The Journal of Rheumatology 2021;xx:xxxx
doi:10.3899/jrheum.210822
First Release September 1 2021

The Journal of Rheumatology

Editorial

The Reproducibility of Patient Self-reported Joint Counts in Rheumatoid Arthritis: A Closer Look at the Results of a Review

Salima Francis Elisabeth van Weely[1]

In the management of rheumatoid arthritis (RA), the systematic evaluation of disease activity is of paramount importance. It is the cornerstone of the "treat-to-target" approach, aiming at disease remission and optimization of quality of life.[1] In times of increasing delivery of remote care, accelerated due to the coronavirus disease 2019 (COVID-19) pandemic, the role of patients in the monitoring of disease activity becomes more topical. Since this decline in personal contact in the clinic could be permanent, patient self-assessment could fill this gap. In addition, patient self-assessment can also improve patient engagement and encourage self-management behavior.[2] A generally accepted and globally used outcome measure to assess disease activity in RA is the Disease Activity Score in 28 joints (DAS28).[3] Next to a laboratory variable (C-reactive protein or erythrocyte sedimentation rate), the DAS28 is composed of weighted values of 28 swollen (SJC) and tender joint counts (TJC) and a patient global assessment. Joint count assessments aimed at detecting clinical synovitis have been shown to be predictive of mortality and are generally done by a healthcare professional (HCP).[4] Rampes, *et al* examined the extent to which self-reported TJC/SJC between patients with RA and HCPs are sufficiently reproducible to be a justified option in calculating disease activity.[2]

Rampes, *et al*[2] conducted a thorough review on the reproducibility of patient self-reported joint counts in RA that updates prior reviews. They analyzed the literature on the measurement properties of patient-reported joint counts in clinical practice and stated that their group was the first to consider agreement. A previous review showed that patient interobserver reliability with HCPs as comparators was better for TJCs (intraclass correlation coefficient [ICC] range 0.31–0.91) compared to SJCs (0.16–0.64).[5] The findings of Rampes, *et al*[2] confirm that the interobserver reliability of joint counts between patients and HCPs varies between moderate to good, and that the reliability for SJC is lower than for TJC.[5,6,7] The interrater reliability of SJC varied from fair to substantial (0.28–0.77), whereas for TJC it varied from moderate to good (0.51–0.85).[2] These findings highlight the potential of patients acting as their own observer in measuring joint counts between clinic visits over time, and of patient self-assessment as an outcome measure in clinical trials.[5,6] The review is a timely topic of relevance to patients, clinical practice, and research, but some limitations specific to this study were identified. By outlining these limitations, future research on the reliability of patient-reported SJC/TJC could be beneficial.

Rampes, *et al*[2] used the term *reproducibility*, which is part of the domain of reliability. It can be divided into the measurement properties reliability and measurement error.[8] Good guidance for definitions of these concepts can be found in the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) taxonomy, which describes the terminology and definitions of these clinimetric concepts (Table 1).[9,10,11] In the setting of this study,[2] the assessment of reliability is about finding the same SJC/TJC score in a patient with RA if you expect the same score despite a changing condition (i.e., a different assessor; in this case, either HCP or patient). An important assumption made in reliability studies (and in studies on measurement error) is that patients are stable regarding the construct to be measured between the repeated measurements; in this case, there should be no changes in symptoms of clinical synovitis.[9,10,11] Further, reliability parameters are highly dependent on the heterogeneity of the study sample, since reliability can also be explained as the ability of a measurement to distinguish between patients. Within a homogeneous group, it is hard to distinguish between patients.[9,10,11] Based on the results of the Quality Appraisal of Diagnostic Reliability (QAREL) checklist and Tables 1–3 in the study of Rampes, *et al*,[2] it could be

[1]*S.F.E. van Weely, PT, PhD, Senior Researcher, Department of Orthopaedics, Rehabilitation and Physical Therapy, Leiden University Medical Center, Leiden, the Netherlands.*

The author declares no conflicts of interest relevant to this article.

Address correspondence to Dr. Salima FE van Weely, LUMC (Leiden University Medical Center), Department of Orthopaedics, Rehabilitation and Physical Therapy, J11, P.O. Box 9600, 2300 RC Leiden, the Netherlands. Email: s.f.e.van_weelij@lumc.nl.

See Patient-reported joint counts, *page xxx*

Table 1. Definition and preferred statistical methods according to COSMIN taxonomy.[9,10,11]

Domain	Measurement Property	Definition	Statistical Method
Reliability		The degree to which the measurement is free from measurement error	
Reliability (extended definition)		The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions (e.g., over time [test-retest]; by different persons on the same occasion [interrater]; or by the same persons [raters or responders] on different occasions [intrarater])	
	Reliability	The proportion of the total variance in the measurements which is due to "true" differences between patients	ICC or weighted κ
	Measurement error	The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured	SEM, LOA, SDC

COSMIN: COnsensus-based Standards for the selection of health Measurement Instruments; ICC: intraclass correlation coefficient; LOA: limits of agreement; SDC: smallest detectable change; SEM: standard error of measurement.

assumed that the patient populations of the 14 included studies were representative of the population of interest, were stable, and contained some degree of heterogeneity. Information on the sample size, age, sex, race, education, and social economic status was given, but information on variables such as disease duration, disease severity, and test setting for the individual studies included seemed to be lacking. Also, the (average) time interval for the test-retest assessment of SJC/TJC was mostly reported as not applicable. The time interval should be long enough to prevent recall bias and short enough to ensure that patients have not changed in the construct to be measured. Although perhaps not likely in this case, symptoms may vary between 2 test situations depending on the time interval, and can affect classification judgment, resulting in a greater difference between HCPs and patients. The lack of information on disease variables and time interval hinders the interpretation of the results, representativeness of the population for generalizability to other populations, and the quality of the reported results. Future studies and/or reviews could improve this with more extensive reporting on these results.

In their results section, Rampes, *et al* reported the correlation coefficients, reliability estimates, and agreement for each study.[2] They found that 13 of 20 studies reported Pearson or Spearman correlation coefficients between the HCPs and the patients' assessments. Historically, Pearson and Spearman correlation coefficients have been used to quantify reliability. These measurements are no longer considered accurate because they do not account for systematic errors and only quantify the strength of an association between 2 parameters, not the reliability.[5,9–13] Pearson or Spearman correlation coefficients may be used if there is evidence that no systematic change has occurred. Rampes, *et al* performed a metaanalysis with the correlation coefficients and provided a summary estimate in a forest plot. Information about possible systematic differences seemed to be lacking. It remains unknown if they were not presented in the original articles or if this information was not extracted. The uninformed reader might be confused by the information presented in the forest plot. Future studies or reviews on (intra- or interrater) reliability of patient-reported joint counts may be beneficial if they include only measures of reliability that are currently considered appropriate or if a separate analysis is performed including the studies with appropiate measures such as an ICC only.

In the review of Rampes, *et al*,[2] it was found that 5 studies reported an ICC or κ to substantiate the reliability of SJC/TJC,[14,15,16,17,18] with 1 study[16] not included in an earlier systematic review on this topic.[5] An ICC ranges from 0 to 1 and a score of > 0.70 is required for the comparison of groups, whereas an ICC of > 0.90 is recommended for individual evaluation.[9,10,11] The ICCs for TJC ranged from 0.51 to 0.85, indicating a moderate to good reliability, and for SJC from 0.28 to 0.55, indicating a poor to moderate reliability. Only 2 of the individual studies that reported the use of ICC[14,15] reported a value of > 0.70 for TJC, supporting the use of TJC for comparison in groups. For SJC, this threshold was not reached. Rampes, *et al*[2] did not report an ICC > 0.90 for TJC or SJC in any of the studies; this is the cut-off value that would support their use in individuals. The ICC and the 95% CIs of the ICC were reported, but information on the form or formula of the ICCs was either not reported or not available in the individual studies. Many forms of ICC exist and are appropriate in different situations in the assessments of reliability of a measurement.[12,19] Future studies and reviews on the reliability of joint counts could benefit from the use of ICC or κ as well as more comprehensive information on the statistical methods used to substantiate the reliability.

Rampes, *et al* stated that their review was the first to also consider and assess the agreement of TJC/SJC measures: "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).[2] Providing information about agreement

parameters (i.e., measurement error) of SJC/TJC in addition to a reliability parameter is strongly encouraged in situations where the instrument will be used for evaluation of individuals, as it facilitates defining true change from measurement error. The Bland-Altman plots presented by Rampes, *et al*, in which the mean SJC/TJC was plotted against the difference between measures or raters (i.e., patient and HCP), might benefit from adding 95% limits of agreement (± 1.96 SD intervals)[9,10,11,19] for interpretation. However, information on systematic differences or measurement error between patients and HCPs was provided. The authors state that in the studies (unknown as to which ones; Figure 3)[2] that were included in this specific analysis, patients reported on average 1.1 more tender joints than HCPs, but this discrepancy was found not to be constant. The measurement error was negligible if TJC was < 5 joints, but patient overestimation increased if TJC was > 5. For SJC, the difference was reported to be negligible or trended in the opposite direction, with patients reporting a lower SJC than HCPs.[2] One may wonder whether the interpretation of Rampes, *et al* could also be the other way around: Did HCPs underestimate when the TJC was > 5? Importantly, these results seem to underline an overall (high) variability in reliability and agreement parameters between raters of SJC/TJC, independent of whether they are patients or HCPs, as the interobserver variability between clinicians is not dissimilar, as also reported by Rampes, *et al*.[2] A combination of repeated measurements and the application of training of HCPs and patients may lead to a (small) increase in the reliability and a decrease measurement error.[14,20] When using patient-reported TJC/SJC as part of disease activity indices, more information about agreement parameters is needed for good interpretation; caution should be paid when using patient-reported TJC/SJC as the sole measurement to support clinical decision making. Future studies could provide for this by calculating agreement parameters such as the measurement error more often.

The focus on patient-reported outcome measures fits in with the spirit of the times with more patient-centered care, more remote care, and a greater focus on shared decision making. Several initiatives worldwide give patients a greater role in assessing their disease in the context of the right care in the right place. Rampes, *et al* rightly state that health literacy and patient education—specifically patient-reported SJC/TJC in this case—are likely to affect remote care. Future studies could incorporate aspects of health literacy in their design to better understand their influence on the reliability of these measures. In addition to the benefits of remote care, the added value of face-to-face contact and communication options must also be considered. Especially in the light of potentially limited health literacy skills, face-to-face communication about symptoms and complaints experienced by patients can contribute to shared decision making and a treat-to-target treatment.

To conclude, Rampes, *et al* reported on several studies analyzing the reproducibility of SJC/TJC in RA between HCPs and patients.[2] Overall, studies showed moderate reliability, with higher reliability for TJC than SJC.[2,5,6] The results support the use of SJC/TJC at a group level only, such as in intervention research. Higher reliability and more information about the measurement error (i.e., agreement parameters) are needed for use in individuals. Future studies and reviews could facilitate this by paying attention to appropriate measures of reliability, reporting more comprehensive information about the study population and statistics to interpret the data, and adding analyses and information on measurement error. The review of Rampes, *et al* underscores the will and the possibilities of using patient-reported SJC/TJC in future to facilitate patient self-management behavior.[2] Based on their results, the use of patient-reported joint counts at the individual level to assess disease activity for patients with RA could be encouraged as a discussion tool between patients and HCPs in shared decision making.

REFERENCES

1. Smolen JS, Landewé RBM, Bijlsma JWJ, Burmester GR, Dougados M, Kerschbaumer A, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. Ann Rheum Dis 2020;79:685-99.
2. Rampes S, Patel V, Bosworth A, Jacklin C, Nagra D, Yates M, et al. Systematic review and metaanalysis of the reproducibility of patient self-reported joint counts in rheumatoid arthritis. J Rheumatol 2021;xxxxxxxxx.
3. Prevoo ML, van't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44-8.
4. 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.
5. 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.
6. 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.
7. 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.
8. Dekker J, Dallmeijer AJ, Lankhorst GJ. Clinimetrics in rehabilitation medicine: current issues in developing and applying measurement instruments 1. J Rehabil Med 2005;37:193-201.
9. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, De Vet HC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018; 27:1147-57.
10. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, Terwee CB. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1171-9.
11. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res 2018;27:1159-70.
12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.

13. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use (5 ed.). Oxford: Oxford University Press; 2015.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. Erratum in: J Chiropr Med 2017;16:346.
20. 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.