The Journal of Rheumatology 2021;48:1784–92
doi:10.3899/jrheum.201439
First Release September 1 2021

The Journal of Rheumatology

Systematic Review and Metaanalysis of the Reproducibility of Patient Self-reported Joint Counts in Rheumatoid Arthritis

Sanketh Rampes[1], Vishit Patel[1], Ailsa Bosworth[2], Clare Jacklin[2], Deepak Nagra[3], Mark Yates[3], Sam Norton[3], and James B. Galloway[3]

ABSTRACT. ***Objective***. To assess the reproducibility of patient-reported tender (TJCs) and swollen joint counts (SJCs) of patients with rheumatoid arthritis (RA) compared to trained clinicians.

Methods. We conducted a systematic literature review and metaanalysis of studies comparing patient-reported TJCs and/or SJCs to clinician counts in patients with RA. We calculated pooled summary estimates for correlation. Agreement was compared using a Bland-Altman approach.

Results. Fourteen studies were included in the metaanalysis. There were strong correlations between clinician and patient TJCs (0.78, 95% CI 0.76–0.80), and clinician and patient SJCs (0.59, 95% CI 0.54–0.63). TJCs had good reliability, ranging from 0.51 to 0.85. SJCs had moderate reliability, ranging from 0.28 to 0.77. Agreement for TJCs reduced for higher TJC values, suggesting a positive bias for self-reported TJCs, which was not observed for SJCs.

Conclusion. Our metaanalysis has identified a strong correlation between patient- and clinician-reported TJCs, and a moderate correlation for SJCs. Patient-reported joint counts may be suitable for use in annual review for patients in remission and in monitoring treatment response for patients with RA. However, they are likely not appropriate for decisions on commencement of biologics. Further research is needed to identify patient groups in which patient-reported joint counts are unsuitable.

Key Indexing Terms: joints, patient-reported outcome measures, rheumatoid arthritis, self-report

Rheumatoid arthritis (RA) is characterized by synovial joint inflammation leading to loss of function and disability if untreated. The systematic evaluation of joints by healthcare professionals (HCPs) is crucial in the monitoring and treatment of RA as part of the "treat-to-target" approach recommended in North American and European guidelines.[1,2,3] Regular monitoring of disease activity allows up-titration of medication to achieve and maintain remission.[1] In addition to clinical practice, joints counts are used across clinical research and are included in the Outcome Measures in Rheumatology (OMERACT) RA core dataset.[4]

Physician-measured joint counts have been shown to be predictive of mortality[5] and are included in disease activity indices such as the Disease Activity Score (DAS), Simplified Disease Activity Index (SDAI), and the Clinical Disease Activity Index (CDAI). The DAS in 28 joints (DAS-28) is a composite measure of disease activity that takes into account weighted values of the 28-joint count of swollen and tender joints, patient global assessment, and C-reactive protein (CRP). In the UK, the DAS-28 is the primary tool used for the assessment of RA and is central to National Institute for Health and Care Excellence technology appraisals for RA therapies.[3]

Swollen (SJC) and tender joint counts (TJC) are performed mostly by HCPs. Several studies have evaluated the reproducibility of patient self-reported joint counts in the assessment of disease activity, as these have the potential to increase patient engagement and encourage self-management behavior. Improved self-management has been associated with beneficial outcomes across health status, pain, and fatigue.[6] A 2015 systematic review reported high intra- and interobserver reliability for patient-reported TJCs but a lower intra- and interobserver reliability for patient-reported SJCs.[7] A metaanalysis performed in 2009 reported a summary estimate Pearson correlation coefficient for TJC of 0.61 (95% CI 0.47–0.75) and for SJC of 0.44 (95% CI 0.15–0.73).[8] The key measure of reproducibility for patient-reported joint counts is agreement. Reproducibility is an umbrella term for the concepts of agreement and reliability.[9] Specifically, agreement is concerned with measurement error, whereas reliability relates to the ability of the assessment to discriminate between people or objects and concerns the ratio of the variability between participants or objects relative to the

MY is funded by Versus Arthritis.

[1]S. Rampes, MA, V. Patel, MSc, Faculty of Life Sciences & Medicine, King's College London, London; [2]A. Bosworth, C. Jacklin, National Rheumatoid Arthritis Society, Berkshire; [3]D. Nagra, MD, M. Yates, PhD, S. Norton, PhD, J.B. Galloway, PhD, Centre for Rheumatic Diseases, King's College London, London, UK.

S. Rampes and V. Patel contributed equally to this work.

The authors declare no conflicts of interest relevant to this article.

Address correspondence to Dr. J.B. Galloway, Centre for Rheumatic Diseases, Room 3.46, Third Floor, Weston Education Centre, King's College London, London SE5 9RJ, UK. Email: james.galloway@kcl.ac.uk.

Accepted for publication May 5, 2021.

total variability, including measurement error. While reliability is important, agreement measures are typically preferred for cases in which the instrument is used for evaluation purposes, such as TJCs and SJCs.[9] Previous reviews have neglected agreement, instead focusing on reliability or correlational approaches that ignore systematic intra- and interindividual differences. This review builds on previous work by considering agreement using a Bland-Altman–type approach to assess reproducibility of patient-reported joint counts in clinical practice.

The coronavirus disease 2019 pandemic has rapidly altered the rheumatological landscape, with a transition to video and telephone clinics. It is likely that reduced face-to-face clinical interaction will be sustained. Now more than ever there is a need to understand the reproducibility of self-reported joint counts. We aimed to conduct a systematic review and metaanalysis of the published evidence around the reproducibility of self-reported TJCs and SJCs for use in calculating disease activity indices including the DAS-28, SDAI, and CDAI.

METHODS

Literature search. This study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and registered with the international prospective register of systematic reviews (PROSPERO 2020 CRD42020189116). A systematic search of the EMBASE and MEDLINE databases was performed for studies published between January 1, 1990, and June 1, 2020. After reviewing 2 previously published systematic searches,[7,8] the following search was conducted: "rheumatoid arthritis OR rheumatoid" AND "joint OR joints OR disease activity" AND "patient-report OR patient-assess OR self-report OR self-assess OR self-monitor OR self-monitor* OR self-administ* OR self-evalua* OR self-examin* OR self-rate OR self-rating". Reference lists of previous systematic reviews were assessed for additional eligible studies.

The search was limited to studies with human participants, and the following English-language publication types: article, article in press, clinical trial, comparative study, or observational study.

Eligibility criteria. Eligible studies included a patient and a trained assessor of TJCs and/or SJCs, and a direct comparison between patient and trained assessor joint counts in patients with RA. Review articles, letters to the editor, and conference abstracts were excluded.

Study selection. Titles and abstracts of studies retrieved using the search strategy detailed above, as well as those identified from reference lists of selected publications, were reviewed independently by 2 investigators (VP, SR), and any disagreement was adjudicated by a third reviewer (MY). The data from the eligible studies were extracted into a table.

Data extraction. Two investigators (VP, SR) each extracted data from half the eligible studies. Data extracted included the following: authors, year, country in which the study was performed, study design, number of subjects, number of assessors, blinding, types of assessors compared, patient education level, study inclusion/exclusion criteria, age, sex, number of joints examined, mean SJCs and TJCs, agreement and correlation measures, agreement value, and intraobserver reliability.

Quality assessment. Risk of bias was assessed using the Quality Appraisal of Reliability Studies (QAREL) checklist, which comprises an 11-item checklist that covers 7 key principles in diagnostic reliability studies (Supplementary Table 1, available with the online version of this manuscript).[10] The 7 principles are as follows: spectrum of examiners, spectrum of subjects, examiner blinding, order effects of examination, suitability of the time interval among repeated measurements, appropriate test application and interpretation, and appropriate statistical analysis.[10]

Statistical methods. Correlation coefficients for TJCs and SJCs were tabulated, with a separate metaanalysis performed for each measure using a random effects model (Supplementary Data 1, available with the online version of this article). Fisher Z transformation of correlation coefficients was used for metaanalysis and results were displayed graphically using forest plots. Given that the Pearson and Spearman methods are on the same metric, these were combined in the same metaanalysis, with the Pearson method used where both were reported. Sensitivity analyses stratified by correlation method confirmed no difference in the estimates. Statistical heterogeneity was described using the I^2 statistic.

As detailed by de Vet, *et al*,[9] we distinguished the use of agreement and reliability parameters, which collectively can be referred to as "reproducibility parameters." Agreement parameters assess the closeness of repeated measurements by estimating measurement error. Reliability parameters assess whether study objects can be distinguished from each other despite measurement error and are related to variability between study objects.[9] Intraclass correlation coefficient (ICC) and Cohen κ are reliability parameters. Agreement parameters are expressed on the actual scale of measurement, whereas reliability parameters are expressed as a dimensionless value between 0 and 1.[9] It is also important to note that Spearman and Pearson correlation coefficients are neither measures of reliability nor agreements—they are measures of association.

Studies that reported reliability estimates, in terms of interclass correlations or Cohen κ, were not used in the metaanalysis but were described narratively in the results.[11,12,13,14] We performed subgroup analysis exploring whether self-reported joints obtained by text or mannequin format affected correlation or reliability. At the study level, agreement between patient and HCP joint counts were compared using study means following a Bland-Altman–type approach.[15] Analyses were performed using Stata 16 (StataCorp LLC).

RESULTS

Search results. The electronic database search identified 1530 articles following removal of duplicates. Title and abstract review removed 1486 articles, leaving 42 articles for full-text review. Reference searching identified 2 additional publications. A further 24 articles were excluded after full-text review, leaving 20 eligible publications. Further details can be found in the flow diagram (Figure 1).

Study and patient characteristics. Details on included studies can be found in Table 1.[11–14,16–31] The median sample size was 64 (range 30–447). The mean age of patients ranged from 49 to 65 years old, with 60–92% being female. Five studies reported race, with a range of 75–97% White.[13,14,20,22,29]

All 20 studies collected data on TJCs, with 15 also considering SJCs. A total of 14/20 (70%) studies utilized a 28-joint count. Of the remaining 6 studies, 1 study had a lower number of joints included, with only 20 joints assessed, and the remaining 5 studies used joint counts > 28 (range 36–60).

Joint counts were measured by physicians only in 14 studies, by trained nurses only in 1 study, by trained assessor/research assistant only in 2 studies, and combined nurse or research assistant and physician in 3 studies. Nine studies detailed how patients were trained to measure joint counts, which was typically a 5- to 10-minute session.

Five studies detailed patient education level or socioeconomic position.[14,19,22,23,28] Education levels ranged from 8 to 13 years spent in education, with a wide representation of educational backgrounds. Studies from Peru and Colombia had a higher

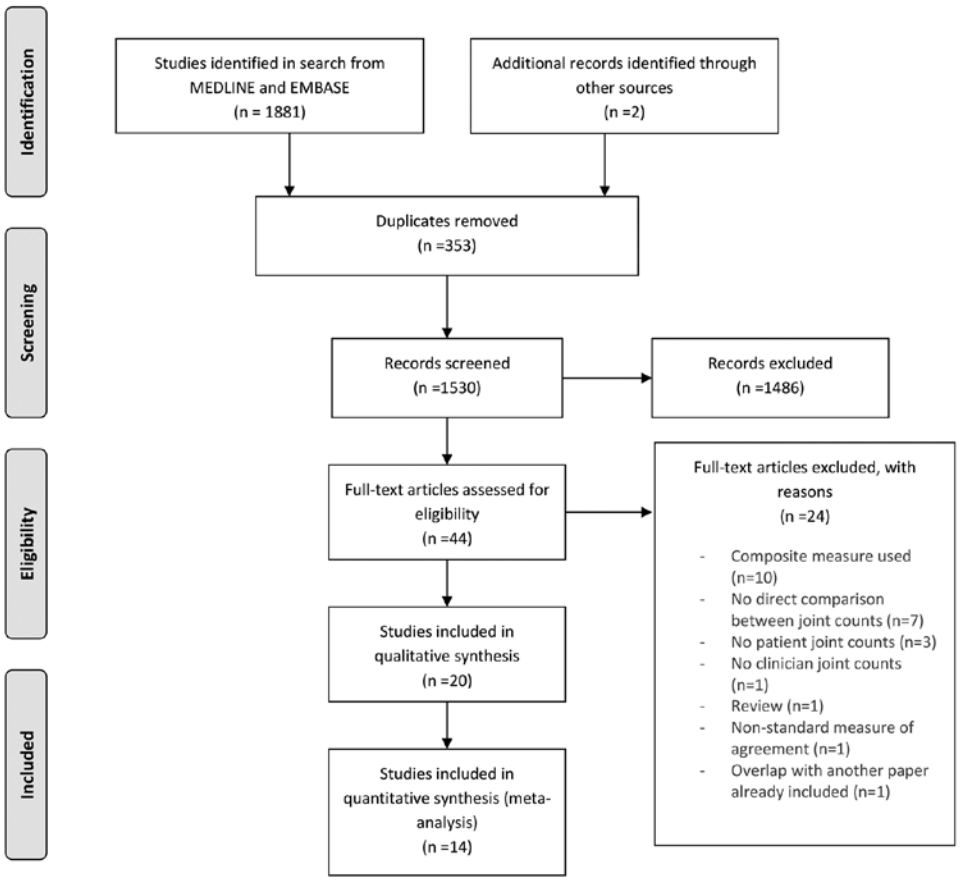


Figure 1. Flow chart of studies included in systematic review and metaanalysis.

proportion of patients with a low educational background. The association between joint counts was comparable to the overall analyses.

Summary correlation coefficients, reliability estimates, and agreement. For TJCs, 7 studies used Pearson and 6 studies used Spearman correlation coefficients. One study reported a correlation coefficient without specifying the type. For TJC reliability, 4 studies reported ICC, 1 reported Cohen κ, and 1 reported Kendall *W* (Table 2). For SJCs, 5 studies used Pearson and 5 studies used Spearman correlation coefficients. For SJC reliability, 3 studies reported ICC, 1 reported Cohen κ, and 1 reported Kendall *W* (Table 3).

The correlation coefficients between patient-reported and clinician joint counts are detailed in Table 2 and Table 3. There was a strong correlation between clinician and patient TJCs of 0.78 (95% CI 0.76–0.80; I^2 = 83.4%) and a strong correlation between clinician and patient SJCs of 0.59 (95% CI 0.54–0.63; I^2 = 72.4%). The summary estimates are presented in a forest plot in Figure 2.

TJC correlation coefficients ranged from moderate to strong (range 0.37–0.94), whereas SJC correlation coefficients ranged from weak to strong (range 0.16–0.93). There was higher correlation for TJCs/SJCs measured on a mannequin compared with text format. For TJCs, mannequin correlation was strong (range 0.60–0.94), whereas text correlation ranged from moderate to strong (range 0.37–0.89). For SJCs, mannequin correlation ranged from moderate to strong (range 0.43–0.93) whereas text correlation ranged from weak to strong (range 0.16–0.58).

Reliability coefficients (ICC, Cohen κ, Kendall *W*) for patient- and clinician-reported joint counts are detailed in Table 2 and Table 3. When reporting the strength of reliability coefficients, we reported as described by Landis and Koch in 1977.[32] For SJC reliability, values ranged from fair to substantial (range 0.28–0.77), whereas for TJC reliability, values ranged from moderate to near perfect (range 0.51–0.85). Higher reliability for TJC was found for mannequin compared with text format, whereas no studies measured SJC reliability using a text format. For TJCs, mannequin reliability ranged from moderate to near perfect (range 0.51–0.85), whereas text reliability was moderate (0.55).

Bland-Altman plots were used to visualize data from across all studies that provided mean TJCs and/or SJCs with limits of agreement calculated to provide an estimate of measurement error (Figure 3). These provide additional insight into the reliability measures from the studies that explicitly reported them. From the more inclusive analysis, it was apparent that agreement was better for SJCs, and agreement for TJCs was better for lower

Table 1. Characteristics of studies included.

Study	Country of Study	Sample Size	No. of Joints	No. of Assessors	Blinding	Reliability[a]	TJC	SJC
Heegaard, 2013[16*]	Denmark	31	28	1	Y	Both	Y	Y
El Miedany, 2010[17*]	UK	82	28	U	U	Both	Y	N
Levy, 2007[18*]	US	60	28	1	Y	Inter	Y	Y
Wong, 1999[19*]	Worldwide	60	· Mannequin: 50 TJCs, 48 SJCs · Text: 20 TJCs, 18 SJCs	U	Y	Inter	Y	Y
Hanly, 1996[20*]	Canada	61	20	4	Y	Inter	Y	Y
Prevoo (Site 1), 1996[21*]	Netherlands	141	28	U	U	Inter	Y	Y
Prevoo (Site 2), 1996[21*]	Netherlands	101	28	U	U	Inter	Y	Y
Abraham, 1993[22*]	US	32	50 clinician; 42 patient	1	Y	Inter	Y	N
Radner, 2012[11]	Austria	209	28	3	Y	Inter	Y	Y
Janta, 2013[12]	Spain	102	28	1	Y	Inter	Y	Y
Amaya-Amaya, 2012[23]	Colombia	135	28	U	Y	Inter	Y	Y
Cheung, 2010[24]	France	50	28	2	Y	Inter	Y	Y
Figueroa, 2007[25*]	US	82	58 clinician; 60 patient	3	U	Inter	Y	Y
Kavanaugh, 2010[13]	US	447	28	U	Y	Inter	Y	Y
Riazzoli, 2010[26*]	Sweden	47	28	1	U	Inter	Y	Y
El Miedany and Palmer, 2008[27]	UK	148	28	U	U	Inter	Y	N
Alarcón, 1999[14]	US	67	36	1	Y	Inter	Y	N
Calvo, 1999[28*]	Peru	60	36	1	U	Inter	Y	N
Greenwood, 2006[29*]	UK	45	28	1	U	Inter	Y	Y
Inderjeeth, 2019[30*]	Australia	52	28	U	Y	Inter	Y	Y
Houssien, 1999[31*]	UK	100	28	1	Y	Inter	Y	Y

[a] Reliability denotes whether studies compared patient joint counts to clinician joint counts (inter); in addition, some studies also compared intrapatient joint counts (both). * Included in metaanalysis. SJC: swollen joint count; TJC: tender joint count; N: No; U: Unknown; Y: Yes.

joint counts. On average, patients report 1.1 more painful joints than clinicians, but because joint counts are skewed, this bias was not constant. Specifically, bias was negligible for TJCs < 5, whereas when the clinician TJCs increased to > 5 joints, the bias caused by overestimation by patients increased. For SJCs, the difference was negligible and any bias appears to go in the opposite direction, with patients tending to report a lower value than clinicians.

Reliability of individual joints. Three studies assessed reliability for individual joints between patients and physicians[13,20,31] and 1 study assessed reliability for individual joints between patients and a trained assessor.[13] Reliability was measured by Cohen κ or Kendall *W*. Reliability varied substantially with a median (IQR) of 0.49 (0.39–0.63) for TJC and 0.26 (0.20–0.38) for SJC, with substantial variation between different joints. There were no obvious trends, but reliability appeared to be higher in larger joints (shoulders, knees, and elbows).

Intrapatient reproducibility of TJCs and SJCs. Two studies reported intrapatient correlation coefficients and 2 studies reported reliability coefficients for TJCs and SJCs.[16,19,24,25] For TJCs, correlation ranged from 0.87 to 0.96 and reliability ranged from 0.90 to 0.94. For SJCs, correlation ranged from 0.87 to 0.97 and reliability ranged from 0.56 to 0.89. Due to small numbers, summary estimates were not performed. The interval between repeated joint counts ranged from 30 minutes to 7 days (Supplementary Tables 2 and 3, available with the online version of this manuscript).

Effect of training. Two studies analyzed the effect of patient training.[11,18] Radner, *et al*[11] reported a paradoxical reduction in the ICC after training for TJCs (0.75–0.59) and a small improvement for SJCs (0.32–0.35). Levy, *et al*[18] observed an improvement in Pearson correlation for both TJCs and SJCs (0.79–0.94 and 0.41–0.93, respectively).

Risk of bias. No study met all the QAREL checklist criteria. Common themes included lack of clarity as to whether participants (patients or HCPs) were blinded to discriminatory information such as inflammatory markers. Few studies explicitly commented on blinding of results between assessors. No details were available regarding sequence of assessments (HCP vs patient, TJC vs SJC; Supplementary Table 1, available with the online version of this manuscript)

DISCUSSION

We present the most comprehensive metaanalysis of self-reported joint count reproducibility to date, to our knowledge, describing measures of correlation, reliability, and agreement between patients and HCPs. The key finding is that the existing evidence supports self-reported joint counts as a reasonable measure to aid clinical decision making as part of disease activity indices such as the DAS, SDAI, and CDAI, although there are important caveats.

It is important to highlight the difference between measures of correlation and agreement. Assessing agreement assumes that 2 measures are comparing a common construct. In contrast,

Table 2. TJC correlation coefficients.

Study	Sample Size	Interval	Joint count	Assessor	Measure	Value[a]
Inderjeeth (Physician)[30]	52	NA	28[b]	Clinician	Pearson	0.59
Inderjeeth (Nurse)[30]	52	NA	28[b]	Nurse	Pearson	0.83
Heegaard[16]	31	Baseline	28 (Mannequin)	Clinician	Pearson	0.88
Heegaard[16]	31	7 days	28 (Mannequin)	Clinician	Pearson	0.93
Levy[18]	60	Baseline	28 (Mannequin)	Clinician	Pearson	0.79
Levy (Trained)[18]	30	Mean: 50 days	28 (Mannequin)	Clinician	Pearson	0.94
Prevoo (Site 1)[21]	141	NA	28 (Mannequin)	Clinician	Pearson	0.62
Prevoo (Site 2)[21]	101	NA	28 (Mannequin)	Clinician	Pearson	0.60
Radner[11]	78	Baseline, trained	28 (Mannequin)	Clinician	ICC	0.66
Radner[11]	131	Baseline, untrained	28 (Mannequin)	Clinician	ICC	0.85
Radner[11]	78	3 months, trained	28 (Mannequin)	Clinician	ICC	0.84
Radner[11]	131	3 months, untrained	28 (Mannequin)	Clinician	ICC	0.81
Janta[12]	69	NA	28 (Mannequin)	Clinician	ICC	0.51
Amaya-Amaya[23]	135	NA	28 (Mannequin)	Clinician	Kendall *W*	0.75
Cheung[24]	50	NA	28 (Mannequin)	Clinician	Cohen κ	0.64
Kavanaugh[13]	447	NA	28 (Mannequin)	Clinician	ICC	0.78
Riazzoli[26]	47	Baseline	28 (Mannequin)	Clinician	Spearman	0.87
Riazzoli[26]	47	3 months	28 (Mannequin)	Clinician	Spearman	0.87
Greenwood[29]	45	NA	28 (Mannequin)	Nurse	Spearman	0.92
El Miedany[17]	82	NA	28 (Mannequin)	Clinician	Pearson	0.84
Houssien[31]	100	NA	28 (Mannequin)	Clinician	Spearman	0.88
El Miedany and Palmer[27]	148	NA	28 (Mannequin)	Clinician	Correlation coefficient	0.82[c]
Wong[19]	27	Baseline	50 (Mannequin)	Clinician	Spearman	0.61
Wong[19]	33	Baseline	20 (Text)	Clinician	Spearman	0.37
Wong[19]	22	2 days	50 (Mannequin)	Clinician	Spearman	0.69
Wong[19]	28	2 days	20 (Text)	Clinician	Spearman	0.45
Hanly[20]	61	NA	20 (Text)	Clinician	Pearson	0.57
Abraham[22]	32	NA	Clinician 58/patient 60 (Text)	Clinician	Pearson	0.89
Alarcón[14]	67	NA	36 (Mannequin)	Clinician	ICC	0.64
Alarcón[14]	67	NA	36 (Text)	Clinician	ICC	0.55
Figueroa[25]	82	NA	Clinician 50/patient 42 (Text)	Clinician	Spearman	0.78
Calvo[28]	60	NA	36 (Mannequin)	Clinician	Spearman	0.77
Calvo[28]	60	NA	36 (Text)	Clinician	Spearman	0.75

[a] Correlation coefficient or reliability estimate. [b] Unknown whether text or mannequin. [c] Correlation coefficient not specified. ICC: intraclass correlation coefficient; NA: not applicable; TJC: tender joint count.

correlation can be used to describe unrelated constructs, and correlation can be high even if agreement is low. For example, if a patient consistently scored their SJCs lower than an HCP, correlation could be very good, but with low agreement. Few studies evaluated agreement, despite agreement representing a vital component of reproducibility.

Correlation between HCPs and patients was strong, although it was higher for TJCs than SJCs, consistent with a previous metaanalysis.[8] One explanation may be greater difficulty for individuals to discriminate a truly swollen joint as opposed to a bony deformation or swelling of other structures nearby.[26] Tenderness is reliant on symptoms, whereas swelling relies more on an objective measure from the assessor.[12,13]

Few studies reported measures of reliability (ICCs) but saw a similar pattern with lower reliability for SJCs than TJCs. The Bland-Altman plots offer additional insight into the reproducibility of the assessments, drawing upon data from all included studies. Across the studies, mean differences between patient- and clinician-reported scores were lower for SJCs than TJCs. The mean differences were stable across the range of SJC values, whereas for TJCs, agreement was excellent for low values but diminished as TJC values increased, demonstrating a positive bias for pain-dependent responses. This could be interpreted as evidence that self-reported joint assessments are in greater agreement at lower values, but as disease activity rises, the agreement of the self-reported joint count reduces. The clinical interpretation of this could be that a self-reported DAS that demonstrates low disease activity or remission is suitable for decision making; however, caution is needed when interpreting moderate or high DAS scores based upon self-reported joint counts. The latter point is relevant to decisions about biologic or targeted immune modulation therapies, whereby the use of patient-reported joint counts may be unsuitable.

An important question is whether differences in patient- and HCP-reported counts are above a clinically significant threshold. Detection of swollen joints may be more

Table 3. Swollen joint count correlation coefficients.

Study	Sample Size	Interval	Count	Assessor	Measure	Value[a]
Inderjeeth (Physician)[30]	52	NA	28[b]	Clinician	Pearson	0.42
Inderjeeth (Nurse)[30]	52	NA	28[b]	Nurse	Pearson	0.69
Heegaard[16]	31	Baseline	28 (Mannequin)	Clinician	Pearson	0.43
Heegaard[16]	31	7 d	28 (Mannequin)	Clinician	Pearson	0.60
Levy[18]	60	Baseline	28 (Mannequin)	Clinician	Pearson	0.60
Levy[18]	30	Mean: 50 days	28 (Mannequin)	Clinician	Pearson	0.93
Prevoo (Site 1)[21]	141	NA	28 (Mannequin)	Clinician	Pearson	0.61
Prevoo (Site 2)[21]	101	NA	28 (Mannequin)	Clinician	Pearson	0.65
Radner[11]	78	Baseline, trained	28 (Mannequin)	Clinician	ICC	0.33
Radner[11]	131	Baseline, untrained	28 (Mannequin)	Clinician	ICC	0.39
Radner[11]	78	3 months, trained	28 (Mannequin)	Clinician	ICC	0.48
Radner[11]	131	3 months, untrained	28 (Mannequin)	Clinician	ICC	0.41
Janta[12]	69	NA	28 (Mannequin)	Clinician	ICC	0.28
Amaya-Amaya[23]	135	NA	28 (Mannequin)	Clinician	Kendall *W* test	0.77
Cheung[24]	50	NA	28 (Mannequin)	Clinician	Cohen κ	0.56
Kavanaugh[13]	447	NA	28 (Mannequin)	Clinician	ICC	0.55
Riazzoli[26]	47	Baseline	28 (Mannequin)	Clinician	Spearman	0.75
Riazzoli[26]	47	3 months	28 (Mannequin)	Clinician	Spearman	0.58
Greenwood[29]	45	NA	28 (Mannequin)	Nurse	Spearman	0.67
Houssien[31]	100	NA	28 (Mannequin)	Clinician	Spearman	0.63
Wong[19]	27	Baseline	48 (Mannequin)	Clinician	Spearman	0.58
Wong[19]	33	Baseline	18 (Text)	Clinician	Spearman	0.58
Wong[19]	22	2 days	48 (Mannequin)	Clinician	Spearman	0.61
Wong[19]	28	2 days	18 (Text)	Clinician	Spearman	0.55
Hanly[20]	61	NA	20 (Text)	Clinician	Pearson	0.16
Figueroa[25]	82	NA	50 clinician; 42 patient (text)	Clinician	Spearman	0.34

[a] Correlation coefficient or reliability estimate. [b] Unknown whether text or mannequin. ICC: intraclass correlation coefficient; NA: not applicable; SJC: swollen joint count.

important than tender joints, as it is the persistence of objective inflammatory disease that predicts radiographic progression.[33] We are unable to describe how accurately self-reported joint counts could classify people into discrete disease activity bands, such as remission, low, moderate, or high activity. From the information we present, it is likely that accuracy of self-reported joint counts will be better for remission and low disease activity states compared to more active disease.

A subsequent question to ask is whether variability between patients and HCPs varies from interobserver differences between HCPs. There is a paucity of published data on interobserver variability for joint counts performed by HCPs, but available research suggests that interobserver variability is not dissimilar between clinicians and that agreement is also worse for SJCs than TJCs, with similar magnitudes in difference.[34,35]

There are limitations to our metaanalysis. Studies have been included from several decades, and over this time understanding of the effect of health literacy on health outcomes and equity within health care has evolved, potentially adding confounding over time. The studies were heterogeneous in design and risk of bias was substantial. For example, it was often unclear how many assessors were involved, or whether they were blinded to objective measures of disease activity (inflammatory markers or imaging results) at the time of performing joint counts.

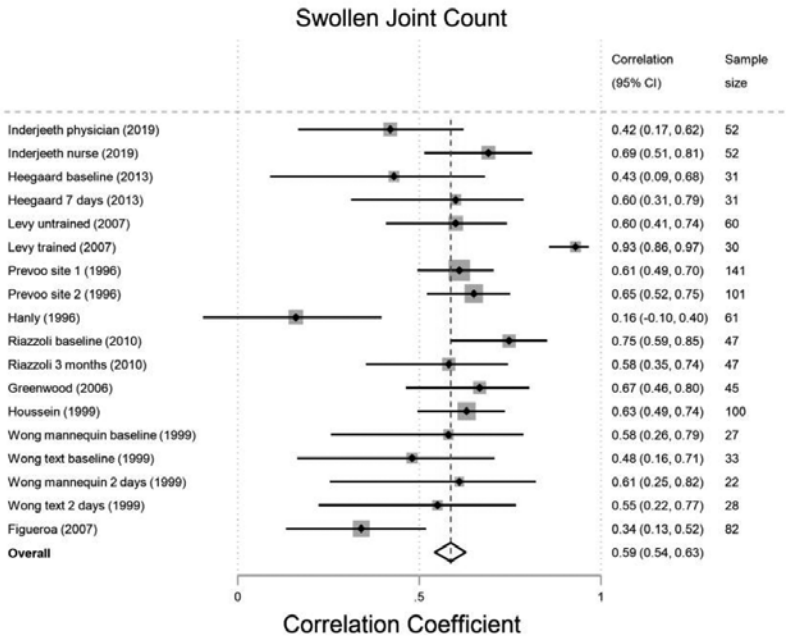
The studies lacked detailed information on patient socioeconomic position, educational background, or prior health awareness. In clinical research, more educated patients tend to volunteer to participate in studies.[36] This is a pertinent issue as health literacy and patient educational level may have an effect on the reliability of patient-reported joint counts. Future studies should aim to capture health literacy level and ensure inclusion of a diverse patient population representative of patients seen within clinical practice.

Finally, concomitant fibromyalgia (FM) was not accounted for. Patients with FM and RA report higher TJCs and pain scores but not SJCs.[37]

The increased use of remote monitoring in RA management requires a greater understanding of the reliability and agreement of self-reported disease activity measures. We present evidence to inform the use of self-reported joint counts. There is good correlation between patient- and clinician-reported joint counts. Reliability is lower than correlation, although fewer studies reported on this. As per our use of the Bland-Altman–type plot, agreement was better for lower values of TJCs. Self-reported joint counts in RA without concomitant FM now have sufficient reproducibility to justify their use in routine practice.

ONLINE SUPPLEMENT

Supplementary material accompanies the online version of this article.



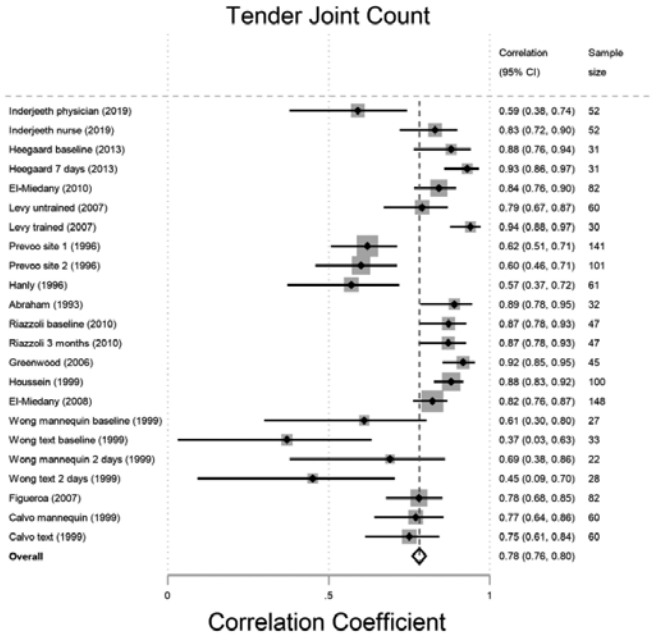


Figure 2. Forest plots for correlation coefficients of tender and swollen joint counts.

REFERENCES

1. Smolen JS, Landewé R, Bijlsma J, Burmester G, Chatzidionysiou K, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. Ann Rheum Dis 2017;76:960-77.
2. Singh JA, Saag KG, Bridges SL Jr, Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Rheumatol 2016;68:1-26.
3. National Institute for Health and Care Excellence (NICE). Rheumatoid arthritis in adults: management. NICE guideline [NG100]. [Internet. Accessed July 16, 2021.] Available from: www.nice.org.uk/guidance/ng100
4. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core

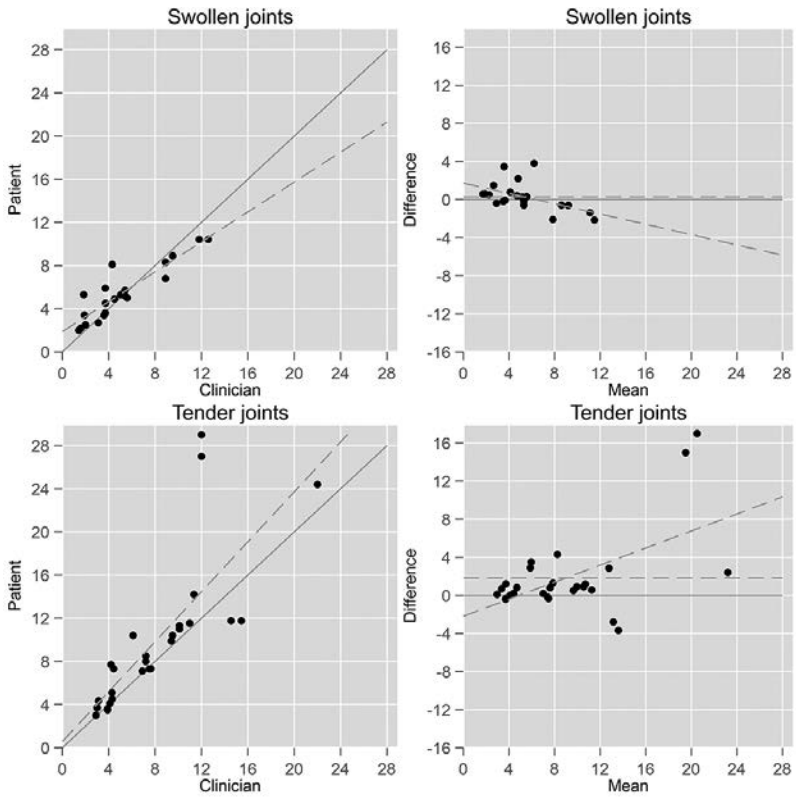


Figure 3. Comparison of mean tender and swollen joint counts between patient and HCP. Bland-Altman–type plots for mean patient and mean HCP tender and swollen joint counts recorded in individual studies included in this systematic review. The 2 plots on the left show mean patient plotted against mean HCP joint counts. The 2 plots on the right show the mean of patient and HCP joint count plotted against the difference (patient – HCP) in joint count. HCP: healthcare practitioner.

set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.

5. Pincus T, Brooks RH, Callahan LF. Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. Ann Intern Med 1994;120:26-34.
6. Osborne RH, Wilson T, Lorig KR, McColl GJ. Does self-management lead to sustainable health benefits in people with arthritis? A 2-year transition study of 452 Australians. J Rheumatol 2007;34:1112-7.
7. Cheung PP, Gossec L, Mak A, March L. Reliability of joint count assessment in rheumatoid arthritis: a systematic literature review. Semin Arthritis Rheum 2014;43:721-9.
8. Barton JL, Criswell LA, Kaiser R, Chen YH, Schillinger D. Systematic review and metaanalysis of patient self-report versus trained assessor joint counts in rheumatoid arthritis. J Rheumatol 2009;36:2635-41.
9. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.
10. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol 2010;63:854-61.
11. Radner H, Grisar J, Smolen JS, Stamm T, Aletaha D. Value of self-performed joint counts in rheumatoid arthritis patients near remission. Arthritis Res Ther 2012;14:R61.
12. Janta I, Naredo E, Martínez-Estupiñán L, Nieto JC, De la Torre I, Valor 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 2013;52:2243-50.
13. Kavanaugh A, Lee SJ, Weng HH, Chon Y, Huang XY, Lin SL. Patient-derived joint counts are a potential alternative for determining Disease Activity Score. J Rheumatol 2010;37:1035-41.
14. Alarcón GS, Tilley BC, Li S, Fowler SE, Pillemer SR. Self-administered joint counts and standard joint counts in the assessment of rheumatoid arthritis. MIRA Trial Group. Minocycline in RA. J Rheumatol 1999;26:1065-7.
15. Giavarina D. Understanding Bland Altman analysis. Biochem Med 2015;25:141-51.
16. Heegaard C, Dreyer L, Egsmose C, Madsen OR. 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.
17. El Miedany Y, El Gaafary M, Youssef SS, Palmer D. Incorporating patient reported outcome measures in clinical practice: development and validation of a questionnaire for inflammatory arthritis. Clin Exp Rheumatol 2010;28:734-44.
18. Levy G, Cheetham C, Cheatwood A, Burchette R. Validation of patient-reported joint counts in rheumatoid arthritis and the role of training. J Rheumatol 2007;34:1261-5.
19. Wong AL, Wong WK, Harker J, Sterz M, Bulpitt K, Park G, et al. Patient self-report tender and swollen joint counts in

early rheumatoid arthritis. Western Consortium of Practicing Rheumatologists. J Rheumatol 1999;26:2551-61.
20. Hanly JG, Mosher D, Sutton E, Weerasinghe S, Theriault D. Self-assessment of disease activity by patients with rheumatoid arthritis. J Rheumatol 1996;23:1531-8.
21. Prevoo ML, Kuper IH, van't Hof MA, van Leeuwen MA, van de Putte LB, van Riel PL. Validity and reproducibility of self-administered joint counts. A prospective longitudinal followup study in patients with rheumatoid arthritis. J Rheumatol 1996;23:841-5.
22. Abraham N, Blackmon D, Jackson JR, Bradley LA, Lorish CD, Alarcón GS. Use of self-administered joint counts in the evaluation of rheumatoid arthritis patients. Arthritis Care Res 1993;6:78-81.
23. Amaya-Amaya J, Botello-Corzo D, Calixto OJ, Calderón-Rojas R, Domínguez AM, Cruz-Tapias P, et al. Usefulness of patients-reported outcomes in rheumatoid arthritis focus group. Arthritis 2012;2012:935187.
24. Cheung PP, Ruyssen-Witrand A, Gossec L, Paternotte S, Le Bourlout C, Mazieres M, et al. Reliability of patient self-evaluation of swollen and tender joints in rheumatoid arthritis: a comparison study with ultrasonography, physician, and nurse assessments. Arthritis Care Res 2010;62:1112-9.
25. Figueroa F, Braun-Moscovici Y, Khanna D, Voon E, Gallardo L, Luinstra D, et al. Patient self-administered joint tenderness counts in rheumatoid arthritis are reliable and responsive to changes in disease activity. J Rheumatol 2007;34:54-6.
26. Riazzoli J, Nilsson JÅ, Teleman A, Petersson IF, Rantapää-Dahlqvist S, Jacobsson LT, 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 2010;49:2098-103.
27. El Miedany Y, Palmer D. Can standard rheumatology clinical practice be patient-based? Br J Nurs 2008;17:673-5.
28. Calvo FA, Calvo A, Berrocal A, Pevez C, Romero F, Vega E, et al. Self-administered joint counts in rheumatoid arthritis: comparison with standard joint counts. J Rheumatol 1999;26:536-9.
29. Greenwood MC, Hakim AJ, Carson E, Doyle DV. Touch-screen computer systems in the rheumatology clinic offer a reliable and user-friendly means of collecting quality-of-life and outcome data from patients with rheumatoid arthritis. Rheumatology 2006; 45:66-71.
30. Inderjeeth CA, Inderjeeth AJ, Raymond WD. A multicentre observational study comparing patient reported outcomes to assess reliability of swollen and tender joint assessments and response to certolizumab treatment as compared to clinician assessments in rheumatoid arthritis. Int J Rheum Dis 2019;22:73-80.
31. 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 1999;38:48-52.
32. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33:363-74.
33. Aletaha D, Smolen JS. Joint damage in rheumatoid arthritis progresses in remission according to the Disease Activity Score in 28 joints and is driven by residual swollen joints. Arthritis Rheum 2011;63:3702-11.
34. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol 2001; 28:892-903.
35. Marhadour T, Jousse-Joulin S, Chalès G, Grange L, Hacquard C, Loeuille D, et al. Reproducibility of joint swelling assessments in long-lasting rheumatoid arthritis: influence on Disease Activity Score-28 values (SEA-Repro study part I). J Rheumatol 2010;37:932-7.
36. Katz MG, Jacobson TA, Veledar E, Kripalani S. Patient literacy and question-asking behavior during the medical encounter: a mixed-methods analysis. J Gen Intern Med 2007;22:782-6.
37. Pollard LC, Kingsley GH, Choy EH, Scott DL. Fibromyalgic rheumatoid arthritis and disease assessment. Rheumatology 2010;49:924-8.