Rheumatoid arthritis

REVIEW

Systematic review of patient-reported outcome measures (PROMs) for assessing disease activity in rheumatoid arthritis

Jos Hendrikx,1,2 Marieke J de Jonge,1 Jaap Fransen,2 Wietske Kievit,3 Piet LCM van Riel1

ABSTRACT

Patient assessment of disease activity in rheumatoid arthritis (RA) may be useful in clinical practice, offering a patient-friendly, location independent, and a time-efficient and cost-efficient means of monitoring the disease. The objective of this study was to identify patient-reported outcome measures (PROMs) to assess disease activity in RA and to evaluate the measurement properties of these measures. Systematic literature searches were performed in the PubMed and EMBASE databases to identify articles reporting on clinimetric development or evaluation of PROM-based instruments to monitor disease activity in patients with RA. Two reviewers independently selected articles for review and assessed their methodological quality based on the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) recommendations. A total of 424 abstracts were retrieved for review. Of these abstracts, 56 were selected for reviewing the full article and 34 articles, presenting 17 different PROMs, were finally included. Identified were: Rheumatoid Arthritis Disease Activity Index (RADAI), RADAI-5, Patient-based Disease Activity Score (PAS) I & II, Patient-derived Disease Activity Score with 28-joint counts (Pt-DAS28), Patient-derived Simplified Disease Activity Index (Pt-SDAI), Global Arthritis Score (GAS), Patient Activity Score (PAS) I & II, Routine Assessment of Patient Index Data (RAPID) 2–5, Patient Reported Outcome-index (PRO-index) continuous (C) & majority (M), Patient Reported Outcome CLinical ARthritis Activity (PRO-CLARA). The quality of reports varied from poor to good. Typically 5 out of 10 clinimetric domains were covered in the validations of the different instruments. The quality and extent of clinimetric validation varied among PROMs of RA disease activity. The Pt-DAS28, RADAI, RADAI-5 and RAPID 3 had the strongest and most extensive validation. The measurement properties least reported and in need of more evidence were: reliability, measurement error, cross-cultural validity and interpretability of measures.

INTRODUCTION

Traditionally, the monitoring of rheumatoid arthritis (RA) in clinical trials and treat to target strategies has been based on indices, such as the Disease Activity Score (DAS), Disease Activity Score with 28-Joint Counts (DAS28), Clinical Disease Activity Index (CDAI) or Simplified Disease Activity Index (SDAI), involving formal joint counts performed by trained professionals.1–3 Formal joint counts, though valued for their information, have been criticised for their use in daily practice because of their time-consuming

Summary points

What is already known about this subject?

▸ Over the past years many Patient Reported Outcome Measures (PROMs) have been developed to measure disease activity in Rheumatoid Arthritis (RA), though information on these measures has been spread over numerous reports.

What does this study add?

▸ This study provides an overview of: available PROMs to measure disease activity in RA, which measurement properties have been assessed and the level of evidence of validation efforts.

▸ Of all patient-reported outcome measures in this review, Patient-derived Disease Activity Score with 28-joint counts (Pt-DAS28), Rheumatoid Arthritis Disease Activity Index (RADAI), RADAI-5 and Routine Assessment of Patient Index Data (RAPID) 3 had the strongest and most extensive validation.

▸ The measurement properties least reported and in need of more evidence are: reliability, measurement error, cross-cultural validity and interpretability of measures.

How might this impact on clinical practice?

▸ Physicians should be cautious when interpreting patient-reported outcome measures for disease activity and when comparing results of these instruments across different countries.
nature. With an increasing focus on patient-centred care, rising costs in healthcare and accompanying decreases in resources, patient-reported outcome measures (PROMs) might offer a patient-friendly, location independent, and time-efficient and cost-efficient means of monitoring chronic diseases such as RA. PROM research in rheumatology spans over 30 years, during which period various measures have been developed. These cover a broad spectrum of health domains, reflecting useful information from patients’ perspectives on the effectiveness of therapies tested in clinical trials. The Health Assessment Questionnaire Disability Index (HAQ), Rheumatoid Arthritis Disease Activity Index (RADAI) and Routine Assessment of Patient Index Data (RAPID) are well-known examples of PROMs for RA that are used in trials as well as in practice. Recently, though, other patient-reported measures reflecting their ‘physician-based’ counterparts, such as the Patient-derived Disease Activity Score with 28-joint counts (Pt-DAS28) or Patient-derived Simplified Disease Activity Index (Pt-SDAI), have also been developed. Until now, information about the measurement properties of all these patient-reported disease activity measures has been spread over numerous reports, hindering the comparison and choice of PROMs to monitor RA disease activity.

In order to understand how we can make good use of PROMs in daily practice, the first step needed is to have an overview of instruments suited to this task. Second, the level of evidence for the various measurement properties of each PROM has to be determined in order to make recommendations for clinical use. The objective of this study was therefore to identify PROMs to assess disease activity in RA and to evaluate their measurement properties.

METHODS

Consensus-based Standards for the selection of health Measurement Instruments (COSMIN http://www.cosmin.nl) were applied in this systematic review. The first step in the methodology recommended by COSMIN is the development of a search strategy. This strategy is a combination of five elements: a construct search, a population search, an instrument search, a validated PubMed filter for measurement properties and an exclusion filter. To retrieve as many PROM-based instruments as possible, a search strategy was developed with the emphasis on sensitivity rather than specificity. PubMed and EMBASE were searched to identify articles published between January 1994 and May 2014. Studies eligible for inclusion in the search results met the following criteria: English language, published in an international peer-reviewed journal, an adult RA population, a focus on clinimetric properties of PROM-based (without a formal professional joint count) instruments aimed at capturing disease activity or focused on the association of PROM-based instruments and disease activity measures. The focus on PROMs specifically addressing disease activity, rather than PROMs measuring other consequences of disease, was chosen in order to collect a comparable set of measures with respect to construct validity. The search strategy was refined with MeSH terms, keywords and free-text words, until a test-set of 11 target publications covering different PROM-based instruments was fully covered. A full specification of the search strategies is presented in online supplementary appendices I and II.

The second step of the review process involved independent evaluation by two assessors (WK and JH) of abstracts found by the search strategies. The selection criteria were as follows:

Inclusion criteria:
The article describes psychometric/clinimetric development or evaluation of a PROM-based instrument, without a formal joint count, for assessing disease activity in RA.

Exclusion criteria:
1. The article describes the above specifically for a juvenile population.
2. The article describes the above specifically for a population other than RA.
3. The article only describes results already presented in earlier articles.

Any discordance in abstract selection was discussed in a consensus meeting. Two assessors (MJdJ and JH) then read the full text of the remaining articles as a final check of eligibility.

In the third step of the review, the methodological quality of each included study was checked by two assessors (MJdJ and JH) independently using COSMIN checklists with a four-point rating scale ranging from poor to excellent. Each measurement property, out of a possible 10, was scored in a separate box containing 5–18 items referring to quality aspects for the respective measurement property (eg, sample size, description of missing items or statistical method used). The guidance given to rate each item of the reported measurement properties was followed and any existing discordance in scores between the assessors was relieved in a second consensus meeting. As recommended, a final overall rating for each measurement property, described in each study, was determined by taking the lowest rating of any item in the respective box. Additionally, the second lowest score was reported to give insight into the possibility of a single low score in a respective category determining the total score.

Finally, the study characteristics and clinimetric data were extracted from the included studies (see table 1 and online supplementary appendix III). For the interpretation of statistical measures being reported in studies, several suggestions have been stated. According to Nunnally and Bernstein, a Cronbach’s α of 0.8 is sufficient for research purposes and a value of 0.9 is recommended in case individual decisions are based on specific test scores. As a rule of thumb, Hinkle et al have...
those above 0.9 represent high accuracy. For poor accuracy, those from 0.7 to 0.9 are moderate and under the curve (AUC) values from 0.5 to 0.7 represent relation. For Cohen’s κ as a measure of agreement, several different categorisations have been proposed, though can largely be regarded as: <0.4 poor, 0.4–0.6 fair/moderate, 0.60–0.80 substantial/good, 0.80–1.00 excellent/almost perfect. According to Swets, area under the curve (AUC) values from 0.5 to 0.7 represent poor accuracy, those from 0.7 to 0.9 are moderate and those above 0.9 represent high accuracy. For the overall overview of measurement properties across the included studies (table 2), the following values were considered as positive indicators of the respective measurement property: Cronbach’s α ≥0.80, correlation coefficients ≥0.60, Cohen’s κ ≥0.60, AUCs ≥0.70. Since there is a lack of guidance for categorisation of the magnitude of measurement error, we considered the measurement error to be positive if it was on par or smaller than similar physician-reported measures (eg, DAS28 or SDAI) that were reported in the same study. The overall quality and consistency of evidence for the measurement properties of each instrument (evaluated over multiple

Table 1 Characteristics of the study population of included studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Size</th>
<th>Age (mean ± SD)</th>
<th>Female (%)</th>
<th>RF+ (%)</th>
<th>Disease duration (mean ± SD)</th>
<th>DAS28 at baseline (mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blanchais et al</td>
<td>26</td>
<td>56.6 (9.5)</td>
<td>NR</td>
<td>NR</td>
<td>16.6 (10) y</td>
<td>NR</td>
</tr>
<tr>
<td>Bossert et al</td>
<td>200</td>
<td>57 (11.5)</td>
<td>75.5</td>
<td>78</td>
<td>13.8 (3) y</td>
<td>3.61 (1.43)</td>
</tr>
<tr>
<td>Castréjon et al</td>
<td>39</td>
<td>56.8 (13.9)</td>
<td>90</td>
<td>NR</td>
<td>3.5 (NR) y</td>
<td>NR</td>
</tr>
<tr>
<td>Castréjon et al</td>
<td>720</td>
<td>48.2 (12.4)</td>
<td>76.2</td>
<td>NR</td>
<td>4.8 (2.9–7.0)* m</td>
<td>5.1 (1.3)</td>
</tr>
<tr>
<td>Choy et al</td>
<td>322</td>
<td>60.3 (23–87)</td>
<td>76</td>
<td>81</td>
<td>9.13 (0–48)† y</td>
<td>NR</td>
</tr>
<tr>
<td>Fransen et al</td>
<td>92</td>
<td>52 (13)</td>
<td>83</td>
<td>NR</td>
<td>9 (4–14)* y</td>
<td>NR</td>
</tr>
<tr>
<td>Fransen et al</td>
<td>584</td>
<td>59 (12)</td>
<td>72</td>
<td>69</td>
<td>8 (3–15)* y</td>
<td>4.3 (1.4)</td>
</tr>
<tr>
<td>Fujisawa et al</td>
<td>250</td>
<td>59.3 (14)</td>
<td>78.4</td>
<td>NR</td>
<td>10.35 (9.83) y</td>
<td>NR</td>
</tr>
<tr>
<td>Harrington</td>
<td>185</td>
<td>63 (22–88)</td>
<td>NR</td>
<td>NR</td>
<td>18 (2–51)* y</td>
<td>NR</td>
</tr>
<tr>
<td>Heegaard et al</td>
<td>30</td>
<td>60 (15)</td>
<td>77</td>
<td>70</td>
<td>15 (6) y</td>
<td>3.5 (1.0)§</td>
</tr>
<tr>
<td>Houssien et al</td>
<td>100</td>
<td>57.7 (12.2)</td>
<td>78</td>
<td>NR</td>
<td>11.5 (8.3) y</td>
<td>4.24 (1.3)</td>
</tr>
<tr>
<td>Janta et al</td>
<td>69</td>
<td>60.12 (13.16)</td>
<td>76.8</td>
<td>NR</td>
<td>11.9 (8.6) y</td>
<td>2.55 (1.0)</td>
</tr>
<tr>
<td>Kavanaugh et al</td>
<td>218</td>
<td>54.3 (21–88)</td>
<td>NR</td>
<td>81.7</td>
<td>8.5 (NR) y</td>
<td>5.4 (1.3)</td>
</tr>
<tr>
<td>Leeb et al</td>
<td>169</td>
<td>57 (19–78)</td>
<td>79.8</td>
<td>50</td>
<td>7.2 (0.2–46)† y</td>
<td>3.51 (0.28–6.67)†</td>
</tr>
<tr>
<td>Leeb et al</td>
<td>108</td>
<td>59.5 (24–87)</td>
<td>77.7</td>
<td>54</td>
<td>NR</td>
<td>2.95 (0.43–6.24)†</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>63</td>
<td>58.5 (19.7)*</td>
<td>ref</td>
<td>ref</td>
<td>3.5 (8.8)* y</td>
<td>ref</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>557</td>
<td>54.6 (20.9)*</td>
<td>ref</td>
<td>ref</td>
<td>6.82 (NR)/6.83(NR)</td>
<td>ref</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>278</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>6.89 (NR)/6.88(NR)</td>
<td>ref</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>1384</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>982</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>557</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>227</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>285</td>
<td>57.4 (14.6)</td>
<td>73</td>
<td>NR</td>
<td>9.7 (9.0) y</td>
<td>3.4 (1.7)</td>
</tr>
<tr>
<td>Pincus et al</td>
<td>200</td>
<td>53.4 (16.2)</td>
<td>81</td>
<td>NR</td>
<td>11.6 (10.8) y</td>
<td>3.7 (1.5)</td>
</tr>
<tr>
<td>Riazzoli et al</td>
<td>47</td>
<td>50 (13)</td>
<td>79</td>
<td>86</td>
<td>9.4 (8.6) y</td>
<td>5.4 (1.2)</td>
</tr>
<tr>
<td>Rintelen et al</td>
<td>392</td>
<td>61 (20–87)</td>
<td>82.1</td>
<td>59.4</td>
<td>62 (3.545)† m</td>
<td>3.26 (0.49–8.09)†</td>
</tr>
<tr>
<td>Rintelen et al</td>
<td>705</td>
<td>62.7 (13.4)</td>
<td>75.9</td>
<td>54.4</td>
<td>97.3 (98.0) m</td>
<td>3.31 (1.37)</td>
</tr>
<tr>
<td>Salaffi et al</td>
<td>191</td>
<td>56.6 (12.2)</td>
<td>82.7</td>
<td>NR</td>
<td>5.1 (5.5) y</td>
<td>6.02 (1.15)</td>
</tr>
<tr>
<td>Salaffi et al</td>
<td>196</td>
<td>56.7 (12.1)</td>
<td>83.1</td>
<td>78</td>
<td>5.1 (5.9) y</td>
<td>3.94 (2.03)</td>
</tr>
<tr>
<td>Singh et al</td>
<td>247</td>
<td>58.1 (11.2)</td>
<td>80.1</td>
<td>76</td>
<td>6.2 (6.6) y</td>
<td>5.31 (0.28)</td>
</tr>
<tr>
<td>Singh et al</td>
<td>200</td>
<td>42.2 (NR)</td>
<td>83</td>
<td>NR</td>
<td>4.9 (NR) y</td>
<td>5.2 (1.6)</td>
</tr>
<tr>
<td>Stucki et al</td>
<td>55</td>
<td>60.0 (14.6)</td>
<td>62</td>
<td>NR</td>
<td>5.1 (1.3–10.7)* y</td>
<td>NR</td>
</tr>
<tr>
<td>Sullivan et al</td>
<td>740</td>
<td>57 (13.7)</td>
<td>83</td>
<td>63.8</td>
<td>14.3 (12.3) y</td>
<td>4.05 (1.5)§</td>
</tr>
<tr>
<td>Uhlig et al</td>
<td>28</td>
<td>61.1 (6.2)</td>
<td>64</td>
<td>64</td>
<td>16.6 (10.4) y</td>
<td>3.12 (1.27)</td>
</tr>
<tr>
<td>Veenhof et al</td>
<td>191</td>
<td>54.5 (13.3)</td>
<td>71</td>
<td>NR</td>
<td>7.0 (3–17) y</td>
<td>5.42 (1.07)</td>
</tr>
<tr>
<td>Wolfe et al</td>
<td>9078</td>
<td>62.2 (12.6)</td>
<td>78.2</td>
<td>NR</td>
<td>16.2 (10.9) y</td>
<td>NR</td>
</tr>
</tbody>
</table>

*Median(IQR): Values are mean (standard deviation) unless otherwise indicated.†Mean (range).‡Median (range).§Median(SD).CRP, C-reactive protein; DAS28, Disease Activity Score with 28-Joint Counts; m, months; NR, not reported; ref, reference to results in earlier publication; RF+, rheumatoid factor positive; y, years.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Internal consistency</th>
<th>Reliability</th>
<th>Measurement error</th>
<th>Content validity</th>
<th>Structural validity</th>
<th>Cross-cultural validity</th>
<th>Hypothesis testing</th>
<th>Criterion validity</th>
<th>Responsiveness</th>
<th>Interpretability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.84–0.91 ( \alpha )</td>
<td>0.92 ICC</td>
<td>14.9%maxSDD</td>
<td>?</td>
<td>++</td>
<td>?</td>
<td>++</td>
<td>0.48–0.83s</td>
<td>0.77 AUCresponder</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.76–0.88</td>
<td></td>
<td></td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>0.89s</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Pt-DAS28</td>
<td>0.92p</td>
<td>23.2%meanSDD</td>
<td></td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>++</td>
<td>0.73–0.94p</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Ft-SDAI</td>
<td>?</td>
<td>+</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>0.87s 0.93p</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>RAPID 2</td>
<td>++</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>0.64–0.91s</td>
<td>0.80 AUCresponder</td>
<td>0.43–0.53 wκ</td>
</tr>
<tr>
<td>RAPID 3</td>
<td>0.87 ( \alpha )</td>
<td>0.90</td>
<td>14.8%maxSDD</td>
<td>?</td>
<td>++</td>
<td>?</td>
<td>+</td>
<td>0.69–0.71s</td>
<td>0.70s</td>
<td>?</td>
</tr>
<tr>
<td>PRO-CLARA</td>
<td>+</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>0.84s</td>
<td>0.82 AUCresponder</td>
<td>?</td>
</tr>
<tr>
<td>PRO-Index C/M</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>0.54s</td>
<td>0.50s</td>
<td>?</td>
</tr>
</tbody>
</table>

+, positive result; −, negative result (for explanation of the categorisation of the levels, see table 3); %maxSDD, SDD as percentage of maximum value of outcome; %meanSDD, SDD as percentage of mean value of the outcome; \( \alpha \), Cronbach’s \( \alpha \); \( \tau \), Kendall’s \( \tau \); AUCflare, area under the curve for patients with flare versus no flare; AUCresponder, area under the curve for patients responding to therapy versus not responding; GAS, Global Arthritis Score; ICC, Interclass correlation coefficient; MDJC, Medical Doctor Joint Count; \( p \), Pearson’s correlation coefficient; PAS, Patient Activity Score; PDAS, Patient-based Disease Activity Score; PRO-Index C/M, Patient Reported Outcome Index continuous (C) & majority (M); Pt-DAS28, Patient-derived Disease Activity Score with 28-joint counts; PRO-CLARA, Patient Reported Outcome Clinical ARthritis Activity; PJUC, Patient Joint Count; Pt-SDAI, Patient-derived Simplified Disease Activity Index; RADAI, Rheumatoid Arthritis Disease Activity Index; RAPID, Routine Assessment of Patient Index Data; s, Spearman’s correlation coefficient; SDD, smallest detectable difference; wκ, Weighted Kappa.
studies shown in online supplementary appendix III) was summarised using a method originally proposed by the Cochrane Back review group and that has been used by others since (table 3). Depending on the presence of either one or more studies of fair, good or excellent methodological quality, and the consistency of findings across studies, the level of overall evidence ranges from unknown to strong (table 3).

RESULTS

The search strategy resulted in 358 articles in PubMed and 275 articles in EMBASE. The two search strategies had a 32% overlap, resulting in 424 articles to be reviewed (figure 1). Independent assessment of the abstracts resulted in 94% concordance and consensus was reached after discussing the remaining abstracts. Discordance was mostly due to discussion if the article was aimed at validating PROM-based instruments intended to measure disease activity. After consensus, 56 abstracts were included for full review and 368 were excluded.

Of the 56 articles that were retrieved for full-text review, 22 were excluded. Reasons for exclusion were as follows: the reported instrument was not specifically developed to assess disease activity; the reported instrument was not PROM based; the article reviewed results of earlier publications; the report did not focus on a clinimetric evaluation or the report did not provide subgroup analyses for the RA subpopulation.

The 34 articles included for full review described the following instruments: Pt-SDAI, Patient-derived Disease Activity Score with 28-joint counts (Pt-DAS28), Patient-Based Disease Activity (PDAS) I, PDAS II, RADAI, RADAI-5, Patient Activity Score (PAS) I, PAS II, RAPID 3, Routine Assessment of Patient Index Data 4-Patient Joint Count (RAPID 4-PJC), Routine Assessment of Patient Index Data 4 Medical Doctor Joint Count (RAPID 4-MDJC), RAPID 5, Patient Reported Outcome-index (PRO-index) majority (M), continuous (C), Patient Reported Outcome CLinical ARthritis Activity (PRO-CLARA). An overview of the basic study characteristics is given in table 1. Most reports focused on 2 or 3 out of 10 possible measurement properties (see online supplementary appendix III). Aspects of validity and responsiveness were evaluated most frequently, whereas aspects of interpretation, cross-cultural validity, content validity, measurement error and reliability were seldom or not investigated. The quality of individual studies ranged from poor to excellent. Most noted reasons for poor scores were: not reporting missing items, not reporting how missing items were dealt with and poor choice of statistical measures.

Levels of evidence, over multiple studies, for each of the 17 instruments are shown in table 2. Overall, most instruments had limited or moderate levels of evidence for 3–5, out of a possible 10, measurement properties. The four instruments with the most extensive validations and strongest levels of evidence were: Pt-DAS28, RADAI, RADAI-5 and RAPID 3.
DISCUSSION
In this study, PROM-based instruments for disease activity in RA were identified and their measurement properties were systematically reviewed based on the COSMIN method. There is a large body of research related to patient-reported outcomes with inconsistent usage of terms describing outcome measures that are patient reported and terms describing different aspects of clinimetric validation. A lot of work has been carried out validating several PROM-based instruments to capture disease activity, though none of the identified instruments have good quality validation studies covering all clinimetric domains (table 2). All the information gathered in this review will be taken up in the European League Against Rheumatism (EULAR) Outcomes Measures Library (OML; http://oml.eular.org/) in order to create an openly accessible database of PROM-based measures, which can be updated with new information as it becomes available.

The first part of this review involved identifying reports which described the clinimetric/psychometric evaluation of PROM-based instruments to assess disease activity in RA. The two search strategies (PubMed and EMBASE) resulted in a substantial amount of unique candidate articles and 32% overlap in search hits (see figure 1, online supplementary appendices I and II). This demonstrates the value of not restricting search efforts to only one major referencing database.

In the second review round, 56 reports were included for full-text evaluation of PROM-based instruments and their measurement properties. Most candidates did not meet the inclusion criterion that the article described in a clinimetric/psychometric evaluation of a PROM-based instrument. This was to be expected as the search strategies (see online supplementary appendices I and II) were developed with a focus on sensitivity not specificity, due to a lack of consistent terminology for describing clinimetric evaluations and PROMs in the literature until now. It is notable that the Rheumatoid Arthritis Impact of Disease (RAID) instrument was not selected for this systematic review. This decision was made because this instrument was designed to capture ‘patients’ perception of the impact of the disease on domains of health. This covers a broader construct of health, including, for example, emotional well-being, when compared with the more ‘biologically’ oriented clinical indices of disease activity. Additionally, the focus of the RAID is on the impact of disease, which can be moderated by coping. These differences make the RAID heterogeneous to the instruments specifically focusing on RA disease activity, which this review was aimed at, and therefore less comparable, especially with regard to assessment of validity.

The third part of this review involved rating the level of evidence for each measurement property reported in the 34 articles selected in step 2. As recommended by COSMIN, the level of evidence was determined by the lowest score of all quality items for each measurement property. Almost all articles failed to report the number of missing items, or did not describe how missing data were handled, reducing the quality rating of the evidence (see online supplementary appendix III). In order to adequately evaluate an instrument, it is important to know if certain items are often missing and why this is so. The issue of failure to report missing data and their handling is not restricted to clinimetric evaluations; between 2006 and 2014, it has been reported as 1 of the 14 most frequently given review comments in the *Annals of Rheumatic Diseases*. We encourage authors and journals to place more emphasis on clear reporting of the occurrence and handling of missing data in the respective methods and results sections of the reports. Another aspect that was not clearly reported was the measurement model of the instrument. The COSMIN guide differentiates between reflective and formative models. Reflective models consist of items which are a manifestation of the same underlying construct. Also known as effect indicators, these items are expected to be highly correlated and interchangeable. Formative models consist of items that together form the construct. These items do not need to be correlated and internal consistency is therefore not relevant. Judging by the content of most instruments, these are probably based on formative models. Since there was no clear description of the measurement models, we scored internal consistency measurement properties as suggested by COSMIN guidelines. It should, however, be noted that these scores are most likely not relevant to the reported instruments and should therefore not be taken into account when judging its clinimetric quality. Of further note, some authors chose to refer to earlier reports for the description of the study population, which we would not advise. This hampers readers from adequately judging reported instruments, as the diversity of the study population can severely impact the evaluation of measurement properties.

In the fourth and final step of the review, all the available evidence of each instrument was compiled into table 2. It can be seen that the result of using the classification method proposed in table 3, that it is not necessary the case that the PROMs which have been most published on, such as RADAI or RAPID 3, are thereby automatically the ‘best’ scores with regard to evidence. This is due to the quality of each individual validation study or the presence of conflicting findings across studies. Furthermore, explicit judgement on which are the ‘best’ scores is not given because that is reliant on the purpose for which the instruments are to be used. Some physicians might want to trade off ease of use against accuracy of an instrument, while others might not. Therefore, an overview with regard to the evidence available of these measures is provided and the choice of instruments is up to the reader/user, for they will be the best judges given the intended use.

To the best of our knowledge, this is the first systematic review based on the COSMIN method for PROM-based...
instruments assessing disease activity in patients with RA. It adds to earlier reviews of physician-based/professional-based instruments for disease activity assessment.\(^2\)\(^6\) We limited the review to studies published in the past 20 years, because these instruments are the most likely to be relevant to current clinical trials and daily practice. The aim of the search strategy was to focus on sensitivity. The search used a validated PubMed filter in conjunction with many free-text terms and identified all 11 test-set articles plus an additional 23 other relevant validation studies. This strengthens our belief that the search strategy was indeed sensitive; however, it is still possible that some validation studies were not found due to the high heterogeneity of terms used in the literature. To ensure the uptake of evidence concerning clinimetric evaluation of PROMs, we recommend that authors pay close attention to choosing appropriate keywords in the title, abstract and keyword section, and that they make use of consistent terminology suggested by COSMIN.\(^16\) As part of the EULAR OML strategy, authors of the identified instruments will be contacted and encouraged to provide any evidence that might have been missed by the search strategy to further enhance sensitivity. The OML will be periodically updated with new evidence by refining and rerunning the search strategies (see online supplementary appendices I and II).

Clinical implications of this systematic review can be deduced from table 2. It is clear that until now the most effort has gone into the measurement properties concerning validity aspects (hypothesis testing, criterion validity, responsiveness). Other clinimetric domains such as reliability and interpretability are in need of more evidence. If, for instance, the measurement error or minimal important change is not well known, this impedes the use of a measure. The clinical implication of this is that without these measurement properties physicians cannot judge if differences in scores are due to chance and if they are truly important to their patients. In addition to this, evaluations of cross-cultural validity and direct comparison studies are needed in order to facilitate comprehension of instrument scores across different studies and different countries. Without evidence of formal validations of instruments in the language of their choice, physicians should be cautious of using instruments, comparing scores or generalising results of clinical studies using instruments in languages other than their patients’.

In conclusion, this systematic review of PROM-based instruments identified 17 measures aimed at monitoring disease activity in RA. The quality and extent of clinimetric validation varied among reports. The measurement properties least reported and in need of more evidence were: reliability, measurement error, cross-cultural validity and interpretability of measures. In general, the Pt-DAS28, RADAI, RADAI-5 and RAPID 3 had the strongest and most extensive validation. We hope this systematic review will aid professionals in the choice of PROM-based tools for disease activity assessment in RA. It is a first step in enhancing standardisation and clinimetric evaluation of these measures for disease activity in RA, and ultimately for supporting their use in clinical trials and daily practice.

**Author affiliations**
1. Department of IQ Healthcare, Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands
2. Department of Rheumatology, Radboud University Medical Center, Nijmegen, The Netherlands
3. Department for Health Evidence, Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands

**Twitter** Follow Jos Hendriks at @josshendriks

**Contributors** JH, JF, WK and PLvR drafted the initial research idea and methods. JH developed the search strategy. JH and WK reviewed all abstracts for inclusion. JH and MJdJ reviewed all full-text articles and extracted the results for publication. All authors declare having read and made a substantial contribution to the final manuscript.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

**REFERENCES**
5. Maska L, Anderson J, Michaud K. Measures of functional status and quality of life in rheumatoid arthritis: Health Assessment Questionnaire Disability Index (HAQ), Modified Health Assessment Questionnaire (MHAQ), Multidimensional Health Assessment Questionnaire (MDHAQ), Health Assessment Questionnaire II (HAQ-II), Improved Health Assessment Questionnaire (Improved HAQ), and Rheumatoid Arthritis Quality of Life (R-AQLQ). *Arthritis Care Res (Hoboken)* 2011;63(Suppl 11):S54–13.
6. Anderson JK, Zimmerman L, Caplan L, et al. Measures of rheumatoid arthritis disease activity: Patient (PtGA) and Provider (PrGA) Global Assessment of Disease Activity, Disease Activity Score (DAS) and Disease Activity Score with 28-Joint Counts (DAS28), Simplified Disease Activity Index (SDAI), Clinical Disease Activity Index (CDAI), Patient Activity Score (PAS) and Patient Activity Score-II (PASII). Routine Assessment of Patient Index Data (RAPID), Rheumatoid Arthritis Disease Activity Index (RADAI) and Rheumatoid Arthritis Disease Activity Index-5 (RADI-5), Chronic Arthritis Systemic Index (CASII), Patient-Based Disease Activity Score With ESR (PDAS1) and Patient-Based Disease Activity Score without ESR (PDAS2), and Mean Overall Index for Rheumatoid Arthritis (MOI-RA). *Arthritis Care Res (Hoboken)* 2011;63(Suppl 11):S14–36.
RMD Open

8


McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short-Form 36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis Care Res (Hoboken) 2011;63(Suppl 11):S240–52.

8. Hewlett S, Dures E, Almeida C. Measures of fatigue: Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire (BRAF MDQ). Bristol Rheumatoid Arthritis Fatigue Numerical Rating Scales (BRAF NRS) for severity, effort, and coping, Chalder Fatigue Questionnaire (CFQ), Checklist Individual Strength (CIS20R and CIS30R), Fatigue Severity Scale (FSS), Functional Assessment Chronic Illness Therapy (Fatigue) (FACT-F), Multi-Dimensional Assessment of Fatigue (MAF), Multi-Dimensional Fatigue Inventory (MFI), Pediatric Quality Of Life (PedsQL) Multi-Dimensional Fatigue Scale, Profile of Fatigue (ProF), Short Form 36 Vitality Subscale (SF-36 VT), and Visual Analog Scales (VAS). Arthritis Care Res (Hoboken) 2011;63(Suppl 11):S263–286.


23. Pincus T, Bergman MJ, Yaqzzi Y, et al. An index of only patient-reported outcome measures, routine assessment of patient index data 3 (RAPIDS), in two abatacept clinical trials: similar results to disease activity score (DAS28) and Short RAPIDS scores that include physician-observed disease (RAPIDS) 2008;47:345–9.


47. Pincus T, Furer V, Keystone E, et al. RAPID3 (Routine Assessment of Patient Index Data 3) severity categories and response criteria: similar results to DAS28 (Disease Activity Score) and CDAI (Clinical Disease Activity Index) in the RAPID 1 (Rheumatoid Arthritis Prevention of Structural Damage) clinical trial of certolizumab pegol. Arthritis Care Res (Hoboken) 2011;63:1142–9.


52. Pincus T, Swearingen CJ, Bergman MJ, et al. RAPID3 (Routine Assessment of Patient Index Data) on an MDHAQ (Multidimensional Health Assessment Questionnaire): agreement with DAS28 (Disease Activity Score) and CDAI (Clinical Disease Activity Index) activity categories, scored in five versus more than ninety seconds. Arthritis Care Res (Hoboken) 2010;62:181–9.


