RMD Open
Rheumatic & Musculoskeletal Diseases

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 Riel[1]

To cite: Hendrikx J, de Jonge MJ, Fransen J, *et al*. Systematic review of patient-reported outcome measures (PROMs) for assessing disease activity in rheumatoid arthritis. *RMD Open* 2016;**2**:e000202. doi:10.1136/rmdopen-2015-000202

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/rmdopen-2015-000202).

Received 25 October 2015
Revised 25 January 2016
Accepted 13 February 2016

For numbered affiliations see end of article.

Correspondence to
Dr Jos Hendrikx; Jos.Hendrikx@radboudumc.nl

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. 2 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 (PDAS) 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.

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.

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

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.[4–8] 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.[2 5 6] 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.[9–13] 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.[14–18] 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.[19] 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.[9 10 12 13 20–26] 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 criterion:

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.[18] 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).[27–30] For the interpretation of statistical measures being reported in studies, several suggestions have been stated. According to Nunnally and Bernstein,[31] 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

Table 1 Characteristics of the study population of included studies

Study	Size	Age	Female (%)	RF+ (%)	Disease duration	DAS28 at baseline
Blanchais *et al*[38]	26	56.6 (9.5)	NR	NR	16.6 (10) y	NR
Bossert *et al*[39]	200	57 (11.5)	75.5	78	13 (8.3) y	3.61 (1.43)
Castréjon *et al*[40]	39	56.8 (13.9)	90	NR	3.5 (NR) y	NR
Castréjon *et al*[41]	720	48.2 (12.4)	76.2	NR	4.8 (2.9–7.0)* m	5.1 (1.3)
Choy *et al*[9]	322	60.3 (23–87)†	76	81	9.13 (0–48)† y	NR
Fransen *et al*[42]	92	52 (13)	83	87	9 (4–14)* y	NR
Fransen *et al*[20]	584	59 (12)	72	69	8 (3–15)* y	4.3 (1.4)
Fujiwara *et al*[43]	250	59.3 (14)	78.4	NR	10.35 (9.83) y	NR
Harrington[44]	185	63 (22–88)†	NR	NR	18 (2–51)‡ y	NR
Heegaard *et al*[45]	30	60 (15)	77	70	15 (6)y	3.5 (1.0)§
Houssien *et al*[10]	100	57.7 (12.2)	78	NR	11.5 (8.3) y	4.24 (1.3)
Janta *et al*[11]	69	60.12 (13.16)	76.8	NR	11.9 (8.6) y	2.55 (1.08)
Kavanaugh *et al*[12]	218	54.3 (21–88)†	81.7	NR	8.5 (NR) y	5.4 (1.3)
	229	54.7 (19–82)†	76.9		7.2 (NR) y	5.0 (1.3)§
Leeb *et al*[21]	169	57 (19–78)†	79.8	50	7.2 (0.2–46)† y	3.51 (0.28–6.67)†
Leeb *et al*[22]	108	59.5 (24–87)‡	77.7	54	NR	2.95 (0.43–6.24)‡
Pincus[46]	63	58.5 (19.7)*	ref	ref	3.5 (8.8)* y	ref
	30	54.6 (20.9)*			2.9 (8.9)* y	
Pincus *et al*[23]	557	ref	ref	ref	ref	6.82 (NR)/6.83(NR)
	278					6.89 (NR)/6.88(NR)
Pincus *et al*[47]	1384	ref	ref	ref	ref	ref
Pincus *et al*[48]	982	ref	ref	ref	ref	ref
Pincus *et al*[49]	557	ref	ref	ref	ref	ref
	227					
Pincus *et al*[50]	Ref	ref	ref	ref	ref	ref
Pincus *et al*[51]	285	57.4 (14.6)	73	NR	9.7 (9.0) y	3.4 (1.7)
Pincus *et al*[52]	200	53.4 (16.2)	81	NR	11.6 (10.8) y	3.7 (1.5)
Riazzoli *et al*[13]	47	50 (13)	79	86	9.4 (8.6) y	5.4 (1.2)
Rintelen *et al*[24]	392	61 (20–87)‡	82.1	59.4	62 (3.545)‡ m	3.26 (0.49–8.09)‡
Rintelen *et al*[53]	705	62.7 (13.4)	75.9	54.4	97.3 (98.0) m	3.31 (1.37)
Salaffi *et al*[54]	191	56.6 (12.2)	82.7	NR	5.1 (5.5) y	6.02 (1.15)
Salaffi *et al*[55]	196	56.7 (12.1)	83.1	78	5.1 (5.9) y	3.94 (2.03)
	247	58.1 (11.2)	80.1	76	6.2 (6.6) y	
Singh *et al*[56]	200	42.2 (NR)	83	NR	4.9 (NR) y	5.2 (1.6)
Stucki *et al*[57]	55	60.0 (14.6)	62	NR	5.1 (1.3–10.7)* y	NR
Sullivan *et al*[58]	740	57 (13.7)	83	63.8	14.3 (12.3) y	4.05 (1.5)§
Uhlig *et al*[59]	28	61.1 (6.2)	64	64	16.6 (10.4) y	3.12 (1.27)
Veehof *et al*[25]	191	54.5 (13.3)	71	NR	7.0 (3–17)* y	5.42 (1.07)
Wolfe *et al*[26]	9078	62.2 (12.6)	78.2	NR	16.2 (10.9) y	NR

*Median(IQR): Values are mean (standard deviation) unless otherwise indicated.
†Mean (range).
‡Median (range).
§DAS28-CRP.
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.

proposed the following categorisation for correlational measures: 0.1–0.29 no or negligible correlation, 0.30–0.49 low correlation, 0.50–0.69 moderate correlation, 0.70–0.89 high correlation and 0.9–1.0 very strong correlation.[32] 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.[33–36] 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.[37] 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 2 Overall levels of evidence of measurement properties per instrument across all included studies

Instrument	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Cross-cultural validity	Hypothesis testing	Criterion validity	Responsiveness	Interpretability
RADAI[10 20 21 25 42 54 55 57–59]	++ 0.84–0.91 α	? 0.92 ICC	? 14.9%$_{max}$SDD	?	++ 1 factor	?	+++	−− 0.48–0.83s	++ 0.77 AUC$_{responder}$	?
RADAI-5[21 22 24 39 53]	++ 0.91–0.92 α	?	?	?	++ 1 factor	?	++	++ 0.62–0.66s	− 0.589s/0.295κ	?
PDAS I[9]	− 0.5 α	+ 0.76–0.88	?	?	?	?	+	+ 0.89s	?	?
PDAS II[9]	− 0.4 α	+ 0.76–0.88	?	?	?	?	+	+ 0.76s	?	?
Pt-DAS28[10–13 45]	?	+ 0.92p	+ 23.2%$_{mean}$SDD	?	?	?	++	++ 0.73–0.94p	±	?
Pt-SDAI[11 45]	?	+ 0.90p	+ 59.9%$_{mean}$SDD	?	?	?	+	++ 0.87s 0.93p	?	?
PAS I[26 54]	?	?	?	?	?	?	+	?	+ 0.79 AUC$_{responder}$	?
PAS II[26]	?	?	?	?	?	?	+	?	?	?
RAPID 2[23]	?	?	?	?	?	?	?	?	?	?
RAPID 3[22 38–41 43 46 48–52 54–56 59]	++ 0.87 α	? 0.90	? 14.8%$_{max}$SDD	?	++ 1 factor	?	++	++ 0.64–0.91s	± 0.80 AUC$_{responder}$ 0.43–0.53 wκ	?
RAPID 4 PtJC[23 38 51]	?	?	?	?	?	?	?	++ 0.65s	?	?
RAPID 4 MDJC[23]	?	?	?	?	?	?	?	?	?	?
RAPID 5[23 51 58]	?	?	?	?	?	?	?	++ 0.69–0.71s	+ 0.70s	?
PRO-CLARA[54 55]	+ 0.89 α	?	?	?	+ 1 factor	?	+	+ 0.84s	+ 0.82 AUC$_{responder}$	?
PRO-Index C/M[47]	?	?	?	?	?	?	?	?	+	?
GAS[44 58]	?	?	?	?	?	?	?	−− 0.54s	− 0.50s	?

+, positive result; −, negative result (for explanation of the categorisation of the levels, see table 3); %$_{max}$SDD, SDD as percentage of maximum value of outcome; %$_{mean}$SDD, SDD as percentage of mean value of the outcome; α, Cronbach's α; τ, Kendall's τ; AUC$_{flare}$, area under the curve for patients with flare versus no flare; AUC$_{responder}$, 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; PtJC, 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.

Table 3 Levels of evidence for the overall quality of measurement properties per instrument across all included studies[20–23]

Level	Rating	Criteria
Strong	+++ or −−−	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
Moderate	++ or −−	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+ or –	One study of fair methodological quality
Conflicting	±	Conflicting findings
Unknown	?	No studies or only studies of poor methodological quality

+, positive result; –, negative result.

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).[20–23] 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-PtJC), Routine Assessment of Patient Index Data 4 Medical Doctor Joint Count (RAPID 4-MDJC), RAPID 5, Patient

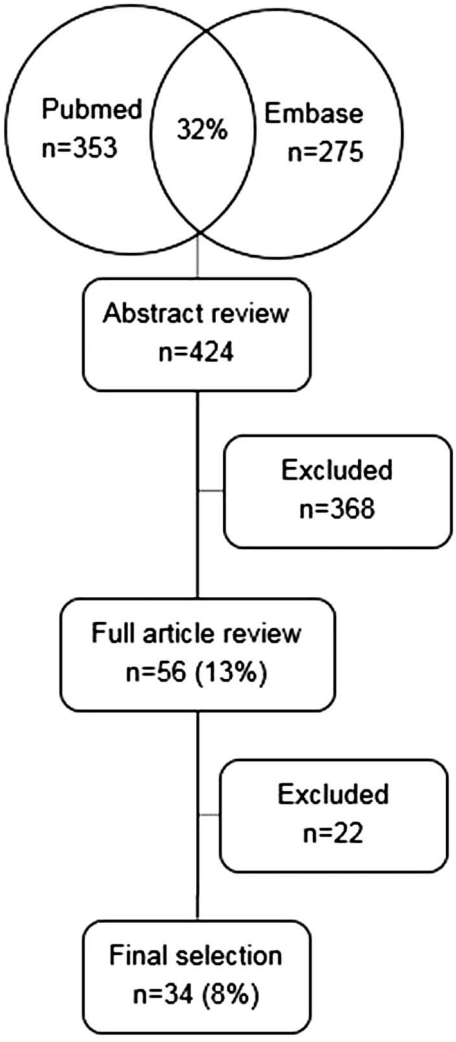

Figure 1 Search results PubMed/EMBASE, overlap, exclusion based on abstract review, exclusion based on full review.

Reported Outcome-index (PRO-index) majority (M), continuous (C), Patient Reported Outcome CLinical ARthritis Activity (PRO-CLARA).[9–13 20–26 38–59] 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.[14–19 27 29 60] 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.[61]

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.[16 19] 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.[62] 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.[63] 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*.[64] 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.[65] 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.[23 47–50] 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 necessarily 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 Hendrikx at @joshendrikx

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.

REFERENCES

1. Aletaha D, Landewe R, Karonitsch T, *et al.* Reporting disease activity in clinical trials of patients with rheumatoid arthritis: EULAR/ACR collaborative recommendations. *Arthritis Rheum* 2008;59:1371–7.
2. Anderson J, Caplan L, Yazdany J, *et al.* Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. *Arthritis Care Res (Hoboken)* 2012;64:640–7.
3. Stoffer MA, Schoels MM, Smolen JS, *et al.* Evidence for treating rheumatoid arthritis to target: results of a systematic literature search update. *Ann Rheum Dis* 2016;75:16–22.
4. Katz PP. Introduction to special issue: patient outcomes in rheumatology, 2011. *Arthritis Care Res (Hoboken)* 2011;63(Suppl 11):S1–3.
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 (RAQoL). *Arthritis Care Res (Hoboken)* 2011;63(Suppl 11):S4–13.
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 (RADAI-5), Chronic Arthritis Systemic Index (CASI), 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.
7. Hawker GA, Mian S, Kendzerska T, *et al.* Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form

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, effect, and coping, Chalder Fatigue Questionnaire (CFQ), Checklist Individual Strength (CIS20R and CIS8R), Fatigue Severity Scale (FSS), Functional Assessment Chronic Illness Therapy (Fatigue) (FACIT-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.
9. Choy EH, Khoshaba B, Cooper D, *et al*. Development and validation of a patient-based disease activity score in rheumatoid arthritis that can be used in clinical trials and routine practice. *Arthritis Rheum* 2008;59:192–9.
10. Houssien DA, Stucki G, Scott DL. A patient-derived disease activity score can substitute for a physician-derived disease activity score in clinical research. *Rheumatology (Oxford)* 1999;38:48–52.
11. Janta I, Naredo E, Martinez-Estupinan L, *et al*. Patient self-assessment and physician's assessment of rheumatoid arthritis activity: which is more realistic in remission status? A comparison with ultrasonography. *Rheumatology (Oxford)* 2013;52:2243–50.
12. Kavanaugh A, Lee SJ, Weng HH, *et al*. Patient-derived joint counts are a potential alternative for determining Disease Activity Score. *J Rheumatol* 2010;37:1035–41.
13. Riazzoli J, Nilsson JÅ, Teleman A, *et al*. Patient-reported 28 swollen and tender joint counts accurately represent RA disease activity and can be used to assess therapy responses at the group level. *Rheumatology (Oxford)* 2010;49:2098–103.
14. Mokkink LB, Terwee CB, Gibbons E, *et al*. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Med Res Methodol* 2010;10:82.
15. Mokkink LB, Terwee CB, Knol DL, *et al*. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010;10:22.
16. Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
17. Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010;19:539–49.
18. Terwee CB, Mokkink LB, Knol DL, *et al*. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;21:651–7.
19. Terwee CB, Jansma EP, Riphagen II, *et al*. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;18:1115–23.
20. Fransen J, Langenegger T, Michel BA, *et al*. Feasibility and validity of the RADAI, a self-administered rheumatoid arthritis disease activity index. *Rheumatology (Oxford)* 2000;39:321–7.
21. Leeb BF, Haindl PM, Maktari A, *et al*. Patient-centered rheumatoid arthritis disease activity assessment by a modified RADAI. *J Rheumatol* 2008;35:1294–9.
22. Leeb BF, Sautner J, Mai HT, *et al*. A comparison of patient questionnaires and composite indexes in routine care of rheumatoid arthritis patients. *Joint Bone Spine* 2009;76:658–64.
23. Pincus T, Bergman MJ, Yazici Y, *et al*. An index of only patient-reported outcome measures, routine assessment of patient index data 3 (RAPID3), in two abatacept clinical trials: similar results to disease activity score (DAS28) and other RAPID indices that include physician-reported measures. *Rheumatology (Oxford)* 2008;47:345–9.
24. Rintelen B, Haindl PM, Sautner J, *et al*. The rheumatoid arthritis disease activity index-5 in daily use. Proposal for disease activity categories. *J Rheumatol* 2009;36:918–24.
25. Veehof MM, ten Klooster PM, Taal E, *et al*. Psychometric properties of the Rheumatoid Arthritis Disease Activity Index (RADAI) in a cohort of consecutive Dutch patients with RA starting anti-tumour necrosis factor treatment. *Ann Rheum Dis* 2008;67:789–93.
26. Wolfe F, Michaud K, Pincus T. A composite disease activity scale for clinical practice, observational studies, and clinical trials: the patient activity scale (PAS/PAS-II). *J Rheumatol* 2005;32:2410–15.
27. van Tulder M, Furlan A, Bombardier C, *et al*. Editorial Board of the Cochrane Collaboration Back Review G. Updated method guidelines for systematic reviews in the Cochrane collaboration back review group. *Spine (Phila Pa 1976)* 2003;28:1290–9.
28. Schellingerhout JM, Heymans MW, Verhagen AP, *et al*. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. *BMC Med Res Methodol* 2011;11:87.
29. Furlan AD, Pennick V, Bombardier C, *et al*. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. *Spine (Phila Pa 1976)* 2009;34:1929–41.
30. Schellingerhout JM, Verhagen AP, Heymans MW, *et al*. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. *Qual Life Res* 2012;21:659–70.
31. Nunnally J, Bernstein I. *Psychometric theory*. 3rd edn. New York: MCGraw-Hill, 1994.
32. Hinkle DE, Wiersma W, Jurs SG. *Applied statistics for the behavioral sciences*. Boston, MA: Houghton Mifflin, 2003.
33. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 4th edn. Oxford, UK: Oxford University Press, 2008.
34. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic* 1981;86:127–37.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
36. Fleiss JL. *Statistical methods for rates and proportions*. 2nd edn. New York: John Wiley & Sons, 1981.
37. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285–93.
38. Blanchais A, Berthelot JM, Fontenoy AM, *et al*. Weekly home self-assessment of RAPID-4/3 scores in rheumatoid arthritis: a 6-month study in 26 patients. *Joint Bone Spine* 2010;77:582–7.
39. Bossert M, Prati C, Vidal C, *et al*. Evaluation of self-report questionnaires for assessing rheumatoid arthritis activity: a cross-sectional study of RAPID3 and RADAI5 and flare detection in 200 patients. *Joint Bone Spine* 2012;79:57–62.
40. Castrejón I, Bergman MJ, Pincus T. MDHAQ/RAPID3 to recognize improvement over 2 months in usual care of patients with osteoarthritis, systemic lupus erythematosus, spondyloarthropathy, and gout, as well as rheumatoid arthritis. *J Clin Rheumatol* 2013;19:169–74.
41. Castrejón I, Dougados M, Combe B, *et al*. Can remission in rheumatoid arthritis be assessed without laboratory tests or a formal joint count? Possible remission criteria based on a self-report RAPID3 score and careful joint examination in the ESPOIR cohort. *J Rheumatol* 2013;40:386–93.
42. Fransen J, Häuselmann H, Michel BA, *et al*. Responsiveness of the self-assessed rheumatoid arthritis disease activity index to a flare of disease activity. *Arthritis Rheum* 2001;44:53–60.
43. Fujiwara M, Kita Y. Reexamination of the assessment criteria for rheumatoid arthritis disease activity based on comparison of the Disease Activity Score 28 with other simpler assessment methods. *Mod Rheumatol* 2013;23:260–8.
44. Harrington JT. The uses of disease activity scoring and the physician global assessment of disease activity for managing rheumatoid arthritis in rheumatology practice. *J Rheumatol* 2009;36:925–9.
45. Heegaard C, Dreyer L, Egsmose C, *et al*. Test-retest reliability of the disease activity score 28 CRP (DAS28-CRP), the simplified disease activity index (SDAI) and the clinical disease activity index (CDAI) in rheumatoid arthritis when based on patient self-assessment of tender and swollen joints. *Clin Rheumatol* 2013;32:1493–500.
46. Pincus T. RAPID3, an index of only 3 patient self-report core data set measures, but not ESR, recognizes incomplete responses to methotrexate in usual care of patients with rheumatoid arthritis. *Bull Hosp Jt Dis (2013)* 2013;71:117–20.
47. Pincus T, Chung C, Segurado OG, *et al*. An index of patient reported outcomes (PRO-Index) discriminates effectively between active and control treatment in 4 clinical trials of adalimumab in rheumatoid arthritis. *J Rheumatol* 2006;33:2146–52.
48. 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.

49. Pincus T, Hines P, Bergman MJ, *et al*. Proposed severity and response criteria for Routine Assessment of Patient Index Data (RAPID3): results for categories of disease activity and response criteria in abatacept clinical trials. *J Rheumatol* 2011;38:2565–71.
50. Pincus T, Strand V, Koch G, *et al*. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the Disease Activity Score (DAS) in a rheumatoid arthritis clinical trial. *Arthritis Rheum* 2003;48:625–30.
51. Pincus T, Swearingen CJ, Bergman M, *et al*. RAPID3 (Routine Assessment of Patient Index Data 3), a rheumatoid arthritis index without formal joint counts for routine care: proposed severity categories compared to disease activity score and clinical disease activity index categories. *J Rheumatol* 2008;35:2136–47.
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.
53. Rintelen B, Sautner J, Haindl P, *et al*. Remission in rheumatoid arthritis: a comparison of the 2 newly proposed ACR/EULAR remission criteria with the rheumatoid arthritis disease activity index-5, a patient self-report disease activity index. *J Rheumatol* 2013;40:394–400.
54. Salaffi F, Ciapetti A, Gasparini S, *et al*. The comparative responsiveness of the patient self-report questionnaires and composite disease indices for assessing rheumatoid arthritis activity in routine care. *Clin Exp Rheumatol* 2012;30:912–21.
55. Salaffi F, Migliore A, Scarpellini M, *et al*. Psychometric properties of an index of three patient reported outcome (PRO) measures, termed the CLinical ARthritis Activity (PRO-CLARA) in patients with rheumatoid arthritis. The NEW INDICES study. *Clin Exp Rheumatol* 2010;28:186–200.
56. Singh H, Gupta V, Ray S, *et al*. Evaluation of disease activity in rheumatoid arthritis by Routine Assessment of Patient Index Data 3 (RAPID3) and its correlation to Disease Activity Score 28 (DAS28) and Clinical Disease Activity Index (CDAI): an Indian experience. *Clin Rheumatol* 2012;31:1663–9.
57. Stucki G, Liang MH, Stucki S, *et al*. A self-administered rheumatoid arthritis disease activity index (RADAI) for epidemiologic research. Psychometric properties and correlation with parameters of disease activity. *Arthritis Rheum* 1995;38:795–8.
58. Sullivan MB, Iannaccone C, Cui J, *et al*. Evaluation of selected rheumatoid arthritis activity scores for office-based assessment. *J Rheumatol* 2010;37:2466–8.
59. Uhlig T, Kvien TK, Pincus T. Test-retest reliability of disease activity core set measures and indices in rheumatoid arthritis. *Ann Rheum Dis* 2009;68:972–5.
60. Mokkink LB, Terwee CB, Stratford PW, *et al*. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Qual Life Res* 2009;18:313–33.
61. Castrejón I, Gossec L, Carmona L. The EULAR Outcome Measures Library: an evolutional database of validated patient-reported instruments. *Ann Rheum Dis* 2015;74:475–6.
62. Gossec L, Dougados M, Rincheval N, *et al*. Elaboration of the preliminary Rheumatoid Arthritis Impact of Disease (RAID) score: a EULAR initiative. *Ann Rheum Dis* 2009;68:1680–5.
63. Gossec L, Paternotte S, Aanerud GJ, *et al*. Finalisation and validation of the rheumatoid arthritis impact of disease score, a patient-derived composite measure of impact of rheumatoid arthritis: a EULAR initiative. *Ann Rheum Dis* 2011;70:935–42.
64. Lydersen S. Statistical review: frequently given comments. *Ann Rheum Dis* 2015;74:323–5.
65. Mokkink LB, Terwee CB, Patrick DL, *et al*. Cosmin Manual. [pdf] 2012 [cited 2015 10-07-2015]; 2012:[COnsensus-based Standards for the selection of health Measurement INstruments checklist manual]. http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20manual%20v9.pdf