Patient-assessed health in ankylosing spondylitis: a structured review

K. L. Haywood, A. M. Garratt¹ and P. T. Dawes²

Objective. To review evidence relating to the measurement properties for all disease-specific, multi-item, patient-assessed health instruments in patients with ankylosing spondylitis (AS).

Methods. Systematic literature searches were made to identify instruments, using predefined criteria relating to reliability, validity, responsiveness and precision.

Results. Twelve AS-specific and three arthritis-specific instruments met the inclusion criteria. Three AS-specific instruments that measure health-related quality of life (HRQL) were reviewed. The Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), the Bath Ankylosing Spondylitis Functional Index (BASFI) and the Dougados Functional Index (DFI) had the greatest amount of evidence for reliability, validity and responsiveness across a range of settings. Four instruments lacked evidence for test–retest or internal consistency reliability. Most were assessed for validity through comparisons with other instruments, global judgements of health, mobility or clinical and sociodemographic variables. Most were assessed for responsiveness through mean score changes. Three instruments lacked evidence of responsiveness.

Conclusion. This review provides a contribution to AS assessment. AS-specific multi-item measures specific to the assessment of pain, stiffness, fatigue and global health were not identified; where assessed, these domains were largely measured with single-item visual analogue scales. Single items may provide a limited reflection of these important domains. The BASFI and DFI remain the instruments of choice for functional assessment. HRQL is recommended as a core assessment domain. Further concurrent evaluation is recommended.

Key words: Ankylosing spondylitis, Patient-assessed health, Measuring properties.

Ankylosing spondylitis (AS) is an incurable, inflammatory disease, primarily affecting the pelvis and spine, but often with involvement of peripheral joints, entheses and extra-articular sites [1]. Often affecting individuals from an early adult age, the disease can have a profound impact on life quality in terms of physical, social and psychological well-being [2]. It is widely accepted that the patients’ perception of disease impact and the outcomes of health-care should be included in clinical trials and similar forms of evaluative study. This has resulted in a significant increase in the availability of patient-assessed health instruments which aim to measure aspects of health from the perspective of the patient [3]. Structured reviews of measurement properties and accompanying professional consensus are helpful in supporting instrument selection and standardization [3, 4].

To encompass the multidimensional impact of AS, five core evaluative domains and, with the exception of functional disability, associated single-item scales have been recommended by the Assessment in AS International Working Group (ASAS; www.asas-group.org): pain [intensity; visual analogue scale (VAS)], spinal stiffness/inflammation (morning stiffness duration; VAS), functional ability [Bath Ankylosing Spondylitis Functional Index (BASFI) [5] or Dougados Functional Index (DFI) [6]], patient global assessment of health status (VAS; last week) and spinal mobility [7, 8]. Instrument selection was informed by a literature review (1988–1995; Medline database and citation searches), expert opinion and professional consensus agreement.

Single items, such as those recommended by ASAS, may not allow patients to appropriately report the wide impact of disease or treatment, providing a limited reflection of health [4]. The resulting summary judgement of health status limits measurement validity and score interpretation [4, 9]. Patient-assessed health instruments usually take the form of questionnaires containing multiple items, or questions, to reflect the broad nature of health status, disease or injury [3, 4]. These instruments aim to provide an accurate assessment of health or disease from the patient’s perspective, which contribute to validity and score interpretation [3, 4, 9]. This structured review presents an updated and more extensive review of published evidence for AS- and arthritis-specific multi-item, patient-assessed instruments that measure any aspect of health or health-related quality of life (HRQL) in this disease (1988–2004). The review will inform instrument selection and future research within this field.

Methods

Identification of studies

The search strategy was designed to retrieve references relating to the development and evaluation of multi-item patient-assessed...
health instruments, including reviews. The first AS-specific patient-assessed health instrument was developed in 1988 [6], so the search strategy was restricted to the period from 1988 to August 2004. All searches included terms specific to AS combined with the measurement of health outcome and instrument measurement properties [4, 10]. Search strategies used medical subject headings (MeSH terms) and free text searching. Further searches used names of identified instruments.

Databases searched included AMED, CINAHL, the Cochrane Controlled Trials Register, EMBASE, Medline and Psychlit. The Patient-assessed Health Instruments (PHI) bibliographic database (http://phi.uhce.ox.ac.uk/), hosted by the National Centre for Health Outcomes Development at the University of Oxford, was also searched. This database is based on systematic searches of the literature and contains over 7000 records relating to published instrument evaluations found on the major electronic databases (from database inception to September 2003) [3].

The reference lists of included articles were reviewed for additional articles. Relevant journals were hand-searched, including *Annals of the Rheumatic Diseases*, *Arthritis and Rheumatism*, *Rheumatology* and the *Journal of Rheumatology*. Texts and compendia were consulted [11–13]. The reference lists of existing reviews [8, 14–17] and manuscripts discussing the clinical management of AS [18, 19] were also reviewed.

**Inclusion criteria**

Titles and abstracts of all articles were assessed for inclusion/exclusion by two independent reviewers (K.L.H., A.M.G.) and agreement was checked. Published articles were included if they provided evidence of measurement properties for AS or arthritis-specific, multi-item, patient-assessed health instruments following completion by adults with AS. Generic or domain-specific multi-item instruments (not specific to AS or arthritis) were excluded. Clinician-assessed instruments, single-item and mobility measures, radiographic and imaging techniques were excluded. The review included instrument evaluations in non-English-speaking populations that were published in an English language journal. Instruments without evidence of reliability or validity were excluded.

**Data extraction**

Data extraction followed predefined criteria considered important in the evaluation of patient-assessed health instruments [4, 9, 11] and included patient characteristics, type of instrument, the domain focus, scaling, length, and evidence of measurement properties. The summary of evidence follows that of previous reviews [10, 11]. Evidence for measurement properties was assessed using accepted criteria [4, 9, 11]. Reliability assesses measurement stability over time, and for multi-item instruments, internal consistency [4, 9]. Test–retest reliability assesses score temporal stability and is assessed following instrument completion at two time points; it assumes no change in underlying condition [4]. Internal consistency reliability assesses the ability of items to measure a single underlying domain, and is assessed following a single application. Evidence for test–retest and internal consistency reliability, specifically test–retest correlation coefficients and Cronbach’s α, are presented. The reliability estimate reflects two components: a true score and an underlying level of error [9]. Reliability estimates range between 0 and 1.0; the closer the score to 1.0, the lower the error [4, 9]. The reliability of an instrument has implications for whether it is suitable for application in group or individual evaluation. For the evaluation of individuals high levels of reliability, above 0.90, have been recommended [4, 9]. For group comparisons, levels over 0.70 are recommended [9].

Validity assesses whether an instrument measures what is intended [4], and is referred to as ‘truth’ within the OMERACT filter [20]. Content and face validity are assessed through an appraisal of item content; evidence for the source of instrument items. Evidence for these forms of qualitative validity is presented. Evidence for external construct validity requires comparison of instrument scores with those for other measures of health, clinical, socio-demographic and health service use variables [4, 10]. Construct validity may also be assessed by group or divergent validity, where, based on theory or existing evidence, we can state that one group will possess more or less of a construct [9]. For example, compared with the general population, people with AS may be expected to report greater levels of pain and worse HRQL. The results of these comparisons are presented. Evidence derived from statistical methods such as factor analysis to describe instrument dimensionality or internal construct validity is presented [4].

Responsiveness has been described as the ability of an instrument to measure clinically important change over time when change is present [4], and is a necessary property of instruments intended for application in evaluative studies [21, 22]. Two broad approaches to evaluating responsiveness include those that are distribution-based and those that are anchor-based. Distribution-based approaches relate changes in instrument score to some measure of variability, the most common being the effect size statistic [4]. Effect size statistics provide a standardized unit of expression of the size and meaning of score change, supporting the comparison of instrument performance; the most responsive instruments have larger effect sizes. Anchor-based approaches assess the relationship between changes in instrument score and an external variable [23]. This includes health transition items or global judgements of change. Responses to transition items have also been compared to instrument score change using correlation. Data extraction covered the full range of approaches to measuring responsiveness and included descriptive statistics.

**Data summary**

The summary of reviewed evidence was informed by previous instrument reviews [11, 24, 25] (Table 1). The thoroughness (that is, the range of evaluations) and results of evaluations are considered separately. A summary scale from 0 to ++++ is presented, where 0 is no evidence for the underlying criteria and ++++ indicates a wide range of testing and good evidence of measurement properties (Table 1).

**Results**

**Identification of studies**

The literature searches produced 138 articles covering 12 AS-specific and three arthritis-specific multi-item, patient-assessed health instruments with evidence of reliability or validity following completion by patients with AS (Table 2).
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Developer</th>
<th>Domains (items)</th>
<th>Response scale</th>
<th>Score</th>
<th>Origin</th>
<th>Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symptoms and disease activity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bath AS Disease Activity Index (BASDAI)</td>
<td>Garrett <em>et al.</em> (1994) [28]</td>
<td>Disease activity (6)</td>
<td>VAS; adjectival anchors</td>
<td>0–10; 0 disease activity</td>
<td>UK</td>
<td>Multiple</td>
</tr>
<tr>
<td>Body Chart</td>
<td>Dziedzic (1997) [29]</td>
<td>Pain (1)</td>
<td>4-point descriptive</td>
<td>No max score; 0 no pain</td>
<td>UK</td>
<td>–</td>
</tr>
<tr>
<td><strong>Function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASAQ</td>
<td>Nemeth <em>et al.</em> (1987) [37]</td>
<td>Function/spinal mobility (2)</td>
<td>8-point adjectival</td>
<td>0–11; 0 best function</td>
<td>UK</td>
<td>–</td>
</tr>
<tr>
<td>BASFI</td>
<td>Calin <em>et al.</em> (1994) [5]</td>
<td>Function (10)</td>
<td>VAS; adjectival anchors</td>
<td>0–10; 0 best function</td>
<td>UK</td>
<td>Multiple</td>
</tr>
<tr>
<td>DFI</td>
<td>Dougados <em>et al.</em> (1988) [6]</td>
<td>Function (20)</td>
<td>3-point categorical</td>
<td>0–40; 0 best function</td>
<td>France</td>
<td>Multiple</td>
</tr>
<tr>
<td>DFI</td>
<td>Dougados <em>et al.</em> (1988) [6]</td>
<td>Function (20)</td>
<td>3-point categorical (revised to 5-point)</td>
<td>0–40; 0 best function</td>
<td>France</td>
<td>Multiple</td>
</tr>
<tr>
<td>HAQa</td>
<td>Fries (1980) 32</td>
<td>Arthritis-specific (core): Disability (20), Discomfort (1)</td>
<td>4-point categorical; 1 × 15 cm VAS</td>
<td>0–3; 0 best function</td>
<td>USA</td>
<td>Multiple</td>
</tr>
<tr>
<td>HAQ-S</td>
<td>Daltroy <em>et al.</em> (1990) [33]</td>
<td>(HAQ + AS-specific items) Disability (23), Discomfort (2)</td>
<td>4-point categorical; 2 × 15 cm VAS</td>
<td>0–3; 0 best function</td>
<td>USA/UK</td>
<td>Dutch</td>
</tr>
<tr>
<td>RLDQ</td>
<td>Abbott <em>et al.</em> (1994) [38]</td>
<td>Function (16)</td>
<td>4-point categorical</td>
<td>0–48; 0 best function</td>
<td>UK</td>
<td>Swedish</td>
</tr>
<tr>
<td><strong>Global well-being</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAS-G</td>
<td>Jones <em>et al.</em> (1996) [30]</td>
<td>Global well-being (2)</td>
<td>VAS; adjectival anchors</td>
<td>0–10; 0 best well-being</td>
<td>UK</td>
<td>–</td>
</tr>
<tr>
<td><strong>Health-related quality of life</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASQoL</td>
<td>Doward <em>et al.</em> (2003) [39]</td>
<td>HRQL (18)</td>
<td>Yes/No</td>
<td>0–18; 0 best HRQL</td>
<td>UK, NL</td>
<td>Dutch</td>
</tr>
<tr>
<td>AIM5a</td>
<td>Meenan <em>et al.</em> (1980) [40]</td>
<td>9 domains (45)</td>
<td>2–6 point categorical</td>
<td>0–10; 0 best health</td>
<td>USA</td>
<td>Multiple</td>
</tr>
<tr>
<td>AIM5a</td>
<td>Meenan <em>et al.</em> (1992) [41]</td>
<td>12 domains (57)</td>
<td>5 point categorical</td>
<td>0–10; 0 best health</td>
<td>USA</td>
<td>–</td>
</tr>
<tr>
<td>AS-AIMS2</td>
<td>Guillemin <em>et al.</em> (1999) [31]</td>
<td>13 domains (63)</td>
<td>5 point categorical</td>
<td>0–10; 0 best health</td>
<td>France</td>
<td>–</td>
</tr>
<tr>
<td>PETb</td>
<td>Bakker <em>et al.</em> (1995) [34]</td>
<td>Single index (15)</td>
<td>7-point descriptive</td>
<td>0–49; 0 best health</td>
<td>Canada/NL</td>
<td>–</td>
</tr>
<tr>
<td>PGI</td>
<td>Haywood <em>et al.</em> (2003) [36]</td>
<td>Single index (7)</td>
<td>10-point descriptive plus weighting</td>
<td>0–10; 10 best HRQL</td>
<td>UK</td>
<td>–</td>
</tr>
</tbody>
</table>

*Arthritis-specific instruments; b* interview-administered.
**Patient characteristics**

The number of respondents ranged from 14 [26] to 4282 [27]. All patients had a clinical diagnosis of AS. Mean respondent ages ranged from 30 to 57 yr. Mean disease duration ranged from 5.5 to 32.4 yr.

**Patient-assessed health instruments**

The Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) [28], BASFI [5] and DFI [6] have undergone the greatest amount of testing for all measurement properties, with 72, 70 and 46 published articles respectively (Table 3).

The shortest instruments were the Body Chart [29] and the Bath Ankylosing Spondylitis Global score (BAS-G) [30], with one and two items respectively, although a single item is often reported for the BAS-G (Table 1). The longest was the AS-specific Arthritis Impact Measurement Scale (63 items) [31]. Most instruments produce index scores; that is, item scores are summed to produce a single score. The Health Assessment Questionnaire (HAQ) [32] and the HAQ-S [33] produce profile scores; that is, item scores are summed within separate domains, providing a reflection of health across these domains. The Patient Elicitation Technique (PET) [34, 35] and Patient-Generated Index (PGI) [36] are individualized measures. With the exception of the PET, which requires interview administration [35], all instruments have been self-completed.

Instruments are grouped in Tables 2 and 3 according to the domains that they purport to measure: symptoms and disease activity (BASDAI, Body Chart); function (Assessment in Ankylosing Spondylitis Questionnaire (ASAQ) [37], BASFI [5], DFI [6], HAQ [32], HAQ-S [33] and the Revised Leeds Disability Questionnaire (RLDQ) [38]); global well-being (BAS-G [30]); and HRQL (ASQoL [39], AIMS [40], AIMS2 [41], AS-AIMS2 [31], PET [35] and PGI [36]).

### Table 3. Summary of evaluations for AS- and arthritis-specific patient-assessed health instruments

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Published evaluations (n)</th>
<th>Reliability</th>
<th>Validity</th>
<th>Responsiveness</th>
<th>Thoroughness</th>
<th>Results</th>
<th>Reliability</th>
<th>Validity</th>
<th>Responsiveness</th>
<th>Thoroughness</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symptoms and disease activity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BASDAI</td>
<td>72</td>
<td>17</td>
<td>37</td>
<td>15 (3)</td>
<td>25</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Body Chart</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0 (2)</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td><strong>Function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASAQ</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>BASFI</td>
<td>70</td>
<td>19</td>
<td>29</td>
<td>18</td>
<td>26</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>DFI</td>
<td>46</td>
<td>16</td>
<td>18</td>
<td>15 (1)</td>
<td>14</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>HAQ</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>1 (1)</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>HAQ-S</td>
<td>22</td>
<td>4</td>
<td>13</td>
<td>10 (2)</td>
<td>2</td>
<td>+</td>
<td>+</td>
<td>+++</td>
<td>++</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>RLDQ</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>3 (1)</td>
<td>0</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td><strong>Global well-being</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAS-G</td>
<td>18</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td><strong>Health-related quality of life</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASQoL</td>
<td>10</td>
<td>4</td>
<td>6</td>
<td>2 (3)</td>
<td>2</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>AIMS</td>
<td>8</td>
<td>0</td>
<td>7</td>
<td>2 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>AIMS2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AS-AIMS2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>PET</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>PGI</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0 (1)</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

---

**AS- or arthritis-specific, multi-item patient-assessed measures**

AS- or arthritis-specific, multi-item patient-assessed measures specific to the assessment of pain, stiffness or fatigue were not identified; where assessed, these domains were largely measured with single-item visual analogue scales.

**Reliability**

The BASDAI and BASFI have the greatest evidence of reliability (Table 2). Seven instruments have evidence of internal consistency reliability: BASDAI, BASFI, DF1, HAQ-S, RLDQ, ASQoL and AS-AIMS2. Alpha levels for studies evaluating the English BASFI [42], RLDQ [38, 43], ASQoL [39, 43] and a range of instrument translations (for example, the Turkish BASFI [44], Dutch [45] and Finnish versions of the DFI [46]) exceed 0.90, the criterion recommended for individual patients [4, 9]. Evidence for the BASDAI [42, 43, 47–49] and the Austrian HAQ-S [50] exceed 0.70, the level recommended for groups [4]. Evidence of item-total correlation for the BASDAI, RLDQ and the ASQoL support scale homogeneity [9]. Tests of internal consistency are not appropriate for the instruments that are not based on summated rating scales: Body Chart, BAS-G, PET and PGI.

Ten instruments have evidence of test–retest reliability: BASDAI, Body Chart, ASAQ, BASFI, DFI, RLDQ, BAS-G, ASQoL, AS-AIMS2 and PGI. All reliability estimates for the RLDQ [38, 43] and the ASQoL [39, 43] exceed criteria necessary for individual assessment. Several reliability estimates for the BASFI [46, 51, 52] and the DFI [5, 45, 46, 53] exceed the criteria necessary for individuals. Most estimates for remaining instruments exceed the criteria necessary for groups [9]. Low levels of test–retest reliability have been reported for the BASDAI (range 0.53–0.64) [54]. The retest period ranged from 1 day to 6 weeks. Few authors indicate if reliability is assessed in patients reporting no change in health over the retest period.
Four instruments have evidence of both forms of reliability—BASDAI, BASFI, DFI and RLDQ—where reliability estimates support application at the group and, in some instances, the individual level. One-week test–retest reliability and associated 95% limits of agreement were calculated for the BASDAI, BASFI, BAS-G and DFI [51]. High reliability estimates were associated with wide score ranges; this was interpreted as indicating poor instrument reliability [51].

Four instruments do not have evidence of reliability in patients with AS: HAQ, AIMS, AIMS2 and PET.

Validity
The ASQoL [39], PET [34] and PGI [36] involved patients in item generation. The BASDAI [28], BASFI [6], RLDQ [38] and AS-AIMS [31] incorporated the opinion of health-care professionals and patients but the role of patients is not made explicit. The DFI included three rheumatologists and the HAQ-S a survey of functional disability in patients with AS. With the exception of the ASAQ [37] and BAS-G [30], for which item generation is not described, the literature and existing instruments provided the major source of items for the remaining instruments.

Evidence supports the internal construct validity of live instruments, the results of which support the single domains of the BASDAI [43, 49], DFI [6], RLDQ [43] and ASQoL [43]. Although the developers of the AS-AIMS2 describe 13 domains, 12 domains were found following principal component analysis [31].

All instruments have evidence for validity through comparison with instruments that measure similar or related constructs, and/or with measures of mobility. This is most extensive for the BASDAI, BASFI and DFI (Table 3). With the exception of the HAQ, AIMS, AS-AIMS2 and PET, all instruments have evidence to support their ability to discriminate between groups of patients with AS defined by clinical, radiographic and socio-demographic variables. This is most extensive for the BASDAI, BASFI and DFI. The BASDAI, BASFI, DFI and HAQ-S have evidence to support their ability to discriminate between groups defined by health service use.

Responsiveness
With the exception of the ASAQ, AIMS2 and AS-AIMS, all instruments have some evidence of responsiveness following completion by patients with AS (Table 3). With the exception of the Body Chart, RLDQ, AIMS, PET and PGI, all instruments have some evidence of responsiveness following both drug therapy and physical therapy. This is most extensive for the BASDAI and BASFI. Effect size statistics were reported for all instruments. Correlation of change scores with change in other variables was reported for the BASDAI, BASFI, DFI, HAQ-S, BAS-G, ASQoL and AIMS. With the exception of the AIMS, Body Chart, RLDQ, PET and PGI, seven instruments had evidence of group discrimination over time. Statistical significance was frequently reported but the clinical significance of score change was rarely addressed. There is limited evidence for the Body Chart, HAQ, HAQ-S, RLDQ and instruments that measure HRQL. Five studies reporting evidence for the HAQ-S describe different stages in a trial of physical therapy [34, 55–58].

Moderate to strong levels of responsiveness were reported for the BASDAI [for example, 49, 59–63] and BASFI (for example, 49, 53, 59–66) following a range of placebo-controlled trials and the longitudinal evaluation of active drugs. Mean score changes greater than 2.0 (scale 0–10) were reported for both the BASDAI and BASFI following the evaluation of anti-TNF therapy with 6- to 52-week follow-up periods.

Few studies reported effect size statistics for the evaluation of instrument responsiveness following physical therapy; most reported mean score change. Moderate and small effect sizes were reported for the BASDAI and BASFI respectively following the longitudinal evaluation of in-patient rehabilitation [67]. Small effect sizes were reported for the BASDAI and BASFI following combined spa and exercise therapy [68, 69]. The BASDAI was responsive to improvement or deterioration in health following the evaluation of usual care [43]. Mean score change for the BASDAI [28, 54, 67–75] and BASFI [5, 54, 67–70, 72–74, 76, 77] did not exceed 1.9 and 1.3 respectively following all physical therapy interventions within a 2- to 40-week follow-up period. BASFI score change of less than 0.7 (scale 0–10) was reported following similar in-patient rehabilitation programmes of 3 weeks’ duration [67, 72, 74]; score change of less than 0.6 was reported following a 6-week out-patient exercise programme [78]. Non-statistically significant score change was reported following a long-term, prospective evaluation of function [72, 75].

Moderate to strong levels of responsiveness were reported for the DFI following a range of drug therapy evaluations. Lower levels of responsiveness and poor group discrimination have been reported following both physical therapy [5, 67]. There is limited evidence of responsiveness for a modified five-point response scale [54, 68, 74].

The HAQ has limited evidence of responsiveness following drug therapy evaluation in patients with AS but large score changes were reported following the longitudinal evaluation of Infliximab in AS [79]. Evidence suggests that the HAQ-S is not responsive to change in functional ability following physical therapy [34, 55, 58, 68, 80].

Larger score changes have been reported for the single BAS-G items following drug therapy evaluation [79, 82] than following physical therapy [67, 73]. Similarly, the ASQoL was more responsive to change following drug therapy (etanercept) [83, 84] than following physical therapy [68, 69] or usual care [43].

Precision
Floor effects were reported for the HAQ-S (25% scored 0) [56] and RLDQ [43]. Skewed score distributions were also reported for the DFI [16, 85]. Floor effects may be a function of limited item content and/or response options [4], and limit measure- ment discrimination and responsiveness. Normal score distributions were reported for the ASQoL, BASDAI, BASFI and PGI. Score distributions were not reported for the remaining instruments.

Discussion
To provide the most effective management in the care of individuals with AS it is important to determine how the disease and treatment affect health from the patient’s perspective. The application of patient-assessed health instruments has become increasingly important within the assessment of health-care [3, 4] and more specifically within rheumatology [3, 86]. Significant progress in the field has been made since the initial ASAS recommendations, which acknowledged that they could change following new research evidence [7, 8]. This review provides a timely assessment of these recommendations and the first detailed review of AS- and arthritis-specific multi-item, patient-assessed health instruments. The review will inform the appropriate selection of multi-item, patient-assessed health instruments to be used in clinical practice and research.

ASAS recommended five core assessment domains: pain, stiffness, function, global well-being and spinal mobility. From the 15 reviewed multi-item instruments, one assessed pain intensity (Body Chart); five AS-specific instruments (ASAQ, BASFI, DFI,
HAQ-S and RLDQ) and one arthritis-specific (HAQ) instrument assessed functional ability; and one instrument assessed global well-being (BAS-G). The BASDAI, an AS-specific measure of disease activity, and three AS-specific measures of HRQL (AS-AIMS2, ASQoL, PGI) include items to assess pain, stiffness and fatigue.

Pain, stiffness and fatigue are frequently described as important symptoms by patients [36, 87] and clinicians [68]. The Body Chart had poor evidence of measurement properties and is not recommended without further evaluation. Although evidence of completion rates were mixed, measurement properties support consideration of the BASDAI as a measure of disease activity. However, other important issues, such as item content and response format, should also be considered.

Six measures of functional ability were reviewed. The ASAQ, HAQ and HAQ-S have limited evidence of measurement properties following completion by patients with AS; the HAQ and HAQ-S also have evidence of poor data quality. The RLDQ has good completion rates and a very high level of reliability, but it is not recommended for application due to poor data quality, the limited range of functional disability assessed, and limited responsiveness.

Both the BASFI and DFI have acceptable levels of reliability. Although both have good evidence for construct validity and similar levels of responsiveness following drug therapy evaluation, the BASFI has better content validity and is more responsive following physical therapy. Further, evidence suggests that the BASFI is more responsive than the DFI in the early stages of treatment with anti-TNF therapies; comparable levels of responsiveness were found after 4 months of treatment [66]. This may be a function of the limited response options for the original DFI. The modified categorical response scale [74, 85] may improve responsiveness. However, patients often experience difficulty in completing VAS, the format of the BASDAI, BASFI and BAS-G, and reservations have been expressed about the interpretation, acceptability and feasibility of VAS scales [4, 9, 54]. Both patients [4, 11, 54] and clinicians [88, 89] have expressed preferences for categorical rating scales. A comparative evaluation of response formats by patients with AS supported a preference for categorical rating scales (49%), followed by numerical rating scales (38%) and visual analogue scales (9%) [54]. The initial ASAS recommendations for functional assessment suggested use of the BASFI or DFI. Evidence suggests that the BASFI and DFI differ quite broadly in terms of item content, response format, levels of patient (and clinician) acceptability and measurement properties in different settings. This suggests that reports of functional ability from these two instruments may not be directly comparable. Further consensus is required to recommend a single instrument for AS-specific functional assessment.

Evidence supports the ASAS recommendation of the BAS-G as a single-item assessment of global well-being [8]. However, single items provide a limited assessment of global health [4]. Since 1999, three AS-specific measures of HRQL have been identified: ASQoL, AS-AIMS2 and PGI. Three arthritis-specific measures of HRQL were also reviewed: AIMS, AIMS2 and PET. These lack evidence of reliability in AS, have limited evidence of validity and cannot be recommended for application in AS assessment without further evaluation. Patients were explicitly involved in item generation for the ASQoL, PET and PGI. Patient participation enhances content validity [4]. Although overlap between ASQoL items and PGI areas was expected, a concurrent evaluation found that several of the areas most frequently nominated by patients completing the PGI (body image, walking and work outside the home) are not addressed by items within the ASQoL [36]. The ASQoL purports to measure AS-related quality of life but does not include some of the 10 most important and frequently mentioned patient concerns [36].

The ASQoL has good completion rates, satisfactory data quality and scaling assumptions, high evidence of validity, and good evidence of responsiveness following drug therapy, but small to moderate levels following physical therapy [68]. The ASQoL has a dichotomous response scale which often fails to support sufficiently detailed descriptions of health [4, 90, 91]. The PGI and AS-AIMS2 are two new measures of HRQL. As a result, both have limited evidence for their measurement properties and further evaluation is recommended.

Instrument measurement properties are context-specific attributes that can differ across populations, settings and interventions [21, 22, 91]. ASAS identified three treatment settings: disease-controlling anti-rheumatic therapy (DC-ART), symptom-modifying antirheumatic drugs (SMARDs) and physical therapy, and clinical record keeping [7, 8]. Evidence for measurement and practical properties across these settings should be considered alongside item content and appropriateness to the clinical or research question when selecting an instrument. Although ASAS made several instrument recommendations, the majority of these were for single-item scales which may not provide an adequate assessment of health [4, 9]. Patient-assessed health instruments should provide an accurate assessment of disease impact and healthcare from the patient’s perspective [3]. The inclusion of multiple items within a questionnaire will provide a more detailed and informative assessment of the health domain or concept [4]. Although quick to complete, single items may be a poor surrogate for a patient’s perception of disease impact; content validity is likely to be lower and important information may be lost which hinders data interpretation and usefulness to clinical decision-making [4, 9].

Low levels of test–retest reliability, which did not support application in group assessment, have been reported for a domain-specific, multi-item measure of fatigue [Multidimensional Fatigue Inventory (MFI)] and a single-item VAS for fatigue severity in patients with AS [92]. MFI domains ranged from 0.57 (physical fatigue) to 0.75 (reduced motivation; mental fatigue); fatigue VAS 0.60. Although associated with moderate to good levels of responsiveness, further evaluation of measurement properties is required before either instrument can be recommended for use. Further concurrent evaluations between AS-specific and domain-specific, multi-item, measures of pain, fatigue and stiffness are recommended. Where appropriate, refinement of existing instruments is required before the development of new instruments; seeking the views of people with AS with regard to instrument format, relevance and mode of completion is strongly recommended [9].

Although a range of comparative evaluations exists, further comparative evaluations are required of multi-item AS-specific instruments, such as the modified DFI and BASFI, the newly developed measures of HRQL, and widely-used generic instruments, across different treatment settings. Particular attention should be paid to the evaluation of instrument responsiveness over longer periods, score interpretation and the role of patient-assessed health instruments in clinical decision-making and communicating treatment benefit. Recent quality improvement initiatives within the NHS recognize the importance of evaluating health-care organizations and technologies in terms of patient-assessed outcomes. For instance, the measurement of HRQL is central to the NICE technology appraisal process. With the increasing availability of new and expensive therapies in rheumatology, the challenge to provide more informative, responsive and relevant patient-based assessment of disease impact and treatment efficacy becomes ever more important.

All reviewed studies reported instrument evaluations following completion by groups of patients participating in clinical trials or longitudinal evaluations of care. Although widely accepted in clinical trials [4], there is currently little evidence to support the effectiveness of including patient-assessed health instruments in
The assessment of functional disability, consensus is required for the BASFI and DFI remain the instruments of choice for content and appropriateness of response formats. Although measure of disease activity, but there are issues relating to item domains. These domains are currently inadequately assessed largely measured with single-item visual analogue scales. Single health, were not identified; where assessed, these domains were specific to the assessment of pain, stiffness, fatigue or global improvement of 20 (scale 0–100), have been proposed following treatment with biologics [94]. Improvement criteria have not been recommended following physical therapy or routine practice. Smaller score changes following intensive physical therapy have been reported for the BASDAI and BASFI, with score change not exceeding 19 and 12 respectively (scale 0–100) [69–71]. Several studies reported mean BASFI score change of less than 7 (scale 0–100) [67, 72–74]. Moreover, in contrast to single-item scales, percentage score change in multi-item instruments may be an inappropriate simplification of score change interpretation, providing erroneous indications of treatment effectiveness or instrument responsiveness [95]. It is possible that a 20% score change in the middle of a scale may be quite different to a 20% change at the ends of the scale [22, 95]. Recent applications of item response theory have shown that a number of instruments have items that cluster around the middle of the scale hierarchy [95, 96]. This makes it ‘easier’ for patients scoring in the middle of the scale to register score changes following real changes in health relative to those positioned nearer the ends of the scale [95]. The level of change in health that is important to patients, the minimal clinically important difference (MCID), has not been widely reported and should be addressed in future research. Instruments should be administered longitudinally before and after treatment known to improve HRQL, and health transition questions should be included as external criteria of change [21]. Although generally less responsive than disease-specific instruments [96], generic instruments have been recommended for use alongside disease-specific instruments, and their performance in the assessment of patients with AS should also be considered. The SF-36 [97] is the most widely used generic health profile both generally [3] and in studies of AS patients. The SF-36 is the only generic instrument with evidence of responsiveness following both drug therapy and physical therapy in AS. The physical component summary (PCS) score and associated domains were both responsive [62, 66, 84] and had evidence of discriminative validity [62, 65, 66] following anti-TNF therapies. A concurrent evaluation following anti-TNF therapy reported comparable levels of responsiveness between the SF-36 PCS and physical function domain, the BASFI and patient’s global assessment of health [66]. In conclusion, this extensive review provides information necessary for the appropriate selection of AS-specific, multi-item, patient-assessed health instruments. Such instruments, specific to the assessment of pain, stiffness, fatigue or global health, were not identified; where assessed, these domains were largely measured with single-item visual analogue scales. Single items may provide a limited reflection of these important domains. These domains are currently inadequately assessed by AS-specific, multi-item patient-assessed health instruments. The BASDAI has acceptable measurement properties as a measure of disease activity, but there are issues relating to item content and appropriateness of response formats. Although the BASFI and DFI remain the instruments of choice for the assessment of functional disability, consensus is required for the recommendation of a single instrument. The domain of HRQL is important to patients and should also be considered as a core assessment domain.

<table>
<thead>
<tr>
<th>Rheumatology</th>
<th>Key messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Several core assessment domains are inadequately assessed by AS-specific, multi-item patient-assessed health instruments. Where assessed, these domains are largely measured using single-item scales.</td>
<td></td>
</tr>
<tr>
<td>- Single-item scales may provide a limited reflection of core assessment domains.</td>
<td></td>
</tr>
<tr>
<td>- Consensus is required for the assessment of functional ability.</td>
<td></td>
</tr>
<tr>
<td>- Instrument-concurrent evaluation in different settings with long-term follow-up and support for score interpretation is required.</td>
<td></td>
</tr>
</tbody>
</table>

The authors have declared no conflicts of interest.

Acknowledgements

The authors thank Dr Kryisia Dziedzic, Keele University, and Dr Jonathan Packham, Staffordshire Rheumatology Centre, for their helpful comments on an earlier version of this review. We would also like to thank the anonymous reviewers for their insightful comments.

References


69. Heikkila S. The Dougados functional index with the 5-point Likert scale is sensitive to change due to intensive physiotherapy and exercise in spondyloarthropathy. Clin Exp Rheumatol 2002;20:689–92.


87. Heikkila S. The Dougados functional index with the 5-point Likert scale is sensitive to change due to intensive physiotherapy and exercise in spondyloarthropathy. Clin Exp Rheumatol 2002;20:689–92.


Clinical Vignette

Bilateral calcaneal osteonecrosis in a patient with systemic lupus erythematosus

Patients with systemic lupus erythematosus (SLE) are at high risk of the development of osteonecrosis. Common sites of involvement are the femoral head, femoral condyles and proximal humerus. Calcaneal osteonecrosis is an extremely rare disorder.

The patient is a 41-yr-old man of African origin with SLE treated with prednisolone, azathioprine and hydroxychloroquine. In 2003 he presented with pain in his lower legs, maximal in his heels. On examination there were features of bilateral reflex sympathetic dystrophy and peripheral neuropathy.

Radiographs demonstrated osteopenia in several of the small bones of the feet. Magnetic resonance imaging showed areas of osteonecrosis within both calcanei posteriorly; there was also oedema and atrophy within the lower leg and intrinsic hind foot muscles, which was symmetrical. A neurophysiological study confirmed a severe symmetrical, sensorimotor axonal polyneuropathy of the lower limbs.

This is the first clinical case report of a patient with SLE who developed calcaneal osteonecrosis. Two cases with SLE have been described previously but in a radiological report that contained few clinical details [1].

Risk factors in our patient for the development of this rare complication included treatment with high-dose corticosteroids, active SLE, peripheral neuropathy and reflex sympathetic dystrophy.

M. W. Seymour, A. W. M. Mitchell, C. G. Mackworth-Young

Correspondence to: M. W. Seymour. E-mail: matthew.seymour@chelwest.nhs.uk