Assessment of active inflammation in juvenile dermatomyositis: a novel magnetic resonance imaging-based scoring system

Warren R. Davis¹, James E. Halls², Amaka C. Offiah³, Clarissa Pilkington⁴, Catherine M. Owens³ and Karen Rosendahl³

Abstract

Objective. To assess the reproducibility of a novel scoring system that we have developed for the objective assessment of acute inflammatory change in JDM. This system defines markers of inflammatory change in four muscle groups and the surrounding soft tissues.

Methods. Forty-eight children (33 girls) underwent retrospective assessment of their MRI studies by two musculoskeletal paediatric radiologists for the presence of disease activity. Each observer performed the readings on two separate occasions. The degree of concordance between the two observers and between the two readings was assessed using kappa analysis. The reproducibility of the total score was determined using Bland–Altman analysis.

Results. There was fair to moderate agreement between the two observers for all the examined disease activity markers in all muscle groups. There was good intra-observer agreement between the two readings. There was no difference according to the side evaluated. The mean total score (out of 20) for Observer 1 was 7.9 and for Observer 2 was 7.5 (mean difference −0.4, 95% limits of agreement −6.8 to 6.0), while the mean total scores for Observer 1 were 9.0 for the first reading and 7.9 for the second reading (mean difference 1.0, 95% limits of agreement −2.6 to 4.6).

Conclusion. Markers of inflammatory change in JDM can be observed on MRI in a reliable fashion and have been used to make a reliable and objective scoring system. The accuracy of the proposed scoring system is acceptable for the single reader, although there is more variability between two different individuals.

Key words: Juvenile dermatomyositis, Magnetic resonance imaging, Scoring system, Reproducibility.

Introduction

JDM is a rare autoimmune disease that occurs in childhood. It is a systemic vasculopathy characterized clinically by proximal muscle weakness, raised muscle enzymes and characteristic skin rashes, including Gottron’s papules over the extensor surfaces and a heliotropic rash over the eyelids.

JDM is the most common inflammatory myopathy in childhood [1, 2]. The incidence of JDM in the UK and Ireland is estimated to be 1.9 per million children under the age of 16 years, with a median age of onset of 6.8–7.7 years. These estimates are similar to those from the USA (2.4 per million under 15) [3]. Girls are more commonly affected than boys (ratio 2.2:1 to 5.0:1) [4, 5].
The only validated diagnostic criterion currently used in JDM was described by Bohan and Peter in the 1970s [6, 7]. Definite diagnosis of JDM is made by the presence of the characteristic skin rash and three of four other diagnostic criteria: systemic proximal muscle weakness, elevated muscle enzymes, characteristic histopathological changes on muscle biopsy and myopathic electromyography. The presence of the rash and two criteria indicates probable disease, while the combination of a characteristic rash and one criterion is classified as possible disease.

The Bohan and Peter criteria do not, however, represent current clinical practice and, as such, revised criteria including the use of MRI studies are currently being developed. Indeed, MRI was identified as the most useful of a number of new proposed diagnostic criteria in a recent international consensus survey, and was available to many of the clinicians responding to this survey [8].

Before the introduction of treatment with steroids, the mortality of JDM was high, with one-third of the cases dying and a further third having severe permanent disability [9]. The early diagnosis and aggressive treatment of JDM with steroids and immunosuppressive drugs such as MTX has dramatically improved the morbidity and mortality of the condition [10]. The results of two recently published national registers only reported one death out of a combined total of 195 patients [5, 11].

MRI has been shown to be a useful imaging modality in the management of JDM, and has been used to help diagnose the condition, monitor treatment and identify sites for muscle biopsy [1, 12–14]. Increased signal intensity from inflammation within muscle correlates better with disease activity than do elevated creatine kinase levels [12, 15, 16]. The diagnostic sensitivity for MRI has been reported as being 76%, compared with 64% for creatine kinase [5].

Various different MRI sequences are used when imaging JDM. A combination of T1-, T2-weighted and sequences that use fat suppression techniques all provide useful information. T1-weighted images are used to demonstrate muscle atrophy and fatty infiltration in the presence of chronic muscle damage [12, 16]. T2-weighted sequences are useful in active JDM, as the increased signal intensity in the affected tissues [15, 17] reflects the presence of oedema in these tissues. T2-weighted relaxation times in muscle have been shown to be significantly different in patients with active JDM compared with inactive JDM and healthy children [12]. T2-weighted relaxation time can be used as a quantitative measure of muscle inflammation and has a good correlation with other measures of disease activity [12, 13]. T2-weighted fat-suppressed techniques and short time inversion recovery (STIR) sequences have also been shown to correlate with global disease activity and improve visualization of muscular abnormalities when compared with standard sequences [18, 19]. STIR sequences have also been shown to be useful in the detection of inflammatory changes in the skin, subcutaneous tissue and fascia, changes that are often undetected by standard assessment [17].

At present, there is no standardized or validated method for reporting the MRI changes present in active JDM. The purpose of this study was to develop a new scoring system for the quantitative assessment of active inflammation in JDM and to check the reproducibility of this scoring system.

Methods

Forty-eight children (33 girls) aged between 18 months and 15 years with a known diagnosis of JDM were included. The average age at presentation was 6 years and 11 months.

The MRI examinations were performed using a 1.5 T Siemens Magnetom Avanto (Siemens, Erlangen, Germany). The study protocol included three axial sequences through the thighs, namely a T1-weighted spin echo (SE) sequence [time to echo (TE) 20.0, time to repetition (TR) 599.0, field of view (FOV) 225 × 300, acquisition time 2 min 50 s], a STIR sequence [time to inversion (TI) 150.00, TE 94.0, TR 5900.0, FOV 233 × 310, acquisition time 3 min 56 s] and a T2-weighted SE sequence (TE 94.0, TR 3160.0, FOV 300, acquisition time 2 min 12 s) with a slice thickness of 8 mm. No sedation was used.

The studies were anonymized before being analysed off-line on an advanced workstation (Siemens Leonardo) by two paediatric radiologists with a special interest in musculoskeletal diagnostics, masked to previous findings. Each observer scored the studies on two separate readings, with an interval of at least 3 weeks.

Several markers for active JDM were assessed and a score derived for each of four muscle groups (gluteal muscles, hamstrings, quadriceps and adductors) based on the STIR-weighted axial images. STIR sequences were used because they demonstrated acute inflammatory change more clearly than the other sequences.

The first marker of disease activity was the degree of muscle inflammation; this was based on a 4-point score (none = 0, mild = 1, moderate = 2 and severe = 3). The degree of muscle inflammation was assessed by taking the overall impression of the entire muscle group, as opposed to obtaining the score from the worst affected focal area of change (Figs 1–4). We also scored the presence of soft-tissue oedema (absent = 0, present = 1) and perifascicular oedema (absent = 0, present = 1). A total score of 20 was possible for each side (Table 1). Standardization of the scoring was performed a priori, by scoring and discussing five examinations not included in the present study.

Statistical analysis

The degree of agreement between the two observers and between the two readings was assessed for inflammatory change using unweighted Cohen’s kappa (0–0.2 = poor agreement, 0.2–0.4 = fair, 0.4–0.6 = moderate, 0.6–0.8 = substantial, 0.8–1.0 almost perfect). The concordance between the total scores generated using our 20-point scoring system was examined using mean differences and 95% limits of agreement as advised by Altman and
Bland [20] (results are given for right thigh only, since no differences were seen between the two sides). All analyses were performed using SPSS software version 17. The institutional guidelines for research ethics were followed. Separate ethical approval for our study was not required as all patients were part of our institution’s JDM cohort study, which has blanket ethical and consent approval.

**Fig. 1** Axial STIR MRI sequence of both thighs demonstrating bilateral mild muscular oedema in the quadriceps muscles. The soft tissues and perifascicular layers are normal.

![Mild muscular oedema](image1)

**Fig. 2** Axial STIR MRI sequence of both thighs. Moderate increased signal is demonstrated in the quadriceps muscles bilaterally.

![Moderate muscular oedema](image2)
Results

Forty-one out of 48 examinations were considered pathological by both the observers, while Observer 2 considered an additional five examinations to be mildly pathological. The total mean (s.d.) inflammation scores were 7.9 (5.1) and 7.5 (6.5) for the two observers (right thigh), respectively. Table 2 lists the observer variability

Fig. 3 Axial STIR sequences of both thighs demonstrating mild, moderate and severe bilateral muscle oedema and perifascicular oedema. There is involvement of both the quadriceps muscles and the hamstrings. The score for muscle oedema is based on the overall impression of each muscle group, as opposed to the worst affected area. For example, the right quadriceps muscle would be scored 3, as the majority of the muscle group exhibits severe change.

Fig. 4 Axial STIR sequences demonstrating bilateral severe muscle oedema, perifascicular oedema and soft-tissue oedema.
for the total inflammation scores and for each muscle group separately. There was fair to moderate agreement between the two observers for all the examined markers (Tables 2–4). Kappa assessment was not possible for perifascicular oedema for the gluteal muscles on both readings and for the adductors on the second reading due to the small number of positive cases. No difference was demonstrated between the two sides. Good intra-observer agreement was also demonstrated between the two readings (Tables 2–4).

The mean total scores, with mean differences and 95% limits of agreement, are given in Table 5. The mean difference between the scores generated by the two observers was small (−0.4); however, the 95% limit of agreement was relatively wide. The mean difference in the intra-observer mean scores was also small (1.2 and −0.1), but the s.d. and limits of agreement were more consistent.

Since the inter-observer agreement for the gluteal and adductor muscles were only fair for all three inflammatory markers, we estimated a second total score based on the quadriceps and hamstring muscles alone, yielding mean total scores of 5.2 and 4.2, respectively, for the two observers, with a mean difference of 0.9, 1.7 and 95% limits of agreement of −2.5 to 4.3 (Table 6).

### Discussion

There is no current standardized method for the assessment of JDM on MRI. Kimball et al. [17] used a scoring system based on STIR sequences that assessed changes in muscle, skin, fascia and subcutaneous tissues, but did not examine the reproducibility of the scores generated by this system. In view of the anticipated establishment of MRI as a new key diagnostic criterion, we were keen to develop an objective scoring system for the assessment of active JDM on MRI. We set out to grade some of the key markers of acute disease activity and see if these markers could be assessed in a reproducible fashion. Our scoring system was developed from this assessment.

We have shown that several MRI markers of disease activity in JDM can be assessed in a reproducible fashion, with κ-values ranging from fair to moderate agreement between observers, and moderate to substantial agreement between observers in all assessed muscle groups. A relatively low inter-rater and high intra-rater variability of the total inflammation score, however, underscore the need for thorough standardization.

There was no significant difference between the three observed markers of disease activity, and thus all were used to help produce the final score. We also found that there was no difference in the scores generated by each side, reflecting the symmetrical nature of the changes, and therefore our proposed scoring system is confined to the right thigh.
Although our study showed that there was at least fair inter-observer agreement for all markers in all muscle groups, the $\kappa$-values were not as high as we would have liked, particularly for the inter-observer assessment of the gluteal and adductor muscles. Omitting these from the total score did improve the accuracy to some degree, however, at the cost of a reduced sensitivity, since three patients had involvement of these muscle groups alone. Again, this underscores the importance of an a priori standardization of the scoring technique.

Although the unweighted kappa analysis does not take into account the level of disagreement between observers, we chose this conservative and slightly more robust approach to demonstrate the agreement between and within observers. Using weighted kappa analysis, both linear and quadratic, would typically have given $\kappa$-values between 0.1 and 0.3 points higher, and should be accounted for when interpreting our results. Overall, the degree of intra-observer agreement was higher than the inter-observer agreement for all markers. In addition, the Bland-Altman analysis of the generated scores demonstrated that although the mean difference between the two readers was small (−0.4), the s.d. was large, with a relatively wide 95% limit of agreement. The Bland-Altman analysis of the intra-observer mean scores mirrored the kappa analysis and showed a better consistency for both readers when compared with the inter-observer agreement.

The increased variability in the scores generated between the two observers when compared with the intra-observer variability is intuitive. However, the difference also reflects the relatively limited standardization that was performed using only five cases. Despite this, the kappa and Bland-Altman analysis show that the scoring system itself is reliable if performed by a single observer. We feel that a more rigorous process would have produced greater inter-observer agreement. If the scoring system is to be adopted into clinical practice, we would advise that individuals using the system would need thorough training on how to use the system or, alternatively, the scoring be done by a single individual with access to reference images. Although this would be impractical for more common conditions, the small number of cases of JDM means that it would be possible for a single individual to score the disease activity in a single tertiary referral centre.

**Table 5** Comparison between the mean scores between and within two readers assessing active inflammation

<table>
<thead>
<tr>
<th>Inter-observer variation (reproducibility)</th>
<th>Score of Observer 1</th>
<th>Score of Observer 2</th>
<th>Mean diff., s.d. (95% limits of agreement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>7.9</td>
<td>7.5</td>
<td>−0.4, 3.2 (−6.8 to 6.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intra-observer variation (repeatability)</th>
<th>First score</th>
<th>Second score</th>
<th>Mean diff., s.d. (95% limits of agreement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score of Observer 1</td>
<td>9.0</td>
<td>7.9</td>
<td>1.0, 1.8 (−2.6 to 4.6)</td>
</tr>
<tr>
<td>Total score of Observer 2</td>
<td>7.4</td>
<td>7.5</td>
<td>0.1, 1.8 (−3.7 to 3.5)</td>
</tr>
</tbody>
</table>

Maximum possible score is 20.

**Table 6** Comparison between the mean scores based on the hamstring and quadriceps muscle alone

<table>
<thead>
<tr>
<th>Inter-observer variation (reproducibility)</th>
<th>Score of Observer 1</th>
<th>Score of Observer 2</th>
<th>Mean diff., s.d. (95% limits of agreement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>5.2</td>
<td>4.2</td>
<td>0.9, 1.7 (−2.5 to 4.3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intra-observer variation (repeatability)</th>
<th>First score</th>
<th>Second score</th>
<th>Mean diff., s.d. (95% limits of agreement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score of Observer 1</td>
<td>5.4</td>
<td>5.2</td>
<td>0.2, 1.1 (−2.0 to 2.4)</td>
</tr>
<tr>
<td>Total score of Observer 2</td>
<td>4.1</td>
<td>4.2</td>
<td>0.1, 1.3 (−2.5 to 2.7)</td>
</tr>
</tbody>
</table>

Maximum possible score is 10.
All our studies were scored using axial STIR sequences. Ostergaard et al. [21] showed that agreement rates for the assessment of RA were higher when two imaging plains were used. It is possible that our κ-values may have improved if coronal and axial sequences were used to generate the scoring systems. In addition, all of our studies were performed using a 1.5 T MR scanner. Subsequent technical developments such as 3T MR scanners and newer sequences are potentially more accurate and the use of this technology could, in theory, improve the reproducibility of the scoring system.

The current study is a comparatively small retrospective analysis of the cases referred to our unit. The small number of cases means that a power analysis was deemed inappropriate. However, 48 patients is a relatively large sample for a rare condition, and accounts for approximately one-third of the cases seen at this centre, and about one-fifth of the cases in the UK and Ireland National Registry.

The two observers in our study were both experienced paediatric radiologists, with a specialist interest in musculoskeletal imaging. Although the reproducibility of the findings would undoubtedly be lower in studies assessed by non-specialist radiologists, in practice the small number of patients means that the images could be reviewed and scored by specialist staff at a tertiary referral centre.

The current study has examined whether the scores generated by the proposed system are reproducible. The next step is to validate the scoring system. Validation could be performed by correlating the generated score with other markers of disease activity, including clinical parameters such as muscle strength, biochemical markers and histological data on the severity of the condition in the individual. The scoring system could only be used in a clinical setting once this validation process was completed. Comparison with long-term outcomes should also be performed. Our system could also potentially be used for prognostic purposes on initial presentation and to monitor the response to treatment.

In summary, markers of acute disease activity in JDM can be identified in a relatively reliable fashion, particularly by the same observer. In addition, the proposed scoring system produces relatively consistent results if used by a single observer. The relatively wide limit of agreement between the two observers highlights the need for a thorough period of standardization before the scoring system is used. Since no difference in total score was seen according to side, we recommend a scoring system based on the right side.

Rheumatology key messages

- Markers of inflammatory change in JDM can be observed on MRI in a reliable fashion.
- The proposed 20-point scoring system for JDM can be used to grade its severity.

Disclosure statement: The authors have declared no conflicts of interest.

References


