC r= 0.99

Table 2.39 Reliability of anthropometric assessment of fingertip to floor distance (anterior flexion). [8] References defined in table 2.38. Abbreviations defined in Glossary.

[8] Fingertip to floor	Physical tests / signs	Disability assessment	Other
Tape measure, standing on floor [1]	(n= 3) AS out-patients [1] Smythe (sum) r= 0.86	-	(n= 73) male AS in-patients[11] Age-adjusted XR change lumbar spine [(12)] FFD r= 0.18 (range Cexp r= -0.15 to Schober index r= -0.66)
	(n= 5) AS out-patients [4] Significant difference between PSMS and FFD (tape measure) p= 0.0006.		
Vertically mounted ruler, standing on floor [2]	-	(n= 45) AS out-patients [13] LDQ sections: Reaching up/neck movement r= 0.62, Mobility r= 0.43, Bending down r= 0.42, Posture r= 0.35	(n= 61) AS out-patients [14] SASSS XR lumbar spine r= 0.35
			(n= 42) AS out-patients [15] Significant difference between smokers and non-smokers FFD p< 0.02
Ruler / tape measure, standing on stool [3]	(n= 33) healthy young adults [3] Trunk flexion r= -0.85, hip flexion r= -0.79, vertebral flexion r= 0.10	(n= 46) AS out-patients [16] HAQ-S (total) r= 0.32	(n= 46) AS out-patients [16] pain (VAS) r= 0.22, stiffness (VAS) r= 0.27
	(n= 44) AS out-patients [16] Smythe (sum) r= -0.71, cervical rotation r= -0.69, chest expansion r= -0.37		
Portable Spinal Mobility Scale [4]	(n= 5) AS out-patients [4] Significant difference between PSMS and FFD (tape measure) p= 0.0006. PSMS measured average 2.1cm shorter distance	(n= 46) AS out-patients [16] Age r= 0.42 , AS duration r= 0.38, height r= 0.07	-
	(n= 5) AS out-patients [4] PSMS measurement of OWD r= 0.38		

Table 2.40 Validity of anthropometric assessment of fingertip to floor distance (anterior flexion). [8] References defined in table 2.38. Abbreviations defined in Glossary.

[a] Fingertip to floor	Used in trial with known efficacy
Tape measure, standing on floor [1]	Non-RCT: Retrospective analysis 3-week rehabilitation (n= 505) AS in-patients. Measurements pre and post treatment. Statistically significant change all measures ($p< 0.001$). Greatest % change in FFD 36.6% (excluding patients able to touch floor)(range Schober test 12.4 to FFD) [17]
	Non-RCT: 3-week rehabilitation (n= 141) AS in-patients. 15-month follow-up. Significant improvement short term all variables (baseline to end of course)($p< 0.001$). Significant improvement long term (baseline to 15-months) for FFD ($p< 0.001$), cervical rotation and Astrand Fitness Index [9]
	RCT: Sulphasalasine and placebo, 36-week follow-up (n= 264) AS out-patients, active disease. No statistically significant improvement in FFD for treatment or control group (FFD $p= 0.99$). FFD unable to discriminate between groups ($p= 0.58$) [18]
Vertically mounted ruler, standing on floor [2]	Non-RCT: 3-week rehabilitation (n= 180) AS consecutive in-patients. Assess day 0 and 18. Mean change in scores. Statistically significant improvement all variables ($p< 0.001$). FFD $p<0.0005$ [2]
	Non-RCT: Retrospective analysis 3-week rehabilitation (n= 52) AS consecutive in-patients. Significant improvement in FFD over 3-weeks: mean improvement 5.4cm ($p< 0.0001$). Significant improvement in FFD over long term (5-year follow-up): mean improvement 2.9cm ($p< 0.047$). Improvement not influenced by disease duration. Although statistically significant change, authors suggest that absolute value may not have clinical significance (largest mean change 2.9cm) [6]
Ruler / tape measure, standing on stool [3]	-
Portable Spinal Mobility Scale [4]	RCT: Home exercise and education vs control group (n= 53) AS out-patients. 4-month follow-up. Baseline and follow-up scores compared. Highly significant improvement in FFD in experimental group (-8.3cm). Deterioration in control group (+2.0cm). FFD able to discriminate between groups ($p< 0.001$). Highly significant improvement also for patient-assessed functional activity (TADLQ) in experimental group (3.92), with minimal change in control group (-0.19). No significant difference in other variables for either group (OWD, Pain) [19]
	RCT: Cross-over continuation of RCT home exercise and education vs control group [19] (n= 53) AS out-patients. Further 4-month follow-up (8-months total). No statistically significant change in FFD (3.6cm ,$p=0.14$), although authors indicate a change in range of 3cm to be clinically important (expert opinion) [20]

Table 2.41 Responsiveness of anthropometric assessments of fingertip to floor distance (anterior flexion).

[a] References defined in table 2.38. Abbreviations defined in Glossary.

follow up. However, Roberts et al (1989) suggest that the long term change in range may not have clinical significance (mean change 2.89cm). Limited evidence suggests that FFD is not responsive following a placebo-controlled dug trial (Clegg et al, 1996). There is also limited evidence to support the responsiveness of the PSMS following physical therapy, and to suggest that the measurement can distinguish between patients receiving more active physiotherapy (Kragg et al, 1990).

Acceptability

When comparing acceptance of several mobility measures patients (n= 3) indicated that FFD (tape measure) was the most convenient, followed by the Modified Schober Index (Macrae and Wright, 1969), the Smythe technique (Miller et al, 1984) and finally goniometer assessment (Miller et al, 1984). In addition, many patients with AS cite fear of falling as a concern (Chapter 3) and it is unlikely that they would accept a methodology requiring them to bend forward whilst standing on a stool (Kippers and Parker, 1987; PSMS - Stokes et al, 1988), challenging their ability to remain balanced as opposed to reflecting trunk mobility.

Feasibility and application

Fingertip to floor distance is widely referred to in published articles and more than 50% of clinicians usually or always measure it during routine clinical evaluation (Bellamy et al, 1998, 1999). The most frequently cited methodologies use a tape measure or mounted ruler with the patient standing on the floor.

Lumbar mobility

Ten methods for the assessment of lumbar mobility have been identified: three flexion, one extension, two flexion and / or extension, and four lateral flexion. All approaches require special equipment, but tape measures have been adopted for the evaluation of most movements (Macrae and Wright 1969; Miller et al, 1984; Moll et al, 1971). A general description and methodology is summarised in table 2.42.

a. Lumbar flexion

General description

Two methodologies adopt a tape measure: the Schober Index (von Schober, 1937) and the Modified Schober Index (MSI)(Macrae and Wright, 1969). The third, the Lumbar Flexion Index (LFI) requires a flexible rule (Adrichem and van der Korst, 1973).

[a] Lumbar mobility assessment			Methodology			
Earliest application - AS		Identified need	Land markings and special equipment	Starting position	Application	Time
Flexion						
Schober 10cm Index [1]	Macrae & Wright (1969) [2]	-	Lumbosacral junction: hands width proximal [1], or 10cm [2]. Initially marked with fingers and thumb and visually estimated [1]. Subsequently measured [2]	Standing	Maximal flexion. Distance between landmarks estimates [1], or measured with a tape measure [2]	-
Modified Schober Index [2]	Macrae & Wright (1969) [2]	Rapid, simple convenient method [2]	Identify lumbosacral junction: mark skin 10cm proximal and 5cm distally (15cm).	Standing	Maximal flexion. Measure distance between landmarks (tape measure)	30 seconds to 2 minutes [6]
Lumbar Flexion Index [3]	Franssen etal (1986) [11]	Simple, reliable, quantitative measure [3]	Identify L5 level. Second mark 15cm proximal using flexible rule.	Standing	Maximal flexion. Measure distance between landmarks with flexible rule	-
Extension						-
Dunham Spondylometer [4]	Dunham (1949) [4]	-	Spondylometer - upper cushion of protractor base level with iliac crests. Proximal knob placed over vertebra prominens (C7)	Standing	Readings taken in standing, full flexion, and extension (degrees)	
Plumb-line extension [5]	Moll et al (1972b) [5]	Simple, reliable, accurate method [5]	Proximal - intersection of xiphisternum with coronal line. Distal - intersection high point iliac crest with coronal line. Brass plumbline suspended-20cm thread held at proximal mark, point coincide distal mark	Standing	Maximal extension. Distance traversed by pointer marked on skin of flank and measured (tape, cm)	-
Smythe technique [6]	Miller et al (1984) [6]	-	Lumbosacral junction identified. 3 proximal marks at 10cm intervals (tape). Measurement: upper (U), mid (M), lower (L) or sum (S) of segments.	Standing	Distance between marks re-measured in max flexion (standing) and extension (from prone)	2 - 3 minutes [6]
Lateral flexion						-
Skin distraction [7]	Moll et al (1972a) [7]	Reliable and objective measure [7]	Proximal- intersection of xiphisternum with coronal line. Distal- intersection of high point iliac crest with coronal line. Distance measured (tape, cm)	Standing	Max side flexion. Approximated distance of marks subtracted from neutral distance. Mark distraction favoured in obsese patients.	-
Fingertip to fibula [8]	Bellamy et al (1991a) [12]	-	In standing, arms by side, mark drawn on homolateral thigh at tip of 3rd finger. Distance between mark and head of fibula measured (tape, cm)	Standing	Max side flexion. Distance between tip of 3rd finger and head of fibula measured. Difference calculated (both sides)	
Fingertip lateral thigh [9]	Creemers et al (1996) [13]	-	In standing, arms by side, mark drawn on homlateral thigh at tip of 3rd finger. Height of patient asked (not measured).	Standing	Max side flexion. Position of tip of 3rd finger marked. Difference calculated (both sides). Result calculated as % of patients height. Mean value - 13%, lower limit 10%	-
Fingertip to floor (ruler) [10]	Pile et al (1991) [10]	-	Standing, tip of middle finger resting on top of horizontal slide of vertically mounted ruler. Position of slide on ruler recorded.	Standing	Max side flexion. Difference between start and end point of slide recorded. Mean calculated.	-

Table 2.42 General description and structure of anthropometric assessments of lumbar mobility. Note: Patient unclad to buttocks for all methodologies.

[a] References defined in table 2.48. Abbreviations defined in Glossary.

Reliability

Evidence of the reliability of lumbar flexion is shown in table 2.43, and high levels of inter and intra-observer reliability have been reported. The greatest amount of evidence is in support of the MSI with the highest levels of reliability generally reported. However, Pile et al (1991) reported a lower level of inter-observer reliability ($r= 0.78$) in a population of 10 AS out-patients, and indicated that there was a significant probability that underlying observer mean values were different ($p= 0.007$).

Validity

Evidence of the validity of lumbar flexion is shown in table 2.44, but limited evidence is available for the Schober Index and the LFI. High correlation between the Schober Index and lumbar flexion radiographs ($r= 0.90$) and a moderate to high correlation with age-adjusted radiological change in the lumbar spine ($r= -0.66$) have been reported. Moderate correlation between the LFI and the measurement of chest expansion and vital capacity, and small correlation with thoracic pain, disease duration and laboratory based measures of disease activity have been reported (Fransen et al, 1986). Several studies report moderate to high correlation between the MSI and lumbar spine radiographic change (Kennedy et al, 1995; Averns et al, 1996a; Viitenan et al, 1999), and additional radiographic analysis strengthens the role of the MSI as a measure of lumbar spine movement ($r= 0.97$) devoid of hip involvement (Macrae and Wright, 1969). Moderate to strong correlation with anthropometric assessment of axial status and chest expansion have been reported, but low to moderate levels of correlation with AS-specific measures of functional disability (Dougadas et al, 1988; Abbott et al, 1994). A small correlation with age, and a moderate correlation with disease duration has also been reported for the MSI.

Responsiveness

Evidence of responsiveness of lumbar flexion is shown in tables 2.45. Evidence from two longitudinal evaluations of physiotherapy with short term follow-up suggest that the Schober Index is not responsive. The MSI has been applied in several longitudinal evaluations of physiotherapy and placebo-controlled drug therapy trials. Accumulated evidence suggests that it is not responsive over the short or longer term, and is unable to discriminate between patients receiving active or placebo drugs. The LFI has only been applied in a single trial of two active drugs in patients with active

[a] Lumbar mobility	Inter-observer	Intra-observer
Flexion		
Schober 10cm Index [1]	2 trained observers (n= 39) male AS in-patients [14] ICC 0.88 ?observers (n=59) AS out-patients, ?type of reliability [13] r= 0.85	
Modified Schober Index[2]	2 trained observers (n= 10, 4 AS) 'mixed' patients [15] r= 0.59 2 trained observers (n= 15) AS out-patients (axial disease)[16] ICC 0.99 5 trained observers, 2 repetitions per patient (n= 10) AS out-patients (minimal disease activity) [10] r= 0.78 Probability that underlying assessor means are different: p= 0.007 3 trained observers (n= 20) AS out-patients [17] r= 0.96 p< 0.001 ?observers (n=59) AS out-patients, ?type of reliability [13] r= 0.87 3 trained observers (n= 3) AS out-patients [6] Non-significant difference	1 trained observer, 1-week retest (n= 15) AS out-patients (axial involvement) [16] r= 0.86 ?observer (n=) AS out-patients [18] CV (%) 6.0 5 trained observers, 4 repetitions (n= 10) AS out-patients (minimal disease activity) [10] Mean 0.6cm, median 0.5cm, 90th percentile 1.5cm 3 trained observers, 1-day retest (n= 20) AS out-patients [17] r= 0.99 1 trained observer, 3-day retest (n= 5) AS in-patients (paired t-test) [19] Non-significant difference (2.0mm+/- 4.5mm)
Lumbar Flexion Index [3]	-	?observers, ? retest period (n=59) AS out-patients [13] r= 0.88
Extension		
Dunham Spondylometer[4]	2 trained observers (n= 19) normal subjects [46] total range r= 0.79 2 observers (n= 10, 4 AS) mixed patients[15] flexion r= 0.76, extension r= 0.87, total range r= 0.88 6 trained observers, random order, 2-day retest (n= 7) male AS out-patients [12] Methodology standardisation: pre - flex r= 0.47, extension r= 0.94; post - flex r= 0.81, extension r= 0.97	1 observer, (n=1) normal subject, 10 observations, ?retest [46] CV(%)-flexion 3.5, extension 2.6. 1 observer, (n=1) normal subject, 10 observations, ?retest [15] CV(%)-flexion 7.01, extension 12.65, total range 3.27 ?observer (n=33) AS out-patients [18] CV (%) - flexion 6.0, extension 12.0
Plumb-line extension [5]	2 observers (1 inexperienced), blinded,?retest (n=14) out-patients (2 AS,1 PID,11 normal)[5] r=0.94 2 observers (n= 10, 4AS) mixed patients [15] r= 0.75 6 trained observers, random order, 2-day retest (n= 7) male AS out-patients [12] Methodology standardisation: pre r= 0.93, post r= 0.98	-
Smythe technique [6]	3 trained observers, (n= 3) AS out-patients[6] No significant difference between U/M/L 10cm segment method in flexion. Not reported for sum change. 6 trained observers, random order, 2-day retest (n= 7) male AS out-patients [12] Methodology standardisation: Smyth (L) flexion$_{\theta}$ - pre r= 0.89, post r= 1.00	1 trained observer (n= 19) AS out-patients, 48-hour retest [47] ICC 0.91 1 trained observer, 4 repetitions over 1-hour, reps 3 and 4 compared (n= 10) AS out-patients[48] Smythe (S) r= 0.97, L r= 0.98, M r= 0.90, U r= 0.82

Table 2.43 Reliability of anthropometric assessments of lumbar mobility (flexion, extension).

[a] References defined in table 2.48. Abbreviations defined in Glossary.

Lumbar mobility	Physical tests / signs	Disability assessment	Other
Flexion Schober 10cm Index [1]	-	(n= 216) AS out-patients[21](r) HAQ-S 0.36, DFI 0.20, HAQ 0.28, AIMS2 0.24	(n= 342) normal population [2] Lateral X-ray: inclination of lumbar spine with skin distraction Schober r= 0.90 (n= 73) male AS in-patients[14] Age-adjusted XR change lumbar spine [(20)] r= -0.66
Modified Schober Index[2]	(n= 85) AS out-patients [22] (r) LLF 0.77, Lumbar extension 0.62, Cexp 0.60 (n= 33, 4 AS) mixed patients[15] Goniometer- lumbar flex r= 0.87, total flex/extension r= 0.83. Spondylometer flexion r= 0.77 (n= 33) AS out-patients [18] Spondylometer flexion r= 0.95, lumbar extension r= 0.78, Cexp r= 0.60 (n=3) AS out-patients [6] Smythe (lower 10cm) r= 0.77	(n= 80) AS out-patients (active disease) [16] DFI r= -0.19 (n= 42) AS out-patients[23] Sub-sections of LDQ - Reaching up / neck movement r= -0.55 (n= 42) AS out-patients [27] Turkish DFI r= 0.52	(n= 1) AS patient, bamboo spine[2] XR - no hip movement (MSI) (n= 342) normal population [2] Lateral X-ray: inclination of lumbar spine with skin distraction MSI r= 0.97 (n= 83) AS out-patients[22] Age r=-0.18 AS duration r= -0.40 (n= 53) AS in-patients [24] Lumbar XR r= 0.68 (n= 61) AS out-patients[25] SASSS XR lumbar spine r=-0.69 (n= 42) AS out-patients[26] Significant difference between smokers and non-smokers MSI p< 0.01 (n= 11) AS in-patients [28] Objective sleep r= 0.62 (n= 65) AS out-patients [29] No relationship with TMJ change (n= 70) AS out-patients [30] SASSS XR lumbar spine r= -0.69
Lumbar Flexion Index [3]	(n= 33) male AS out-patients[11] Vital capacity r=0.45, Cexp r=0.56. Small correlation thoracic pain, AS duration, ESR, IgA.	-	-
Extension			
Dunham Spondylometer[4]	(n= 33, 4 AS) mixed subjects [15] Goniometer (C7 -Sacrum)(total range) r= 0.92. Goniometer lumbar flexion r= 0.91. Goniometer total lateral flex (C7/S1) r= 0.92 (n= 33) AS out-patients [18] Flexion with MSI r= 0.95, p< 0.001	-	(n= 168) AS out-patients[4] Good correlation with XR (no result) (n= 15) AS out-patients, vertebral fracture; (n= 30) age/sex-matched (AS, no fracture)[53] Significant difference between groups (flexion, p< 0.05) and TWD (p<0.005)
Plumb-line extension [5]	(n= 85) AS out-patients [22] Anterior flexion r= 0.62, p<0.001, LLF r= 0.62, p< 0.001, Cexp r= 0.50, p< 0.001 (n= 33, 4 AS) mixed patients[15] Goniometer flexion (C7 - sacrum) r= 0.45. Spondylometer extension r= 0.72	-	(n= 24) out-patients (5 AS) [5] XR spinal extension r = 0.75 (n= 85) AS out-patients[22] Age r= -0.16, AS duration r= -0.34
Smythe technique [6]	(n= 3) AS out-patients [6] Smythe (S) with Goniometer total range r= 0.82, and FFD r= 0.86. Smythe (L) with MSI r= 0.77.	(n= 46) AS out-patients [49] HAQ-S (global) r= 0.33 (n= 144) AS out-patients [50] Overall health (VAS) r= -0.32 (n= 119) AS out-patients [51] PET r= -0.13 (n= 59) AS out-patients [52] Rating scale r= 0.26, Standard gamble r= 0.36 (n= 216) AS out-patients[21] HAQ-S r=0.27, DFI r=0.13, HAQ r=0.17, AIMS2 r=0.12	-

Table 2.44 Validity of anthropometric assessments of lumbar mobility (flexion, extension). * References defined in table 2.48. Abbreviations defined in Glossary.

[a] Lumbar mobility	Used in trial with known efficacy
Flexion Schober 10cm Index [1]	Non-RCT: Retrospective analysis of 3-week rehabilitation (n= 505) AS in-patients. Statistically significant change all measures (p< 0.001). Schober 12.4% (range Schober 12.4% to FFD 36.6%)[31]
	Non-RCT: 4-week rehabilitation (n= 79) male AS in-patients. Measurements day 1 and 18. Effect size: Schober 0.24 (range TLflexion (0.23) to TLrot (0.73) and FFD(0.71))[14]
	Non-RCT: Longitudinal evaluation (usual treatment)(n= 14) AS out-patients. Baseline and 15-months. Non-significant trend for a reduction in all anthropometric measures [32]
	Non-RCT: Longitudinal evaluation of methotrexate (n= 10) AS out-patients, active disease. Baseline and 1-year. Non-significant change in all variables [33]
Modified Schober Index[2]	Non-RCT: 3-week rehabilitation (n= 180) consecutive AS in-patients. Measurement day 0 and 18. Mean change in scores. Statistically significant improvement all variables. MSI p< 0.0005 [31]
	Non-RCT: Retrospective analysis rehabilitation programme (n=52) AS in-patients. 3-weeks significant improvement MSI (mean 0.37cm, p< 0.0001).5-years: non-significant difference MSI (0.25cm, p=0.03). No significant difference between disease duration groups. Conclude MSI is not useful for evaluation of short or long term change [34]
	RCT: Piroxicam (active) and placebo (n= 80) AS out-patients, active disease. 6-week follow-up. Discriminant analysis. Non-significant difference between groups for MSI[16]
	RCT: Sulphasalasine and placebo (n= 89) AS out-patients. 3-year follow-up. Both groups- tendency for MSI to improve(non-significant). No statistically significant improvement short or long term[35]
	RCT: 3 physiotherapy regimes (n= 44) AS out-patients. Assess baseline, 6-weeks, 6-months. Non-significant difference in MSI all treatment groups at 6-weeks and 6-months (F=1.84, p=0.18) [19, 23]
	Non-RCT: Longitudinal evaluation of balneotherapy (2-week duration)(n= 14) AS out-patients. Follow-up 4, 8, 12-weeks No significant change in MSI from baseline to all follow-up [36]
	Non-RCT: 30-minute cycle ergometry (n= 11) male AS out-patients. Assess baseline, 15-minutes, 3 and 5-hours post-exercise. Significant improvement in MSI at 15 minutes only (MSI p< 0.05)[37]
	RCT: 3 active drugs, no placebo (n= 90) AS out-patients, active disease. 26-week follow-up. No significant difference in MSI between treatments or over time[38]
	RCT: Sulphasalasine and placebo (n= 264) AS out-patients, active disease. 36-week follow-up. No significance difference in MSI for either treatment group (p= 0.21), or between groups (p= 0.25) [39]
	RCT: active drug (ACTH) vs placebo; plus physiotherapy (n= 14) male AS out-patients.Assess baseline, 2-weeks.Significant improvement ACTH group (p< 0.001). No difference between groups [40]
	RCT: Sulphasalasine and placebo (n= 62) AS out-patients. 48-week follow-up. No significant difference between treatment groups and no significant difference over time for either group [41]
	RCT: 2 x NSAID (no placebo)(n= 310) AS out-patients. Baseline, 15/30/60/90 days follow-up. MSI 16-21% improvement (0.7mm+/- 0.1mm (p< 0.01). Non-significant difference between groups[42]
	Non-RCT: Open study- 10-week course of rifamycin SV infiltrations to all large peripheral joints vs oral administration (n= 22) AS out-patients, active disease. Clinical improvement at 10-weeks persisted for 12-week follow-up (p< 0.006). Non-significant improvement for oral administration [43]
	RCT: Efficacy and tolerability of Sulphasalasine vs placebo (n= 134) AS out-patients. 6-month duration. Assess baseline, 3, 6-months. No statistically significant difference in MSI[44]
LFI [3]	RCT: 2 active drugs (n= 33) AS male out-patients, active disease. 12-weeks treatments, 36-week follow-up. Significant improvment in LFI, Cexp, pain and AS activity at 12 and 48-weeks (p< 0.01)[11]
Extension Spondylometer[4]	Non-RCT: Longitudinal evaluation of sulphasalasine (36-weeks follow-up) (n= 20) AS out-patients with active disease. No significant change in spondylometer measurements at follow-up [54]
Plumb-line extension [5]	RCT: active drug (ACTH) vs placebo; plus physiotherapy (n= 14) male AS out-patients.Assess baseline, 2-weeks.Significant improvement ACTH group (p< 0.001). No difference between groups [40]
Smythe technique [6]	RCT: Home exercise vs home plus group exercise (n= 144) AS out-patients. 9-month follow-up (treatment period). Statistically significant difference between groups for Smythe and patient global assessment of health (VAS)(p< 0.01). Non significant difference in cexp, cervical rotation or self-assessed function [55,57]
	Non-RCT: 6-week before and after evaluation of supervised individual physiotherapy (n= 144) AS out-patients. Non-significant change in Smythe (p= 0.92) [47]
	RCT: Supervised vs unsupervised therapy (n= 68) AS out-patients. 9-month follow-up. Non-significant difference between groups, and non-significant change for Smythe and all ROM[56]
	Non-RCT[b]: Longitudinal evaluation group physiotherapy (n= 67) AS out-patients. 9-month follow-up. Correlation with change in patient assessed global health (VAS) r= 0.15 [45]

Table 2.45 Responsiveness of anthropometric assessments of lumbar mobility (flexion and extension). [b] correlation of scale change with change in other variables.
[a] References defined in table 2.48. Abbreviations defined in Glossary.

disease and a significant improvement in range was reported for both short and long-term follow-up (Fransen et al, 1986).

b. Lumbar extension

General description

One approach to specifically assess thoracolumbar extension has been described which requires a plumb-line and tape measure (Moll et al, 1972b). Two additional approaches measure both flexion and / or extension of the lumbar, thoracic or total spine: Dunham spondylometer (Dunham, 1949) and the Smythe Technique (Miller et al, 1984). The Smythe technique requires a tape measure and a treatment couch. A general description and methodology for these techniques is summarised in table 2.42.

Several authors have referred to the use of the MSI in the assessment of lumbar extension but without further methodological detail. Macrae and Wright (1969) do not describe the measurement of lumbar extension with the MSI and this methodology has not been included in the review.

Reliability

Evidence of the reliability of lumbar extension is shown in table 2.43. Good levels of inter-observer reliability are reported for the spondylometer assessment of both flexion and extension, although low inter-observer reliability pre-standardisation of flexion methodology was reported (Bellamy et al, 1991a). There is limited evidence of intra-observer reliability of the spondylometer in patients with AS. A good level of inter-observer reliability for the plumb-line extension technique (Moll et al,1972b) has been reported by several studies, but intra-observer reliability has only been reported in healthy subjects (Moll et al, 1972b; Reynolds, 1975). Evidence of reliability for the Smythe technique suggests good inter and intra-observer reliability for the sum assessment (sum-S) and individual segmental assessment (lower-L, middle-M, upper-U). However, evidence for the Smythe technique is limited due to poor clarification for the segmental assessment recorded.

Validity

Evidence of the validity of lumbar extension is shown in table 2.44. High levels of correlation between the spondylometer and the MSI have been reported, and a 'good' level of correlation with spinal radiographic change (Dunham, 1949). A significant

difference in range of movement between AS patients with vertebral fracture and those without was also reported (Ralston et al, 1990). However, investigators do not always clarify if movement refers to spinal flexion, extension or both. A strong relationship between plumb-line extension and radiographic evaluation of spinal extension (r= 0.75)(Moll et al, 1972b), and moderate to strong relationships with other measures of spinal range of movement have been reported. A low relationship with age and disease duration was also reported. Strong correlation between the Smythe (sum) and other assessments of total spinal mobility and with FFD have been reported. Small correlation with several assessments of disability have been reported suggesting a small relationship with functional disability and patient-assessed health.

Responsiveness

Limited evidence of the responsiveness of lumbar extension is shown in table 2.45. A single, small sample study suggests that plumb-line extension may be responsive to change following active drug therapy combined with physiotherapy. However, the spondylometer and the Smythe technique appear not to be responsive following drug therapy and physiotherapy respectively.

c. Lumbar lateral flexion

General description

Four approaches for the assessment of lumbar lateral flexion (LLF) have been described. Three methodologies require only a tape measure, and the fourth requires a vertically mounted ruler with a horizontal slide. A general description and methodology is shown in table 2.42.

Reliability

Evidence of the reliability of LLF is shown in table 2.46. There is no consensus on the level of inter-observer reliability of the skin distraction technique (Moll et al, 1972a) and a range of values have been reported, ranging from small to good (r= 0.31 to 0.97), and Pile et al (1991) report a significant probability that the observer means are different (p< 0.02). A moderate level of inter-observer reliability was reported for the measurement of fingertip to fibula distance (r= 0.79), a value that deteriorated post-standardisation of procedure (r= 0.77)(Bellamy et al, 1991a). There is limited evidence of the reliability of the methodology described by Domjan et al (1990) in patients with AS. Creemers et al (1996) report a good level of intra-observer

Lumbar mobility	Inter-observer	Intra-observer
Lateral flexion		
Skin distraction [7]	2 observers (1 inexperienced) (n= 17) out-patients (2 AS, 1 PID, 1 scoliosis, 13 normal) [7] r = 0.68	2 observers, (n= 10, 4 AS) mixed patients [15] CV (%) right 15.75, left 12.91
	2 trained observers, (n= 10, 4 AS) mixed patients [15] right r= 0.41, left r= 0.31	5 trained observers, 4 reps, random order (n= 10) AS out-patients, minimal AS activity [10] Right (mean 0.8cm, median 0.5cm, 90th percentile 1.5cm). Left (mean 0.7cm, median 0.5cm, 90th percentile 2.0cm)
	5 trained observers, 2 repetitions (n= 10) AS out-patients (minimal disease activity)[10] right r=0.72, left r= 0.59. Probability that underlying assessor means are different: right p= 0.003, left p= 0.016	
	6 trained observers, random order, 2-day retest, (n= 7) male AS out-patients [12] Methodology standardisation: pre r= 0.97, post r= 0.97	
Fingertip to fibula [8]	6 trained observers, random order, 2-day retest, (n= 7) male AS out-patients [12] Methodology standardisation: pre r= 0.79, post r= 0.77	-
Fingertip lateral thigh [9]	-	? observers ? retest (n= 59) AS out-patients [13] r= 0.88
Fingertip to floor (ruler) [10]	5 trained observers, 2 repetitions (n= 10)AS out-patients (minimal disease activity) [10] right r=0.83, left r= 0.79. Probability that underlying assessor means are different: right p= 0.029, left p=non-significant	5 trained observers, 4 reps, random order (n= 10) AS out-patients, minimal AS activity [10] Right (mean 2.3cm, median 2.0cm, 90th percentile 5.0cm). Left (mean 3.0cm, median 1.6cm, 90th percentile 5.7cm).
	3 trained observers (n= 20) AS in-patients [17] r= 0.94	3 trained observers, 24-hour retest (n= 20) AS in-patients [17] r= 0.99

Table 2.46 Reliability of anthropometric assessments of lumbar mobility (lateral flexion).

[a] References defined in table 2.48. Abbreviations defined in Glossary.

reliability (r= 0.88), but do not describe the number of observers or retest period. The strongest support is for fingertip to floor distance following LLF (Pile et al, 1991), with good inter (r= 0.79 to 0.84) and intra-observer (r= 0.99) reliability reported by different investigators. Evidence for the intra-observer reliability of other methodologies in AS is limited.

Validity

Evidence of the validity of LLF is shown in table 2.47. There is no evidence of the validity of fingertip to fibula distance (Little, 1986) and the method described by Domjan et al (1990) in AS. Most evidence supports the skin distraction technique, and a strong relationship with radiographic evaluation of LLF was reported (r= 0.79)(Moll et al, 1972a). Moderate correlation with a sub-section of the LDQ (Abbott et al, 1994), and moderate to strong correlation with various anthropometric spinal measures. Low to moderate correlation with age and disease duration was also reported (Moll et al, 1973). A moderate relationship between LLF (mounted ruler) and lumbar radiographic change (r= 0.59)(Kennedy et al, 1995) was reported.

Responsiveness

Limited evidence of the responsiveness of LLF is shown in table 2.48, and suggests that LLF measured by skin distraction is unable to distinguish between patients receiving active drug or placebo in a controlled trial (Calcraft et al, 1974). There is no evidence for other methodologies.

Acceptability of measures of lumbar mobility

The skin distraction techniques were designed to reduce the potential danger or inconvenience to patients observed in other measurement approaches (Macrae and Wright, 1969; Moll et al, 1972a; Moll and Wright, 1973). Patients described the MSI as the most convenient assessment of lumbar mobility (Miller et al, 1984).

Feasibility and application of measures of lumbar mobility

The measurement of lumbar flexion with the MSI is one of the most frequently cited measurements in published articles, and in routine clinical practice (Bellamy et al, 1998; 1999). The Spondylometer has not been widely applied over the review period. The need for costly equipment that may not be readily portable may have supported the decline in use.

[a] Lumbar mobility	Physical tests / signs	Disability assessment	Other
Lateral flexion			
Skin distraction [7]	(n= 85) AS out-patients [22] Cexp r=0.56, MSI r=0.77, lumbar extension r=0.62 (n= 33, 4 AS) mixed patients [15] range: lumbar extension (goniometer) r= 0.53 to spondylometer flexion (total) r= 0.95	(n=45)AS out-patients[23] LDQ - significant correlation with Reaching up /neck movements r= -0.62, Posture r= -0.38	(n= 85) AS out-patients [22] Age r=0.30, AS duration r=0.51 (n= 43) mixed subjects (7 AS out-patients, 36 normal) [7] XR lumbar side flexion (T9-L5) r= 0.79 (n= 65) AS out-patients[29] No correlation with TMJ change (n= 11) AS
Fingertip to fibula [8]	-	-	-
Fingertip lateral thigh [9]	-	-	-
Fingertip to floor (ruler)[10]	-	(n= 42) AS out-patients [27] Turkish DFI r= 0.50	(n= 53) AS in-patients [12] LLF (component of BASMI) with lumbar radiographic analysis r= 0.56, p< 0.001

Table 2.47 Validity of anthropometric assessments of lumbar mobility (lateral flexion). [a] References defined in table 2.48. Abbreviations defined in Glossary.

Lumbar mobility	Used in trial with known efficacy
Lateral flexion	
Skin distraction [7]	RCT: placebo controlled trial of azapropazone (n= 22) AS out-patients. 6-week follow-up. LLF unable to discriminate between treatment and placebo groups [58] Non-RCT: Longitudinal evaluation (usual treatment)(n= 14) AS out-patients. Baseline and 15-months. Non-significant trend for a reduction in all anthropometric measures [32]
Fingertip to fibula [8]	-
Fingertip lateral thigh [9]	RCT: 2 x NSAID (no placebo)(n= 310) AS out-patients. Baseline, 15/30/60/90 days follow-up. LLF 6-10% improvement (0.6mm+/- 0.1mm (p< 0.05). Non-significant difference between groups [42]
Fingertip to floor (ruler)[10]	-

Table 2.48 Responsiveness of anthropometric assessments of lumbar mobility (lateral flexion). Abbreviations defined in Glossary.

References: [1] von Schober (1937), [2] Macrae & Wright (1969), [3] Adrichem & van der Korst (1973), [4] Dunham (1949), [5] Moll et al (1972b), [6] Miller et al (1984), [7] Moll et al (1972a), [8] Little (1986), [9] Domjan et al (1990), [10] Pile et al (1991), [11] Franssen et al (1986), [12] Bellamy et al (1991a), [13] Creemers et al (1996), [14] Viitenan et al (1995b), [15] Reynolds (1975), [16] Dougadas et al (1988), [17] Jenkinson et al (1994a), [18] Fisher et al (1990), [19] Helliwell et al (1996), [20] Dale & Vinje (1985), [21] Ward & Kusiz (1999), [22] Moll & Wright (1973), [23] Abbott et al (1994), [24] Kennedy et al (1995), [25] Averns et al (1996a), [26] Averns et al (1996b), [27] Dalyan et al (1999), [28] Jamieson et al (1995), [29] Ramos-Remus et al (1997), [30] Dawes (1999), [31] Viitenan et al (1992), [32] Lee et al (1997), [33] Ostendorf et al (1998), [34] Roberts et al (1989), [35] Kirwan et al (1993), [36] Tishler et al (1995), [37] Carbon et al (1996), [38] Taggart et al (1996), [39] Clegg et al (1996), [40] Pearcy et al (1985), [41] Corkhill et al (1990), [42] Battle-Gualda et al (1996), [43] Caruso et al (1992), [44] Dougadas et al (1995), [45] Hidding & van der Linden (1995), [46] Sturrock et al (1973), [47] Hidding et al (1993b), [48] Roberts et al (1988), [49] Daltroy et al (1990), [50] Hidding et al (1994b), [51] Bakker et al (1995), [52] Bakker et al (1994a), [53] Ralston et al (1990), [54] Fraser & Sturrock (1990), [55] Hidding et al (1993a), [56] Hidding et al (1994a), [57] Bakker et al (1994b), [58] Calcraft et al (1974).

Posture

General description

Three methodologies for the assessment of upper cervical or spinal posture have been identified. One approach measures tragus to wall distance (TWD) with a perspex t-square (Tomlinson et al, 1986). The remaining two methods measure occiput to wall distance (OWD), one approach using a tape measure (Americam Rheumatology Association - ARA, 1984) and the second a PSMS (Stokes et al, 1988). A general description and methodology is summarised in table 2.49.

Reliability

Evidence of the reliability of the measurement of spinal posture is shown in table 2.50. High levels of inter and intra-observer reliability have been reported for all measurement techniques, but evidence for the PSMS is limited to a single study from the developers. Several studies report high levels of reliability for OWD with a tape measure (inter-observer ICC> 0.92, intra-observer ICC> 0.99), and TWD with a t-square (inter-observer r> 0.97, intra-observer r> 0.99).

Validity

Limited evidence of the validity of the measurement of spinal posture is shown in table 2.51. Small to moderate correlation between OWD (tape measure) and lumbar spine age-adjusted radiological change (r= 0.49)(Viitenan et al, 1995a), and with AS-specific radiographic change in the cervical and lumbar spine (range r= 0.37 to 0.38) have been reported (Taylor et al, 1991a). A strong correlation between TWD (t-square) and cervical spine radiographic assessment was reported (r= 0.61)(Kennedy et al, 1995). Moderate to high correlation between OWD (tape measure) and other anthropometric assessments was reported, but a small correlation between OWD and FFD both measured with the PSMS (Stokes et al, 1988).

Responsiveness

Evidence of the responsiveness of the measurement of spinal posture is shown in table 2.52. There is limited evidence for all methodologies and no consensus can be clearly drawn. Evidence suggests that the PSMS is not responsive, but that OWD (tape measure) and TWD (t-square) may demonstrate low levels of responsiveness following physiotherapy.

[a] Spinal posture		Methodology			
	Earliest application - AS	Identified need	Land markings	Starting position	Application and special equipment	Time
Tape - OWD [1]	Stokes et al (1988) [3]	-	Standing, back to wall, heels, buttocks and shoulders touching wall, knees in extension. Tragus of ear identified	Standing	Patient pushes head back to wall, chin parallel to floor. If unable to touch occiput-to-wall (OWD), OWD measured. Metal end of tape measure placed flush to wall and extended horizontally to occiput[2]	-
T-square - TWD[2]	Tomlinson et al (1986) [2]	-	Standing, back to wall, heels, buttocks and shoulders touching wall, knees in extension. Tragus of ear identified	Standing	Patient attempts to place occiput against wall, chin tucked in, eyes level. Horizontal distance tragus-to-wall (TWD)(both sides) measured with transparent 'T'-square. Mean value calculated	-
PSMS - OWD [3]	Stokes et al (1988) [3]	Reliable and valid measure	Standing, back to wall, heels, buttocks and shoulders touching wall, knees in extension. Occiput identified	Standing	Base of Portable Spinal Mobility Scale (PSMS) positioned against wall, long arm rests horizontally on patients head (perpendicular to the base and 'small' arm). Patient pushes head back to wall, chin parallel to the floor. Small arm of scale is moved to meet the occiput. OWD measured	-

Table 2.49 General description and structure of anthropometric assessment of spinal / upper cervical posture. [a] References defined in table 2.52. Abbreviations defined in Glossary.

[a] Spinal posture	Inter-observer	Intra-observer
Tape - OWD [1]	6 trained observers, random order (n= 7) male AS out-patients [5] Methodology standardisation: pre r= 0.98, post r= 0.99	2 trained observers, 2-hour retest (n= 38) AS in-patients [7] ICC 0.99
	2 trained observers (n= 39) AS in-patients [7] ICC 0.92	
	2 experienced observers (n= 39) male AS in-patients [8] ICC 0.92	
	2 trained observers, no warm-up (n= 52) AS out-patients [9] ICC range 0.89 - 0.98 (cervical mobility)(no specific result)	
T-square - TWD[2]	5 trained observers, 2 repetitions (n= 10) AS out-patients, minimal AS activity [4] r= 0.97 Non-significant difference between assessor means.	5 trained observers, 4 repetitions (n= 10) AS out-patients, minimal AS activity [4] mean 0.5cm, median 0.5cm, 90th percentile 1.5cm
	6 trained observers, random order (n= 7) male AS out-patients [5] Methodology standardisation: pre r= 1.00, post r= 1.00	3 trained observers, blinded, 1 repetition, 24-hour retest (n= 20) AS out-patients [6] r= 0.99
	3 trained observers, blinded, 1 repetition (n= 20) AS out-patients [6] r= 0.99	
PSMS - OWD [3]	5 trained observers, random order, (n= 5) AS out-patients [3] Non-significant difference between observers p= 0.20	5 trained observers, random order, (n= 5) AS out-patients [3] ICC 0.97

Table 2.50 Reliability of anthropometric assessment of assessment of spinal / upper cervical posture. [a] References defined in table 2.52. Abbreviations defined in Glossary.

[a] Spinal posture	Physical tests / signs	Disability assessment	Other
Tape - OWD [1]	(n= 32) AS out-patients [11] Total spinal flexion, chest expansion, FFD. Range r= 0.60 – 0.65	(n= 216) AS out-patients [12] HAQ-S r= 0.33, DFI r= 0.22, HAQ r= 0.23, AIMS2 r= 0.21	(n= 73) male AS in-patients [8] Age-adjusted XR change lumbar spine[(13)] r= 0.49
			(n= 32) AS out-patients [7] XR spinal score (range) r= 0.37 – 0.47
			(n= 52) AS out-patients [9] AS-specific XR spinal change[(13)] cervical r=0.38, thoracic r= 0.22 lumbar r= 0.37, sacroiliac joint r= 0.20
			(n= 61) AS out-patients[15] SASSS XR evaluation lumbar spine r= 0.50
			(n= 42) AS out-patients[16] Significant difference between smokers and non-smokers p<0.02
			(n= 65) AS out-patients [17] No relationship with TMJ condylar erosion (CT scan)
T-square - TWD[2]	-	-	(n= 53) AS in-patients [10] Cervical spine XR r= 0.61
PSMS - OWD [3]	5 trained observers, random order, (n= 5) AS out-patients [3] Relationship between OWD and FFD both measured with PSMS r= 0.38	-	-
	Significant difference between PSMS-OWD and OWD (tape measure) p= 0.0138 [3]		

Table 2.51 Validity of anthropometric assessment of spinal / upper cervical posture. [a] References defined in table 2.52. Abbreviations defined in Glossary.

[a] Spinal posture	Used in trial with known efficacy
Tape - OWD [1]	Non-RCT: Retrospective analysis of 3-week rehabilitation (n= 505) AS in-patients. Statistically significant change in all measures (p< 0.001). % change OWD 30.8% (range Schober test 12.4% to FFD 36.6%) [19]
	Non-RCT: 3-week rehabilitation programme (n= 79) male AS in-patients. Measurements assessed on day 1 and day 18. Effect size (ES) calculated OWD ES= 0.25. Most responsive thoracolumbar rotation and FFD (ES= 0.73). Least responsive thoracolumbar flexion (ES= 0.23) and Schober test (ES= 0.24) [8]
T-square - TWD[2]	Non-RCT: 3-week rehabilitation (n= 180) AS consecutive in-patients. Measurements assessed on day 0 and end of programme. Mean change in scores calculated. Statistically significant improvement in all variables (p< 0.001). TWD p< 0.0005[2]
PSMS - OWD [3]	RCT: physiotherapy (home exercise and education) vs control group (n= 53) AS out-patients. 4-month follow-up. Baseline and follow-up scores compared. No significant difference in OWD, Smythe technique, or Pain (VAS). Significant improvement in FFD and functional ability in treatment group (p< 0.001) [14,18]

Table 2.52 Responsiveness of anthropometric assessments of spinal / upper cervical posture. Abbreviations defined in Glossary.

References: [1] American Rheumatology Association (ARA) Glossary (1984), [2] Tomlinson et al (1986), [3] Stokes et al (1988), [4] Pile et al (1991), [5] Bellamy et al (1991a), [6] Jenkinson et al (1994a), [7] Viitenan et al (1995b), [8] Viitenan et al (1995a), [9] Viitenan et al (1998), [10] Kennedy et al (1995), [11] Taylor et al (1991a), [12] Ward & Kuzis (1999), [13] Dale & Vinje (1985), [14] Kragg et al (1990), [15] Averns et al (1996a), [16] Averns et al (1996b), [17] Ramos-Remus et al (1997), [18] Kragg et al (1994), [19] Viitenan et al (1992).

Acceptability of measures of spinal / upper cervical posture

Acceptability to patients has not been reported.

Feasibility and application of measures of spinal / upper cervical posture

There has been wide application of the measurement of both OWD and TWD in published studies, but the PSMS has only been referred to in two studies beyond the developers. Bellamy et al (1998, 1999), report that the measurement of OWD is routinely applied in clinical practice by more than 50% of clinicians, but the methodology is not defined.

BASMI

Purpose and conceptual base

Jenkinson et al (1994a) determined the need to identify the minimum number of clinically appropriate measures to accurately assess axial status in AS. Axial status was described by mobility of the cervical, thoracic and lumbar spine, hips and pelvic region.

General description

The BASMI consists of five items, as shown in table 2.53: cervical rotation (simple inclinometer - Klaber-Moffett et al, 1989), TWD (t-square - Tomlinson et al, 1986), LLF (mounted ruler - Pile et al, 1991), lumbar flexion (MSI - Macrae and Wright, 1969) and hip mobility (intermalleolar distance - Calin, 1985). Items were selected from an initial pool of 20 measures, supported by a literature search, clinical opinion and an assessment of all measures in a sample of in-patients (n= 43). The extent of the literature search, item selection criteria (beyond the reflection of axial status), and the patient assessment was not detailed. A high level of concurrent validity between the five selected measures and the original 20 measures was reported (r= 0.94). The original three-point response scale, determined by the specific range of movement available for each measure, was replace by an 11-point response scale due to poor sensitivity (Jones et al, 1995). An open-ended range of possible movement values for each measure is divided into 11 equal sections (0-10), determined by expert opinion. Each item response is totalled (0-50), and the mean value reflects the BASMI score (0-10); a lower score indicating less restricted axial mobility.

	[a] BASMI	
Author, year of publication	Jenkinson 1994a	
Earliest application in AS	Jenkinson 1994a	
Identified need for instrument	No standardisation for measurement of axial movement. Aim to determine the minimum number of clinically appropriate and reliable measurements to assess axial status in AS.	
Methodology		
Land markings and special equipment	i.	Cervical rotation – gravity action goniometer / simple inclinometer (AAOS 1965, Klaber-Moffett et al 1989)
	ii.	Tragus to wall – tragus of ear. Perspex t-square (Tomlinson et al 1986)
	iii.	Lateral lumbar flexion - finger tip to floor distance. Mounted ruler with slide (Plie et al 1991)
	iv.	Modified Schober Index – lumbosacral junction, 5cm distal and 10cm proximal (Macrae & Wright 1969). Tape measure.
	v.	Intermalleolar Distance – Distance between medial malleoli (Calin 1985). Tape measure.
Patient position	i.	Supine with goniometer placed on forehead.
	ii.	Standing with heels and buttocks against a wall, knees extended, shoulder and head back as far as possible.
	iii.	Standing with arms down by side – tip of 3rd finger resting onto of the slide of the ruler.
	iv.	Standing erect.
	v.	Supine, legs together, knees extended and feet pointing towards the ceiling.
Method of application	i.	Maximal rotation to left and to right. Mean of right and left calculated.
	ii.	Horizontal distance between tragus and wall with a ruler (Not indicate if both sides measured).
	iii.	Maximal side flexion, without forward flexion, or knee flexion. Mean of right and left calculated
	iv.	Maximal forward flexion.
	v.	Maximal bilateral abduction of the hips.
Time for administration	7 minutes	

Table 2.53 General description and scale structure of BASMI.

[a] References defined in table 2.54. Abbreviations defined in Glossary.

Reliability

As shown in table 2.54 high levels of inter (r> 0.94) and intra-observer (r> 0.98) reliability of the individual measurements included in the BASMI have been reported, but the test-retest reliability of the index has not been reported.

Validity

Moderate to strong levels of correlation between the total BASMI index (r= 0.74) and most individual items (range LLF r= 0.56 to MSI r= 0.69) in the index with radiographic analysis of related spinal areas were reported and accepted as evidence for the validity of the index as a reflection of axial status, as shown in table 2.55. The lowest correlation was between intermalleolar distance and a hip radiograph (r= 0.27). A moderate correlation between the index and disease duration was reported (r=0.44). The developers suggest that the BASMI should be considered the new 'gold standard' for the evaluation of mobility, replacing radiographic evaluation (Jenkinson et al, 1994a; Kennedy et al, 1995).

Responsiveness

Limited evidence of responsiveness suggests that the BASMI is responsive to change following a short period of intensive physiotherapy (p< 0.001)(Band et al, 1997). Significant score improvement following a longitudinal drug study is limited to the evaluation of only 8 patients (Maksymowych et al, 1998)(table 2.56).

Acceptability

Acceptability to patients has not been reported.

Feasibility and application

The BASMI takes 7-minutes to administer but several items of equipment are required: inclinometer, t-square, mounted ruler, tape measure and a treatment couch. Clinical feasibility has not been reported. The BASMI has been identified in 5 published articles and 1 letter, all by members of the development team.

Normal values

No identified anthropometric measure is purely AS-specific, and the impact of normal aging or pathological insult should be considered as independent facets when assessing mobility (Viitenan et al, 1995a). Normal movement covers a wide range of

	Inter-observer	Intra-observer
BASMI[1]	3 trained observers, blinded, 1 repetition (n= 20) AS in-patients [1] cervical rotation r = 0.99, tragus to wall r = 0.99, lumbar lateral flexion r = 0.98, Modified Schober Index r = 0.99, intermalleolar distance r = 0.99	3 trained observers, blinded, 24-hour re-test (n= 20) AS in-patients (different sample to inter-observer) [1] cervical rotation r = 0.98, tragus to wall r = 0.99, lumbar lateral flexion r = 0.94, Modified Schober Index r = 0.96, intermalleolar distance r = 0.98

Table 2.54 Reliability of BASMI. Abbreviations defined in Glossary.

References: [1] Jenkinson et al (1994a), [2] Kennedy et al (1995), [3] Taylor et al (1998), [4] Band et al (1997), [5] Maksymowych et al (1998).

[a] Physical tests / signs	Radiographic change	Other
(n= 43) AS in-patients [1] BASMI with 'parent' instrument (20 measures) r = 0.92 (n= 54) AS in-patients [2] BASMI with 'parent' instrument (20 measures) r = 0.94	(n= 53) AS in-patients [2] BASMI (total) and separate items with spinal XR. BASMI (total) + XR r= 0.74 Cervical rotation + cervical XR r= 0.59 Tragus to wall + cervical XR r= 0.61 Lumbar side flexion + lumbar XR r= 0.56 Modified Schober Index + lumbar XR r= 0.68 Intermalleolar distance + hip XR r= 0.27	(n= 43) AS in-patients[2] Age p< 0.01, Disease duration p< 0.004 (n= 393) AS out-patients[3] AS duration r= 0.44

Table 2.55 Validity of BASMI. [a] References defined in table 2.54. Abbreviations defined in Glossary.

[a]	Used in trial with known efficacy
BASMI[1]	Non-RCT: 3-week rehabilitation (n= 56) AS in-patients. Comparison of mean value at baseline (3.34, sd 2.71) and 3-weeks (2.16, sd 2.42). Approximate 30% score improvement. Of total population 71% improved, 29% the same (n= 16), 0% deteriorated. Of those staying the same n=8 had baseline score of 0, n=4 scored< 3 [1]
	Non-RCT: 2-week rehabilitation (n= 236) AS in-patients. Comparison day 0 and 14. Significant improvement in BASMI p< 0.001; 22% improvement (BAS-G 27%, BASFI 25%, BASDI 18%) [4]
	Non-RCT: Longitudinal open analysis of pamidronate (n= 8) AS out-patients, active disease (3-month intervention; baseline, 3 / 6 / 9-month follow-up). Comparison between baseline and follow-up - significant improvement in BASMI (p= 0.01), BASDAI (p= 0.03), and ESR scores at 6-months (paired t-test) and for BASMI at 3-months (p= 0.007) [5]

Table 2.56 Responsiveness of BASMI. [a] References defined in table 2.54. Abbreviations defined in Glossary.

values that may overlap with ranges observed in specific pathologies, thereby supporting the assessment of mobility against a range of values instead of a single average value (Moll et al, 1972b; Moll and Wright, 1973). There is no consensus on the clinical significance associated with change in most anthropometric assessments. In the evaluation of FFD, Roberts et al (1989) reported a statistically significant mean change of 2.86cm following the long term follow-up of physiotherapy but suggested that this was not clinically significant. However, Kragg et al (1990, 1994) suggest that a change of 3.0cm is clinically significant.

Only the BASMI provides a battery of measurements that, when combined, provide a score specific to the evaluation of AS. The response scale for each section of the instrument was based on clinical knowledge of AS and not on published evidence of normal values.

Commentary

Traditionally, evaluation of axial movement was based on visual estimation (AAOS, 1965). However, the lack of reliable and valid data led to the development of more objective means for anthropometric assessment. The purpose and conceptual base of all reviewed anthropometric measures has practical appeal and clinical relevance related to the impact of AS on the axial skeleton. All measurements have been developed through clinical knowledge and experience with AS, often following a pragmatic approach to measurement. The majority of anthropometric measures assessed purport to measure axial mobility. Tragus or occiput to wall distance provides a representation of spinal posture, as opposed to spinal movement. All methodologies are examination-based and place different demands on patient and clinician in terms of time and ease of administration, convenience and need for special equipment. Tables 2.57 to 2.63 provide a summary of the data synthesis.

A full description or reference to support the methodology adopted is essential to allow data to be combined. It was often difficult to identify articles describing the original development of a methodology (level I articles) and the evidence for several anthropometric measures commence with the first clear description of application in AS (level II articles). In addition, where articles reflecting initial development of a methodology were identified detail was often limited. Difficulties in clearly identifying methodological approaches was a common problem in the review. Many

investigators simply indicate that, for example, lumbar flexion was assessed. The review has illustrated that many similar methodologies exist for measuring the same movement. However, reliability and standardisation effects are not the same for closely related techniques and the 'most appropriate technique from competing alternatives' (Bellamy et al, 1991a) should be identified. Also, although methodologies may not differ substantially, standardisation is essential to support comparison of results (Viitenan et al, 1995a) and to support the synthesis of evidence. The acceptability, feasibility and measurement properties of an anthropometric measure can only be inferred for the protocol described, and therefore, despite the large number of articles indicating the use of anthropometric measurements the availability of data has been limited.

Many investigators do not indicate the time of day the measurements were taken. Morning stiffness is a significant feature of AS (Dziedzic, 1998), and due to stiffness resolution measurements taken in the morning may differ to those taken later in the day. This could influence estimates of instrument reliability, validity and responsiveness. Mobility may also be influenced by the level of exercise pre-measurement and standardisation of warm-up or exercise is recommended (Roberts et al, 1988) but not often reported by investigators.

The relative experience or training and number of observers was often not detailed. Application of clear methodology and the need to identify specific bony landmarks for certain measurements implicates the need for training and, or standardisation of approach between observers. Bellamy et al (1991a) highlight the influence of methodology standardisation on several anthropometric measurements, generally resulting in an improvement in inter-observer reliability.

Notwithstanding these issues, together with a range of test-retest intervals and the lack of consideration for disease-stability shown by many investigators, evidence of the inter and intra-observer reliability for many frequently adopted measures seems satisfactory and supports the application of most methods in group evaluation. Many methods achieve higher levels of reliability and are suitable for individual evaluation, for example, MSI and TWD. However, there is a lack of evidence for the reliability of several measures, for example, inclinometer assessment of cervical flexion and extension. These measures are not recommended for use without some evidence in

support of their reliability in AS. Greater evidence of reliability with clearly defined and appropriate retest periods, clarification of the number and experience of observers, defined patient populations, and external evaluation of condition stability is required for all measurements.

There is limited evidence to support the validity of most anthropometric measures, and all investigators have failed to state a priori hypothesised relationships between variables, therefore logically weakening the results (McDowell and Jenkinson, 1996). A 'gold standard' evaluation of spinal movement is not available for routine use, although radiographic evaluation may provide a valid representation of vertebral joint movement (Kennedy et al, 1995). The relationship between several anthropometric measures and radiographic evaluation of regional joint movement or of AS-specific change has been reported. A strong relationship between certain measures, for example, between the MSI and the lumbar spine, and between TWD and the cervical spine, have been accepted as evidence of the validity of the anthropometric measure as a reflection of spinal mobility or disease progression.

Direct comparison between anthropometric measures presenting different approaches to measuring the same range of movement have been reported for measurements of lumbar and total spinal flexion, and strong correlation reported. For example, between the MSI, spondylometer and goniometer. Evidence of the acceptance, feasibility and reliability of methods with evidence to suggest that they measure a similar movement may further assist in the selection of a specific methodology. Few studies of direct comparison between the many available methods for measuring different ranges of cervical mobility, or the different ranges of lumbar mobility, for example, rotation versus flexion, have been identified and may further assist appreciation of the inter-relationship between movements. For example, it may not be necessary to include measurements of lumbar flexion and lateral flexion in an evaluation if a strong relationship between the two measurements exists.

Although small to strong levels of correlation have been reported between selected anthropometric measures and AS-specific measures of functional disability, comparison with more established measures of HRQL have not been undertaken and are recommended with a priori hypotheses stated. This may provide a wider appreciation of the role of anthropometric assessment in patient evaluation.

Many anthropometric measures do not have evidence of responsiveness. Accumulated evidence suggests that cervical rotation and fingertip to floor distance (anterior flexion) (FFD) are responsive to change in mobility following physiotherapy, but not following drug therapy. Chest expansion may be responsive to change in thoracic mobility following physiotherapy, but the extent of change may not be clinically significant (Moll and Wright, 1973). Evidence suggests that lumbar flexion and lateral lumbar flexion are not responsive to change in mobility following physiotherapy or drug therapy over the short term. However, radiographic evaluation of AS-specific change in the lumbar spine is strongly correlated with the MSI. Radiographic change is often used as a long term end-point in the evaluation of AS (Dawes, 1999), and the MSI may therefore play an important role in the long-term evaluation of AS, reflecting structural change in axial status.

Few authors report on the acceptability of measures. However, the frequent application of certain measures in research and in routine practice suggests high feasibility. The anthropometric measures used by more than 70% of clinicians in routine longitudinal evaluation require only a tape measure, and are quick and easy to perform (MSI, TWD, FFD and chest expansion) (Bellamy et al, 1998, 1999). In addition, cervical rotation is the most frequently measured cervical movement in published articles, and several authors have indicated that cervical rotation and lateral flexion are both clinically useful measures of cervical mobility (Pile et al, 1991; Jenkinson et al, 1994a).

Cervical mobility

Irrespective of methodology, evidence suggests that the assessment of cervical rotation is more reliable and responsive than the measurement of cervical lateral flexion, flexion or extension, as shown in table 2.57, and is the measurement selected for inclusion in the longitudinal study. However, selection of a specific methodology was not self-evident.

Although universal goniometers are readily accessible in most clinics, the protractor (Calcraft et al, 1974; Roberts et al, 1988) is bulky and less available, and evidence of poor reliability suggests that neither methodology should be adopted in AS evaluation. There was often a lack of clarity for the type of inclinometer adopted by

Anthropometric measures													
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
Cervical spine [a]	Scale	Special equipment	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
Rotation													
Large protractor[1]	degrees	Protractor	Research	Trained observer	1	3	0	+	+	++	++	0	0
Universal goniometer [2]	degrees	Universal goniometer	Research and clinical practice	Trained observer	1	0	7	+	+	+	+	++	+/++
Simple inclinometer[3]	degrees	Inclinometer	Research and clinical practice	Trained observer	0	5	2	+	++	+	++	++	++
Myrin inclinometer[4]	degrees	Myrin inclinometer	Research and clinical practice	Trained observer	0	1	3	++	++	+	++	++	++
Tape measure[5]	centimetres		Research and clinical practice	Trained observer	1	0	0	+/++	++	+	++	0	0
Lateral flexion													
Universal goniometer [2]	degrees	Universal goniometer	Research and clinical practice	Trained observer	1	1	0	+	+	+	+	0	0
Tragus to ACJ[2]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	1	0	+	+	+	++	0	0
Simple inclinometer[3]	degrees	Inclinometer	Research and clinical practice	Trained observer	0	2	2	+	+	+	+	++	+
Myrin inclinometer[5]	degrees	Myrin inclinometer	Research and clinical practice	Trained observer	0	1	0	+	+	+	++	0	0
Flexion / extension													
Universal goniometer [2]	degrees	Universal goniometer	Research and clinical practice	Trained observer	1	1	0	+	+	0	0	+	+
Occiput to C7 [6]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	1	0	+	+	0	0	0	0
Chin to chest [6]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	3	2	++	++	+	++	+	++
Simple inclinometer[3]	degrees	Inclinometer	Research and clinical practice	Trained observer	0	2	2	+	+	+	+	+	+

Table 2.57 Cervical mobility - summary of data evaluation. [a] References: [1] Cheshire (1957), [2] AAOS(1965), [3] O'Driscoll et al (1978), [4] Viitenan et al (1992), [5] Viitenan et al (1998), [6] Calcraft et al (1974).
Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

investigators and although all operate on a similar principle, different types may possess different levels of reliability and clarity is needed (Mellin et al, 1994). However, although inclinometers are expensive and not widely accessible, good levels of reliability and evidence of validity have been reported. Good evidence of reliability and validity also exists for the evaluation of cervical rotation with a tape measure: a cheap, quick, portable and widely accepted item of equipment. Due to the difficulty in selecting between these two methodologies the evaluation of cervical rotation with a tape measure and with a simple inclinometer were both selected for inclusion in a pre-pilot evaluation (Chapter 4), to support selection for the comparative study. Difficulty in selecting between available methodologies for the assessment of cervical mobility due to inadequate methodological detail and poor study design has been reported by other investigators (Jordan, 2000), and recommendation to adopt the tape measure as a feasible, cheap and acceptable methodology for cervical assessment was made.

Chest expansion

Limitation in chest expansion, relative to normal values and adjusted for age and sex, is included in the diagnostic criteria for AS (van der Linden et al, 1984), and it is a widely adopted measurement in routine practice (Bellamy et al, 1998, 1999). However, evidence suggests poor standardisation of methodology, poor inter-rater reliability and a level of responsiveness following physiotherapy that may not be clinically significant (Moll and Wright, 1973)(table 2.58). Therefore, although important diagnostically, the measurement is not recommended for the longitudinal evaluation of AS.

Thoracolumbar flexion

Although evidence for reliability is good, there is poor standardisation of methodology, validity is limited and accumulated evidence suggests poor responsiveness (table 2.59). Therefore, the measurement is not recommended for evaluative purposes in AS.

Thoracic rotation

The original methodology requires a large and cumbersome frame, which is not easily portable and limits the feasibility of application in research or clinical practice. Despite promising evidence of measurement properties, the approach has not been

Anthropometric measures													
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
Chest expansion [a]	Scale	Special equipment	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
Nipple level[1]	centimetres	Tape measure	Research, diagnosis and clinical practice	Trained observer	1	2	0	+	+	0	0	0	0
4th intercostal space [2]	centimetres	Tape measure	Research, diagnosis and clinical practice	Trained observer	1	2	7	++	+	++	++	++	+/++
Xiphisternum[3]	centimetres	Tape measure	Research, diagnosis and clinical practice	Trained observer	0	7	8	++	+	++	++	++	+/++

Table 2.58 Chest expansion - summary of data evaluation.

[a] References: [1] Hart et al (1963), [2] Moll & Wright (1972), [3] Tomlinson et al (1986).
Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

					Anthropometric measures								
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
Thoracolumbar mobility [a]	Scale	Special equipment	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
Flexion													
C7 to iliac crest [1]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	1	2	+	+	+	+	+	+
C7 to 10cm proximal to LSJ [2]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	1	0	+	+	0	0	0	0
C7 to SC junction [3]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	0	2	0	0	+	+	+	+
C7 to dimples of Venus [4]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	1	0	+	+	0	0	0	0
C7 to S1 [5]	centimetres	Tape measure	Research and clinical practice	Trained observer	0	1	2	+	++	0	0	+	+
C7 to L5 [6]	centimetres	Tape measure	Research and clinical practice	Trained observer	0	1	1	0	0	+	+	0	0
Rotation													
Rotation frame [7]	degrees	Rotation frame	Research	Trained observer	1	3	2	++	++	++	++	+	+
Pavlaka method [8]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	0	0	+	+	+	+	0	0

Table 2.59 Thoracolumbar mobility - summary of data evaluation.

[a] References: [1] AAOS (1965), [2] Calcraft et al (1974), [3] Hyde (1980), [4] Armstrong et al (1984), [5] Viitenan et al (1992), [6] Averns et al (1996a), [7] Viitenan et al (1993), [8] Viitenan et al (1999)
Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

adopted by investigators outside the development team and was not recommended for inclusion in the longitudinal study (table 2.59). Limited evidence suggests that the Pavlaka method (Viitenan et al, 1999) has good reliability and validity, and improved feasibility.

Fingertip to floor distance

The popularity of the measurement of FFD (anterior flexion) is demonstrated by the large number of published studies including it in evaluation (table 2.60), and the report that more than 60% of clinicians usually or always include it in the longitudinal evaluation of AS patients (Bellamy et al, 1998, 1999). Although originally recommended as a measure of thoracolumbar vertebral flexion (AAOS, 1965), validity is disputed. Many investigators indicate that it does not reflect pure spinal movement due to the influence of soft tissues and hip mobility and is therefore 'clinically unhelpful' in assessment (Pile et al, 1991). However, proponents suggest that it is a useful measure of trunk mobility, and analysis in healthy adults, further supported by radiographic analysis in AS patients (Viitenan et al, 1995a; Averns et al, 1996a) has reported a strong association with 'trunk' and hip flexion, and a low association with vertebral flexion (Kippers and Parker, 1987).

The highest levels of reliability are reported for the simplest methodologies requiring only a tape measure or ruler with the patient standing on the floor (table 2.60). Although only a low to moderate association with AS-specific measures of functional disability have been reported, Kragg et al (1990, 1994) reported that FFD and functional ability were the only two measures to demonstrate significant improvement during a four-month follow-up of physical therapy in AS, and accumulated evidence suggests that it is responsive to change following intensive physiotherapy, but not following drug therapy. The evaluation of FFD with a tape measure and with a vertically mounted ruler were both selected for inclusion in a pre-pilot evaluation (Chapter 4) to support selection for the comparative study.

Lumbar flexion

Limitation in anterior flexion of the lumbar spine, relative to normal values and adjusted for age and sex, is included in the AS diagnostic criteria (van der Linden et al, 1984), and the Modified (15cm) Schober Index (MSI)(Macrae and Wright, 1969) is the preferred approach. Measurement of lumbar flexion with the MSI is the most

					Anthropometric measures								
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
FFD (anterior flexion) [a]	Scale	Special equipment	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
Tape measure, standing on floor [1]	centimetres	Tape measure	Research and clinical practice	Trained observer	0	5	3	++	++	++	+/++	++	+
Vertically mounted ruler, standing on floor [2]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	3	3	++	+/++	++	+/++	++	+
Ruler / tape measure, standing on stool [3]	centimetres	Tape measure	Research	Trained observer	1	1	1	+	+	++	++	0	0
Portable Spinal Mobility Scale [4]	centimetres	Tape measure	Research	Trained observer	1	0	2	+	+	+	+	++	+

Table 2.60 Fingertip to floor distance (anterior flexion) - summary of data evaluation.

[a] References: [1] Miller et al (1984), [2] Tomlinson et al (1986), [3] Kippers & Parker (1987), [4] Stokes et al (1988).
Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

frequently applied measurement in published articles, and in routine clinical practice (Bellamy et al, 1998, 1999). Accumulated evidence suggests that it is a reliable and valid measure of lumbar spine flexion, but it is not responsive to change over the short term following either physiotherapy or drug intervention (table 2.61). The MSI was selected for inclusion in the longitudinal study as a reflection of lumbar spine mobility and long term AS-specific change in the lumbar spine (Chapter 4).

Lumbar extension

Although evidence for the reliability and validity of lumbar extension measured with the skin distraction technique (Moll et al, 1972b) is good, the developers indicate that lumbar flexion (MSI) is a more sensitive index of AS (Moll et al, 1972b), and the method has not been widely applied in the published literature or clinical practice (Bellamy et al, 1998, 1999)(table 2.61). The Dunham spondylometer and Smythe technique have not been widely applied and evidence of measurement properties and feasibility is varied, although a strong association with the MSI has been reported. The evaluation of spinal extension is not recommended as part of routine clinical evaluation and has not been included in the longitudinal study.

Lateral lumbar flexion

Although evidence of reliability is limited, LLF (skin distraction)(Moll et al, 1972a) is widely applied in the published literature. Evidence suggests that the most reliable method is LLF measured with a mounted ruler (Pile et al, 1991)(table 2.61). Radiographic evaluation of the lumbar spine correlates strongly with both techniques, supporting the validity of the methodologies as measurements of LLF. Based on published evidence alone it was difficult to select between these methodologies, and both were selected for inclusion in a pre-pilot evaluation (Chapter 4).

Posture

Occiput (OWD) or tragus to wall distance (TWD) are frequently assessed measures of spinal posture in published studies and in clinical practice, with more than 40% of clinicians usually or always recording the value in routine practice (Bellamy et al, 1998, 1999). Evidence suggests good reliability, and radiographic assessment of the cervical spine demonstrated a moderate correlation of OWD and strong correlation with TWD. Evidence of responsiveness is not conclusive and suggests a low responsiveness following physiotherapy. Measurement of TWD with a t-square was

					Anthropometric measures								
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
Lumbar mobility [a]	Scale	Special equipment	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
Flexion													
Schober 10cm Index [1]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	4	3	+	+	++	++	++	+
Modified Schober Index[2]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	12	18	+++	++/+++	+++	+++	+++	++
Lumbar Flexion Index [3]	centimetres	Flexible ruler	Research	Trained observer	1	1	1	+	+	+	+	+	+
Extension													
Dunham Spondylometer[4]	degrees	Spondylometer	Research	Trained observer	1	2	3	+	+	+	+	+	+
Plumb-line extension [5]	centimetres	Plumb-bob and tape measure	Research and clinical practice	Trained observer	1	3	1	+	+	++	++	+	+
Smythe technique [6]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	2	10	++	++	++	++	++	+
Lateral flexion													
Skin distraction [7]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	5	4	++	+/++	+++	+++	++	+
Fingertip to fibula[8]	% of body height	Tape measure	Research and clinical practice	Trained observer	1	1	0	+	+	0	0	0	0
Fingertip lateral thigh [9]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	1	1	+	+	0	0	+	+
Fingertip to floor (ruler) [10]	centimetres	Mounted ruler and slide	Research and clinical practice	Trained observer	0	3	1	++	++	++	++	0	0

Table 2.61 Lumbar mobility - summary of data evaluation.

[a] References: [1] von Schober (1937), [2] Macrae & Wright (1969), [3] Adrichem & van der Korst (1973), [4] Dunham (1949), [5] Moll et al (1972b), [6] Miller et al (1984), [7] Moll et al (1972a), [8] Little (1986), [9] Domjan et al (1990), [10] Pile et al (1991).
Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

					Anthropometric measures								
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
Upper cervical / spinal posture [a]	Scale	Special equipment	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
Tape - OWD [1]	centimetres	Tape measure	Research and clinical practice	Trained observer	1	5	5	++	++	++/+++	++	+	+
T-square - TWD[2]	centimetres	T-sqaure	Research and clinical practice	Trained observer	0	4	1	++	++	+	++	+	+
PSMS - OWD [3]	centimetres	PSMS and block	Research	Trained observer	1	0	2	+	+	+	+	+	+

Table 2.62 Upper cervical / spinal posture - summary of data evaluation.

[a] References: [1] American Rheumatology Association (ARA) Glossary (1984), [2] Tomlinson et al (1986), [3] Stokes et al (1988).
Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

					Anthropometric measures [a]								
					Published articles (n) - Stage of instrument development [b]			Reliability [c]		Validity [c]		Responsiveness [c]	
	Scale	Special equipment	Application	Administered by (time)	I	II	III	Thoroughness	Results	Thoroughness	Results	Thoroughness	Results
BASMI[1]	Varied - centimetres and degrees	Simple inclinometer, t-square, mounted ruler with slide, tape measure and treatment couch.	Research and clinical practice	Trained observer 7-minutes	1	1	3	+/++	++	++	++	+/++	++

Table 2.63 BASMI - summary of data evaluation.

[a] Reference: [1] Jenkinson et al (1994a)
Superscript: [b] Stage of instrument development I original development, II further testing, III wider application and testing (table 2.5); [c] Reliability / Validity / Responsiveness: 0 = no evidence to +++ = strong evidence.

selected for inclusion in the comparative study (table 2.62) (Chapter 4).

BASMI

Items included in the BASMI were selected as a result of expert opinion and a literature review of evidence. However, this process was very limited, and the developers do not detail selection methdology. Based on the limited published evidence at the time of instrument selection for the comparative study, there was acceptable evidence of the reliability, validity and responsiveness of the BASMI. However, the responsiveness of an index score that includes both measures of reversible and irreversible change in axial status should be addressed, and would be unlikely to be high. In addition, evidence of the acceptability and feasibility of the approach was lacking and several specialised items of equipment are required for administration (table 2.63). Therefore, the BASMI was not selected.

Conclusion

There is no strong recommendation for any anthropometric measure identified and evidence is limited for most methodologies. Many anthropometric measures have limited evidence for reliability, validity and, or responsiveness. Most vary in the need for special equipment and the time and inconvenience for the investigator, clinician or patient. The clinical feasibility and level of acceptability to the patient are important features for all methodologies but are rarely reported by investigators. In addition, there is no consensus on the value of measuring axial status in the longitudinal evaluation of patients with AS (Pile et al, 1991). Four of the nine measures selected for inclusion in a pre-pilot evaluation mirror those included in the BASMI. However, the feasibility, acceptability and reliability of the different approaches for measuring the selected ranges of movement will be assessed (Chapter 4) before a final selection for the comparative study is made.

2.7 Discussion

This review has followed the guidelines for performing a systematic review of randomised controlled trials and the effectiveness of interventions described by the Centre for Reviews and Dissemination (CRD Report 4, 1996) to provide as systematic a review of patient-based and anthropometric measures of outcome as possible. Although structured reviews of outcome measures have been identified, reviews with a similar systematic format to that described in this chapter had not been identified at

the time of the initial search. In 1999 a report, described as a systematic literature review, was detailed by Ruof and Stucki (1999a). Although limited in the extent of the literature search and assessing only two AS-specific measures of functional disability, the method of data extraction would appear to be systematic and the quality of both instruments was assessed in terms of evidence for instrument development and measurement properties. However, detail in the published article is lacking, no reference to standard texts for performing systematic reviews of evidence is made, and there is no attempt to explicitly appraise instrument quality.

In conducting a systematic review of outcome measures it is important that all, or nearly all articles, and all outcome measures of relevance to the focus of the review are identified to limit a biased assessment (Greenhalgh and Meadows, 1999). The review has produced the first comprehensive list of patient-based and anthropometric measures of outcome applied in published studies of AS. The reviews described by Bakker et al (1993b) and extended by van der Heijde et al (1997) focussed only on AS-specific measures of outcome and included single-item measures. The described search was as comprehensive as possible and it is highly unlikely that any major patient-based or anthropometric outcome measures have not been identified. The search was limited by the exclusion of non-English articles and non-Anglicised patient-based measures of outcome, a necessary limitation within the study resources, but one that limits the generalisability of the review. Several English-language articles describe instruments developed in a different language. For example, the Dougadas Functional Index (Dougadas et al, 1988). These instruments have subsequently been applied in English-speaking populations without indication of translation and re-testing of measurement properties. Likewise, other investigators apply English-developed instruments in non-English speaking populations (Hidding et al, 1993a,b; Bakker et al, 1995) without further indication of translation or consideration of cross-cultural differences in item content (Anderson et al, 1995). Where referral was made to the original English-based article describing instrument development, these articles have been included in the review of evidence. However, this highlights the need for investigators to clarify the version of an instrument adopted and any translation and subsequent re-testing of the instrument performed. Translation and cross-cultural differences may result in different measurement properties and requires careful attention when instruments are applied across cultures and languages (Anderson et al, 1995).

Peripheral joint assessment was also excluded from the review. Evidence of the role of peripheral joint assessment in AS evaluation is required, particularly for the hip joint. Several investigators have described the importance of hip joint pathology in advancing AS and its relationship with functional disability and future disease prognosis (Braun et al, 1998 ; Dalyan et al, 1999), and is recommended for inclusion in a further review of anthropometric measurement in AS.

Although anthropometric measurement has been referred to in published articles for many years and recommendation to include lumbar mobility in AS assessment was made in 1968 (Bennett and Wood, 1968), the first application of most measures identified in the review is supported by articles from the late 1980's. This reference indicates the first article where a clear methodological approach could be described, and not the first time that a measurement was referred to in a published article. Evidence supporting the measurement properties of anthropometric measures has been limited by the inadequate or incorrect indexing of measures. Without methodological clarity, evidence supporting application or measurement properties could not be included in the review and has been restricted to that obtained from articles with clear methodological description or reference. The lack of standardisation of anthropometric measurement in AS was first highlighted in 1991 (Laurent et al, 1991), and the review has indicated that this remains a problem. A further recommendation for standardisation of technique and methodological clarity or referencing of anthropometric measurement is made.

Evidence in support of anthropometric measurement could have been enhanced by a structured literature search of electronic databases pre-1990. However, in light of the paucity of referencing or methodological description in both pre and post-1990 literature this would be unlikely to result in a large number of additional articles supporting defined anthropometric measures. This extended search is not necessary for the patient-based measures of outcome because the first AS-specific measure was published in 1987 and all published references have been obtained. Alternatively, the investigators for each article could be contacted to clarify the measurement approaches adopted. This would greatly enhance the availability of data for each anthropometric assessment described and would clarify the version of patient-based measures adopted, but was not feasible within the study resources. Instrument

developers could also be contacted to request information about the application of the respective instrument. However, use of the AS-specific patient-based measures of outcome and all anthropometric measures identified does not require the permission of instrument developers and this approach is unlikely to identify further references beyond the exhaustive literature search described.

Searching the grey literature could have identified a wider range of publications and outcome measures, but the focus of the review was to identify patient-based and anthropometric measures of outcome used in the published literature. However, communication with experts identified several newly developed measures, yet to be published: the Body Chart (Dziedzic K.- personal communication, 1997) provides a disease-specific measure of bodily pain, and the AS Quality of Life questionnaire (ASQoL), an AS-specific measure of HRQL (Doward L. (Galen Research) and Helliwell P.- personal communication, 1998). Both instruments are considered further in Chapter 4.

Following data evaluation no strong recommendation for any patient-based or anthropometric measure of outcome could be made. Evidence of instrument development, measurement properties and levels of acceptability and feasibility of adopting the instruments in clinic based or postal evaluation was limited for most instruments (Haywood et al, 1998). However, instrument selection was based on a systematic, rigorous and explicit methodological approach (Mulrow, 1995; Sutton et al, 1999) and results and conclusions may differ from other reviews of outcome measures due to differences in methodological and selection criteria (Pile et al, 1991; Jenkinson et al, 1994a; van der Heijde et al, 1999a,b,c).

An important step in synthesising the evidence of systematic reviews of randomised clinical trials (RCT) is to assess the quality of each trial (Jadad et al, 1998). This process requires a definition of the quality construct and the adoption of an assessment tool. However, these are controversial issues (Moher et al, 1995; Jadad et al, 1998). The current systematic review has modified this process for the purpose of appraising the evidence in support of instrument quality and has focussed on measurement properties. That is, an evaluative instrument requires evidence in support of instrument reliability, validity and responsiveness (Kirshner and Guyatt, 1985). A scale originally described by McDowell and Newell (1996) in the

evaluation of evidence to support the reliability and validity of reviewed instruments, was adopted for the current study to provide a quantitative summary of the level of evidence to support the thoroughness and results of testing for instrument measurement properties (section 2.4.4). However, in common with scales adopted in the quality assessment of RCTs this scale has not undergone rigorous testing and the summary based upon this scale should be viewed with caution (Moher et al, 1995). In addition, the scale has not been applied by different assessors in the current review and appraisal of assessor agreement was not possible (Jadad et al, 1998). The review has therefore, provided detailed and explicit qualitative evidence (both text and tabular) which may validate the quantitative assessment, whilst providing the reader the opportunity to assess the evidence, and possible instrument hierarchy, irrespective of quality score (Cook et al, 1997; Jadad et al, 1998; Sutton et al, 1999). In addition, the review also considered the development, acceptability and feasibility of each instrument in the data synthesis, an aspect not included in the quality scale assessment.

A different approach to identifying and evaluating instruments for use in AS was adopted by the ASAS group (van der Heijde et al, 1997; 1999a,b,c) and the systematic review. Both methodologies are summarised and compared in table 2.64.

Recommendations made by ASAS were based on the consensus opinion of gathered experts who considered the relevance and feasibility of the instrument (table 2.65) (van der Heijde et al, 1999a,b,c). Selection was not based on a systematic review and explicit evaluation of gathered evidence relating to instrument development, measurement properties, acceptability and feasibility.

Although the systematic review described considered all measurement properties and practical requirements in data evaluation and therefore presents a more detailed and explicit instrument appraisal then the ASAS, data extraction and evaluation was performed by a single investigator (KLH) and may exhibit a biased result. This could be lessened by a second reviewer checking the extraction and evaluation of all, or a random sample of, articles and instruments but was not feasible within the study resources (Jadad et al, 1998).

ASAS	Systematic review
1. Identified available outcome measures in AS (1985–1996) (Medline and citation searches only - non exhaustive search).	1. Identified all patient-based and anthropometric measures of outcome applied in AS (1990-2000) (exhaustive search of major electronic databases; hand searching and citation searching).
2. Identification of core domains by expert opinion - based on domains addressed by listed outcome measures.	2. Domains addressed by identified patient-based and anthropometric outcome measured listed.
3. Group consensus (expert opinion) raised issues of domains not initially selected.	3. Domains considered important in evaluation of AS listed following literature review and expert opinion. Gaps identified.
4. Selection of outcome measures based on appreciation of relevance and feasibility - determined by expert opinion and group consensus, not on review of evidence.	4. Quality of disease-specific and anthropometric measures of outcome appraised in light of systematic evaluation and synthesis of published evidence of development, measurement properties, feasibility and acceptability. Instruments selected to fulfill domains considered important in the evaluation of AS patients.
5. No consideration of generic instruments or measures of HRQL.	5. Literature search and contact with experts identified additional generic and non-published disease-specific instruments to fulfill gaps in domains - specifically HRQL.
6. Outcome measures recommended for use in evaluative studies of AS. Evidence of construct validity, reliability and responsiveness not considered in making recommendation (relevance and feasibility only). ASAS report that this is the next step in the selection process.	6. Selected instruments the focus of comparative empirical study (Chapter 4). Primary objective to further evaluate the practical requirements and measurement properties of identified instruments. Recommendations will be made in support of a standardised package of evaluative patient-based and anthropometric measures of outcome.

Table 2.64 Comparison of approaches adopted by ASAS group and the systematic review for identification of outcome measures in AS.

The literature search conducted by ASAS was not exhaustive and may have failed to identify all available outcome measures applied in AS. Generic measures of HRQL applied in the evaluation of patients with AS have not been identified, and the domains assessed are biased towards the evaluation of impairment and disability. The ASAS included single item measures in the evaluation process and several have been recommended for evaluative purposes (table 2.65). The systematic review explicitly excluded single item measures due to the poor measurement properties inherent in such measures (Fitzpatrick et al, 1998a). However, single item measures are widely applied in both research and clinical practice and a further review to establish the variety of single items adopted, and the available evidence in support of measurement properties would be beneficial to support or reject the use of these measures.

Instruments recommended by ASAS and following the systematic review are summarised in table 2.65.

ASAS (1999)		Systematic review (1998)	
Domains	Instrument	Domains	Instrument
Function	Dougadas Functional Index BASFI	Clinical Functional status	Revised Leeds Disability Questionnaire (Dougadas Function Index)
Pain	i. VAS, last week, spine, at night, due to AS. ii. VAS, last week, spine, due to AS.	Clinical Pain	i. Body Chart-global pain, now.
Spinal mobility	Chest expansion Modified Schober (10cm) Occiput to wall	Clinical Spinal mobility	Cervical rotation Modified Schober Index (15cm) Tragus to wall Fingertip to floor distance: i. forward flexion ii. lateral lumbar flexion
Patient global	VAS, last week	Patient global	Addressed in PGI-AS[a]
Stiffness	Duration morning stiffness, spine, last week	Stiffness	Addressed in ASQoL[b], BASDAI[b] and PGI-AS[a]
Peripheral joints and entheses	Number of swollen joints (44 joint count); No preferred instrument for assessment of entheses	Peripheral joints and entheses	Not assessed. Review - non-conclusive evidence for entheses indices
Fatigue	No preferred instrument available	Fatigue	Addressed in ASQoL[b], BASDAI[b] and PGI-AS[a]
-	-	Clinical Disease activity	BASDAI
-	-	Disease-specific HRQL psychometric	Ankylosing Spondylitis – Quality of Life (AS-QoL)[b]
-	-	Disease-specific HRQL individualised	Patient Generated Index – Ankylosing Spondylitis: (PGI-AS)[a]
-	-	Generic Psychometric	SF-12[c]
-	-	Generic Utility	EuroQol[c]

Table 2.65 Domains and measures of outcome identified by ASAS (van der Heijde et al, 1999a) and the systematic review.

Superscript: a Development of PGI-AS described in Chapter 3;
b,c: ASQoL, Body Chart and generic instruments (EuroQol and SF-12) discussed in Chapter 4.

Conclusion

The selection of patient-based and anthropometric measures of outcome proposed by the current study presents the first evidence-based selection of instruments based on a methodologically rigorous and explicit systematic review of the literature.
Limitations of the review have been identified and recommendations for future systematic reviews of measures of outcome made. The measurement properties of the selected instruments will be further evaluated in the comparative study described in Chapter 4 before recommendations for a package of patient-based and anthropometric measures of outcome suitable for use in routine clinical evaluation of AS and research are made.

Chapter 3 The Development of a Patient Generated Index for AS (PGI-AS)

3.1 Introduction

This chapter describes the development of the first individualised measure of AS-related quality of life, the Patient-Generated Index for AS (PGI-AS). Section 3.2 discusses the measurement of HRQL in AS and section 3.3 describes the stages in the development of the PGI-AS. The chapter closes with a discussion.

3.2 Health-related quality of life in Ankylosing Spondylitis

The review of outcome measures in AS (Chapter 2) has demonstrated that the most usual means of evaluation in published research and routine practice (Bellamy et al, 1998, 1999) remain focused towards the measurement of impairment (disease activity, anthropometric measurement) and disability (functional disability questionnaires). Although several investigators have involved patients in the development of patient-based measures of functional disability this involvement has been limited and dominated by expert opinion. Instrument content therefore remains biased towards the beliefs of health professionals and does not necessarily address aspects of AS considered important by patients.

Although several generic measures of HRQL and domain-specific instruments that evaluate individual domains related to the overall concept of HRQL were identified (table 2.9) an AS-specific measure of HRQL was not identified in the initial review. The core evaluative domains recommended by the ASAS group (van der Heijde et al, 1997, 1999a,b,c) were based upon knowledge of the available outcome measures and not surprisingly are heavily weighted towards the evaluation of impairment and disability. Although the group subsequently acknowledged the significance of quality of life in evaluation, it was not recommended as a core domain due to the 'novelty' of the measurement in AS and 'uncertainty over the best measurement technique' (van der Heijde et al, 1997). The evaluation of quality of life should, in theory, consider a very broad concept of life that does not only focus on the impact of ill-health and several authors suggest that the focus of HRQL evaluation is towards aspects of life that might, in principle, be influenced by health and health care (Ware, 1997; Jenkinson et al,1998).

The HRQL of a patient with a chronic, incurable disease such as AS is considered by many to be an important indicator of disease impact at an individual level (Guyatt et al, 1993; Aronson, 1997). Although many developers involve patients in item generation to ensure the representation of patient concerns, most instruments adopt a closed completion format and patients respond to all listed items. Although these instruments often have good measurement properties such highly standardised instruments may omit issues of importance to individual patients (Stratford et al, 1995; Carr et al, 1996) whilst containing items of little relevance to others, thus introducing noise to the evaluation (Tugwell et al, 1987). Items may appear detached from the contextual setting thus loosing the social or personal significance that may be afforded by a more individualised evaluation (Carr and Thompson, 1994). HRQL is specific to an individual, to their priorities, expectations and experience of life and ill-health. Therefore, many investigators suggest that evaluation should be individually tailored to provide a more meaningful assessment of patient-specific HRQL. A patient-centred evaluation fosters appreciation of the impact of ill-health on an individuals expectations and aspirations as well as recording the physical and psychosocial impact of disease (Carr, 1996). It may therefore be more sensitive to a patients needs, demands and change in status (Barlow et al, 1993a).

Several instruments have been identified which attempt to provide a more patient-centred and individualised evaluation of HRQL. For example, the Chronic Respiratory Distress Questionnaire (CRD - Guyatt et al, 1987a), Disease Repercussion Profile (DRP - Carr and Thompson, 1994) and the Measure Yourself Medical Outcomes Profile (MYMOP - Paterson, 1996). The Schedule for the Evaluation of Individual Quality of Life (SEIQoL - O'Boyle et al, 1993) and the shorter SEIQoL-DW (direct weighting)(Hickey et al, 1996) provide two patient centred generic measures of quality of life. Patients are generally asked to individually nominate specific items adversely affected by ill-health or to identify areas of life that they consider to be important. Items are then rated to represent the extent of disease impact. Some instruments, such as the SEIQoL-DW, take a further step and ask patients to weight the relative impact or importance of each item. All of these instruments require interview-administration and some require a considerable completion time. Although representing an important development in the evaluation of HRQL many of these instruments have not been widely accepted into research or

routine practice. Many are still in development and further evidence of feasibility, acceptability and measurement properties are required.

3.3 The Patient Generated Index (PGI)

An additional patient-centred instrument, the Patient Generated Index (PGI) has also been described (Ruta et al, 1994a). The PGI offers a generic approach to the evaluation of HRQL which is made disease-specific by the inclusion of a disease-specific trigger list. Although a single item addresses non-health areas of life, the PGI is more specifically focussed towards the assessment of disease-related quality of life due to the instruction for patients to consider the most important areas of life affected by a specific disease.

In the management and evaluation of patients medicine must not forget patient autonomy. The PGI supports this autonomy by providing the opportunity or freedom for patients to identify areas of their life that they deem to be of greatest importance when considering disease impact. The PGI quantifies individual disease-related priorities in terms of the effect of disease on a patients day-to-day life. The conceptual base behind the PGI defines quality of life as:

> 'the extent to which our hopes and ambitions are matched and fulfilled by experience'
> (Calman, 1984).

If effective health care is viewed as an attempt to improve a patients HRQL, the result may be a reduction in the gap between a patients 'hopes and expectations and what actually happens' (Ruta et al, 1994a). This may be of particular relevance in a disease such as AS where a cure is not possible and management must focus on the control of disease activity and symptomology, and thus in reducing or minimising disease impact. AS can have an unpredictable progress and patients require monitoring and medical care for the remainder of their life span (Rigby, 1991; Barlow et al, 1993a). Patients are required to make considerable psychological, emotional and physical adjustments and management may be required to assist patients in a revision of life expectation, thus affecting a reduction in the gap between expectation and reality which, it would be hoped, could lead to an improvement in patient-centred HRQL. Successful adaptation to chronic illness has been shown to be positively influenced by a patient's perception of chronic disease and the impact on their life (Carr, 1996).

The PGI was chosen as the basis for a new AS-specific measure of disease-related quality of life for several reasons. Firstly, it offers a patient-centred approach to the evaluation of ill-health which allows issues considered important by a patient to be incorporated in the evaluation. Second, it can be adapted to provide an AS-specific evaluation of disease-related quality of life. Third, it has good evidence of development and measurement properties in varied patient populations (Ruta et al, 1994a; Herd et al, 1997; Ruta et al, 1999; Jenkinson et al, 1998b) and in populations with similar pathologies to AS (McArthur, 1997, cited by Macduff and Russell, 1998). Particular strengths have been reported to lie with the content validity and instrument responsiveness (Ruta, 1998). Finally, it has been administered in both self-administered (postal) and interview-based evaluations and good levels of feasibility and acceptability have been reported. This was an important consideration for the comparative study (Chapter 4).

Since the PGI was first published in 1994 (Ruta et al, 1994a) the instrument has been revised to improve feasibility and patient acceptance (Cotton et al, 1993; Ruta, 1998). In addition, several 'hybrid' versions have been proposed. Patients with multiple health problems of an unrelated nature reportedly had difficulty completing the original PGI and a disease-specific format was proposed to allow patients to consider health problems not directly associated with the specific focus of the PGI (Ruta, 1998). AS covers a wide clinical spectrum but it was considered important to allow patients to report the impact of non-AS related health issues and the disease-focused PGI was selected for the study. An additional single item in the disease-focused PGI allows the patient to consider the impact of health problems not related to the specific disease.

Completion of the PGI is in three steps (Appendix 2 - PGI-AS). The first step asks the patient to identify the most important areas of life that are affected by *the specific disease*, for example AS. The patient may write up to five areas in the boxes provided. To assist in item generation a trigger list of important areas commonly mentioned by other people with the specific disease is provided on an adjacent page for ease of reference, together with a completed example of the PGI. The last two boxes ask the patient to consider the impact of health problems other than AS (box 6),

and all other non-health related areas of life (box 7). Items are not written in these boxes but the areas are scored in step 2.

In the second step the patient is asked to score the areas mentioned in step 1. The score illustrates the impact of the identified area on the patient over the previous one month. A scale from 0 ('the worst you could imagine') to 10 ('exactly as you would like to be'), with verbal descriptors for each level is provided, and each identified area must be scored. Should the patient not experience any other health problems (box 6) they may indicate 'none' in writing without a further need to score this box (as illustrated in the completed example). However, this box should not be left blank. Box 7 ('all other non-health related areas of life') must be scored by all patients.

In the final step, step 3, the patient is asked to consider that any or all of the areas of their life could be improved. Points are spent to reflect the relative importance of each identified area with more points spent on areas where an improvement would be most valued. Points do not need to be spent in each listed area and all 14 points may be spent in one area if so desired. If no AS-specific items are identified all points must be spent between box 6 and / or 7.

An index score is generated using the following equation:

$$\text{(Step 2 score} \times \frac{\text{Step 3 points)}}{14}$$

This is calculated for each item where points are spent. The result is then totalled (0-10), where a lower score represents a wider gap between expectation and reality, and a lower patient generated disease-related quality of life. The score relates to the response scale in step 2.

3.3.1 Developing a trigger list for the PGI-AS

The first stage in adapting the PGI for use in AS was to develop a trigger list of the most important areas of life affected by AS as determined by a representative population of AS patients.

40 patients were randomly selected from the AS database at the Staffordshire Rheumatology Centre (SRC)(33 male; mean age 46.19 years, SD 10.10; range 28 - 69

years). Although the severity of a condition is suggested not to be synonymous with a patients level of quality of life (Whalley et al, 1997; Aronson, 1997) the random sample was intended to provide a representative sample of patients with AS.

A letter was sent to all patients inviting them to attend the SRC for a 'chat' about the affect of AS on day-to-day life. The letter detailed the purpose and expected duration of the interviews. It was stressed that patients were under no obligation to participate and that the study would not impact upon normal management. Appointment times ranging from two to four weeks after the letter was posted were listed and patients asked to nominate a date and time (a tick box option) or to indicate a preferred appointment in writing, and to return both informed consent and appointment forms to the lead investigator (KLH) in the reply-paid envelope. Appointments were confirmed by telephone. Patients not wishing to participate were asked to return the pre-coded consent form. Non-responders were sent reminders and revised appointments at two weeks and again after four weeks.

29 patients participated in the semi-structured qualitative interviews with the lead investigator (KLH)(table 3.1). 24 patients were male (82.8%)(mean age 48.41 years, SD 10.12, range 31-69 years), with a mean duration of AS diagnosis of 11 years (SD 10.68, range 2-41 years), suggesting a broad spectrum of disease presentation covered by the population. The structure and objectives of the pilot interviews did not differ from the main interviews and information was incorporated in the generation of the trigger list.

	Pilot (n= 4)		Main (n= 36)		Total (n= 40)	
	n	%	n	%	n	%
Response rate	4	100	25	69.4	29	72.5
Refusal	0	-	6	16.6	6	15.0
Non-response	0	-	5	13.8	5	12.5

Table 3.1 Response rates for qualitative interviews.

To test for response bias patients who failed to respond or refused to participate in the interviews were compared with respondents by age and gender. Patients not taking part were significantly younger than responders (mean age non-responders 40.33 years, SD 7.68, range 28-52 years)(t-test p= 0.02), but there was no significant difference in gender (Fisher's exact test, p= 0.64).

The aim of the interview was to elicit a patients free responses about the impact of AS on their everyday life and the importance of the areas affected by AS. Life priorities affected by AS were considered. Interviews were performed in a private room at the SRC and lasted between 30 minutes and one hour. The sample size was supported by work by other investigators (de Jong et al, 1997; Jenkinson et al, 1998a) and no new significant themes emerged during the last few interviews.

With the permission of the patient interviews were audio-recorded and later transcribed (Whalley et al, 1997). Verbatim statements were listed that readily identified important and common themes related to the HRQL of patients with AS. A total of 99 areas of life affected by AS were identified. In the first instance the frequency endorsement of individual items was determined (Guyatt et al, 1987a). The top 22 most frequently mentioned areas are shown in table 3.2.

	Items	Frequency endorsement
Impact on ability to work	Ability to pursue chosen hobbies or sports	24
Difficulty driving		19
Relationship with partner	Worry about the future impact of AS.	17
Feeling depressed		
Worry about deterioration in condition	Difficulty walking	16
Feeling tired / fatigued	Ability to remain physically mobile / Reduced spinal movement	15
Constant pain	Loss of independence	14
Ability to do things / jobs around the home	Social life	13
Ability to do 'D.I.Y'		
Disturbed sleep	Ability to do housework	12
Loss of motivation to do things / Feelings of lethargy		
Increased dependency on partner	Ability to lift heavy weights / carry the shopping	11
Feeling that life is controlled by AS		

Table 3.2 Frequency endorsement of the most important areas of life affected by AS (maximum = 29).

Data analysis allowed for the generation of conceptual categories (Bowling 1997). Related items were highlighted, grouped together and organised by category. This listing was discussed with a member of the research advisory group (RAG)(AG) and scrutinised for repetition and ambiguity. It was possible for certain items to be

represented in more than one category and the final decision for placement was taken by the lead investigator only (KLH). When related items were grouped together 16 main categories were described (table 3.3). The hierarchy of the list is based purely on the frequency of endorsement for each category (range 5-66).

Category	Frequency endorsement	Category	Frequency endorsement
Relationship with family / partner	66	Self-esteem	44
Pain	57	Worry about the future	43
Functional activities	56	Tiredness / Fatigue	39
Control over life	55	Depression / moody	33
'Jobs' around the home	50	Leisure activities	25
Level of independence	47	Driving	22
Impact on work	45	Limited spinal movement	21
Social life / friendships	45	Mental agility	5

Table 3.3 Frequency endorsement of areas of life affected by AS by category.

The final trigger list reflects these categories and contains 37 items frequently mentioned by patients with AS (table 3.4). The number of items was guided by the available space in the formatted instrument (Appendix 2) and selection was based upon the frequency of endorsement. Where possible the list includes verbatim statements made by patients. If this was not possible statements closely resemble those made by patients.

3.3.2 Pre-pilot study

The pre-pilot study involved interview-administration of the index in a clinic environment (n= 10)(9 male; mean age 47.8 years, SD 8.75; range 28-58 years) and patient self-completion in their home in the form of a postal response (n=10)(9 male; mean age 47.7 years, SD 12.91; range 29-69 years). Participants consisted of a further random sample of patients from the SRC AS database. Completion of the index was followed by semi-structured interviews with the lead investigator at the SRC to identify any ambiguities in the index, to ensure that it could be easily understood and completed and to invite the patients to comment on the content and use of the trigger list. Patients participating in the postal survey were asked to attend the SRC for a follow-up interview at a time convenient to them.

Trigger list			
Impact on work	Disturbed sleep	Increased time to do things	Walking
Relationship with husband / wife / partner	Pain	Control over life	Difficulty sitting down / standing / lying down
Ability to play with children / grandchildren	Feelings of low self-esteem	Ability to plan ahead	Ability to remain physically active
Sex life	Embarrassment	Enjoyment of life	Fear of Falling
Family life	Poor self body-image	Worry about the future	Dressing
Worry over 'letting people down'	Fatigue	Pursuing chosen hobbies	Washing
Level of independence	Feeling Tired	Sporting activities	Ability to do jobs around the home
Relationship with friends	Loss of motivation		Limited spinal movement
Social Life	Depression		Difficulty 'getting going' in the morning
Driving	Moody		
	Mental activity		

Table 3.4 PGI-AS trigger list.

Five patients were interviewed following interview-administration of the instrument (mean age 43.00 years, SD 10.12; range 28-55 years) and a further six following the postal survey (mean age 46.00 years, SD 15.68; range 29-64 years). There was no significant difference in gender (Fisher's exact test, p= 0.50) or age between responders and non-responders for either the clinic-based (t-test p= 0.08) or the postal survey (p= 0.70). Verbatim statements were recorded. No substantial amendment to the PGI-AS was required in light of the evaluation and only minor modification to the wording in step 3. Positive feedback suggested that the trigger list was beneficial in assisting patients to identify the most important areas of life affected by the disease.

3.4 Discussion

The open and dynamic nature of the PGI-AS places the patient at the centre of the evaluative process and is proposed as the first individualised measure of AS-related quality of life. The PGI-AS was developed to provide a sufficiently short and simple instrument that would be feasible for application following self-completion in postal surveys and following interview-administration within a clinic setting. However, there is no clear consensus on the most appropriate format to adopt for follow-up completion. That is, whether patients should complete the instrument blind to areas identified at baseline completion ('blind'), informed of areas identified, but allowed to

change areas if desired ('informed and open') or informed of the areas, but not allowed to change them ('closed'). Evaluation of these different formats is essential to determine which approach has the greatest relevance to patients, researchers or clinicians (Jenkinson et al, 1998c).

It is acknowledged that the frequency with which a concept is mentioned during semi-structured interviews does not necessarily equate with the social significance of the topic (Bowling, 1997). However, the trigger list acts only as a prompt for patients completing the PGI-AS. Item selection and subsequent weighting is specifically individualised. In addition, the content of the PGI-AS trigger list addresses a wide diversity of areas such as relationships with family, fear of falling, ability to plan ahead and the level of social embarrassment associated with poor posture and reduced mobility. It also captures patients concern about the future direct and indirect consequences of the disease. For example, the impact of disease on the ability to work and the resulting financial impact; plus the impact on marital and family relationships. Many items are distinctively associated with AS although many may differ over time and between patients, a feature common with other patient-centred measures of HRQL (Carr, 1996; Jenkinson et al, 1998a). The conceptual base of the PGI-AS suggests that the patient whose life is being assessed is most qualified to judge its quality (Ruta, 1998), and comments such as:

> 'That's me! That's Spondy in a nut-shell'
> and
> 'the list made me feel that I am not alone with the AS and with the problems that I experience.'

support the content validity of the PGI-AS. In addition, this support for the trigger list and the individuality of instrument completion suggests that the age difference between responders and non-responders in developing the trigger list should not adversely influence instrument completion.

Before the PGI-AS can be recommended for use in the evaluation of AS patients in clinical research or routine practice evidence for the measurement properties, acceptability and feasibility of the instrument is required. The comparative study described in Chapter 4 has provided the first evidence for these properties.

Chapter 4 Developing a package of outcome measures for use in AS

4.1 Introduction

This chapter describes the methodology and results of the patient recruitment and data collection for both clinic-based and postal surveys of the comparative study. The pre-pilot evaluation of the anthropometric measures identified in the systematic review is described in section 4.2 and the selected study instruments are summarised in section 4.3. The aims and objectives of the comparative study are described in section 4.4 and patient inclusion and exclusion criteria is described in section 4.5. Sections 4.6 and 4.7 describe the survey methodology and results of the clinic-based and postal surveys respectively, on which the empirical work that follows is based. The chapter concludes with a discussion of the results.

4.2 Pre-pilot evaluation of anthropometric measures

Nine anthropometric measures were selected in the systematic review (Chapter 2). A pre-pilot evaluation of these measures assessed the clinical feasibility as determined by the expert opinion of three observers (two experienced physiotherapists, one 'trained' non-physiotherapist), and the intra and inter-observer reliability. From a random sample of 20 AS out-patients from the SRC database, 12 agreed to participate in the study (n= 12 males; mean age 48.50 years, SD 9.13; range 36-69 years; symptom duration 4-43 years, mean 21.25 years, SD 12.51; duration of diagnosis 3-42 years, mean 12.9 years, SD 11.04) (Haywood et al, 1999). There was no significant difference in age between responders and non-responders (t-test p= 0.12), or for gender (Fisher's exact test, p = 0.49).

Four methodologies were excluded due to poor clinical feasibility: cervical rotation (inclinometer), Fingertip to floor distance (FFD)(vertically mounted ruler), Lateral lumbar flexion (LLF)(skin distraction) and Tragus to wall distance (TWD)(t-square). The inclinometer was rejected due to the necessity for patients to lie supine, a position often found to be painful and difficult for patients with AS and therefore difficult to standardise. The vertically mounted ruler used to measure FFD (and LLF) was bulky, limiting movement of the most flexible patients and those of shorter stature. The identification and standardisation of landmarks for the LLF skin distraction technique was time consuming and difficult in obese patients and those with more pronounced spinal deformity. Measurement of FFD and LLF as fingertip to floor distance using a

retractable steel tape measure was more acceptable, easier and quicker. Finally, the t-square used to measure TWD failed to reach the tragus of the most severely affected patients and was replaced with a retractable steel tape measure. The measurement of FFD (anterior flexion), LLF (fingertip to floor distance) and TWD with a steel, retractable tape measure offered a compromise between the use of a tape measure, and the rigidity of the mounted ruler. The solid base of the steel tape measure ensured that the position of the ruler, perpendicular to the floor (or to the wall in measuring TWD) could be maintained, whilst offering a cheap, quick, readily portable and adaptable instrument. The tape measure assessment of cervical rotation (Viitenan et al, 1998) was modified, identifying the tip of the nose as a more fixed facial landmark than the chin. High levels of reliability were found for the selected measurements (ICC > 0.85) and all five were retained for the comparative study (table 4.1).

4.3 Measures of health outcome

Instruments were selected for inclusion in the comparative study to represent a core set of health domains considered important in the evaluation of patients with AS. These domains were supported by the work of the ASAS working group (table 2.1)(van der Heijde et al, 1997; van der Linden and van der Heijde, 1998), a further search of the literature relating to the measurement of health outcome (Fitzpatrick, 1993a; Carr, 1996; Jenkinson et al, 1998a) and expert opinion (RAG).

Disease-specific

Two disease-specific patient-based instruments to measure functional disability and disease activity and five anthropometric measures of axial status were selected as a result of the systematic review (Chapter 2) and the pre-pilot evaluation (section 4.2) (table 4.1).

Expert opinion

The systematic review did not identify disease-specific instruments to reflect all important domains: that is, pain and HRQL, and communication with measurement experts in rheumatology and AS identified two, as yet unpublished AS-specific patient-based instruments: the Body Chart and the AS Quality of Life questionnaire (ASQoL). The Body Chart is a measure of global bodily pain and is routinely administered in the clinical assessment of AS patients at the SRC (Dziedzic, 1997). The instrument is interview-administered and consists of a body manikin (anterior and

Category	Instrument	Description
Clinical Functional status	Revised Leeds Disability Questionnaire – RLDQ (Abbott et al, 1994)	16 items - AS-specific functional disability 4-point ordinal response scale - perceived activity completion Response 'Yes, with no difficulty' (0) to 'Unable to do' (3). Items totalled. Score 0-48; 0 is better functional ability
Clinical Disease status	Bath Ankylosing Spondylitis Disease Activity Index – BASDAI (Garrett et al, 1994)	6 items - AS disease activity. Response - 6x 10cm horizontal visual analogue scale (VAS) (anchors 'none' and 'very severe'). Mean of items 5 and 6, plus items 1-4; total divided by 5. Score 0-10; 0 is less disease activity.
Clinical Pain	Body Chart in Ankylosing Spondylitis (Dziedzic, 1997)	Global bodily pain. Patient sketches area/areas of 'current or present pain'. Each area scored (range '1 – mild' to '4 – very severe'). Final score is sum of pain intensity scores. Score 0+; 0 is no bodily pain.
Disease specific Psychometric	Ankylosing Spondylitis Quality of Life - ASQoL (Doward et al, 1998)	18 items - AS-specific HRQL Response 'Yes' or 'No'. Score 0-18; 0 is better HRQL
Disease Specific Individualised	Patient Generated Index – Ankylosing Spondylitis PGI-AS	Up to 5 'most important' areas of life affected by AS identified. Areas scored on 10 point descriptive scale (range 0 -'the worst you could imagine' to 10 -'exactly as you would like to be'). Points spent to reflect relative importance of each problem. Index score calculated (0-100); 0 is a worse level of AS-related quality of life.
Generic Psychometric	Short Form-12 Health Survey – SF-12 (Ware et al, 1995)	Shortened version of the SF-36. 12 items with Likert type responses - produce physical and mental component scales. Score 0-100; 0 is worse HRQL.
Generic Utility	EuroQol Quality of Life Scale – EuroQol (The EuroQol Group, 1990)	EQ-5D: expresses HRQL in a single index score. Covers 5 dimensions of health – mobility, self-care, role/main activity, family and leisure activities, pain and mood. Score -0.59-1.0; 1.0 is best HRQL. Thermometer: measure of health status separately represented on a vertical 20cm 'thermometer' scale. Score 0-100; 0 worst perceived possible health state.
Clinical Lumbar flexion	Modified Schober Index (15cm) (Macrae & Wright, 1969)	Distance between two marks placed 15cm apart in standing (10cm proximal and 5cm distal to the PSIS) following maximal forward flexion of the spine (plastic tape measure)
Clinical Trunk mobility	Fingertip to floor distance (trunk forward flexion)	Distance between tip of right middle finger and the floor following maximal lumbar flexion, whilst maintaining knee extension. Measured with a retractable steel tape measure.
Clinical Lumbar mobility	Lumbar lateral flexion	Distance between tip of ipsilateral middle finger and floor following maximal lateral flexion, maintaining heel contact with floor and without trunk rotation. Measured with a retractable steel tape measure.
Clinical Upper cervical / spinal posture	Tragus to wall distance	Horizontal distance between right tragus and wall in standing, knees extended and chin drawn in. Measured with a retractable steel tape measure.
Clinical Cervical mobility	Cervical rotation	The difference between tip of the nose and ACJ in sitting (neutral): difference between neutral position and maximal rotation to ipsilateral side calculated for right/left rotation. Measured with plastic tape measure.

Table 4.1 Patient-based and anthropometric study instruments.

posterior views) onto which patients sketch or draw the area or areas of 'current or present' pain (Appendix 3). Each area is scored from a four-point ordinal scale (1=mild pain to 4= very severe pain). Areas are totalled; a lower score indicates less bodily pain. There is no maximum score. The Body Chart has been tested in a clinic-based study and preliminary evidence suggests satisfactory measurement properties, acceptability and feasibility (Dziedzic, 1997). Although not previously applied in a

self-completed format, the instrument is not complicated and it was considered suitable for inclusion in the postal survey.

The ASQoL is a newly developed AS-specific measure of HRQL following a closed item format and needs-based model of assessment (Helliwell P.- personal communication, 1998; Doward L. (Galen Research) - personal communication, 1998)(Reynolds et al, 1999). The instrument was developed following patient-based interviews but published detail is not yet available. It consists of 18 items with dichotomous response options (yes / no) (Appendix 3). Items are totalled (0-18); a lower score indicating a better level of AS-specific HRQL. This is the first AS-specific measure of HRQL to be identified and was therefore, with the permission of the developers (Galen Research), incorporated into the package of outcome measures.

The PGI-AS (Chapter 3) was incorporated in the package of instruments to provide the first individualised measure of disease-related quality of life in AS. The choice of PGI-AS format for follow-up completion in the clinic or postal survey was dictated by the practicalities of entering baseline data into the follow-up questionnaires. The 'closed' and 'informed and open' formats required all baseline areas (step 1) to be manually entered into the follow-up questionnaire. This was the responsibility of the lead investigator (KLH). Fewer patients participated in the clinic based test-retest survey; therefore, all patients completed the 'informed and open' format of the PGI-AS. Respondents to the two-week postal survey were randomly assigned to receive 'blind' or 'closed' formats. Respondents to both clinic and postal surveys at six months completed either 'blind' or 'informed and open' formats. A decision not to include the 'closed' format at six months was made to reflect the open nature of the PGI-AS and is discussed further in Chapter 7.

Generic instruments

Generic instruments have not been widely applied in AS, but six generic patient-based instruments including three health profiles and three utility measures were identified in the review (table 2.9). There is limited evidence for their measurement properties in AS and all instruments are very long, and therefore not acceptable for inclusion in the patient-completed package of instruments proposed for the comparative study. For example, the Sickness Impact Profile (SIP)(Bergner et al, 1976) contains 136 items and the Short-Form 36-item Health Survey (SF-36)(Ware, 1997) contains 36

items. The utility measures are specifically designed for cost-benefit studies and involve complex and time-consuming procedures (Carr, 1996) and are not practical for use in the proposed study.

The need for a short, comprehensive instrument required a further literature search and the Short Form 12-item Health Survey (SF-12)(Ware et al, 1995) and the EuroQol (EuroQol group, 1990) were identified. The two instruments represent quite different approaches to assessing overall health: the SF-12 is a health profile and the EuroQol (EQ-5D) a utility measure. There is debate over which is the most appropriate in evaluation (Jenkinson et al, 1996) and so both were included in the study. Although not applied in patients with AS both instruments have good evidence of development and measurement properties and have been applied in the evaluation of patients with disease of similar nature to AS (Hurst et al, 1997, 1998; Ruta et al, 1998; Coons et al, 2000). The SF-12 is a shortened version of the SF-36, containing only 12-items with Likert type responses. It is based on the psychometric approach to instrument construction and produces two summary scales (physical and mental health) with a score range based on the general population (range 0-100, mean 50, standard deviation10), where a higher score indicates a better HRQL. It takes approximately one to two minutes to complete and covers two sides of A4 paper. The EuroQol is based on a model including health state valuations and has greater potential for application in economic evaluation (Garratt, 2000). It contains two sections; the first (EQ-5D) has five items covering the domains of mobility, self-care, usual activity, pain / discomfort and anxiety / depression. Each item has a three-point response scale (1= no problems to 3= inability / extreme problems). In total 243 possible health states (3^5) are reflected by the EQ-5D and weighted values generated by a healthy population have been calculated. The index score rates HRQL on a continuum between -0.59 and 1.00, where 1.00 is perfect health, but a score of less than 0 is a state worse than death. The second section includes a vertical thermometer on which the patient records their overall perceived health 'today' (0 = worst imaginable to 100 = best imaginable). The multi-dimensional generic nature of the SF-12 and EuroQol cover many of the issues addressed by the several domain-specific instruments also listed in table 2.9, and so these instruments have not been considered further.

The choice of generic instruments has recently been supported by a comparative review of evidence supporting the development and measurement properties of the

most widely applied generic 'quality of life' instruments (Coons et al, 2000). A high level of evidence for the SF-36 beyond that available for other profile instruments such as the SIP, the Nottingham Health Profile (NHP)(Hunt et al, 1981) and COOP charts (Nelson et al, 1987) was found, and the EuroQol was found to be comparable to other utility instruments.

The same version of each patient-based instrument was incorporated into a self-administered questionnaire to be completed in both clinic and postal surveys (Appendix 3). Demographic information relating to age, gender, AS symptom duration, year of AS diagnosis, employment status, extent of education, and marital and housing status was requested. In addition, patients were asked to report if they had required any assistance in completing the questionnaire. The same versions of the PGI-AS and Body Chart were interview-administered in the clinic survey. Anthropometric measures were included in the clinic survey only.

4.4 Aims and objectives

The study aims to evaluate the selected patient-based and anthropometric measures of outcome in patients with AS to determine if each instrument fulfills its proposed role given the established evidence and claims of the developers. The study will also provide the first evidence in support of the measurement properties, acceptability and feasibility of the PGI-AS.

The primary aim of the study is to provide an empirical comparison of evidence in support of the measurement properties, acceptability and feasibility of all selected instruments in the described population. The objectives are addressed in subsequent chapters and are as follows:

1) To assess the data quality, scaling assumptions and reliability of the study instruments (Chapter 5)
2) To assess the validity of the study instruments (Chapter 6)
3) To assess the responsiveness of the study instruments (Chapter 7)

A further aim of the study will be to provide additional evidence in support of a standardised and evidence-based package of patient-based and anthropometric measures of outcome for use in AS evaluation in clinical research and routine

practice. Results will be considered in light of the state of health care evaluation in general and specifically within AS.

4.5 Patient population

Patients with AS are generally referred to specialist centres for diagnosis and management. All rheumatology centres approached for participation in the survey were identified as centres of excellence in the management of AS. The study population is expected to represent a wide disease spectrum of diagnosed disease.

4.5.1 Inclusion and exclusion criteria

Patients diagnosed by rheumatology specialists have a greater likelihood of conforming to established diagnostic criteria (Gran & Husby, 1998) and all rheumatologists confirmed a primary diagnosis of AS (Modified New York Criteria - van der Linden et al, 1984) (table 4.2). 'Probable' or 'possible' AS was excluded.

Modified New York Criteria for AS			
Clinical Criteria			Grading
Low back pain > 3 months	Limitation of the lumbar spine in frontal and sagittal planes	Limitation of chest expansion (age/sex related)	Definite – x-ray plus 1 clinical criteria Probable – 3 clinical criteria
X-Ray criteria			Grading
Sacroiliitis >/= grade 2 bilaterally or grade 3-4 unilaterally			Definite – x-ray plus 1 clinical criteria Probable – x-ray criteria

Table 4.2 Diagnostic criteria (van der Linden et al, 1984).

Patients were also excluded if they fulfilled any of the conditions listed in table 4.3. Unaided questionnaire completion was required for both postal and clinic surveys, and an understanding of the English language was important

Exclusion criteria	Reason in support of exclusion
'Probable' or 'possible' AS	Unclear diagnosis
Pregnancy	Impact of pregnancy on HRQL
Inability to comprehend the English language Learning difficulties	Inability to self-complete questionnaire
Less than 18 years of age	Juvenile expression of AS
More than 75 years of age	? influence of co-morbidity on HRQL ? ability to complete questionnaires unaided

Table 4.3 Survey exclusion criteria.

4.6 Clinic survey

The clinic survey comprised a six-month longitudinal study and a separate two-week test-retest study. The patient population is described in section 4.6.1 and section 4.6.2 describes the process of attaining ethical approval. Section 4.6.3 describes the pilot study and sections 4.6.4 and 4.6.5 describe patient recruitment for the longitudinal clinic and test-retest survey respectively. The response rates for both surveys are reported in sections 4.6.6 and 4.6.7 respectively.

4.6.1 Patient population

100 AS patients were required for the longitudinal clinic survey and 50 additional AS patients for the test-retest study. The number of patients provided a figure that was achievable within study resources but was also comparable to, and in excess of many, other studies evaluating the performance of outcome measures in AS (Chapter 2). Resource limitations were influenced by the necessity for the lead investigator (KLH) to recruit all patients, perform the majority of baseline assessments and all two-week and six month assessments. Also, assessments were limited by the availability of clinic space at the SRC and the four month recruitment period (December 1998-March 1999).

The SRC provided the focus for patient recruitment for all clinic-based stages of the study, the pilot postal survey and partial recruitment for the postal survey. The SRC has close ties with the community and has a large AS patient database. Permission to locate the study at the SRC was granted by Dr Peter Dawes, Consultant Rheumatologist and Clinical Director of the Locomotor Directorate (North Staffordshire Hospital Trust). As a check for the validity of the SRC database the computer-based records of all patients were checked by the lead investigator (KLH) for the confirmed diagnosis of AS before including patients in the sampling frame.

4.6.2 Ethical approval

Once permission to approach patients had been gained from all consultant rheumatologists at the SRC ethical approval for the study was sought. The North Staffordshire local research ethics committee (LREC) was initially approached to grant study approval for the development and testing of the PGI-AS (Chapter 3) and the clinic survey. Permission was granted in April 1998 (Appendix 4).

4.6.3 Pilot study

Before the main study the self-completed and interview-administered questionnaires were tested for ambiguity, acceptability and feasibility. The assessment was intended to be a reflection of normal practice and an acceptable level of respondent and clinician burden was sought.

To retain the integrity of the normal clinic format patients were required to attend the clinic 30 minutes before their designated appointment, during which time informed consent was attained and the self-administered questionnaire completed (table 4.1). Questionnaire completion was based upon times reported by other investigators. Patients could request assistance from the lead investigator (KLH) and any assistance was noted.

Interview-administration of the study instruments was developed to closely correspond with normal physiotherapy clinic practice at the SRC (Dziedzic K.- personal communication, 1998). During a standard 20-minute period a selection of patient-based and anthropometric measures are completed and other issues deemed necessary by the patient and therapist discussed. The research assessment was incorporated without detracting from information normally gained or adding to the time requirement. The Body Chart was already incorporated into the regular assessment at the SRC, further justifying its inclusion in the study. The addition of the PGI-AS was the only major change to the usual format and pre-pilot evaluation had supported the feasibility of its incorporation (Chapter 3). The five anthropometric measures (table 4.1) differed slightly to those routinely incorporated in the clinic assessment. However, available evidence and a pre-pilot evaluation of their feasibility supported their use in the study (section 4.2).

Doctor and physiotherapist led clinics run in parallel during the weekly SRC seronegative spondyloarthropathy clinics and provided the focus for the longitudinal clinic survey. All patients attending the physiotherapy clinic were assessed by the senior physiotherapist (JW) and all patients attending the doctors clinic were assessed by the lead investigator (KLH) after visiting the doctor.

Six consecutive AS patients were identified from the computer-based list of appointments and asked to participate in the pilot study (mean age 42.5 years, SD 8.26; range 31-56 years). The same approach for patient recruitment was adopted in both the pilot and main study and is detailed in section 4.6.4. Five patients responded, but one patient was unable to take part due to illness unrelated to AS (n= 4; mean age 43.75 years, SD 10.62; range 31-45 years). There was no significant difference in age between responders and non-responders (t-test p= 0.12). Three patients attended the physiotherapy clinic and one patient attended the doctors clinic.

Following completion of the pilot study patients and the physiotherapist (JW) were subsequently interviewed. No problems were found and the main study followed the same format as that described for the pilot study. Therefore, patients were included in the data collection for the main clinic survey and followed-up at six months.

4.6.4 Patient recruitment - longitudinal clinic survey

At the SRC patients are usually followed up over a six or 12-month period, as the patients condition dictates (Dziedzic K.-personal communication, 1998). Therefore, the clinic survey involved patient assessment at baseline and six months with an experienced physiotherapist (Baseline JW and KLH; six-months KLH).

Consecutive patients were identified from the computer-based list of clinic appointments. This list was effectively a random sample of patients registered with the SRC and attending the clinic during the four-month recruitment period. The initial approach to all patients was from a consultant rheumatologist from the SRC (Dr. Peter Dawes). A patient information letter, information sheet, informed consent form (Appendix 5) and a reply-paid envelope was sent to patients attending the clinic four weeks before their appointment date. All envelopes, patient names and investigator signatures were individually hand-written. Non-responders were sent reminders after two weeks and again one week before the appointment date (McColl et al, 1998). Patients not wishing to participate were asked to return the pre-coded consent form in the reply-paid envelope.

An additional random sample of patients was taken from the SRC database to increase recruitment ('research clinic'). These patients reflected those not included on the above clinic list and those not included in the registers of the local branch of the

National Ankylosing Spondylitis Society (NASS) and the weekly AS exercise group identified below (section 4.6.5). Appointment times ranging from two to four weeks after the letter was posted were listed and patients asked to nominate a date and time (a tick box option) or to indicate a preferred appointment in writing. Appointments were confirmed by telephone. Patients not wishing to participate were asked to return the pre-coded consent form. Non-responders were sent reminders and revised appointments at two weeks and again after four weeks. The lead investigator (KLH) performed all baseline and six-month follow-up assessments with these patients.

For the duration of the study all patients received their usual care. Patients were assured that they were not obliged to participate in the study and were free to leave the study at any time should they choose to do so.

Six-month follow-up appointments were designed to coincide with other clinic appointments. A letter was sent to all patients four weeks before the clinic appointment date, and at two-weeks and one-week before the appointment for non-responders. Where pre-arranged clinic dates were not available letters were sent four weeks before the six-month point, with appointments ranging from one week before to one week after the six-month date (tick-box format). Patients were asked to nominate a date and time or to indicate a preferred appointment in writing. All appointments were confirmed by telephone. Non-responders were sent reminders with revised appointments after two weeks and again after four weeks. Patients no longer wishing to participate in the study were asked to return the pre-coded consent form.

All six-month assessments were performed by the lead investigator (KLH) following the same format as baseline assessments and were performed at a similar time of day. Two six-month health transition questions relating to general health and to AS-specific health were included in the self-administered questionnaire ('Compared to six-months ago how would you rate your health in general / Ankylosing Spondylitis now: much better, somewhat better, about the same, somewhat worse, much worse?').

4.6.5 Patient recruitment - test-retest survey

To assess the two-week test-retest reliability of the study instruments in a clinic environment, a group of AS patients whom regularly attended the SRC was identified.

This sampling frame consisted of members of the Stoke-on-Trent branch of NASS attending weekly self-help group meetings and patients regularly attending twice-weekly AS exercise classes. The president of the local NASS branch (Mr. Ted Brown) was approached in writing to ask permission to approach members of the group for participation in the study and to request a copy of the branch register. All members were also registered with the SRC and diagnosis of AS was confirmed. This sampling frame represents a select group of patients with AS who were active participants in regular exercise and self-help groups. This was considered the most feasible and cost-effective approach to both patients and the study and one representative of normal practice.

A simple random sample of patients was identified from the combined register and patients approached in writing to request their participation in the study. A choice of baseline and two-week appointments were listed and patients encouraged to identify both appointments, preferably at the same time of day. Non-respondents were sent reminders with revised appointment dates at two and four weeks. Patients not wishing to participate in the study were asked to return the blank pre-coded consent form. Appointments were designed to coincide with the weekly Tuesday/Thursday AS exercise groups and the Wednesday evening NASS group held at the SRC. To limit the potential interruption of the exercise classes appointments were also offered before and after classes. All appointments were confirmed by telephone. Patients were reminded of their two-week appointment at the baseline assessment.

At the two week follow-up assessment health transition questions relating to general health and AS-specific health were included in the self-administered questionnaire ('Compared to two-weeks ago how would you rate your health in general / Ankylosing Spondylitis now: much better, somewhat better, about the same, somewhat worse, much worse?').

4.6.6 Response rates - longitudinal clinic survey

In total 189 patients were identified, 102 from the additional research clinic, 42 from the doctors clinic, 39 from the physiotherapy clinic and 6 from the pilot study (table 4.4). Of these 36 (19.0%) failed to respond and 45 (23.8%) refused to take part. Thus, 108 patients agreed to take part in the baseline survey giving a final response rate of 57.1%. The majority of patients were male (n= 87, 80.6%) with a mean age of

49.62 years (SD12.51, range 20 to 74 years)(table 4.5). AS symptom duration ranged from 1 to 58 years (mean 20.34 years, SD 10.14) and the duration of diagnosis ranged from 1 to 49 years (mean 15.5 years, SD 11.78)(n= 100).

	Pilot study		Doctors clinic		Physiotherapist clinic		Additional 'research' clinic		Total	
	n	%	n	%	n	%	n	%	n	%
Total population	6		42		39		102		189	
Non-response	1	16.6	9	21.4	4	10.2	22	21.5	36	19.0
Refusal	1	16.6	7	9.5	10	25.6	27	26.5	45	23.8
Patients taking part	4	66.7	26	61.9	25	64.1	53	52.0	108	57.1

Table 4.4 Baseline response rate for longitudinal clinic survey

To test for response bias, patients who failed to respond or refused to participate were compared with respondents in age and gender (table 4.5). 69 (85.2%) of the non-responders were male (Chi-square p= 0.41, non-significant difference). The mean age of the non-responders was 42.40 years (SD12.44, range 20-75) and although a similar age range was covered a statistical difference in the ages of responders and non-responders was calculated (t-test p< 0.0001).

	Responders (n= 108)	Non-responders (n= 81)
Gender		
- male (n)	87 (80.6%)	69 (85.2%)
- female (n)	21 (19.4%)	12 (14.8%)
Age (years)		
- mean (SD)	49.62 (12.51)	42.40 (12.44)
- median	49.50	42.0
- range	20 - 74	20 - 75

Table 4.5 Responders and non-responders to the longitudinal clinic survey at baseline.

Almost 80% percent of the clinic population were married or co-habiting (table 4.6).

	Total	
Clinic population	n	%
Married / co-habiting (n= 77)	60	77.9
Employed / self-employed	60	56.6
Retired	26	24.5
Not working due to ill-health	15	14.2
Continued education after minimum school leaving age	48	45.3
Degree or equivalent	25	23.6

Table 4.6 Demographic information for longitudinal clinic survey (n= 106).

56.6% of the population were in employment and 24.5% were retired, with only 14.2% of patients unable to work due to ill-health. These figures are comparable to the demographic data from the larger postal survey (section 4.7.6).

Although six-month follow-up appointments were intended to coincide with pre-arranged clinic dates this proved to be difficult due to the forced re-arrangement of many clinics and alternative research appointments to facilitate attendance were offered. Of the 108 patients sent an appointment 13 (12.0%) failed to respond and 7 (6.5%) refused to take part giving a final response rate of 88 patients (81.5%)(table 4.7). A good response was observed for all groups, with the highest response rate in those patients identified in the additional research clinic (86.8%).

6-month Response rate	Pilot study		Doctors clinic		Physiotherapist clinic		Additional research clinic		Total	
	n	%	n	%	n	%	n	%	n	%
Total population	4		26		25		53		108	
Non-response	0	0	4	15.4	5	20.0	6	11.3	13	12.0
Refusal	0	0	2	7.7	2	8.0	1	1.9	7	6.5
Patients taking part	4	100	20	77.0	18	72.0	46	86.8	88	81.5

Table 4.7 Six-month response rate for longitudinal clinic survey.

When compared to the baseline responders a very similar gender ratio (male n= 71, 80.6%) and a slightly lower mean age of 47.02 years (SD12.58; range 20 - 74) was observed for the six-month population. 88 of the 108 patients measured at baseline also participated in the six-month follow-up clinic survey. The mean age of responders to both baseline and six-month surveys was 49.97 years (SD 12.87; range 20-74 years). A non-significant difference to those who only participated in the baseline survey was calculated (n= 20; mean age 48.10 years, SD 10.96, t-test p=0.55). There was no significant difference in gender (Fisher exact test, p= 0.58).

4.6.7 Response rates - test-retest survey

In total 88 patients were identified to take part in the test-retest survey. Six patients were found to be participants in the larger clinic study and so were omitted from the test-retest sampling frame. Unfortunately one patient had died but the hospital records had not been updated. In addition, two patients had moved house and had no forwarding address and one further patient agreed to participate but was unable to

identify a date within the time frame of the study. A corrected total sampling frame of 80 with a corrected response rate of 51 (63.7%) was achieved (table 4.8).

Response rate	Baseline		2-week	
	n	%	n	%
Total population	80		51	
Non-response	6	7.5	3	5.9
Refusal	23	28.7	3	5.9
Patients taking part	51	63.7	45	88.2

Table 4.8 Response rate for test-retest clinic survey.

45 of the 51 patients measured at baseline returned for their two-week assessment. Three of the six patients who failed to attend at two weeks contacted the lead investigator to excuse themselves due to illness. The additional three patients failed to contact the research team.

The majority of patients participating in this study were male (baseline n= 45 (90%), 2-weeks n= 41 (91%)) and the age range covered by the baseline population was 27 to 70 years (mean 47.7, SD 11.34). There was no significant difference between the mean ages of patients taking part in the baseline or two-week assessments. Patients reported a wide range of symptom duration (2 - 49 years; mean 21.4, SD 10.9) suggesting a broad spectrum of disease presentation covered by the population.

To test for response bias, patients who failed to respond or refused to participate in the survey were compared with respondents in age and gender. 27 of the 29 non-responders were male and the mean age of non-responders was 43.6 years (SD 9.21). There was no significant difference between responders and non-responders in gender (Fisher exact test, p= 0.49) and age (t-test p =0.12).

4.7 Postal survey

Section 4.7.1 describes the postal survey patient population and the pilot study is described in section 4.7.2. Sections 4.7.3 and 4.7.4 describe the multi-centre nature of the study and acquisition of ethical approval respectively. Patient recruitment is described in section 4.7.5 and response rates are detailed in section 4.7.6.

4.7.1 Patient population

A population of more than 400 patients were required for the postal survey. There are no published sample size tables for developing and testing patient outcome measures. Hence the planned size of this study has been based on the published work of my supervisors, Dr. Andrew Garratt and Professor Ian Russell (Garratt et al, 1996a; Ruta et al, 1999). In addition, the number of patients presented a figure that was achievable within the study resources and one that was comparable to, and in excess of many, other postal surveys of patients with AS, in particular those evaluating measures of outcome (Chapter 2). The number was also comparable to the number used in the PGI developmental work with a population of patients with low back pain (Ruta et al, 1994a). Resource limitations were also influenced by the necessity for the lead investigator (KLH) to recruit all patients (February - May 1999), to send, collate and record all questionnaires, and to input all data.

A multi-centre study was required to identify sufficient patient numbers. The main sampling frame is described by a random sample of patients with diagnosed AS registered with specialist centres of rheumatology in England and Scotland. Patient inclusion and exclusion criterion (section 4.5.1) was confirmed by participating consultant rheumatologists. As a check for the validity of the patient databases, the records of all patients from two participating centres (Southmead Hospital, Bristol – Dr. Paul Creemer; SRC – KLH) were evaluated to confirm fulfillment of diagnostic criteria.

4.7.2 Approach to specialist centres of rheumatology

Seven established rheumatology centres, identified due to their internationally recognised research activity in rheumatology and AS, were targeted for participation in the postal survey. It was considered that such centres would possess accessible patient databases and would provide a population that was not confined to one particular geographical region.

Letters describing the purpose of the study and requesting the participation of the centre were sent to an identified consultant rheumatologist for each centre. Consultants who were willing to permit access to their patient database were asked to complete a short questionnaire (Appendix 6). This requested information about the patient population, the type of database used and accessibility of information.

Confirmation that other rheumatologists within the centre were happy for their patients to be involved in the proposed study and that relevant management approval would be granted was also requested.

Following agreement to participate in the study by the consultant rheumatologist, contact was made with a senior chartered physiotherapist within the same department. This was considered beneficial to provide an accessible contact for all patient participants, should such contact be required. The physiotherapist would act as a source of help in cases of difficulty or where advise was required. For the duration of the study each physiotherapist was contacted on a weekly basis to identify any patient queries that required follow-up by the lead investigator (KLH) or to identify any problems with the study.

4.7.3 Ethical approval

Once the consultant rheumatologists had granted permission for access to their patient database, ethical approval was sought. The research project took place within six centres with different geographical boundaries and approval by the Multi-Centre Research Ethical Committee (MREC) was required. Following approval from the Northern and Yorkshire MREC in December 1998 approval from Local Research Ethical Committees (LREC) for each participating centre was sought. To reduce consultant burden all LREC applications were completed by the lead investigator (KLH) and signed by the lead rheumatologist. The postal survey commenced following LREC approval (February to November 1999).

Following MREC approval the lead investigator (KLH) made a personal visit to the lead rheumatologist and participating physiotherapist for each centre. The meeting provided a forum within which the relationship between participating centres and the lead investigator could be developed. The aims, objectives and study requirements were discussed and the patient population for each centre identified.

4.7.4 Pilot study - postal survey

The postal survey self-administered questionnaire was the same as the clinic survey with the addition of the Body Chart and PGI-AS (table 4.1). Evidence and experience from the clinic survey (section 4.6) suggested a completion time of approximately 30-minutes.

Six AS patients were randomly selected from the SRC database (mean age 46.00 years, SD 10.12; range 29-58 years). A patient information letter, consent form and self-completed questionnaire was mailed to each patient. Patients were asked to complete the questionnaire and to return it in a reply-paid envelope to the SRC. Non-responders were sent reminders after two and four weeks. A complete package of information was sent on each occasion. Patients not wishing to participate were asked to return the pre-coded blank consent form. All patients responded (January 1999). However, one patient was unable to participate due to family illness and a second due to work commitments. All four remaining patients completed the questionnaire correctly (mean age 43.75 years, SD 10.62; range 29-52 years) and a non-significant difference in age between responders and non-responders was calculated ($p= 0.50$). The high response rate suggested an acceptable self-administered format and no changes were made.

4.7.5 Patient recruitment

From a total of seven centres approached, six expressed their willingness to participate in the study (Appendix 7). For centres with a large register (Glasgow, South Cleveland, Cambridge) a simple random sample of patients was identified (table 4.9). For the remaining centres (Bristol and Cannock) the total population were included in the study. Those patients from the SRC not previously selected for participation in earlier stages of the study were included in the postal evaluation.

Centre	Total AS population (n)	Random sample (n)	Corrected population (n)
Bristol	65	65	65
Cambridge	140	110	110
Cannock	45	45	45
Glasgow	180	110	100
South Cleveland	195	110	103
Stoke	373	29	28
Total	998	469	451

Table 4.9 Patient population registered with postal survey rheumatology centres.

Of the 469 randomly identified patients, a total of 15 questionnaires could not be delivered by the post office (Glasgow n= 8, South Cleveland n= 7), and two patients from Glasgow and one from Stoke had died but had not been removed from the database. The corrected baseline total population is 451.

The initial approach to patients was from the consultant rheumatologist acting as lead contact for each hospital. A letter and patient information sheet summarising the study requirements were sent to each patient to explain the purpose of the study and to request their participation (Appendix 8). The letter requested that should the patient be in agreement to participate in the study, their name and address could be released to the lead investigator (KLH).

All envelopes and patient names were hand-written. All baseline letters were signed by the respective consultant rheumatologist. The initial approach to patients also included the self-completed questionnaire (table 4.1) and a patient informed consent form (Appendix 8). Patients were asked to return the completed consent form and questionnaire to the SRC in the reply-paid envelope. The act of returning the completed items indicated that the patient was willing to participate in the study further. Subsequent patient contact was addressed from the lead investigator, although reference to the respective rheumatology team was always made. Patients not wishing to participate were asked to return the pre-coded blank questionnaire and consent form in the reply-paid envelope. Non-responders were sent reminders after two and again after four weeks. Evidence suggests that the low response rates associated with single approaches to patients may invalidate the results of a postal survey (Bowling, 1997).

Follow-up questionnaires were sent to all participants at two weeks to assess test-retest reliability and at six months to assess responsiveness. Colour coded questionnaires were sent to clarify the different assessments. Two-week and six month health transition questions relating to general health and AS-related health were included in the respective questionnaires (sections 4.6.5 and 4.6.4 respectively). At six months questionnaires were sent to all baseline responders (n= 348).

4.7.6 Response rate - postal survey

From a corrected baseline population of 451, 71 (15.7%) failed to respond and 31 (6.9%) refused to take part. Thus, 349 patients agreed to take part in the study, giving a baseline response rate of 77.4% (table 4.10).

Baseline	Bristol		Cambridge		Cannock		Glasgow		South Cleveland		Stoke		Total population	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Total population	65		110		45		100		103		28		451	
Non-response	7	10.8	11	10.0	9	20.0	13	13.0	26	25.2	5	17.9	71	15.7
Refusal	4	6.2	5	4.5	0	0	10	10.0	6	5.8	6	21.4	34	6.9
Patients taking part	54	83.1	94	85.5	36	80.0	77	77.0	71	69.0	17	60.7	349	77.4

Table 4.10 Postal survey baseline response rate

At two weeks 303 patients agreed to take part in the survey (87.1%)(table 4.11). The Post Office returned a questionnaire from a patient from South Cleveland who had participated in the baseline evaluation but provided no forwarding address. The corrected population total at two weeks is 348.

2-week	Bristol		Cambridge		Cannock		Glasgow		South Cleveland		Stoke		Total population	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Total population	54		94		36		77		70		17		348	
Non-response	7	12.9	7	7.4	2	5.5	5	6.4	8	11.4	5	29.4	34	9.8
Refusal	2	3.7	4	4.3	0	0	2	2.5	3	4.3	0	0	11	3.2
Patients taking part	45	83.3	83	88.3	34	94.4	70	90.9	59	84.3	12	70.6	303	87.1

Table 4.11 Postal survey 2-week response rate

All 348 baseline responders were sent follow-up questionnaires at six months. 45 (13.2%) failed to respond and 14 (4.0%) refused to take part, giving a response rate of 289 patients (82.8%)(table 4.12).

6-month response	Bristol		Cambridge		Cannock		Glasgow		South Cleveland		Stoke		Total population	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Total population	54		94		36		77		70		17		348	
Non-response	8	14.8	13	13.8	3	12.0	5	6.5	13	19.7	3	17.6	45	13.2
Refusal	3	5.6	4	4.2	0	0	4	5.1	2	2.8	1	5.9	14	4.0
Patients taking part	43	79.6	77	81.9	33	91.6	68	88.3	55	77.5	13	76.5	289	82.8

Table 4.12 Postal survey six-month response rate

The response rates for all postal centres at baseline, two-weeks and six-months is summarised in table 4.13.

	Total population	Patients taking part		Non-response	
Recruitment	n	n	%	n	%
Baseline	454	349	76.9	105	23.1
2-weeks	348	303	87.1	45	13.0
6-months	348	289	82.8	59	17.2

Table 4.13 Summary of postal survey response rates

The ages of patients participating in the postal survey at baseline ranged from 18 to 75 years (mean 46.09, SD 12.58). AS symptom duration ranged from 1 to 56 years (mean 19.8 years, SD 11.76), and duration of diagnosis from six months to 52 years (mean 13.63 years, SD 11.26)(n= 336)(table 4.14). This suggests that the population covers a wide spectrum of disease duration. The mean age of patients from Cambridge was younger than all other centres (mean 42.07 years, SD 11.04), and the oldest population was Glasgow (mean 49.79 years, SD 11.61). A statistically significant difference between mean age was observed between Bristol, Cambridge, Glasgow and South Cleveland (p= 0.002) at baseline. The response rate by gender supported a 3:1 male to female ratio (74.2% males) for all centres combined and closely resembles the reported gender ratio of between 2.5 to 4:1 in the British population (Kennedy et al, 1993). There was no statistically significant difference between centres in gender ratio.

The combined age range and gender ratio did not differ significantly between responders at baseline, two weeks and six months (table 4.14).

Description	Baseline (n= 349)	2-week (n= 303)	6-month (n= 289)
Gender (male)	259 (74.2%)	223 (73.6%)	214 (74.0%)
Age (years)			
- mean (SD)	46.09 (12.58)	46.37 (12.70)	46.87 (12.75)
- median	47.0	46.0	47.0
- range	18 - 75	18 - 75	18 - 75

Table 4.14 Descriptive data for postal responders

To test for response bias at baseline, patients who failed to respond or refused to participate in the postal survey were compared with respondents in age and gender

(table 4.15). Patients not taking part were significantly younger than responders (t-test p= 0.001), with a non-significant difference in gender (Chi square p= 0.09).

	Responders (n= 349)	Non-responders (n= 105)
Gender (male)	259 (74.2%)	84 (82.4%)
Age (years)		
- mean (SD)	46.09 (12.58)	41.36 (12.66)
- median	47.0	40.0
- range	18 - 75	18-71

Table 4.15 Responders and non-responders to postal survey at baseline.

The youngest mean age for non-responders was reported for Cambridge and South Cleveland, a similar pattern to that observed for responders (table 4.16).

Baseline	Bristol	Cambridge	Cannock	Glasgow	South Cleveland	Stoke	Total
Non-response	7	11	9	13	25	6	71
Refusal	4	5	0	10	8	4	31
Total non-response	11	16	9	23	33	10	102
Gender (n= male)	10 (90.9%)	12 (75%)	9 (100%)	16 (69.5%)	27 (81.8%)	10 (100%)	84 (82.4%)
Age (years)							
- mean (SD)	39.18 (13.7)	39.9 (10.9)	45.7 (13.02)	45.3 (15.3)	38.9 (12.3)	43.5 (7.5)	41.36 (12.6)
- median	42.0	40.0	47.0	41.5	37.0	44.5	40.0
- range	18 - 61	18 - 64	20 - 64	20 - 71	18 - 66	32 - 53	18 - 71

Table 4.16 Baseline non-responders to postal survey by postal centre.

289 of the 349 patients measured at baseline also completed the 6-month follow-up questionnaire. The mean age of responders to both baseline and 6-month surveys was 46.8 years (SD 12.77), and a statistically significant difference to those who only completed the baseline questionnaire was calculated (mean age 42.46 years, SD 11.06, t-test p=0.013). There was no significant difference in gender (Chi-square p= 0.81).

Response by occupation indicated that over 50% of the total population of respondents were in employment (54.3%)(table 4.17) and is in keeping with reports from other authors (Ward, 1998). However, when centres were examined independently the majority of patients from Cambridge (76.6%) and Bristol (60.4%) were employed, whereas only a minority from Glasgow were in paid employment (35.1%). A large percentage of the population from Glasgow (23.4%) and Cannock (22.2%) reported being unable to work due to ill-health. For the total population this percentage was much lower (15.2%).

Demographic data	Total	
	n	%
Marital status		
- Married / co-habiting (n= 306)	222	72.6
- Single	54	15.5
Employment status		
- Employed / self-employed (n= 348)	189	54.3
- Retired	73	21.0
- Not working due to ill-health	53	15.2
Educational status		
- Continued education after minimum school leaving age (n=345)	176	51.0
- Degree or equivalent (n= 341)	102	29.9

Table 4.17 Demographic data for postal respondents.

4.8 Discussion

The most economic approach to completion of patient-based measures of outcome is postal self-administration within a patients home (Bowling, 1997). This represents a relatively inexpensive means of collecting data and the ability of an instrument to lend itself to such a format has been suggested as an important feature when intended for regular use in health evaluation (Ruta et al, 1994a). Although often a feature of research studies (Guyatt et al, 1987b; Abbott et al, 1994; Garratt et al, 2000), there is little evidence that postal self-administration is regularly accepted as part of the routine practice in AS. Self-completion or interview-administration of instruments with in a clinic environment may represent a more pragmatic reflection of normal practice within AS evaluation, and both postal and clinic-based completion has been addressed in this study. However, less than 50% of rheumatologists incorporate patient-based instruments in routine clinic-based AS evaluation, the evaluation being dominated by anthropometric assessment (Bellamy et al, 1998, 1999).

This study describes the first AS-specific comparative evaluation of a systematically identified and evidence-based package of patient-based and anthropometric measures of outcome in both a postal and clinic environment. The size of both populations is comparable to many and larger than most other AS-specific studies with a specific intention to evaluate the quality and performance of outcome measures (Chapter 2).

The response rate for the clinic-based surveys was satisfactory. A greater baseline response rate was observed for patients participating in the test-retest study (63.0%) than for participants in the longitudinal study (57.1%). The test-retest population was approached with the knowledge that patients regularly attended the SRC for self-help

exercise classes and the greater response rate may reflect the increased motivation of members of a self-help group (Barlow et al, 1993a) or the relative convenience of participating in a study. The two-week response for the test-retest study was also high (88.2%), and may further reflect the ease with which appointments could be arranged to coincide with usual attendance.

High response rates were observed for the physiotherapy and doctor clinics (range 61.9 - 66.7%) and may reflect the convenience of participating in a study run as part of a routine clinic. The lowest response rate was for the additional research clinic (52.0%), the majority of patients attending the clinic in their own time. However, a high follow-up response rate was observed for this group at six months which may reflect the limited impact of the study on the working day. Patients may also represent a particularly motivated group (McColl et al, 1998). Several patients indicated that they supported the research but were unable to find the time to attend the clinic. In light of this, the overall response rate is good and compares favourably with other studies (Fitzpatrick et al, 1993c; Lubrano et al, 1998). The six month response rate for the clinic survey was good (81.5%), although loss to follow-up may have been associated with difficulty in arranging follow-up research appointments to coincide with pre-arranged clinic appointments. No patients were lost due to change of address or pregnancy.

Both groups of clinic patients experienced the same research assessment under similar circumstances, thus enabling baseline results to be combined. This provides a revised baseline population total of 159 patients (59.0% response rate). Both populations cover a similar age range and duration of symptom severity. However, a greater percentage of the test-retest survey population were male. This may be a reflection of the nature of the exercise classes, having a greater attraction to males with AS.

The response rate for the postal survey was very good at all stages of recruitment and follow-up, with some individual centres exceeding a 90% response rate. This result compares favourably to other studies (Bindman et al, 1990; Ruta et al, 1994b; Jenkinson et al, 1994b). However, there was a statistically significant age difference between responders and non-responders at baseline for both the main clinic survey (49.62 versus 42.40 years respectively) and the postal survey (46.09 versus 41.36 years respectively), with non-responders being younger. There was no significant

difference in gender. Other studies have reported a similar difference between responders and non-responders (Garratt et al, 1993). The peak incidence of disease onset is between 25-34 years of age (Carbone et al, 1992) and the bias of responders towards the older age group may reduce the generalisability of the result. For example, the results may have a greater relevance to patients with more severe disease. Limited resources prevented further contact with non-responders to investigate reasons for non-response.

Completion and acceptability of questionnaires may be influenced by many factors. For example, time to complete, legibility and understanding of items, appearance and complexity of the questionnaire and the possibility of distress when completing sensitive items (Fitzpatrick et al, 1998b). Three questionnaires (2 from Glasgow, 1 from Stoke) were returned blank with covering notes written by carers to indicate that the patient was blind and unable to self-complete the questionnaire, and is an issue that should be considered as future exclusion criteria for studies of AS patients requiring questionnaire self-administration.

The saliency of questionnaire content is an important factor influencing response rates in mailed surveys, and the level of importance attributed to the questionnaire by the respondent may have a greater influence over response than actual questionnaire length (McColl et al, 1998). These issues are of equal relevance to both surveys, and have been reflected in the response rates. The self-completed questionnaire mailed to patients, or alternatively completed in the clinic, was very long (14 pages) but contained specific items considered to be of importance to most patients with AS (Appendix 3). The introductory letter to patients indicated the importance of the information to be gained from completion of the questionnaire, helping to improve understanding of the effect of AS on people with the disease (Appendices 5 and 8). The excellent response rates to both surveys supports the fact that many patients were happy to share their experiences of the disease and to complete the questionnaire. However, several baseline postal questionnaires were returned blank with a covering note to indicate that the patient considered their AS to be 'too mild', or 'no longer caused them difficulty', and they therefore felt unable to add to the survey (n= 5). This would suggest a perception of irrelevance associated with some items, perhaps addressing issues associated with more severe disease, and patients representing the less severe spectrum of disease may not be catered for by included items. Several

patients returned blank questionnaires at follow-up of the postal survey with a covering note to indicate that they 'were the same as the last time', and did not provide information that could be used in the analysis. Likewise, some patients may have been lost to follow-up due to the burden of completing three long questionnaires, or the belief that they no-longer had relevant information to add to the study.

Increasing the saliency of the questionnaire to respondents is an important factor influencing response rates (McColl et al, 1998), but there is no consensus on the best approach to encourage patients to complete questionnaires when they believe that their contribution may be of little value. However, the order of questionnaires within a self-completed package may influence completion. For example, patients experiencing minimal disease-related difficulties may find an enhanced affinity to items contained within generic instruments as opposed to those of disease-specific questions, thus facilitating questionnaire completion. Therefore, a revision of the order of instruments within the described package should be considered. Similar findings have been observed in other studies evaluating the performance of patient-based instruments (Grampian Health Outcomes Study (GHOST), Garratt AM.-personal communication, 2000). This is an important issue and warrants further investigation for measures of outcome intended for evaluative purposes. The following chapters consider the measurement properties and relative performance of the study instruments within the questionnaire.

Chapter 5 Reliability

5.1 Introduction

This chapter presents the reliability testing for all study instruments. A definition of reliability is given in section 5.2. Criteria that should be considered in the assessment of data quality and scaling assumptions at both item and scale level and in the overall assessment of reliability is presented. Section 5.3 considers the criteria for data quality, scaling assumptions and reliability for all study instruments. After the results are presented in section 5.4 the chapter closes with a discussion.

It is recommended that both practical considerations and measurement properties are addressed in pursuit of evidence to support the acceptability of multi-item instruments and the standardisation of high quality data (Kosinski et al, 1999a). The development and testing of instruments should give consideration to whether data are of sufficient quality and scaling assumptions are being met (Gandek et al, 1998a; Ware and Gandek, 1998b). Data quality refers to the data completeness and end effects. Scaling assumptions refers to empirical evidence relating to the inclusion of items within hypothesised scales. These tests have implications for item reduction and should take place alongside the tests of internal consistency reliability.

5.2 Reliability and measures of health outcome

Reliability is associated with correctness, referring to the ability of a measure of outcome to produce consistent, reproducible and accurate results both at one point in time and over time when the underlying condition has not changed (Streiner and Norman, 1995). The expression of reliability reflects two components: a true score alternatively referred to as a 'signal', and an underlying level of error or 'noise' (Streiner and Norman, 1995). Reliability estimates the extent to which an instrument is free from error thus reflecting a true score (Fitzpatrick et al, 1998a).

Two forms of reliability have been widely applied to patient-based instruments: internal consistency and test-retest reliability. Items within a multi-item instrument intended to address a specified domain of health should all relate to this domain and internal consistency reliability describes the level of item homogeneity (Nunnally and Bernstein, 1994). The test-retest reliability of an instrument refers to the temporal

stability of the resulting scores over time assuming that the underlying condition has not changed (McDowell and Newell, 1996).

The analysis of data quality and scaling assumptions at both item and scale level provides evidence of item performance and supports instrument scale construction (Gandek et al, 1998a; Ware and Gandek, 1998b; Kosinski et al, 1999a). Interpretation of item performance plays an important role in establishing the quality of multi-item instruments before consideration is given to instrument internal reliability. For example, items with high levels of missing data are not acceptable to patients and limits analysis to a subset of patients, therefore introducing bias into the measurement. Score distribution at both item and scale level shows the range of the defined domain covered. Items that have end effects, that is, a large percentage of patients are responding in the same way, are of limited value when discriminating between patients. Items within a scale designed to address aspects of the same domain should be related to each other and to the scale. Analysis of the relationship between items and the hypothesised scale can be assessed by principle component analysis and item-total correlation. These analyses of item performance support the inclusion of items within multi-item instruments and should be undertaken before more general measurement properties, including reliability, are assessed. Where appropriate these analyses have been adopted in the current study (table 5.1).

Item performance	Scale performance
Data quality and scaling assumptions	Data quality and scaling assumptions
- item completion rates	- scale completion rates
- distribution of item responses	- distribution of scale scores
- frequency of endorsement	- frequency of endorsement
- end effects (ceiling / floor)	- end effects (ceiling / floor)
- equivalence of item means and standard deviations	
Dimensionality	Internal consistency reliability
- principal component analysis	- Cronbachs alpha
- content validity of dimensions	
Item-total correlation	Test-retest reliability
- corrected item-total correlation	- 2-week test-retest (ICC)
	Inter-observer reliability
	- between two or more observers (ICC)

Table 5.1 Assessment of data quality, measurement performance and reliability of evaluative measures of health outcome. ICC = Intra-class correlation coefficient

The specific tests of data quality, scaling assumptions and reliability of the study instruments follow in sections 5.2.1 and 5.2.2.

Instruments not based on classical test-construction are not amenable to tests of internal reliability. For example, instruments based on decision theory or economic theory, such as the EuroQol (EuroQol Group, 1991). The PGI-AS and anthropometric measures are not based on classical test-construction and are not assessed for internal consistency reliability. The reliability of these instruments is assessed by test-retest reliability.

5.2.1 Data quality, scaling assumptions and internal consistency reliability.

Data quality and scaling assumptions

Data completeness shows the extent to which patients are both willing and able to respond to questionnaire items (Safran et al, 1998). The frequency with which items are omitted or answered incorrectly gives an indication of items that should be considered for removal from the scale due to poor completion rates, or re-written due to poor comprehension, intolerance or ambiguity. The level of missing data has implications for the calculation of scale scores. Some instrument developers are explicit about the number of items that may be omitted whilst still allowing the calculation of a final score. For example, a final score may be calculated for the SF-36 if one-half or fewer items are missing (Ware, 1997).

The aim of measuring an attribute is to assess the extent to which a defined characteristic or domain is present or, for example, the extent to which difficulties with specified activities are encountered (Streiner and Norman, 1995). In the evaluation of functional ability one may wish to estimate 'how much difficulty' a patient experiences in 'tying shoe laces' or 'getting into and out of the bath'. The process of measurement requires quantification or 'scaling' and hence the numerical representation of an attribute in a manner that allows users of outcomes data to appoint both meaning and relevance to the result (Nunnally and Bernstein, 1994). Most features of HRQL should be considered on a continuum, and the level of measurement adopted in scaling will affect the quality and extent of information retrieved from its evaluation. Various scaling approaches have been described. For example, visual analogues scales (VAS), Likert scaling and adjectival scales (Streiner and Norman, 1995). The method of scaling items should provide sufficient

discrimination to demonstrate clinically meaningful change and to facilitate score interpretation (Kirshner and Guyatt, 1985).

At item level categorical response options are chosen for many instruments. However, the final score is often assessed on a continuous scale (Nunnally and Bernstein, 1994). For example, the Revised Leeds Disability Questionnaire (RLDQ) (Abbott et al, 1994) employs a four-item categorical response scale with options scored between 0 and 3. The overall score is reported on a continuous scale between 0 and 48. The transformation of data gathered at one level of measurement into a higher level of measurement is frequently observed for measures of behavioural attributes (Nunnally and Bernstein, 1994). The transformation of data into a continuous format facilitates the representation of score distribution characteristics as mean and standard deviation (SD) values at both item and scale level providing important information in relation to the variability of responses (Safran et al, 1998).

Instruments based on classical test-construction theory produce a final score based on the summation of item scores (Nunnally and Bernstein, 1994). To assess whether standardisation or weighting of item scores is required before the addition of items two aspects of data quality should be assessed (Safran et al, 1998; Kosinski et al, 1999a). First, the distribution and variability of item responses should be considered. An acceptable level of variability in item responses and final scores is necessary for instruments to provide meaningful information for the evaluation of a specified domain and change in this domain (Safran et al, 1998). The distribution of responses at item level provides evidence in support of the discriminative ability of an item. Secondly, the equivalence of means and standard deviations across items indicates if standardisation of item variances is required before the addition of item scores. If items address similar attributes and responses are normally distributed, items contained within the same scale would in general be expected to produce similar mean and standard deviation values (Kosinski et al, 1999a).

An exception to this result would be if items were specifically intended to define the ceiling or floor of the domain. Item performance would be reflected in a skewed distribution of responses and a mean and standard deviation value that departed from other items in the scale. Assuming a representative patient population, response distribution at item level would usually approximate normality. This would describe

the range of response options and the ability for most respondents to describe change (Nunnally and Bernstein, 1994). Skewed data at item level suggests that respondents score among the most or least favourable health states. End-effects, where the majority of item scores accrue at the ceiling or floor of the response scale should be identified (Streiner and Norman, 1995). A ceiling effect would be demonstrated where, for example, more than 80% of responders choose the response option reflecting the best possible health state (Jenkinson et al, 1996). This is of limited use for evaluative purposes due to the inability of a respondent to record any meaningful improvement in health. The opposite is true for an item demonstrating a floor effect (Fitzpatrick et al, 1998a).

Items with a very large or very small level of endorsement for a particular response option add little to an instruments discriminative ability and should be removed (Streiner and Norman, 1995). If a large number of patients respond to an item in the same way, the item contributes little to the variation in the scale scores. In developing the Aberdeen Low Back Pain Scale, Ruta et al (1994b) considered items for rejection if more than 80% of patients gave the same response, suggesting that these items would inadequately discriminate between differing levels of severity.

Dimensionality

The dimensionality of multi-item instruments has been assessed with principal component analysis (PCA)(Joliffe, 1986). McHorney et al (1993) suggest that to further understanding of the impact of health or disease the multi-dimensional evaluation of health is necessary to provide a 'synergistic and comprehensive assessment'. PCA can be used to describe the underlying structure of a multi-item instrument through the identification of components into which items may group (McDowell and Newell, 1996). PCA adds empirical weight to the hypothesised domains, thus informing further on the underlying concept and instrument internal structure (McHorney et al, 1993; McDowell and Newell, 1996). PCA may be used to confirm the presence of hypothesised domains described by instrument developers (Kosinski et al, 1999a). For example, in hypothesising the dimensionality of the SF-36 McHorney et al (1993) proposed that two domains, mental and physical, would underlie the structure of the eight scales and PCA supported this. An evaluation of the content validity of items captured within identified domains further supports the role of items in addressing the underlying conceptual base (Ruta et al, 1994b).

Item-total correlation

Multi-item scales based on classical test construction must consist of homogeneous items that address an underlying domain of health. The inclusion of multiple items addressing the same clearly defined domain generates greater information relating to the construct. If these items relate closely to each other and to the specified domain, reliable information about the domain should be gained (Kosinski et al, 1999a). If items fail to relate closely to each other or inadequately represent the underlying domain error may be introduced into the measurement. Issues of acceptability and feasibility make it impossible to include all possible items in an instrument, and so items are sampled to represent different aspects of the domain, the combination of items fostering finer discrimination between patients.

The relationship between items and the remainder of their scales should be assessed for all instruments based on classical test construction theory (Nunnally and Bernstein, 1994). Item-total correlation correlates an item and the total score for the scale. However, inclusion of the specific item in this calculation artificially inflates the score and its contribution is removed from the total scale score before calculating the 'corrected' item-total correlation (Streiner and Norman, 1995).

When an instrument or separate domains within an instrument consist of a small number of items a greater level of item-total correlation is required to achieve a satisfactory level of reliability. A score of 0.4 has been recommended as a minimal level to support the internal consistency of items (Ware, 1997; Kosinski et al, 1999a). Items with a high level of item-total correlation demonstrate more variance relating to the common domain between items and as such are more discriminating. These items add more to the estimated level of instrument reliability, as measured by Cronbachs alpha, than items with a lower item-total correlation (Nunnally and Bernstein, 1994). Items with low levels of item-total correlation usually demonstrate a poor relationship to the underlying construct and should be considered for rejection. These items may also be ambiguous, or attract very large or small levels of endorsement.

If a rigorous approach to data completeness and item scaling is taken then this should produce satisfactory scale properties for the instrument as a whole. Once issues

relating to item performance have been addressed, and any necessary changes to instrument structure made, instrument properties at scale level should be assessed.

Data quality and scaling assumptions at scale level

Data quality and scaling assumptions include the distribution of total instrument scores and any end-effects. Assuming a representative population of interest, score distribution would usually approximate normality (Nunnally and Bernstein, 1994), and responses reflecting the full range of the domain described by the instrument should be recorded (Ware and Gandek, 1998b). If a skewed result is found the percentage scoring at the ceiling (best health) or floor (worst health) of the scale should be noted since this affects the ability of the instrument to measure change (Fitzpatrick et al, 1998a). Where more than 80% of responders score at the extreme of a scale the instrument may be of limited use in evaluation. If a scale score is skewed and does not cover the important range the performance of items and response options should be reviewed. However, if a patient population represents a specific spectrum of disease a skewed distribution may support instrument validity. Kosinski et al (1999a) reported a skewed distribution of responses towards less favourable health states for SF-36 scales representing disability and distress in patients experiencing an exacerbation of their arthritis.

Internal consistency reliability

The internal consistency reliability of a multi-item instrument is assessed by Cronbach's Alpha (Streiner and Norman, 1995). Alpha assesses the relationship between all items in a scale, reflecting both the total number of items and their average correlation (Nunnally and Bernstein, 1994). Alternatively, alpha may be described as representing the average expected correlation between an alternative scale consisting of the same number of items sampled from the same specific domain (Ware, 1997).

Study instruments with published evidence of internal consistency reliability include the parent instrument of the RLDQ, the Leeds Disability Questionnaire (LDQ)(0.93)(Abbott et al, 1994) and the BASDAI (0.84)(Jones et al, 1996c; Calin et al, 1999a).

5.2.2 Test-retest Reliability

Test-retest reliability addresses the accuracy of results over time assuming no underlying change in condition (Ware, 1997). Consistency of a result repeated on different occasions would be expected if contributing factors to the health state have not changed and there is an absence of random or systematic error.

Test-retest reliability and internal consistency reliability are distinct concepts to the extent that high levels of test-retest reliability in multi-item instruments are not necessarily associated with high levels of internal reliability. Test-retest reliability may not provide a complete picture of the reliability of multi-item instruments based on classical test-construction theory because the result is only partly influenced by internal reliability (Nunnally and Bernstein, 1994). For example, an item may correlate weakly with other scale items but correlate highly with itself over a test-retest period. Thus, high temporal stability with low internal consistency may be demonstrated. However, internal consistency reliability may over-estimate true reliability due to the lack of consideration for different sources of variance (Streiner and Norman, 1995). Internal consistency reliability requires instrument completion on one occasion only and does not consider variation between repeated administrations. For example, variation between observations over time. Test-retest reliability requires consistency over two points in time and may provide a more rigorous measure for evaluative contexts (Kirshner and Guyatt, 1985). Evaluative instruments based on classical test-construction theory should be assessed for both internal consistency and test-retest reliability. The reliability of instruments not based on classical test-construction theory is dependent on establishing support for the temporal stability.

Test-retest reliability generally involves patient self-completion of an instrument on two occasions separated by a suitable time period. Alternatively, an observer may complete the instrument on behalf of the patient on both occasions. For example, most anthropometric measures require assessment by a trained observer. High levels of test-retest reliability, or alternatively intra-observer reliability, reflect low levels of within-person variance (Beurskens et al, 1995).

Test-retest reliability is assessed for patients indicating no change in their condition (Deyo et al, 1991; Fitzpatrick et al, 1993b). Health transition questions to describe

the underlying stability have evidence of validity and ask patients if their health, in general or specific to a particular disease, has changed between administrations of the instrument (Fitzpatrick et al, 1993b; Peto et al, 1998).

There is no agreed time period for test-retest studies. A shorter period may facilitate answer recall thus artificially inflating reliability (Nunnally and Bernstein, 1994; Streiner and Norman, 1995). Alternatively, the fluctuating nature of health could mean that too long an interval results in actual change and a reduced number of patients suitable for inclusion in the analysis. The time period is also influenced by the nature of the health state. For example, very acute disorders may achieve a state of full recovery by two weeks and a shorter retest period is necessary (Beurskens et al, 1995). A window of between two days and two weeks has been recommended for most conditions (Streiner and Norman, 1995).

Inter-observer reliability

Inter-observer reliability assesses the level of agreement between results when different observers are responsible for the assessment. An instrument that produced differing results dependent on the observer would not be useful. Stability of results between observers supports instrument application by different observers, assuming the same format for administration is followed (McDowell and Newell, 1996).

Statistical analysis

The correlation coefficient is the most frequently used statistical method for calculating estimates of test-retest reliability. Pearson's correlation coefficient, r, is commonly used but fails to take account of systematic bias (Deyo et al, 1991; Beurskens et al, 1995; Streiner and Norman, 1995). Repeated observations may be systematically different but still demonstrate a linear relationship with the first set of observations. Therefore, Pearson's correlation may overestimate reliability.

The Intra-class correlation coefficient (ICC) is based on Analysis of Variance and considers the variability between patients as a proportion of the total variability (the sum of between patient and within patient variability) (Jordan, 2000). Therefore, in a heterogeneous population reliability will be higher. The ICC takes account of systematic bias and is recommended in preference to Pearson's correlation coefficient (Streiner and Norman, 1995; Jordan, 2000). However, the main source of

measurement error is usually random and as a consequence the ICC and Pearson's correlation closely approximate each other (Streiner and Norman, 1995). Several forms of ICC have been described and an appropriate method for the assessment of measures of outcome considers observers as a random selection from a wider population of observers (random effects model; ICC(2,1)) (Shrout and Fleiss, 1979).

Reliability of the study instruments

Most of the study instruments have published evidence of test-retest reliability (Chapter 2)(table 5.2). There is no published evidence of the reliability for the ASQoL, PGI-AS or the RLDQ.

Authors	Instrument	n	Time between administrations	Correlation coefficient
Garrett et al (1994)	BASDAI	46	24 hours	0.93
Dziedzic (1997)	Body Chart	14	1 hour	$p = 0.34$ (Wilcoxon matched pairs)
Abbott et al (1994)	LDQ	9	5 days	0.98
		25	24 hours	0.92
Hurst et al (1997)	EuroQol - EQ-5D	93	3 month	0.73
		31	2 week	0.78
Hurst et al (1997)	EuroQol - thermometer	93	3 month	0.70
		31	2 week	0.85
Hurst et al (1998)	SF-12	75	3 month	MCS - 0.71
				PCS - 0.75

Table 5.2 A selection of evidence of test-retest reliability of patient-based study instruments

BASDAI - scored 0 -10; higher scores indicate greater disease activity.
Body Chart - minimum score 0, no maximum score; higher scores greater perceived bodily pain.
LDQ - scored 0-3; higher scores indicate worse functional ability.
EuroQol EQ-5D - scored -0.59 to 1.00; 1.00 the 'best possible health'.
EuroQol thermometer - scored 0-100; higher scores indicate better health status.
SF-12 - uses norm based scoring from the general population. Scales are transformed to have a mean of 50 (sd 10), with a range 0-100. Higher scores indicate better health status. MCS-mental summary scale , PCS-physical summary scale.

The selected generic instruments, EuroQol and SF-12, have not previously been assessed for reliability in AS, but reliability has been described in patients with RA (n= 233) (Hurst et al, 1997,1998; Ruta et al, 1998). Test-retest reliability was assessed for patients indicating 'no change' in RA on a transition question. The EuroQol was assessed for both three-month (n= 91) and two-week (n= 31) test-retest reliability and satisfactory levels were reported (range 0.73 - 0.80) (Hurst et al, 1997). Three-month test-retest reliability of the SF-12 was satisfactory (>0.71)(n= 75).

5.2.3 Standards for estimates of reliability

The reliability of an instrument has implications for whether it is suitable for application in group or individual evaluation. The levels of acceptable reliability have

increased over recent years as the complexity and conceptual base of measures of health outcome have advanced, supporting a reduction in measurement error as the ability to generate information has improved (Ware, 1997). For the evaluation of individuals high levels of reliability, above 0.90, have been recommended (Streiner and Norman, 1995; Ware, 1997). For group comparisons levels over 0.70 have been suggested (Nunnally and Bernstein, 1994; Streiner and Norman, 1995).

5.3 Methods for assessing data quality, scaling assumptions and reliability of the study instruments

When item performance of a multi-item instrument is addressed, requirements for item retention can be considered under three main headings (table 5.3). Following standards recommended by several authors (McHorney et al, 1994; Juniper et al, 1997; Jenkinson et al, 1999a), items should be considered for rejection if they fail to fulfill any of the following criteria:

Rejection of items
Data quality and scaling assumptions
- > 10% missing values
- skewed distribution of responses at item level
- 80% frequency endorsement for response option
- evidence of end-effects (> 80%)
Dimensionality
- items with low component loadings (item loading < 0.4)
- poor content validity
Item-total correlation
- item-total correlation < 0.4

Table 5.3 Evidence to support item rejection from a multi-item instrument.

Where appropriate the item and scale level performance of instruments completed during baseline evaluations was investigated (table 5.1). All instruments were assessed for missing data, the distribution and symmetry of item response scores and the endorsement frequency of response options. Items with high levels of missing data, the presence of end-effects or excessive or minimal levels of endorsement were considered for rejection. The data quality and distribution of scores was subsequently assessed at scale level. Where appropriate internal consistency reliability was assessed. A two-week test-retest reliability evaluation of all study instruments was

performed. The anthropometric measures were also assessed for inter-observer reliability.

5.3.1 Data quality, scaling assumptions, internal consistency reliability and test-retest reliability of the study instruments

PGI-AS

A rating scale (Appendix 9) was completed by the therapist to assess the level of assistance and completion time required by patients during interview-administration. Completion was compared between baseline and six-month assessments. Anomalies in postal self-administration were identified.

Data completeness for each step of the PGI-AS was considered. The number and frequency with which items were mentioned in step 1 and the data quality and scaling assumptions for the unweighted item responses in step 2 were assessed. At scale level the data quality and scaling assumptions for the final index score was assessed.

Following the developers of the PGI (Ruta et al, 1994a) the final score and three steps in instrument completion were assessed for test-retest reliability. Constituent parts of the PGI-AS were assessed to identify any weaknesses in the instrument and to support any need for modification. Three follow-up formats (table 5.4) have been separately assessed for test-retest reliability. Evaluation of the different formats is required to ensure that patients, clinicians and researchers are able to make meaningful interpretations when completing the instrument (Jenkinson et al, 1998c).

Follow-up completion	Step 1: Identifying areas	Step 2: Scoring each area	Step 3: Spending points
Blind	Blind to baseline areas	Blind to baseline score	Blind to baseline points
Closed	Informed of baseline areas. Not allowed to change or add to list	Blind to baseline score	Blind to baseline points
Informed and open	Informed of baseline areas. Allowed to change or retain list as necessary	Blind to baseline score	Blind to baseline points

Table 5.4 PGI-AS follow-up formats.

The variation in follow-up format includes the provision or absence of the areas identified in step 1 at baseline. Patients are not informed of scores or points. The 'blind' format asks the patient to complete the instrument without previous responses.

The 'closed' format lists baseline areas, but the patient is not allowed to change or add to this list. The 'informed and open' format lists baseline areas and patients are allowed to retain or change these as necessary.

Participants in the postal survey were randomly assigned to complete either the 'blind' or 'closed' formats at two-weeks. All clinic participants completed the 'informed and open' format.

An index of change was calculated for step 1 (identifying areas) for the 'blind' and 'informed and open' formats. The 'closed' format does not allow patients to change baseline areas and an index of change cannot be calculated. At baseline patients identify up to five areas of their life affected by AS. An area substituted at follow-up for a baseline area was given one point. An area added at follow-up in place of a blank at baseline was given half a point. An area omitted at follow-up resulting in a reduction in areas was also awarded half a point. The number of 'points' was then summed. The index of change is equal to (5-x) / 5, where x = the number of points (Ruta et al, 1994a). The final index of change ranges from 0 to 5 and three groups are described: first, those making 0 - 1 area changes; secondly, those making 1.5 - 2.5 area changes; and, finally those making 3 - 5 area changes.

In step 2 of the PGI-AS ('scoring each area') the patient scores areas identified in step 1 between 0 ('The worst you could imagine') and 10 ('Exactly as you would like to be'). The summed scores for step 2 is the 'unweighted' PGI-AS score. Two-week test-retest reliability was assessed.

The test-retest reliability of the PGI-AS index score was also assessed. Patients 'spend points' in step 3 to reflect their priorities for improvement producing a 'weighted' score (Ruta et al, 1994a). The weighted and un-weighted reliability coefficients were compared.

Finally, reliability coefficients for the unweighted and weighted PGI-AS were calculated for each group as defined by the 'index of change'. This analysis considers the influence of change in areas on the total score. In summary, test-retest reliability was calculated for several versions of the PGI-AS (table 5.5):

PGI-AS	Completion format of PGI-AS			
	Postal population			Clinic population
	Blind format	Closed format	Total postal population	Informed and open format
PGI-AS Index Score - without index of change	✓	✓	✓	✓
PGI-AS Index Score - with index of change	✓	✗	✗	✓
Unweighted PGI-AS Score - without index of change	✓	✓	✓	✓
Unweighted PGI-AS Score - with index of change	✓	✗	✗	✓

Table 5.5 Summary of test-retest reliability analyses calculated for the PGI-AS.

✓ Test-retest reliability estimated; ✗ test-retest reliability not estimated.

PGI-AS Format: 'Blind' - blind to baseline areas. 'Closed' - informed of baseline areas, not allowed to change. 'Informed and open' - informed of baseline areas, allowed to change or retain areas as necessary.

b) Disease-specific and c) Generic instruments

The dimensionality of the ASQoL, BASDAI, and the RLDQ was assessed using principle component analysis (PCA). Criteria relating to the level of variation explained by components and the level of component loadings were based on previously published studies of measure of health outcome (Hyland et al, 1991; McHorney et al, 1993; Juniper et al, 1997; Jenkinson et al, 1999a). PCA with varimax rotation (Jolliffe, 1986) was used to identify separate components of health within each instrument. Two approaches to PCA were considered. First, the dimensionality of instruments was assessed by evaluating eigenvalues (a statistical measure of the components ability to explain variation between patients) of more than 1.0 (Hyland et al, 1991; Jenkinson et al, 1999a). A further assessment sought the number of dimensions identified by instrument developers (McHorney et al, 1993). For example, if the ASQoL is to maintain the structure recommended by the developers then one dimension should be confirmed by the PCA. The content validity of components identified for all analyses was considered in light of the results of the PCA and the original work presented by the instrument developers.

There is a lack of consensus between studies on the level of component loadings below which items should be considered not to be sufficiently related to the relevant component. Hyland et al (1991) rejected items if loadings were less than 0.3, Juniper et al (1997) removed items which loaded less than 0.4, and Jenkinson et al (1999a) consider loadings equal to or above 0.5 to be important. This lack of consensus meant that component loadings were interpreted with care but for both analyses component

loadings of more than 0.4 were sought (Junniper et al, 1997; Garratt AM.- personal communication, 1999). Items with component loadings less than 0.4 were identified and considered for rejection.

Item-total correlations were calculated at item level and Cronbach's alpha was calculated at scale level for the three instruments based on classical test theory: ASQoL, BASDAI and RLDQ. Levels of item-total correlation above 0.4 were sought (Ware, 1997), and items with low levels of correlation were considered for rejection. Cronbach's alpha greater than 0.7 were sought to support instrument internal consistency reliability (Streiner and Norman, 1995).

Although the SF-12 consists of multi-items, the developers have indicated that the evaluation of internal consistency reliability underestimates the instruments reliability (Ware et al, 1995). This is because the SF-12 items, all selected from the parent instrument the SF-36, were chosen for their relative heterogeneity, having a 'reliable variance of proven value in estimating physical or mental health' (Ware, 1995). Although the SF-36 is based on classic test construction theory, the SF-12 is not. The SF-12 is based on multiple regression analysis of the SF-36 and is not suited to exploratory principle component analysis (Ware et al, 1994). Therefore, only test-retest reliability has been calculated for the SF-12.

All instruments were assessed for test-retest reliability for both clinic and postal modes of administration. A random sample of out-patients participated in the clinic based two-week test-retest survey (section 4.6.5), and patients participating in the postal survey were mailed a second questionnaire two-weeks after returning / completing the first questionnaire.

Two health transition questions were included in the two-week questionnaire (figure 5.1). The first question related to AS-specific and the second to their general health.

'Compared to two-weeks ago, how would you rate your Ankylosing Spondylitis / health in general now?'

Much better than 2-weeks ago	1
Somewhat better than 2-weeks ago	2
About the same as 2-weeks ago	3
Somewhat worse than 2-weeks ago	4
Much worse than 2-weeks ago	5

Figure 5.1 AS-specific and general health transition questions.

The two transition formats were considered important to represent both disease-specific change represented primarily by disease-specific instruments and a broader appreciation of change in HRQL captured by the generic instruments. Test-retest reliability was calculated for those patients indicating that both their AS and general health had remained the same. The PGI-AS represents the impact of both AS-specific and general health and stability in both items is required for the assessment of test-retest reliability for this instrument.

The intra-class correlation coefficient (2,1) (Shrout and Fleiss, 1979; Streiner and Norman, 1995) was used as a measure of agreement between repeated observations.

5.3.2 Test-retest and inter-observer reliability of anthropometric measures

Following completion of the questionnaire five anthropometric measures were assessed in all clinic patients. During the baseline assessment two observers (one experienced physiotherapist - KLH, and one 'trained' non-physiotherapist - KJ) recorded all measurements in selected patients (section 4.6.5). The order of observers was randomised. One observer (KLH) repeated the measurements at two-weeks.

A warm-up period has been recommended when assessing the repeatability of lumbar anterior flexion (FFD) and cervical rotation (Roberts et al, 1988). Up to six repetitions may be required before stability is achieved (Roberts et al, 1988). However, this approach is not feasible in routine practice, and a standardised format was followed in an attempt to reduce performance error and to increase patient familiarity with the measurements. All movements were described to the patient by the first observer and each movement repeated once to ensure correct performance and to act as a 'warm-up'.

5.4 Results of reliability testing

This section presents the results of the assessment of data quality and scaling assumption at both item and scale level and reliability testing in both surveys.

5.4.1 Results of data quality, scaling assumptions, internal consistency reliability and test-retest reliability of the study instruments

Analysis of instrument performance for the ASQoL, BASDAI, EuroQol, RLDQ and SF-12 at both item and scale levels has been based on both clinic and postal surveys.

This maximises the sample sizes which is of particular importance for principle component analysis (PCA) and is justified because completion of the instruments in both surveys followed the same patient self-completed format. Completion of the Body Chart and the PGI-AS followed interview-administration during the clinic survey and self-administration during the postal and data will be assessed separately for each survey. Anthropometric measures were only assessed in the clinic survey.

a) PGI-AS

Completion of the PGI-AS was considered to be correct if all 3 steps were completed and allocation of points in step 3 was correct (Ruta et al, 1999). A possible total of seven areas may be scored and subsequently points may be spent. Area 7 ('All other non-health areas of your life') is the only 'box' that must be scored, although points need not be spent in this area if priorities are described elsewhere. However, area 6 ('areas affected by health problems other than AS') must not be left blank. The patient must indicate 'none' in writing (as illustrated in the completed example) or score the item.

Additional patient assistance during interview administration was described on a therapist completed rating scale (Appendix 9)(table 5.6) and compared for baseline and six-month assessments. A similar level of assistance was required for completion of all steps at baseline: the most for step 2 ('scoring each area') and the least for step 1 ('identifying areas'). Overall, less assistance was required at six-months.

However, at six-months more assistance was required to complete step 1 ('identifying areas') and many patients were more inquisitive about areas to identify. It is possible that the baseline completion had caused them to think at greater length about the impact of AS on their life and this was reflected in the increase in questions asked.

Baseline completion time ranged from less than 5-minutes (28.6%) to approximately 25-mintues (4.8%) with the majority of patients taking between 5 and 10 minutes (44.0%). 72.6% required less than 10-minutes. Follow-up completion was much quicker. At six-months 44.0% of patients required less than 5-minutes completion time and 83.3% less than 10-minutes. At baseline one patient required maximum assistance and more than 25-minutes to complete the instrument. This was influenced by the difficulty in maintaining a focus on instrument completion. The interview

provided an opportunity for issues to be addressed that the patient had felt previously unable to discuss. However, this patient required much less assistance and a reduced completion time at six-months. At six-months no patient required more than 15-minutes completion time.

PGI-AS steps of completion		Rating Scale [a]									
		None at all		A little assistance		Moderate assistance		Significant assistance		Maximum assistance	
	Assessment	n	%	n	%	n	%	n	%	n	%
Step 1 - Identifying areas	Baseline	45	53.6	26	31.0	9	10.7	3	3.6	1	1.2
	6months	1	1.2	67	79.8	8	9.5	6	7.1	2	2.4
Step 2 - Scoring each area	Baseline	52	61.9	16	19.0	12	14.3	3	3.6	1	1.2
	6months	1	1.2	71	84.5	6	7.1	4	4.8	2	2.4
Step 3 - Spending points	Baseline	51	60.7	19	22.6	8	9.5	5	6.0	1	1.2
	6months	68	81.0	8	9.5	6	7.1	2	2.4	0	0.0
Overall clinician reported level of assistance	Baseline	43	51.2	24	28.6	12	14.3	4	4.8	1	1.2
	6months	61	72.6	14	16.7	7	8.3	2	2.4	0	0.0
		5 mins		6-10 mins		11-15 mins		16-20 mins		21-25 mins	
Time taken to complete	Baseline	24	28.6	37	44.0	13	15.5	6	7.1	3	3.6
	6months	37	44.0	33	39.3	14	16.7	0	0.0	0	0.0

Table 5.6 Rating scale of assistance required to complete the interview-administered PGI-AS (n= 84) (Pilot study excluded).

[a] Rating scale: None at all = standard instructions only
A little assistance = repetition of instructions
Moderate assistance = repetition of instructions with increased reference to completed example
Significant assistance = increased reference to example. Possible re-wording of instructions
Maximum assistance = maximum reference to 'trigger list' and example. Examples read aloud to facilitate completion.
Prolonged completion time.

Various anomalies following postal self-administration were identified (table 5.7). Most difficulties involved 'spending points' in step 3 (n=24, 6.8%). To allow for the calculation of an index score a total of 14 points must be spent in areas that have been scored in step 2. Occasionally patients spent 13 or 15 points suggesting a simple mistake in calculation, although some patients used an obscure number of points suggesting a level of confusion.

Difficulties with step 2 of the PGI-AS most frequently involved score omission for items 6 and 7. Of the 303 patients for whom a baseline index score could be calculated a total of 23 patients (7.6%) failed to score box 6 ('areas affected by health problems other than AS'). However, 90 (29.7%) patients indicated no additional health problems by writing 'none' in the box, and under this circumstance a score is not required. 14 patients (4.0%) failed to score box 7 ('all other non-health areas of your life') and 10 patients (2.8%) failed to score both boxes.

	n	%
Completed PGI-AS correctly – index score calculable (area 6 / 7 may or may not be scored)	303	87.5
Completed PGI-AS correctly – index score calculable (area 6 / 7 completed correctly)	276	79.8
Omitted to score box 6	13	3.7
Omitted to score box 7	4	1.1
Omitted to score boxes 6 and 7	10	2.9
Incorrect points spent step 3	24	6.9
Partially completed PGI-AS	6	1.7
Evidence of misunderstanding in completion - index score not computable	6	1.7
Left PGI-AS blank	7	2.0
Questionnaire printing error	3	0.8

Table 5.7 PGI-AS completion difficulties encountered in the baseline postal survey (n= 346).

Completion of PGI-AS considered to be incorrect if all 3 steps were not completed and allocation of points in step 3 incorrect.

Rather than scoring boxes 6 and 7 in step 2, eight patients (2.3%) indicated in writing 'areas affected by other health problems' or 'other non-health areas of life' respectively. For example, writing 'asthma' into box 6 or 'marriage' into box 7. When this was not accompanied by a score completion was considered incorrect. If points were also spent in these areas in step 3 a final score could not be computed. More frequently scores were simply not entered. 27 patients incorrectly completed areas 6 and/or 7 and of these a final index score could not be calculated for 12 (3.5%).

Four patients made a partial attempt at completing the PGI-AS. All patients entered areas in step 1, but did not complete steps 2 and 3. In a further six patients there was clear evidence of misunderstanding. For example, areas and scores were entered on the sample PGI or trigger list items circled. Seven patients left the PGI-AS blank. Two of these patients added notes to indicate that they did not experience any problems with their AS and so found completion difficult and one patient drew a line through the page.

All patients attempting to complete the PGI-AS in the postal survey entered at least one area in step 1 ('identify areas')(n= 336). Five areas were listed by 68.7% (n= 233) of patients, 87.0% (n= 295) completed four boxes, 94.4% (n= 320) completed three boxes, and 96.4% (n= 327) completed two boxes. However, a final score could only be calculated for 303 (90.2%) patients.

During interview-administration six patients (3.7%) were unable to complete step 1 because they felt so well. However, the remaining 153 patients listed at least one area (96.2%), 84.9% (n=135) identified two areas, 71.7% (n=114) listed three areas, 57.2% (n= 91) patients listed four areas and 43.4% (n=69) listed a total of five areas. Those patients unable to list any areas were directed to score areas 6, if applicable, and area 7 only, and to spend all 14 points between these two areas. When patients felt it unnecessary to complete step 1 it became apparent that the questionnaire was limited in guiding patients to completing boxes 6 and 7 only, and further guidance for self-completion is required. However, the results are an improvement on non-completion due to lack of problems reported by Ruta et al (1999).

On the other hand, patients with severe disease or those capable of identifying many areas of life affected by their AS expressed difficulty in choosing the five most important. This did not prevent index completion but increased completion time. During the pilot study analysis of the postal PGI-AS one patient circled additional areas in the trigger list and indicated difficulty in limiting his selection to five. This problem has been reported by other investigators (Cotton et al, 1993).

Each item in the PGI-AS is scored between 0 and 10, where lower scores indicate more difficulty with the identified item. Similar mean values at item level ('between poor and fair') were found for both postal and clinic populations and for items one to four (table 5.8). Item response distribution approximated normality. The lowest item mean was found for item 5 ('poor, but not the worst you could imagine'). Responses to this item were slightly skewed towards the lower levels of disease-related quality of life. The largest floor effect was for item 5 in the clinic population (7.4%). Although patients were not requested to list the areas in any particular order the fifth item in step 1 produced the lowest score for both populations.

Higher mean values were found for areas 6 and 7, suggesting that patients in general experienced 'fair' health in relation to other health problems and a 'good' level of non-health related quality to their life respectively (table 5.8). Item responses were skewed slightly towards better levels of the underlying attribute. However, no item produced end-effects of greater than 80%.

Postal Survey Item / Scale data quality	% Missing		Mean (SD)		% floor		% ceiling		Intra-class correlation coeffcient (95% CI)	
	Postal	Clinic	Postal	Clinic	Post.	Clin.	Post.	Clin.	Postal	Clinic
PGI-AS	12.5	0.6	4.05 (1.65)	4.68 (2.25)	2.6	4.4	0.7	4.4	0.83 (.81 - .87)	0.85 (.69 - .93)
Item 1	3.4	0.6	4.22 (2.05)	4.88 (2.69)	3.0	6.6	0.6	2.6	-	-
Item 2	3.4	0.6	4.26 (2.09)	4.72 (2.11)	2.4	2.2	0.9	0.7	-	-
Item 3	3.7	0.6	4.28 (2.10)	4.33 (2.04)	2.8	3.5	1.6	0.9	-	-
Item 4	3.4	0.6	4.16 (1.98)	4.22 (2.23)	1.4	4.4	1.4	1.1	-	-
Item 5	4.3	0.6	3.81 (1.92)	3.76 (2.26)	2.6	7.4	0.9	2.9	-	-
Item 6	7.6	0.6	5.65 (2.55)	5.53 (2.62)	2.2	1.0	6.7	3.8	-	-
Item 7	4.6	0.6	6.92 (5.47)	6.90 (2.15)	0.3	1.3	4.7	6.6	-	-

Table 5.8 Item and scale properties of the PGI-AS. Results from the postal (n= 349) and clinic surveys (n= 158).

PGI-AS index score and single items are scored 0-10; higher scores indicate better disease-related quality of life.

Completion rates for the clinic survey were excellent (99.4%, n= 158). The completion rate for the postal survey was satisfactory (87.5%, n= 303). Although not clearly indicated by the developers of the PGI a final index score should not be calculated if step 2 for areas 6 and 7 is incorrectly completed or omitted (Garratt AM. - personal communication, 1999). When non-completion of areas 6 and 7 was considered a final score could be computed for 276 patients (79.8%). When a final score was calculated for the PGI-AS, with or without correct completion of areas 6 and 7, completion rates were much higher than those reported by authors using earlier versions of the PGI (table 5.9):

Authors	Patient population	n	Survey	Version of PGI	% Correct completion of PGI
	AS	349	Postal	PGI-AS	87.5
Current study	AS	349	Postal	PGI-AS (corrected)	79.8
	AS	159	Clinic	PGI-AS	99.4
Ruta et al (1994a)	Low back pain	571	Postal	Original PGI	63.0
McArthur (1997)	Rheumatoid arthritis	151	Postal	Original PGI	77.0
Herd et al (1997)	Atopic dermatitis	56	Clinic	Original PGI	100.0
McDuff &Russell (1998)	Limiting long term illness	71	Postal	Revised PGI (corrected)	62.0
Jenkinson et al (1998b)	Obstructive sleep apnoea	89	Clinic	Original PGI	100.0
Ruta et al (1999)	Low back pain Mennorrhagia Suspected peptic ulcer Varicose vein	672	Postal	Original PGI	51.0

Table 5.9 Completion rates for different versions of the PGI.

'Corrected' indicates that index score only calculated for patients completing items 6 and 7 correctly.

The PGI-AS index is scored between 0 and 10, where higher scores indicate better levels of the underlying attribute. The scores were wide ranging in both populations, and an approximately normal score distribution was observed. The mean values for both clinic and postal populations were similar (table 5.10), indicating that patients in both groups experienced similar levels of disease-related quality of life when assessed by the PGI-AS.

PGI-AS Scale	Frequency endorsement (%)	
	Postal (n= 303)	Clinic (n= 158)
floor 0 - 0.9	2.6	4.4
1.0 - 1.9	7.0	5.7
2.0 - 2.9	15.2	10.8
3.0 - 3.9	26.4	16.2
4.0 - 4.9	20.7	17.8
5.0 - 5.9	14.9	17.7
6.0 - 6.9	8.6	10.7
7.0 - 7.9	2.9	5.7
8.0 - 8.9	1.0	6.4
9.0 - 10.0 ceiling	0.7	4.4
% missing	12.5	0.6
Mean (sd)	4.05 (1.7)	4.7 (2.3)

Table 5.10 PGI-AS scale properties. Results from postal and clinic surveys.
PGI-AS is scored 0-10; higher scores indicate better disease-related quality of life.

A small percentage of postal respondents scored at the ceiling, indicating that their disease-related quality of life was 'exactly as they would like it to be' (0.7%), compared to 4.4% of clinic patients (table 5.10).

The test-retest reliability of the three follow-up formats of the PGI-AS were compared for un-weighted and weighted scores (table 5.11). Test-retest reliability greater than 0.80 was found for the PGI-AS index for all formats and the highest levels were found for the 'closed' (0.87) and 'informed and open' (0.85) formats. Un-weighted reliability was lower than the weighted reliability on all occasions (range 0.70 - 0.81). The highest un-weighted reliability was also found for the 'closed' format (0.81).

The impact of step 1 area changes on PGI-AS reliability was assessed by calculating an index of change (tables 5.12 and 5.13).

PGI-AS	Step 2 - Un-weighted score		Step 3- Index score	
	n	ICC (95% CI)	n	ICC (95% CI)
Postal survey				
Blind and Closed formats combined	143	0.76 (0.68 - 0.82)	144	0.83 (0.77 - 0.87)
Blind	71	0.71 (0.57 - 0.80)	75	0.82 (0.73 - 0.88)
Closed	72	0.81 (0.72 - 0.88)	69	0.87 (0.81 - 0.92)
Clinic survey				
Informed & open	25	0.84 (0.67 - 0.92)	25	0.85 (0.69 - 0.93)

Table 5.11 Test-retest reliability of PGI-AS for index and un-weighted scores; postal and clinic surveys.

Blind - blind to baseline areas; Closed - informed of baseline areas, not allowed to change; Informed and open - informed of baseline areas and allowed to change or to retain list.

PGI-AS 'Blind'	Step 2 - Un-weighted score		Step 3- Index score	
Index of change	n	ICC (95% CI)	n	ICC (95% CI)
0 to 1 area changes	24	0.84 (0.66 - 0.93)	25	0.91 (0.80 - 0.96)
1.5 to 2.5 area changes	22	0.75 (0.49 - 0.90)	24	0.80 (0.57 - 0.91)
3 to 5 area changes	25	0.46 (0.10 - 0.72)	26	0.56 (0.23 - 0.78)
Total population	71	0.71 (0.56 - 0.81)	75	0.82 (0.72 - 0.88)

Table 5.12 Postal survey PGI-AS test-retest reliability by area changes (blind).

Following 'blind' completion high levels of reliability were found for patients not changing any areas or scoring up to 1-point on the index of change (0.84 - 0.91). These patients may have substituted one area for another, omitted up to two areas, or added up to two additional areas. When 1.5 to 2.5 area changes were made reliability was also high (range 0.75 - 0.80). When more area changes were made reliability was low (0.46 - 0.56). On all occasions reliability was greater for the weighted index than for the un-weighted score.

All clinic patients completing the 'informed and open' format of the PGI-AS at two-weeks and indicating no change in health made between 0 and 1 area changes when assessed by the index of change (n= 23), and test-retest reliability was high (> 0.84) for both the un-weighted and index score (table 5.13). Two patients completed the PGI-AS 'blind' and have been excluded from the analysis.

PGI-AS 'Informed & open'	Step 2 - Un-weighted score		Step 3- Index score	
Index of change	n	ICC (95% CI)	n	ICC (95% CI)
0 to 1 area changes	23	0.84 (0.66 - 0.93)	23	0.85 (0.69 - 0.93)
1.5 to 2.5 area changes	0	-	0	-
3 to 5 area changes	0	-	0	-

Table 5.13 Clinic survey PGI-AS test-retest reliability by area changes (informed & open).

b) Disease-specific

ASQoL

The item and scale properties for the ASQoL are shown in table 5.14. The most frequently omitted items were 9 ('I have unbearable pain')(4.3%) and 11 ('I am unable to do jobs around the house')(4.1%). However, these items were not omitted with sufficient frequency to consider them suitable for rejection based purely on this criterion.

Scale / Item	% Missing	Mean (SD)	Response options: No % ceiling	Response options: Yes % floor	Item-total correlation	Cronbach's Alpha	ICC 95% CI (n= 166)
ASQoL	3.4	8.35 (5.6)	9.0	4.5	-	0.92	0.96 (.94-.97)
1. Limits places I can go	2.4	-	49.5	48.1	0.63	-	-
2. Sometimes feel like crying	2.0	-	57.2	40.8	0.55	-	-
3. Difficulty dressing	1.6	-	59.4	39.1	0.57	-	-
4. Struggle - jobs around home	1.8	-	43.2	55.0	0.72	-	-
5. Impossible to sleep	2.6	-	78.5	21.5	0.48	-	-
6. Activities - friends / family	3.2	-	53.8	46.2	0.65	-	-
7. Tired all the time	2.8	-	49.5	50.5	0.55	-	-
8. Stopping to rest	2.0	-	43.5	56.5	0.70	-	-
9. Unbearable pain	4.3	-	79.4	20.6	0.50	-	-
10. Time to get going - morning	3.0	-	47.0	53.0	0.63	-	-
11. Unable to do jobs at home	4.1	-	71.6	28.4	0.56	-	-
12. Tired easily	3.4	-	30.8	69.2	0.61	-	-
13. Often get frustrated	2.4	-	38.4	61.6	0.64	-	-
14. Pain is always there	1.6	-	32.9	67.1	0.57	-	-
15. Miss out on a lot	2.6	-	54.3	45.7	0.72	-	-
16. Difficult to wash my hair	2.8	-	73.4	26.6	0.59	-	-
17. Condition gets me down	2.4	-	41.4	58.6	0.68	-	-
18. Worry - letting people down	2.4	-	55.8	44.2	0.67	-	-

Table 5.14 Item and scale properties of the ASQol. Results from the postal and clinic surveys combined (n= 507).

ASQoL scored 0-18; lower scores indicate better HRQL.

ICC = Intraclass correlation coefficient; 95% CI = 95% confidence interval

During the clinic evaluation several patients asked for assistance. Most frequently patients were unable to make a clear distinction between 'yes' and 'no' response options. Several patients responded by placing a 'tick' between the two boxes, often

supplementing this with a comment such as 'sometimes' or 'it depends'. This difficulty was encountered particularly with items relating to pain (item 9, 1.3%) and the 'time to get going in the morning' (item 10, 0.4%), and prevented item scoring. Several patients (0.8%) had read through the questionnaire and 'ticked' one box only, or one box on each page.

The ASQoL employs a dichotomous response format. Both responses were covered for all items. Items 5 ('It's impossible to sleep')(78.5%), 9 ('I have unbearable pain') (79.4%) and 16 ('I find it difficult to wash my hair')(73.4%) had high levels of endorsement for the 'No' response. Items 5 and item 9 were borderline for removal but neither item exceeded the proposed criteria (> 80%), and analysis of the ASQoL dimensionality by PCA maintained the number of items proposed by the developers.

The first PCA of the ASQoL selected components with eigenvalues above 1.0, producing a three component solution to explain 55.8% of the variance (table 5.15).

	Postal and Clinic Surveys Combined (n= 452)				
	Eigenvalues > 1.0			One component	
Hypothesised scale / item	C1	C2	C3	C1	Item -total correlation
ASQoL[a]					
1. Limits places I can go	0.77			0.69	0.63
2. Sometimes feel like crying		0.62		0.60	0.55
3. Difficulty dressing	0.63			0.63	0.57
4. Struggle - jobs around home	0.71			0.77	0.72
5. Impossible to sleep			0.77	0.54	0.48
6. Activities - friends/family	0.70			0.70	0.65
7. Tired all the time		0.74		0.61	0.55
8. Stopping to rest	0.60	0.48		0.75	0.70
9. Unbearable pain			0.70	0.55	0.50
10. Time to get going -morning	0.40	0.48		0.68	0.63
11. Unable to do jobs at home	0.55			0.62	0.56
12. Tired easily		0.75		0.66	0.61
13. Often get frustrated		0.63		0.69	0.64
14. Pain is always there		0.51		0.57	0.52
15. Miss out on a lot	0.65			0.72	0.67
16. Difficult to wash my hair	0.56		0.43	0.59	0.54
17. Condition gets me down		0.72		0.68	0.64
18. Worry letting people down		0.55		0.67	0.62

Table 5.15 Principle component analyses of ASQoL and item-total correlation.

[a] instrument scoring summarised in table 5.14.

With the exception of item 10 ('It takes a long time to get going in the morning') all items had component loadings above 0.5, and the majority were above 0.6. All other items load clearly onto one of the three components. Component 1 (C1) comprises items relating to the impact of disease on functional activities and social life. Component 2 (C2) is described by items relating to the emotional impact of disease. For example, feelings of tiredness, frustration and depression. Item 10 ('It takes a long time to get going in the morning') fails to load clearly between components one or two. It may be related to component 1 due to the impact of morning stiffness on the functional activity, or to component 2 due to influence of tiredness on the ability to 'get going'. Component 3 (C3) contains two items only that may relate to extreme pain (item 5: 'It's impossible to sleep'; item 9: 'I have unbearable pain').

A further PCA of the ASQoL assessed whether the instrument was uni-dimensional, the structure recommended by the developers (table 5.15). The single component explained 42.8% of the variance. All items had component loadings above 0.50, with the majority above 0.60. The majority of items in the one component solution (11/18) have item loadings equal to or greater than the loadings in the three component solution demonstrating a strong relationship for all items within a single underlying domain.

All items demonstrated acceptable levels of item-total correlation with the remainder of the hypothesised domain (table 5.15), and further supports the relationship of all items with the underlying construct. There is little evidence to suggest that any items should be rejected from the instrument when all aspects of item performance are considered. These results suggest that the ASQoL should be considered as a single domain measure of the impact of AS on HRQL.

The completion rate for the ASQoL was excellent, and similar for both postal and clinic surveys (table 5.16), a pattern observed also in the completion of the EuroQol thermometer and the RLDQ. A final score is calculable for the ASQoL if no more than 3 items are omitted. A total of 490 patients (96.6%) completed the ASQoL correctly.

ASQoL scores for the total population covered the full range possible (0 - 18), a lower score representing better HRQL (table 5.14). The mean value was 8.35 (SD 5.60)

and scores approximated normality. This indicates that the ASQoL is capable of detecting a full range of scores and that the majority of patients included in this analysis scored in the mid-range of values representing HRQL as measured by the ASQoL. The great majority of patients would therefore be able to record improvement or deterioration in score at a follow-up assessment.

	% Completion rate							
	ASQoL	BASDAI	Body Chart	EuroQol EQ-5D	EuroQol therm	PGI-AS	RLDQ	SF-12
Combined surveys (n= 507)	96.6	89.3	-	96.8	97.8	-	96.6	92.7
Postal (n= 349)	97.1	91.0	88.8	97.4	98.3	86.8	98.0	94.8
Clinic (n= 159)	95.0	88.0	99.4	94.8	96.3	99.4	96.2	87.4

Table 5.16 Completion rates of instruments by survey.

ASQoL internal consistency reliability as estimated by Cronbach's Alpha was 0.92, and the test-retest reliability was 0.96 in the larger postal survey (n= 166)(table 5.14).

BASDAI

The item and scale properties for the BASDAI are shown in table 5.17. The most frequently omitted item was item 6 ('How long does your morning stiffness last from the time you wake up?')(21.3%). Patients are requested to identify the duration of 'morning stiffness' on a VAS (anchored '0 hours' to '2 or more hours'). Many patients failed to complete this item (n= 22, 4.4%) or completed it incorrectly (n= 86, 16.9%). For example, writing the duration of morning stiffness alongside the scale or placing obscure marks alongside the line, making it impossible to score. All other items were omitted with a similar high frequency of approximately 10%.

The level of missing data influences the calculation of scale scores. However, the developers of the BASDAI make allowance for the omission of one item from items 1 to 4, and the omission of either item 5 or 6 (Calin A. - personal communication, 1999). The final score is the average of items 1 to 4 added to the average of items 5 and 6. If one of items 1 to 4 is omitted the average of three completed items is calculated. If either item 5 or 6 is omitted the calculation is taken as the single scored item. Therefore, although many patients omitted item 6, item 5 was less frequently omitted and a final score could be calculated for the majority of patients.

Scale / Item	% Missing	Mean (SD)	% ceiling (0 - 0.9)	% floor (9.0 -10.0)	Item-total correlation	Cronbach Alpha	ICC 95% CI (n= 150)
BASDAI[a]	10.7	4.59 (2.31)	5.1	2.2	-	0.87	0.87 (.83-.90)
1. Level of fatigue/tiredness	10.6	6.00 (2.65)	6.2	8.6	0.66	-	-
2. Level of AS neck, back or hip pain	10.8	5.60 (2.68)	6.0	10.6	0.77	-	-
3. Level of pain / swelling - other than neck, back, hips	10.6	3.45 (2.94)	26.5	5.1	0.60	-	-
4. Level of discomfort - tender to touch or pressure	10.1	3.86 (3.07)	21.5	7.2	0.72	-	-
5. Severity of morning stiffness from waking	10.2	4.75 (3.05)	13.4	11.4	0.82	-	-
6. Duration of morning stiffness from waking	21.3	4.64 (3.37)	15.0	17.8	0.68	-	-

Table 5.17 Item and scale properties of the BASDAI. Results from the postal and clinic surveys combined (n= 507).

[a]BASDAI scored 0-10; lower scores indicate a less active disease state.

ICC = intraclass correlation coefficient; 95% CI = 95% confidence interval.

The BASDAI employs six 10cm horizontal VAS. When each VAS is divided into ten 1cm segments, the responses to each item cover the full available range (table 5.18).

	Frequency Endorsement (%)						
Scale (0 – 10.0)	Fatigue	Pain	Swelling	Discomfort	Morning Stiffness severity	Morning Stiffness duration	BASDAI[a] Total
0 – 0.9	6.2	6.0	26.5	21.5	13.4	15.0	5.1
1.0 – 1.9	6.4	6.4	15.4	13.6	10.3	11.3	9.9
2.0 – 2.9	10.4	9.9	10.6	13.6	10.6	14.1	13.8
3.0 – 3.9	6.8	8.2	7.3	6.3	9.9	8.0	15.4
4.0 – 4.9	8.4	9.1	6.4	8.4	8.1	4.0	12.7
5.0 – 5.9	10.4	12.8	11.7	8.7	10.1	13.5	11.5
6.0 – 6.9	15.9	13.1	5.5	6.4	7.9	3.3	13.5
7.0 – 7.9	19.2	14.8	6.9	8.3	9.9	8.2	10.8
8.0 – 8.9	7.7	9.1	4.6	6.6	8.4	4.8	4.9
9.0– 10.00	8.6	10.6	5.1	7.2	11.4	17.8	2.2
Mean	5.43	5.44	3.45	3.86	4.75	4.64	4.59
SD	2.64	2.67	2.94	3.07	3.05	3.37	2.31
n	453	452	453	456	455	399	453
Missing	54	55	54	51	52	108	54

Table 5.18 BASDAI - Frequency endorsement at item and scale level. Results from postal and clinic surveys combined (n= 507).

[a] instrument scoring summarised in table 5.17. SD = standard deviation.

Item response did not approximate normality. Responses to items 1 ('Fatigue / Tiredness') and 2 ('AS neck, back or hip pain') are skewed towards more severe levels

of the attribute, and responses to items 3 ('Pain / swelling in joints other than neck, back or hips') and 4 ('Discomfort from areas tender to touch or pressure') towards less severe levels. Responses to items 5 (Morning stiffness - severity') and 6 ('Morning stiffness - duration') both demonstrated a multi-modal distribution. The greatest levels of endorsement were for the lowest levels of 'swelling' (item 3, 26.5%) and 'discomfort' (item 4, 21.5%). No item produced an end effect greater than 80%. There was little equivalence across item means and standard deviations for all items (table 5.18). However, there was similarity between item means and standard deviations for items 1 ('Fatigue') and 2 ('Pain'), for items 3 ('Swelling') and 4 ('Discomfort'), and for items 5 ('Morning stiffness - severity') and 6 ('Morning stiffness - duration').

At item level all BASDAI items failed on the criterion of item completion, but were all retained so that measurement properties could be assessed in the group of patients completing the instrument.

The PCA of the BASDAI selected components with eigenvalues above 1.0, producing a one component solution to describe 64.5% of the variance (table 5.19). All items had component loadings of more than 0.70 with three items above 0.80. Items 5 ('Morning stiffness - severity') demonstrated the highest component loading (0.88). The lowest loading was for item 3 ('Pain / swelling in joints other than neck, back or hips'). The single dimension and high component loadings follow the developers findings that all items relate to the same underlying domain.

BASDAI	Postal and clinic surveys combined (n= 374)	
Item	Eigenvalues > 1.0 CI	Item-total correlation
1. Level of fatigue / tiredness experienced	0.76	0.66
2. Level of AS neck, back or hip pain	0.85	0.77
3. Level of pain/swelling in joints other than neck, back or hips	0.71	0.60
4. Level of discomfort from any areas tender to touch or pressure	0.81	0.72
5. Severity of morning stiffness from waking	0.88	0.82
6. Duration of morning stiffness from waking	0.79	0.68

Table 5.19 Principle component analysis of the BASDAI and item-total correlation.

[a] instrument scoring summarised in table 5.17.

All items demonstrated high levels of item-total correlation with the remainder of the hypothesised domain further supporting the relationship of all items with the single underlying domain.

The completion rate of the BASDAI was reasonable. A total of 453 patients (89.3%) completed sufficient items in the BASDAI to produce a final score. For those instruments where data for both surveys could be combined the BASDAI completion rate was the lowest (table 5.16). The response rate was comparable to that of the Body Chart and the PGI-AS in the postal survey and the SF-12 in the clinic survey.

The BASDAI score is represented by a scale of 0-10, where lower scores indicate a less active disease state. A wide range of scores was covered, but the lowest score of '0' ('no' disease activity) was not calculated. However, when the response scale was assessed in intervals of '1.0', 5.1% of scores were at the ceiling (range 0-0.9)(table 5.18). A lower endorsement for the highest possible scores ('very severe') was observed with 2.2% of scores at the floor of the scale.

The mean score was 4.59 (SD 2.31) and scale scores approximated normality. This suggests that the BASDAI is capable of recording a full range of scores, and that the majority of patients were described by the 'mid-range' of disease activity as measured by the BASDAI. The majority of patients would therefore be able to record improvement or deterioration in score at a follow-up assessment.

The internal consistency reliability of the BASDAI as estimated by Cronbach's Alpha was 0.89, and test-retest reliability was 0.87 (n= 150) in the larger postal survey.

Body Chart

The body chart asks patients to indicate areas of bodily pain on a body manikin and to score each area. The scoring range has a lower limit of 'zero' but no upper limit, with a higher score indicating a greater level of perceived body pain. The instrument does not consist of individual items and is considered at scale level only.

A final score was computable for 99.4% of the clinic population, but a much lower completion rate of 89.9% was calculated following self-completion in the postal survey (table 5.16 and 5.20).

Body Chart Scale / Item (response range)	% Missing	Mean (SD)	Median	% ceiling	% floor	ICC (95% CI)
Postal survey (n=310)	11.1	16.16 (16.20)	11.0	2.3 (0)	0.3 (122)	0.86 (.81-.90) (n= 142)
Clinic survey (n= 158)	0.6	10.89 (11.19)	8.00	8.3 (0)	0.6 (56)	0.87 (.73-.94) (n= 25)

Table 5.20 Item and scale properties of the Body Chart.

Body Chart: minimum score 0, where 0 is no bodily pain (ceiling). No limit to maximum score.

ICC = Intraclass correlation coefficient; 95% CI = 95% confidence interval.

Features of incorrect completion in the postal survey (n= 36, 10.3%) included shading various parts of the body chart, with a subsequent failure to score areas. If more than two areas were not scored a final score could not be awarded (Dziedzic K. - personal communication, 1999). Several patients shaded the manikin with great precision but circled a single score, thus preventing scoring. Only 3 patients (0.8%) left the body chart blank. Such problems were not encountered during interview-administration.

A much greater range of scores was observed in the postal population (range 0 - 122) when compared to the clinic population (0 - 56). However, the distribution of scores was skewed towards the lower levels of pain for both populations with a greater percentage of the clinic population scoring at the ceiling of the range. 8.3% of the clinic population indicated that they were not experiencing body pain, whereas only 2.3% of the postal population reported 'no pain' (table 5.20). Mean values were greater for the postal population (mean 16.15, SD 16.20) than for the clinic population (mean 10.89, SD 11.19).

Reporting the associated confidence intervals for the ICC as an estimation of test-retest reliability is based on the assumption that data is approximately normally distributed. Therefore, due to the positively skewed distribution of the Body Chart data, data was logarithmically transformed to yield a lognormal distribution (Altman, 1996)(table 5.20). Similar levels of test-retest reliability of the Body Chart were estimated in both postal (0.86, n= 142) and clinic surveys (0.87, n= 25).

RLDQ

The most frequently omitted item from the RLDQ was item 4a ('Walking on your heels') omitted by 7.7% of patients (table 5.21). During self-completion of the RLDQ

in the clinic, several patients requested advise with this item and responded with comments such as:

> 'I've never walked on my heels before – I don't know if I can do it'
>
> and
>
> 'Walk on your heels - what does this mean?'

In addition, several patients were observed 'walking on their heels' along the clinic corridor, putting their 'functional ability' to the test. Item 4b ('Cough or sneeze') was also frequently omitted (5.3%).

			Response options [b]						
Instrument Scale / Item	% Missing	Mean (SD)	0 % ceiling	1	2	3 % floor	ITC	Alpha	ICC 95% CI (n=166)
RLDQ[a]	3.4	13.56 (9.9)	5.1	58.5	32.7	3.7	-	0.93	0.94 (..92-.96)
1. Mobility								-	-
a. Into and out of the bath	3.7	0.74 (0.87)	48.7	33.7	12.1	5.5	0.76	-	-
b. Into and out of the car	2.8	0.80 (0.75)	39.5	41.5	18.6	0.4	0.67	-	-
c. Up/ out of bed-morning	4.1	0.71 (0.69)	42.1	44.8	12.9	0.2	0.61	-	-
d. Rolling over in bed	3.3	0.78 (0.70)	36.0	51.3	10.8	1.8	0.63	-	-
2. Bending Down								-	-
a. Wiping yourself - toilet	3.0	0.42 (0.64)	65.1	29.0	4.7	1.2	0.57	-	-
b. Put on / take off socks	3.1	0.85 (0.82)	37.8	44.3	13.2	4.7	0.73	-	-
c. Put on shoes & tie laces	4.7	0.89 (0.90)	38.6	41.1	12.6	7.6	0.74	-	-
d. Cut your toe nails	3.0	1.18 (1.11)	32.5	38.1	8.1	21.3	0.77	-	-
3. Neck Movements								-	-
a. Opening high windows	3.3	0.92 (1.01)	42.2	36.5	8.4	13.0	0.77	-	-
b. Look both ways before crossing the road	3.1	0.73 (0.86)	48.4	35.0	11.6	5.1	0.72	-	-
c. Look at what you are reaching on a high shelf	3.1	0.95 (1.05)	42.7	34.6	8.1	14.6	0.72	-	-
d. Drink from a small glass or can	3.5	0.56 (0.83)	61.8	24.7	9.4	4.1	0.62	-	-
4. Posture								-	-
a. Walk on your heels	7.7	0.55 (0.92)	65.4	23.2	2.1	9.2	0.61	-	-
b. Coughing or sneezing	5.3	0.54 (0.60)	51.4	43.7	4.6	0.4	0.43	-	-
c. Sleep on your back	3.1	1.08 (1.09)	35.6	41.1	3.5	19.9	0.47	-	-
d. Sleep on your stomach	4.5	1.70 (1.21)	19.6	34.2	3.1	43.1	0.59	-	-

Table 5.21 Item and scale properties of the RLDQ. Results from postal and clinic surveys combined (n=507).

[a] RLDQ scored 0-48; lower scores indicate better functional ability.

[b] Each item is scored 0-3. Response options: 0='Able to do without difficulty', 1='Able to do with difficulty', 2='Only able to do using unusual movements or gadgets', 3='Unable to do'.

ITC = Item-total correlation; Alpha = Cronbach's Alpha; ICC = Intra-class correlation coefficient; 95% CI = 95% confidence interval.

Although the full range of response options was found for all items, three items had very low levels of endorsement for the option 'Unable to do' (table 5.21). Less than 0.5% of patients indicated an inability to perform activities described by items 1b ('Getting into and out of the car'), 1c ('Getting up and out of the bed in the morning'), and 4b ('Coughing or sneezing').

A large proportion of patients scored at the ceiling of the range for several items indicating that the activity could be performed 'without difficulty': items 2a ('Wiping yourself after using the toilet')(65.1%), 3d ('Drinking from a small glass or can')(61.8%) and item 4a ('Walk on your heels')(65.4%). However, no item produced an end-effect of greater than 80.0%.

The skewed distribution of item responses suggests that most respondents experience no or only moderate limitation in functional activities described by the RLDQ. This is reflected in the low mean values for all items (item range 0 - 3). Nine items have similar mean values (range 0.73 - 0.95; items 1a-d, 2 b-c, and 3 a-c). Four items have very low mean values (< 0.55; items 2a, 3d and 4a-b). The remaining three items have marginally higher mean values (> 1.00; items 2 d, and 4 c-d).

Analysis of RLDQ dimensionality by PCA retained the number of items proposed by the developers. The first PCA selected components with eigenvalues above 1.0, producing a three component solution to describe 65.5% of the variance (table 5.22). All items had component loadings of more than 0.5, except for item 4a ('Walk on your heels')(0.44), with the majority above 0.60. All components correspond to recognisable aspects of functional ability. Component 1 (C1) contains items representative of bending activities, consisting of all items from section 2 ('Bending down') with the addition of items 1a ('Into and out of the bath') and items 4a ('Walk on your heels'). The relationship between 'walking on your heels' and items describing bending activities such as 'wiping yourself after using the toilet' is not clear. Component 2 (C2) is described by activities that impose demands upon neck mobility, consisting of all items from section 3 ('Neck movements'), with the addition of item 4d ('Sleep on your stomach'). The relationship between these items has clinical validity. To lie prone requires an adequate range of neck rotation combined with extension. Component 3 (C3) contains activities representative of general mobility and functional activities, containing three items from section 1 (b-d) and two items

from section 4 (b and c). The relationship between 'coughing or sneezing' and items describing functional activities such as 'rolling over in bed' is not clear.

	Postal and Clinic Surveys Combined (n= 490)				
Instrument Scale / Item	Eigenvalues > 1.0			One component	
RLDQ [a]	C1	C2	C3	C1	Item -total correlation
1. Mobility					
a. Into and out of the bath	0.61	0.46		0.81	0.76
b. Into and out of the car	0.47		0.51	0.73	0.67
c. Up/ out of bed-morning	0.41		0.68	0.67	0.61
d. Rolling over in bed			0.68	0.69	0.63
2. Bending Down					
a. Wiping yourself - toilet	0.64			0.63	0.57
b. Put on / take off socks	0.85			0.78	0.73
c. Put on shoes & tie laces	0.85			0.79	0.74
d. Cut your toe nails	0.74	0.41		0.81	0.77
3. Neck Movements					
a. Opening high windows	0.42	0.74		0.80	0.77
b. Look both ways - cross road		0.80		0.76	0.72
c. Look at what reaching-high shelf		0.86		0.76	0.72
d. Drink from small glass/can		0.82		0.66	0.62
4. Posture					
a. Walk on your heels	0.44			0.66	0.61
b. Coughing or sneezing			0.72	0.48	0.43
c. Sleep on your back			0.58	0.51	0.47
d. Sleep on your stomach	0.57	0.40		0.63	0.59

Table 5.22 Principle component analyses of RLDQ and item-total correlation

[a] instrument scoring summarised in table 5.21.

The RLDQ produces a single index score representing a single underlying domain of AS-specific functional disability (Abbott et al, 1994). Scores are not generated for individual sections. Therefore, a further PCA of the RLDQ assessed instrument uni-dimensionality (table 5.22). All items had component loadings above 0.50, except for item 4b ('Cough or sneeze')(0.48). The majority of items in the one component solution (14/16) have item loadings equal to or greater than 0.6 supporting the uni-dimensionality.

All items demonstrated acceptable levels of item-total correlation with the remainder of the hypothesised domain of functional ability and further supports instrument unidimensionality. When all aspects of item performance are considered there is limited evidence to support the rejection of specific items from the RLDQ.

A final score was calculable for the RLDQ if no more than 2 items per section of the instrument were omitted. A total of 490 patients (96.4%) completed the RLDQ adequately to give a final score. A wide range of scores was observed, although the maximum score of 48 was not achieved (range 0 - 41). The mean score for the total population was 13.56 (9.59) and scale score distribution was skewed towards the better levels of functional ability.

5.1% of patients scored at the ceiling. No patients scored at the floor of the scale, but 3.7% of scores were at the lower range of the scale (33 - 41 points) indicating great difficulty with the majority of activities. The most frequently endorsed levels of functional ability described mild (score range 1-16)(58.5%) to moderate (score range 17 - 32)(32.7%) levels of difficulty with some or all functional activities.

The internal consistency reliability of the RLDQ as estimated by Cronbach's Alpha was 0.93, and test-retest reliability was 0.94 in the larger postal survey (n= 166).

c) Generic

EuroQoL

The item properties of the EuroQol EQ-5D are shown in table 5.23. The most frequently omitted were items 2 ('Self-care')(1.4%) and 4 ('Pain / Discomfort')(1.4%).

			Response options		
EuroQol EQ-5D Item level	% Missing	Mean (SD)	% ceiling No problems (1)	Some problems (2)	% floor Extreme problems (3)
1. Mobility	0.6	1.58 (0.50)	42.5	57.1	0.4
2. Self-care	1.4	1.37 (0.51)	64.6	34.2	1.2
3. Usual activities	0.4	1.78 (0.57)	29.7	62.4	7.9
4. Pain /Discomfort	1.4	2.08 (0.53)	10.4	71.0	18.6
5. Anxiety/Depression	1.2	1.60 (0.60)	45.5	48.5	6.0

Table 5.23 Item properties of the EuroQol EQ-5D. Results from combined postal and clinic surveys (n= 507).

The majority of patients responded using the first two response options. However, no item produced an end effect or level of endorsement of greater than 80%.

The mean values (SD) for items 1 - 3 and 5 were very similar. The mean value for item 4 ('Pain / Discomfort') was slightly higher.

Completion of the EQ-5D was good and a final index score could be calculated for 97.0% of patients. The full range of index scores was found (table 5.24).

		Score range			
EQ-5D Scale level	% Missing	Mean (SD)	% ceiling (1.00)	% floor (-0.59)	ICC (95% CI) (n= 165)
Index score	3.0	0.54 (0.33)	7.7	0.4	0.85 (.80 -.89)

Table 5.24 Scale properties of the EuroQol EQ-5D. Results from combined postal and clinic surveys (n= 507).

EuroQol EQ-5D is scored -0.59 - 1.0; -0.59 represents a state worse than death, and 1.0 the best possible health.
ICC = intraclass correlation coefficient; 95% CI = 95% confidence interval.

The distribution of index scores were slightly skewed towards a more positive health state. 38 patients (7.7%) scored at the ceiling of the scale range ('best possible health state') and two patients scored at the floor (0.4%)('worst possible health state').

Test-retest reliability of the EQ-5D was 0.85 in the larger postal survey (n= 165).

Few problems with completion of the EuroQol thermometer were experienced. The main problem concerned patients simply striking a line through the thermometer instead of 'drawing a line from the box' toward the thermometer. Completion rates were very good (97.6%)(table 5.25).

Thermometer Scale / Item (response range)	% Missing	Mean (SD)	% floor (0-10)	% ceiling (90-100)	ICC (95% CI) (n= 166)
Thermometer score	2.4	59.56 (21.36)	1.2	8.1	0.83 (.78 -.87)

Table 5.25 Scale properties of the EuroQol Thermometer. Results from combined postal and clinic surveys (n= 507).

EuroQol Thermometer is scored 0-100; higher scores indicate better HRQL.

The wide range of scores (0 to 100) suggests that the thermometer was able to detect all health states from 'worst imaginable' to 'best imaginable' in this population. The results were slightly skewed towards a more positive health state, with a mean value of 59.56 (21.36). When the scale was assessed in segments of '10-points', 1.2% of patients scored at the floor of the scale (between 0 and 10) and 8.1% of patients scored at the ceiling (between 90 and 100).

The test-retest reliability of the thermometer was 0.83 in the larger postal survey (n= 166).

SF-12

The item and scale properties of the SF-12 are shown in table 5.26. The most frequently omitted items were items 7 ('Didn't do work or other activities as carefully as usual' - emotional)(3.5%) and item 5 ('Were limited in the kind of work or other activities' - physical)(3.3%). Items 6 ('Accomplished less than you would like' - emotional)(2.0%) and item 4 ('Accomplished less than you would like' - physical) (1.8%) were also frequently omitted.

			Response options		
SF-12 Item / Scale	% Missing	Mean (SD)	% ceiling	% floor	ICC (95% CI) (n= 156)
Mental component scale	7.3	46.33 (10.43)	0.0	0.0	0.79 (.73 -.84)
Physical component scale	7.3	36.95 (10.04)	0.0	0.0	0.90 (.86 -.92)
1. General Health	1.2	2.80 (1.31)	4.4	11.0	-
2. Moderate activities	1.2	2.42 (1.80)	24.3	27.7	-
3. Climb flights of stairs	1.4	2.53 (2.04)	33.7	32.5	-
4. Accomplished less- physical	1.8	1.86 (2.42)	37.3	62.7	-
5. Limited in work/activities - physical health	3.3	1.92 (2.44)	38.5	61.5	-
6. Accomplished less- emotional	2.0	2.92 (2.47)	58.4	41.6	-
7. Work/activities as carefully - emotional	3.5	0.76 (0.61)	60.8	39.2	-
8. Pain - normal work	1.0	2.59 (1.45)	8.5	11.5	-
9. Calm and peaceful	1.0	2.58 (1.37)	5.6	6.8	-
10. Lot of energy	0.8	2.90 (1.46)	6.5	15.7	-
11. Downhearted and blue	1.0	3.37 (1.23)	18.9	2.8	-
12. Physical/Emotional health - social activities	1.0	3.40 (1.49)	35.8	4.4	-

Table 5.26 Item and scale properties of the SF-12. Results from postal and clinic surveys combined (n=507). All items recoded (0-5).

The SF-12 uses norm-based scoring from the general population. Scales are transformed to have a mean of 50 (sd=10), with a range 0-100; higher scores indicate a better level of HRQL.

Items recoded 0-5, where 5 indicates a better level of HRQL.

To allow direct comparison of data quality and scaling assumptions between items, all response options were re-coded to equivalent scales (0-5). The response options for all items were covered and no item produced an end effect greater than 80%.

There was equivalence of item means (SD) for seven items (items 1-3,6,8-9), with an approximately normal item response distribution. Item means were also similar for

items 4 (accomplished less than would like - physical) and 5 (limited in kind of work / activities - physical), with a slightly skewed distribution of responses towards lower levels of HRQL. Items 11 (downhearted and blue) and 12 (physical / emotional health - social activities) had similar mean values with a slightly skewed distribution of responses towards better HRQL. Item 7 (work / activities as carefully as usual - emotional) had a very small mean value with no equivalence to other items and responses skewed towards better HRQL.

A total of 470 (92.7%) patients completed all items in the SF-12. Mental (MCS) and physical component summary (PCS) scores were calculated for these patients. A wide range of scores were observed although the full possible range of either scale was not covered (MCS range 17.00 to 66.83; PCS range 12.04 to 63.34). The mean score for the MCS was 46.34 (SD 10.43), a value that was closer to the mean population score (50, SD 10) than the PCS mean value of 36.95 (SD 10.03). The scores for the MCS approximated normality. Scale sores for the PCS were slightly skewed towards the less physically abled scores.

Levels of test-retest reliability were greater for the SF-12 PCS (0.90, n= 156) than for the MCS (0.79).

Summary of test-retest reliability

Similar levels of test-retest reliability were found for both surveys (table 5.27). The lowest levels of reliability for both surveys was found for the SF-12 MCS (range 0.72 - 0.79). All other instruments in the postal survey had levels of reliability above 0.80 with three instruments producing estimates above 0.90 (ASQoL, RLDQ, SF-12 PCS). Three instruments in the clinic survey had estimated reliability levels greater than 0.90 (RLDQ, Body Chart and EuroQol-thermometer). Only the RLDQ consistently demonstrated this high level of reliability in both surveys.

5.4.2 Results of data quality and test-retest reliability of anthropometric measures

All patients participating in the clinic survey were assessed for their available range of movement by five different anthropometric measurements. Scale properties are shown in table 5.28.

Instrument / Score	Postal survey		Clinic survey	
	n	ICC (95% CI)	n	ICC (95% CI)
ASQoL	166	0.96 (.94 - .97)	23	0.87 (.73 - .94)
BASDAI	150	0.87 (.83 - .90)	21	0.89 (.74 - .95)
Body Chart (log)	142	0.86 (.81 - .90)	25	0.87 (73 - .94)
EuroQol EQ-5D	165	0.85 (.80 - .89)	23	0.83 (.63 to .92)
EuroQol thermometer	166	0.83 (.78 - .87)	22	0.92 (.82 to .97)
PGI-AS blind and closed combined results	144	0.83 (.77 - .87)	-	-
PGI-AS - blind	75	0.81 (.72 - .90)	-	-
PGI-AS - closed	69	0.87 (.81 - .92)	-	-
PGI-AS informed & open	-	-	25	0.85 (.69 to .93)
RLDQ	166	0.94 (.92 - .96)	23	0.93 (.83 - .97)
SF-12 - MCS	156	0.79 (.73 - .84)	21	0.72 (.75 - .95)
SF-12 - PCS	156	0.90 (.86 - .92)	21	0.89 (.42 - .87)

Table 5.27 Test-retest reliability for study instruments (postal (n= 173) and clinic surveys (n=25)).

ASQoL scored 0-18; lower scores indicate better HRQL.
BASDAI scored 0-10; higher scores indicate greater disease activity.
Body Chart scored from 0, with no maximum score. Higher scores indicate greater perceived body pain.
EuroQol - EQ-5D scored -0.59 - 1.0; 1.0 is the best possible health.
EuroQol - thermometer scored 0-100; higher scores indicate better health states.
PGI-AS scored 0-10; higher scores indicate better disease-related quality of life. Blind - blind to baseline areas; Closed - informed of baseline areas, but not allowed to change; Informed & open - informed of baseline areas, and allowed to change.
RLDQ scored 0-48; higher scores indicate greater functional disability.
SF-12 uses norm-based scoring from the general population. Scales are transformed to have a mean of 50 (sd 10), with a range 0-100; higher scores indicate better levels of HRQL. MCS - mental summary scale; PCS - physical summary scale.
ICC = intra-class correlation coefficient. 95% CI = 95% confidence interval

The measurements of cervical rotation, fingertip to floor distance (FFD) and the Modified Schober Index (15cm)(MSI) may all record minimum values of 0cm. This describes maximal limitation, or the 'floor', of cervical rotation and the MSI, but the 'ceiling' of FFD (maximal range equates to touching the floor). This score is not possible for lateral lumber flexion (LLF) or tragus to wall distance (TWD). It is not possible to describe a true end effect for these measurements, there being no pre-defined maximal or minimal limit to the available range of movement. The number of patients scoring the maximum or minimum range of movement for the each measure is reported (table 5.28).

Values for cervical rotation approximated normality, with equal range and similar mean values for both right and left rotation. Very few patients scored at the floor (0.0 - 1.0cm) (n= 1, 0.6% left rotation; n= 2, 1.3% right rotation) or at the ceiling of the range (14.0-15.0cm) (n= 2, 0.6% left rotation; n=4, 1.9% right rotation).

Clinic Anthropometric measures	Mean (SD) (cm)	Minimum score cm	Minimum score %	Maximum score cm	Maximum score %	ICC (n= 25)
Cervical rotation						
- left	7.36 (3.74)	0.0 0.0 - 1.0	0.6 4.3	15.00 14.0-15.0	0.6 1.2	0.87
- right	7.28 (3.57)	0.0 0.0 - 1.0	0.6 2.4	15.50 14.0-15.0	2.4 1.9	0.95
Fingertip to floor distance	19.71 (15.73)	0.0	22.2	60.8	0.6	0.98
Lateral lumbar flexion						
- left	53.10 (6.59)	25.8	0.6	68.4	0.6	0.95
- right	52.21 (6.30)	21.3	0.6	65.0	0.6	0.98
Modified Schober Index (15cm)	4.02 (2.30)	0.0 0.0 - 1.0	0.6 10.8	9.0 8.0-9.0	0.3 3.3	0.95
Tragus to wall Distance	17.94 (7.11)	9.1 9.1-10.0	0.6 3.2	44.5 35.0-44.5	0.6 3.2	0.98

Table 5.28 Scale properties of anthropometric measures. Clinic survey (n= 159)

Cervical rotation - distance between tip of nose and acromioclavicular joint measured in neutral and maximum ipsilateral rotation. Difference between two positions calculated, where a smaller difference indicates a more restricted range.

Fingertip to floor distance - distance between tip of right middle finger and floor following maximum trunk flexion, where the smaller distance indicates greater movement.

Lateral lumbar flexion - distance between tip of middle finger and floor measured following maximum ipsilateral lateral flexion, where the smaller distance indicates greater movement.

Modified Schober Index (15cm) (Macrae and Wright, 1969) - distance between two marks placed 15cm apart in standing (10cm proximal and 5cm distal to posterior superior iliac spine). Distance after maximum trunk flexion, where a larger difference indicates greater lumbar movement.

Tragus to wall distance - Distance between right tragus and wall measured in standing, where a larger distance indicates worse spinal / upper cervical posture.

A wide range of results was observed for FFD (0cm to 60.0cm). A large percentage of patients were capable of touching the floor, indicating very good trunk flexibility (n=35, 22.2%). When these patients were taken into account, the results remained slightly skewed towards the smaller values, indicating better movement. When assessed in ranges of 10cm most patients were described by the range between 10-20cm FFD (27.2%). 11.4% of the population achieved measurements of more than 40cm FFD, with only one patient (0.6%) scoring at the floor of the recorded range (60.8cm).

Lateral lumbar flexion was slightly skewed towards more limited movement. A wide range of results was found, with most patients recording ranges between 50-59cm (55.1% left, 55.7% right). The distance was less than 40cm in very few patients (left 2.5%, n= 4; right 3.8%, n= 5).

A wide range of values were found for the MSI (range 0 - 9.00cm). The measurement described a bi-modal distribution with a cluster of patients recording very limited movement (range 0-2.9cm), and a second cluster with greater range of movement (3.0-9.0cm). One patient scored at the floor of the available range (0.0cm) and two patients scored at the ceiling for this population (9.0cm, 1.3%). When the movement is considered in 1cm segments, 10.7% (n= 17) of patients scored in the range 0-1.0cm, and 3.3% (n= 7) patients scored between 8.0 - 9.0cm.

TWD provides an indication of the standing posture of the patient. Smaller values indicate a better posture. An increasingly stooped posture, often a hallmark of disease progression, increases TWD. A large range of values were found (range 9.1 to 44.5cm), with results greatly skewed towards the lower values. However, 7.6% of the population recorded values in excess of 30.0cm.

High levels of test-retest reliability were found for all measurements (table 5.29), and levels greater than 0.90 were found for all measurements, except left cervical rotation.

	Test-retest reliability		Inter-observer reliability (baseline values)	
Anthropometric measures [a]	n	ICC (95% CI)	n	ICC (95% CI)
Cervical rotation: starting position				
- left	-	-	51	0.65 (.25 - .82)
- right				0.68 (.50 - .80)
Cervical rotation: difference				
- left	25	0.87 (.94 - .73)	51	0.94 (.89 - .96)
- right		0.95 (.89 - .98)		0.90 (.84 - .94)
Fingertip to floor distance	25	0.98 (.96 - .99)	51	0.96 (.94 - .98)
Lateral lumbar flexion				
- left	25	0.95 (.89 - .98)	51	0.96 (.94 - .98)
- right		0.98 (.95 - .99)		0.98 (.97 - .99)
Modified Schober Index (15cm)	25	0.95 (.90 - .98)	51	0.90 (.83 - .94)
Tragus to wall distance	25	0.98 (.96 - .99)	51	0.98 (.97 - .99)

Table 5.29 Clinic survey test-retest reliability (AS and general health transition the same, n= 25), and inter-observer reliability (2 observers) of anthropometric measures.

[a] a description of each measurement can be found in table 5.28.
ICC = intraclass correlation coefficient; 95% CI = 95% confidence intervals.

High inter-observer reliability was also found (table 5.29). The only exception was for the starting positions for cervical rotation (0.65 - 0.68). All other measurements achieved levels greater than 0.90.

5.5 Discussion

Reliability may be defined as the extent to which an instrument produces consistent results when the underlying condition has not changed, and relates to consistency of results at one point in time or over time. For multi-item instruments based on classical test-construction theory internal consistency reliability describes the relationship between items measuring aspects of the same domain and is assessed at one point in time. Test-retest reliability assesses instrument temporal stability.

Before instrument reliability is considered, performance both at item and scale level should be assessed through tests that relate to data quality and scaling assumptions. These tests have implications for item reduction and have been considered alongside the tests of internal consistency reliability. Where appropriate the study has demonstrated more extensive testing of the data quality and scaling assumptions, internal consistency reliability and test-retest reliability for all study instruments than previously reported. Also, a larger population of patients with AS than previously reported has been involved which increases the generalisability of the result.

Two unresolved issues around the design of test-retest studies can be described: first, the most appropriate time period between repeat administrations; and secondly, how to account for patients who have actually changed. Published studies reporting the test-retest reliability of the study instruments have adopted various retest periods, varying from 1-hour to 3-months. Also, many fail to describe an external assessment of change. The current research has adopted a two-week retest period (Streiner and Norman, 1995) and may provide a more appropriate assessment of test-retest reliability than previously described for many study instruments. Researchers have attempted to describe change in health using clinical criteria, clinical opinion and health transition questions (Deyo et al, 1991; Fitzpatrick et al, 1993b). Health transition questions should address the underlying domain measured by the instrument and in the current study questions inquiring about change in both disease-specific and general health were adopted, with stability in both items being sought. The PGI-AS incorporates both disease-specific and generic issues in the final score and inclusion of both transition questions was particularly relevant.

The data quality, scaling assumptions and reliability of individual instruments were assessed for both clinic and postal data individually. Results for both populations

were similar. Sample sizes for principal component analysis (PCA) were maximised by combining the two data-sets which produced components that were more readily interpretable. Results are discussed and recommendations for the modification and application of instruments made.

a) PGI-AS

This study has provided the first evidence for the data quality, scaling assumptions and test-retest reliability of the PGI-AS. Completion rates for the PGI-AS improve on levels reported by previous authors applying the original PGI (Ruta et al, 1999) which could be due to changes in the instrument structure. The PGI-AS trigger list is more extensive than those included in earlier versions of the PGI and closer affinity to items may have helped improve completion. During interview administration several patients remarked on the relevance of areas to their own life, gaining solace from the realisation that other patients with AS experienced similar problems.

The test-retest reliability of the constituent parts of the PGI-AS and for three different formats was assessed. This deconstruction of the instrument promotes increased understanding of the contribution of each stage to overall reliability and any weaknesses can be identified. The three formats all had levels of reliability greater than 0.80 for the index score supporting their use in group evaluation (Ware, 1997). This is an improvement on all previous reliability estimates (Ruta et al, 1994a; MacDuff and Russel 1998; Ruta et al, 1999).

High levels of reliability were found for patients not changing any areas in step 1, or scoring up to 1-point on the index of change when completing the 'blind' (0.91) or 'open and informed' (0.85) formats. Although only a small sample size, this supports application of the 'blind' format in individual evaluation (>0.90). However, increased areas changes reduced reliability, and when more than three area changes were made reliability was not acceptable for group evaluation. When considered irrespective of the index of change 'blind' completion produced the lowest reliability (blind 0.81; informed and open 0.85; closed 0.87).

Most patients completing the 'informed and open' format and indicating no change in health at two-weeks retained all baseline areas. Following the pattern observed for the 'blind' format it may be hypothesised that reduced reliability would be associated

with an increase in area changes. Repetition of the study with a larger sample size may provide evidence to address this hypothesis.

For all variants of the PGI-AS the un-weighted score reliability was lower than that estimated for the weighted index. The results satisfied levels recommended for group analysis (range 0.71 - 0.84) except where the index of change was high. This result suggests that the 'points' spent in step 3 are adding to instrument reliability.

Evidence suggests that the most reliable versions of the PGI-AS are the 'closed' and 'informed and open' formats. In choosing the most appropriate follow-up format the trade-off between reliability and content validity should be considered. The assessment of disease-related quality of life in AS is a relative heterogeneous issue and the open nature of the PGI-AS baseline completion ensures the individuality of content to support the instrument conceptual base. The results suggest that during follow-up completion when informed of baseline areas and given the opportunity to change areas most patients reporting no change in their underlying condition choose not to make large changes to areas listed. Therefore, 'blind' completion may introduce 'noise' to follow-up completion of the PGI-AS leading to a reduction in reliability, and the 'informed and open' format may not be necessary. Therefore, keeping selected areas the same at follow-up completion may improve the clinical validity of the instrument for evaluative purposes, without threatening content validity.

However, this result focuses only on patients indicating no change in condition at two-weeks and patients reporting change may not demonstrate the same stability of areas. The acceptability of providing patients with their baseline responses without providing the format whereby items can be changed or retained as necessary fails to support the potentially dynamic and reflective nature of the PGI. This is an issue that will be returned to in the evaluation of instrument responsiveness (Chapter 7).

Patients completing the PGI-AS in the clinic survey requested greatest assistance with step 2 (scoring areas). A reduction in the number of response options from 10 to, for example, 5 may reduce the complexity of this step thereby improving acceptability. It would also reduce the variance associated with the score and may improve reliability.

The majority of scoring anomalies during self-completion of the PGI-AS involved the spending of points in step 3. Many instrument developers adjust for missing data and a 'window of error' could be calculated for the PGI-AS so that a score is not jeopardised by a possible miscalculation of points. For example, spending 13 to 15 points. Several patients indicated that '14' points was an unusual number and found division difficult. A more rounded number such as '10' or '100' was felt to be more acceptable. Other patients indicated that there were 'just not enough points', finding prioritisation difficult. A revision of points in step 3 to '10' may be more acceptable and may facilitate prioritisation of areas.

b) Disease-specific

Completion rates for the ASQoL and the RLDQ were very good and most items in both instruments had only minimal levels of missing data. No items were omitted by more than 10% of patients. Several patients (0.8%) had read through the ASQoL and 'ticked' one box only which could be in response to the request 'tick the 'one' response that best applies', and further clarification may be beneficial. Also, several patients had difficulty with the dichotomous response. The most frequently omitted item from the RLDQ was item 4a ('Walk on your heels')(7.7%).

There were unacceptable levels of missing data for the BASDAI, in particular for item 6 ('How long does your morning stiffness last from the time you wake up?'). Strict application of the proposed criteria for item rejection would have resulted in removing the instrument from the study. However, the BASDAI was retained so that its measurement properties could be compared to the other instruments albeit for the subset of patients that completed the instrument adequately.

The suggestion by the BASDAI developers that VAS offer highly sensitive response scales capable of detecting small degrees of change (Garrett et al, 1994) should be considered in light of the many completion inaccuracies encountered in this study. A revision of the response scale for the whole instrument is strongly recommended. For example, replacing the VAS with five or seven-point adjectival scales with discrete or continuous responses (Streiner and Norman, 1995).

Completion rates for the Body Chart were greater for the clinic than for the postal survey (99.4% vs 88.8%). Improved completion in the clinic survey may be

explained by the ability of the therapist to further clarify instructions. Furthermore, the routine use of the instrument at the SRC may have helped to familiarise patients with the instrument. A larger score range was found in the postal population. Patients may have taken more time to consider their body pain and to complete the chart without feeling pressurised by the presence of the therapist. Alternatively, clinic patients often requested clarification about the time period over which to consider their pain and the therapist may have clarified the request for 'current or present pain' only. A greater percentage of the clinic population scored at the ceiling. However, this may reflect the failure of the instrument to clearly indicate the procedure for self-completion if 'no pain' is experienced. Body charts devoid of shading and scoring following self-completion may indicate non-completion due to lack of understanding or lack of pain.

The expression of bodily pain on the body chart is limited by the approach adopted in shading the manikin. If the manikin is shaded as 'one area' and a single score of 4 awarded this low score would be in contrast to the patient shading the manikin as several areas and giving each area a score of 4. Although both patients describe pain traversing the whole body the final score does not describe the equality in pain distribution or severity. The low completion rate following self-completion indicates that modifications to enhance completion are required. For example, further clarification of the scoring procedure, both when minimal or extreme pain is present and when multiple areas of pain are identified.

A check of the measurement properties is appropriate when an instrument is modified or applied in a new setting (Bjorner et al, 1998), and this study provides the first evidence of the data quality, scaling assumptions and reliability of the revised RLDQ. Although all response options at item level were covered and end-effects did not exceed critical levels, the distribution of responses was skewed towards better levels of functional ability. The descriptors used in the ordinal response scale may fail to provide clear options for fine discrimination between levels of functional ability at item level with the majority of responders endorsing 'Able to do' and 'Able to do with difficulty'. Patients may find difficulty distinguishing between these options when, during the slow progress of the disease they learn to adapt to their environment. If response options and descriptors are limited a major change in functional ability is

required before a change in response is recorded, and if responses fail to cover the full range of limitations in AS the instrument may have limited validity.

However, increasing the number of response options or changing the descriptors may not improve score distribution if the majority of patients score at the end of a scale. Therefore, there is also a case for using different items representing 'harder' functional activities that offer better discrimination between patients. Several items (2d, 4c-d) had a less skewed distribution of responses with higher mean values which would suggest that they addressed 'harder' functional activities. Two of these items (4c-d) also had lower levels of item-total correlation than all other items. Therefore, although appearing less efficient when assessed by item-internal consistency they may accomplish an objective of raising the scale ceiling and improving content validity (Ware and Gandek, 1998b). The final scale score of the RLDQ demonstrated a more normally distributed response although the full range of responses reflecting increased functional difficulty was not recorded. The need to include more arduous functional activities in the parent instrument, the LDQ, was described in Chapter 2. A similar problem with the revised RLDQ has been described by the current study.

The assessment of dimensionality and item-total correlation supports the unidimensionality of the ASQoL, BASDAI and RLDQ proposed by the instrument developers. The results provide evidence in support of the data quality and scaling assumptions of the ASQoL at both item and scale level and of the BASDAI and RLDQ at scale level.

Published evidence of the reliability of both the ASQoL and RLDQ is not available, and study estimates of both internal consistency reliability and test-retest reliability are very good supporting their use in individual evaluation. Greater levels of test-retest (0.97) reliability than internal consistency reliability (0.92) were calculated for the ASQoL. Internal consistency reliability (0.93) and test-retest reliability (0.94) for the RLDQ were similar, and were similar to levels reported for the parent instrument, the LDQ (Abbott et al, 1994).

Test-retest reliability of the BASDAI was satisfactory and equivalent to the internal consistency reliability (0.87). The estimate of internal consistency reliability improved on published levels (Jones et al, 1996c; Calin et al, 1999a), although test-

retest reliability was lower (Garrett et al, 1994) and did not support application in individual evaluation. Satisfactory test-retest reliability was found for the Body Chart (> 0.86). The BASDAI and Body Chart have levels of reliability acceptable for use in group evaluation.

c) Generic

This study has provided the first evidence for the reliability of the EuroQol and the SF-12 in AS. The EuroQol EQ-5D and thermometer had the lowest levels of missing data and hence the best completion rates for all study instruments. Completion rates of the SF-12 were also good. Items relating to limitation in work or usual activities due to emotional or physical health problems were most frequently omitted. Non-completion of these items in the SF-12 and in the parent instrument, the SF-36, has been previously reported in patients with RA (Hurst et al, 1998; Ruta et al, 1998). Although the omission of individual items was not high (range 0.8 to 3.5%), the omission of a single item prevents the calculation of a final score in the SF-12 (7.3%). The SF-36 allows patients to omit up to half of the items in the instrument without jeopardising a final score. It is suggested that final score calculation of the SF-12 could be improved if the criterion for item completion was not so strict and the role of mean score computation could be assessed. The subsequent impact of this on measurement properties would need to be addressed.

Test-retest reliability for both parts of the EuroQol (EQ-5D 0.85; thermometer 0.83) and the SF-12 MCS (0.79) was good, supporting use in group evaluation. The result for the EuroQol improves on those reported by authors applying the instrument in patients with RA (EQ-5D 0.78; thermometer 0.83) (Hurst et al, 1997). A high level of test-retest reliability was found for the SF-12 PCS (0.90), supporting application in individual evaluation. The two-week test retest reliability of the SF-12 has not been reported in a UK patient population with a similar status to AS.

d) Anthropometric measures

Both fingertip to floor distance following trunk anterior flexion (FFD) and lateral lumbar flexion (LLF) were measured purely as a reflection of the distance between the tip of the middle finger and the floor following the described movement. Measurement did not consider patient height or the starting distance in neutral. Therefore, a tall patient with large range of movement may record a similar fingertip

to floor distance as a patient of smaller stature with limited movement. In light of this anomaly the results for both measures are informative at the individual level only. Calculation of the score as a percentage of the starting position distance may provide a more standardised result suitable for group comparison.

A wide range of values were recorded for the MSI (mean 4.02cm (SD 2.30); range 0 - 9.00cm). Moll et al (1971) investigated the available range of lumbar flexion in 'normal' subjects when measured by the MSI and found an approximately normal distribution of results (n= 237; mean 4.93cm (SD 0.90) to 7.23cm (SD 0.92) dependent on age and sex groupings; range 3.5-10.0cm). The population described in the current study therefore covered a broad spectrum of lumbar mobility ranging from extreme limitation to levels comparable with the normal population.

All anthropometric measures had estimates of test-retest reliability greater than 0.90, achieving the more stringent levels of reliability for individual evaluation (Nunnally and Bernstein, 1994).

Conclusion

When data quality and scaling assumptions are assessed, all instruments, except for the BASDAI, demonstrate adequate properties at both item and scale level. The BASDAI demonstrates adequate properties at scale level in those patients completing the instrument adequately to receive a score but modification of the response format to improve item completion is strongly recommended.

All measures of outcome are sufficiently reliable for group evaluation. The ASQoL, RLDQ, SF-12 PCS and all anthropometric measures have estimates of reliability that support application in individual evaluation (Nunnally and Bernstein, 1994). Modifications to the BASDAI, Body Chart, PGI-AS and RLDQ have been suggested that may lead to improved data quality, scaling assumptions and levels of reliability. Modifications to the scoring of FFD and LLF have been made that may support adoption in group evaluation.

Chapter 6 Validity

6.1 Introduction

This chapter presents the validity testing for the study instruments. Section 6.2 describes the different types of validity used to assess measures of health outcome. Section 6.3 presents the methods of validity testing for the study instruments. After the results are presented in section 6.4 the chapter closes with a discussion.

6.2 Validity and measures of health outcome

Validity considers how well an instrument measures what it purports to measure in the settings in which the instrument may be applied (Nunnally and Bernstein, 1994). Evidence of the validity of applying an instrument in different situations, to evaluate underlying theories, to investigate different relationships and to further establish confirmatory information in support of the instruments purpose is an on-going process. The process of validity testing facilitates score interpretation and furthers appreciation of what change in scores actually mean (Ware, 1997).

6.2.1 Methods of validity testing

Three forms of validity are most frequently applied in the evaluation of measures of health outcome: content, criterion and construct validity (Ware, 1997). Evidence to support the extent to which an instrument is capable of expressing an underlying and defined domain is common to all forms of validity. Although methodologically the three forms vary, all may have the concept of construct validity as their base, construct validity representing the 'basic meaning of validity' (Streiner and Norman, 1995).

Content validity

Content validity does not involve statistical analysis, but does require a more qualitative appreciation of instrument content. Judgement of domain coverage and the inclusion of representative items requires an accepted definition or theoretical description of instrument purpose. Appreciation of methods adopted in item selection may also be beneficial (Fitzpatrick et al, 1998a). However, instrument developers often omit to detail these important issues (McDowell and Newell, 1996; Chapter 2). If the purpose and conceptual base is not clarified or instrument content fails to cover the defined domain invalid inferences may be made.

Criterion validity

The assessment of criterion validity requires the presence of external criteria for which known and independent evidence in support of the criteria is available and against which the results of the instrument may be compared (Ware, 1997). If criterion information is produced concurrently then it is concurrent validity. If it is produced in the future then it is predictive validity.

Circumstances where evidence in support of concurrent validity may be observed in the evaluation of measures of health outcome are where refined versions of an instrument are assessed alongside the original. It would be hoped that the modified version would retain the important attributes of the original and a strong relationship would be predicted. In response to the need to produce a shorter patient-based measure of outcome the SF-12 was developed from the SF-36 (Ware et al, 1995). Multiple regression analysis was used to select twelve items from the eight scales of the parent measure that best represented the physical (PCS) and mental (MCS) component summary scales (Ware et al, 1994; Gandek et al, 1998b). The performance of the SF-12 is comparable to that of the SF-36 in several different patient populations (Ware et al, 1996; Hurst et al, 1998), whilst having the advantage of being shorter, quicker to complete and therefore potentially more acceptable to clinical practice and research.

In the field of health outcomes it is unlikely that an instrument is available for which evidence of criterion validity or a 'gold standard' is available. Ware (1997) suggests that most criteria are infact variables with a conceptual relationship to the domain of concern. They are not criteria in the sense of an independent 'gold standard' but they are capable of providing meaningful information for empirical tests evaluating instrument application. This requires that underlying relationships between different variables are considered. Most assessments of the validity of measures of health outcome will rely on establishing evidence in support of the hypothesised relationship between variables which is defined as construct validity.

Construct validity

Construct validity considers the relationship between a measure and a hypothetical construct. A construct may be defined as a non-observable variable which by

definition address a more general hypothesis than that supported by a specific behaviour (Nunnally and Bernstein, 1994). In the field of health outcomes variables relating to, for example, HRQL may be considered constructs due to the inability to relate such concepts to a concrete and observable theory. Theories to describe the hypothesised relationship between such variables, for example, between HRQL and functional disability, may be described as 'hypothetical constructs'. Instruments designed to measure health outcome attempt to address an aspect of this hypothetical construct (Streiner and Norman, 1995).

The first step in establishing evidence to support the construct validity of an instrument is to clearly define the domain addressed. Once a domain and a construct have been defined the relationship between the instrument and different variables can be investigated to test underlying theories (Ware, 1997). The support for construct validity lies in the accumulation of evidence from various experiments to test a range of hypothesised relationships.

Two forms of validity have been described which form the basis of construct validity. These are convergent and discriminant validity where the statistical relationship between variables is usually correlational (Ware, 1997). Convergent validity addresses the relationship between two different measures of the same construct. It would be hypothesised that these measures would produce similar results and demonstrate a strong correlation. Alternatively, discriminant validity estimates the relationship between two instruments measuring unrelated constructs. These instruments should not be related and small correlation would be expected. The adoption of the multitrait-multimethod matrix (MTMM)(Campbell and Fiske, 1959) allows a number of validity issues, in particular convergent and discriminant validity, to be addressed simultaneously. In circumstances where different methods of assessing the same domain have been included in the evaluation or where different domains have been measured with several instruments the convergent or discriminant nature of the relationships may be assessed.

McDowell and Jenkinson (1996) stress the importance of predicting the strength of relationship to be expected between instruments *a priori*, and of describing both convergent and discriminant relationships 'to permit validity to be *dis*proved'. Although it is difficult to predict relationships between constructs the level of

statistical correlation provides evidence in support of the hypothesised relationship. If it is hypothesised that a new instrument addresses a major part of the domain addressed by an established instrument a large correlation would be predicted (Streiner and Norman, 1995). Too high a correlation would suggest that the instruments were measuring very similar attributes and that the new instrument was offering little more than a different approach. If the new instrument addressed only one aspect of a multi-dimensional domain addressed by the established instrument then a small correlation would be expected.

In assessing validity the level of instrument reliability and criteria to which it is being compared should be considered as this places an upper limit on the expected level of association (Streiner and Norman, 1995). McHorney et al (1993) provide guidance to interpretation of the level of correlation between variables: a correlation of more than 0.70 would describe a large relationship, 0.50 a moderate relationship, and less than 0.30 a small relationship.

Construct validity can also be assessed by the use of 'extreme groups' which theorises that one group will possess more or less of the defined attribute to be measured by the instrument (Streiner and Norman, 1995). In support of the construct validity of the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), scores of a population of AS in-patients and AS out-patients were compared (Garrett et al, 1994). The authors theorised that the in-patient population would have higher scores on the BASDAI than the out-patient population. This was the association which was evidence for the validity of the BASDAI.

Many instrument developers investigate the relationship between health and socio-deomographic characteristics as a basis for evidence in hypothesis testing for construct validity. Several authors have reported on the positive relationship between health and socioeconomic status (Garratt, 1997).

Published evidence in support of the validity of all disease-specific study instruments (Chapter 2) is summarised in table 6.1. Most developers fail to provide *a priori* hypotheses. Rather, a list of correlation coefficients is provided which the authors indicate as evidence for instrument validity.

Author	Instrument	AS population (n)	Method	Correlation coefficient
Garrett et al (1994)	BASDAI	46 in-patients 108 out-patients	Comparison between scores for in-patient and out-patients	Higher scores for in-patients (p=0.005)
		46 in-patients	Correlation between BASDAI and Bath Disease Activity Index (Bath DAI = original BASDAI)	0.75
Jones et al (1996a)	BASDAI	200 out-patients	Correlation between the BASDAI and the Bath AS Global scale (BAS-G)	0.73
Dziedzic (1997)	Body Chart	69 out-patients	Correlation between Body Chart and:	
			Patient perceived pain (VAS)	0.32
			Night pain (VAS)	0.55
			Stiffness (VAS)	0.54
			Enthesitis scores	0.50 to 0.53
			Pain disability questionnaire	0.50
			Cervical mobility (tape measure)	-0.30 to -0.42
			Fingertip to Floor Distance	0.36
			Lumbar spine side flexion	0.30 to 0.39
			Modified Schober Index (flexion)(15cm)	-0.30
			Tragus to Wall distance.	0.31
Abbott et al (1994)	LDQ	42 out-patients	Two groups according to LDQ score - 'best' and 'worse' function. Comparison between groups for age, disease duration, ROM and posture.	'Worse' functional group - older, longer disease duration, limited ROM / posture.
			Correlation between sections of LDQ and:	
			Cervical mobility (goniometer)	-0.27 to -0.73
			Chest expansion	-0.33 to -0.54
			Fingertip to Floor Distance	0.35 to 0.62
			Lumbar spine extension	-0.31 to -0.60
			Modified Schober Index (15cm)	-0.38 to -0.62
			Tragus to Wall distance	0.30 to 0.57

Table 6.1 Evidence of the validity of disease-specific study instruments.

There is no published evidence in support of the ASQoL, RLDQ and PGI-AS in patients with AS.

6.3 Methods for assessing the validity of the study instruments

The measurement properties of the EuroQol and SF-12 have not previously been assessed in patients with AS. However, both have well documented evidence of development and validity testing, and have evidence supporting their validity as generic measures of HRQL in patients with chronic disorders similar to AS (table 6.2) (Ware et al, 1995; Hurst et al, 1997; Hurst et al, 1998). Evaluation in patients with RA supported the validity of the EuroQol as a measure of HRQL (Hurst et al, 1997), the SF-12 PCS as a measure of physical health and of the SF-12 MCS as a measure of mental health (Hurst et al, 1998). The SF-12 MCS was reportedly a better measure of mental health than the SF-36 MCS, the longer version demonstrating that in part it was measuring the impact of RA on physical status (Ruta et al, 1998).

Author	Instrument	Patient population	Validation
Hurst et al (1994)	EuroQol	55 RA out-patients	Correlation between EQ-5D and EQ-thermometer with HAD; HAQ and clinical measures
Hurst et al (1997)	EuroQol	233 RA out-patients	Correlation between EQ-5D and MHAQ, HAD and socio-economic variables
Ruta et al (1998)	SF-36	233 RA out-patients	SF-36 PCS and MCS scores and eight scales with SF-12 PCS and MCS scores, MHAQ, core disease activity measures, HAD
Hurst et al (1998)	SF-12	233 RA out-patients	SF-12 PCS and MCS scores with SF-36 PCS and MCS scores, eight SF-36 subscales, MHAQ and the HAD

Table 6.2 Studies assessing the validity of the EuroQol, SF-36 and SF-12 in patients with Rheumatoid Arthritis.

Key: MHAQ - Modified Health Assessment Questionnaire; HAD - Hospital Anxiety and Depression questionnaire.

Therefore, the EuroQol and SF-12 can contribute to the validation of other study instruments and both are expected to correlate more strongly with other measures of HRQL, disability and pain than with measures of impairment and disease process.

6.3.1 Validity of the study instruments

The content and conceptual base of a multi-item instrument provides a strategy by which the construct validity may be considered (Keller et al, 1999b). The content validity of all patient-based instruments has been summarised in Chapters 2 and 5, and now consideration is given to the expected relationships between all study instruments. Figure 6.1 summarises the tests of validity performed.

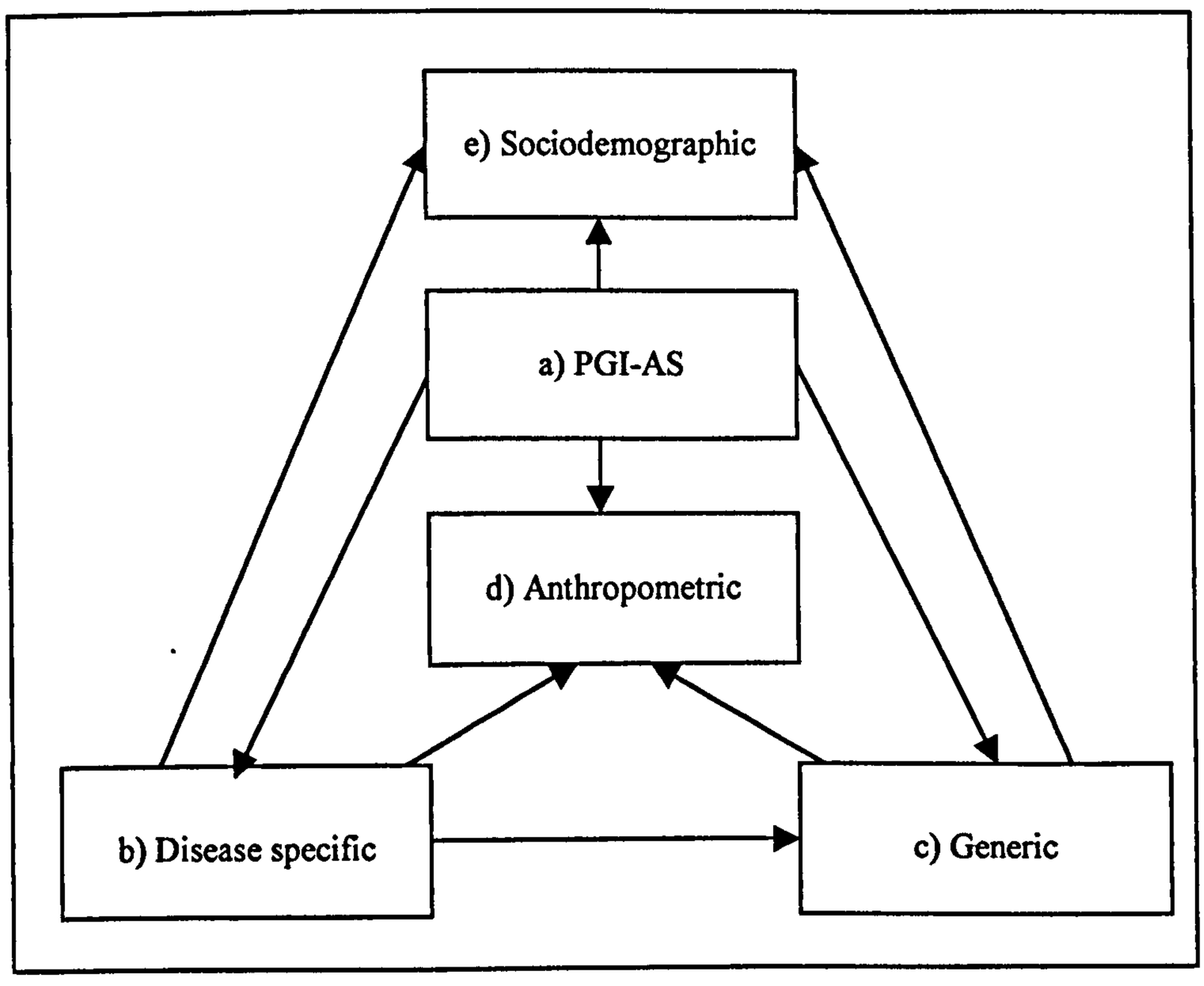


Figure 6.1 Tests of validity for the study instruments

The PGI-AS is the most recent instrument to be applied in AS. Therefore, it is correlated with disease-specific, generic and anthropometric measures of outcome, and compared against sociodemographic variables. There are a number of disease-specific instruments within AS and those selected for inclusion in the study are correlated with each other to assess the convergent validity of related dimensions. They are also compared to the generic instruments, the anthropometric measures and against sociodemographic variables. The generic instruments are used to test the construct validity of the PGI-AS and the disease-specific instruments, and are therefore assessed through comparison with anthropometric measures and against sociodemographic variables. The anthropometric measures are assessed for convergent validity through correlation with other anthropometric measures.

To support the generation of theoretical relationships between patient-based instruments the underpinning purpose and item content of all instruments is considered (tables 6.3 and 6.4). Although the defined purpose is important in considering the relationship between instruments, item content and content overlap play the most important role in supporting hypothesis generation. The Body Chart is a disease-specific measure of current or present bodily pain, and does not contain individual items.

Hypothesised theoretical relationships between all domains were considered a priori (table 6.5) and are detailed in the following section. Where possible, relationships tested previously by instrument developers were re-tested.

a) PGI-AS

The PGI-AS is a disease-specific measure of disease-related quality of life, represented as the gap between a patients expectations and reality (after Ruta et al, 1994a). The PGI-AS reflects the extent to which a patients' experience of a disease may detract from fulfillment of life quality.

Content validity

The frequency with which trigger list and supplementary items are mentioned will be considered support for the adequacy to which issues akin to AS-related quality of life are measured by the PGI-AS, hence supporting the content validity of the trigger list.

Disease-specific instruments				
ASQoL - HRQL	BASDAI - disease activity	PGI-AS - trigger list disease-related quality of life		RLDQ - functional disability
Limits places	Fatigue / tiredness	Impact on work	Life enjoyment	Getting into and out of the bath
Feel like crying	AS neck, back or hip pain	Worry about future	Worry-letting people down	Getting into and out of the car
Difficulty dressing	Pain / swelling - joints other than neck, back, hips	Relationship with wife / husband / partner	Independence	Getting up and out of bed - morning
Struggle with jobs at home	Discomfort - tender to touch / pressure	Unable to plan ahead	Difficulty dressing / washing	Rolling over in bed
Impossible to sleep	Morning stiffness (severity)	Feelings of low self-esteem	Ability - jobs around the home	Wiping yourself after using the toilet
Unable to join in activities with friends / family	Morning stiffness (duration)	Ability to play with children	Relationship with friends	Putting on / take off socks
Tired all of the time		Sex life	Social life	Put on shoes / tie laces
Have to keep stopping to rest		Family life	Embarrassment	Cut toe nails
Unbearable pain		Pain	Self body image	Open high window
Time to get going - in the morning		Disturbed sleep	Worry - future	Looking both ways - crossing road
Unable to do jobs around the home		Difficulty 'getting going' in morning	Fatigue	Look at what reaching - high shelf
Tired easily		Walking	Feeling tired	Drinking - small glass
Often get frustrated		Difficulty sitting/ standing / lying down	Loss of motivation	Walk on heels
Pain always there		Physical activity	Depression	Cough or sneeze
Miss out on a lot		Fear of falling	Moody	Sleep - back
Difficult-wash hair		Increased time to do things	Hobbies	Sleep - stomach
Condition gets me down		Control over life	Sporting activities	
Worry-letting people down		Ability to plan ahead	Driving	
			Limited spinal mobility	
			Mental activity	

Table 6.3 Definition of purpose and item content of disease-specific instruments.

Generic instruments			
EuroQol 5-D HRQL	EuroQol Thermometer health status	SF-12 MCS mental component scale	SF-12 PCS physical component scale
Mobility 'No problems walking about' to 'confined to bed'	Range: 'Worst imaginable health state' to	Accomplished less - role emotional	General health - (EVGFP rating)
Self-care 'No problems' to 'Unable to wash or dress myself'	'Best imaginable health state'	Not careful - work / activities - role emotional	Moderate activities - physical function
Usual activities (e.g., work, study, housework, family or leisure activities) 'No problems' to 'Unable to perform'		Calm and peaceful	Climb several stairs
		Energy	Accomplished less - role physical
Pain / discomfort 'No pain / discomfort' to 'extreme pain / discomfort'		Downhearted and blue	Limited in kind of work / activities - role physical
Anxiety / depression 'Not anxious or depressed' to 'extremely anxious or depressed'		Social time - physical health or emotional problems	Bodily pain interfere with normal work - includes work outside home and housework

Table 6.4 Definition of purpose and item content of generic instruments.

The item content of the PGI-AS will be compared to that of the only other AS-specific measure of HRQL, the ASQoL. It was hypothesised that a broader range of areas listed by patients completing the PGI-AS might confer greater validity as a measure of HRQL in AS above that of the ASQol (Herd et al, 1997).

Construct validity

In support of the conceptual base of the PGI Garratt (1997) suggests that 'health is not usually valued for its own sake, but for the extent to which it influences our ability to enjoy life'. Rather than simply deriving a nominal score for health and non-health related issues determined important by a patient, the PGI-AS considers the relationship between these issues and the ability to fulfill ones expectations of life. Therefore, to the extent that health or disease influences life quality and the PGI-AS is a valid measure of disease-related quality of life, a moderate level of correlation would be hypothesised between the PGI-AS and other disease-specific and generic instruments (table 6.5).

From a total of seven items that may be considered in completion of the PGI-AS five are disease-specific. Item 6 relates to general health status and the remaining item refers to non-health issues. It is therefore likely that the majority of PGI-AS content will be disease-specific and a larger (moderate) correlation with disease-specific instruments than seen with generic instruments is hypothesised. However, the inclusion of 'non-health' items is hypothesised to reduce the expected relationship of the PGI-AS with both disease-specific and generic instruments.

The PGI-AS and ASQoL are disease-specific measures of HRQL with similar item content (table 6.3), but the generic and non-health components of the PGI-AS suggests that there will be a moderate, as opposed to a large level of correlation. Herd et al (1997) compared an original version of the PGI with a disease specific measure of HRQL (Dermatology Quality Life Index - DQI) and found a moderate correlation (-0.52).

The main issues measured by the BASDAI (disease activity) and Body Chart (bodily pain) are within the PGI-AS trigger list (table 6.3). It is hypothesised that the BASDAI and Body Chart would have moderate levels of correlation with the PGI-AS.

	Hypothesised association [a]												
	ASQoL	BASDAI	Body Chart	EuroQol EQ-5D	EuroQol - therm	PGI-AS	RLDQ	SF-12 MCS	SF-12 PCS	Cervical rotation	FFD	Lat lsp flex	Modified Schober
BASDAI	+++												
Body Chart	+++	+++											
EuroQol EQ-5D	++	++	++										
EuroQol - therm	++	++	++	++									
PGI-AS	++	++	++	++	++								
RLDQ	+++	++	++	++	++	++							
SF-12 - MCS	++	+	+	++	++	++	+						
SF-12 - PCS	++	++	++	++	++	++	++	+					
Cervical rotation	+	+	+	+	+	+	++	+	+				
FFD	+	+	+	+	+	+	++	+	+	++			
Lat lsp flex	+	+	+	+	+	+	+	+	+	+	+		
Modified Schober	+	+	+	+	+	+	++	+	+	+	+	++	
TWD	+	+	+	+	+	+	++	+	+	++	+	+	++

Table 6.5 Hypothesised associations between all study instruments and anthropometric measures.

[a] Scale of association: +++ Large (> 0.70), ++ Moderate (0.50), + Small (0.30).

Key: EuroQol therm - thermometer; FFD - fingertip to floor distance; Lat lsp flex - lateral lumbar flexion; Modified Schober - Modified Schober Index; TWD - tragus to wall distance.

Several issues measured by the RLDQ (functional ability) are within the PGI-AS trigger list (table 6.3). However, to the extent that additional items not included in the trigger list may be described by patients completing the PGI-AS a moderate correlation between these two instruments is expected.

Previous research has demonstrated small to moderate levels of correlation between the original version of the PGI and scales of the SF-36 in patients with low back pain (Ruta et al, 1999). The largest correlations were between the PGI and the pain (0.47) and social functioning scales (0.38). Although the PGI-AS is disease-specific and the EuroQol and SF-12 generic measures of HRQL, several issues addressed by the generic instruments are within the PGI-AS trigger list (tables 6.3 and 6.4). Therefore, a small to moderate correlation is hypothesised between the PGI-AS and both generic instruments.

Anthropometric measures of impairment are not expected to have large correlations with patient-based instruments. Where movement limitation is potentially reversible such limitations are expected to influence ones ability to fulfill the demands and expectations of life, and a small to moderate correlation with the patient-based instruments are expected. Measures reflecting irreversible change in axial status and spinal mobility are associated with disease progression and changes occur over prolonged periods of time (van der Linden and van der Heijde, 1995; Calin et al, 1999c). A patient may learn to compensate for movement limitation with a consequent change in expectation. A small relationship with measures of HRQL is expected.

It is hypothesised that the PGI-AS would: 1) have moderate levels of correlation with the ASQoL, BASDAI, Body Chart, and the RLDQ; 2) have small to moderate levels of correlation with the EuroQol and SF-12; 3) have small levels of correlation with cervical rotation and fingertip to floor distance (FFD); 4) have little or no correlation with other anthropometric measures (table 6.5).

b) Disease-specific instruments

The ASQoL (HRQL), BASDAI (disease activity), Body Chart (bodily pain) and the RLDQ (functional ability) all measure related aspects of HRQL. The main issues measured by the BASDAI and Body Chart, and several items measured by the RLDQ,

are within the item content of the ASQoL (table 6.3). Therefore, a large level of correlation is hypothesised between the ASQoL and the BASDAI, with moderate to large levels of correlation between the ASQoL and both the Body Chart and RLDQ (table 6.5).

The BASDAI and Body Chart measure closely related aspects of health and have a similar item content (table 6.3), and a large correlation is hypothesised. There is minimal overlap of item content between the BASDAI or Body Chart and the RLDQ (table 6.3). However, all instruments measure related aspects of health and active disease and pain have an impact on normal function. Moderate levels of correlation between these three instruments are hypothesised.

A similarity of item content exists between the ASQoL and the EuroQol EQ-5D and SF-12 PCS (tables 6.3 and 6.4). However, items in the ASQoL are not clearly anchored to AS and a moderate to large correlation is hypothesised with the EQ-5D and SF-12 PCS, and a moderate correlation with the EuroQol thermometer. If ASQoL items were more clearly anchored to the impact of AS a more moderate correlation would be hypothesised. Bodily pain is one aspect of HRQL measured by the SF-12 PCS and EQ-5D (table 6.4), and so a moderate correlation is hypothesised between the BASDAI and the Body Chart and these generic instruments, and small to moderate correlation with the EuroQol thermometer. The EuroQol and SF-12 are generic instruments developed to measure multiple health domains related to HRQL and so a moderate correlation with instruments measuring varied disease-specific domains rather than a very large correlation with any single domain is hypothesised (table 6.5).

Pain, tiredness and fatigue may be associated with adverse mental health (Jenkinson et al, 1999a). A similar item content between the ASQoL, BASDAI and SF-12 MCS exists (tables 6.3 and 6.4), and so a moderate correlation with the SF-12 MCS is hypothesised. A small to moderate correlation is hypothesised between the Body Chart and the SF-12 MCS.

Both the RLDQ and SF-12 PCS have similar item content (tables 6.3 and 6.4), but the disease-specific and generic nature suggests that there will only be a moderate level of correlation. Similarly, a moderate correlation is hypothesised between the RLDQ and

both sections of the EuroQol. The RLDQ and SF-12 MCS address different issues and therefore a small correlation is hypothesised.

Limited mobility affects the ability to satisfy ones needs as reflected by items in the ASQoL and a small relationship with anthropometric measures which may reflect reversible change is hypothesised. Pain influences mobility and so a small correlation between the BASDAI, the Body Chart and these anthropometric measures is also hypothesised. However, anthropometric measures representative of irreversible change in mobility are not expected to have any relationship with measures of HRQL.

The developers of the RLDQ reported moderate to large correlations between sections of the parent instrument, the LDQ, and a selection of anthropometric measures (table 6.1) (Abbott et al, 1994). The largest correlation was between cervical rotation and section 3 (Neck mobility). However, the correlation between the LDQ index score and anthropometric measures was not reported.

Activities measured by the RLDQ (table 6.3), are influenced by limited mobility and a moderate association with most anthropometric measures is hypothesised. Several items reflect cervical mobility and a moderate to large correlation with cervical rotation and tragus to wall distance (TWD) is hypothesised. A moderate correlation with FFD and the Modified Schober index (MSI) is expected due to the impact of reduced flexibility and limited spinal mobility on several activities measured. Lateral lumbar flexion (LLF) is not expected to influence many activities measured and a small correlation is hypothesised.

As summarised in table 6.5, it is hypothesised that all disease-specific instruments would: 1) have a moderate to large level of correlation with other disease-specific instruments; 2) have a moderate (to large) level of correlation with the EuroQol (EQ-5D and thermometer) and SF-12 PCS; 3) have a small to moderate correlation with the SF-12 MCS; 4) have a small correlation with cervical rotation and FFD, and 6) have little or no correlation with remaining anthropometric measures. Only the RLDQ is expected to have a moderate to large correlation with cervical rotation, FFD, the MSI and TWD.

c) Generic instruments

The EuroQol and SF-12 have similar item content (table 6.4), and so a moderate correlation is hypothesised between the EuroQol and the SF-12 PCS, and a small to moderate correlation with the SF-12 MCS (table 6.5). Evidence suggests that a small to moderate correlation is expected between the EQ-5D and thermometer (Hurst et al, 1994), and a small correlation between the two components of the SF-12 (Ware, 1997).

Specific spinal mobility has little association with the domains measured by the generic instruments (table 6.4), and little or no correlation is expected. However, a small correlation with measures reflecting reversible change is hypothesised. Evidence supporting the relationship between anthropometric measures and generic instruments of HRQL has not been identified.

d) Anthropometric measures

A large correlation between the MSI and LLF and progressive AS-specific radiographic change in the lumbar spine, and between TWD and change in the cervical spine has been reported (Kennedy et al, 1995; Dawes, 1999). A limited MSI and LLF are key features of AS and are included in the diagnostic criteria (van der Linden et al, 1984). Increased TWD is also a feature of progressive disease (Dziedzic, 1998). These findings support the inference that these measures are reflective of structural and irreversible change in AS. Changes in cervical rotation and FFD may, to a certain extent, be reversible (Roberts et al, 1988).

It is hypothesised that cervical rotation will have a moderate correlation with FFD and TWD. Cervical rotation and FFD will have small levels of correlation with the other anthropometric measures. The MSI will have a moderate correlation with lateral lumbar flexion and TWD (table 6.5).

e) Sociodemographic variables

The validity of all patient-based instruments was further assessed in relation to socioeconomic status reflected by level of post-school education, occupational status, and housing tenure. T-tests were used to test for differences between these groups of patients.

Level of income and extent of education has a positive relationship with health (Hay, 1988). Patients not continuing education beyond minimum school leaving age and those without a degree or equivalent qualification were hypothesised to have scores reflecting worse levels of health than their counterparts on all instruments.

Patients reporting an inability to work due to ill-health were expected to have scores reflecting worse levels of health on all instruments than their counterparts who were in employment. Housing tenure was applied as a proxy for socioeconomic status with patients from a lower social class expected to report lower levels of health on all instruments (Hurst et al, 1998).

6.4 Results of validity testing

The results of validity testing are presented for the PGI-AS, disease-specific and generic instruments, anthropometric measures and finally all instruments compared to sociodemographic variables.

6.4.1 Results of validity testing of the study instruments

Analysis of the ASQoL, BASDAI, RLDQ, EuroQol and the SF-12 has been based on the combined results of postal and clinic surveys to maximise the sample size (table 6.6). This is justified because completion in both surveys followed the same self-completed format. The Body Chart and PGI-AS followed different completion formats and data is assessed separately for each survey (tables 6.7 and 6.8). Anthropometric measures were only assessed in the clinic survey (table 6.8).

	ASQoL	BASDAI	EuroQol - EQ - 5D	EuroQol - therm	RLDQ	SF-12 MCS
BASDAI	0.75					
EQ - 5D	-0.75	-0.69				
EuroQol -therm	-0.68	-0.59	0.61			
RLDQ	0.68	0.62	-0.59	-0.49		
SF-12 - MCS	-0.53	-0.40	0.49	0.39	-0.26	
SF-12 - PCS	-0.70	-0.57	0.58	0.52	-0.64	0.20

Table 6.6 Correlation between scores for patient-based measures of outcome. Combined postal and clinic survey (n= 398).

	ASQoL	BASDAI	Body Chart	EuroQol - EQ 5D	EuroQol - therm	RLDQ	SF-12 MCS	SF-12 PCS
Body Chart	0.67	0.72	-	-0.62	-0.56	0.55	0.38	0.60
PGI-AS	- 0.53	-0.52	-0.46	0.50	0.57	-0.42	0.46	0.51

Table 6.7 Correlation between the Body Chart and PGI-AS with other patient-based study instruments. Postal survey (n= 224).

a) PGI-AS

Content validity

Patients in the postal survey listed a total of 68 areas including all 37 trigger list items. The frequency with which items were selected is shown in table 6.9. The most frequently mentioned area of importance affected by AS was 'work', often described as 'impact on ability to work' or 'threat to work'. This was followed by the impact of AS on 'sleep'.

All but one of the items within the ASQoL are addressed by the PGI-AS trigger list (table 6.10). The item not within the list, item 13 ('I often get frustrated'), was identified as important by four (1.2%) patients in the postal survey.

All correlations between the PGI-AS and other instruments were in the hypothesised directions and moderate correlations between the PGI-AS and all disease-specific and generic instruments were found. As hypothesised, a hierarchy of association with the largest correlations between the PGI-AS and disease-specific instruments followed by the generic instruments and finally the anthropometric measures was found. Similar results were calculated for both postal and clinic data.

The largest (moderate) correlation was with the ASQoL (-0.53 postal, -0.70 clinic), and the smallest (moderate) correlation with a disease-specific instrument was with the RLDQ (-0.41 postal, -0.55 clinic). Moderate levels of correlation were found with the EQ-5D (0.50 postal, 0.54 clinic) and SF-12 MCS (0.46 postal, 0.57 clinic). Slightly larger correlations were found with the EuroQol thermometer (0.57 postal, 0.42 clinic) and SF-12 PCS (0.51 postal, 0.60 clinic). The small levels of correlation with cervical rotation and FFD and the weak correlations with the remaining anthropometric measures were hypothesised.

	ASQoL	BASDAI	Body Chart	EuroQol EQ - 5D	EuroQol - therm	PGI-AS	RLDQ	SF-12 MCS	SF-12 PCS	Csp rot - left	Csp rot - right	FFD	LLF-left	LLF - right	Modified Schober
Body Chart	0.61	0.62	-	-0.56	-0.46	-0.54	0.44	-0.40	-0.57						
PGI-AS	-0.70	-0.61	-0.54	0.54	0.42	-	-0.56	0.57	0.60						
Csp rot - left	-0.48	-0.44	-0.34	0.34	0.33	0.34	-0.69	0.20	0.45						
Csp rot - right	-0.52	-0.42	-0.36	0.36	0.46	0.34	-0.65	0.20	0.48	0.89					
FFD	0.30	0.28	0.28	-0.31	-0.36	-0.29	0.41	-0.07	-0.38	-0.41	-0.49				
LLF - left	0.25	0.21	0.22	-0.23	-0.24	-0.19	0.31	-0.01	-0.29	-0.31	-0.31	0.46			
LLF - right	0.25	0.23	0.22	-0.27	-0.27	-0.18	0.29	-0.06	-0.28	-0.25	-0.27	0.47	0.89		
Modified Schober	-0.19	-0.17	0.07	-0.20	0.26	0.20	-0.47	0.01	0.26	0.50	0.55	-0.50	-0.58	-0.52	
TWD	0.17	0.12	-0.03	-0.17	-0.17	0.08	0.48	-0.02	-0.28	-0.53	-0.49	0.33	0.42	0.33	-0.68

Table 6.8 Correlation between patient-based and anthropometric measures of outcome. Clinic survey (n= 105).

Key: Csp rot - cervical spine rotation; FFD - fingertip to floor distance; LLF - lateral lumbar flexion; Modified Schober - Modified Schober Index; TWD - tragus to wall distance.

Trigger list	Frequency	%	Additional areas cited	Frequency	%
Impact on / unable to Work	134	39.5	Difficulty sitting	24	7.1
Disturbed Sleep	101	29.8	Difficulty lying down	12	3.5
Worry about the future	89	26.2	Gardening	11	3.2
Sporting activities / exercise	76	22.4	Difficulty standing / standing for long periods	9	2.6
Feeling tired	69	20.3	Ability to complete tasks / do simple tasks	9	2.6
Difficulty with housework / DIY / lifting	58	17.1	Named body part (other than back/knees or hands)	9	2.6
Walking	57	16.8	Travelling / travel distances	6	1.8
Ability to remain physically active / mobility in general	55	16.2	'Back'	6	1.8
Poor self body image / posture / embarrassment / self-conscious	50	14.7	Impact of medication / side effects / efficacy	5	1.5
Pain / discomfort	50	14.7	Shopping	5	1.5
Feelings of depression	50	14.7	Morning stiffness / stiffness	5	1.5
Fatigue / loss of energy / lethargy / stamina	48	14.1	Vision / iritis	5	1.5
Social life / holidays / relationship with friends	48	14.1	'Normal activities'	5	1.5
Relationship with wife / husband / partner	45	13.3	General fitness (physical)	4	1.2
Driving / into and out of car	44	13.0	Ability to relax / relaxation	4	1.2
Limitations to joint / spinal movement-mobility specific	41	12.1	Frustration /anxiety	4	1.2
Feelings of low self-esteem / confidence	41	12.1	Financial impact	3	0.9
Ability to play with / look after children/ grandchildren	38	11.2	Difficulty with transfers - crouch to standing / out of chairs / out of bath	2	0.6
Pursuing chosen hobbies / past-times / leisure activities	32	9.4	Concern over weight gain	2	0.6
Sex life	31	9.1	Fear of being knocked or bumped/standing in crowds	2	0.6
Getting going in the morning	28	8.2	Quality of life	2	0.6
Family life / relationship with family and children	27	7.9	'Health'	2	0.6
Ability to plan ahead	25	7.4	Hands	2	0.6
Level of independence / dependency on others	25	7.4	Knees	2	0.6
Loss of motivation	20	5.9	Breathing	2	0.6
Feeling moody / miserable / irritable	15	4.4	Crossing the road	1	0.3
Difficulty sitting / standing / lying down	14	4.1	Reaching above head	1	0.3
Control over life / life in general / daily living	11	3.2	Getting out of bed / turning over in bed	1	0.3
Fear of falling	10	3.0	Drinking	1	0.3
Letting people down / meeting commitments	8	2.3	Inability to defend oneself / ones partner physically	1	0.3
Enjoyment of life	8	2.3	Sneezing	1	0.3
Dressing and bathing / personal hygiene	8	2.3	Concern over childbirth / future childbirth	1	0.3
Slow to do things	7	2.1	Impact on choice of footwear	1	0.3
Lack of spontaneous thought / Mental concentration	7	2.1			

Table 6.9 Frequency endorsement of areas mentioned in step 1 of PGI-AS. Baseline postal survey (n= 339).

Trigger list of PGI-AS	ASQoL - item content
Impact on / unable to work	
Disturbed sleep	Item 5: It's impossible to sleep
Worry about the future	
Sporting activities / exercise	
Feeling tired	Item 7: I am tired all the time Item 12: I get tired easily
Difficulty with housework / DIY / lifting	Item 4: I struggle to do jobs around the house Item 11: I am unable to do jobs around the house
Walking	
Ability to remain physically active / mobility in general	
Poor self body image/ posture/ embarrassment/ self-conscious	
Pain / discomfort	Item 9: I have unbearable pain Item 14: The pain is always there
Feelings of depression	Item 17: My condition gets me down
Fatigue / loss of energy / lethargy / stamina	Item 8: I have to keep stopping what I am doing to rest
Social life / holidays / relationship with friends	Item 15: I feel I miss out on a lot
Relationship with wife / husband / partner	
Driving / getting into and out of car	
Limitations to joint / spinal movement-mobility specific	
Feelings of low self-esteem / confidence	
Ability to play with / look after children/ grandchildren	
Pursuing chosen hobbies / past-times / leisure activities	
Sex life	
Getting going in the morning	Item 10: It takes a long time to get going in the morning
Family life / relationship with family and children	Item 6: I am unable to join in activities with my friends/ family
Ability to plan ahead	
Level of independence / dependency on others	
Loss of motivation	
Feeling moody / miserable / irritable	Item 2: I sometimes feel like crying
Difficulty sitting / standing / lying down	
Control over life / life in general / daily living	Item 1: My condition limits the places I can go
Fear of falling	
Letting people down / meeting commitments	Item 18: I worry about letting people down
Enjoyment of life	
Dressing and bathing / personal hygiene	Item 3: I have difficulty dressing Item 16: I find it difficult to wash my hair
Slow to do things	
Lack of spontaneous thought / mental concentration	
Additional item cited by patients:	
Frustration / anxiety	Item 13: I often get frustrated

Table 6.10 Comparison of items included in PGI-AS trigger list and ASQoL.

b) Disease specific

The results of the correlations between the disease-specific instruments, generic instruments and anthropometric measures are shown in tables 6.6 and 6.8.

All correlations between disease-specific instruments were in the predicted direction and the majority of correlations agreed with *a priori* hypotheses supporting the hypothesised convergent validity. The largest correlation was between the ASQoL and BASDAI (0.75). Moderate to large correlations were also found between the ASQoL and the RLDQ (0.68) and Body Chart (0.67 postal, 0.61 clinic).

A large correlation between the BASDAI and Body Chart (0.72 postal, 0.62 clinic), and a moderate to large correlation between the BASDAI and RLDQ (0.62) was as hypothesised. A moderate correlation between the Body Chart and RLDQ was found for the larger postal survey (-0.55) although a small to moderate correlation was found for the clinic survey (-0.44).

Correlation between the ASQoL and both generic instruments was slightly larger than hypothesised. A large correlation with the EQ-5D (-0.75) and moderate to large correlations with the SF-12 PCS (-0.70) and EuroQol thermometer (-0.68) were found. A moderate correlation with the SF-12 MCS (-0.53) was found as hypothesised.

As hypothesised, moderate to large correlations of the BASDAI, Body Chart and the RLDQ with the EQ-5D, EuroQol- thermometer and SF-12 PCS, and a small to moderate correlation with the SF-12 MCS were found (tables 6.6 and 6.8).

Small correlations of the ASQoL, BASDAI and Body Chart with the anthropometric measures of cervical rotation and FFD, and very small correlations with the remaining measures were found as hypothesised.

Also hypothesised was the moderate to large correlation between the RLDQ and cervical rotation and moderate correlations with FFD, the MSI and TWD, and a very small correlation with lateral lumbar flexion (table 6.8).

c) Generic instruments

The correlations between the generic instruments and anthropometric measures were in the predicted direction (tables 6.6 and 6.8).

The moderate to large correlation between the EQ-5D and thermometer (0.61) was larger than hypothesised. A moderate correlation between both sections of the EuroQol and the SF-12 PCS (EQ-5D 0.58, thermometer 0.52), and a small to moderate correlation with the SF-12 MCS (EQ-5D 0.49, thermometer 0.39) was found, as hypothesised, as was the very small correlation between the two sections of the SF-12 (0.20).

All correlations between the generic instruments and anthropometric measures were in the hypothesised direction (table 6.8). A small to moderate correlation of the EuroQol and SF-12 PCS with cervical rotation (EQ-5D 0.34 to 0.36; thermometer 0.33 to 0.46; SF-12 PCS 0.45 to 0.48), and a small correlation with FFD (EQ-5D -0.31; thermometer -0.36; SF-12 PCS -0.38) was found. Small correlations with the remaining anthoropometric measures and a small correlation between the SF-12 MCS and all anthropometric measures (< 0.19) were hypothesised.

d) Anthropometric measures

The correlations between the anthropometric measures were in the predicted direction (table 6.8).

As hypothesised, a small to moderate correlation between cervical rotation and TWD (range -0.53 to -0.49), and with FFD (range -0.31) was found. A larger than hypothesised relationship was found between FFD and lateral lumbar flexion (LLF) (range 0.46 to 0.47). Moderate correlations were found between the MSI and all anthropometric measures (range 0.50 to 0.55), as hypothesised, except for the large correlation with TWD (-0.68) which was larger than expected. Small levels of correlation were found between LLF and both cervical rotation and TWD, as hypothesised. The largest correlations were between right and left cervical rotation, and right and left LLF (0.89).

e) Sociodemographic Data

The results of tests of validity relating to sociodemographic variables are shown in table 6.11.

Compared to patients leaving school at the minimum leaving age their counterparts reporting a continuation of education rated their health as better for all instruments.

	Post-school education?			Degree or equivalent			Occupational Status			Housing tenure		
Instrument[a]	No (n= 97)	Yes (n= 125)	t-value	No (n= 147)	Yes (n= 75)	t-value	Employed (n= 134)	Not work - ill-health (n= 30)	t-value	Owner (n= 196)	Rented (n= 34)	t-value
ASQoL	9.79 (5.46)	7.26 (5.19)	-3.53	8.94 (5.34)	7.26 (5.48)	-2.18	6.58 (4.77)	13.66 (3.93)	-7.56	8.09 (5.32)	9.00 (5.35)	-0.90
BASDAI	4.88 (2.06)	3.91 (2.12)	-3.40	4.50 (2.02)	4.01 (2.34)	-1.59	3.74 (1.86)	6.39 (1.59)	-7.21	4.24 (2.17)	4.56 (1.71)	-0.92
Body Chart	2.67 (0.84)	2.25 (0.93)	-3.52	2.52 (0.86)	2.25 (0.99)	-2.11	2.22 (0.88)	3.03 (0.75)	-4.63	2.39 (0.92)	2.67 (0.71)	-1.91
EuroQol EQ-5D	0.49 (0.35)	0.60 (0.30)	2.50	0.52 (0.34)	0.61 (0.29)	2.04	0.65 (0.26)	0.20 (0.34)	6.79	0.56 (0.32)	0.53 (0.32)	0.43
EuroQol - thermometer	53.80 (18.9)	64.86 (19.9)	4.19	57.96 (18.93)	64.29 (22.14)	2.22	64.60 (18.98)	43.63 (18.91)	5.47	59.68 (20.47)	58.65 (20.33)	0.27
PGI-AS	4.06 (1.44)	4.28 (1.72)	0.90	4.07 (1.47)	4.41 (1.82)	1.48	4.37 (1.68)	3.39 (1.21)	3.01	4.18 (1.60)	4.50 (1.82)	-1.00
RLDQ	15.35 (9.65)	11.80 (9.68)	-2.65	13.88 (9.63)	12.58 (10.04)	-0.94	9.54 (7.78)	20.62 (8.18)	-6.97	12.92 (9.97)	15.70 (10.89)	-1.47
SF-12 MCS	46.03 (11.9)	46.50 (11.4)	0.30	45.76 (11.75)	47.27 (11.26)	0.92	47.43 (11.32)	39.56 (13.94)	3.29	46.20 (11.29)	46.58 (11.38)	-0.18
SF-12 PCS	33.53 (9.69)	38.25 (11.1)	3.32	35.20 (10.17)	38.20 (11.58)	1.48	39.60 (9.90)	26.37 (6.95)	6.93	36.43 (11.18)	35.57 (10.50)	0.41

Table 6.11 Mean (standard deviation) instrument scores according to education level, occupational status and housing tenure. Results from postal survey.

[a]ASQoL: scored 0-18, where lower scores indicate better health related quality of life.
BASDAI: scored 0-10, where higher scores indicate greater disease activity.
Body Chart: scored from 0 upwards, with no maximum score limit. Higher scores indicate greater levels of perceived body pain.
EuroQol EQ-5D: scored -0.59-1.0, where -0.59 is the worst and 1.0 the best possible health.
EuroQol - thermometer: scored 0-100, where higher scores indicate better health states.
PGI-AS: scored 0-10, where higher scores indicate better health related quality of life.
RLDQ: scored 0-48, where higher scores indicate increased limitation in functional ability.
SF-12 uses norm-based scoring from the general population. Scales are transformed: mean of 50 (sd=10), range 0-100.

All results were in the direction hypothesised and the majority were significant (7/9). The most significant difference was calculated for the EuroQol thermometer. There was no significant score difference for the PGI-AS or SF-12 MCS.

Compared to those without a degree or equivalent qualification patients with a degree or equivalent have better scores on all instruments. All results were in the direction hypothesised although few were significant (4/9). Significant differences were found for the ASQoL, Body Chart and the EuroQol; the EuroQol thermometer produced the most significant result.

Compared to those unable to work due to ill-health patients in work have significantly better levels of health for all disease-specific and generic instruments ($p < 0.01$); the ASQoL produced the most significant score difference. The EQ-5D and SF-12 PCS demonstrated similar levels of significance, but the results were less significant than for three of the disease-specific instruments, the ASQoL, BASDAI, RLDQ.

Compared to patients living in rented accommodation patients living in their own home reported better levels of health for the majority of instruments (7/9). The majority of results, apart from the SF-12 PCS, were in the direction hypothesised, but none were significant.

6.5 Discussion

Validity is the extent to which an instrument measures what it reports to measure, thus providing evidence in support of inferences and resulting scores (Fitzpatrick et al, 1998a). Most measures of patient outcome describe phenomena that are not directly observable and as such rely on establishing evidence in support of their construct validity. McHorney et al (1993) indicate that validity 'is not as simple as whether or not a health status scale is valid', rather, validity testing is a process of accumulating evidence that contributes to further understanding of inferences that may be made, and is essential for instruments that are new, modified or applied in a different context (Garratt, 1997; Bjorner et al, 1998).

The construct validity of the study instruments was assessed by relating scores to other disease-specific instruments, to other more established instruments and to sociodemographic variables. The study has demonstrated more extensive testing of

instruments and involved a larger population of patients with AS than previously reported which increases the generalisability of the results. No disease-specific instrument has previously been assessed against instruments with such well documented and established levels of validity as those for the EuroQol and SF-12. Previous attempts to provide evidence of construct validity for many study instruments has relied upon the correlation with other disease-specific or anthropometric instruments, often with little evidence to support the validity of these instruments. Further more, no study has constructed hypothetical relationships between instruments and set out to test these hypotheses. Rather, correlation was undertaken and a list of results provided as evidence of validity. None of the disease-specific study instruments have been previously compared against sociodemographic variables.

a) PGI-AS

The PGI-AS is a new instrument for the evaluation of disease-related quality of life in AS, adopting a modified version of the original PGI (Ruta et al, 1994a) (Chapter 3). This study has provided the first evidence for the validity of the PGI-AS and for the modified instrument.

The majority of PGI-AS trigger list items were selected more frequently than supplementary items introduced by patients and supports the content validity of the list. The open nature of the PGI-AS addresses individuals' own concerns rather than imposing pre-determined items that may have less relevance, and the addition of supplementary items to the trigger list highlights the diversity and individuality of disease-related quality of life as a concept.

The ASQoL includes many items frequently endorsed by patients completing the PGI-AS, but several issues considered important to the HRQL of patients with chronic disease, and included in the PGI-AS trigger list, are omitted by the ASQoL (Fitzpatrick, 1993a; Ware, 1998). In particular, the impact of AS on 'work' is not measured. This was the most frequently endorsed item in the PGI-AS and is an important issue in AS (Guillemin et al, 1990; Ward, 1998). Multiple items measuring tiredness, jobs around the home and pain are included in the ASQoL and distinguish between the severity or frequency of a symptom. These items relate to three of the most frequently endorsed items in the PGI-AS. The inclusion of two items to address

one issue may enhance the precision of information provided (Gandek et al, 1998a,b). Alternatively, item repetition may reduce instrument content (Streiner and Norman, 1995). The comparison of item content between the PGI-AS and ASQoL supports the hypothesis that the PGI-AS confers greater content validity as a measure of HRQL and justifies the individualised and open nature of the instrument.

Evidence for the validity of the PGI-AS was gained from comparisons with widely used disease-specific and generic instruments, with anthropometric measures traditionally used in clinical practice, and with sociodemographic variables. All correlations were in the hypothesised direction. Moderate correlations with the disease-specific and generic instruments were found, which were generally larger for the former. Correlation with anthropometric measures reflecting reversible change was small.

The more moderate levels of correlation with both disease-specific and generic instruments were hypothesised and may be a function of the alternative approach to measuring HRQL presented by the PGI-AS. In particular, due to the role of explicit weighting and the influence of items relating to 'other health' and 'non-health' issues on the score. A similar low to moderate correlation between the original PGI and the SF-36 was reported by Garratt (1997). However, the level of correlation between the PGI-AS and generic measures of HRQL found in this study were greater than those reported by other authors comparing the original version of the PGI to the SF-36 (Ruta et al, 1994a, 1999; Garratt, 1997). The improved correlation supports the improved validity of the modified version of the PGI in evaluating disease-related quality of life. Alternatively, the result could be due to AS having a greater impact on general health than the disorders assessed in earlier studies. That is, mennorhagia, varicose vein, peptic ulcer and low back pain.

The correlation of the PGI-AS with the SF-12 was larger than that found between the uni-dimensional disease-specific instruments and the SF-12, and further supports the hypothesis that the PGI-AS measures a broader domain of HRQL. However, to further explore the ability of the PGI-AS to provide a multidimensional and comprehensive assessment of disease-related quality of life, the relationship with the eight dimension profile of HRQL derived from the SF-36 is suggested. Although the disease-specific and generic instruments included in the evaluation measure a wide

range of domains the SF-36 profile gives a broader picture of disability, role and social functioning.

Finally, the PGI-AS behaved as hypothesised for the majority of socio-demographic variables, but the differences in scores were small and statistical significance was only found when discriminating between patients unable to work due to ill-health and those able to work.

The results represent good evidence for the validity of the PGI-AS and suggests that it is sensitive to the effects of AS on aspects of HRQL measured by all disease-specific and generic instruments and by certain anthropometric measures and sociodemographic variables.

b) Disease specific

Evidence for the construct validity of the four additional disease-specific instruments, the ASQoL, BASDAI, Body Chart and RLDQ, was gained from comparisons between instruments, with generic instruments and anthropometric measures and with sociodemographic variables.

The majority of correlations between the disease-specific instruments were of a moderate to large size and were all in the hypothesised direction. The only correlation that was slightly less than hypothesised was between the Body Chart and the RLDQ in the clinic survey (0.44), but for the larger postal survey this correlation was moderate (0.55). The largest correlations were between the ASQoL and all other disease-specific instruments, and the largest was with the BASDAI.

The disease-specific instruments were all compared to the EuroQol and SF-12. All correlations were in the hypothesised direction and all disease-specific instruments were found to have moderate to large levels of correlation with the EuroQol and the SF-12 PCS, and a small to moderate correlation with the SF-12 MCS. The largest correlations were between the ASQoL and both generic instruments, and these correlations were larger than the correlations found between both generic instruments.

All correlations with anthropometric measures were in the hypothesised direction. Small to moderate levels of correlation between anthropometric measures reflecting

reversible change (cervical rotation and FFD) and most disease-specific instruments were found. A large correlation with cervical mobility and a moderate correlation with other measures, except for LLF, was found for the RLDQ. These results suggest that the RLDQ is more sensitive to the effects of AS on aspects of spinal mobility represented by anthropometric measures included in the study, than the other disease-specific instruments.

Finally, all disease-specific instruments behaved as hypothesised for the majority of socio-demographic variables. In particular the ASQoL was found to discriminate well for all sociodemographic variables, and results had greater statistical significance than found for both generic instruments.

The results represent good evidence for the validity of the ASQoL as a disease-specific measure of HRQL and suggest that it is very sensitive to the effects of AS on aspects of HRQL measured by all disease-specific and generic instruments and certain anthropometric measures.

The results also provide good evidence for the validity of the BASDAI, Body Chart and RLDQ as AS-specific measures of disease activity, bodily pain and functional disability respectively. The results suggest that the instruments are sensitive to the effects of AS on aspects of HRQL measured by all disease-specific and generic instruments and certain anthropometric measures.

c) Generic

Evidence for the validity of the EuroQol and SF-12 was gained from comparison between both instruments, with anthropometric measures and against sociodemographic variables. All correlations were in the hypothesised direction and a moderate to large correlation between both instruments and a small correlation with anthropometric measures measuring reversible change was found. A larger correlation than hypothesised was found between the EQ-5D and the thermometer (0.61) suggesting that they are both addressing similar aspects of HRQL. As hypothesised, a weak correlation between the SF-12 PCS and MCS was found suggesting that the two components measure very different aspects of HRQL in patients with AS.

The EuroQol thermometer was the generic instrument most capable of discriminating between patients on most sociodemographic variables (3/4), but the SF-12 PCS had the most significant difference in discriminating between patients able to work and those unable to work due to ill-health.

Overall, results from the tests converged with study hypotheses and represent good evidence for the validity of the EuroQol and SF-12 as generic measures of HRQL in AS.

d) Anthropometric measures

Evidence for the validity of the anthropometric measures was gained from comparison between all measures. All correlations were in the hypothesised direction. A moderate to large correlation between the MSI and all other anthropometric measures was found suggesting that the MSI is sensitive to aspects of mobility assessed by cervical rotation, FFD, LLF and TWD. As hypothesised, small to moderate correlations were also found between cervical rotation and both FFD and TWD. Further correlations between measures were very small.

Conclusion

The study has investigated a broad pattern of relationships between disease-specific and generic instruments and sociodemographic variables in a population of AS patients. Evidence of validity to further support inferences that may be made about the study instruments has been demonstrated. Evidence to support the application of the PGI-AS as a new individualised measure of AS-related quality of life has been provided. The best performing disease-specific instruments were the ASQoL and the PGI-AS, and the best anthropometric measures were the Modified Schober Index and cervical rotation. Additionally, the use of the EuroQol and SF-12, two well-developed and tested generic measures of HRQL has not previously been reported in patients with AS. This study provides evidence to support their validity in this patient population.

Chapter 7 Responsiveness

7.1 Introduction

This chapter presents the tests of responsiveness for the study instruments. The evaluation of responsiveness in the field of outcome measurement is discussed in section 7.2. Section 7.3 discusses the methods adopted in the current study and the results of the responsiveness analysis are presented in section 7.4. The chapter closes with a discussion.

7.2 Responsiveness and measures of health

Responsiveness refers to the ability of an instrument to detect clinically important change over time, when change is present, and together with evidence of reliability and validity is an essential requirement of an evaluative instrument (Kirshner and Guyatt, 1985). However, despite this important role, it is a relatively neglected measurement property (Fitzpatrick et al, 1993c; Liang, 1995). This neglect was identified in Chapter 2, in the assessment of AS-specific measures of outcome.

Interpretation of change in instrument score, and the ability to identify relevant or clinically important change is important in evaluation and may inform patient management and clinical decision making (Fortin et al, 1995; Redelmeier et al, 1996). However, the definition of change is problematic and no gold standard has been recommended. Real change in the underlying state is the 'signal' that assessment of responsiveness aims to detect. The influence of an error variance that cannot be attributed to real change is referred to as 'noise' and represents the influence of random and systematic error. This may be due to non-specific influences such as natural variation in the underlying state, variation between patients, or variation from the effects of an intervention (Liang, 1995). Evaluation of responsiveness seeks to detect change that occurs above and beyond that due to random and systematic error (Deyo et al, 1991).

Variability in the level of responsiveness calculated for different instruments in the same group of patients with rheumatological disorders of a similar nature to AS has been reported (Kazis et al, 1989; Fitzpatrick et al, 1993a). This illustrates the concern that the use of multiple instruments may result in misleading, inconsistent and contradictory results (Ziebland, 1994). However, the use of a single instrument could be equally misleading. Knowledge of instrument responsiveness should aid selection

and assist prioritisation or reduction in the number selected for evaluative purposes. It may also assist in sample size estimation to enhance the statistical power of clinical trials (Deyo et al, 1991; Stucki et al, 1995).

Responsiveness should be appraised in relation to the patient population and setting in which the evaluation was performed (Fitzpatrick et al, 1998a,b), and as for other measurement properties, should be viewed as a continual process of establishing evidence to support instrument application in different evaluative contexts. The conceptual base and item content will influence levels of responsiveness found in different circumstances. For example, an intervention to reduce disease activity is likely to result in greater responsiveness for a disease-specific measure of disease activity than would be found for a generic measure of HRQL, particularly if the intervention had little impact on HRQL. This highlights the importance of selecting instruments appropriate to the aims of the intervention and setting prior hypotheses of the expected behaviour of instruments under consideration (Garratt, 1997).

7.2.1 Methods of responsiveness testing

There is no clear consensus on the preferred method for assessing responsiveness and several methods have been proposed including responsiveness statistics, paired t-tests and correlation with other change scores (Deyo et al, 1991; Fitzpatrick et al,1998a). Garratt (1997) describes a two step approach: the first step involves the specification of external criteria by which change in health may be judged and the second step, the quantification of responsiveness.

The choice of external criteria by which change in health or HRQL is estimated is important and will affect instrument relative responsiveness (Deyo et al, 1991). Health transition questions, clinical variables or measures of disease process have been used to describe change. Patient or physician reported health transition questions which describe the magnitude and direction of change in health over a given time period provide a valid approach to measuring change and have been widely used as external criteria in the evaluation of instrument responsiveness (Fitzpatrick et al, 1993b; Keller et al, 1999a,b). To the extent that a patient-based instrument is a valid measure of health and is capable of measuring change, a strong association with a patient reported health transition item would be expected (Garratt et al, 1996a), but will be dependent on the specific instrument domain and the form of transition

question. For example, a disease-specific transition question would be expected to have a stronger relationship with a disease-specific instrument than a generic instrument.

Correlation of change scores between instruments and other well-established clinical variables or selected instruments provides evidence of whether an instrument demonstrates change in score over time that is consistent with change in other variables (Deyo et al, 1991; Fitzpatrick et al, 1998a; Kosinski et al, 1999b). This relationship has been described as longitudinal validity and requires that evaluative instruments demonstrate longitudinal within-patient score change that bears the expected relation to changes in external criteria (Kirshner and Guyatt, 1985). The measure of responsiveness is the relative strength of correlation between the change in instrument score and external criteria, and a correlation with other change scores of more than 0.40 has been described as a 'substantial' relationship (Stucki et al, 1995; Keller et al, 1999). It has been suggested that instruments demonstrating strong cross-sectional validity, should also be valid for measuring within-person change over time (Katz et al, 1992; Ware, 1997). However, the requirements for these two measurement properties are different and both should be assessed for evaluative instruments (Kirshner and Guyatt, 1985; Deyo et al, 1991).

The responsiveness of the unweighted EuroQol EQ-5D in patients with Rheumatoid Arthritis (RA) was assessed by relating mean score change to categories of a health transition question ('Compared to three months ago is your arthritis better, the same or worse?') (Hurst et al, 1997). Significant association between each category, except for the anxiety/depression item, supported the responsiveness of the EQ-5D profile in this population.

An alternative external criterion theorises that a treatment of known efficacy will result in a score improvement in instruments measuring domains targeted by the intervention. The study would aim to identify score change beyond that due to non-specific variance in patients otherwise described as stable. The responsiveness of the Leeds Disability Questionnaire (LDQ) was assessed in patients randomly assigned to receive one of three different physiotherapy treatment regimes with known efficacy (Abbott et al, 1994). Using a paired t-test, significant improvement in score for all patients supported instrument responsiveness.

Following the application of criteria to describe change, responsiveness may be quantified to enable judgement about instrument relative responsiveness. Three effect size statistics have been described (Fitzpatrick et al, 1998a)(table 7.1).

Responsiveness statistic	Method
Effect size (Kazis et al, 1989)	Change in instrument score divided by standard deviation of baseline score
Standardised response mean (SRM) (Liang et al, 1990)	Change in instrument score divided by standard deviation of change score
Modified standardised response mean (MSRM) (Guyatt et al, 1987)	Change in instrument score divided by standard deviation of change score for stable patients

Table 7.1 Methods for calculating a responsiveness statistic (after Fitzpatrick et al, 1998a).

All calculations use the same numerator, but differ in the denominator adopted. The Effect size (ES) statistic divides the change in score by the standard deviation of baseline scores, therefore using the variation in baseline scores as a reference against which to assess change (Liang, 1995). The standardised response mean (SRM) uses the standard deviation of the change score to incorporate the response variance in change scores and is therefore more appropriate for comparisons of instrument responsiveness (Fitzpatrick et al, 1998a). However, both the ES and SRM may be influenced by natural variance in the underlying state and by measurement error (Liang, 1995; Hurst et al, 1997). The modified standardised response mean (MSRM) addresses the inherent natural variance that may occur in patients who otherwise report their health as unchanged and non-specific score change by using the standard deviation of change in patients who are defined as stable, and is therefore considered the preferred responsiveness statistic (Deyo et al, 1991). In demonstrating responsiveness to clinically important change, instruments should detect change above the non-specific change incorporated in the MSRM (Deyo et al, 1991).

Each effect size, or responsiveness statistic provides a quantitative and standardised unit of expression of the size and meaning of change to support instrument comparison (Fitzpatrick et al, 1998a). Guidance for data interpretation has been proposed: a score of more than 0.8 represents a large level of responsiveness, a score of 0.5 moderate, and a score of 0.2 a small level (Cohen, 1977; Fitzpatrick et al,

1998a). For example, a MSRM of more than 1.00 would indicate a change in score that is greater than the standard deviation of score change in stable patients.

Most evidence suggests that disease-specific instruments are more responsive than generic instruments (Guyatt et al, 1993), and evidence of comparative responsiveness is important when selecting evaluative instruments. Hurst et al (1997;1998) and Ruta et al (1998) assessed the measurement properties of three generic instruments, the EuroQol, SF-36 and SF-12 and selected disease-specific instruments in patients with RA (table 7.2).

Author	Instrument	RA population	Methodology	Result
Hurst et al (1997)	EuroQol EQ-5D	56 out-patients	Self-reported improvement in RA (3months) Mean change SRM MSRM	 +0.22 0.70 1.00
Hurst et al (1997)	EuroQol thermometer	56 out-patients	Self-reported improvement in RA (3months) Mean change SRM MSRM	 +12.2 0.71 1.00
Hurst et al (1998)	SF-36	42 out-patients	Self-reported improvement in RA (3months) Mean change SRM	 PCS +4.30 MCS +3.50 PCS 0.61 MCS 0.35
Ruta et al (1998)	SF-36	233 out-patients	Self-reported improvement in RA (3months) Mean change, SRM	Mean change SRM PF +8.2 0.43 SF +12.9 0.49 RL-P +15.0 0.36 RL-E +10.1 0.27 MH +4.7 0.33 BP +18.9 0.90 Vit +10.6 0.50 GH +5.8 0.36 PCS +0.43 0.61 MCS +0.35 0.35
Hurst et al (1998)	SF-12	42 out-patients	Self-reported improvement in RA (3months) Mean change SRM	 PCS +4.10 MCS +2.60 PCS 0.52 MCS 0.31

Table 7.2 Studies assessing the responsiveness of the EuroQol, SF-36 and SF-12 in patients with Rheumatoid Arthritis.

Key: PF- Physical functioning; SF- Social functioning; RL-P- Role limitations physical; RL-M- RL mental; MH- Mental health; BP-: Bodily pain; Vit- Energy and fatigue; GH- General health;, PCS- Physical component summary scale; MCS - Mental component summary scale.

A statistically significant improvement in EuroQol scores associated with patient-reported improvement in RA, but a non-statistically significant reduction in EQ-5D scores associated with deterioration in RA was found (Hurst et al, 1997). However, both sections of the EuroQol were more responsive than the Modified Health Assessment Questionnaire (MHAQ), a disease-specific measure of functional ability (Kirwan and Reeback, 1983). The authors conclude that the EuroQol is very

responsive to patient reported change in RA and is capable of reflecting clinically important change.

The responsiveness of the SF-36 and the SF-12 was equivalent in the same patient population reporting improvement in RA (Hurst et al, 1998), and was greater than that found for the MHAQ. The size of change in both the SF-36 and disease-specific instrument was greater in patients reporting an improvement than in those reporting a deterioration, and the investigators suggest that this may reflect a larger 'biological change' in patients who perceive their arthritis as better, as opposed to being a reflection of instrument sensitivity (Ruta et al, 1998).

For the AS-specific patient-based study instruments for which evidence of responsiveness is available (table 7.3), the study methodology adopted is lacking and restricted to the impact of physical therapy on change.

Author	Instrument	AS population	Methodology	Result
Garrett et al (1994)	BASDAI	47 in-patients	Known efficacy: 3-week intensive physiotherapy. Comparison of mean scores day '0' and day '18'.	BASDAI: Mean change -0.85 (p= 0.009) 16.4% score improvement Bath DAI: Mean change -1.22 (p= 0.002) 22.8% score improvement
Dziedzic (1997)	Body Chart	41 out-patients	Usual care. Assessed 3, 6, 9-months. Area under the curve - relationship between variables (Spearmans)	Statistically significant relationship between Body Chart and all variables: pain, night pain, stiffness, enthesitis indices, disability questionnaire
Abbott et al (1994)	LDQ	42 out-patients	Known efficacy: RCT - 3 physiotherapy regimes . Assessments pre-randomisation, post-treatment (6-weeks), 6-months.	Significant improvement in score for all patients at 6-weeks (paired t-test: t= 2.79, p< 0.01). Interpretation of 6-month results difficult - poor response.

Table 7.3 Evidence of the responsiveness of the patient-based study instruments.

These findings may partly reflect the methodological difficulties in assessing responsiveness (Fitzpatrick et al, 1993c). There is no published evidence in support of the responsiveness of the ASQoL, RLDQ and PGI-AS.

7.3 Methods for assessing the responsiveness of the study instruments

Responsiveness was assessed using data collected from baseline and six-month administration of all instruments. The external criteria by which responsiveness was judged was by self-reported health transition at six-months.

7.3.1 Assessing responsiveness using self-reported health transition

Two separate health transition questions were included in the patient-completed questionnaire, relating to AS-specific or general health ('Compared to six-months ago, how would you rate your AS / general health now - much better, somewhat better, about the same, somewhat worse, much worse ?'). To maximise sample size responsiveness was assessed for patients indicating that they were 'better' or 'worse'.

For the purpose of assessing longitudinal validity instrument scores were compared to self-reported health transition. The level of concordance between change in instrument score and patient response to both transition items was calculated and assessed for a liner trend (Garratt, 1997).

In addition, the responsiveness statistic, which is equal to the mean change in scores divided by the standard deviation of the score differences in stable patients (MSRM)(Fitzpatrick et al, 1998a; Garratt et al, 2000), was calculated for patients reporting an improvement or deterioration on specific transition questions.

PGI-AS

Two of the three follow-up formats of the PGI-AS were completed at six-months for both clinic and postal surveys: blind, and informed and open (table 7.4). These approaches are detailed in Chapter 3.

Follow-up completion	Step 1: Identifying areas	Step 2: Scoring each area	Step 3: Spending points
Blind	Blind to areas identified at baseline	Blind to baseline score	Blind to baseline points
Informed and open	Informed of areas identified at baseline. Allowed to change or retain list as necessary	Blind to baseline score	Blind to baseline points

Table 7.4 PGI-AS completion formats at six months.

The variation in format addresses the provision or absence of baseline areas in step 1. The lead investigator (KLH) was responsible for entering data into all follow-up questionnaires and the choice of format was made due to the limited study resources. The 'blind' format did not require the inclusion of baseline areas. The 'informed and open' format was selected in preference to the 'closed' format due to the individualised and open nature reflected at all stages of instrument completion. This selection was

made prior to any data analysis. Patients were randomly assigned to receive either format at six-months.

7.4 Results of responsiveness testing

The results of the tests of longitudinal validity which compared change in instrument scores with self-reported change in AS and general health transition are shown in tables 7.5 to 7.10.

The change scores of the ASQoL, BASDAI, RLDQ, EuroQol and SF-12 have been based on the combined results of postal and clinic surveys to maximise sample size. Of the 254 patients who attempted the AS-specific transition and completed all instruments at baseline and six months, 56 (22.0%) perceived their AS-specific health as better, 125 (49.2%) stated that it was the same, and 73 (28.7%) indicated that their AS-specific health was worse than six months earlier (table 7.5).

	AS health transition			
Instrument	Better (n= 56)	About the same (n= 125)	Worse (n= 73)	F-test for linearity
ASQoL	-1.00 (3.46)	-0.03 (3.33)	1.21 (3.27)	7.16 **
BASDAI	-1.00 (1.44)	0.08 (1.63)	0.37 (2.11)	10.78 **
RLDQ	-2.01 (4.84)	-0.12 (6.29)	1.87 (6.06)	6.86 **
EuroQol EQ-5D	0.13 (0.25)	-0.005 (0.25)	-0.12 (0.34)	9.55 **
EuroQol - thermometer	23.44 (91.22)	2.13 (17.83)	-4.56 (22.2)	12.46 **
SF-12 MCS	3.72 (10.18)	-0.18 (9.21)	-1.45 (10.52)	4.53 *
SF-12 PCS	3.90 (7.40)	0.34 (7.70)	-1.92 (6.85)	9.82 **

Table 7.5 Mean change (standard deviation) in instrument scores by 6-month AS health transition. Combined postal and clinic data (n= 254).

* significant at $p< 0.05$; ** significant at $p< 0.01$
ASQoL: scored 0-18, where lower scores indicate better HRQL.
BASDAI: scored 0-10, where higher scores indicate greater disease activity.
RLDQ: scored 0-48, where higher scores indicate increased functional disability.
EuroQol EQ-5D: scored -0.59-1.0, where -0.59 is the worst and 1.0 the best possible HRQL.
EuroQol thermometer: scored 0-100, where higher scores indicate better health states.
SF-12: norm based scoring: mean of 50 (sd=10), range 0-100. MCS-mental component scale; PCS- physical component scale.

Of the 248 patients completing the general health transition, 51 (20.6%) perceived their general health as better, 135 (54.4%) indicated that it was the same, and 62 (25.0%) perceived their general health as worse than six months earlier (table 7.6).

	General health transition			
Instrument *	Better (n= 51)	About the same (n= 135)	Worse (n= 62)	F-test for linearity **
ASQoL	-1.57 (2.53)	0.46 (2.88)	0.74 (4.21)	8.14
BASDAI	-1.13 (1.41)	-0.0001 (1.66)	0.72 (2.08)	16.10
RLDQ	-2.30 (4.93)	-1.34 (6.20)	2.12 (6.15)	7.77
EuroQol EQ-5D	0.16 (0.27)	-0.03 (0.26)	-0.11 (0.35)	13.31
EuroQol - thermometer	27.00 (95.07)	-2.08 (16.33)	-7.06 (23.57)	9.18
SF-12 MCS	6.06 (10.41)	-0.74 (9.11)	-1.39 (10.35)	10.65
SF-12 PCS	3.84 (7.82)	0.04 (6.71)	-1.34 (8.64)	7.25

Table 7.6 Mean changes (standard deviation) in instrument scores by 6-month general health transition. Combined postal and clinic data (n= 248).

** all significant at $p < 0.01$. * instrument scoring summarised in table 7.5

Analysis of the Body Chart and the PGI-AS was based on the results of the larger postal survey only (tables 7.7 and 7.8). Completion during the clinic survey followed an interview-administered format and could not be combined with the postal results. Also, once replies to the transition questions were taken into account the sample sizes for the clinic survey were too small to draw inferences from the data.

	AS health transition			
Instrument	Better (n= 36)	About the same (n= 85)	Worse (n= 44)	F-test for linearity
Body Chart (log)	-0.46 (0.85)	0.13 (0.86)	0.44 (0.71)	11.98 **
PGI-AS (combined)	0.51 (1.80)	0.10 (1.58)	-0.30 (1.43)	2.57 (p= 0.08)
- blind	0.52 (1.62) (n= 16)	0.32 (1.58) (n= 48)	-0.10 (1.49) (n= 23)	0.35
- informed/open	0.75 (1.63) (n= 20)	-0.17 (1.56) (n= 37)	-0.79 (1.24) (n= 21)	5.19 (p= 0.08)

Table 7.7 Mean changes (standard deviation) in instrument scores by 6-month AS health transition. Postal data (n= 165).

** significant at $p < 0.01$
Body Chart: scored from 0, with no maximum score limit. Higher scores indicate greater levels of perceived body pain.
PGI-AS: scored 0-10, where higher scores indicate better disease-related quality of life. Combined - combined results for both formats; Blind - not informed of baseline areas; Informed and open - informed of baseline areas and allowed to change.

Of the 165 patients who attempted the AS-specific transition and completed all study instruments at baseline and six months, 36 (21.8%) perceived their AS-specific health as better, 85 (51.5%) stated that it was the same, and 44(26.6%) indicated that their AS-specific health was worse than six months earlier (table 7.7). Of the 162 patients completing the general health transition, 29 (17.9%) perceived their general health as

better, 95 (58.6%) indicated that it was the same, and 38 (23.4%) perceived their general health as worse than six months earlier (table 7.8).

	General health transition			
Instrument*	Better (n= 29)	About the same (n= 95)	Worse (n= 38)	F-test for linearity
Body Chart (log)	-0.33 (1.18)	0.10 (0.84)	0.32 (0.58)	4.71 **
PGI-AS (combined)	-0.86 (1.93)	0.13 (1.49)	-0.61 (1.38)	7.44 **
- blind	0.75 (1.47) (n= 14)	0.39 (1.56) (n= 52)	-0.25 (1.56) (n= 20)	1.80
- informed/open	1.34 (1.87) (n= 15)	-0.18 (1.34) (n= 43)	-1.00 (1.13) (n= 18)	11.27 **

Table 7.8 Mean changes (standard deviation) in instrument scores by 6-month general health transition. Postal data (n= 162).

** significant at $p < 0.01$. *instrument scoring summarised in table 7.7

The change scores for both formats of the PGI-AS reflect the categories of specific and generic health transition and is evidence for the longitudinal validity of the instrument (tables 7.7 and 7.8): those patients who indicate that their AS or general health is better over the six months have an average improvement in PGI-AS score; and those where AS or general health is worse have an average deterioration in score. The largest levels of change were found for the informed and open format of the PGI-AS on the general health transition: patients who say that their general health is better have an average improvement in their PGI-AS score of 1.34 (on a scale 0-10, where 10 is the best disease-related quality of life); those whose general health is about the same have an average deterioration of -0.18; whilst those whose general health is worse have an average deterioration of -1.00. A strong relationship with AS health transition was also found for the informed and open format (table 7.7), although this did not quite reach statistical significance. Although reflecting the same trend, smaller levels of change and a non-significant relationship between the blind format of the PGI-AS and both transition questions was found. When both PGI-AS formats were considered together a reduced association with both transition questions, than seen for the informed and open format alone, was found.

The informed and open format of the PGI-AS produced the largest F-statistic for linearity when compared to the blind format or combined PGI-AS results, and this format has the strongest relationship with both health transition formats. The strength of the relationship between the informed and open format of the PGI-AS and general

health transition (11.27) was comparable to the large levels of change and strength of the relationship between the Body Chart and AS transition (11.98)(table 7.7). These two instruments were assessed in a smaller data set and the results, although significant, are not directly comparable to the larger combined data set (tables 7.5 and 7.6).

The change scores for all disease-specific patient-based measures of outcome also reflect the categories of specific and generic health transition and is evidence for the longitudinal validity of the instruments (tables 7.5 to 7.8). The largest levels of change were found for the BASDAI on both AS and general health transition: patients who say that their AS is better have an average improvement in score of -1.00 (on a scale of 0-10, where 0 is the lowest level of disease activity), and where general health is better an average score improvement of -1.13 is found; those whose AS is worse have an average deterioration in score of 0.37, and where general health is worse an average score deterioration of 0.72.

The BASDAI also produced the largest F-statistic for linearity for both AS (10.78) and general health transition (16.10), demonstrating the strongest relationship with AS health transition when compared to other disease-specific instruments from the same data set, and the strongest relationship with general health transition when compared to disease-specific and generic instruments.

Although significant, the smallest levels of change scores for disease-specific instruments were found for the ASQoL and RLDQ on both transition questions, the smallest levels of change seen in the RLDQ. However, F-statistics for linearity were greater for both instruments than those found for the SF-12 MCS on AS transition and the SF-12 PCS on general transition, suggesting a stronger relationship with health transition.

The change scores for all generic instruments reflect the categories of specific and generic health transition and is evidence for the longitudinal validity of the instruments (tables 7.5 to 7.6). The EuroQol and the SF-12 MCS have larger levels of change and a stronger linear relationship with responses to the generic transition than with the AS transition question: for example, patients who say that their general health is much better over the six months have an average improvement in EuroQol

thermometer score of 27.00 (on a scale 0-100, where 100 is the best possible health state), and an improvement of 23.44 when AS health is better; those whose general health is the same have an average reduction of -2.08 (general health) or and average increase of 2.13 (AS); and those whose general health is worse have an average deterioration of -7.06, and an average deterioration of -4.56 when AS is worse. However, the SF-12 PCS has similar levels of change and linear relationship with responses to both transition questions. The EuroQol thermometer produces the largest F-statistic for the relationship with AS health transition (12.46) than all other disease-specific and generic instruments, supporting the strongest relationship with AS health transition. Although significant, the SF-12 PCS has a weaker linear relationship with general health transition than all other instruments. The SF-12 MCS has a weak association with AS transition.

Anthropometric measures were completed during the clinic survey only. The mean changes for anthropometric measures at six months in self-reported AS-specific and general health transition are shown in tables 7.9 and 7.10.

	AS health transition			
Instrument	Better (n= 7)	About the same (n= 32)	Worse (n= 15)	F-test for linearity
Cervical rotation - left	1.48 (2.17)	0.39 (1.65)	0.05 (2.22)	1.39
Cervical rotation - right	0.45 (2.11)	0.06 (1.45)	-13.57 (54.98)	1.22
FFD	-5.87 (11.26)	2.52 (6.85)	1.22 (6.20)	3.74 *
LLF - left	-0.31 (2.09)	1.87 (3.98)	1. 08 (4.84)	0.86
LLF - right	-0.58 (3.16)	1.00 (3.28)	1.10 (4.11)	0.64
MSI	0.43 (0.66)	-0.04 (0.75)	0.08 (0.85)	1.11
TWD	-0.63 (1.35)	-0.33 (1.47)	0.25 (1.52)	1.13

Table 7.9 Mean changes (standard deviation) in instrument scores by 6-month AS health transition. Clinic data (n= 54).

* significant at $p < 0.05$

Cervical rotation - distance between tip of nose and acromioclavicular joint measured in neutral and maximum ipsilateral rotation. Difference between two positions calculated, where a smaller difference indicates a more restricted range.

FFD - Fingertip to floor distance - distance between tip of right middle finger and floor following maximum trunk flexion, where the smaller distance indicates greater movement. 0 is the maximum possible range (touching floor).

LLF - Lateral lumbar flexion - distance between tip of middle finger and floor measured following maximum ipsilateral lateral flexion, where the smaller distance indicates greater movement.

MSI - Modified Schober Index (15cm) (Macrae and Wright, 1969) - distance between two marks placed 15cm apart in standing (10cm proximal and 5cm distal to posterior superior iliac spine). Distance after maximum trunk flexion, where a larger difference indicates greater lumbar movement.

TWD - Tragus to wall distance - Distance between right tragus and wall measured in standing, where a larger distance indicates a worse spinal / upper cervical posture.

Of the 54 patients who attempted the AS-specific transition and completed all study instruments at baseline and six months during the clinic survey, 7 (12.9%) perceived their AS-specific health as better, 34 (62.9%) stated that it was the same, and 13 (24.2%) indicated that their AS-specific health was worse than six months earlier (table 6.9). Of the 54 patients completing the general health transition, 7 (12.9%) perceived their general health as better, 37 (68.5%) indicated that it was the same, and 10 (18.6%) perceived their general health was worse than six months earlier (table 7.10).

	General health transition			
Instrument[a]	Better (n= 7)	About the same (n= 32)	Worse (n= 15)	F-test for linearity
Cervical rotation - left	0.94 (2.32)	0.27 (2.00)	0.74 (1.47)	0.45
Cervical rotation - right	0.72 (2.43)	-0.06 (2.00)	-16.62 (61.56)	0.38
FFD	-8.22 (12.89)	2.49 (6.77)	0.57 (5.51)	4.90 *
LLF - left	0.04 (2.45)	1.70 (3.81)	1. 80 (4.06)	0.38
LLF - right	-1.04 (3.73)	0.87 (3.13)	1.42 (4.43)	0.89
MSI	0.42 (0.66)	0.08 (0.81)	-0.22 (0.61)	1.40
TWD	-0.82 (1.55)	-0.25 (1.38)	0.19 (1.74)	0.88

Table 7.10 Mean changes (standard deviation) in instrument scores by 6-month general health transition. Clinic data (n= 54).

* all significant at $p < 0.05$
[a] instrument scoring summarised in table 7.9

The change scores for four of the five anthropometric measures do not clearly reflect the categories of the specific or generic health transition questions (tables 7.9 and 7.10). The only measure to produce a significant relationship with both transition questions is fingertip to floor distance following anterior trunk flexion (FFD): patients who say that they are better at six months have an average improvement in FFD of -5.87cm (AS) or -8.22cm (general health)(on a scale where 0 is the best possible score; that is, fingers touch the floor); those whose AS health is worse have an average deterioration in score of 1.22cm, and where general health is worse an average score deterioration of 0.57cm. The F-statistic for linearity produced a significant result on both transition questions (AS health 3.74; general health 4.90). However, this was less than that found for most patient-based measures of outcome.

The tests of responsiveness are shown in tables 7.11 to 7.13.

For patients reporting an improvement in AS-specific health the largest MSRM was found for the EuroQol thermometer (1.30), followed by the BASDAI (-0.60). Moderate MSRMs were found for the Body Chart (-0.53), EQ-5D (0.50), SF-12 (PCS 0.51, MCS 0.41) and the PGI-AS (informed and open)(0.42). Small MSRMs were found for the PGI-AS (blind)(0.33), the RLDQ(-0.32) and the ASQoL (-0.29). However, sample sizes for the Body Chart and PGI-AS, and particularly when the two versions of the PGI-AS were assessed, were smaller than the sample size for the analyses conducted for the other instruments and direct comparison of results is difficult.

For patients reporting a deterioration in AS-specific health the largest MSRM was found for the Body Chart (0.51), followed by the PGI-AS (informed and open)(-0.48). A moderate MSRM was also found for the EQ-5D (-0.46), but for all other patient-based instruments a small MSRM was found. The lowest MSRMs were found for the ASQoL and RLDQ for patients reporting both improvement or deterioration in AS, suggesting that they show little responsiveness to AS-specific change in health.

For patients reporting an improvement in general health the largest MSRM was found for the EuroQol thermometer (1.65), followed by the PGI-AS (informed and open)(0.74). Large MSRMs were also found for the SF-12 MCS (0.67), the BASDAI (0.67) and the EQ-5D (0.64). Moderate to large MSRMs were found for the SF-12 PCS (0.57) and the ASQoL (0.55). All other MSRMs were small, the smallest was found for the Body Chart (0.13).

For patients reporting a deterioration in general health the largest MSRM was calculated for the PGI-AS (informed and open)(-0.75) followed by the EQ-5D (-0.44), the EuroQol thermometer (-0.43) and the BASDAI (0.43). MSRMs for all other patient-based instruments were small.

For patients reporting an improvement in AS (-0.86) or general health (-1.21) a large MSRM was produced for the anthropometric measurement of fingertip to floor distance (FFD) (table 7.13). However, this was associated with a very small MSRM in patients reporting a deterioration in health (AS 0.18; general health 0.08) suggesting that the measurement is not sensitive to deterioration in health. Anomalies in the cervical rotation results were found with very low and very high responsiveness

	AS health transition				General health transition			
	Better (n= 56)		Worse (n= 73)		Better (n= 51)		Worse (n= 62)	
Instrument[a]	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM
ASQoL	-1.00 (3.50)	-0.29	1.18 (3.32)	0.35	-1.57 (3.53)	-0.55	0.74 (4.21)	0.26
BASDAI	-0.98 (1.45)	-0.60	0.37 (2.16)	0.22	-1.12 (1.41)	-0.67	0.72 (2.08)	0.43
EuroQol EQ-5D	0.13 (0.25)	0.50	-0.12 (0.34)	-0.46	0.16 (0.26)	0.64	-0.11 (0.35)	-0.44
EuroQol - thermometer	23.23 (92.0)	1.30	-5.02 (22.2)	-0.28	27.00 (95.1)	1.65	-7.06 (23.6)	-0.43
RLDQ	-2.05 (4.87)	-0.32	-1.82 (6.12)	0.28	-2.29 (4.93)	-0.37	2.12 (6.15)	0.34
SF-12 MCS	3.82 (10.24)	0.41	-1.36 (10.7)	-0.15	6.06 (10.41)	0.67	-1.39 (10.3)	-0.15
SF-12 PCS	3.95 (7.44)	0.51	-2.04 (6.67)	-0.26	3.84 (7.82)	0.57	-1.33 (8.63)	-0.20

Table 7.11 Mean score changes (standard deviations) and modified standardised response mean (MSRM) at 6-months. Postal and clinic data combined.

[a] instrument scoring summarised in table 7.5

	AS health transition				General health transition			
	Better (n= 36)		Worse (n= 44)		Better (n= 29)		Worse (n= 38)	
Instrument[a]	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM
Body Chart	-0.46 (0.85)	-0.53	0.44 (0.71)	0.51	-0.33 (1.18)	0.13	-0.16 (0.37)	-0.11
PGI-AS (combined)	0.51 (1.79)	0.33	-0.30 (1.43)	-0.20	0.86 (1.93)	0.58	-0.61 (1.38)	-0.41
PGI-AS	0.52 (1.62)	0.33	-0.09 (1.49	-0.06	0.75 (1.48)	0.48	-0.41 (1.71)	-0.15
- blind	(n= 16)		(n= 23)		(n= 14)		(n= 20)	
PGI-AS	0.66 (1.68)	0.42	-0.75 (1.63)	-0.48	1.34 (1.87)	0.74	-1.00 (1.13)	-0.75
- informed and open	(n= 20)		(n= 21)		(n= 15)		(n= 18)	

Table 7.12 Mean score changes (standard deviations) and modified standardised response mean (MSRM) at 6-months. Postal data.

[a] instrument scoring summarised in table 7.7

	AS health transition				General health transition			
	Better (n= 7)		Worse (n= 13)		Better (n= 7)		Worse (n= 10)	
Instrument[a]	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM	Mean Change (SD)	MSRM
Csp rotation - left	1.48 (2.17)	0.89	0.05 (2.22)	0.03	0.94 (2.32)	0.47	0.74 (1.47)	0.37
Csp rotation - right	0.46 (2.12)	0.32	-13.57 (54.9)	-9.36	0.72 (2.43)	0.36	-16.62 (61.6)	-8.31
FFD	-5.87 (11.26)	-0.86	1.23 (6.20)	0.18	-8.22 (12.8)	-1.21	0.57 (5.51)	0.08
Lat lsp flex - left	-0.31 (2.09)	-0.08	1.80 (4.84)	0.45	0.04 (2.46)	0.01	1.80 (5.33)	0.47
Lat lsp flex - right	-0.58 (3.16)	-0.18	1.10 (4.11)	0.33	-1.04 (3.74)	-0.33	1.42 (4.43)	0.45
MSI	0.43 (0.66)	0.57	0.08 (0.85)	0.11	0.42 (0.66)	0.52	-0.22 (0.62)	-0.27
TWD	-0.63 (1.35)	-0.43	0.25 (1.53)	0.17	-0.82 (1.56)	-0.59	0.19 (1.74)	0.14

Table 7.13 Mean score changes (standard deviations) and modified standardised response mean (MSRM) at 6-months. Clinic data.

[a] measurement scoring summarised in table 7.9.

statistics calculated for right and left rotation. Sample sizes for the assessment of responsiveness for the anthropometric measures were very small, and interpretation of the results is therefore difficult.

Although small to moderate MSRMs were found for lateral lumbar flexion (LLF) in patients reporting a deterioration in AS (0.33 to 0.45) or general health (0.45 to 0.47), and for the Modified Schober Index (MSI) and tragus to wall distance (TWD) in patients indicating an improvement in AS (MSI 0.57, TWD -0.43) or general health, the associated mean change in actual movement may not be of clinical significance (Moll et al, 1971)(table 7.13).

7.5 Discussion

Responsiveness describes the ability of an instrument to detect change in health over time and is an essential measurement property for evaluative instruments (Kirshner and Guyatt, 1985). Although several study instruments have satisfactory evidence in support of their reliability and validity, most have little published evidence of their responsiveness in patients with AS (Chapter 2). This study has demonstrated more extensive testing of instrument responsiveness than previously reported, and it has involved a larger sample of AS patients which improves generalisability. None of the disease-specific and anthropometric measures of outcome have previously been compared to change in patient-reported health transition as a reflection of longitudinal validity. The comparative responsiveness of instruments, particularly in relation to more established generic instruments such as the EuroQoL and the SF-12, has not previously been assessed. This provides important information to support the selection of evaluative instruments in a disease such as AS where the beneficial effects of management in routine care may not result in large changes in HRQL (Liang et al, 1985; Fitzpatrick et al, 1993c).

In the current study most patients had established and well-controlled AS. All patients underwent usual care and although fluctuations in state would be anticipated (Fortin et al, 1995) large improvements or deterioration in HRQL were not expected. Under these circumstances, it may be difficult to distinguish between no real underlying change in condition and lack of instrument responsiveness (Ruta et al, 1998; Peto et al, 1998), and narrowly focussed disease-specific instruments were expected to be more responsive to any change that occurred. However, both generic

and disease-specific patient-based measures of outcome demonstrated satisfactory evidence in support of their responsiveness. Although the majority of patients reported their health as on average the same as six months earlier, the strong relationship between both AS-specific and general health transition over the six month period and changes in health as measured by all patient-based instruments is evidence of the validity of both the disease-specific and generic instruments as measures of health outcome in AS. However, most anthropometric measures had a very poor relationship to both transition questions, suggesting poor validity as evaluative measures of outcome in AS when assessed against the described external criteria over a six month period.

A stronger relationship between the generic instruments and the general health transition was expected but only found for the EuroQol. The PGI-AS (informed and open) and the BASDAI had a strong relationship with both transition questions, but had the strongest relationship with general health transition when compared to all other instruments. Likewise a stronger relationship between the specific transition and the disease-specific instruments was expected, but again a stronger relationship with the EuroQol thermometer was found in comparison to all other instruments. A similar result was reported by investigators assessing the responsiveness of the EuroQol in patients with RA (Hurst et al, 1997) and in patients with angina (Garratt et al, 2000).

The EuroQol thermometer was found to be more responsive to improvement in both AS and general health than all other study instruments when the responsiveness statistic was calculated. Although not as responsive to deterioration in health, moderate MSRMs comparable to most disease-specific instruments were found for the EQ-5D for deterioration in AS-specific health, and for both sections of the EuroQol to deterioration in general health. The Body Chart and the PGI-AS (informed and open) produced large responsiveness statistics for deterioration in health on AS transition. The PGI-AS (informed and open) also produced a large responsiveness statistic for detecting both improvement and deterioration in general health.

The small sample size for the anthropometric measures made data interpretation difficult. However, FFD appeared to be the most responsive measure to improvement

in AS or general health, but was unable to reflect deterioration in health, and cervical rotation appears worthy of further investigation. However, most other measures appear not to be responsive to change in health over six months.

Apart from the performance of the EuroQol thermometer, there is no clear trend that disease-specific or generic instruments are more responsive to patient-reported change in health, but the anthropometric measures would appear not to be as responsive as patient-based instruments to change in health over the limited study period.

Item content and overlap between generic and disease-specific instruments, for example, pain and functional activity, and the influence of both specific and general health on these items may explain why all patient-based instruments demonstrated a strong relationship with both transition questions. Alternatively, the instruments demonstrating greater responsiveness may reflect the goals of usual care (Guyatt et al, 1999). Pain is a dominant feature of AS, and pain management is often an important objective in routine practice (Dziedzic, 1998). The strong influence of pain on a patients perception of short term change in RA has been described (Fitzpatrick et al, 1993b). Therefore, instruments with a focus on pain may be expected to demonstrate a stronger relationship with patient reported health transition following routine management over a six month period. Pain is a dominant feature of both the BASDAI and the Body Chart, it may be nominated by patients completing the PGI-AS and is included in the EuroQol. These instruments all demonstrated strong levels of correlation with health transition, supporting the importance of including the evaluation of pain in the short term assessment of change in AS.

a) PGI-AS

Evidence for the responsiveness of the PGI-AS (informed and open) in patients with AS is satisfactory. The blind format was not responsive to change and did not correlate strongly, or significantly with AS or general health transition. The closed format was not assessed for responsiveness in the current study. The stronger relationship between the PGI-AS and general health transition than observed with the specific transition may indicate a narrowness of context addressed by the AS health transition when compared to the scope of the PGI-AS. Patients may relate change in disease symptomology more clearly to AS-specific transition, whereas the PGI-AS addresses a broader concept that is more clearly reflected in the general transition.

Alternatively, the relationship may indicate that the areas picked by patients have wide implications for general health. Validity testing of the PGI-AS found a stronger relationship with both generic instruments than found for most other disease-specific instruments, supporting the hypothesis that the PGI-AS measures a broad domain of HRQL (Chapter 6). This may also indicate that AS has a large impact on general health.

Further information of the role of the different formats in patients indicating change in health is required. An index of change for areas mentioned in step 1 of the PGI-AS (identifying areas), as calculated for test-retest reliability (Chapter 5), may inform on the nature of change in areas over the six month period in patients indicating improvement or deterioration in health, and the relative responsiveness of the PGI-AS associated with this change. Following completion in the clinic survey, blind to areas mentioned at baseline, a patient listed identical areas despite indicating that he was much better than six months earlier. All areas had better scores. When shown his baseline areas he commented:

> 'These are always the most important elements of my life affected by AS, whether the disease is active or in remission - an underlying feeling of tiredness is always present which affects everything else'

However, the blind completion format may introduce 'noise' when patients do not list the same areas, thus reducing reliability, and may be a factor in the reduced responsiveness calculated for the blind format. When patients remaining the same at two-weeks were given the opportunity to change areas, all patients retained the original list. The impact of change in health on this decision requires further assessment.

Sample sizes for the evaluation of all measurement properties of the PGI-AS, in particular the responsiveness and reliability testing for the different formats, were small and reduces confidence in the results. Further work to assess the measurement properties, acceptability and feasibility of the PGI-AS should be performed with a larger sample size for all three formats to improve confidence in the results and to support the recommendation of one format for evaluative purposes.

b) Disease-specific

Items within the ASQoL are not directly anchored to the impact of AS on HRQL, and the stronger association with change in general health than with change in AS-specific health, and the stronger than predicted relationship with the generic instruments in the assessment of validity (Chapter 6), suggests that the ASQoL more effectively evaluates change in general HRQL than change in AS-specific HRQL. Specific anchoring of items towards the impact of AS may improve AS-specificity. However, evidence of responsiveness was poor and does not support its adoption in the routine evaluation of AS patients. Large changes in HRQL are required before a change in the dichotomous response format can be expected, and the ASQoL may therefore be insensitive to small, but important changes in HRQL. A revision of the response scale to include more response options may improve responsiveness (Streiner and Norman, 1995).

In patients for whom a final score could be calculated, high levels of responsiveness were found for the BASDAI. Although items within the BASDAI are symptom based, the stronger relationship with general health transition than with AS-specific transition may indicate that the items have a greater implication for general health. Validity testing of the BASDAI generally found stronger levels of correlation with the generic instruments than hypothesised (moderate to high)(Chapter 6), and further supports the hypothesis that AS has a wide impact on general health.

As predicted, the Body Chart has a strong relationship with AS transition, but a weak relationship with general transition, supporting the importance of specific pain evaluation in AS. A satisfactory level of responsiveness was found in the postal survey for patients reporting an improvement or deterioration in health. The sample size for the clinic survey was too small to draw inferences from the data, and this evaluation should be repeated with a larger sample size.

Responses to the RLDQ generally corroborate a patient's perception of change on both transition questions. However, the ceiling effect described in Chapter 5 suggests that the instrument may underestimate improvement in function, which is supported by the poor responsiveness found in the current study. Documentation of improvement in patients with excellent health may be less of a concern than documenting deterioration in all patients, but especially in those already experiencing

poor health (Bindman et al, 1990). However, the responsiveness of the RLDQ to deterioration in health also appears to be limited. The four response options of the RLDQ present a compressed response range and revision to the response format is recommended, in addition to a revision of items (Chapter 5).

c) Generic

The EuroQol (EQ-5D and thermometer) was found to be more responsive than all other disease-specific and generic instruments, and combined it is able to detect improvement or deterioration of AS-specific or general health. The SF-12 was also responsive to change in both general and AS-specific health, with results that were comparable to most disease-specific instruments. However, the SF-12 was less able to detect deterioration in health.

These results challenge the recommendation to include disease-specific instruments in evaluation for their greater responsiveness when compared to generic instruments. The study does not provide sufficient information to determine why the EuroQol was more responsive than disease-specific instruments, but several hypotheses can be proposed. The item content of the EuroQol is biased towards the measurement of functional disability and so has a close affinity to the problems experienced by patients with AS (Bakker et al, 1995). In conditions with a similar impact on function such as RA (Hurst et al, 1997) and angina (Garratt et al, 2000), the EuroQol has also demonstrated greater levels of responsiveness than disease-specific instruments. However, in conditions with a lesser impact on functioning, such as obstructive sleep apnoea, the EuroQol has not demonstrated such satisfactory responsiveness (Jenkinson et al, 1998b). Alternatively, certain disease-specific study instruments may have inadequate measurement properties, a suggestion supported by the results of further evaluations conducted in this study. The suitability of the health transition questions included as external criteria in the evaluation of responsiveness should also be considered. However, both generic and disease-specific transition items were assessed separately and the EuroQol demonstrated a stronger relationship with both items than most other instruments. Finally, the role of external valuations in calculating the index score for the EuroQol, as opposed to item summation, may also support the high level of responsiveness found in patients with AS.

d) Anthropometric

Most anthropometric measures were not responsive to change over six months. Fingertip to floor distance (FFD) was the only measure to demonstrate a significant relationship with transition questions and a moderate level of responsiveness associated with improvement in health. However, FFD methodology does not account for the starting position or patient height and should not be recommended for group comparison (Chapter 5,6). Interpretation of the responsiveness of cervical rotation was difficult. A strong correlation with radiographic change in the lumbar spine (Kennedy et al, 1995) supports the role of the Modified Schober Index (MSI), lateral lumbar flexion (LLF) and tragus to wall distance (TWD) in reflecting long term, irreversible change in spinal status, but these measures do not reflect clinically important change in range of movement over six-months.

Assessment of validity suggests that most anthropometric measures bear little relation to patient-based measures of HRQL, indicating that they measure a different concept of disease impact. This result therefore challenges the appropriateness of the health transition questions included in the assessment of the anthropometric measures. Change in disease-specific or general health may have little impact on change in range of movement and more appropriate external critieria may be radiographic change or change in alternative clinical criteria.

The domains addressed by many study instruments may not be expected to change over the relatively short period of the study in patients with stable AS. However, the six month period reflects normal practice in the routine evaluation of AS in many rheumatology centres (Dziedzic, 1998; Lubrano et al, 1998). The level of responsiveness found in several patient-based instruments, particularly the EuroQol, PGI-AS (informed and open) and the BASDAI may make them suitable for routine monitoring of health outcome in the longitudinal evaluation of AS, where routine management may result in subtle change in HRQL. However, levels of reliability (Chapter 5) suggest that these instruments are only suitable for group assessment. Instruments with the highest reliability, the ASQoL, RLDQ and the SF-12 (PCS), where individual assessment was supported, demonstrated poor responsiveness.

The choice of patients and external criteria may influence responsiveness and the use of health transition questions may overestimate responsiveness when compared to

trials of known efficacy because the latter will also include patients who have not improved (Deyo et al, 1991). Therefore, further evaluation of instrument responsiveness following a trial of known efficacy in AS is recommended. Limited evidence of the responsiveness of disease-specific study instruments has been generated following trials of physical therapy only and responsiveness statistics have not been calculated. Evidence of the impact of drug therapy should also be considered in future evaluations. It is suggested that comparison of instruments in randomised controlled trials may provide the strongest evidence of the differential measurement properties of evaluative instruments (Guyatt et al, 1999).

The patient population represents a wide range of disease presentation, with similar features to those reported in other hospital based studies (Garrett et al, 1994; Lubrano et al, 1998). However, patients with newly diagnosed disease were not widely represented. This may be a suitable subset of patients in which to assess the impact of diagnosis and early stages of management on HRQL. Evaluation of measurement properties in a variety of patient groups improves confidence in instrument performance and the generalisability of application (Bindman et al, 1990).

Conclusion

Although the results from the evaluation of the PGI-AS are based on a smaller sample size, evidence suggests that the PGI-AS (informed and open) and the BASDAI are the most responsive disease-specific instruments in the current study.

The generic instruments were not markedly less sensitive to change over time when compared to disease-specific instruments, and the EuroQol was generally the most responsive instrument on both methods of assessment, and when both improvement or deterioration in health was considered. These results challenge the assumption that disease-specific instruments are more responsive than generic instruments and hypotheses to explain these findings have been proposed.

Based on the available evidence no anthropometric measure can be clearly recommended for evaluation of AS. The measurement of cervical rotation and FFD require further exploration as measures capable of reflecting short term and reversible change. The MSI may reflect structural and irreversible change in the long term

evaluation of outcome in AS (Kennedy et al, 1995; Dawes, 1999). Further empirical evidence of the role of anthropometric measurement in AS is required.

The expectation of including patient-based and anthropometric measures of outcome in the evaluation of AS, the relevance of change in score to clinicians and patients, and the minimal clinically important difference that may influence clinical decision making must be addressed.

Chapter 8 Summary and Discussion

8.1 Introduction

This chapter provides a summary of the current state of research into the measurement of health outcome, and the contribution of the described study to this body of knowledge. The limitations of this research are discussed and an agenda for future research is presented both within the context of AS health outcome measurement and for the measurement of health outcome in general. Section 8.2 summarises the current state of research into the measurement of health outcome, and specifically within AS, and Section 8.3 discusses the main findings of the preceding Chapters. Section 8.4 addresses the general implications and main conclusions from the study. Study limitations and suggestions by which deficiencies in the study might be addressed in future research are discussed in Section 8.5. The chapter concludes with Section 8.6.

8.2 The measurement of health outcome

Accurate measurement of outcome across the wide spectrum of health and disease has become an important medical and social issue (Ware, 1998). Whereas traditional methods of measurement focussed on the presence or absence of disease or the measurement of impairment, the changing prevalence of disease dictated a change in emphasis in management and the methods of evaluation (McDowell and Newell, 1996). The role of the patient is increasingly seen as central to this process and is reflected in the increasing availability of patient-based measures of outcome and the emergence of individualised measures, for example, the PGI-AS.

A wide range of evaluative instruments can be described (McDowell and Newell, 1996; Bowling, 1997). However, there is little standardisation in measurement practice and instruments often have inadequate evidence describing their development, testing and practical features which makes instrument selection for routine practice or clinical research very difficult (McDowell and Newell, 1996). Selection has often been guided by historical precedence (Jenkinson et al, 1994a) or more recently by expert opinion (van der Heijde et al, 1999a,b,c), and the need to improve the quality of patient evaluation by the adoption of instruments with clear evidence to support the development, measurement properties, acceptability to

patients and feasibility for the required application is an important requirement for all fields of health care (Kirshner and Guyatt, 1993; Fitzpatrick et al, 1998a).

Within several areas of rheumatology recommendations for domains to include in patient evaluation and instruments to fulfil these domains have been made (Calin et al, 1999b). However, many instruments were developed for research application (Bellamy et al, 1998) and recommendations often appear to have little relevance to routine practice. Not surprisingly the application of patient-based measures of outcome has been greatest in clinical and health services research and the role, integration and acceptance of these instruments in routine practice now demands greater attention. Clinicians demand that instruments are quick, simple and easy to score whilst requiring information about the benefits of including patient-based instruments alongside traditional methods of evaluation in routine practice and medical audit (Bellamy et al, 1998). For example, the benefit to clinical decision-making, individual patient outcome and quality of care, resource allocation, purchasing decisions and health policy. Although the importance of these qualities is widely recognised (Ware, 1997), the feasibility of using this information to inform on these attributes is poorly understood (Garratt, 1997; Greenhalgh and Meadows, 1999).

8.3 Health outcomes in AS: summary of findings

The current research consists of three main stages: first, a systematic review and evaluation of patient-based and anthropometric measures of outcome applied in evaluative studies of AS (1990-2000). Second, the development of the first AS-specific individualised measure of disease-related quality of life, the PGI-AS. And third, an empirical comparison of generic and disease-specific patient-based and anthropometric measures of outcome in patients with AS.

The systematic review of the entire range of patient-based and anthropometric measures of outcome applied in published studies of AS, and the evaluation of all disease-specific and anthropometric measures represents the first detailed and explicit synthesis of evidence relating to the development, measurement properties, acceptability and feasibility of outcome measures applied in current practice in AS. Data evaluation supported the adoption of two disease-specific patient-based and five anthropometric measures in the third stage of the study. Selected instruments reflected domains considered important in the evaluation of AS (van der Heijde et al,

1997; RAG - expert opinion, 1998), and patient-based measures were acceptable for self-completion in a postal survey.

Although domain-specific instruments capable of evaluating a wide range of issues related to HRQL were identified, disease-specific measures of pain or of HRQL were not identified by the original review. The first AS-specific, individualised measure of disease-related quality of life, the PGI-AS, was therefore developed. The PGI was identified as a basis for the new instrument due to the individualised approach to evaluation described (Ruta et al, 1994a). Before the PGI-AS can be recommended for the evaluation of patients with AS in the United Kingdom (UK), it must demonstrate acceptable levels of acceptability, feasibility and measurement properties. The third stage of the study has provided initial evidence for these properties. Communication with measurement experts in rheumatology subsequently identified the AS Quality of Life questionnaire (ASQoL), an unpublished AS-specific measure of HRQL, (Doward L.- personal communication, 1998), and the Body Chart, an AS-specific measure of global bodily pain. Both instruments were included in the comparative study.

Several generic measures of HRQL were identified in the review but these were long and not suitable for application in self-completed format. Therefore, a further literature search identified two additional generic instruments, the EuroQol and the SF-12, which had good evidence of measurement properties in patients with similar disorders to AS and were brief and suitable for self-completion (Hurst et al, 1997, 1998; Coons et al, 2000).

The third stage of the study describes the first comparative evaluation of the measurement and practical properties of a broad ranging and evidence-based package of instruments in the same population of AS patients. The instruments have also been assessed for data quality and scaling assumptions, the results of which have previously not been reported for the disease-specific instruments. In addition, the ASQoL, BASDAI and RLDQ were assessed for dimensionality, a property not previously reported.

The comparative study involved a longitudinal evaluation of instruments in both a clinic based and postal survey. The baseline response rate to the clinic survey was acceptable (n = 159, 59.0%) and comparable to other studies (Lubrano et al, 1998).

Postal evaluation involved completion of all patient-based instruments in a self-completed questionnaire using a multi-centre study design with participants from the North, Midlands and South of the UK. The baseline response rate (n= 349, 77.4%) was satisfactory. The study represents the largest clinic-based and one of the largest multi-centre postal evaluations of outcome measures in UK-based AS patients. The study also describes the most rigorous process of instrument testing previously reported in AS. This enhances confidence in the results, which are more generalisable to the evaluation of the wider AS population.

When data quality and scaling assumptions were assessed all instruments, except for the BASDAI and RLDQ, demonstrated adequate properties at both item and scale level. Items in the BASDAI had high levels of missing data and modification of the response format is strongly recommended. Although all response options were covered, and end-effects did not exceed recommended levels, responses to items of the RLDQ were skewed towards better levels of functional ability. Increasing the number of response options, changing descriptors to improve discrimination between options, or adding further items that represent more difficult functional activities are suggested as ways to improve the data quality of the RLDQ. The BASDAI demonstrated adequate properties at scale level for those patients for whom a score could be calculated, and the distribution of responses to the RLDQ at scale level were approximately normally distributed, although still failing to cover the extreme range of disability described by the instrument. Assessment of the dimensionality and item-total correlation of the ASQoL, BASDAI and RLDQ supports the uni-dimensional structure proposed by the developers.

Although completion of the PGI-AS following interview-administration was excellent, acceptability of the self-completed format may be improved by simplifying the spending of points in step 3 (spending points), and allowing for minor errors in summation of these points. Also, more explicit guidance for completion by patients not experiencing AS-specific problems is required.

All instruments were assessed for test-retest reliability in patients indicating no change in both AS and general health at two-weeks, and all instruments exceeded levels recommended for group evaluation (Nunnally and Bernstein, 1994). The ASQoL, RLDQ, SF-12 PCS and all anthropometric measures achieved levels that

support their use in individual evaluation (Streiner and Norman, 1995). Where appropriate, tests of internal consistency reliability also exceeded levels recommended for group analysis (BASDAI), and for the ASQoL and RLDQ exceeded levels for individual evaluation.

An index of change for step 1 (identifying areas) of the PGI-AS was calculated to assess the impact of area changes on test-retest reliability. In patients completing the instrument blind to baseline areas, when few areas were changed reliability supported its use in individual evaluation (0.91). However, when more than three areas were changed reliability levels suggested that the format was not suitable for use in group evaluation (0.56). All patients completing the informed and open format retained their original list and reliability supported its use in group evaluation (0.85). When considered irrespective of the index of change, the highest level of reliability was calculated for the closed format (0.87), and the lowest for the blind format (0.81). It is suggested that the blind format may introduce noise into the assessment of reliability, and the open and informed format may not be necessary. Therefore, keeping the selected areas the same at follow-up completion (closed format) may improve the reliability and clinical validity of the instrument, without a threat to the content validity. However, this result only applies to patients indicating no change to health at two weeks, and does not consider the role of the PGI-AS in patients indicating change. Recommendations to reduce the number of response options in stage 2 of the PGI-AS (scoring areas) may also help to improve reliability.

Construct validity was assessed by relating instrument scores to other more established instruments, to other disease-specific instruments and to sociodemographic variables. No disease-specific instrument has previously been assessed against instruments with such well-established and documented levels of validity as observed in the EuroQol and SF-12. In addition, no study evaluating measures of outcome in AS has constructed hypothetical relationships between instruments and set out to test these hypotheses. All correlations between instruments were in the hypothesised directions. The results represent good evidence for the validity of the PGI-AS as a measure of disease-related quality of life and support the validity of all disease-specific patient-based measures of outcome. The ASQoL and the PGI-AS were the best performing disease-specific measures. However, the ASQoL had a stronger than predicted relationship with the generic measures of

HRQL, which may suggest inadequate anchoring of items to the specific impact of AS. Evidence further supported the validity of both the EuroQol and the SF-12 as generic measures of HRQL in AS. The EuroQol was the generic instrument and the ASQoL the disease-specific instrument most capable of discriminating between patients on most sociodemographic variables. As hypothesised, very weak associations between all anthropometric measures and patient-based measures were found supporting a minimal association between limitation in specific spinal mobility and the impact of AS on various aspects of HRQL. The strongest, but weak, associations were observed between cervical rotation and fingertip to floor distance (FFD) and all patient-based measures.

All instruments were assessed and compared for responsiveness to change at six months by two criteria: first, the linear relationship between change in instrument score and patient reported change on health transition; and secondly, a responsiveness statistic was calculated. Although most patients indicated that on average their AS-specific and general health was the same, all patient-based instruments demonstrated strong and significant linear relationships with both AS and general health transition questions, which supported instrument longitudinal validity. The strongest relationships were found for the EuroQol, the PGI-AS (informed and open) and the BASDAI on both AS-specific and general health transition. The changes in scores for four of the five anthropometric measures did not reflect the categories of AS or general health transition. A significant relationship between fingertip to floor distance (FFD) and both transition questions was found, and although the results apply to a smaller sample size, this was a smaller relationship than observed for all patient-based instruments.

Instrument responsiveness was compared for all patients reporting an improvement or deterioration on health transition using the Modified Standardised Response Mean (MSRM)(Fitzpatrick et al, 1998a). The EuroQol was generally the most responsive instrument, with both sections of the instrument able to detect improvement or deterioration in AS-specific or general health. The thermometer and the EQ-5D were the most responsive to improvements and deterioration in health respectively. The PGI-AS (informed and open) also performed well in measuring improvement or deterioration in AS-specific or general health. The lowest levels of responsiveness were found for the ASQoL, the RLDQ and PGI-AS (blind) for change in AS-specific

health and for the Body Chart, the RLDQ and PGI-AS (blind) for change in general health. Although the sample size was greater than most identified studies evaluating the measurement properties of anthropometric measures in AS, sample sizes were too small to allow clear inferences to be drawn from the data. However, FFD and cervical rotation demonstrate small levels of responsiveness. The results suggest that the PGI-AS (informed and open) and the BASDAI are the most responsive disease-specific instruments. The EuroQol (EQ-5D and thermometer) is the most responsive generic instrument, and in the majority of comparisons was more responsive than the disease-specific instruments.

Two formats of the PGI-AS were completed at six months; the informed and open format was the most responsive. It is suggested that the noise that influences the reliability of the blind format may also influence responsiveness. The closed format was not assessed for responsiveness.

8.4 General implications and conclusions

The systematic review highlighted the wide diversity of outcome measures applied in the evaluation of AS. It also described the lack of standardisation, the limited evidence describing the measurement properties or comparative performance, acceptability and feasibility of the majority of patient-based and anthropometric measures of outcome adopted in AS, and the focus towards measures of impairment and disability. The review supported the need to recommend a standardised and evidence-based package of patient-based and anthropometric measures of outcome that would be suitable for application in routine practice and clinical research, and to extend measurement practice to consider the role of both disease-specific and generic measures of HRQL in AS.

The subsequent empirical evaluation has provided the first comparative evidence for the measurement properties, acceptability and feasibility of the PGI-AS, an evidence-based selection of disease-specific patient-based and anthropometric measures, and two widely applied generic measures of HRQL in a large population of AS out-patients. Combined with the systematic review and data evaluation, this provides important information against which the relationship between instruments can be judged in terms of necessary measurement properties (McHorney and Tarlov, 1995; Beaton et al, 1997).

Most patients with AS present with multiple, coexisting problems and in order to develop a thorough understanding of the impact of AS, a multi-dimensional approach to evaluation is required. However, in making recommendations consideration must be given to what the clinician or investigator wishes to measure, and what is important to the patient. It is therefore convenient to describe measures of outcome in terms of the domains considered important in the evaluation of AS.
Recommendations for instruments to fulfill these domains, based upon the available empirical evidence of measurement properties, acceptability and feasibility are made, together with a recommendation for application in routine practice or research.

Individualised disease related quality of life

Available evidence suggests that the PGI-AS (informed and open) may be used in the evaluation of groups of patients with AS in routine practice or clinical research following an interview-administered format. Although interview-administration will reduce acceptance in clinical research, minimal interviewer training is required and after the initial average completion time of 10-minutes, most patients complete subsequent formats in approximately 5-minutes. Although completion rates for the modified version of the PGI reported here are a great improvement on those previously reported, further recommendations are made to reduce missing data, which may improve acceptance as a self-completed instrument in clinical research or routine practice. Evidence suggests that the informed and open format is both more reliable and more responsive to change at six months than the blind format, but evidence of the responsiveness of the more reliable closed format is required before recommendations for a particular format for purposes of evaluation can be made.

The PGI-AS offers a unique approach to the evaluation of disease-related quality of life by allowing patients to nominate individualised areas of life affected by AS. The moderate correlation with other patient-based and generic measures of HRQL suggest that it is measuring different aspects of HRQL, not covered by the more conventional patient-based instruments included in the study. When applied in combination with patient-based and anthropometric measures of outcome the PGI-AS has the potential to provide more individualised information relating to health outcome, but further refinement of the instrument is recommended to improve levels of reliability and to enhance its role in individual evaluation. Evidence also supports the hypothesis that

the individually tailored evaluation of disease-related quality of life described by the PGI-AS provides a more responsive patient-specific evaluation than found for most other conventional disease-specific instruments included in the study. The original versions of the PGI were also found to be generally more responsive to change than disease-specific and generic instruments in several different patient populations (Ruta et al, 1994; Garratt, 1997; Ruta et al, 1999).

AS-specific HRQL

Although the ASQoL has good completion rates, satisfactory data quality and scaling assumptions, and a very high level of reliability, it is not recommended for the routine evaluation of AS patients due to its poor level of responsiveness. Anchoring of items to the impact of AS may improve the validity of the instrument as an AS-specific measure of HRQL, and revision of the dichotomous response scale may improve discrimination and responsiveness to change.

Disease Activity

High levels of missing data following self-completion of the BASDAI prohibits the recommendation for use as a self-completed instrument in clinical practice or research. Although interview-administration may reduce these levels, this is a time consuming process and reduces instrument acceptance in clinical trials (Fitzpatrick, 1999). However, satisfactory measurement properties in the current study for those patients completing the instrument adequately to receive a score, support a recommendation for its use in the evaluation of groups, and strongly indicate that a revision of the response scale is necessary before recommendation for self-completion in clinical practice or research can be made.

Body Pain

The Body Chart was the only disease-specific instrument to demonstrate a much stronger association with change in AS-specific health than general health over the six months, being responsive to both improvement and deterioration in AS. The instrument also had satisfactory levels of reliability and validity and can be recommended for the evaluation of groups following interview administration in routine practice or research. Although not originally designed for self-completion, further clarification of the self-completed format is required which may improve instrument acceptance in clinical research. Improved instrument reliability may be

achieved by a more standardised approach to shading the body manikin and the awarding of points.

Functional Disability

Although the RLDQ has good completion rates and a very high level of reliability, it is not recommended for evaluative purposes in AS due to its poor data quality at item level, the inadequacy of the instrument to provide a broad reflection of functional disability in AS, and the low level of responsiveness. Recommendations to revise the response format and item content have been made which may improve data quality, the coverage of functional disability and responsiveness.

Generic HRQL - Utility measure

Both sections of the EuroQol had high completion rates and levels of reliability that support application of the instrument in group evaluation. Evidence supports the validity of the instrument as a measure of generic HRQL in AS, and both sections of the instrument, when applied together provide a generic instrument that is both responsive to improvement or deterioration in AS-specific and general health over six months. Where an index of generic HRQL is acceptable the EuroQol is recommended for the evaluation of groups in both routine practice and clinical research in AS.

Generic HRQL - Health profile

Completion rates of the SF-12 were satisfactory although lower than the EuroQol, and modification of the treatment of missing values has been recommended to allow for item omission without jeopardising a final score. Reliability of the SF-12 PCS supports application in individual evaluation, but the MCS should only be used in the evaluation of groups. Evidence of validity supports its role as a generic measure of HRQL in AS. The levels of responsiveness were small to moderate, were less than those observed for the EuroQol, and the instrument was less able to detect deterioration in general or AS-specific health. Where a limited health profile is acceptable and respondent burden is a factor, the SF-12 may be acceptable for individual evaluation in routine practice or research. However, the measurement error associated with the MCS and the limited ability to detect deterioration in health may limit this role.

All patient-based instruments are readily available and most are easy to score without the use of a computer-based programme. However, due to the external weighting of the EuroQol and the scoring algorithms of the SF-12 computer-based scoring of these instruments is recommended.

Anthropometric measures

The results suggest that the choice of anthropometric measures could be reduced to the measurement of cervical rotation, as a short term reflection of reversible change in spinal mobility and the MSI as a long term reflection of irreversible change in spinal status (Kennedy et al, 1995; Dawes, 1999). Both measurements can be completed in less than 5 minutes, and the only instrumentation required is a relatively inexpensive plastic tape measure. Routine use of the approaches in clinical practice would suggest that they already have a level of clinical acceptance (Bellamy et al, 1998, 1999; Lubrano et al, 1998). However, a further evaluation of the responsiveness of both cervical rotation and a revised methodology of FFD over the short term, and of the MSI over the long term (that is, more than two-years (Calin et al, 1999c)), with revised external criteria, is required in a larger patient population.

In conclusion, no study instrument fulfilled the required measurement properties for the evaluation of individual patients. That is, instruments with adequate levels of reliability (> 0.90)(ASQoL, RLDQ, SF-12 PCS, all anthropometric measures) were not sufficiently responsive to change, and instruments that were responsive over the six month period did not have sufficiently high levels of reliability for individual assessment (BASDAI, Body Chart, PGI-AS, EuroQol). Although instrument selection should be made following consideration of all measurement and practical properties (Fitzpatrick et al, 1998a), reliability may be the most important issue when identifying an instrument for individual evaluation (McHorney and Tarlov, 1995). However, some clinicians may be willing to accept lower levels of reliability, and the associated increase in measurement error, if the adoption of a patient-based measure includes areas of HRQL not covered by traditional methods of evaluation (McHorney and Tarlov, 1995). For example, the identification of mental health problems or functional difficulties. In many cases the usefulness of an instrument at the individual level has been constrained by the requirements for higher levels of reliability (Garratt, 1997).

Recommendations have been made to make modifications to all disease-specific study instruments, and to the measurement of FFD, and for further empirical evaluations of instrument properties once changes have been made. Therefore, recommendations may change in light of the proposed modifications and further evaluation of measurement properties.

8.5 Limitations and criticisms

The first evidence-based selection of instruments fulfilling domains considered important in the evaluation of AS were described following the systematic review, communication with experts, an additional literature overview, and the development of the first individualised measure of AS-related quality of life (PGI-AS). However, further appraisal in the first comparative study of its type in AS indicated that no instrument adequately fulfilled the required measurement properties and practical criteria considered necessary for use in individual evaluation (McHorney and Tarlov, 1995). Several instruments can be recommended for use in group analysis and in clinical trials, but because the ASQoL and RLDQ cannot be recommended, all relevant domains are not described, that is, disease-specific HRQL and functional disability. Also, the acceptability of the BASDAI is limited without modification to the response format. The measurement of spinal mobility requires further evidence to support the responsiveness of cervical mobility and the MSI before any recommendation can be affirmed.

A second disease-specific measure of HRQL has recently been published: the AS-Arthritis Impact Measurement Scale 2 (AS-AIMS2)(Guillemin et al, 1999). Although developed in French, early evidence suggests satisfactory measurement properties, acceptability and feasibility (Guillemin et al, 1999). Following English translation direct comparison with the ASQoL and with generic measures of HRQL, for example, the EuroQol and SF-36, is recommended to assess the measurement properties of the instrument.

Two additional AS-specific measures of functional disability have been recommended by the ASAS group (van der Heijde et al, 1999a,b): the Dougadas Functional Index (DFI)(Dougadas et al, 1988; Spoorenberg et al, 1999a) and the Bath AS Functional Index (BASFI)(Calin et al, 1994). Following suggested modifications to the RLDQ, direct instrument comparison is recommended to support the selection of a single AS-

specific measure of functional disability. Alternatively, the modified RLDQ may so closely resemble the revised DFI, that the more widely used DFI is accepted as the most appropriate instrument. The response scales of the BASFI are identical to the BASDAI and completion in a patient population unfamiliar with the instrument, as described in the current research, may provide important information relating to the acceptability of the response scales.

In recommending a single format of the PGI-AS for evaluative purposes in AS a full appreciation of the measurement properties and practical considerations for all versions is required and was not provided by the current study. The closed format was not assessed for responsiveness and sample sizes were very small when the different formats were considered. It is recommended that the study is repeated with a larger sample size.

Although the PGI-AS was readily incorporated into the routine 20-minute clinic assessment, the views of the clinical physiotherapist following administration to 26 baseline patients were not obtained. This was due to the unavailability of the physiotherapist (JW), who was subsequently unable to participate in the six-month follow-up. However, the clinical relevance to health care professionals and feasibility of adopting the instrument in routine practice or in research should be addressed. Also, due to the individualised nature of the PGI-AS, consideration of the format with the greatest relevance to patients may further assist in recommending a particular format. For example, different versions could be used in a clinic setting followed by focus group interviews where patients and clinicians are asked which version they prefer. It is suggested that the closed format may be more acceptable to research where the investigator wishes to evaluate change in disease-related quality of life in relation to areas listed at baseline.

An additional follow-up format of the PGI has been adopted in patients with Multiple Sclerosis (MS)(Pimm J.- personal communication, 1999). Following completion blind to baseline areas patients are shown a copy of their original areas and are asked to score these if they differ from the new list. Although published evidence of the measurement properties of this format are not available the approach may have enhanced clinical relevance, whereby a clinician is willing to accept the incumbent

time implications. This format was not suitable for self-completion and was not included in the current research.

Although not addressed in the current research, anecdotal reports suggest that the PGI may have an important role to play in directing patient management (Ruta et al, 1999). Theoretically, the incorporation of patients needs and expectations into the management of incurable and chronic disease such as AS could be of great benefit. However, modifications to reduce the measurement error associated with the present format of the PGI-AS are required before it can be recommended for individual evaluation.

Direct comparison of the PGI-AS with an alternative individualised approach, for example, the SEIQoL-DW (Hickey et al, 1996) would allow the measurement properties of instruments with a similar conceptual base to be compared. This may provide more relevant information than an assessment of validity against instruments following a more traditional approach to the evaluation of HRQL, and may lend further support to the validity of the PGI-AS as a broad measure of disease-related quality of life.

A recent concept proposed as an important consideration for patient-based measures of outcome is that of response shift (RS), defined as a change in the meaning of an individuals self-evaluation of a target construct (Schwarz and Sprangers, 1999). This may be a result of three interrelated concepts: scale recalibration, concept redefinition or a change in the patients' internal values. Although all study instruments are susceptible to RS, it is of particular relevance to the PGI-AS where patents nominate areas for inclusion in the evaluation. RS may influence the PGI-AS format adopted for follow-up evaluation and the setting in which the instrument is completed. For example, due to the influence of social comparison. Patients completing the PGI-AS in the clinic often required more assistance at six months than at baseline, with several patients indicating that since the original completion they had thought at greater length about the impact of AS on their life. Therefore, although the impact of AS at six months may not have changed, scale recalibration by which the patient evaluated disease impact, or a redefinition and reconceptualisation of disease impact may have resulted in a RS. It is possible that such a RS could partially account for the improved responsiveness of the PGI-AS at six months when compared to other disease-specific

instruments, as opposed to reflecting a real change. It may also partially describe the poor responsiveness of the blind format.

There is no empirical evidence to support the existence of RS in individualised instruments. However, recent evidence has described the role of RS in administering the Health Assessment Questionnaire (HAQ), an arthritis-specific measure of functional ability, to patients with osteo-arthritis (OA)(Daltroy et al, 1999). Patients with a recent onset of health problems may have an inflated perception of functional difficulties due to a shift in their internal standards of measurement. Physical performance of an activity before instrument completion influenced self-reported function, possibly due to a redefinition of the concept. In the current research anthropometric assessment followed completion of patient-based instruments and should not have unduly influenced the results.

RS may be a useful construct to investigate in future research to further address the most suitable format of the PGI-AS to adopt. Change in patients perceptions, values and priorities over time are important considerations to the conceptual base of the PGI (that is, the interaction between expectation and reality), and are concepts that relate strongly to that of a response shift where patients may 'rethink and reframe' the impact of disease on HRQL (Wilson, 1999).

Few studies have compared the relationship between the EuroQol and the SF-12 (Johnson and Coons, 1999). Although both measure generic HRQL the moderate correlation between instruments in the current study suggests that they measure different aspects of HRQL. The SF-12 produces a limited profile of HRQL describing only mental and physical component summary scores, and selection of the EuroQol (EQ-5D and thermometer) is supported by the results of the current study. Although the relationship between both instruments and other disease-specific instruments is similar the EuroQol has greater discriminatory power when sociodemographic variables are assessed. The EuroQol also demonstrates a greater responsiveness to change in AS or general health over six months than the SF-12, and the inability of the SF-12 to detect deterioration in health is an important consideration if the instrument is used to assess AS relative to other disorders within health care. However, the SF-12 PCS had a level of reliability that supported its use in individual evaluation. The SF-12 was identified for the current study in preference

to the parent instrument, the SF-36, due to the reduced respondent burden. However, the SF-36 describes both component summary scores and a profile across eight domains of health. Further studies should consider the advantages of the additional information provided by the SF-36 against instrument acceptability and feasibility. The measurement properties of the SF-36 have not been rigorously tested in AS and it should be compared to the EuroQol to provide a further assessment of the role and usefulness of generic profile and utility measures in AS.

Anthropometric assessment is one component (impairment) of a multidimensional group of outcome measures considered relevant in the evaluation of patients with AS. The results of the current study present a view that challenges the role and usefulness of anthropometric assessment described by clinical investigators (Lubrano et al, 1997; Dziedzic, 1997; Lubrano and Helliwell, 1999). However, the widespread adoption of anthropometric assessment in both routine practice (Bellamy et al, 1998, 1999) and in clinical research (Chapter 2) demands that both viewpoints should be considered in recommending anthropometric assessment in AS evaluation.

When empirical evidence of the acceptability, feasibility and measurement properties of selected anthropometric measures is appraised in light of the systematic review and comparative study, this suggests that although reliable and suitable for use in individual evaluation, most measures are not responsive to change in general or disease-specific health over a six-month period. In addition, most measures have little relationship with measures of HRQL and patient perception of disease impact. However, certain measures have a strong relation to disease-specific radiographic change of the spine (Kennedy et al, 1995; Dawes, 1999). Therefore, based on this empirical evidence most anthropometric measures appear to have a limited role in the evaluation of short-term change in patients with AS. However, clinical experience and further empirical evidence of validity suggests that they may have a different role in AS evaluation to patient-based measures of outcome.

The clinical viewpoint suggests that anthropometric assessment is essential to ascertain clinical outcome in AS, to provide an insight into the natural history and serial progression of disease and to identify sub-groups of patients (Lubrano and Helliwell, 1999; Dawes, 1999; van der Heijde and Spoorenberg, 1999). An important consideration in the role of anthropometric assessment is the ability of a measure to

reflect the irreversible or reversible nature of AS, and in selecting a measure consideration should be paid to what the clinician or investigator expects from the assessment.

Structural damage is considered an important outcome in the evaluation of AS, and one that is often measured by radiographic assessment (van der Heijde and Spoorenberg, 1999). The strong relationship between AS-specific radiographic change and certain anthropometric measures, for example, the MSI, LLF and TWD (Kennedy et al, 1995), supports the ability of these measures to reflect the structural change and the irreversible nature of AS, whilst reducing the need for radiographic exposure in the serial evaluation of disease progression. However, radiographic assessment is unable to detect structural change over periods of less than two-years (Calin et al, 1999c; van der Heijde and Spoorenberg, 1999), and therefore, as demonstrated, these anthropometric measures would not be expected to be responsive to change over the six-month period of the current study. Although this result could reflect the insensitivity of the measures, it is more likely to represent the inappropriateness of the six-month follow-up period and the slow rate of AS disease progression in patients with stable disease. Therefore, significant change in these measures over the short term plays an important role in clinical decision making, and would act as a trigger for further investigations (Dawes P. and Dziedzic K.-personal communication, 2000). Over longer periods of time (six to 14 years), a gradual deterioration in range of movement has been reported in AS patients, irrespective of initial disease severity or level of exercise (Sturrock et al, 1973; Lubrano and Helliwell, 1999), and may describe more suitable periods of time over which to assess the responsiveness of anthropometric assessments reflective of irreversible change in AS.

Certain anthropometric measures have a lower correlation with radiographic change and may be more reflective of reversible change in AS (Roberts et al, 1988). For example, cervical rotation and FFD. Results from the current study suggest that these measures may be responsive to change in range of movement over six-months and may be acceptable in evaluative practice to capture the short-term effects of management. However, evidence is limited and requires further evaluation of measurement properties in larger populations of patients with AS, particularly in relation to responsiveness to change.

Clinicians may be willing to reduce the number of anthropometric measures included in an evaluation in light of empirical evidence, but may not be willing to forgo anthropometric assessment completely in favour of patient-based assessment. Additional information intuitively gained from the recording of such measures in routine practice, for example, the quality of patient movement (hesitant movement), the state of joints and skin quality (joint swelling, muscle wasting, psoriasis), the ability of the patient to dress independently (functional disability), is often purely qualitative. However, the clinical interpretation of this information associated with a theoretical appreciation of disease process is often used to assist in clinical decision making and to direct patient management. These combined issues reflect the reality of evaluation in routine practice, a reality that may not be quite so apparent in clinical trials and one that places different demands on instruments recommended for evaluative purposes.

The relevance of anthropometric assessment to patients should also be considered. Although anecdotally clinicians suggest that patients consider change in range of movement an important indicator of progression, this may be due to a lack of viable alternative that considers a patients' viewpoint of change. For example, the PGI-AS or other patient-based instrument. Several patients interviewed during the pre-pilot evaluation of the PGI-AS (Chapter 3) indicated that they were empowered by the opportunity to identify important areas of life impacted by AS, as opposed to the 'mechanical' nature of anthropometric measurement:

> 'To be told that my measurements are the same as last time, when I already know this - the quality of my life is far more important - it is very real'.

However, patients also indicated that the relevance of such instruments would be enhanced if they lead to the development of a plan of action, and did not just result in a 'paper exercise'.

The selection of anthropometric measures recommended as a result of the review and empirical study differ from those recommended by the ASAS group (table 2.65). Clear guidance in the selection of measures is required to support standardisation and

to ensure that anthropometric measures with inadequate measurement properties are not included in routine evaluative practice.

Dynamic movement has recently been assessed in AS using the Fastrack (FK). Evidence suggests that the FK more accurately describes shoulder and cervical movement than goniometer or tape measure assessment and was able to distinguish between sub-groups of disease severity (Jordan K.-personal communication, 2000). The computer generated imaging allows the visual comparison of movement against normal values and against previous assessments. However, the feasibility of including the FK in routine practice or research requires further evaluation.

Alternatively, there is little evidence to support the effectiveness of including patient-based measures of outcome in routine practice (Greenhalgh and Meadows, 1999). Clinicians may view the information as difficult to interpret and possibly as irrelevant (Bellamy et al, 1998, 1999; Chesson, 1998). However, when used alongside traditional measures score comparison may serve to enhance data interpretation of the patient-based instruments by a form of 'calibration' (Deyo and Carter, 1992; McHorney and Tarlov, 1995). Eventually, an intuitive feel for these instruments may develop. Comparison of patient-based instruments against anthropometric measures in the current research described a very small correlation providing minimal support for data interpretation. Further comparison of the patient-based instruments against other traditional measures, for example, laboratory based measures and radiographic assessment may provide further support for data interpretation.

The current research did not include laboratory based and radiographic assessment in the evaluation and identification of instruments for additional domains identified by ASAS was not possible (table 2.1) (van der Heijde et al, 1997). Further research should address the role of these forms of evaluation in AS and their association with patient-based instruments. The ASAS group have recently appraised the evidence for two laboratory based measures (ESR and CRP) but no consensus was reached on their role in AS (Ruof and Stucki, 1999b; Spoorenberg et al, 1999b). Several instruments for radiographic evaluation, including the Stoke AS Spine Score (SASSS)(Averns et al, 1996a; Dawes, 1999) and the Bath AS Radiographic Index (BASRI)(MacKay et al, 1998) have also been assessed and suggest good reliability, but evidence for responsiveness is limited (Spoorenberg et al, 1999c).

Associated with data interpretation and instrument application is the ability of an instrument to describe the full range of a domain (Fitzpatrick et al, 1998a). In the current research an inconsistency between items in the RLDQ and the wide impact of AS on functional disability was observed. That is, most items were easily performed by most patients, and very few items were considered difficult. Therefore, a score at the ceiling of the RLDQ may be interpreted as perfect function, when the instrument is more suited to measuring severe functional dysfunction and is unable to capture more subtle levels of dysfunction (McHorney and Tarlov, 1995).

Item contribution can be further investigated by Item Response Theory (IRT) including Rasch models analysis (Nunnally and Bernstein, 1994). Rasch models assume that items within a uni-dimensional instrument are equally discriminative, and evaluates items in terms of their level of difficulty and interval location within a structured hierarchy (Raczek et al, 1998; Fitzpatrick et al, 1998a). This analysis identifies gaps in the domain described by the instrument, placing greater emphasis on the ability to describe a range of difficulties so as to maximise information as opposed to maximising internal consistency reliability (Nunnally and Bernstein, 1994; Raczek et al, 1998). Rasch models analysis may therefore improve instrument performance whilst producing a linear, interval-level score which is easier to interpret and more precise than the ordinal level of measurement produced by traditional summated rating scales (Raczek et al, 1998). It may also more accurately estimate scores when data is missing by utilising expected response information for each item, as opposed to substituting a person-specific estimate (Raczek et al, 1998). However, this analysis is only suitable for multi-item instruments decribing the range of a single domain. For example, the ASQoL and the RLDQ. It is not clear that the BASDAI contains items covering the range of AS-specific symptomlogy. The Body Chart, PGI-AS and EuroQol would not be suitable because they incorporate explicit weightings and certain domains of the SF-12 are described by a single item. However, separate scales of the SF-36 have been assessed by Rasch analysis supporting the unidimensional and hierarchical nature of the physical functioning scale (Fitzpatrick et al, 1998a; Raczek et al, 1998). The RLDQ is currently being assessed by Rasch models analysis to evaluate the extent to which items cover the range of AS functional disability (Helliwell P.- personal communications, 2000).

Data interpretation may be further enhanced by describing a score range against which real change may be assessed. Calculating the 95% limits of agreement as an estimation of test-retest reliability describes a range of values that is expected to describe the agreement between two observations for most patients indicating no change in health (Bland and Altman, 1986; Altman, 1996). The range acknowledges that few repeat observations will be identical due to random error. However, there has not been a wide spread adoption of this methodology and few authors describe what they consider to be a minimally important change in instrument score. Therefore, interpretation of the limits described is difficult.

The elderly (over 75 years) and children (less than 18 years) were excluded from the current study for reasons related to difficulty in instrument comprehension and self-completion, and due to the impact of co-morbidity (elderly) or the juvenile expression of AS (children) on the data. However, AS does not markedly reduce life expectancy and with the ageing population many patients in routine practice will be older than 75 years. To obtain valid information from this important section of the population the acceptance of questionnaires by these patients must be addressed. Likewise, the ability to clearly record the impact of disease in children is important. In circumstances where self-completed patient-based instruments are not acceptable the use of proxy respondents in the form of carers or health care professionals has been described (Guyatt et al, 1993; Garratt, 1997). In the current research several patients were unable to complete the questionnaire due to blindness caused through persistent iritis, a severe complication of AS (Dziedzic, 1998). These patients may represent a section of the population with a high risk for poor outcome and the role of proxy respondents requires further investigation.

Patients unable to comprehend the written English language were also excluded from the study and may reduce the generalisability of the result. The ability to apply patient-based instruments cross-nationally supports meta-analysis of data from multi-national studies and supports comparison of results across countries and patient-groups (Anderson et al, 1995; Ware and Gandek, 1998b,c). Several disease-specific instruments have evidence of formal translation into a non-English format, and initiatives to support the translation and cross-cultural adaptation of generic measures of HRQL have been identified: the International Quality of Life Assessment (IQoLA:

SF-36; Ware and Gandek, 1998a); World Health Organisation Quality of Life (WHOQoL)(Anderson et al, 1995).

The non-response bias of the current research towards the younger age group, a feature described by other investigators (Garratt et al, 1993), and the return of several questionnaires by patients describing a perception that their AS was too mild to be of benefit to the study, suggests that the generalisability of the result may be reduced and focussed towards older patients with more severe disease. The impact of this form of bias on the evaluation of instrument measurement properties is not clear, but may not be as important as non-response bias when information from outcomes is used to evaluate the effectiveness of competing interventions, and ultimately to direct resource allocation within the health care system (Garratt, 1997). It was not possible to contact a random sample of non-responders, for example, by telephone, due to the limited study resources. However, this may have provided beneficial information relating to AS and general health and further supported interpretation of the non-responder bias. Providing patients the opportunity to indicate that they experience minimal symptoms, and directing these patients to complete the generic instruments or a summary of items may enhance data retrieval from patients otherwise lost to the study.

8.6 Conclusion

The thesis makes three important contributions to the evidence base for outcome measurement in AS. First, the study was in a unique position to describe the first evidence-based and systematic selection of patient-based and anthropometric measures to fulfil a defined selection of domains considered important in the evaluation of AS. Secondly, the development of the first AS-specific individualised measure of disease-related quality of life, the PGI-AS, is described. Thirdly, the first comparative evaluation of the PGI-AS and an evidence-based selection of disease-specific, anthropometric and generic measures of outcome in patients with AS was performed.

The methodology adopted for the systematic review and evaluation may provide a format whereby the wide range of outcome measures available in other diseases can be identified and explicitly appraised to support instrument recommendation. For example, in neurology, respiratory care or oncology. Alternatively, the results may

highlight gaps in the availability of evaluative instruments for specific domains of health, or the need for further research to evaluate the measurement properties of widely used instruments, as described in the current research.

Systematic reviews of RCTs consider the quality of measures of outcome when assessing the quality of trials included in the review (Jadad et al, 1998). However, no previous attempt to systematically and explicitly assess the quality of measures of outcome has been described. This study describes the first attempt to describe instrument quality within a specific disease, and should be considered an important addition to systematic reviews of RCTs in AS. The methodology described may support the development of guidelines for future reviews of patient-based and anthropometric measures of outcome.

The first comparative evaluation of a wide ranging selection of outcome measures in AS supports appreciation of instrument measurement properties, acceptability and feasibility, and the recommendation of instruments to include, and not to include in the evaluation of groups in routine practice or clinical research. Unfortunately, no instrument had sufficient measurement properties to support recommendation for use in individual evaluation and no clear recommendation for the disease-specific evaluation of HRQL, disease activity, functional disability or spinal mobility could be made.

Following suggested modifications, the next step for the current research is to address the feasibility of including the identified package of instruments in routine practice or clinical research, and a re-evaluation of measurement properties and practical criteria following these applications. It is not proposed that all instruments should be included in clinical practice or research due to the burden of completion and administration. The list is not intended to be prescriptive, but to provide an evidence-based and standardised approach to evaluation in AS. The selected instruments describe different although often complimentary domains of health, and instrument selection for use in routine practice or research will be influenced by several factors: the multi-dimensional nature of AS; the objectives of the clinician, investigator or possibly the patient; the objectives of the intervention or management; the patient population; and available resources for administration, scoring and data interpretation (Read, 1987). Application of the package of instruments in a clinical trial will

provide further information on the performance of instruments and support data interpretation on a range of scores between different arms of a trial.

Although instrument administration in the clinic-based survey aimed to reflect the demands and requirements of routine clinical practice (Dziedzic, 1998; Lubrano et al, 1998), the role of patient self-completion of instruments prior to a consultation and the feasibility of scoring, interpretation and inclusion of these results within the consultation period requires further appraisal. The role of postal self-completion of instruments in routine practice may also be worthy of further investigation, but the costs and benefits of such an exercise would also require careful evaluation.

However, the question remains 'Why?' at the start of the 21st Century in patients with chronic and incurable disease such as AS for which the HRQL and psychosocial impact of disease have been described as major considerations (Barlow et al, 1993a,b), does evaluation in clinical practice appear to remain focused towards recording impairment? Whereas, evaluation in clinical trials has for the last decade included patient-based evaluation addressing wider issues of HRQL (after Brooks and Kamberg, 1987).

This chapter has identified several issues that may support the reasons for poor acceptance of patient-based instruments in routine practice. Despite the increasing evidence to support the measurement properties of patient-based instruments and the increasing importance of including patient-based assessment in evaluation, clinicians often cite a lack of intuitive feeling for the interpretation of the data (Fitzpatrick et al, 1992) and a perceived lack of relevance of these instruments to patient management and intervention (Chesson, 1998).

Improving the awareness of health care professionals to patient-based measures of outcome and facilitating the incorporation into routine practice is an important requirement if these instruments are to gain acceptance. Through the combined input of several chartered physiotherapists with a specialist interest in outcome measurement, including the lead investigator of the current research (KLH), the Chartered Society of Physiotherapy (CSP), the professional body of physiotherapists in the UK, has produced a database of outcome measures for access by the membership. The database is housed at the National Institute for Clinical Excellence

(NICE) and can be accessed by the Internet (http://www.nice.org.uk), by post or by telephone. The main objective of the site was to provide a facility where up to date information relating to available patient-based instruments, their measurement properties, acceptability, feasibility and cost of administration could be accessed by clinicians interested in incorporating these instruments into routine practice. To have an increased relevance to physiotherapists in routine practice, as opposed to research, the database has focussed on measures of disability, handicap and HRQL. Feedback to the CSP indicated that further information relating to instrument selection, incorporation into routine practice and score interpretation was required, and several workshops have been proposed to support the acquisition of this knowledge.

Further work to evaluate the most effective approach supporting the integration of patient-based instruments into routine practice and to support standardisation in routine practice and research is required. From a starting point of a package of evidence-based evaluative instruments which describe domains of health considered important in the evaluation of HRQL focus group interviews with clinicians and patients may provide additional information to support this process. However, evidence of the benefit to clinical decision making, resource allocation and health policy that may be gained by the inclusion of these instruments in routine practice is also required (Greenhalgh and Meadows, 1999). Following initial evidence of satisfactory measurement properties, the role of individualised measures such as the PGI-AS in routine practice and research, in clinical decision making and directing patient-centred management also demands further investigation.

Appendix 1 Data extraction sheet

Publication details	**Review number:**
	Author (s) Article title Source (journal or conference), year, volume, part, pages Country of origin Institutional affiliation (first author) and contact address
Outcome measure	
Stage of development described by article	I, II, III
Name of measure	
Definition of purpose	
Conceptual base	
General description and scale structure	General description Number of items Length Response possibilities (ordinal / interval / nominal / yes - no / VAS) Method of administration and scoring (time and training required) Item development Sub-scales Capacity / performance based Application (clinical / research / survey / screening) Time specificity Comments
Study design	
Type of study	Study design Aims clarified Sample size; do numbers add up at end of study period? Study duration - follow-up and adequacy of - were relevant outcome measures ignored?
Population investigated	Patient population Setting and methods for recruitment Inclusion and exclusion criteria Diagnosis and co-morbidity Disease duration (symptoms and diagnosis) and severity Patients broken down by age (range, mean, SD), sex, other variables
Reliability testing	
Internal consistency reliability (multi-item scales)	Item total correlation Overall correlation between items – Cronbachs alpha Result
Test-retest reliability	Test-retest; intra-observer; inter-observer Number of patients – who; how selected Number of observers – who; how selected; experience / training Number of repetitions Retest period Blinding of observers / patients Time of day Which measurement recorded? – mean; last attempt Statistical tests; description of method Result
Validity testing	
Construct validity	Use of 'extreme groups' Hypothesis stated a priori (or implicit) Correlation of measurements: physical tests / signs, pain measures, psychosocial findings, disability assessments, other. Quality of measures correlated against Statistical tests applied Result
Responsiveness testing	
Longitudinal validity	Use of Health Transition Questions
Used in trial of known efficacy	Effect size statistic Correlation of scale change with changes in other measures
Acceptability / Practicality	
	Application - research; clinical practice; other (time, convenience, confidence, training, cost, etc.) Acceptability to clinician / patient
Use in published articles	Original developers Others
Commentary	

Appendix 2 The Patient-Generated Index for AS (PGI-AS)

Your answers to the following steps will tell us how your life is affected by your Ankylosing Spondylitis and how you would like to see your life improved.

STEP 1: Identifying Areas

We would like you to think about the **most important areas of your life** that are affected by your **ANKYLOSING SPONDYLITIS (AS)**

Please write up to **FIVE** areas in the boxes below. **You don't have to write an area in each box and they don't have to be in order of importance.**

The last two boxes have been completed for you

On the next page is a list of 'important areas' mentioned by other people with AS. There is also an example of a completed form. You may find this helpful when filling in the boxes.

Areas affected by health problems other than Ankylosing Spondylitis

All other non-health areas of Your life

STEP 2: Scoring Each Area

In this part we would like you to score the areas you mentioned in step1. This score should show how badly affected you were over the past **MONTH**.
Please score each area out of 10 using this scale:

10 = Exactly as you would like to be
9 = Close to how you would like to be
8 = Very good, but not how you would like to be
7 = Good, but not how you would like to be.
6 = Between good and fair.
5 = Fair
4 = Between poor and fair
3 = Poor, but not the worst you could imagine
2 = Very poor, but not the worst you could imagine
1 = Close to the worst you could imagine
0 = The worst you could imagine

Please don't forget to score the last two boxes

You must score the last box

STEP 3: Spending Points

We want you to imagine that any or all of the areas of your life could be improved.

You have '**14**' imaginary points to spend to show which areas you would most like to see improved.

Spend more points on areas you would most like to see improve and less on areas that are not so important.

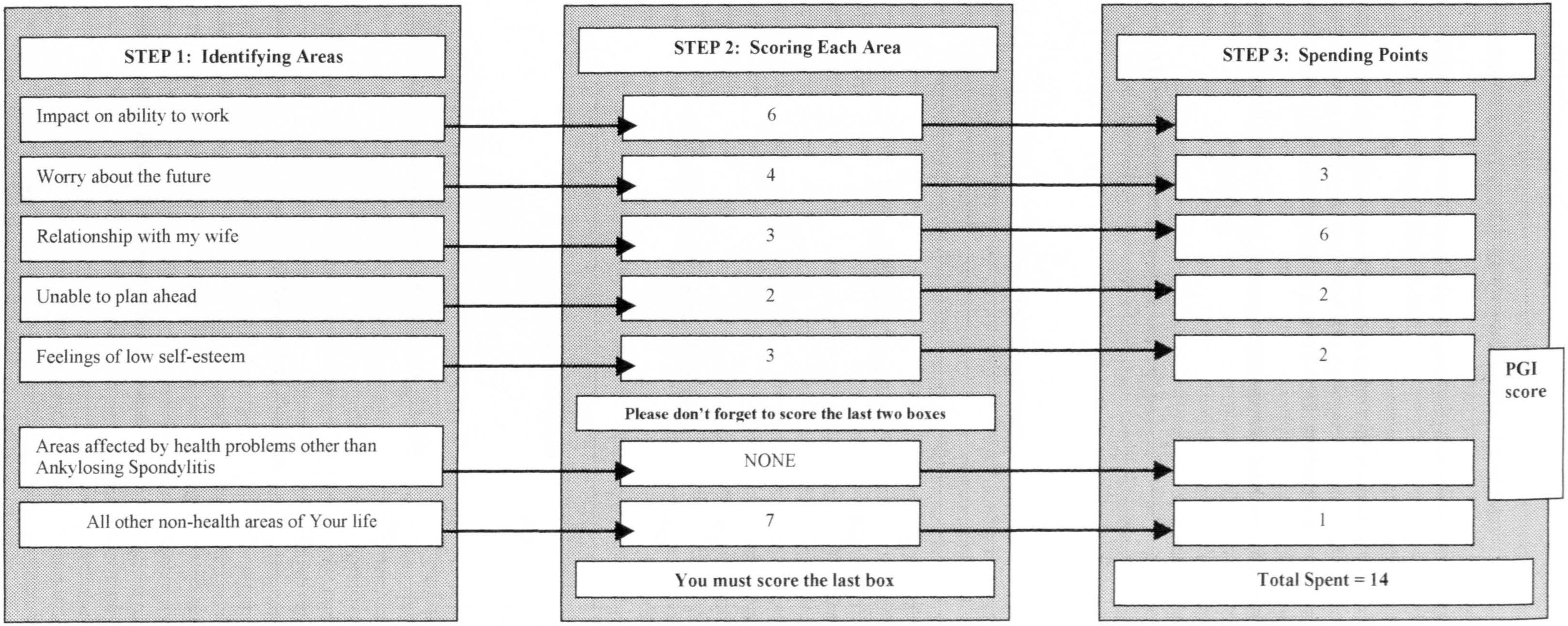
You don't have to spend points in every area.
You can't spend more than **'14'** points in total.
Remember the total for this column must add up to '14'

PGI score

Total Spent = 14

TRIGGER LIST OF AREAS COMMONLY MENTIONED BY PEOPLE WITH ANKYLOSING SPONDYLITIS (AS):

Relationship with partner, Ability to play with children, Sex life, Family life, Pain, Disturbed Sleep, Difficulty 'getting going' in the morning, Walking, Difficulty sitting down/ standing/ lying down, Ability to remain physically active, Fear of Falling, Increased time to do things, Control over life, Ability to plan ahead, Enjoyment of life, Worry over 'letting people down', Level of Independence, Dressing, Washing, Ability to do jobs around the home, Impact on Work, Relationship with friends, Social Life, Feelings of low self-esteem, Embarrassment, Poor self body-image, Worry about the future, Fatigue, Feeling Tired, Loss of motivation, Depression, Moody, Pursuing chosen hobbies, Sporting activities, Driving, Limited spinal movement, Mental activity.



Appendix 3 Patient-based Measures of Outcome

Ankylosing Spondylitis Quality of Life Questionnaire (ASQoL)

On the following pages you will find some statements which have been made by people who have Ankylosing Spondylitis.

Please read each statement carefully. We would like you to tick **'Yes'** if you feel that statement applies to you and tick **'No'** if it does not.
Please choose the response that best applies to you **AT THE MOMENT**

Please read each item carefully and tick the *one* response that applies best to you *at the moment.*

1. My condition limits the places I can go. — Yes ☐ No ☐
2. I sometimes feel like crying. — Yes ☐ No ☐
3. I have difficulty dressing. — Yes ☐ No ☐
4. I struggle to do jobs around the house. — Yes ☐ No ☐
5. It's impossible to sleep. — Yes ☐ No ☐
6. I am unable to join in activities with my friends / family. — Yes ☐ No ☐
7. I am tired all the time. — Yes ☐ No ☐
8. I have to keep stopping what I am doing to rest. — Yes ☐ No ☐

Please read each item carefully and tick the <u>one</u> response that applies best to you <u>at the moment.</u>

9. I have unbearable pain.
   - Yes ☐
   - No ☐

10. It takes a long time to get going in the morning.
   - Yes ☐
   - No ☐

11. I am unable to do jobs around the house.
   - Yes ☐
   - No ☐

12. I get tired easily.
   - Yes ☐
   - No ☐

13. I often get frustrated.
   - Yes ☐
   - No ☐

14. The pain is always there.
   - Yes ☐
   - No ☐

15. I feel I miss out on a lot.
   - Yes ☐
   - No ☐

16. I find it difficult to wash my hair.
   - Yes ☐
   - No ☐

17. My condition gets me down.
   - Yes ☐
   - No ☐

18. I worry about letting people down.
   - Yes ☐
   - No ☐

Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)

PLEASE PLACE A MARK ON EACH LINE BELOW TO INDICATE YOUR ANSWER TO EACH QUESTION, RELATING TO THE PAST WEEK

1. How would you describe the overall level of fatigue / tiredness you have experienced?

NONE _____ VERY SEVERE

2. How would you describe the overall level of AS neck, back or hip pain you have had?

NONE _____ VERY SEVERE

3. How would you describe the overall level of pain / swelling in joints other than neck, back or hips you have had?

NONE _____ VERY SEVERE

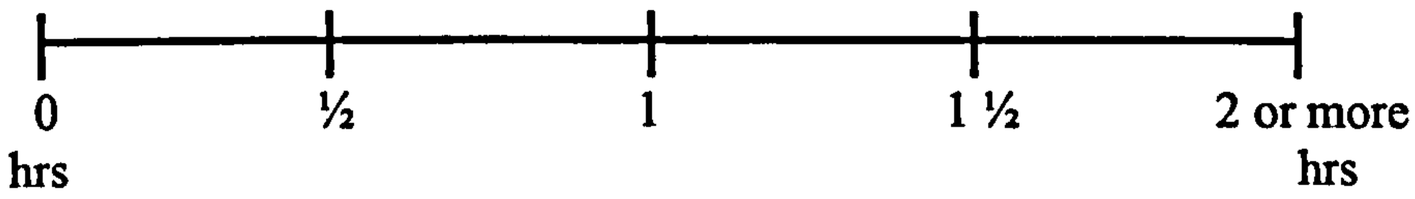
4. How would you describe the overall level of discomfort you have had from any areas tender to touch or pressure?

NONE _____ VERY SEVERE

5. How would you describe the overall level of morning stiffness you have had from the time you wake up?

NONE _____ VERY SEVERE

6. How long does your morning stiffness last from the time you wake up?



The Body Chart

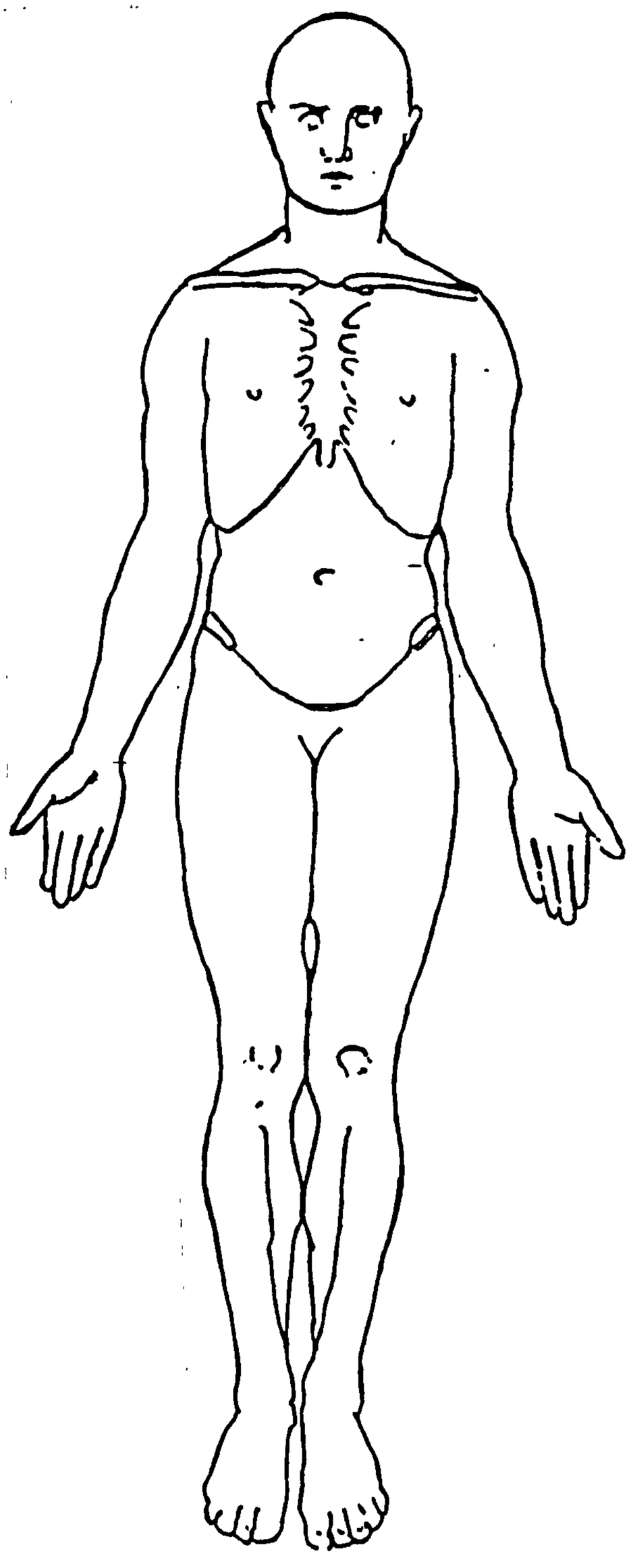
The following two line drawings represent a 'body chart'. The pictures illustrate the front and the back of a person. Please look at these line drawings.

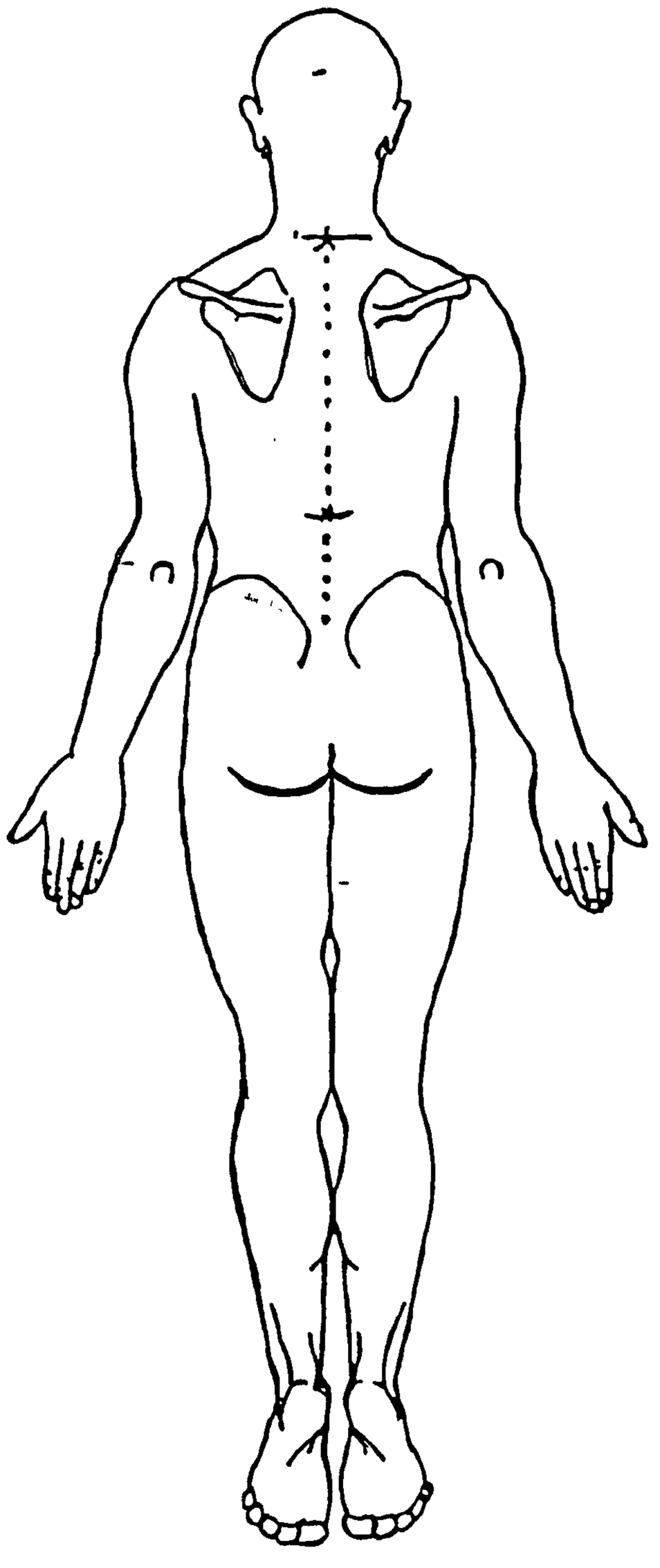
Using a pen please shade in the body chart to show the area or areas where you are experiencing pain.
This refers to your current or present pain.

Now score these areas of pain as:

1 = mild pain
2 = moderate pain
3 = severe pain
4 = very severe pain

Please make sure that each individual area or areas of pain are scored.





Revised Leeds Disability Questionnaire

PLEASE TICK THE ONE RESPONSE WHICH BEST DESCRIBES YOUR ABILITIES
OVER THE PAST WEEK

PLEASE PAY CAREFUL ATTENTION TO COLUMN 3 (USING UNUSUAL MOVEMENTS). For example, if you only get out of a car by pulling yourself up with your hand on the roof, then tick this column in response to question 1b.

	Able to do without difficulty	Able to do with difficulty	Only able to do using unusual movements or gadgets	Unable to do
1. Mobility				
a. Getting into and out of the bath				
b. Getting into and out of the car				
c. Getting up and out of bed in the morning				
d. Rolling over in bed				
2. Bending Down				
a. Wiping yourself after using the toilet				
b. Putting on and taking off your socks				
c. Putting on your shoes and tying your laces				
d. Cutting your toe nails				
3. Neck Movements				
a. Opening high windows				
b. Looking both ways before crossing the road (e.g. do you have to move your feet)				
c. Looking at what you are reaching on a high shelf				
d. Drinking from a small glass or can (e.g Do you have to bend your knees?)				
4. Posture				
a. Walk on your heels				
b. Coughing or sneezing				
c. Sleep on your back				
d. Sleep on your stomach				

Total score

The EuroQoL - EQ-5D

By placing a tick (' ✓') in one box in each group below, please indicate which statement best describes your own health state **today.**

Do not tick more than one box per question.

1. **Mobility:**
   - I have no problems in walking about ☐
   - I have some problems in walking about ☐
   - I am confined to bed ☐

2. **Self-Care:**
   - I have no problems with self-care ☐
   - I have some problems washing or dressing myself ☐
   - I am unable to wash or dress myself ☐

3. **Usual Activities** (e.g. work, study, housework, family or leisure activities):
   - I have no problems with performing my usual activities ☐
   - I have some problems with performing my usual activities ☐
   - I am unable to perform my usual activities ☐

4. **Pain / Discomfort:**
   - I have no pain or discomfort ☐
   - I have moderate pain or discomfort ☐
   - I have extreme pain or discomfort ☐

5. **Anxiety / Depression:**
   - I am not anxious or depressed ☐
   - I am moderately anxious or depressed ☐
   - I am extremely anxious or depressed ☐

Compared with my **general level of health** over the **past 6-months**, my health state today is:

Please tick one box:

- Better ☐
- Much the same ☐
- Worse ☐

The EuroQol - Thermometer

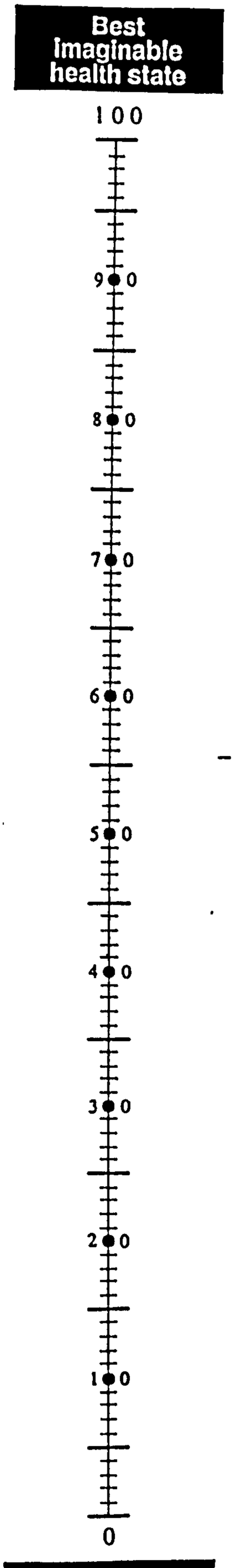
Your own health state today

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked by 100 and the worst state you can imagine is marked by 0.

We would like you to indicate on this scale **how good or bad is <u>your own health today</u>, in your opinion.**

Please do this by <u>drawing a line from the box below</u>, to whichever <u>point on the scale indicates how good or bad your current health state is **today**</u>.

Your own health state TODAY



Please do not write in this box:

Short-Form 12-Item Health Survey Questionnaire (SF-12)

The final twelve questions ask for your views about your health. This information will help keep track of how well you are able to do your usual activities.

Please answer every question by marking one box. If you are unsure about how to answer, please give the best answer you can.

1. In general, would you say your health is:

☐	☐	☐	☐	☐
Excellent	**Very Good**	**Good**	**Fair**	**Poor**

The following items are about activities you might do during a typical day. Does <u>your health now limit you</u> in these activities? If so, how much?

	Yes, Limited A Lot	**Yes, Limited A Little**	**No, Not Limited At All**
2. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling or playing golf.	☐	☐	☐
3. Climbing **several** flights of stairs	☐	☐	☐

During the <u>past 4-weeks</u>, have you had any of the following problems with your work or other regular daily activities <u>as a result of your physical health?</u>

	YES	**NO**
4. **Accomplished less** than you would like	☐	☐
5. Were limited in the **kind** of work or other activities	☐	☐

During the <u>past 4-weeks</u>, have you had any of the following problems with your work or other regular daily activities <u>as a result of any emotional problems</u> (such as feeling depressed or anxious) ?

	YES	**NO**
6. **Accomplished less** than you would like	☐	☐
7. Didn't do work or other activities as **carefully** as usual	☐	☐

8. During the past 4-weeks, how much did pain interfere with your normal work (including both work outside the home and housework) ?

☐ **Not at all** ☐ **A little bit** ☐ **Moderately** ☐ **Quite a bit** ☐ **Extremely**

These questions are about how you feel and how things have been with you during the past 4-weeks. For each question, please give the one answer that comes closest to the way you have been feeling.
How much of the time during the past 4-weeks –

	All of the Time	**Most of the Time**	**A Good Bit of the Time**	**Some of the Time**	**A Little of the Time**	**None of the Time**
9. Have you felt calm and peaceful?	☐	☐	☐	☐	☐	☐
10. Did you have a lot of energy?	☐	☐	☐	☐	☐	☐
11. Have you felt downhearted and blue?	☐	☐	☐	☐	☐	☐

12. During the past 4-weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.) ?

☐ **All of the time** ☐ **Most of the time** ☐ **Some of the time** ☐ **A little of the time** ☐ **None of the time**

Appendix 4 Ethical Approval

North Staffordshire Health Authority
Heron House
Great Fenton Business Park
Grove Road
Stoke-on-Trent
Staffordshire ST4 4LX
Tel: 01782 298000
Fax: 01782 298298
Minicom Tel: 01782 298035

North Staffordshire Health

PLEASE REPLY TO: RESEARCH ETHICS COMMITTEE, HUMAN RESOURCES DIRECTORATE, ROYAL INFIRMARY, HARTSHILL, STOKE-ON-TRENT, ST4 7PS

Ext. 4530 (Direct Dial 01782 554530)

VGH/JEC

1st June 1998

PRIVATE AND CONFIDENTIAL

Ms. K.L. Haywood
12 Haig Road
Catterick Garrison
Richmond
North Yorkshire
DL9 3AH

Dear Ms. Haywood

Project 884
The evaluation of health outcome in Ankylosing Spondylitis

I am pleased to inform you that the above project was approved at the meeting of the Research Ethics Committee on the 27th May 1998.

Yours sincerely

Dr. V.G. Hughes
Chairman
Local Research Ethics Committee

Appendix 5 Clinic Survey - Patient information and consent form

All Letters printed on appropriately headed paper.

Dr Peter Dawes, Consultant Rheumatologist.
Date:

Dear .....

How does Ankylosing Spondylitis Affect You?

We are interested in finding out how peoples day-to-day lives are affected by AS.

The Staffordshire Rheumatology Centre and the University of York are carrying out an important study of people with Ankylosing Spondylitis (AS). The research will involve a research physiotherapist (Kirstie Haywood), Jackie Waterfield (Senior Physiotherapist) and myself (Dr. Peter Dawes, Consultant Rheumatologist).

We are writing to you to ask if you would please consider taking part in this study. Your participation will help to improve our understanding of the way in which people with AS are affected by the disease, and how this can be measured.

The research physiotherapist would like to meet with you when you attend for your pre-arranged clinic appointment with the doctor at the Haywood Hospital out-patients rheumatology clinic. Your routine appointment with the doctor will not be affected should you be happy to take part in the research study.

Should you be happy to take part in the research study we would like you to attend the clinic 30-minutes before your pre-arranged clinic appointment time. During this time we would like you to complete a questionnaire. The questionnaire will ask you about your AS, the areas of your life affected by the disease, and the importance of these areas to you. In addition, the physiotherapist would like to take several measurements of your back and neck movement. These will be similar to the measurements taken when you attend for a normal clinic appointment with the physiotherapist. A gown will be provided for your comfort, or you may wish to bring a pair of shorts with you. These measurements may be taken following your assessment with the doctor and will take about 10-minutes. All information will be completely confidential and will only be used for research purposes. The research will not alter any part of your usual care and does not require you to come into hospital for any additional appointments.

If you are able to take part in the study you will be invited to attend for a follow-up assessment when you attend for your next 'routine' clinic assessment in six-months time. This assessment will follow the same format as the initial research assessment. If you choose to withdraw from the study you may do so freely at any time without having to give your reasons.

If you feel able to take part in the study please complete the enclosed 'consent form' to indicate your willingness to meet with the physiotherapist for the research assessment. Once completed, the form should be returned to the physiotherapist in the reply-paid envelope. Please keep this letter for your future reference and make a note of your appointment time.

Please be assured that you are under no obligation to take part in the study. If you choose not to take part in the study please still attend the clinic for your pre-arranged clinic appointment. If you choose not to take part in the study, your usual management will not be affected. Even if you are unable to assist in this study we would be grateful if you would return the 'consent form' in the reply-paid envelope to indicate you would not like to take part in the research assessment.

We hope that you will feel able to help us with this important study and look forward to your response. If you would like to know more about this study please do contact the research physiotherapist (Kirstie Haywood) on the following number: xxxxxxx (daytime, evening and weekends).

Yours sincerely,

Dr. Peter Dawes.

Consultant Rheumatologist, Staffordshire Rheumatology Centre.

Routine Clinic Appointment: Wednesday ..... 1999:Time: .....

Research Assessment: Please arrive **30-minutes** before your clinic appointment:
Time: .....

Should you decide to take part in the study, please indicate to the receptionist when you arrive at the clinic that you are taking part in the 'AS Study' with Kirstie Haywood (Research Physiotherapist).

Clinic Study: Informed Consent Form

Name: ..... Date of Birth: .....

Address: .....

.....

Post Code: ..... Telephone Number: .....

- **Are you able to take part in the research study when you attend for your next routine clinic appointment?**
  Please place a '✓' in the chosen box:

Routine Clinic Appointment: Wednesday .....1999: Time: .....

Research Assessment: Please arrive **30-minutes** before your clinic appointment:
Time: .....

Yes – I am happy to take part in the study ☐

No – I do not wish to take part in the study ☐

If 'NO':
Many thanks for your time and consideration.
Please return this page in the reply-paid envelope.

Please still attend the clinic for your 'routine' clinic appointment.

If 'YES':

- Please sign below to confirm your willingness to participate in the study before returning the completed form.

- I ..... of the above address hereby fully and freely consent to participate in the study: **'How does Ankylosing Spondylitis Affect You?'.**

- I understand and acknowledge that the study is designed to improve medical understanding of the way in which people with AS are affected by the disease, and how this can be measured.

- I note that I may withdraw my consent to take part at any stage in the study.

- I have received a written explanation of the study and understand the requirements of my participation.

Signed: ..... Date: .....

Many thanks for your assistance.
The information that you provide will be very helpful to us.

Please return this page in the pre-paid envelope.

Clinic - Patient Information Leaflet

How Does Ankylosing Spondylitis Affect You?

We are interested in finding out how peoples day-to-day lives are affected by AS. We are seeking people with AS to take part in this study. Before deciding whether you would like to take part, please read this information leaflet carefully. It tells you about the study and what you would be expected to do.

What is the study for?

Your participation in the study will help to improve our understanding of the way in which people with AS are affected by the disease, and how this can be measured.

What will I be required to do if I take part in the study?

If you are happy to take part in the study you will be invited to attend the Haywood Hospital rheumatology clinic on two separate occasions to participate in a research assessment. The research assessment will follow a similar format to the assessment when you attend for a normal clinic appointment. The assessment will take approximately 45-minutes. During this time we would like you to complete a questionnaire. The questionnaire will ask you about your AS, the areas of your life affected by the disease, and the importance of these areas to you. In addition, the physiotherapist would like to take several measurements of your back and neck movement. These will be similar to the measurements taken when you attend for a normal clinic appointment. A gown will be provided for your comfort, or you may wish to bring a pair of shorts with you.

If you are able to take part in the study you will be invited to attend for a follow-up assessment in six-months time. This assessment will follow the same format as the initial research assessment.

Who will be taking part in the study?

We are writing to over 120 people with diagnosed AS to ask about the disease and how it affects their day-to-day life. The study would not be suitable for pregnant women. If you are pregnant you should not take part in the study, but please do return the consent form to prevent further letters being sent to you.

What are the risks/benefits to me from taking part in the study?

There are no known risks or side effects from participating in the research study. The study will not alter your usual care in any way.

What about confidentiality?

All information will be deemed to be confidential and your identity will not be made known to other individuals. In all instances your confidentiality will be respected and maintained.

Can I withdraw from this study?

Yes. You may withdraw from the study freely at any time without having to give your reasons.

Further information

The research will involve the research physiotherapist (Kirstie Haywood) and (Dr. Peter Dawes). If you choose to take part in the study further contact will be made with you by Kirstie Haywood (Research Physiotherapist). Should you have any questions or problems about this study, please do contact: Kirstie Haywood, Research Physiotherapist, on *xxxxxxxx*.

Appendix 6 Postal Survey - Database Questionnaire (Consultant completion)

The Evaluation of Health Outcome in Ankylosing Spondylitis - Postal Survey

I would be very grateful if you would complete and return the following questionnaire. A stamped-and-self-addressed-envelope has been included for your reply.

The information is intended to determine the accessibility of a random sample of patients with diagnosed AS under your care.

Name: *Consultant Rheumatologist.* **Hospital:** *Department of Rheumatology.*

i). Would you be willing for a random sample of subjects with AS under your care to be approached in a mail-base survey?

Yes ☐ No ☐

If **'no'**, please do not continue. *Many thanks for your time.*

If **'yes'**, please answer the following questions.

ii). Is there access to a database of subjects with AS within your department?

Yes ☐ No ☐

If **'no'**, please do not continue. *Many thanks for your time.*

If **'yes'**:

a). How often is this database updated?

Frequency (in weeks or months): Unsure ☐

b). In what format does this database exist?

c). Is the database easily accessible?

Yes ☐ No ☐

iii). Is there a clear indication of individual patient diagnosis within the database?

Yes ☐ No ☐

iv). Is it possible to identify the number of individuals with diagnosed AS (Modified New York Criteria – van der Linden et al, 1984) on the database?

Yes ☐ No ☐

v). How many patients with diagnosed AS (Modified New York Criteria) are there on the database?

Number: Unsure ☐

vi). Would it be possible to take a 'random sample' from the patients identified with diagnosed AS on the database?

Yes ☐ No ☐

vii). Would you be willing for me to identify a random sample from the identified database?

Yes ☐ No ☐

viii). Is it possible to identify the 'disease spectrum' of AS covered by the database?

Yes ☐ No ☐

ix). Is it possible to readily distinguish between 'newly diagnosed' and 'established' disease on the database?

Yes ☐ No ☐

x). Is it possible to readily distinguish between males and females on the database?

Yes ☐ No ☐

xi). Could you please indicate the process for obtaining ethical approval for the proposed postal evaluation.

Appendix 7 Rheumatology centres and contacts for postal survey

Rheumatology centre	Consultant contact	Physiotherapist contact
Addenbrooke's Hospital, Cambridge	Professor JSH Gaston	Mrs J Isaacson
Cannock Chase Hospital, Stafford	Dr T Price	Mrs C David Mrs L Preston
Glasgow Royal Infirmary, University Hospital Trust	Professor R Sturrock	Ms F Gough
South Tees Acute Hospitals NHS Trust, South Cleveland	Professor I Haslock Dr M Plant	Mrs K West
Southmead Hospital, Bristol	Dr P Creamer	Mrs R Lewis
Staffordshire Rheumatology Centre (SRC), Stoke-on-Trent	Dr PT Dawes	KLH

Appendix 8 Postal Survey - Patient information and consent form

All Letters printed on appropriately headed paper.

Name of Consultant Rheumatologist.
Date:

Dear .....

How does Ankylosing Spondylitis Affect You?

We are interested in finding out how peoples day-to-day lives are affected by AS. The *Department of Rheumatology, xxx Hospital* and the Staffordshire Rheumatology Centre are carrying out an important study of people with Ankylosing Spondylitis (AS). The research will involve a research physiotherapist (Kirstie Haywood), *local Physiotherapist contact and myself (Consultant Rheumatologist).*

I am writing to you to ask if you would please consider taking part in this study. Your participation will help to improve our understanding of the way in which people with AS are affected by the disease, and how this can be measured. We are writing to over 400 people with AS to ask if they will take part.

If you are able to help in this study please complete the enclosed questionnaire. The questions ask you about your AS, the areas of your life affected by the disease, and the importance of these aspects to you.

Once completed, the questionnaire should be returned to the research physiotherapist (Kirstie Haywood) at the Staffordshire Rheumatology Centre in the reply-paid envelope. All information will be completely confidential and will only be used for research purposes. The research will not alter any part of your usual care. The return of the completed questionnaire will indicate that you are happy for your name and address to be given to the physiotherapist for the purpose of the study.

If you are happy to take part in the study you will be invited to complete a second questionnaire 2-weeks after the return of the first questionnaire, with a final questionnaire being sent to you in 6-months time. Each questionnaire will be completed at home, which will take about 30-minutes, and then returned in the reply-paid envelopes provided. You will not be required to come into hospital for any additional appointments. If you choose to withdraw from the study you may do so freely at any time without having to give your reasons. If you

feel able to take part in the study please complete the attached 'consent form' and return it with the completed questionnaire. Please keep this letter for your future reference.

Please be assured that you are under no obligation to take part in the study. If you choose not to take part in the study, your usual care will not be affected. Even if you are unable to assist in this study I would be grateful if you would return the blank questionnaire and consent form in the reply-paid envelope to indicate that you would not like to take part.

If you would like to know more about this study please read the enclosed 'Patient Information Leaflet' or contact the physiotherapist on the telephone number indicated.

I hope that you will feel able to help with this important study and look forward to your response.

Yours sincerely,

Consultant Rheumatologist

Postal Evaluation – Informed Consent Form

How Does Ankylosing Spondylitis Affect You?

- **Are you able to help with the study by completing the enclosed questionnaire?**
  Please place a '✓' in the chosen box:

  Yes – I am happy to take part in the study ☐

  No – I do not wish to take part in the study ☐

If 'NO':

Many thanks for your time and consideration.
Please return this page and the 'blank' questionnaire in the reply-paid envelope. You do not need to write your name and address on this letter.

If 'YES':

- Please sign below to confirm your willingness to participate in the study, and complete your name and address, before returning this form with the completed questionnaire.

- I ..... of the below address fully and freely consent to participate in the postal study entitled: **'How does Ankylosing Spondylitis Affect You?'.**

- I understand and acknowledge that the study is designed to improve medical understanding of the way in which people with AS are affected by the disease, and how this can be measured.

- I note that I may withdraw my consent to take part at any stage in the study.

- I have received a written explanation of the study and understand the requirements of my participation.

Signed: .....Date: .....

Name: ..... **Date of Birth: .....**

Address: .....

.....

Post Code: ..... **Telephone Number:.....**

Many thanks for your assistance. The information that you provide will be very helpful to us.
Please return this page in the pre-paid envelope with the completed questionnaire

You will receive a second questionnaire in 2-weeks time.

Postal - Patient Information Leaflet
How Does Ankylosing Spondylitis Affect You?

We are interested in finding out how peoples day-to-day lives are affected by AS.

We are seeking people with AS to take part in this study. Before deciding whether you would like to take part, please read this information leaflet carefully. It tells you about the study and what you would be expected to do.

What is the study for?

Your participation in the study will help to improve our understanding of the way in which people with AS are affected by the disease, and how this can be measured.

What will I be required to do if I take part in the study?

If you are happy to take part in the study you will be invited to complete a questionnaire on three separate occasions. The first questionnaire is included with this letter. A second questionnaire will be sent to you 2-weeks after the return of the first questionnaire, with a final questionnaire being sent to you in 6-months time. Each questionnaire will be completed at home, which will take about 30-minutes, and then returned in reply-paid envelopes. You will not be required to come into hospital for any additional appointments.

Who will be taking part in the study?

We are writing to over 400 people with diagnosed AS to ask about the disease and how it affects their day-to-day life. The study would not be suitable for pregnant women. If you are pregnant you should not take part in the study, but please do return the blank questionnaire to prevent further questionnaires being sent to you.

What are the risks/benefits to me from taking part in the study?

There are no known risks or side effects from completing the questionnaires. The study will not alter your usual care in any way.

What about confidentiality?

All information we receive will be deemed to be confidential and your identity will not be made known to other individuals. In all instances your confidentiality will be respected and maintained. The return of a completed questionnaire and consent form will indicate that you are happy for your name and address to be released to the research physiotherapist for the purpose of the study.

Can I withdraw from this study?

Yes. You may withdraw from the study freely at any time without having to give your reasons.

Further information

The research will involve ourselves (*Physiotherapist and Consultant Rheumatologist*) and a research physiotherapist (Kirstie Haywood, Research Physiotherapist, Staffordshire Rheumatology Centre). If you choose to take part in the study further contact will be made with you by Kirstie Haywood. Should you have any questions or problems about this study, please do contact: *Physiotherapist, on: xxxxxxxx*

Appendix 9 PGI-AS Rating Scale

Step 1 'Identify Areas'

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Coding:

None At all	Standard instructions read out only. No repetition of instructions.
A little assistance	Repetition of elements of standard instructions required. For example, repeat reference made to the 'trigger list' (but no examples read out).
Moderate assistance	Increased repetition of standard instructions with increased reference to usefulness of 'trigger list' and the completed form.
Significant assistance	Increased reference to 'trigger list' and completed form. Possible minimal re-wording of standard instructions to improve understanding.
Maximum assistance	Maximum reference to 'trigger list' and completed form. Random examples from the trigger list and completed form read out aloud to patient to facilitate identification of 'most important areas of life'. Interviewer may add that the areas chosen by the patient may 'be quite personal and only have meaning to you, or it may be quite a big thing'. Re-wording of standard instructions to improve understanding. Prolonged period of time to identify areas.

Step 2 'Score Areas'.

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Coding:

None At all	Standard instructions read out only. No repetition of instructions.
A little assistance	Repetition of elements of standard instructions required. For example, repeat reference made to the scale (no repetition of standard example).

Moderate assistance	Increased repetition of standard instructions with increased reference to the use of the scale and the completed form. Moderate assistance is considered to have been offered if attention needs to be drawn to the scoring of the 'last two boxes', beyond that provided in the standard instructions.
Significant assistance	Increased reference to use of the scale and the example of the completed form. Possible minimal re-wording of standard instructions to improve understanding.
Maximum assistance	Maximum reference to the format of the scale and the possible assistance from the completed example. Random examples from the trigger list and completed form read out aloud to the patient to facilitate identification of 'most important areas of life' and scoring of these areas. Re-wording of standard instructions to improve understanding. Prolonged period of time to identify areas.

Step 3 'Spend Points'.

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Coding:

None at all	Standard instructions read out only. No repetition of instructions.
A little assistance	Repetition of elements of standard instructions required. For example, repeat reference made to the number of points to be spent.
Moderate assistance	Increased repetition of standard instructions with increased reference to the method and purpose of spending points, with reference to the completed example.
Significant assistance	Increased reference to method and purpose of spending points, the number of points to be spent, and the example of the completed form. Possible minimal re-wording of standard instructions to improve understanding.
Maximum assistance	Maximum reference to the reason for spending points, the number of points to be spent and the possible assistance from the completed form. Examples of how the points may be spent in relation to the areas identified by the subject will be made. Re-wording of standard instructions to affect understanding. Prolonged period of time to spend points.

Step 4 Overall: Clinician-reported level of assistance required to complete PGI-AS.

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Step 5 Time taken to complete PGI-AS:

☐	☐	☐	☐	☐	☐
5 mins	**6-10 mins**	**11-15 mins**	**16-20 mins**	**21-25 mins**	**>25 mins**

Step 6 Additional Patient Comments
(in relation to PGI-AS or other instruments)

PGI-AS Rating Scale (Clinic)

Date of PGI-AS Completion: _____ ☐☐☐☐

Number of Times PGI-AS completed _____

Step 1 'Identify Areas'

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Step 2 'Score Areas'.

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Step 3 'Spend Points'.

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Step 4 Overall: Clinician-reported level of assistance required to complete PGI-AS

☐	☐	☐	☐	☐
None at all	**A little assistance**	**Moderate assistance**	**Significant assistance**	**Maximum assistance**

Step 5 Time taken to complete PGI-AS

☐	☐	☐	☐	☐	☐
5 mins	**6-10 mins**	**11-15 mins**	**16-20 mins**	**21-25 mins**	**> 25 mins**

Step 6 Additional patient comments (in relation to PGI-AS or other questions)

Glossary

List of Abbreviations

Acromioclavicular Joint	ACJ
American Academy of Orthopaedic Surgeons	AAOS
American Rheumatology Association	ARA
Andrew Garratt	AG
Ankylosing Spondylitis	AS
Ankylosing Spondylitis Quality of Life Questionnaire	ASQoL
Arthritis Impact Measurement Scale	AIMS
Assessment in Ankylosing Spondylitis group	ASAS
Assessment in Ankylosing Spondylitis Questionnaire	ASAQ
Bath Ankylosing Spondylitis Disease Activity Index	BASDAI
Bath Ankylosing Spondylitis Functional Index	BASFI
Bath Ankylosing Spondylitis Global Index	BAS-G
Bath Ankylosing Spondylitis Metrology Index	BASMI
Bath Ankylosing Spondylitis Radiological Index	BASRI
Bath Disease Activity Index	Bath - DAI
Cervical rotation	C.rot
Cervical spine - 7th spinous process	C7
Chest expansion	C.exp
Computerised Tomographic Scan	CT
Correlation of Variation (%)	CV (%)
C-reactive protein	CRP
Dougadas Functional Index	DFI
Effect Size	ES
Erythrocyte Sedimentation Rate	ESR
Fingertip to floor distance (anterior flexion)	FFD
Forced Vital Capacity	FVC
Health Assessment Questionnaire - Spondyloarthropathies	HAQ-S
Health Related Quality of Life	HRQL
Human Leucocyte Antigen - B27	HLA-B27
Intraclass Correlation Coefficient	ICC
Jackie Waterfield (physiotherapist)	JW
Kirstie Louise Haywood (lead investigator)	KLH
Lateral Lumbar Flexion	LLF
Leeds Disability Questionnaire	LDQ
Lumbar Flexion Index	LFI
MacMaster and Toronto Questionnaire for AS / Patient Elicitation Technique	MACTAS / PET
Mental component summary scale (SF-36; SF-12)	MCS
Modified Schober Index (15cm)	MSI
Modified Standardised Response Mean	MSRM

Newcastle Enthesitis Index	NEI
Non-steroidal anti-inflammatory drug	NSAID
Occiput to wall distance	OWD
Outcome Measures in Rheumatology and Clinical Trials	OMERACT
Patient Generated Index	PGI
Physical component summary scale (SF-36; SF-12)	PCS
Portable Spinal Mobility Scale	PSMS
Quality of Life	QoL
Radiograph	XR
Randomised Controlled Trial	RCT
Research Advisory Group	RAG
Response Shift	RS
Revised Leeds Disability Questionnaire	RLDQ
Rheumatoid Arthritis	RA
Schedule for the Evaluation of Individual Quality of Life (direct weighting)	SEIQoL - DW
Short Form 36-item Health Survey Questionnaire	SF-36
Short Form 12-item Health Survey Questionnaire	SF-12
Smythe technique (S); upper / middle / lower	S: U / M / L
Staffordshire Rheumatology Centre	SRC
Standard deviation	SD
Standardised Response Mean	SRM
Stoke Ankylosing Spondylitis Spinal Score	SASSS
Sickness Impact Profile	SIP
Stoke Enthesitis Index	SEI
Temperomandibular Joint	TMJ
Thoracolumbar flexion	TLF / TLflex
Toronto Activities of Daily Life Questionnaire	TADLQ
Tragus to Wall distance	TWD
Visual Analogue Scale	VAS
World Health Organisation	WHO

References

Abbott CA, Helliwell, PS, Chamberlain MA.(1994) Functional assessment in Ankylosing Spondylitis - Evaluation of a new self-administered questionnaire and correlation with anthropometic variables. British Journal of Rheumatology. 33:1060-1066.

Adrichem, JAM, van der Korst, JK. (1973) Assessment of the flexibility of the lumbar spine - A pilot study in children and adolescents. Scandinavian Journal of Rheumatology. 2:87-91.

Albert E, Scholtz S. (1987) Immunogenetics and rheumatic disease. Clinical and Experimental Rheumatology. 5:S30-S31.

Albrecht GL. (1994) Subjective health assessment. In: Jenkinson C (Ed). Measuring health and medical outcomes. London: UCL Press. 7-26.

Altman DG. Practical Statistics for Medical Research. (1996) Chapman and Hall, London.

American Academy of Orthopaedic Surgeons. Joint Motion - Method of measuring and recording. 1966. Churchill Livingstone Press. Edinburgh.

American Rheumatology Association (ARA) Glossary Committee: Dictionary of Rheumatic Diseases. Vol.1: Signs and Symptoms. New York: Contact Press, 1982.

Anderson KL, Burckhardt CS. (1999) Conceptualisation and measurement of quality of life as an outcome variable for health care intervention and research. Journal of Advanced Nursing. 29(2):298-306.

Anderson RT, Aaronson NK, Wilkin D. (1995) Critical review of the international assessments of health-related quality of life: generic instruments. In: Shumaher SA, Benton R. (Eds). International Assessment of Health-Related Quality of Life: Theory, Translation, Measurement and Analysis. Rapid Communications of Oxford, Ltd. Oxford.

Armstrong RD, Laurent R, Panayi GS. (1984) A comparison of indoprofen and indomethacin in the treatment of Ankylosing Spondylitis. Pharmatherapeutica. 3(10): 637-641.

Aronson KJ. (1997) Quality of life among persons with Multiple Sclerosis and their caregivers. Neurology. January.48:74-80.

Astrand PO, Rodahl K. (1977) Textbook of Work Physiology. Physiological Bases of Exercise. New York, McGraw-Hill.

Averns HL, Oxtoby J, Taylor HG, Jones PW, Dziedzic K, Dawes PT. (1996a) Radiological Outcome in Ankylosing Spondylitis: use of the Stoke Ankylosing Spondylitis Spine Score (SASSS). British Journal of Rheumatology. April 35(4):373-6.

Averns HL, Oxtoby J, Taylor HG, Jones PW, Dziedzic K, Dawes PT. (1996b) Smoking and Outcome in Ankylosing Spondylitis. Scandinavian Journal of Rheumatology. 25(3):138-42.

Badley E, Wagstaff S, Wood PHN. (1984) Measures of functional ability (disability) in arthritis in relation to impairment of range of joint movement. Annals of the Rheumatic Diseases. 43:563-9.

Bakker C, Boers M, van der Linden S. (1993b) Measures to assess Ankylosing Spondylitis: taxonomy, review and recommendations. The Journal of Rheumatology. 20(10):1724-1732.

Bakker C, Hidding A, van der Linden S, Doorslaer E. (1994a) Cost effectiveness of group physical therapy compared to individualised therapy for Ankylosing Spondylitis. A randomised controlled trial. The Journal of Rheumatology. 21(2) 264-268.

Bakker C, Rutten-van-Molken M, Hidding A, van Doorslaer E, Bennett K, van der Linden S. (1994c) Patient utilities in Ankylosing Spondylitis and the association with other outcome measures. The Journal of Rheumatology. 21(7):1298-1304.

Bakker C, Rutten M, van Doorslaer E, Bennett K, van der Linden S. (1994c) Feasibility of utility assessment by rating scale and standard gamble in patients with Ankylosing Spondylitis. The Journal of Rheumatology. 21(2):269-274.

Bakker C, van der Linden SJ, van Santen-Hoeufft M, Bolwijn P, Hidding A. (1995) Problem elicitation to assess patient priorities in Ankylosing Spondylitis and Fibromyalgia. The Journal of Rheumatology. Jul.22(7):1304-10.

Band DA, Jones SD, Kennedy LG, Garrett SL, Porter J, Gay L, Richardson J, Whitelock HC, Calin A. (1997) Which patients with ankylosing spondylitis derive most benfit from an in-patient management program? The Journal of Rheumatology. Dec.24(12):2381-4.

Band D, Calin A. (1998) How does a generic health status measure (SF-36) compare with disease-specific instruments in relation to sensitivity to change in patients with Ankylosing Spondylitis (AS)? British Journal of Rheumatology. 37.Abstracts supplement 1:44.

Barlow JH, Macey SJ, Struthers G. (1992) Psychosocial factors and self-help in Ankylosing Spondylitis patients. Clinical Rheumatology. 11(2):220-225.

Barlow JH, Macey SJ, Struthers GR. (1993a) Health locus of control, self-help and treatment adherence in relation to ankylosing spondylitis patients. Patient Education and Counseling. May.20(2/3):53-66.

Barlow JH, Macey SJ, Struthers GR. (1993b) Gender, depression, and Ankylosing Spondylitis. Arthritis Care and Research. March. 6(1):45-51

Barlow JH, Barefoot J. (1996) Group education for people with arthritis. Patient Education and Counseling. April.27(3):257-67.

Barlow JH, Williams B, Wright C. (1996) The Generalised Self-Efficacy Scale in people with arthritis. Arthritis Care and Research. 9(3):189-196.

Batlle-Gualda E, Figuero M, Ivorra J, Raber A. (1996) The efficacy and tolerability of aceclofenac in the in the treatment of patients with Ankylosing Spondylitis : a multicenter controlled clinical trial. Aceclofenac Indomethacin Study Group. The Journal of Rheumatology. Jul.23(7):1200-6.

Beaton DE, Hogg-Johnson S, Bombardier C. (1997) Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. Journal of Clinical Epidemiology. 50(1):79-93.

Bell MJ, Bombardier C, Tugwell P. (1990) Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis and Rheumatism. 33:591-601.

Bellamy N. (1999) Clinimetric concepts in outcome assessment: the OMERACT filter. The Journal of Rheumatology. April.26(4):948-50.

Bellamy N, Buchanan WW, Esdaile J.M, Fam AD, Kean WF, Thompso, JM, Wells GA, Campbell J. (1991a) Ankylosing Spondylitis Antirheumatic Drug Trials I: Effects of standardisation procedures on observer dependent outcome measures. The Journal of Rheumatology. 18(11):1701-1708.

Bellamy N, Kaloni S, Pope J, Coulter K, Campbell J. (1998) Quantitative rheumatology: a survey of outcome measurement procedures in routine rheumatology outpatient practice in Canada. The Journal of Rheumatology. May.25(5):852-8

Bellamy N, Muirden KD, Brooks PM, Barraclough D, Tellus MM, Campbell J. (1999) A survey of outcome measurement procedures in routine rheumatology outpatient practice in Australia. The Journal of Rheumatology. Jul.26(7):1593-9

Bennett PH, Wood PHN. (Editors). Population Studies in the Rheumatic Diseases. Proceedings of the third International Symposium, New York (1966)(p.456). Excerpta Medica Foundation, Amsterdam: 1968.

Bennet K, Torrance G, Tugwell P. (1991) Methodological challenges in the development of utility measures of health-related quality of life in Rheumatoid Arthritis. Controlled Clinical Trials. 12:118S-128S.

Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. (1976) The Sickness Impact Profile : Conceptual formulation and methodology for the development of a health status measure. International Journal of Health Services. 6(3):393-415.

Beurskens AJ, de Vet H, koke AJ, van der Heijden GJ, Knipschild PG. (1995) Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine. 20(9):1017-1028.

Biasi D, Carletto A, Caramaschi P, Bambar, LM. (1996) An update on the Bath Ankylosing Spondylitis Disease Activity and Functional Indices (BASDAI, BASFI) : excellent Cronbach's alpha scores. The Journal of Rheumatology. Feb.23(2):407-8.

Bindman AB, Keane D, Lurie N. (1990) Measuring health changes among severely ill patients. Medical Care. 28(12):1142-1152.

Bjorner JB, Damsgaard MT, Watt T, Groenvold M. (1998) Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. Journal of Clinical Epidemiology. 51(11):1001-1011

Blake DJ, Maisiak R, Alarcon GS, Holley HL, Brown S. (1997) Sexual quality of life of patients with arthritis compared to arthritis free controls. The Journal of Rheumatology. 14:570-6.

Bland JM, Altman DG. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. I:307-310.

Boers M, Brooks P, Strand V, Tugwell P. (1999) Introduction: OMERACT IV: Fourth International Consensus Conference on Outcome Measures in Rheumatology. The Journal of Rheumatology. 26(1):199-200.

Borg G. Subjective effort in relation to physical performance and working capacity. Psychology: From Research to Practice. New York, Plenum Publishing, 1978.

Bowker P. (1998) Instrumented measurement: its joys and sorrows. Physiotherapy. April.84(4):187-89.

Bowling A. Measuring Disease. A review of disease-specific quality of life measurement scales. Open University Press, Buckingham, Philadelphia. 1996.

Bowling A. Research Methods in Health. Investigating health and health services. Open University Press, Buckingham, Philadelphia. 1997.

Braun J, Bollow M, Sieper J. (1998) Radiologic diagnosis and pathology of the spondyloarthropathies. Rheumatic Disease Clinics of North America. 24(4)November:697-735.

Brettle AJ, Long AF, Grant MG, Greenhalgh J. (1998) Searching for information on outcomes: do you need ot be comprehensive? Quality in Health Care. 7:163-167.

British Society for Rheumatology Handbook (BSR). Oxford University Press, London: 1995-1996.

Brooks RH, Kamberg CJ. (1987) General health status measures and outcome measurement: a commentary on measuring functional status. Journal of Chronic Diseases. 40.

Supplement:131S-136S.

Brown GMM, Dare CM, Smith PR, Meyers OL. (1987) Important problems identified by patients with chronic arthritis. South African Medical Journal. 72:126-128.

Bruton A, Conway JH, Holgate ST. (2000) Reliability: What is it, and how is it measured? Physiotherapy. Feb.86(2):94-99.

Burgos-Vargas R. (1990) Isolated juvenile onset HLA-B27 associated peripheral arthritis - reply. The Journal of Rheumatology. 17(4):568.

Calcraft B, Tildesley G, Evans KT, Gravelle H, Hole D, Lloyd N. (1974) Azapropazone in the treatment of Ankylosing Spondylitis: a controlled clinical trial. Rheumatology and Rehabilitation. 13(1):23-29

Calin A. Ankylosing Spondylitis. In: Kelly WN, Harris ED, Ruddy S, Sledge CB. (Eds). Textbook of Rheumatology. 2nd Edition. Philadelphia: WB Saunders, 1985:993-1005.

Calin A, Elswood J, Rigg S, Skevington SM. (1988) Ankylosing Spondylitis - An analytical review of 1500 patients: the changing pattern of the disease. The Journal of Rheumatology. 15(8):1234-1238.

Calin A, Edmunds, Kennedy LG. (1993) Fatigue in Ankylosing Spondylitis - Why is it ignored? The Journal of Rheumatology. 20(6):991-95

Calin A. (1995a) The Dunlop-Dottridge Lecture: Ankylosing Spondylitis: defining disease status and the relationship between radiology, metrology, disease activity, funciton and outcome. The Journal of Rheumatology. April.22(4):740-4.

Calin A. (1995b) The Individual with Ankylosing Spondylitis: defining disease status and the impact of the illness. (The Margaret Holroyd Essay). British Journal of Rheumatology. 34:663-672.

Calin A, Elswood J. (1990) Retrospective analysis of 376 irradiated patients with Ankylosing Spondylitis and non-irradiated controls. The Journal of Rheumatology. 16:1443-1445.

Calin A, Elswood J. (1990) A prospective nationwide cross-sectional study of NSAID Usage in 1331 Patients with Ankylosing Spondylitis. The Journal of Rheumatology. 17(6):801-803.

Calin A, Garrett S, Whitelock H, Kennedy GL, O'Hea J, Mallorie P, Jenkinson T. (1994) A new approach to defining functional ability in Ankylosing Spondylitis: the development of the Bath Ankylosing Spondylitis functional index. The Journal of Rheumatology. 21(12):281-2285.

Calin A, Jones SD, Garrett SL, Kennedy LG. (1995) Bath Ankylosing Spondylitis Functional Index (BASFI). British Journal of Rheumatology. Aug.34(8):793-4.*letter*

Calin A, Gueguen A, Nakache JP, Zeidler H. (1997) Discriminant capacity of clinical and biological variables in ankylosing spondylitis. Review of Rheumatology (English Edition). 64(B92):744

Calin A, Nakache J.P, Guegen A, Zeidler H, Mielants H, Dougadas M. (1999a) Defining disease activity in ankylosing spondylitis: is a combination of variables (Bath Ankylosing Spondylitis Disease Activity Index) an appropriate instrument? Rheumatology. 38:878-882

Calin A, Nakache JP, Gueguen A, Zeidler H, Mielants H, Dougadas M. (1999b) Outcome variables in ankylosing spondylitis: evaluation of their relevance and discriminant capacity. The Journal of Rheumatology. April.26(4):975-9.

Calin A, Mackay K, Santos H, Brophy S. (1999c) A new dimension to outcome: application of the Bath Ankylosing Spondylitis Radiology Index. The Journal of Rheumatology. April.26(4):988-92.

Calman KC. (1984) Quality of life in cancer patients - an hypothesis. Journal of Medical Ethics. 10: 124-127.

Campbell DT, Fiske DW. (1959) Convergent and discriminant validation by the multitrai-multimethod matrix. Psychological Bulletin. 56:81-105.

Carbon RJ, Macey MG, McCarthy DA. (1996) The effect of 30 minute cycle ergometry on Ankylosing Spondylitis. British Journal of Rheumatology. Feb.35(2):167-77.

Carbone LD, Cooper C, Michet CJ, Atkinson EJ, O'Fallon WM, Melton LJ. (1992) Ankylosing Spondylitis in Rochester, Minnesota, 1935-1989: Is the epidemiology changing? Arthritis and Rheumatism. 35(12):1476-1482.

Carette S, Graham D, Little H, Rubestein J, Rosen P. (1983) The natural disease course of Ankylosing Spondylitis. Arthritis and Rheumatism. 26:186-90.

Carlson R, Levy N. (1968) Brief method for assessing social-personal orientation. Psychology Reports. 23:911-914.

Carr AJ. A patient-centred approach to evaluation and treatment in rheumatoid arthritis: the development of a clinical tool to measure patient-perceived handicap. (1996) British Journal of Rheumatology. 35:921-932.

Carr AJ, Thompson PW. (1994) Towards a measure of patient-perceived handicap in Rheumatoid Arthritis. British Journal of Rheumatology. 33:378-382.

Carr AJ, Thompson PW, Kirwan JR. (1996) Quality of Life Measures. British Journal of Rheumatology. 35:275-281.

Caruso I, Cazzola M, Santandrea S. (1992) Clinical improvement in Ankylosing Spondylitis with rifamycin SV infiltrations of peripheral joints. Journal of International Medical Research. April.20(2):171-81.

Cats A, van der Linden SJ, Goei The HS, Khan MA. (1987) Proposals for diagnostic criteria of Ankylosing Spondylitis and allied disorders. Clinical and Experimental Rheumatology. 5:167-171.

Centre for Reviews and Dissemination (CRD) Report 4. Undertaking Systematic Reviews of Research on Effectiveness: CRD Guidelines for Those Carrying Out or Commissioning Reviews. NHS Centre for Reviews and Dissemination, University of York (1996).

Chapman CA, Zwillich SH. (1994) Olsalazine in Ankylosing Spondylitis: a pilot study. The Journal of Rheumatology. 21(9):1699-1701.

Cheshire MW. (1957) New apparatus: a device for measuring rotation of the neck. Archives of Physical Medicine. 38:592-7.

Chesson R. (1998) Psychosocial aspects of measurement. Physiotherapy. September.84(9):435-38.

Claudepierre P, Sibilia J, Goupille P, Flipo RM, Wendling D, Eulry F, Clerc D, Berthelot JM, Vergne P, Roudot-Thoraval F, Larget-Piet B, Chevalier X. (1997) Evaluation of a French version of the Bath Ankylosing Spondylitis Disease Activity Index in patients with spondyloarthropathy. The Journal of Rheumatology. Oct.24(10):1954-8.

Clegg DO, Reda DJ, Weisman MH, Blackburn WD, Cush JJ, Cannon GW, Mahowald ML. (1996) Comparison of sulphasalazine and placebo in the treatment of Ankylosing Spondylitis. A department of Veterans Affairs cooperative study. Arthritis and Rheumatism. Dec.39(12):2004-12.

Cohen J. (1977) Statistical power analysis for the behavioural sciences. New York: Academic Press.

Cook DJ, Greenhold NL, Ellrodt AG, Weingarten SR. (1997) The relation between systematic reviews and practice guidelines. Annals of Internal Medicine. 127(3):210-216.

Coons SJ, Rao S, Keininger DL, Hays RD. (2000) A comparative review of generic quality of life instruments. Pharmoeconomics. Jan.17(1):13-35.

Corkhill MM, Jobanputra P, Gibson T, Macfarlane DG. (1990) A contrlled trial of sulphasalazine treatment of of chronic Ankylosing Apondylitis; failure to demonstrate a clinical effect. British Journal of Rheumatology. 29:41-5.

Cotton S, Ruta D, Garratt AM, Russell EM. (1993) The problem of non-completion in quality of life measurement: making a quality of life measure more user friendly. University of

Aberdeen.

Creemers MCW, van't Hof MA, Fransse MJAM, van de Putte LBA, Gribnau FWJ, van Riel PLCM. (1994) A Dutch Version of the Functional Index for Ankylosing Spondylitis: development and validation in a long term study. British Journal of Rheumatolgy. 33(9):842-846.

Creemers MC, van't Hof MA, Franssen MJ, van der Putte LB, Gribnau FW, van Riel PL. (1996) Disease activity in Ankylosing Spondylitis: selection of a core set of variables and a first step in the development of a disease activity score. British Journal of Rheumatology. Sep.35 (9):867-73.

Cronstedt H, Waldner A, Stenstrom CH. (1999) The Swedish version of the Bath Ankylosing Spondylitis functional index. Reliability and validity. Scandinaian Journal of Rheumatology. Suppl.111:1-9

Dale K, Vinje O. (1985) Radiography of the Spine and Sacro-iliac joints in Ankylosing Spondylitis and Psoriasis. Acta. Radiological Diagnosis. 26:145-60.

Daltroy LH, Larson MG, Roberts WN, Liang MH. (1990) A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. The Journal of Rheumatology. 17(7): 946-950.

Daltroy LH, Larson MG, Eaton HM, Phillips CB, Liang M. (1999) Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. Social Science and Medicine. 48:1549-1561.

Dalyan M, Guner A, Tuncer S, Bilgic A, Arasil T. (1999) Disability in ankylosing spondylitis. Disability and Rehabilitation. Feb.21(2):74-9.

Dawes P. (1999) Stoke Ankylosing Spondylitis Spine Score. The Journal of Rheumatology. April. 26(4):993-6.

Dawes PT, Sheeran TP, Beswick EJ, Hothersall TE. (1987) Enthesopathy index in Ankylosing Spondylitis. Annals of the Rheumatic Diseases. 46:717.

Dawes PT, Sheeran TP, Beswick EJ, Hothersall TE. (1988) Chest pain - a common feature of Ankylosing Spondylitis. Postgraduate Medical Journal. 64:27-29.

de Jong Z, van der Heijde D, McKenna SP, Whalley D. (1997) The reliability and construct validity of the RAQoL: a Rheumatoid Arthritis-specific quality of life instrument. British Journal of Rheumatology. 36:878-883.

de Jong-Gierveld J, Kamphuis F. (1985) The development of a Rasch-type Loneliness Scale. Applied Psychological Measurement. 9: 89-99

De Witte LP. (1991) After the rehabilitation centre. A study into the course of functioning after discharge from rehabilitation. IRVSeries in Rehabilitation Research. Vol.3 Amsterdam: Swets & Zeitlinger.

Deyo RA, Inui TS. (1984) Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. Health Service Research. 19:275-90.

Deyo RA, Diehr P, Patrick DL. (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clinical Trials. 12:142S-158S

Deyo RA, Carter WB. (1992) Strategies for improving and expanding the application of health status measures in clinical setting. A researcher-developer viewpoint. Medical Care. May.30(5) Supplement:MS176-MS186.

Dickersin K, Scherer R, Lefebvre C. (1995) Identifying relevant studies for systematic reviews. In: Chalmers I, Altman DG. (Eds) Systematic Reviews. BMJ Publishing Group.

Domjan L, Nemes T, Balint GP, Toth Z, Gomor B. (1990) A simple method for measuring lateral flexion of the dorsolumbar spine. The Journal of Rheumatology. 17(5):663-665.

Dougadas M, van der Heijde D. (1999) Evaluation of functional capacity in Ankylosing Spondylitis. The Journal of Rheumatology. Jan.26(1):4-6

Dougadas M, Boumier P, Amor, B. (1986) Sulphasalasine in Ankylosing Spondylitis : a double blind controlled study in 60 patients. British Medical Journal. 11 October:293.

Dougadas M, Gueguen A, Nakache JP, Nguyen M, Mery C. Amor B. (1988) Evaluation of a functional index and an articular index in Ankylosing Spondylitis. The Journal of Rheumatology. 15(2):302-307.

Dougadas M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B. (1990) Evaluation of a functional index for patients with Ankylosing Spondylitis. The Journal of Rheumatology. 17;1254-5.

Dougadas M, Nguyen M, Caporal R, Legeais J, Bouxin-Sauzet A, Pellegri-Guegnault B, Gomeni C. (1994) Ximoprofen in Ankylosing Spondylitis. A double blind placebo controlled dose ranging study. Scandinavian Journal of Rheumatology. 23:243-8.

Dougadas M, van der Linden SJ, Leirisalo-Repo M, Huitfeldt B, Juhlin R, Veys E, Zeidler H, Kvien TK. (1995) Sulfasalazine in the treatment of spondylartropathy. A randomised, multi-center, double-blind, placebo-controlled study. Arthritis and Rheumatism. May.38(5):618-27.

Dougadas M, Gueguen A, Nakache JP, Velicitat P, Veys EM, Zeidler H, Calin A. (1999a) Ankylosing Spondylitis: what is the optimum duration of a clinical study? A one year versus a 6 weeks non-steroidal anti-inflammatory drug trial. Rheumatology. Mar.38(3):235-44.

Dunham WF. (1949) Ankylosing Spondylitis: Measurement of Hip and Spine Movements. The British Journal of Physical Medicine. Sept-Oct.(12):126-129.

Dziedzic KSG. The Body Chart: A further sketch towards a fuller picture of Ankylosing Spondylitis. Ph.D. Thesis. 1997. University of Keele, Staffordshire.

Dziedzic K. (1998) Chapter 10 - Ankylosing Spondylitis. In: David C, Lloyd J (Eds) Rheumatological Physiotherapy. Mosby, London. 97-114.

Edmunds L, Elswood J, Kennedy LG, Calin A. (1991) Primary Ankylosing Spondylitis, Psoriatic and Enteropathic Spondylarthropathy: A controlled analysis. The Journal of Rheumatology. 18(5):696-698.

EuroQol Group. (1990) EuroQol: a new facility for the measurement of health-related quality of life. Health Policy. 16:199-208.

Ferraz MB, Tugwell P, Goldsmith CH, Atra E. (1990) Meta-analysis of sulphasalazine in Ankylosing Spondylitis. The Journal of Rheumatology. Nov.17(11):1482-6.

Fisher LR, Cawley MID, Holgate ST. (1990) Relation between chest expansion, pulmonary function, and exercise tolerance in patients with Ankylosing Spondylitis. Annals of the Rheumatic Diseases. 49:921-925.

Fitzpatrick R. (1993a) The measurement of health status and quality of life in rheumatological disorders. Ballieres Clinical Rheumatology. June.7(2):297-317

Fitzpatrick R. (1999) Assessment of quality of life as an outcome: finding measurements that reflect individuals' priorities. Quality in Health Care. 8(1):1-2.

Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. (1992) Quality of life measures in health care. I: Applications and issues in assessment. British Medical Journal. 305.31October:1074-77.

Fitzpatrick R, Zieblans S, Jenkinson C, Mowat A, Mowat A. (1993b) Transition questions to assess outcomes in rheumatoid arthritis. British Journal of Rheumatology. Sep.32(9):807-11

Fitzpatrick R, Zieblan S, Jenkinson C, Mowat A., Mowat A. (1993c) A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. The Journal of Rheumatology. 20(3):429-436

Fitzpatrick R, Jenkinson C, Peto V. (1997) Desirable properties for instruments assessing quality if life: evidence from the PDQ-39. Journal of Neurology, Neurosurgery, and Psychiatry. 62:104-107

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. (1998a) Evaluating patient-based outcome measures for use in clinical trials. Health Technology Assessment. 2 (14).

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. (1998b) Chapter 2 - Patient-assessed outcome measures. In: Black N, Brazier J, Fitzpatrick R, Reeves B. (Eds). Health Service Research Methods. A guide to best practice. Part One: Measurement of benefits and costs. London: BMJ Books:13-22.

Fletcher A, Gore S. Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. (1992) Quality of life measures in health care. II: Design, analysis, and interpretation. British Medical Journal. 305.7 November:1145-48.

Fortin PR, Stucki G, Katz JN. (1995) Measuring relevant change: an emerging challenge in rheumatological clinical trials. Arthritis and Rheumatism. 38(8):1027-1030.

Franssen MJ, van Herwaarden CL, van de Putte P.B. (1986) Lung function in patients with Ankylosing Spondylitis. A study of the influence of disease activity and treatment with antiinflammatory drugs. The Journal of Rheumatology. 13:936-40.

Fraser SM, Sturrock RD. (1990) Evaluation of Sulphasalazine in Ankylosing Spondylitis - an interventional study. British Journal of Rheumatology. 29:37-39.

Freemont AJ. (1987) Chapter 1 - Pathology of Ankylosing Spondylitis. In: Calabro JJ, Carson Dick W. (Eds). New Clinical Applications, Rheumatology: Ankylosing Spondylitis. MTP Press, London:1-22.

Fries JF. (1980) Measurement of patient outcome in arthritis. Arthritis and Rheumatism. 23:137-45.

Funch DP. (1986) Assessment of a short scale to measure social support. Social Science and Medicine. 23(3):337-344.

Gandek B, Ware JE Jr. (1998c) Methods for validating and norming translations of health status questionnaires: the IQOLA Project approach. International Quality of Life Assessment. Journal of Clinical Epidemiology. Nov.51(11):953 - 9

Gandek B, Ware JE Jr., Aaronson NK, Alonso J, Apolone G, Bjorner J, Brazier J, Bullinger M, Fukuhara S, Kaasa S, Leplege A, Sullivan M. (1998a) Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countires: results from the IQOLA Project. International Quality of Life Assessment. Journal of Clinical Epidemiology. Nov.51(11):1149-58.

Gandek B, Ware JE, Aaronson NK, Apolone G, Bjorner J, Brazier J, Bullinger M, Kaasa S, Leplege A, Prieto L, Sullivan M. (1998b) Cross-validation of item selection and scoring for the SF-12 health survey in nine countries: Results from the IQOLA project. Journal of Clinical Epidemiology. 51(11): 1171-1178

Garratt AM. (1997) A comparison of four approaches to measuring health outcome. DPhil Thesis (1997). University of Aberdeen.

Garratt AM, Macdonald LM, Ruta DA, Russell IT, Buckingham JK, Krukowski ZH.(1993) Towards measurement of outcome for patients with varicose vein. Quality in Health Care. l(2):5-10

Garratt AM, Ruta DA, Abdalla MI, Russell IT. (1996a) Responsiveness of the SF-36 and a condition-specific measure of health for patients with varicose veins. Quality of Life Research. 5:223-234

Garratt AM, Ruta DA, Russell IT, Macleod K, Brunt P, McKinlay A, Mowat A, Sinclair T. (1996b) Developing a condition-specific measure of health for patients with dyspepsia and ulcer-related symptoms. Journal of Clinical Epidemiology. 49(5):565-571

Garratt AM, Hutchinson A, Russell I. (2000) The UK version of the Seattle Angina Questionnaire (SAQ-UK): reliability, validity and responsiveness. Journal of Clinical Epidemiology. *(in press).*

Garrett S, Jenkinson T, Kennedy G, Whitelock H, Gaisford P, Calin A. (1994) A new approach to defining disease status in Ankylosing Spondylitis: the Bath Ankylosing Spondylitis disease activity index. The Journal of Rheumatology. 1(12):2286-2291.

Gibbons FX. (1999) Social comparison as a mediator of response shift. Social Science and Medicine. 48:1517-1530.

Goodacre JA, Mander M, Carson Dick W. (1991) Patients with Ankylosing Spondylitis show individual patterns of variation on disease activity. British Journal of Rheumatology. 30:336-338.

Gran JT, Husby G. (1993) The epidemiology of ankylosing spondylitis. Seminars in Arthritis and Rheumatism. 22(5):319-334.

Gran JT, Husby G. (1998) Chapter 15 - Ankylosing Spondylitis - Prevalence and Demography. In: Klippel JH, Dieppe PA.(Eds). Rheumatology. Second Edition. Mosby, London. 1:15.1-2.

Greenhalgh J, Meadows K. (1999) The effectiveness of the use of patient-based measures of health in routine practice in imporving the process and outcomes of patient care: a literature review. Journal of Clinical Evaluation in Clinical Practice. 5(4):401-416.

Guillemin F, Briancon S, Pourel J, Goucher A. (1990) Long-term disability and prolonged sick leave as outcome measurements in ankylosing spondylitis. Possible predictive factors. Arthritis and Rheumatism. 33:1001-6.

Guillemin F, Challier B, Urlacher F, Vancon G, Pourel J. (1999) Quality of Life in ankylosing spondylitis: validation of the ankylosing spondylitis Arthritis Impact Measurement Scales 2, a modified Arthritis Impact Measurement Scales Questionnaire. Arthritis Care and Research. 2(3):157-162.

Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. (1987a) A measure of quality of life for clinical trials in chronic lung disease. Thorax. 42:773-778

Guyatt G, Walter S, Norman G. (1987b) Measuring change over time: assessing the usefulness of evaluative instruments. Journal of Chronic Disease. 40(2):171-178.

Guyatt GH, Van Zanten SJO, Feeny DH, Patrick DL. (1989a) Measuring Quality of Life in clinical trials : a taxonomy and review. Canadian Medical Association Journal. 140:1441-8.

Guyatt G, Mitchell A, Irvine EJ, Singer J, Williams N, Goodacre R, Tompkins C. (1989b) A new clinical measure of health status for clinical trials in inflammatory bowel disease. Gastroenterology. 96:804-10

Guyatt GH, Feeny DH, Patrick DL. (1993) Measuring health-related quality of life. Annals of Internal Medicine. 118:622-629

Guyatt GH, King DR, Feeny DH, Stubbing D, Goldstein RS. (1999) Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation. Journal of Clinical Epidemiology. 52(3):187-192

Hart FD, Bogdanovitch A, Nichol W. (1963) The thorax in Ankylosing Spondylitis. Annals of Rheumatic Diseases. 22:11-17.

Haslock I. (1993) Chapter 6 - Ankylosing Spondylitis. Ballieres Clinical Rheumatology. 7(1):99-115.

Hay DI. (1988) Socioeconomic status and health status: a study of males in the Canada Health Survey. Social Science and Medicine. 12:1317-25.

Haywood KL, Garratt AM, Dziedzic K, Dawes PT. (1998) A systematic review of outcome measures in Ankylosing Spondylitis (AS). British Journal of Rheumatology. 37.Abstracts supplement1: no.77:43

Haywood KL, Jordan K, Waterfield J, Dziedzic K, Garratt AM, Dawes PT. (1999) Reliability of Anthropomorphic measures in Ankylosing Spondylitis. Physiotherapy. July.85(7):372-373.

Hazes JMW, Hayton R, Siman AJ. (1993) A reevaluation of the symptom of morning stiffness. The Journal of Rheumatology. 20(7):1138-1142.

Helewa A, Goldsmith CH, Smythe HA. (1982) Independent measurement of functional capacity in Rheumatoid Arthritis. The Journal of Rheumatology. 9(5) Sept-Oct:;794-7.

Helliwell PS, Abbott CA, Chamberlain MA. (1996) A randomised trial of three different physiotherapy regimes in Ankylosing Spondylitis. Physiotherapy. Feb. 82(2): 85-90.

Herd RM, Tidman MJ, Ruta D, Hunter JAA. (1997) Measurement of quality of life in atopic dermatitis: correlation and validation of two different methods. British Journal of Dermatology. 136:502-507.

Hickey AM, Bury G, O'Boyle CA, Bradley F, O'Kelly F, Shannon W. (1996) A new short form individual measure (SEIQoL-DW): application in a cohort of individuals with HIV/AIDS. British Medical Journal. 313.6 July:29-33.

Hidding A, van der Linden SJ, Boers M, Gielen X, Kester A, Vlaeven A. (1992) Fake good test-taking attitude in Ankylosing Spondylitis. Arthritis and Rheumatism. Suppl.35:S244

Hidding A, van der Linden SJ, Boers M, Gielen X, de Witte L, Kester A, Dijkmans B & Moolenburgh D. (1993a) Is group physical therapy superior to individualised therapy in ankylosing spondylitis ? Arthritis Care and Research. 6(3):117-25.

Hidding A, van der Linden S, de Witte L. (1993b) Therapeutic effects of individual physical therapy in Ankylosing Spondylitis related to duration of disease. Clinical Rheumatology. 12(3):334-340.

Hidding A, van der Linden SJ, Gielen X, de Witte L, Dijkmans B, Moolenburgh M. (1994a) Continuation of group physical therapy is necessary in Ankylosing Spondylitis: results of a randomised controlled trial. Arthritis Care and Research. 7(2):90-6.

Hidding A, de Witte L, van der Linden S. (1994b) Determinants of self-reported health status in Ankylosing Spondylitis. The Journal of Rheumatology. 21(2):275-278.

Hidding A, van Santen M, De Klerk E, Gielen X, Boers M, Geenen R, Vlaeyen J, Kester A, van der Linden S. (1994c) Comparison between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. The Journal of Rheumatology. 21(5):818-823.

Hidding A, van der Linden S. (1995) Factors related to change in global health after group physical therapy in Ankylosing Spondylitis. Clinical Rheumatology. 14(3):347-351.

Hunt SM, McKenna SP, McEwan J. (1981) The Nottingham Health Profile: subjective health status and medical consultations. Social Science Medicine. 10:93-7.

Hunt SM, McKenna SP, McKenna J. (1989) The Nottingham Health Profile: user's manual. Revised edition. Manchester, England: (Manuscript).

Hurst NP, Jobabputra P, Hunter M, Lambert M, Lochhead A, Brown H. (1994) Validity of the EuroQol, a generic health status instrument, in patients with Rheumatoid Arthritis. British Journal of Rheumatology. 33:655-662.

Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. (1997) Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). British Journal of Rheumatology. 36:551-559.

Hurst NP, Ruta DA, Kind P. (1998) Comparison of the MOS Short Form-12 (SF-12) health status questionnaire with the SF-36 in patients with rheumatoid arthritis. British Journal of Rhematology. 37: 62-869.

Hyde SA. (1980) Physiotherapy in Rheumatology. Blackwell Scientific Publications. London.

Hyland ME, Finnis S, Irvine SH. (1991) A scale for assessing quality of life in adult asthma sufferers. Journal of Psychosomatic Research. 35(1):99-110

Jadad AR, Moher D, Klassen TP. (1998) Guides for reading and interpreting systematic reviews. Archives of Paediatric and Adolescent Medicine. 152(Aug):812-817.

Jamieson AH, Alford CA, Bird HA, Hindmarch I, Wright V. (1995) The effect of sleep and nocturnal movement on stiffness, pain, and psychomotor performance in Ankylosing Spondylitis. Clinical and Experimental Rheumatology. Jan-Feb.13(1):73-8.

Jenkinson C. (1994) Measuring health and medical outcomes: an overview. In: Jenkinson C (Ed). Measuring health and medical outcomes. London: UCL Press: 1-6.

Jenkinson C. (1995) Evaluating the Efficacy of Medical Treatment: possibilities and limitations. Social Science and Medicine. 41(10):1395-1401.

Jenkinson TR, Mallorie PA, Whitelock HC, Kennedy GL, Garrett SL, Calin A. (1994a) Defining spinal mobility in Ankylosing Spondylitis (AS). The Bath AS Metrology Index. The Journal of Rheumatology. 21(9):1694-1698.

Jenkinson C, Peto V, Coulter A. (1994b) Measuring change over time: a comparison of results from a global single item of health status and the multi-dimensional SF-36 health status survey questionnaire in patients presenting with menorrhagia. Quality of Life Research. 13:317-321.

Jenkinson C, Layte R, Wright L, Coulter A. (1996) The UK SF-36: an analysis and interpretation manual. A guide to health status measurement with particular reference to the Short-Form 36 Health Survey. Health Service Research Unit, Department of Public Health and Primary Care, University of Oxford. March. Joshua Horgan Print Partnership.

Jenkinson C, Fitzpatrick R, Peto V. (1998a) The Parkinson's Disability Questionnaire. User manual for the PDQ-39, PDQ-8 and PDQ Summary Index. Health Services Research Unit, Department of Public Health, University of Oxford. Joshua Horgan Print Partnership.

Jenkinson C, Stradling J, Petersen S. (1998b) How should we evaluate health status? A comparison of three methods in patients presenting with obstructive sleep apnoea. Quality of Life Research. 7:95-100.

Jenkinson C, Ruta D, Peterson S, Mowat A, Stradling J. (1998c) Should respondents be allowed to nominate new areas at follow-up in individualised quality of life assessment? An evaluation of two scoring methods using the Patients Generated Index (PGI). Quality of Life Research. November.7(7): 612

Jenksinson C, Fitzpatrick R, Brennan C, Bromberg M, Swash M. (1999a) Development and validation of a short measure of health status for individuals with amyotrophic lateral sclerosis / moror neurone disease: the ALSAQ-40. Journal of Neurology. 246(Suppl 3):III/16-III/21.

Jenkinson C, Fitzpatrick R., Peto V. (1999b) Health-related quality of life measurement in patients with Parkinson's Disease. Pharmoeconomics. Feb.15(2):157-165

Johnsen K, Gran JT, Dale K, Husby G. (1992) The prevalence of Ankylosing Spondylitis among Norwegian Smis (Lapps). The Journal of Rheumatology. 19(10):1591-1594.

Johnson JA, Coons SJ. (1999) Comparison of the EQ-5D and SF-12 in an adult US sample. Quality of Life Research. 7:155-166.

Joliffe IT. (1986) Principle component analysis. New York: Springer-Verlag.

Jones PW. (1992) Patients' perception of health as a measure of outcome. Medical Audit News. 2(4):57-58.

Jones SD, Porter J, Garrett SL, Kennedy LG, Whitelock H, Calin A. (1995) A new scoring system for the Bath Ankylosing Metrology Index (BASMI). The Journal of Rheumatology. Aug.22(8):1609. *letter.*

Jones SD, Steiner A, Garrett SL, Calin A. (1996a) The Bath Ankylosing Spondylitis Patient Global Score (BAS-G). British Journal of Rheumatology. 35: 66-71.

Jones SD, Koh WH, Steiner A, Garrett SL, Calin A. (1996b) Fatigue in Ankylosing Spondylitis: its prevalence and relationship to disease activity, sleep, and other factors. The Journal of Rheumatology. March.23.3:487-90.

Jones SD, Calin A, Steiner A. (1996c) An Update on the Bath Ankylosing Spondylitis Disease Activity and functional indices (BASDAI, BASFI) : excellent Cronbach's alpha scores. The Journal of Rheumatology. Feb.23(2):407.*letter.*

Jordan K. (2000) Assessment of published reliability studies for cervical spine range of motion measurement tools. Journal of Manipulative and Physiological Therapeutics. 23(3): 180-195.

Juniper EF, Huyatt GH, Streiner DL, King DR. (1997) Clinical impact versus factor analysis for quality of life questionnaire construction. Journal of Clinical Epidemiology. 50(3):233-238.

Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. (1992) Comparative measurement sensitivity of short and longer health status instruments. Medical Care. October.30(10):917-925.

Kazis LE, Anderson JJ, Meenan RF. (1989) Effect sizes for interpreting changes in health status. Medical Care. March. 27(3): Supplement S178-S189.

Keller SD, Maijkut TC, Kosinski M, Ware JE Jr. (1999a) Monitoring health outcomes among patients with arthritis using the SF-36 health survey. Medical Care. 37(5 Suppl):MS1-MS9.

Keller SD, Ware JE, Hatoum HT, Kong SX. (1999b) The SF-36 Arthritis-specific health index (ASHI) II. Tests of validity in four clinical trials. Medical Care. 37(5):MS51-MS60

Kennedy AC, Germain BF, Goldman AL, Vreede PD. (1991) A double-blind, crossover comparison of ketoprofen and indomethacin in Ankylosing Spondylitis. Advances in Therapy. 8(3):148-156.

Kennedy LG, Edmunds L, Calin A. (1993) The natural history of Ankylosing Spondylitis: does it burn out? The Journal of Rheumatology. 20:688-92.

Kennedy LG, Jenkinson TR, Mallorie PA, Whitelock HC, Garrett SL, Calin A. (1995) Ankylosing Spondylitis: the correlation between a new metrology score and radiology. British Journal of Rheumatology. Aug.34(8):767-70.

Kidd B, Mullee M, Frank A, Cawley M. (1988) Disease expression of Ankylosing Spondylitis in males and females. The Journal of Rheumatology. 5(9):1407-1409.

Kind P, Dolan P, Gudex C, Williams A. (1998) Variations in population health status: results froma United Kingdom national questionnaire survey. British Medical Journal. 316:736-72.

Kippers V, Parker AW. (1987) Toe-touch test. A measure of its validity. Physical Therapy. 67:1680-4.

Kirshner B, Guyatt G. (1985) A methodological framework for assessing health indices. Journal of Chronic Diseases. 38:27-36.

Kirwan JR, Reeback JS. (1983) Using a modified Stanford Health Assessment Questionnaire to assess disability in UK patients with Rheumatoid Arthritis. Annals of Rheumatic Diseases.

42:219-20.

Kirwan J, Edwards A, Huitfield B, Thompson P. (1993) The course of established Ankylosing Spondylitis and the effects of sulphazine over 3 years. British Journal of Rheumatology. 32:729-733.

Klaber Moffett JA, Hughes I, Griffiths P. (1989) Measurement of cervical spine movements using a simple inclinometer. Physiotherapy. 75(6):309-312.

Koh WH, Garrett SL, Calin A. (1997a) Cervical spine surgery in ankylosing spondylitis: is the outcome good? Clinical Rheumatology. Sep.16(5):466-70.

Koh WH, Pande I, Samuels A, Jones SD, Calin A. (1997b) Low dose amitriptyline in ankylosing spondylitis: a short term, double blind, placebo controlled study. The Journal of Rheumatology. Nov. 24 (11):2158-61

Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE Jr. (1999a) The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: tests of data quality, scaling assumptions and score reliability. Medical Care. May.37 (5 Suppl):MS10-22

Kosinski M, Keller SD, Ware JE Jr, Hatoum HT, Kong SX. (1999b) The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: relative validity of scales in relation to clinical measures of arthritis severity. Medical Care. May.37(5 Suppl):MS23-39

Kragg G, Stokes B, Groh, J, Helewa A, Goldsmith C. (1990) The effects of comprehensive home physiotherapy and supervision on patients with Ankylosing Spondylitis - A randomised controlled trial. The Journal of Rheumatology. 17(2):228-233.

Kragg G, Stokes B, Groh, J, Helewa A, Goldsmith C. (1994) The effects of comprehensive home physiotherapy and supervision on patients with Ankylosing Spondylitis - an 8 month followup. The Journal of Rheumatology. 21(12):261-263.

Laurent MR, Buchanan WW, Bellamy N. (1991) Methods of assessment used in Ankylosing Spondylitis clinical trials: a review. British Journal of Rheumatology. 30:326-9.

Lee YS, Schlotzhauer T, Ott SM, van Vollenhoven RF, Hunter J, Shapiro J, Marcus R, McGuire JL. (1997) Skeletal status of men with early and late ankylosing. American Journal of Medicine. Sep. 103(3):233-41.

Liang MH. (1995) Evaluating measurement responsiveness. The Journal of Rheumatology. 22(6):1191-1192.

Liang MH, Larson MG, Cullen KE. (1985) Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis and Rheumatism. 28:542-547.

Liang MH, Fossel AH, Larson MG. (1990) Comparisons of five health status instruments for orthopaedic evaluation. Medical Care. July.28(7):632-642.

Little H. (1986) The Neck and Back. In: The Rheumatological Physical Examination. Orlando: Grune and Stratton:56.

Lubrano E, Butterworth M, Heddelden A, Well S, Helliwell P. (1998) An audit of anthropometric measurements by medical and physiotherapy staff in patients with ankylosing spondylitis. Clinical Rehabilitation. June.12(3):216-20.

Lubrano E, Helliwell P. (1999) Deterioration in anthropometric measures over six years in patients with ankylosing spondylitis. Physiotherapy. 85/3:138-143.

Macduff C, Russell E. (1998) The problem of measuring change in individual health-related quality of life by postal questionnaire: use of the patient-generated index in a disabled population. Quality of life Research. 7:761-769.

MacKay K, Mack C, Brophy S, Calin A. (1998) The Bath Ankylosing Spondylitis Radiology Index. Arthritis and Rheumatology. Dec.41(12):2263-70.

Macrae IF, Wright V. (1969) Measurement of Back Movement. Annals of Rheumatic Disease. 28: 584-589.

Maksymowych WP, Jhangri GS, Leclercq S, Skeith K, Yan A, Russell AS. (1998) An open study of pamidronate in the treatment of refractory ankylosing spondylitis. The Journal of Rheumatology. Apr.25(4):714-7.

Mander M, Simpson J, McLellan A, Walker D, Goodacre JA, Carson Dick W. (1987) Studies with and enthesis index as a method of clinical assessment in ankylosing spondylitis. Annals of the Rheumatic Diseases. 46:197-202.

Mari JJ, Williams P. (1985) A comparison of the validity of two psychiatric screening questionnaires (GHQ-12 and SRQ-20) in Brazil using relative operating characteristics (ROC) analysis. Psychology and Medicine. 15:651-9.

McArthur G. (1997) Assessing disability in Rheumatoid Arthritis. Unpublished BmedSci dissertation. University of Aberdeen.

McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Garratt A, Harvey E, Thomas R, Bond J. (1998) Designing and using patient and staff questionnaires. In: Black N, Brazier J, Fitzpatrick R, Reeves B. (Eds). Health Service Research Methods. A guide to best practice. Part One: Measurement of benefits and costs. London: BMJ Books:46-60.

McDowell I, Jenkinson C. (1996) Development of standards for health measures. Journal of Health Service Research Policy. October.1(4):238-246.

McDowell I, Newell C. (1996) Measuring Health. A guide to rating scales and questionnaires. Oxford University Press, Inc.: Second Edition.

McHorney CA, Ware JE, Raczek AE. (1993) The MOS-36-item Short-Form Health Survey (SF-36): II. Psychometric and clinic tests of validity in measuring physical and mental health constructs. Medical Care. Mar.31(3):247-63.

McHorney CA, Ware JE, Lu JF, Sherboourne CD. (1994) The MOS-36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Medical Care. Jan.32(1):40-66.

McHorney CA, Tarlov RA. (1995) Individual-patient monitoring in clinical practice: are available health status surveys adequate? Quality of Life Research. 4:293-307.

Meenan RF, Gertmen PM, Mason JH. (1980) Measuring health status in arthritis. The Arthritis Impact Measurement Scales. Arthritis and Rheumatism. 23:146-52.

Meenan RF, Mason JH, Anderson JJ. (1992) AIMS2: the contents and properties of a revised and expanded Arthritis Impacts Measurement Scales health status questionnaire. Arthritis and Rheumatism. 35:1-10.

Mellin G, Olenius P, Setala H. (1994) Comparison between three different inclinometers. Physiotherapy. 80(9):612-614.

Melzack R. (1975) The McGill Pain Questionnaire:major properties and scoring methods. Pain. 1:277-99.

Miller MH, Lee P, Smythe HA, Goldsmith CH. (1984) Measurement of spinal mobility in the sagittal plane: new skin contraction technique compared with established methods. The Journal of Rheumatology. 11:507-11.

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. (1995) Assessing the quality of randomised controlled trials: an annotated bibliography of scales and checklists. Controlled Clinical Trials. 16:62-73.

Moll JMH, Wright V. (1972) An objective clinical study of chest expansion. Annals of Rheumatic Disease. 31:1-8.

Moll JMH, Wright V. (1973) The pattern of chest and spinal mobility in Ankylosing Spondylitis. Rheumatology and Rehabilitation. 12:115-134.

Moll JMH, Liyange SP, Wright V. (1971) Normal range of spinal mobility - An objective clinical study. Annals of the Rheumatic Diseases. 30:381-386.

Moll JMH, Liyange SP, Wright V. (1972a) An objective clinical method to measure lateral spinal flexion. Rheumatology and Physical Medicine. 11:225-239.

Moll JMH, Liyange SP, Wright V. (1972b) An objective clinical method to measure spinal extension. Rheumatology and Physical Medicine. 11:293-312.

Mulrow CD. (1995) Rationale for Systematic Reviews. In: Chalmers I, Altman DG. (Eds). Systematic Reviews. BMJ Publishing Group.

Nelson EC, Wasson JH, Johnson DJ. (1987) Assessment of function in routine clinical practice: description of the COOP chart method and preliminary findings. Journal of Chronic Diseases. 40(Suppl.1): 55S-63S.

Nemes L. (1991) Psychosocial and physical problems in AS. NASS Newsletter. Spr/Sum:8-10.

Nemeth R, Smith F, Elswood J, Calin A. (1987) Ankylosing Spondylitis (AS) - an approach to measurement of severity and outcome: Ankylosing Spondylitis Assessment Questionnaire (ASAQ) - a controlled study. British Journal of Rheumatology. Supplement 1:69-70.

Nicassio PM, Wallston KA, Callahan LF, Herbert M, Pincus T. (1985) The measurement of helplessness in rheumatoid arthritis. The development of the Arthritis Helplessness Index. The Journal of Rheumatology. 12(3):462-7.

Nunnally JC, Bernstein IH. (1994) Psychometric Theory. Third Edition. McGraw-Hill Series in Psychology, McGraw-Hill, Inc.

O'Boyle CA, McGee H, Hickey A. (1993) The Schedule for the Evaluation of Individual Quality of Life (SEIQoL): Administration Manual. Dublin: Department of Psychology, Royal College of Surgeons in Ireland.

O'Brien BJ, Elswood J, Calin A. (1990) Willingness to accept risk in the treatment of rheumatic disease. Journal of Epidemiology and Community Health. 44:249-52.

O'Driscoll SL, Jatson MIV, Baddeley H. (1978) Neck movements in ankylosing spondylitis and its response to physiotherapy. Annals of the Rheumatic Diseases. 37:64-66.

Ostendorf B, Specker C, Schneider M. (1998) Methotrexate lacks efficacy in the treatment of severe ankylosing spondylitis compared with rheumatoid and psoriatic arthritis. Journal of Clinical Rheumatology. 4(3):129-136.

Oxman A. (1995) Checklists for Review Articles. In: Chalmers I, Altman DG. (Eds). Systematic Reviews. BMJ Publishing Group.

Pal B. (1987) Early diagnosis of Ankylosing Spondylitis. Journal of the Indian Medical Association. 85(9):275-277.

Palferman TG, Webley M. (1991) A comparative study of nabumetone and indomethacin in Ankylosing Spondylitis. Europen Journal of Rheumatology and Inflammation. 11(2):23-29.

Pasero G, Ruju G, Marcolongo R, Senesi M, Serni U, Mannoni A. (1994) Aceclofenac versus naproxen in the treatment of Ankylosing Spondylitis: a double-blind, controlled study. Current Therapeutic Research and Clinical Experimentation. 8(4):833-842.

Paterson C. (1996) Measuring outcomes in primary care: a patient generated measure, MYMOP, compared with the SF-36 health survey. British Medical Journal. 312.20 April:1016-1020.

Patrick DL, Erickson PE. (1993) Health status and health policy: allocating resources to health care. Oxford: Oxford University Press.

Pearcy MJ, Wordsworth BP, Portek I, Mowat AG. (1985) Spinal Movements in Ankylosing Spondylitis and the effect of treatment. Spine. 10:472-4.

Peto V, Jenkinson C, Fitzpatrick R. (1998) PDQ-39: a review of the development, validation and application of a Parkinson's disease quality of life questionnaire and it's associated measures. Journal of Neurology. 245(Suppl 1):S10-S14.

Pile KD, Laurent MR, Salmond CE, Best MJ, Pyle EA, Moloney RO. (1991) Clinical assessment of ankylosing spondylitis: a study of observer variation in spinal measurements. British Journal of Rheumatology. 30:29-34.

Pincus T, Summey JA, Soraci SA Jr. (1983) Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis and Rheumatism. 26:1346-1353.

Radloff DP. (1977) The CES-D Scale : A self-report depression scale for research with the general population. Applied Psychological Measurement. 1(3):385-401.

Ralston SH, Urquhart GDK, Brzeski M, Sturrock RD. (1990) Prevalence of vertebral compression fractures due to osteoporosis in Ankylosing Spondylitis. British Medical Journal. 300. 3 March: 563-565.

Ramos-Remus, C. & Russell, AS. (1992) New Clinical and Radiographic features of Ankylosing Spondylitis. Current Opinion in Rheumatology. Aug.4(4):463-9.

Ramos-Remus C, Major P, Gomez-Vargas A, Petrikowski G, Hernandez-Chavez A, Gonzalez-Marin E, Russell AS. (1997) Temporomandibular joint osseous morphology in a consecutive sample of Ankylosing Spondylitis patients. Annals of Rheumatic Disease. Feb.56(2):103-7.

Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. (1998) Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project. Journal of Clinical Epidemiology. 51(11): 1203-1214.

Read JL, Quinn RJ, Hoffer MA. (1987) Measuring overall health: an evaluation of three important approaches. Journal of Chronic Diseases. 40(1)Supplement:7S-21S.

Redelmeier DA, Guyatt GH, Goldstein RS. (1996) Assessing the minimal importance difference in symptoms: a comparison of two techniques. Journal of Clinical Epidemiology. 49(11):1215-1219.

Reynolds PMG. (1975) Measurement of Spinal Mobility: a comparison of three methods. Rheumatology and Rehabilitation. 14:180-5.

Reynolds S, Doward LC, Spoorenberg A, Helliwell PS, McKenna SP, Tennant A, van der Heijde DMFM, Chamberlain MA. (1999) The development of the Ankylosing Spondylitis Quality of Life Questionnaire (ASQoL). Quality of Life Research. 8(7):651.

Rigby AS. (1991) Ankylosing Spondylitis - review of UK data on the rheumatic diseases - 5. British Journal of Rheumatology. 30:50-53.

Ritchie DM, Boyle JA, McInnes JM, Jasani MK, Dalakos TG, Griebeson O, Buchanan WW. (1968) Clinical studies with an articular index for the assessment of joint tenderness in patients with Rheumatoid Arthritis. Quarterly Journal of Medicine. New Series. 37(147) July:393-406.

Roberts WN, Liang MH, Pallozzi LM, Daltroy LH. (1988) Effects of warming up on reliability of anthropometric techniques in Ankylosing Spondylitis. Arthritis and Rheumatism. 31(4) April: 549-552.

Roberts WN, Larson MG, Liang MH, Harrison RA, Barefoot J, Clarke AK. (1989) Sensitivity of anthropometric techniques for clinical trials in Ankylosing Spondylitis. British Journal of Rheumatology. 28:40-45.

Ruof J, Stucki G. (1999a) Comparison of the Dougadas Funcitonal Index and the Bath Ankylsoing Spondylitis Index. A literature review. The Journal of Rheumatology. April.26(4):955-60.

Ruof J, Stucki G. (1999b) Validity aspects of erythorocyte sedimentation rate and C-reactive protein in Ankylosing Spondylitis: a literature review. The Journal of Rheumatology. April.26(4):966-970.

Ruof J, Sagha O, Stucki G. (1999) Comparative responsiveness of three functional indices in Ankylosing Spondylitis. The Journal of Rheumatology. 26/9:1959-1963.

Russell AS. (1998) Chapter 14 - Ankylosing Spondylitis - History. In: Klippel JH, Dieppe PA. (Eds). Rheumatology. Second Edition. Mosby, London. 1:14.1-2.

Ruta D. (1998) Patient Generated Index (PGI) Web-site. http://www.dundee.ac.uk/epidemiology/PGI

Ruta D, Garratt A, Leng M, Russell IT, Macdonald LM. (1994a) A new approach to the measurement of quality of life - the Patient Generated Index. Medical Care. 32(11):1109-1126.

Ruta DA, Garratt AM, Wardlaw D, Russell IT. (1994b) Developing a valid and reliable measure of health outcome for patients with low back pain. Spine. 19 (17):1887-1896.

Ruta DA, Hurst NP, Kind P, Hunter M, Stunnings A. (1998) Measuring health status in British patients with Rheumatoid Arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). British Journal of Rhematology. 37:425-436.

Ruta DA, Garratt AM, Russell IT. (1999) Patient centred assessment of quality of life for patients with four common conditions. Quality in Health Care. 8(1):22-29.

Safran DG, Kosinski M, Tarlov AR, Rogers WG, Taira DG, Lieberman N, Ware JE. (1998) The Primary Care Assessment Survey: tests of data quality and measurement performance. Medical Care. May.36(5): 728-39.

Santos H, Brophy S, Calin A. (1998) Exercise in ankylosing spondylitis: how much is optimum? The Journal of Rheumatology. Nov. 25(11): 2156-60.

Schwartz CE, Sprangers MAG. (1999) Methodological approaches for assessing response shift in longitudinal health-related quality of life research. Social Science and Medicine. 48:1531-1548.

Shrout PE, Fleiss JL. (1979) Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 86:420-428.

Spoorenberg A, van der Heijde D, de Klerk E, Dougadas M, de Vlam K, Mielants H, van der Tempel H, van der Linden S. (1999a) A comparative study of the usefulness of the Bath Ankylosing Spondylitis Functional Index and the Dougadas Functional Index in the assessment of ankylosing spondylitis. The Journal of Rheumatology. Apr.26(4):961-5.

Spoorenberg A, van der Heijde D, de Klerk E, Dougadas M, de Vlam K, Mielants H, van der Tempel H, van der Linden S. (1999b) Relative value of erythrocyte sedimentation rate an C-reactive protein in assessment of disease activity in ankylosing spondyltiis. The Journal of Rheumatology. Apr. 26(4): 980-4.

Spoorenberg A, de Vlam K, van der Heijde D, de Klerk E, Dougadas M, Mielants H, van der Tempel H, Boers M, van der Linden S. (1999C) Radiological scoring methods in Ankylosing Spondylitis: reliability and sensitivity to change over one year. The Journal of Rheumatology. Apr. 26(4): 997-1002.

Steinbrocker O, Traeger CH, Batterman RC. (1949) Therapeutic Criteria in Rheumatoid Arthritis. Journal of the American Medical Association. 140:659-62.

Stokes BA, Helewa A, Goldsmith CH, Groh JD, Kragg GR. (1988) Reliability of spinal mobility measurements in ankylosing spondylitis patients. Physiotherapy Canada. 40(6): 338-44.

Stratford P, Gill C, Westaway M, Binkley J. (1995) Assessing disability and change on individual patients: a report of a patient specific measure. Physiotherapy Canada. Fall.47(4):258-263.

Streiner DL, Norman GR. (1995) Health Measurement Scales. A practical guide to their development and use. Second Edition. Oxford Medical Publications, Inc.

Stucki G, Liang MF, Fossel AH, Katz JN. (1995) Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. Journal of Clinical Epidemiology. 48(11):1369-1378.

Sturrock RD, Wojtulewski JA, Hart DA. (1973) Spondylometry in a normal population and in ankylosing spondylitis. Rheumatology and Rehabilitation. 12:135-142.

Sutton AJ, Jones DR, Abrams KR, Sheldon TA, Song F. (1999) Systematic reviews and meta-analysis: a structured review of the methodological literature. Journal of Health Service Research Policy. 4(1)January:49-55.

Symmons DP. (1995) Disease assessment indices: activity, damage and severity. Ballieres Clinical Rheumatology. May.9(2):267-85.

Symmons DP. (1996) Mortality in Ankylosing Spondylitis. Rheumatology in Europe. 25(1):15-16.

Taggart A, Gardiner P, McEvoy F, Hopkins R, Bird H. (1996) Which is the active moiety of sulphasalazine in Ankylosing Spondylitis?. A randomised, controlled study. Arthritis and Rheumatism. Aug.39(8):1400-5.

Taylor AL, Balakrishnan C, Caln A. (1998) Reference centile charts for measures of disease activity, functional impairment, and metrology in ankylosing spondylitis. Arthritis and Rheumatism. Jun.41(6):119-25.

Taylor HG, Wardle T, Beswick EJ, Dawes, PT. (1991a) The relationship of clinical and laboratory measurements to radiological change in ankylosing spndylitis. British Journal of

Rheumatlogy. 30:330-335.

Taylor HG, Beswick EJ, Dawes PT. (1991b) Sulphasalazine in Ankylosing Spondylitis. A radiological, clinical and laboratory assessment. Clinical Rheumatology. 10(1):43-48.

Thompson GT, Chalmers IM. (1993) Fiddling while Rome burns: burn out, remission and disease activity measurements in Ankylosing Spondylitis (editorial). The Journal of Rheumatology. April. 20(4):607-9.

Tishler M, Brostovski Y, Yaron M. (1995) Effect of Spa Therapy in Tiberias on patients with Ankylosing Spondylitis. Clinical Rheumatology. Jan.14(1):21-5.

Tomlinson MJ, Barefoot J, Dixon AS. (1986) Intensive in-patient physiotherapy courses improve movement and posture in Ankylosing Spondylitis. Physiotherapy. 72(5):238-40.

Torrence GW. (1976) Social preferences for health states: an empirical evaluation of three measurement techniques. Socieconomic Planning Science. 10:129-36.

Tugwell P, Bombardier C, Buchanan W, Goldsmith CH, Grace E, Hanna B. (1987) The MACTAR Patient Preference Disability Questionnaire. An individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. The Journal of Rheumatology. 14(3):446-451.

van der Heijde D, Spoorenberg A. (1999) Plain radiographs as an outcome measure in ankylosing spondylitis. The Journal of Rheumatology. Apr.26(4):985-7.

van der Heijde D, Bellamy N, Calin A, Dougadas M, Khan MA, van der Linden S. (1997) Preliminary core sets for endpoints in Ankylosing Spondylitis. The Journal of Rheumatology. 24(11):2225-2229.

van der Heijde D, van der Linden S. (1998) Measures of outcome in ankylosing spondylitis and other spondyloarthropathies. Ballieres Clinical Rheumatology. Nov.12(4):683-93.

van der Heijde D, Calin A, Dougadas M, Khan MA, van der Linden S, Bellamy N. (1999a) 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. The Journal of Rheumatology. Apr.26(4):951-4.

van der Heijde D, van der Linden S, Dougadas M, Bellamy N, Russell AS, Edmonds J. (1999b) Ankylosing Spondylitis: plenary discussion and results of voting on selection of domains and some specific instruments. The Journal of Rheumatology. Apr.26(4):1003-5.

van der Heijde D, van der Linden S, Bellamy N, Calin A, Dougadas M, Khan MA. (1999c) Which domains should be included in a core set for endpoints in Ankylosing Spondylitits? Introduction to the Ankylosing Spondylitis module of OMERACT IV. The Journal of Rheumatology. Apr.26(4): 945-7.

van der Linden S, van der Heijde DM. (1995) Ankylosing Spondylitis and other B27 related spondyloarthropathies. Ballieres Clinical Rheumatology. May.9 (2):355-73.

van der Linden S, van der Heijde DM. (1996) Clinical and epidemiological aspects of Ankylosing Spondylitis and spondyloarthropathies. Current Opinion in Rheumatology. Jul.8(4):269-74.

van der Linden S, van der Heijde D. (1998) Ankylosing Spondylitis. Clinical features. Rheumatic Disease Clinics of North America. Nov.24(4):663-76, vii.

van der Linden SJ, Valkenburg HA, Cats A. (1984) Evaluation of Diagnostic Criteria for Ankylosing Spondylitis - a proposal for modification of the New York Criteria. Arthritis and Rheumatism. April.27(4):361-368.

van der Linden SM, Ferraz MB, Tugwell P. (1990) Clinical and functional assessment of Ankylosing Spondylitis. Spine: State of the Art Reviews. Sep.4(3):583-94.

Viitanen JV, Suni J, Kautianen H. (1992) Effect of Physiotherapy on spinal mobility in Ankylosing Spondylitis. Scandinavian Journal of Rheumatology. 21:38-41.

Viitenan JV. (1993) Thoracolumbar rotation in Ankylosing Spondylitis. A new non-invasive measurement method. Spine. 18(7):880-3.

Viitenan JV, Suni J. (1995) Management principles of physiotherapy in Ankylosing Spondylitis - which treatments are effective? Physiotherapy. 81(6):322-9.

Viitenan JV, Kautiainen H, Suni J, Kokko ML, Lehtinen K. (1995a) The relative value of spinal and thoracic mobility measurements in Ankylosing Spondylitis. Scandinavian Journal of Rheumatology. 24(2):94-7.

Viitanen JV, Lehtinen K, Suni J, Kautiainen H. (1995b) Fifteen Months' follow-up of intensive inpatient physiotherapy and exercise in Ankylosing Spondylitis. Clinical Rheumatology. Jul.14(4):413-9.

Viitanen JV, Kokko ML, Lehtinen K, Suni J. (1995c) Correlation between mobility restrictions and radiological changes in Ankylosing Spondylitis. Spine. Feb 15;20(4):492-6.

Viitenan JV, Kautiainen H, Kokko ML, Ala-Peijari S. (1995d) Age and spinal mobility in Ankylosing Spondylitis. Scandinavian Journal of Rheumatology. 24(5):314-5.

Viitanen JV, Kokko ML, Heikkila S, Kautiainen H. (1998) Neck mobility assessment in Ankylosing Spondylitis: a clinical study of nine measurements including new tape methods for cervical rotation and lateral flexion. British Journal of Rheumatology. Apr.37(4):377-81.

Viitanen JV, Kokko ML, Heikkila S, Kautiainen H. (1999) Assessment of thoracolumbar rotation in Ankylosing Spondylitis: a simple tape method. Clinical Rheumatology.

18(2):152-7.

Von Schober P. (1937) The Lumbar Region of the Spine and Lumbago. Munchener Medizinische Wochenschrift.

Waldner A, Cronstedt H, Stenstrom CH. (1999) The Swedish version of the Bath ankylosing spondylitis disease activity index. Reliability and validity. Scandinavian Journal of Rheumatology. Suppl. 111:10-6.

Walker EA, Katon WJ, Jemelka RP. (1993) Psychoatric disorders and medical care utilisation among people in the general population who report fatigue. Journal of General Intern. Medicine. 8:436-440

Wallston KA, Wallston BA, DeVellis R. (1978) Development of the multi-dimensional health locus of control (MHLC) scales. Health Education. Monograph. 6:161-170.

Ward MM, Kuzis S. (1999) Validity and Sensitivity to change of spondylitis-specific measures of functional disability. The Journal of Rheumatology. Jan.26(1):121-7.

Ward MM. (1998) Quality of life in patients with Ankylosing Spondylitis. Rheumatic Disease Clinics of North America. Nov.24(4):815-27, x.

Ware JE, Kosinski M, Keller SD. (1994) SF-36 Physical and Mental Health Summary Scales: A user's manual. Boston, MA: Health Assessment Lab. 5th printing.

Ware JE. (1997) SF-36 Health Survey Manual and Interpretation Guide. The Medical Outcomes Trust. Nimrod Press, Boston, USA.

Ware JE. (1998) Preface - International Quality of Life Assessment (IQOLA) Project. Journal of Clinical Epidemiology. 51(11):891-892.

Ware JE, Gandek B. (1998a) Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. Journal of Clinical Epidemiology. 51(11):903-912.

Ware JE, Gandek B. 91998b) Methods for testing data quality, scaling assumptions, and reliability: The IQOLA project approach. Journal of Clinical Epidemiology. 51(11):945-952.

Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. (1995) Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes study. Medical Care. 33(4) Supplement:AS264-AS279.

Ware JE, Kosinski M, Keller SD. (1995) SF-12: How to score the SF-12 Physical and Mental Health Summary Scales. The Health Institute, New England Medical Centre, Boston, USA.

Ware JE, Kosinski M, Keller SD. (1996) A 12-item short-form health survey. Construction of scales and preliminary tests of reliability and validity. Medical Care. 4(3):220-233.

Whalley D, McKenna SP, de Jong Z, van der Heijde D. (1997) Quality of life in Rheumatoid Arthritis. British Journal of Rheumatology. 36:884-888.

Will R, Edmunds L, Elswood J, Calin A. (1990) Is there Sexual Inequality in Ankylosing Spondylitis? A Study of 498 Women and 1202 Men. The Journal of Rheumatology. 17(12):1649-1652.

Wilson IB. (1999) Clinical understanding and clinical implications of response shift. Social Science and Medicine. 48:1577-1588.

Wolfe F. (1995) Health status questionnaires. Rheumatic Disease Clinics of North America. 21(2):445-464.

Wordsworth BP, Mowat AG. (1986) A review of 100 patents with Ankylosing Spondylitis with particular reference to socio-economic effects. British Journal of Rheumatology. 25: 175-178.

World Health Organisation. (1947) Constitution of the World Health Organisation. Geneva: WHO.

Ziebland S. (1994) Measuring changes in health status. In: Jenkinson C (Ed). Measuring health and medical outcomes. London: UCL Press :42-53.

Zukovskis K, Taylor HG, Beswick E, Dawes PT. (1991) Enthesitis as a measure of disease activity in Ankylosing Spondylitis (AS). British Journal of Rheumatology. Abstracts Supplement.30(81): No.162.