RHEUMATOLOGY

Rheumatology 2016;55:41–48
doi:10.1093/rheumatology/kev282
Advance Access publication 10 August 2015

Original article

Reliability of patient-reported outcomes in rheumatoid arthritis patients: an observational prospective study

Paul Studenic[1], Tanja Stamm[1], Josef S. Smolen[1,2] and Daniel Aletaha[1]

Abstract

Objective. Patient-reported outcomes (PROs) such as pain, patient global assessment (PGA) and fatigue are regularly assessed in RA patients. In the present study, we aimed to explore the reliability and smallest detectable differences (SDDs) of these PROs, and whether the time between assessments has an impact on reliability.

Methods. Forty RA patients on stable treatment reported the three PROs daily over two subsequent months. We assessed the reliability of these measures by calculating intraclass correlation coefficients (ICCs) and the SDDs for 1-, 7-, 14- and 28-day test–retest intervals.

Results. Overall, SDD and ICC were 25 mm and 0.67 for pain, 25 mm and 0.71 for PGA and 30 mm and 0.66 for fatigue, respectively. SDD was higher with longer time period between assessments, ranging from 19 mm (1-day intervals) to 30 mm (28-day intervals) for pain, 19 to 33 mm for PGA, and 26 to 34 mm for fatigue; correspondingly, ICC was smaller with longer intervals, and ranged between the 1- and the 28-day interval from 0.80 to 0.50 for pain, 0.83 to 0.57 for PGA and 0.76 to 0.58 for fatigue. The baseline simplified disease activity index did not have any influence on reliability. Lower baseline PRO scores led to smaller SDDs.

Conclusion. Reliability of pain, PGA and fatigue measurements is dependent on the tested time interval and the baseline levels. The relatively high SDDs, even for patients in the lowest tertiles of their PROs, indicate potential issues for assessment of the presence of remission.

Key words: RA, reliability, patient perspective, test–retest, disease activity, outcomes research, psychometrics, patient reported outcomes.

Rheumatology key messages

- In RA, smallest detectable difference for patient global assessment, pain and fatigue is 25, 25 and 30 mm, respectively.
- A threshold for true change of 20 mm or below is applicable in evaluation of RA patients in clinical remission.

Introduction

Patient-reported outcomes (PROs) are widely employed in clinical studies of RA patients as well as clinical practice, because they are an important part of the assessment of disease activity and response to therapy [1–3]. Key PRO domains include pain, patient global assessment (PGA) of disease activity and fatigue. Typically, all are evaluated on a 100-mm visual analogue scale (VAS) [4–6].

Several concepts are used to characterize these measures for use and interpretation in clinical practice. To map changes on scales to the perception of the patient, the minimal clinically important difference is used [7]. Several studies addressed the determination of minimal clinically important difference in VAS measurements of PROs in various rheumatic diseases [8], as it conveys important information when evaluating response to treatment.

[1]Division of Rheumatology, Department of Internal Medicine 3, Medical University Vienna and [2]2nd Department of Medicine, Hietzing Hospital, Vienna, Austria

Submitted 11 August 2014; revised version accepted 8 July 2015

Correspondence to: Daniel Aletaha, Division of Rheumatology, Department of Internal Medicine 3, Medical University Vienna, 1090 Vienna, Austria. E-mail: daniel.aletaha@meduniwien.ac.at

Fewer data exist on the so-called reliability of an instrument, which characterizes its stability or reproducibility in a test–retest setting [9, 10]. Thus, reliability needs to be determined during stable disease and personal and environmental factors [11]. Statistically, the intraclass correlation (ICC) coefficient and the smallest detectable difference (SDD) can be calculated as relative and absolute measures of reliability [10–12].

In RA, for pain and PGA, moderate to poor reliability was found for both measures in a test–retest setting, with relatively large SDDs [13, 14]. Moreover, pain scores contribute strongly to the patient's estimation of disease activity [15, 16]. Fatigue is a serious symptom of chronic musculoskeletal diseases and is associated with worse functional outcomes, but a proper evaluation of its SDD has not yet been performed [17, 18]. This may become clinically highly relevant if treatment decisions are based on instruments that include these measures, such as when following a treat-to-target approach in RA [19] using disease activity indices or the provisional ACR/EULAR remission criteria as the target. All of these comprise the PGA [20], and it is therefore very relevant to understand which variability is underlying this particular measure [21, 22]. Here, we aimed to determine the ICC and SDD for changes in pain, PGA and fatigue, and to investigate how these properties change over increasing periods of time between two assessments, and how they are influenced if patients have higher or lower initial measurements.

Methods

Study design

Forty consecutive patients from routine clinical care classified as having RA by the ACR 1987 revised criteria [23] or the ACR/EULAR 2010 criteria [24] were randomized by a computerized allocation programme into two different groups. All patients visited our clinic at baseline, at day 28 (follow-up visit 1) and at day 56 (follow-up visit 2). Between these assessments, one group of 20 patients kept daily records of their pain, fatigue and PGA in a diary; they had been encouraged to perform the assessments at the same time every day. The other 20 patients were called daily by telephone at random time points (between 8 a.m. and 4 p.m.), and asked to assess these VASs and to report the results to the study team during the call. At the time of the three clinic visits, pain, fatigue and PGA scores, as well as all other core set variables of disease activity, were obtained from each patient. During these 8 weeks, patients remained on stable treatment with DMARDs, glucocorticoids and NSAIDs.

PGA was assessed on a 100-mm VAS, using no disease activity and highly active disease as anchors. The wording of the question was: how do you estimate your disease activity today? (originally in German: Wie schätzen Sie heute Ihre Krankheitsaktivität ein?). Pain was evaluated on a 100-mm VAS, responding to the question: how severe is your pain today? (originally in German: Wie stark sind Ihre Schmerzen heute?), using no pain and unbearable pain as anchors. Fatigue was also assessed on a 100-mm VAS, asking: how strong was your fatigue today (originally in German: Wie stark war Ihre Müdigkeit heute?), using no fatigue at all and worst imaginable fatigue as anchors. The study patients have not been trained specially on how to answer the three questions; thus, we provided the same information as is provided to any other patient in routine care. There the questionnaire is handed out by a health professional, who briefly explains the use of a VAS and points out the respective anchors. The patient fills out the questionnaire while waiting for the physician. The ethic committee of the Medical University Vienna approved the study, and written consent was obtained according to the Declaration of Helsinki from all patients.

Reliability analyses using ICC and SDD of PGA, pain and fatigue

The ICC can be used to assess the reliability of two or more measurements and results as a value between 0 and 1. An ICC of 1 means that 100% of the variability in the measurements is due to differences between patients (i.e. no error, no within-patient variability: perfect reliability), while an ICC of 0 means that all variability is related to within-patient variability and error. This is based on a very generic formula dividing the true variance (within patient variability) by the observed variance (the total variance) [25–28].

In contrast to the ICC as a relative measure of reliability, the SDD provides a cut-off value for the smallest amount of difference that is needed to reliably distinguish true change from measurement error [8, 9, 29]. The SDD is calculated by multiplying the S.D. of the difference between two assessments by 1.96. Subtracting or adding the SDD to the mean difference is known as the limits of agreement, as described by Bland and Altman [29].

We also calculated standardized response means for PGA, pain and fatigue to assess whether a change over time that is greater than random has occurred [30].

Assessing differences in variability between the telephone and the diary group

Baseline characteristics of the two groups were compared by parametric or non-parametric tests, as appropriate. The course of PGA, pain and fatigue levels were analysed separately for each patient. We calculated ICC and SDD of PGA, pain and fatigue for the various test–retest intervals, separately for the diary and the telephone groups. Thus, we evaluated whether or not the method of obtaining repeated measurements by the patients (diary or telephone report) was an important determinant of results and, consequently, if they could be used jointly for further analyses.

Reliability for increasing intervals between two assessments

To investigate how reliable measurements remain as measurement intervals increase, we calculated ICC for pain, PGA and fatigue separately for 1-, 7-, 14- and

28-day test–retest intervals. In other words, given the 56 days of repeated assessments, we calculated 55 ICCs for the 1-day test–retest interval (days 1–2, 2–3, etc.), 50 ICCs for the 7-day interval, 43 ICCs for the 14-day interval and 29 ICCs for the 28-day interval.

ICC in this test–retest setting was calculated with a two-way mixed design because all patients were evaluated, and the VAS assessments were performed on consecutive days and thus were not purely at random; this is based on the model ICC by Shrout and Fleiss (1979) [31]. In our case, we further assumed an absolute agreement between assessment days, meaning that the model was not adjusted for differences in mean score between days [12, 31]. For the calculations of the SDDs for the three PROs, we proceeded in an analogous way.

Reliability of measurements in patients with different baseline levels of disease activity

To be able to discriminate between patients with higher or lower variability/reliability of PROs, we calculated ICC and SDD separately for certain subgroups of patients. We grouped patients by forming tertiles according to their baseline simplified disease activity index (SDAI), their baseline value of PGA, their baseline pain level and their baseline fatigue level. Further, we divided patients according to whether they had $\leqslant$10 mm PGA at baseline or $>$10 mm PGA. Then we calculated SDDs separately for those groups. SDDs and ICCs were again calculated for 1-, 7-, 14- and 28-day test–retest intervals. In a sensitivity analysis, we excluded the top 10% of patients with the largest changes in the SDAI (improvement and deterioration) during the 2-month study period. Thus, we repeated the above-described analyses in the remaining 80% of patients with more stable disease.

Results

Patient characteristics

Forty RA patients {85% female, 60% RF positive, median SDAI: 13.4 [interquartile range (IQR) 6.5–20.4], median disease duration 9.5 years (IQR 5.0–14.8), Table 1} participated in this study. Despite randomization, there were some numerical (though not statistically significant) differences in baseline and follow-up disease characteristics in pain, PGA and fatigue between the telephone and diary groups. Medication remained stable over the study period, and NSAID use was balanced in each group (75% of patients in each group used NSAIDs); even minor changes in disease activity, as reflected by an SDAI 50% response [32], were found only in 18% of the patients at the first follow-up visit and in 25% at the second follow-up visit. Over the 2-month period, the top 10% of patients in terms of worsening had an SDAI change of 8.3 or more, and the 10% of the patients who improved had a change of −8.9 or less; the median change in SDAI was −0.9 (IQR −2.3 to 2.9). Concerning the PROs, the standardized response means for the various test–retest intervals ranged between −0.012 and 0.029 for pain, between −0.004 and 0.053 for PGA and between −0.06 and 0.01 for fatigue, thus supporting the notion of an absence of true change. Since PGA is an integral part of the SDAI, we also calculated the SDD of the SDAI. For the first month interval the SDD was 10.73, and for the second month interval it was 12.67, which is a mean SDD for the SDAI of 11.7 (S.D. 1.37).

Table 1 Baseline characteristics of total patient group and separately for patients assessed by telephone or by diary

Characteristic	Baseline		
	Total	Diary	Telephone
Patients, *n*	40	20	20
Age, years	52.8 (54.5–61.5)	51.0 (42.25–57.0)	57.5 (50.25–65.0)
Female, %	85.0	85.0	85.0
Duration of disease, years	9.5 (5.0–14.75)	9.0 (5.25–13.75)	11.5 (4.25–20.25)
CRP, mg/dl	0.57 (0.5–1.32)	0.62 (0.50–1.51)	0.5 (0.5–1.15)
ESR, mm/h	16.5 (10.25–30.0)	13.0 (11.25–23.5)	20.0 (10.0–35.75)
Pain (visual analogue scale), mm	16.5 (7.5–28.5)	16.0 (5.5–20.75)	19.5 (9.25–37.0)
Patient Global Assessment, mm	20.0 (10.0–37.25)	18.0 (10.0–37.25)	21.5 (8.35–37.25)
Evaluator Global Assessment, mm	24.5 (18.0–43.75)	25.5 (18.0–47.75)	22.5 (11.5–42.75)
Swollen joint count, 28 joints	3.0 (1.25–8.75)	4.0 (0.25–8.75)	3 (2.0–9.5)
Tender joint count, 28 joints	1.0 (0.0–2.75)	1.0 (0.0–2.75)	1.0 (0.0–2.75)
Health Assessment Questionnaire	0.63 (0.28–1.22)	0.63 (0.28–0.97)	0.5 (0.28–1.34)
RF positive, %	60.0	65.0	55.0
Simplified disease activity index	13.4 (6.5–20.42)	12.5 (6.8–20.0)	14.4 (6.0–21.2)
Clinical disease activity index	11.9 (5.83–18.18)	11.9 (5.9–18.2)	11.8 (5.0–19.5)
DAS28	3.12 (2.72–4.15)	3.1 (2.6–3.9)	3.2 (2.9–4.3)
NSAIDs, %	72.5	75.0	70.0
Glucocorticoids, %	37.5	30.0	45.0

Median (interquartile range), unless indicated otherwise.

Reliability and SDD and the influence of the length of assessment intervals

The overall ICCs for pain, PGA and fatigue in the 1-day/7-day test–retest interval were 0.8/0.67, 0.83/0.71 and 0.76/0.66, respectively (Table 2). Correspondingly, the overall SDDs in millimetres were 18.8/24.5, 19/25 and 25.9/30.2 (Table 3). As expected, higher reliability according to ICC coincided with smaller SDD and vice versa.

Comparing the ICC and the SDD across the various assessment intervals, there was a significant trend towards lower ICC and higher SDD with longer intervals for all three measures (Tables 2 and 3). The SDDs of pain and PGA differed by the same amount between the 1-day test–retest and the 7-day test–retest intervals. This corresponds to an increase of 6 mm from 19 mm (S.D. 4.7) to 25 mm (S.D. 4.7) for pain and from 19 mm (S.D. 4.7) to 25 mm (S.D. 6) for PGA. The same difference was then observed between the 7- and 28-day test–retest intervals [increase by further 5–30 mm (S.D. 5.8) for pain and 30 mm (S.D. 6.1) for PGA]. SDDs for fatigue were generally higher, starting with 25.9 mm (S.D. 5.7) for the 1-day interval, which increased by 4.3 mm at the 7-day test–retest interval. The difference in SDDs of 3.7 mm between the 7- and the 28-day test–retest intervals was again about the same as between the 1- and 7-day intervals. The differences between ICCs and between SDDs comparing the 28-day interval with the 1-day test–retest intervals were 0.30 and 10.8 mm, respectively, for pain, 0.27 and 10.9 mm for PGA and 0.17 and 8.1 mm for fatigue.

Differences between telephone and diary groups

The reliability as expressed by the ICC for the 1-day test–retest intervals was very similar in both the diary and the telephone group [0.78 (S.D. 0.11)/0.80 (S.D. 0.10) for pain; 0.82 (S.D. 0.09)/0.83 (S.D. 0.10) for PGA; 0.77 (S.D. 0.11)/0.72 (S.D. 0.13) for fatigue]. Differences between the separately calculated SDDs for each group over the various test–retest intervals were only 1–3 mm for pain, 0.9–4.6 mm for PGA and 0.1–2.5 mm for fatigue (detailed data not shown). The least differences in SDDs were found in fatigue and in overall PROs for the 28-day test–retest interval. The method of obtaining the data (self-report or telephone call) did not appear to have a strong influence on the 1-day reliability, since differences of <5 mm (corresponding to 5% of the scale range) cannot be considered of significance; thus, we pooled the data for the remaining analyses.

Factors associated with high or low reliability and the SDD

Reliability analyses calculated separately by baseline SDAI tertiles (range: 3–8; >8–17; >17–38) did not reveal any differences that could be used to differentiate between more and less reliable patients concerning their PRO reporting (data not shown). When SDDs were calculated separately for the baseline tertiles of each of the PROs (Table 3), the heterogeneity of differently scored VAS in patients could be reduced, and in the case of PGA, higher baseline values could be shown to be associated with higher SDDs. For pain, SDDs were also lowest in the lower tertile, although no trend was observable across the three tertiles. For fatigue, SDDs differed between the lowest and both other baseline tertiles, with lowest SDDs in patients having a fatigue score of <9 mm and highest SDDs in those with >41 mm. The trend that longer test–retest intervals led to bigger SDDs and smaller ICCs was continued when patients were divided into smaller subgroups (Tables 2 and 3). ICCs

Table 2 Summary statistics of intraclass correlation coefficients of PGA, pain and fatigue

	1-day interval		7-day interval		14-day interval		28-day interval	
By tertiles of	Mean (95% CI)	Median	Mean (95% CI)	Median	Mean (95% CI)	Median	Mean (95% CI)	Median
PGA								
Total	0.83 (0.80, 0.85)	0.86	0.71 (0.67, 0.75)	0.75	0.67 (0.63, 0.72)	0.71	0.57 (0.51, 0.63)	0.58
Lowest	0.72 (0.66, 0.78)	0.78	0.55 (0.47, 0.63)	0.62	0.51 (0.63, 0.59)	0.45	0.49 (0.40, 0.58)	0.47
Middle	0.76 (0.70, 0.82)	0.82	0.55 (0.47, 0.63)	0.60	0.50 (0.40, 0.59)	0.55	0.73 (0.25, 0.49)	0.38
Highest	0.80 (0.76, 0.83)	0.85	0.69 (0.64, 0.75)	0.71	0.64 (0.58, 0.70)	0.66	0.54 (0.45, 0.62)	0.57
Pain								
Total	0.80 (0.78, 0.83)	0.81	0.67 (0.64, 0.71)	0.70	0.58 (0.53, 0.62)	0.61	0.50 (0.44, 0.57)	0.51
Lowest	0.68 (0.63, 0.74)	0.71	0.49 (0.41, 0.57)	0.51	0.36 (0.27, 0.45)	0.38	0.42 (0.32, 0.52)	0.41
Middle	0.73 (0.69, 0.78)	0.78	0.56 (0.50, 0.62)	0.56	0.73 (0.30, 0.45)	0.39	0.26 (0.13, 0.38)	0.27
Highest	0.81 (0.77, 0.84)	0.82	0.69 (0.64, 0.75)	0.71	0.69 (0.64, 0.74)	0.72	0.61 (0.54, 0.68)	0.60
Fatigue								
Total	0.76 (0.73, 0.78)	0.76	0.67 (0.63, 0.71)	0.70	0.65 (0.62, 0.68)	0.69	0.58 (0.54, 0.63)	0.62
Lowest	0.66 (0.59, 0.74)	0.75	0.58 (0.49, 0.66)	0.65	0.53 (0.44, 0.63)	0.59	0.51 (0.38, 0.64)	0.57
Middle	0.63 (0.57, 0.69)	0.63	0.54 (0.47, 0.61)	0.59	0.45 (0.38, 0.52)	0.46	0.30 (0.20, 0.41)	0.35
Highest	0.71 (0.66, 0.76)	0.74	0.55 (0.46, 0.65)	0.67	0.61 (0.54, 0.67)	0.66	0.48 (0.37, 0.59)	0.50

Intraclass correlation coefficient (ICC) is calculated totally for all patients (total) and for patients divided into tertiles of the respective baseline values of the PRO (lowest, middle, highest). Separately calculated for 1-, 7-, 14- and 28-day test–retest intervals. PGA: patient global assessment; PRO: patient-reported outcome.

TABLE 3 Summary of baseline values of PROs and of the smallest detectable differences in PGA pain and fatigue

	Baseline descriptives		Smallest detectable differences							
			1-day interval		7-day interval		14-day interval		28-day interval	
By tertiles of	Mean (95% CI)	Median (range)	Mean (95% CI)	Median	Mean (95% CI)	Median	Mean (95% CI)	Median	Mean (95% CI)	Median
PGA										
Total	19.5 (14.1, 25.0)	18.0 (0–74)	19.0 (18.8, 19.2)	19.1	25.0 (24.7, 25.2)	24.1	26.6 (26.4, 26.9)	25.7	29.9 (29.5, 30.2)	30.1
Lowest	5.6 (5.2, 6.0)	5.0 (0–11)	14.5 (13.9, 15.0)	13.1	19.3 (18.6, 20.1)	17.7	19.9 (19.2, 20.6)	21.1	21.2 (20.3, 22.1)	21.3
Middle	20.4 (20.0, 20.8)	20.0 (15–28)	16.0 (15.4, 16.5)	15.3	23.7 (23.0, 24.5)	23.5	24.5 (23.8, 25.3)	24.2	29.3 (28.4, 30.1)	29.6
Highest	48 (46.4, 49.6)	45.0 (29–74)	22.1 (21.7, 22.6)	21.7	28.0 (27.4, 28.7)	27.3	30.9 (30.1, 31.7)	30.3	34.7 (33.5, 35.9)	38.1
Pain										
Total	18.9 (14, 23.9)	16.5 (0–70)	18.8 (18.6, 19.0)	19.1	24.5 (24.3, 24.7)	24.5	27.8 (27.5, 28.1)	26.9	29.6 (29.3, 29.9)	30.2
Lowest	4.7 (2.5, 6.9)	5.0 (0–10)	15.4 (14.8, 15.9)	13.9	20.8 (20.2, 21.5)	20.5	23.6 (22.9, 24.3)	23.2	24.0 (23.1, 24.8)	26.3
Middle	16.5 (15.0, 18.1)	16.0 (13–20)	19.3 (18.9, 19.8)	17.7	25.7 (25.1, 26.2)	25.0	31.7 (30.9, 32.5)	29.1	34.0 (33.0, 35.0)	32.5
Highest	38.7 (29.2, 48.2)	36.0 (21–70)	18.3 (17.7, 18.8)	18.7	23.3 (22.6, 24.0)	22.9	23.9 (23.2, 24.6)	24.2	26.4 (25.6, 27.2)	28.1
Fatigue										
Total	26.9 (18.9, 34.9)	21.5 (1–100)	25.9 (25.6, 26.1)	25.8	30.2 (29.9, 30.6)	28.3	31.1 (30.8, 31.3)	30.6	33.9 (33.6, 34.3)	33.8
Lowest	4.0 (2.2, 5.9)	3.0 (1–8)	22.3 (21.5, 23.2)	20.7	24.9 (23.9, 25.9)	26.9	26.3 (25.3, 27.4)	27.7	25.7 (24.3, 27.0)	23.3
Middle	25.0 (19.1, 30.9)	24.5 (10–40)	25.1 (24.4, 25.9)	24.6	28.3 (27.4, 29.2)	24.7	30.6 (29.6, 31.2)	29.6	34.9 (33.9, 36.0)	31.5
Highest	54.2 (44.1, 64.4)	49.0 (42–100)	25.4 (24.7, 26.2)	24.5	32.4 (31.2, 33.6)	29.4	30.7 (29.8, 31.6)	29.2	36.2 (34.8, 37.7)	34.8

Smallest detectable differences (in mm) are calculated totally for all patients (total) and for patients divided into tertiles of the respective baseline value of the PRO (lowest, middle, highest). Separately calculated for 1-, 7-, 14- and 28-day test–retest intervals. PGA: patient global assessment; PRO: patient-reported outcome.

were higher for patients with higher PROs, but this does not adequately reflect reliability, because the ranges of scores were bigger in the higher tertiles. Supplementary Table S1, available at *Rheumatology* Online, presents the results obtained when excluding the top 10% with respect to worsening in SDAI and the top 10% with respect to improvement over 2 months. These results were similar to those of the main analyses. SDDs of patients who had a baseline PGA of $\leqslant$10 mm ($n = 12$; remission cut-off point according to the recent ACR/EULAR remission criteria [20]) were significantly smaller than those of other patients: 15 mm (S.D. 8.3) *vs* 20 mm (S.D. 5.1) (using the 1-day test–retest interval); this increased to 22.3 mm (S.D. 9.3) *vs* 32.2 mm (S.D. 7.2) in the 28-day interval (Table 4).

Discussion

This study provides cut-offs for true change in pain and PGA in a representative population of RA patients. Since day-to-day variations in VAS have not to date been explored in depth, we designed this study asking patients to document their pain and PGA levels daily over a period of 56 days. The reliability of pain, PGA and fatigue measurements decreased with longer time intervals, although there was no true change that could actually dilute the measurement and violate reliability assumptions (i.e. no true change between the two assessment times); thus, most variability between the measurements is explainable by within-patient variability and measurement error [28]. Among the three measures assessed here, the PGA is clearly the most relevant for RA disease activity assessment; this is especially so because of its inclusion in RA disease activity composite indices. However, although mostly not directly used for activity assessment, pain and fatigue influence the patient's estimation of disease activity [15, 33], and therefore they were also included in this study. We show here that cut-offs to distinguish true change from noise for measures of PGA, pain and fatigue differ when comparing different time intervals. SDDs for pain and PGA were very similar, but higher SDDs were found for fatigue. The 1-day test–retest cut-off was 19 mm for both pain and PGA. Indeed, reliability seemed

TABLE 4 Summary statistics of smallest detectable differences in PGA calculated separately for patients with baseline PGA ⩽ 10 mm and baseline PGA > 10 mm

	1-day interval		7-day interval		14-day interval		28-day interval	
By baseline PGA	Mean (95% CI)	Median	Mean (95% CI)	Median	Mean (95% CI)	Median	Mean (95% CI)	Median
PGA								
⩽10 mm	15.0 (14.3, 15.6)	13.1	20.3 (19.5, 21.1)	18.5	21.0 (20.2, 21.8)	21.3	22.3 (21.3, 23.2)	22.1
>10 mm	19.8 (19.5, 20.0)	19.8	26.2 (25.8, 26.5)	26.7	28.4 (28.0, 28.7)	28.9	32.1 (31.7, 32.7)	33.1

Smallest detectable differences (in mm) are calculated separately for 1-, 7-, 14- and 28-day test–retest intervals. PGA: patient global assessment.

to progressively decrease from the 1- to the 7-, 14- and 28-day intervals; this was also the case for fatigue, although reliability was somewhat lower. As SDDs do not change much with longer test–retest intervals, a putative threshold value of 25 mm for both pain and PGA could potentially also be valid for even longer intervals of 2–3 months, which represent the typical outpatient visit schedules of RA patients. Fatigue especially seems to show more variability over time, resulting in lower reliability and higher SDDs. Considering that baseline fatigue scores were heterogeneous (ranging from 1 to 100 mm), based on our analysis, an overall SDD threshold of 30 mm seems to be applicable.

A study in RA patients testing a 7-day interval reported 26 mm for PGA and 22 mm for pain as SDD [14]. Test–retest reliability was examined in other studies, providing reliability coefficients ranging between 0.7 and 0.93. Retest evaluation was mostly done a few hours after the initial evaluation [34, 35]. However, a re-evaluation only a few hours later may have a very high recall bias. Lassere *et al*. [13] reported a SDD for pain of 27 mm for a 1-day test–retest interval and 49 mm for a 7-day test–retest interval; the SDD for PGA was 37 mm for a 7-day interval. The ICC for all test–retest intervals and for PGA and pain (which was tested in 24 patients for the 1-day interval and 26 patients for the 7-day interval) was 0.75.

All former studies tested specific intervals, thus reporting point estimates for true change and showing no spread. The strength of our study is that we have multiple evaluations of the same interval, assuming that a 1-day interval between the first and the second day contains the same inherent error as, for example, the interval from the eighth to the ninth day. It can be seen in Table 3 that SDDs of the same time interval show a spread, which we then summarized to one SDD. As a second point, patients were on stable treatment, and no interventions coincided with any study visit, supported by the fact that no change in PROs could be seen (assessed via standardized response means), which is the foundation of a proper evaluation of reliability. Another principle of reliability studies is to use a time interval that is neither too long nor too short. In some ways we have intentionally violated this rule, because we wanted to investigate whether this holds true, and different intervals indeed lead to different results [27]. A limitation in our study could be that there is a dependency in the data, when using day 2 twice for calculating two test–retest intervals, comparing it with day 1 on the one hand and with day 3 on the other. Patients can also get used to daily assessments, and the diary group in particular was not blinded to their previous scores, so that reinforcement over the 56-day period may have taken place [36, 37]. Furthermore, co-morbidities might influence the reliability of or fluctuations in these VAS scores. For example, patients with secondary FM, OA or low back pain experience pain and limitations in daily life, and it can be difficult for the patient to differentiate symptoms from these as opposed to symptoms caused by RA. Although co-morbidities were not formally assessed in our study, their presence must be assumed—at least to some extent—also among our study population [38].

An important aspect in interpreting the ICC is the variation in scores between individual patients. Overall, our patients were rather heterogeneous, resulting in higher ICCs [37], that is, they represent a wide range of individuals with RA. We explored this aspect when we calculated the reliability measures for tertiles of baseline PROs. Lower SDDs were found for patients in the lowest subgroup (VAS at baseline ranging between 0 and 11 mm), compared with higher subgroups. The most interesting subgroup in this respect were patients with a PGA ⩽ 10 mm, since this value constitutes the cut-off point for the Boolean-based ACR–EULAR remission definition [20]. Here, a 13- to 22-mm change set the threshold for true change in PGA. For the other patients, SDDs of 20–33 mm of change are needed, in line with our findings for the total cohort. As the patient global criterion seems to be an important limiting factor for fulfilment of remission criteria [39–42], in particular of the Boolean-based ACR–EULAR remission criteria, the nature of its variability is important. Thus, patients in remission on stable treatment who evaluate their PGA at 2 cm on one clinical visit might further be regarded as being in remission if no other deviation of disease activity is noticeable.

In conclusion, the results of this study suggest that, in stable RA patients, a 25-mm change on the VAS for pain

or PGA and a 30-mm change for fatigue may identify true change; however, this is clearly dependent on (and can be refined based on) the starting measurement level. It is also apparent that in patients who are assessed less frequently, the evaluation of measurement differences as indicating changes is more difficult. After identifying what comprises a true change, in clinical practice, of course, the next important step will be to determine whether the change is clinically relevant. All this in fact speaks for a more global interpretation of disease activity encompassing both patient- and physician-based measures, which in their totality can be a good estimate of a patient's true disease activity.

Acknowledgements

We are indebted to Theresa Kapral and Florian Dernoschnig for their efforts in data acquisition. This is a publication of the Joint and Bone Center for Diagnosis, Research and Therapy of Musculoskeletal Disorders of the Medical University of Vienna.

Funding: This study was supported through Coordination Theme 1 (Health) of the European Community's FP7; Grant Agreement number HEALTH-F2-2008-223404 (Masterswitch).

Disclosure statement: The authors have declared no conflicts of interest.

Supplementary data

Supplementary data are available at *Rheumatology* Online.

References

1 Strand V, Boers M, Idzerda L *et al*. It's good to feel better but it's better to feel good and even better to feel good as soon as possible for as long as possible. Response criteria and the importance of change at OMERACT 10. J Rheumatol 2011;38:1720–7.

2 Kirwan JR, Minnock P, Adebajo A *et al*. Patient perspective: fatigue as a recommended patient centered outcome measure in rheumatoid arthritis. J Rheumatol 2007;34:1174–7.

3 Kirwan JR, Newman S, Tugwell PS *et al*. Progress on incorporating the patient perspective in outcome assessment in rheumatology and the emergence of life impact measures at OMERACT 9. J Rheumatol 2009;36:2071–6.

4 Kirwan JR, Hewlett SE, Heiberg T *et al*. Incorporating the patient perspective into outcome assessment in rheumatoid arthritis—progress at OMERACT 7. J Rheumatol 2005;32:2250–6.

5 Hewlett S, Carr M, Ryan S *et al*. Outcomes generated by patients with rheumatoid arthritis: how important are they? Musculoskeletal Care 2005;3:131–42.

6 Kalyoncu U, Dougados M, Daures JP, Gossec L. Reporting of patient-reported outcomes in recent trials in rheumatoid arthritis: a systematic literature review. Ann Rheum Dis 2009;68:183–90.

7 Tubach F, Ravaud P, Martin-Mola E *et al*. Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: results from a prospective multinational study. Arthritis Care Res 2012;64:1699–707.

8 Stauffer ME, Taylor SD, Watson DJ, Peloso PM, Morrison A. Definition of nonresponse to analgesic treatment of arthritic pain: an analytical literature review of the smallest detectable difference, the minimal detectable change, and the minimal clinically important difference on the pain visual analog scale. Int J Inflamm 2011;2011:231926.

9 Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stats 2007;17:571–82.

10 Bellamy N. Science of assessment. Ann Rheum Dis 2005;64(Suppl 2):ii42–5.

11 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.

12 Doros G, Lew R. Design based on intra-class correlation coefficients. Am J Biostat 2010;1:1–8.

13 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.

14 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.

15 Studenic P, Radner H, Smolen JS, Aletaha D. Discrepancies between patients and physicians in their perceptions of rheumatoid arthritis disease activity. Arthritis Rheum 2012;64:2814–23.

16 Khan NA, Spencer HJ, Abda E *et al*. Determinants of discordance in patients' and physicians' rating of rheumatoid arthritis disease activity. Arthritis Care Res 2012;64:206–14.

17 van Oers ML, Bossema ER, Thoolen BJ *et al*. Variability of fatigue during the day in patients with primary Sjogren's syndrome, systemic lupus erythematosus, and rheumatoid arthritis. Clin Exp Rheumatol 2010;28:715–21.

18 George A, Pope JE. The minimally important difference (MID) for patient-reported outcomes including pain, fatigue, sleep and the health assessment questionnaire disability index (HAQ-DI) in primary Sjogren's syndrome. Clin Exp Rheumatol 2011;29:248–53.

19 Smolen JS, Aletaha D, Bijlsma JW *et al*. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis 2010;69:631–7.

20 Felson DT, Smolen JS, Wells G *et al*. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Ann Rheum Dis 2011;70:404–13.

21 Prince FH, Bykerk VP, Shadick NA *et al*. Sustained rheumatoid arthritis remission is uncommon in clinical practice. Arthritis Res Ther 2012;14:R68.

22 Vermeer M, Kuper HH, van der Bijl AE *et al*. The provisional ACR/EULAR definition of remission in RA: a comment on the patient global assessment criterion. Rheumatology 2012;51:1076–80.

23 Arnett FC, Edworthy SM, Bloch DA *et al*. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.

24 Aletaha D, Neogi T, Silman AJ *et al*. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569–81.

25 Bortz J, Döring N. Forschungsmethoden und Evaluation für Human- und Sozialwissenschafler. 4th edn. Berlin: Springer Medizin Verlag, 2006.

26 Field A. Discovering statistics using SPSS. 3rd edn. London: SAGE Publications Ltd, 2009.

27 Webb N, Shavelson RJ, Haertel EH. Reliability coefficients and generalizability theory. In: Rao CR, Sinharay S, eds. Handbook of statistics. Vol. 26. 1st edn. North Holland: Elsevier, 2006: 81–124.

28 Field A. Intraclass correlation. In: Everitt B, Howell DC, eds. Encyclopedia of behavioral statistics. New York: Wiley, 2005: 948–54.

29 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.

30 Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459–68.

31 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

32 Aletaha D, Martinez-Avila J, Kvien TK, Smolen JS. Definition of treatment response in rheumatoid arthritis based on the simplified and the clinical disease activity index. Ann Rheum Dis 2012;71:1190–6.

33 Khan NA, Spencer HJ, Abda EA *et al*. Patient's global assessment of disease activity and patient's assessment of general health for rheumatoid arthritis activity assessment: are they equivalent? Ann Rheum Dis 2012;71:1942–9.

34 Rohekar G, Pope J. Test–retest reliability of patient global assessment and physician global assessment in rheumatoid arthritis. J Rheumatol 2009;36:2178–82.

35 Clark P, Lavielle P, Martínez H. Learning from pain scales: patient perspective. J Rheumatol 2003;30:1584–8.

36 Matthews JN, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. BMJ 1990;300:230–5.

37 Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th edn. Oxford: Blackwell Publishing, 2002: 704–7.

38 Radner H, Yoshida K, Smolen JS, Solomon DH. Multimorbidity and rheumatic conditions—enhancing the concept of comorbidity. Nat Rev Rheumatol 2014;10:252–6.

39 Curtis JR, Shan Y, Harrold L, Zhang J, Greenberg JD, Reed GW. Patient perspectives on achieving treat-to-target goals: a critical examination of patient-reported outcomes. Arthritis Care Res 2013;65:1707–12.

40 Balogh E, Madruga Dias J, Orr C *et al*. Comparison of remission criteria in a tumour necrosis factor inhibitor treated rheumatoid arthritis longitudinal cohort: patient global health is a confounder. Arthritis Res Ther 2013;15:R221.

41 Masri KR, Shaver TS, Shahouri SH *et al*. Validity and reliability problems with patient global as a component of the ACR/EULAR remission criteria as used in clinical practice. J Rheumatol 2012;39:1139–45.

42 Studenic P, Smolen JS, Aletaha D. Near misses of ACR/EULAR criteria for remission: effects of patient global assessment in Boolean and index-based definitions. Ann Rheum Dis 2012;71:1702–5.