Diagnostic classification of shoulder disorders: interobserver agreement and determinants of disagreement

Andrea F de Winter, Marielle P Jans, Rob J P M Scholten, Walter Devillé, Dirkjan van Schaardenburg, Lex M Bouter

Abstract

Objectives—To assess the interobserver agreement on the diagnostic classification of shoulder disorders, based on history taking and physical examination, and to identify the determinants of diagnostic disagreement.

Methods—Consecutive eligible patients with shoulder pain were recruited in various health care settings in the Netherlands. After history taking, two physiotherapists independently performed a physical examination and subsequently the shoulder complaints were classified into one of six diagnostic categories: capsular syndrome (for example, capsulitis, arthritis), acute bursitis, acromioclavicular syndrome, subacromial syndrome (for example, tendinitis, chronic bursitis), rest group (for example, unclear clinical picture, extrinsic causes) and mixed clinical picture. To quantify the interobserver agreement Cohen's \( \kappa \) was calculated. Multivariate logistic regression analysis was applied to determine which clinical characteristics were determinants of diagnostic disagreement.

Results—The study population consisted of 201 patients with varying severity and duration of complaints. The \( \kappa \) for the classification of shoulder disorders was 0.45 (95% confidence intervals (CI) 0.37, 0.54). Diagnostic disagreement was associated with bilateral involvement (odds ratio (OR) 1.9; 95% CI 1.0, 3.7), chronic complaints (OR 2.0; 95% CI 1.1, 3.7), and severe pain (OR 2.7; 95% CI 1.3, 5.3).

Conclusions—Only moderate agreement was found on the classification of shoulder disorders, which implies that differentiation between the various categories of shoulder disorders is complicated. Especially patients with high pain severity, chronic complaints and bilateral involvement represent a diagnostic challenge for clinicians. As diagnostic classification is a guide for treatment decisions, unsatisfactory reproducibility might affect treatment outcome. To improve the reproducibility, more insight into the reproducibility of clinical findings and the value of additional diagnostic procedures is needed.


Shoulder disorders are associated with pain, restricted range of motion and disability, which in some cases may last for several years. Histology and physical examination are the cornerstones of the diagnosis of shoulder disorders. Differentiation between various shoulder disorders might be an important prerequisite for effective treatment. However, differential diagnosis of shoulder disorders is often difficult, as various extrinsic and intrinsic conditions may underlie shoulder pain. The complexity of the diagnosis of shoulder disorders is illustrated by the lack of consensus on the appropriate diagnostic criteria and the fact that several diagnostic classifications have been proposed.

In 1990, the Dutch College of General Practitioners developed guidelines for the diagnosis and management of shoulder pain, which are largely based on the concepts of Cyriax. According to Cyriax, the anatomical site of the lesion can be identified by a systematic examination of the shoulder joint and the cervical spine. Although his concepts are well known and widely used, formal evaluation is still scarce.

The reproducibility of a diagnostic classification determines, to a large extent, its usefulness for clinical practice and research. Existing data on the interobserver agreement on diagnostic classification of shoulder disorders have mainly been derived from small studies, and have yielded contradictory results. Therefore, in this study the interobserver agreement of the classification of shoulder disorders was assessed in a large population, according to the diagnostic criteria recommended in the guidelines of the Dutch College of General Practitioners. The second objective of this study was to identify the determinants of diagnostic disagreement.
Table 1  Diagnostic classification of shoulder disorders

<table>
<thead>
<tr>
<th>Diagnostic categories</th>
<th>Main criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Capsular syndrome (adhesive capsulitis, arthrosis, frozen shoulder, etc)</td>
<td>Capsular pattern: proportionally larger passive restriction of external rotation to glenohumeral abduction and internal rotation. Pain in C5 dermatome</td>
</tr>
<tr>
<td>2 Acute bursitis</td>
<td>Restriction of active and passive abduction, severe pain in C5 dermatome. Acute onset, no evident trauma</td>
</tr>
<tr>
<td>3 Acromioclavicular syndrome</td>
<td>Restriction of horizontal adduction. Pain in area of acromioclavicular joint and C4 dermatome</td>
</tr>
<tr>
<td>4 Subacromial syndrome (chronic bursitis, tendinitis, rotator cuff tears)</td>
<td>Painful arc during abduction. Pain in C5 dermatome. No restriction in passive range of motion. At least one positive resistance test.</td>
</tr>
</tbody>
</table>
| 5 Rest group                                                                            |Bursectic: variable/slight pain, no or slight muscle weakness  
Tendinitic: pain, no or slight muscle weakness  
Cuff tears: light pain, moderate to severe muscle weakness |
| Unclear clinical picture                                                                | Signs and symptoms do not meet the criteria of one of the other diagnostic categories              |
| Extrinsic causes                                                                       |Extrinsic causes such as cervical spine disorders, thoracic outlet syndrome, shoulder-hand syndrome, referred pain |
| Other intrinsic causes                                                                  |Glenohumeral instability                                                                          |
| No shoulder problems                                                                   |No signs and symptoms indicating intrinsic or extrinsic shoulder disorders*                        |
| 6 Mixed clinical picture                                                                |Signs and symptoms do meet the criteria for two diagnostic categories†                              |

*Between the inclusion by the general practitioners, physicians or rheumatologists and the physical examination by the examiners recovery of the shoulder problems might have occurred. † SDQ score (0–100).

Methods

PATIENTS

During a 20 month period, all consecutive eligible patients with incident or prevalent shoulder pain were invited to participate in this study by 20 general practitioners, two physicians working in an orthopaedic practice, and 20 secondary care rheumatologists. Patients were eligible for participation if they gave informed consent, were between 18 and 75 years of age, and were sufficiently competent to complete questionnaires (for example, no dementia). Patients with shoulder problems attributable to neurological, vascular or internal disorders, systemic rheumatic diseases, fractures or dislocations were not invited to participate.

DESIGN

Two examiners (MPJ and AFW), both physiotherapists with three years and 10 years of clinical experience, respectively, performed the diagnostic procedure, which consisted of standardised history taking, physical examination, and subsequent diagnostic classification. One of the examiners was leading the history taking in the presence of the other. Subsequently, both examiners independently performed a physical examination. In each case, the history taking examiner performed the first physical examination, and within one hour the other examiner performed the second physical examination, after which each examiner independently registered the diagnosis. The sequence of the examiners was randomly assigned. Before the study, the physical examination was standardised and trained and the criteria for the diagnostic classification were established. Moreover, during a pilot study among four patients the feasibility of the diagnostic procedure was tested.

Before undergoing the diagnostic procedure, the participants completed several questionnaires. The examiners were blinded for the results, because the answers given by the participants might have influenced the diagnostic assessment of the shoulder complaints.

Table 2  Main characteristics of participating patients recruited in different settings

<table>
<thead>
<tr>
<th></th>
<th>General practice (n=75)</th>
<th>Orthopaedic practice (n=83)</th>
<th>Clinic for rheumatology and rehabilitation (n=93)</th>
<th>Total (n=201)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (%)</td>
<td>61</td>
<td>52</td>
<td>69</td>
<td>66</td>
</tr>
<tr>
<td>Mean age in years (SD)</td>
<td>44 (13)</td>
<td>49 (11)</td>
<td>51 (11)</td>
<td>48 (12)</td>
</tr>
<tr>
<td>Dominant shoulder affected (%)</td>
<td>57</td>
<td>67</td>
<td>43</td>
<td>54</td>
</tr>
<tr>
<td>Precipitating cause (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unknown</td>
<td>43</td>
<td>24</td>
<td>41</td>
<td>39</td>
</tr>
<tr>
<td>strain/overuse</td>
<td>16</td>
<td>33</td>
<td>34</td>
<td>27</td>
</tr>
<tr>
<td>(minor) injury</td>
<td>17</td>
<td>12</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>sport injury</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>other causes</td>
<td>13</td>
<td>22</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Previous episode(s) of shoulder pain (%)</td>
<td>40</td>
<td>30</td>
<td>43</td>
<td>40</td>
</tr>
<tr>
<td>Duration of current episode (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;3 months</td>
<td>45</td>
<td>9</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>3–6 months</td>
<td>13</td>
<td>12</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>6–12 months</td>
<td>19</td>
<td>27</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>&gt;12 months</td>
<td>23</td>
<td>52</td>
<td>40</td>
<td>34</td>
</tr>
<tr>
<td>Sleep disturbances (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unable to lie on the involved shoulder waking up; cannot fall asleep</td>
<td>44</td>
<td>39</td>
<td>55</td>
<td>51</td>
</tr>
<tr>
<td>Pain at rest (%)</td>
<td>63</td>
<td>58</td>
<td>75</td>
<td>67</td>
</tr>
<tr>
<td>Mean pain score* (SD):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>at night</td>
<td>54 (32)</td>
<td>36 (32)</td>
<td>54 (27)</td>
<td>51 (30)</td>
</tr>
<tr>
<td>during the day</td>
<td>55 (27)</td>
<td>41 (27)</td>
<td>57 (23)</td>
<td>54 (26)</td>
</tr>
<tr>
<td>Mean SDQ score† (SD)</td>
<td>63 (25)</td>
<td>61 (24)</td>
<td>75 (19)</td>
<td>68 (23)</td>
</tr>
</tbody>
</table>

*Visual analogue scale (0–100). †SDQ score (0–100).
restriction of active or passive motion, range of motion (in degrees), presence or absence of a painful arc, presence or absence of a capsular pattern, degree of pain (none, slight, moderate, severe) and degree of muscle weakness (none, slight, moderate, severe). Subsequently, the shoulder complaints were classified into one of six diagnostic categories: capsular syndrome, acute bursitis, acromioclavicular syndrome, subacromial syndrome, rest group, and mixed clinical picture. Table 1 gives the main criteria for the six diagnostic categories. In addition, both examiners estimated independently the severity of the pain on a 100 mm visual analogue scale (VAS) ranging from 0 “no pain” to 100 “very severe pain”. Detailed information on the diagnostic procedure is available on request from the first author.

### Questionnaires

All patients recorded the severity of their pain, both at night and during the day, in the preceding week on a VAS ranging from 0 “no pain” to 100 “very severe pain”. Furthermore, they filled in the Shoulder Disability Questionnaire (SDQ), which consists of 16 questions pertaining to difficulties in performing various daily activities on the previous day. The total score ranges from 0 “no disability” to 100 “difficulty with all applicable items”. Personality traits (anxiety, anger, depression and optimism) were measured by means of the Self-Assessment Questionnaire-Nijmegen (SAQ-N).19-22

### Analyses

Percentage of agreement and Cohen’s κ, including 95% confidence intervals (CI), were calculated to quantify the interobserver agreement.23 24 The κ statistic was computed for the overall classification of shoulder disorders in the six categories and for each diagnostic category separately (dichotomous κ). Multivariate logistic regression analysis was applied to determine whether demographic and clinical characteristics, pain, functional status, and personality traits were determinants of overall diagnostic disagreement. Variables with p ≤ 0.25 for the χ² test were considered as candidates for multivariate logistic regression. The logistic regression model was fitted by backward selection of variables (removal criterion p>0.10). The predictive performance of the logistic model was assessed by means of the Hosmer-Lemeshow test (goodness of fit test; calibration) and the receiver operating characteristic (ROC) curve area (discrimination).25 26 Odds ratios (OR) and 95% CI were calculated.

### Results

Table 2 shows the main characteristics of the 201 patients. The severity of the shoulder problem was greater in patients recruited by rheumatologists than in patients recruited in other settings, expressed by more frequent sleep disturbances and pain during rest, and a lower functional status (higher SDQ-score).

Table 3 presents the diagnostic classification of shoulder disorders according to both examiners. The percentage of agreement was 60%, and the overall κ was 0.45 (95% CI 0.37, 0.54). In cases of disagreement (81 patients; 40%), the examiners frequently disagreed on whether the shoulder pain should be classified as a distinct category or a mixed clinical picture (42 patients; 21%); dichotomous κ = 0.14).

Univariate analyses showed that most indicators of high severity of the complaints were associated with disagreement. Multivariate logistic regression analysis revealed that bilateral involvement, long disease duration (> 6 months), and high pain severity (mean pain score according to both examiners > 7.2) were independently associated with diagnostic disagreement (table 4). The model fitted the data
well (goodness of fit test: $p = 0.46$) and the discrimination was acceptable (area under the ROC curve: 0.68).

Discussion
In this study, the interobserver agreement of the diagnostic classification of shoulder disorders was evaluated in a large number of consecutive patients ($n=201$) with varying severity and duration of complaints. Only a moderate interobserver agreement was found (percentage of agreement = 60%, $\kappa = 0.45$). Especially for patients with high pain severity, chronic shoulder complaints and bilateral involvement, it seems to be difficult to define the anatomical site of the lesion.

The patients for this study were enrolled in three different health care settings. An association between type of health care setting and diagnostic disagreement might be present when patients are more likely to be referred by general practitioners to other health care settings when shoulder disorders are difficult to diagnose. Before the study we expected that the disagreement would vary between the setting of the general practitioners (GPs) and the other two settings. However, the level of disagreement was similar for the patients recruited by the GPs (37.3% disagreement) and the clinic for rheumatology and rehabilitation (39.8%), whereas for the patients ($n=33$) recruited in the orthopaedic practice a higher level of diagnostic disagreement was found (54.5%). We have no clear explanation for this finding. The results of the multivariate logistic regression showed the same pattern when the patients of the orthopaedic practice were excluded from the analysis (data not shown).

In interobserver agreement studies varying experience of examiners might influence the level of reproducibility. In this study the examiners had a different level of clinical experience and this could have influenced the results. To lessen this influence, before the study, the examiners had already achieved theoretical consensus and the diagnostic procedure had been standardised. In routine daily practice less attention will be paid to standardisation of the physical examination. Therefore, it might be argued that, on average, the level of reproducibility in clinical practice would not reach the level in this study.

One earlier study also evaluated interobserver agreement on the basis of the diagnostic classification recommended in the guidelines of the Dutch College of General Practitioners. In this study, which compared the diagnosis of general practitioners with the diagnosis of physiotherapists in routine daily practice, frequent discrepancies were found ($\kappa = 0.32$). This was a disappointing result, considering the fact that the physiotherapists were not blinded for the diagnoses of the GPs.

Two other interobserver studies based on a slightly different diagnostic classification showed varying results. In the study carried out by Bamji et al, three consultant rheumatologists agreed on the diagnosis of shoulder disorders in less than 50% of the cases involved. In contrast, high agreement ($\kappa = 0.88$) between experienced physical therapists who examined 19 patients was found by Pellecchia et al. However, the clinical characteristics of their study population were not presented, so it is unclear whether the study population consisted of consecutive patients or a selected group of patients. Therefore, it is difficult to establish an explanation for the high level of reproducibility found in their study.

The unsatisfactory reproducibility reported in the various studies might be explained by the fact that the diagnostic categories are insufficiently mutually exclusive. If clinical findings are not clearly attributable to one single diagnostic category, clinicians have to decide which clinical findings are most prominent to differentiate between shoulder disorders. In this study, the examiners frequently had difficulties in classifying the shoulder disorders into distinct categories, given the number of cases classified as “mixed” or “unclear clinical picture”. Based on the same diagnostic classification, Sobel and Winters showed that with strict application of the criteria only 3% of the cases could be distinctly classified, whereas with less stringent application of the criteria and additional tests 50% of the cases could be classified into distinct categories.

Insufficient mutual exclusiveness might also explain why patients with high pain severity, chronic complaints, and bilateral involvement represent a diagnostic challenge. These patients probably meet the diagnostic criteria for more than one category. This is understandable, because for patients with severe pain many of the test results will be positive, which makes it difficult to assess the relation between local factors and complaints. The complexity will also increase if various shoulder disorders, or a combination of a shoulder disorder with extrinsic conditions, underlie the shoulder complaints. This increased complexity might explain why it is more difficult to classify patients with bilateral involvement and chronic complaints. Moreover, for chronic complaints it has been suggested that local factors might determine the initial location of complaints, but that reasons for persistence and recurrence may be more general, such as previous episodes and psychosocial factors.

What efforts should be made to improve the reproducibility of diagnostic classification of shoulder disorders? It has been postulated that diagnostic injections with a local anaesthetic or additional tests during physical examination, such as the Neer impingment test, are helpful in establishing a diagnosis. If those diagnostic procedures do offer a solution is unclear. Additional skills will be required to perform certain diagnostic procedures, and some procedures might result in an increase in patient discomfort. It can also be questioned whether imaging techniques, such as ultrasound and magnetic resonance imaging (MRI), are beneficial in the selection of a diagnosis. Ultrasound might be an important diagnostic procedure because it is non-invasive and the costs are low. Recently, a meta analysis was conducted to assess the
accuracy and reliability of ultrasound for shoulder disorders. After evaluation of 58 studies the authors concluded that the accuracy of ultrasound was acceptable and that the reliability of ultrasound is unknown. However, often the accuracy of ultrasound was assessed for the detection of partial or complete tears of the rotator cuff. Consequently, it remains unclear what the accuracy of ultrasound will be in a population with varying severity of shoulder complaints. Although in the medical literature MRI is considered to be a useful diagnostic procedure in the evaluation of shoulder pain, this procedure is time consuming and would increase health care costs.

It has been suggested that the exact localisation of the anatomical site of the lesion is a prerequisite for effective treatment. In most randomised clinical trials, the main selection criterion for patients is the diagnosis, based on history taking and physical examination. It can be questioned whether unsatisfactory treatment outcome in some patients is because of the difficulties involved in localising the lesion. Therefore, future research should demonstrate whether additional diagnostic procedures could increase reproducibility, and thereby also improve the outcome of treatment.

A less complicated diagnostic classification system is also proposed, to reduce the complexity of diagnosing shoulder disorders. In our study the reproducibility was assessed for a diagnostic classification that was based on the concepts of Cyriax. Other diagnostic classifications can be based on different diagnostic criteria and have a different reproducibility. Unfortunately, there are no studies that report on the reproducibility of the various clinical findings that underlie the diagnostic classification of shoulder disorders. It has, however, been shown that various diagnostic labels are applied, even when there is consensus on the clinical findings. More insight into the reproducibility of clinical findings and careful examination of the diagnostic criteria is obviously needed before new classification systems can be adopted.

In conclusion, distinguishing between distinct shoulder disorders on the basis of history taking and physical examination seems to be rather complicated. Especially patients with high pain severity, bilateral involvement, and chronic complaints represent a diagnostic challenge. Serious doubt about the reproducibility of the diagnostic classification of shoulder disorders raises the question whether diagnosis based on history taking and physical examination is actually beneficial in the choice of treatment. Future studies should therefore determine whether additional diagnostic procedures improve diagnostic agreement. Moreover, additional research is needed to investigate the sources of diagnostic disagreement attributable to interobserver differences in clinical findings. This might be helpful in reaching further consensus on the appropriate diagnostic criteria.

AFW designed the study, was responsible for data collection and analysis, and is the first author of the manuscript. MPJ contributed to the design of the study and was also involved in data collection. RJPM, who was Project Leader of the study, supervised the design and conduct of the study, together with De V. WD supervised the data analysis, LMB initiated the study and provided overall supervision. All authors contributed to the revision of the manuscript.

Funding: partially financed by the “Stichting Anna-Fonds” foundation.