The handle http://hdl.handle.net/1887/36591 holds various files of this Leiden University dissertation.

**Author:** Navarro Compan, Maria Victoria  
**Title:** Aspects of methodology in assessing inflammation and damage in rheumatoid arthritis and axial spondyloarthritis  
**Issue Date:** 2015-12-03
Rate of adjudication of radiological progression in rheumatoid arthritis randomized controlled trials depending on preset limits of agreement: a pooled analysis from 15 randomized trials

V. Navarro-Compán, R. Landewé, H.A. Ahmad, C.G. Miller, D. Xu, R. Wolterbeek, D. van der Heijde

Rheumatology (Oxford) 2013;52:1404-7
ABSTRACT

Objective
The aim of this study is to provide data on the adjudication rate for a predetermined threshold of difference in change score between two readers in randomized controlled trials (RCTs).

Methods
Fifteen datasets from RCTs in RA were scored by 13 experienced readers as pairs according to the modified Sharp-van der Heijde method. The theoretical adjudication rates for thresholds of between 3 and 20 units were calculated. We investigated the influence of the number of time points within the same session, the length of the interval and disease duration on the adjudication rates.

Results
A total of 21,295 time points from 7,643 patients from 15 databases were included in the analysis. The adjudication rate was inversely related to the threshold. Higher adjudication rates were observed with a higher number of time points, longer time intervals, and in early versus established RA. The adjudication rates ranged from 0% to 22% depending on the scenario.

Conclusions
With trained and experienced readers, the adjudication rate in RA RCTs is low even with very conservative adjudication thresholds.
INTRODUCTION

Radiographically defined joint damage is an important outcome in RA, as it is the consequence of joint inflammation and is related to disability. Prevention of structural damage is included as one of the claims in the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) Guidelines on clinical investigation of new drugs for treatment of RA.

The gold standard imaging modality to assess the degree of structural damage attributable to RA is conventional radiography. Several methodologies have been developed to quantify the radiographic progression, but the modified Sharp methods are most frequently used for registration purposes of new drugs. The Sharp-van der Heijde (SHS) method is the recommended method by the EMA.

The quality of a scoring method is defined partially by its precision, which is the ability of the measurement to be consistently reproduced. Moreover, the degree of baseline and progression of radiographic damage in patients participating in RA randomized controlled trials (RCTs) is decreasing over time. Precision may be adversely affected by inter-observer variability. Common measures to restrain measurement error include the use of two readers and averaging their scores. If a difference in change score between readers exceeds a certain threshold, adjudication using a third reader or rereading films could be used to further increase precision.

The choice of threshold for adjudication is arbitrary and has historically been set at an inter-reader difference of 7-15 SHS units as compared with baseline. However, there are no published reports regarding the impact of a particular threshold nor which threshold is most optimal for a particular clinical trial. Regulatory authorities have expressed concerns if 20% or more of the cases are the result of adjudication for a given clinical trial. To our knowledge, no previous study has analyzed data to determine the number of cases that would result in adjudication at a predefined difference in change score between readers.

The principal aim of this study is to provide data on how a selected threshold for adjudication results in a specific adjudication rate between two trained readers using the SHS method. The secondary aim is to investigate if additional factors such as the number of time points within the same reading session, the length of the interval between baseline and follow-up visits or disease duration of patients included in the studies influence the adjudication rate.
METHODS

Data were extracted from 15 databases from RCTs for the approval of biologic treatments in patients with RA. We selected these studies because they had been used for product registration purposes using similar methodologies and evaluated according to the SHS method and detailed data were available. Thirteen experienced readers who received the same training had scored all digitized films as pairs on a 21 CFR Part 11 compliant read system deployed by BioClinica, Inc. The readers were blinded to patient identification, treatment, and order of the time points.

Available information for each RCT included the following variables: study duration, number and time interval of films, mean disease duration of the patients and erosion and joint space narrowing raw score per joint according to the SHS method.

The total SHS was calculated per visit and per reader, to obtain the change score per reader between baseline and all follow-up visits. Then, considering all treatment groups as one per trial, the adjudication rate per study was calculated for thresholds between 3 and 20 units for the difference in SHS change score between two readers from baseline to all follow-up visits.

We further investigated whether the number of time points within the same reading session, the length of the interval between baseline and follow-up visits and the disease duration of patients included in the studies influenced the adjudication rate. A cut-off of 3 years for mean disease duration was used to differentiate between RCTs in early and established RA studies.

Statistical analysis

For descriptive purposes, characteristics of the separate RCTs are presented by mean ± standard deviation (SD), while data across studies are shown as medians [interquartile range (IQR)/range]. The datasets were analyzed to yield the theoretical number of cases, represented as a percentage, that would be adjudicated to a third reader based on a predefined threshold for difference of change score between two readers as compared with baseline. For this purpose, we applied a marginal model using generalized estimating equations having identified study as cluster level and patients clustered within studies. Statistical analysis was done using SPSS software version 20.0.
RESULTS

Initially, 23,672 time points and 4,010,390 joint scores from 8,435 patients were included in datasets. From all RCTs, 2 studies had two time points, 10 studies had three time points and 3 studies had four time points analysed. A total of 10,577 (0.26%) joint scores from 131 (1.6%) patients were missing, primarily due to surgery or joint replacement. The difference in change score between the two readers was missing for 792 (9.4%) patients because at least the score of one time point for one of the readers was not available. The median (range) number- and percentage of missing patients for all studies was 48 (0-161) patients and 8.6% (0-19.1), respectively. Finally, a total of 21,295 time points from 7,643 patients were included in this analysis. The median (range) sample size of the studies was 517 (103-901) patients, and the number of time points within one reading session was two for 1,172 patients, three for 5,296 patients and four for 1,175 patients. Median (IQR) baseline radiological damage and progression scores across all studies were 32 (18-48) and 1.05 (0.5-2.0), respectively. Detailed characteristics of studies are shown in supplementary Table 1.

The adjudication rate was inversely related with the threshold for difference in change score from baseline between two readers (Figure 1), and this relationship was stronger as the thresholds decreased. The number of time points within the same reading session influenced the adjudication rate as follows: the higher the number of time points, the higher the adjudication rate for a given threshold.

Figure 1: Adjudication rate in relation to the threshold based on the number of time points (Tp) of the studies.
The length of the interval between baseline and time points also influenced the adjudication rate, especially when the allowed difference in change score from baseline between two readers was strict. The longer the time interval, the higher the adjudication rate, but it remained below 22% for all intervals and thresholds.

Four RCTs were classified as early RA while 11 studies included patients with established RA. For the same threshold, the adjudication rate was higher for the early RA studies versus the established ones, especially for low thresholds (Figure 2). In all cases, the adjudication rate remained below 20%, even with a very conservative threshold selection.

Figure 2: Adjudication rate at first follow-up time point based on disease duration.
DISCUSSION

This study shows the number of cases that would result in adjudication for a predetermined threshold of difference in change score between two readers in RA RCTs using the SHS score. It also provides rational data to select the optimal threshold when designing the imaging plan for a prospective trial. As expected, the percentage of adjudicated cases was inversely related to the threshold, and in most cases it remained below 20%, which has been mentioned by regulatory authorities as a reason for concern, even with a very low threshold.

The number of time points within the same reading session and the length of the interval between baseline and follow-up time points increased the adjudication rate; this was especially true when looking at low thresholds. The selection based on 3 years (in contrast to 1 or 2 years) disease duration might have influenced this result, as well as the relatively low number of trials in this group (n=4). Moreover, the percentage of adjudicated cases was higher in early RA RCTs compared to the established ones. Although the subsequent course of radiological progression in RA is highly variable, usually radiological damage increases over time and patients with a longer disease duration usually have higher SHS scores. However, the adjudication rate also remained below 20% in both groups.

One limitation of this study is that imputed scores were ignored, which may have influenced the adjudication rates. But only less than 0.3% of joints were missing, and we do not feel that ignoring this low number of missing values has influenced the difference in change score between readers. Additionally, all the readers were experienced and had received the same training and all used the SHS scoring method. It is unknown if these results could be extrapolated to other readers using different training or scoring systems. Moreover, all images were scored with unknown time order, so these results cannot be applied when images are scored with known time order. However, the SHS method is one of the methods that is accepted by agencies for the approval of a new treatment in patients with RA, being the setting that we explored in this study. Another possible limitation is that we selected a cut-off period of 3 years to define early RA, while a cut-off of 1 or 2 years is often used. The reason was to include a representative number of patients in both groups.

In conclusion, the results of this study give guidance as to the adjudication rate that can be expected given a predetermined threshold for the difference in change score between two readers in RA RCTs using the SHS score. When using trained and experienced readers, the adjudication rate in RCTs with patients with RA is quite low, even with very conservative thresholds less than 20%, which adds to the validity of the SHS method as a precise and reliable outcome measure in RA.
REFERENCES


### SUPPLEMENTARY MATERIAL

**Supplementary Table 1: Study Characteristics.**

<table>
<thead>
<tr>
<th>Study</th>
<th>Visits (n)</th>
<th>Initial patients (n)</th>
<th>Missed patients (n)</th>
<th>Included patients (n)</th>
<th>Mean disease duration (years)</th>
<th>Baseline SHS (mean ± SD)</th>
<th>SHS progression (mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>546</td>
<td>3</td>
<td>543</td>
<td>6</td>
<td>47.7 ± 55.0</td>
<td>0.91 ± 3.03</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>244</td>
<td>0</td>
<td>244</td>
<td>6</td>
<td>53.4 ± 53.1</td>
<td>2.00 ± 4.91</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1008</td>
<td>87</td>
<td>921</td>
<td>7</td>
<td>31.5 ± 50.2</td>
<td>0.53 ± 3.59</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>509</td>
<td>0</td>
<td>509</td>
<td>7</td>
<td>29.1 ± 48.0</td>
<td>0.71 ± 5.12</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>839</td>
<td>103</td>
<td>736</td>
<td>12</td>
<td>63.6 ± 71.2</td>
<td>1.76 ± 5.35</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>299</td>
<td>4</td>
<td>295</td>
<td>12</td>
<td>55.9 ± 60.5</td>
<td>2.03 ± 7.00</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>607</td>
<td>48</td>
<td>559</td>
<td>1</td>
<td>13.2 ± 26.8</td>
<td>1.05 ± 5.79</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>371</td>
<td>0</td>
<td>371</td>
<td>1</td>
<td>13.2 ± 28.1</td>
<td>1.29 ± 7.05</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>123</td>
<td>20</td>
<td>103</td>
<td>2</td>
<td>17.7 ± 20.1</td>
<td>5.71 ± 9.30</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>636</td>
<td>77</td>
<td>559</td>
<td>6</td>
<td>18.7 ± 32.8</td>
<td>0.70 ± 4.03</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>636</td>
<td>119</td>
<td>517</td>
<td>6</td>
<td>14.4 ± 27.0</td>
<td>0.40 ± 4.15</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>444</td>
<td>85</td>
<td>359</td>
<td>8</td>
<td>29.7 ± 43.6</td>
<td>1.06 ± 4.88</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>994</td>
<td>161</td>
<td>833</td>
<td>6</td>
<td>37.8 ± 46.4</td>
<td>0.35 ± 3.39</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>629</td>
<td>0</td>
<td>629</td>
<td>6</td>
<td>43.8 ± 54.2</td>
<td>0.53 ± 6.83</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>550</td>
<td>85</td>
<td>465</td>
<td>3</td>
<td>43.6 ± 43.3</td>
<td>4.48 ± 10.27</td>
</tr>
</tbody>
</table>