Poor to modest agreement between rheumatoid arthritis response measures in clinical practice

K. Michaud[1,2], T.R. Mikuls[1], S.E. Call[3], A.M. Reimold[4], R. Hooker[4],
G.S. Kerr[5], J.S. Richards[5], L. Caplan[6], G.W. Cannon[3]

[1]VA Medical Center and Nebraska Arthritis Outcomes Research Center, University of Nebraska Medical Center, Omaha, Nebraska; [2]National Data Bank for Rheumatic Diseases, Wichita, Kansas; [3]George E. Wahlen VA Medical Center and University of Utah, Salt Lake City, Utah; [4]VA Medical Center and University of Texas Southwestern Medical Center, Dallas, Texas; [5]VA Medical Center, Washington, DC and Georgetown University Hospital; [6]VA Medical Center and University of Colorado Denver Medical School, Denver, Colorado, USA.

Abstract

Objective

To evaluate the agreement among several rheumatoid arthritis (RA) response measures in a clinical setting.

Methods

529 patients with RA were seen at 2 regular visits where the following response measures were determined: ACR-20, EULAR good or moderate (EULAR-GM), Simplified Disease Activity Index moderate (SDAI-M), Clinical DAI moderate (CDAI-M), and Patient Reported Outcomes Index-M 20 (PRO-IM-20). Each measure was modified to include a "worse" response, i.e. the inverse of the respective guidelines for a positive improvement response.
Introduced for comparison was the Real-time Assessment of Disease Activity in Rheumatoid Arthritis (RADARA), a response measure that registers improvement if the patient's tender and swollen joint counts and HAQ score all improve and worsening if all three increase. Contingency tables comparing the three responses (worse, no change, and improvement) along with Cohen's kappa were calculated.

Results

The mean (SD) baseline characteristics of the patients included: age 66.5 (10.7) years, RA duration 12.9 (11.0) years, 91.3% male, 84.1% rheumatoid factor positive, and a Disease Activity Score-28 of 3.5 (1.3). The percentage of patients who improved/worsened were as follows: ACR-20 4.7/9.1, EULAR-GM 23.4/26.3, SDAI-M 16.1/20.6, CDAI-M 16.3/20.0, PRO-IM-20 22.5/34.4, and RADARA 7.0/11.5. Agreement (kappa) was poor to slight (≤ 0.4) between most of the response measures with the exception of RADARA/ACR-20 which showed substantial agreement (0.67) and SDAI/EULAR-GM and CDAI/EULAR-GM, which showed moderate agreement (0.54 and 0.52, respectively).

Conclusion

RA response measures can be made more informative by the addition of a "worse" response, although even in this case the agreement in the clinic setting is primarily poor to moderate.

Key words

Rheumatoid arthritis, outcome assessment, disease activity, ACR improvement criteria, disease activity score.

Kaleb Michaud, PhD
Ted R. Mikuls, MD, MSPH
Steven E. Call, MD
Andreas M. Reimold, MD
Roderick Hooker, PA, PhD
Gail S. Kerr, MD
John S. Richards, MD
Liron Caplan, MD
Grant W. Cannon, MD

Amgen provided VARA funding to support this study. This work and the VARA project are also supported in part by the Health Services Research and Development Program of the Veterans Health Administration and unrestricted support from Abbott Laboratories and Bristol-Myers Squibb. Dr. Michaud received support through the Arthritis Foundation New Investigator Award.

Please address correspondence to:
Kaleb Michaud, PhD,
986270 Nebraska Medical Center,
Omaha, Nebraska 68198-6270, USA.
E-mail: kmichaud@unmc.edu

Received on October 9, 2008; accepted in revised form on April 24, 2009.

©Copyright Clinical and Experimental Rheumatology 2009.

Competing interests: none declared.

Introduction

The development of effective therapies for RA, including targeted biologic therapies, has produced unprecedented results, particularly when given early in the course of the disease (1, 2). The availability of multiple, effective treatment options emphasizes the need to provide optimal patient-specific therapy in a timely and efficient manner. Objective means of quantifying changes in RA disease activity (including treatment responses and treatment failures) are becoming increasingly critical in everyday clinical practice to guide therapeutic decisions. For example, although up to 50-70% of RA patients treated with approved biologic therapies experience at least 20% improvement based on the American College of Rheumatology criteria (ACR-20), there remain 30-50% of patients who fail to meet this modest threshold of clinical response. Furthermore, non-response is uninformative as to the number of included patients that actually worsen (3-5). In the absence of informative measures of disease activity in clinical practice, it is possible that ineffective therapies are perpetuated despite increased treatment costs and potential risks to patients; likewise, effective therapies may be discontinued due to lack of objective clinical evidence of improvement.

Traditional composite measures of disease activity and treatment response, such as the ACR-20 and the European League Against Rheumatism response (EULAR), have been studied primarily in the context of randomized controlled trials (RCTs). Use of these measures in a clinical practice setting is not common, often due to the lack of necessary data (*i.e.* laboratory measures at the time of the visit) and competing demands on the time of health care providers and clinic personnel (6). This underscores the need for "real time" measures that can be more easily implemented in the clinical setting. Simpler measures of disease activity and responses have been proposed including the Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI), among others (7-9). However, these were primarily developed and validated using major clinical trial databases, potentially limiting their application in the clinical setting.

As a practical exception, the Real-time Assessment of Disease Activity in Rheumatoid Arthritis (RADARA) is a simplified composite response measure that was developed in a clinical practice setting (10). In the present study, using longitudinal clinical data from patients with RA from four US Veterans Affairs (VA) rheumatology practices, we examined the comparability of several composite response measures of disease activity including RADARA, in addition to formal measures traditionally used in RA clinical trials.

Methods and materials

Patient population

Study patients were U.S. veterans over the age of 19 years enrolled in the ongoing Veterans Affairs Rheumatoid Arthritis (VARA) registry. The characteristics of this population have been previously reported (11). In brief, VARA is a multi-center RA registry initiated in 2002 that to date involves active clinical data collection sites at VA Medical Centers in Dallas, Denver, Jackson, Omaha, Salt Lake City, and Washington, DC. With only limited follow-up data from the Denver and Jackson sites available to date, those patients were not included in the present study. All VARA participants are patients fulfilling the American College of Rheumatology (ACR) classification criteria for RA (12).

To be included in this analysis and to calculate changes in measures, the study subjects were required to have had two clinic visits, with the collection of a full ACR core data set at the time of each visit (13). The first visit that met these criteria was designated as the index visit. To ensure that the study was comparable with usual clinical care, and since clinical visits usually take place approximately every 6 months, the second visit was required to have taken place between 3 to 12 months after the index visit and was designated as the follow-up visit. For patients with three or more visits, the first two visits that met these inclusion criteria were used.

To test whether there were important differences between the study patients and those who did not qualify, we used data from the enrollment visit defined as the first time the patients were seen at the clinic, independent of the above-defined index visit.

Study review and approval

Institutional Review Board (IRB) approval has been obtained at each site and all study subjects provided informed written consent prior to their enrollment in VARA. The VARA Scientific and Ethics Advisory Committee also approved this study.

Clinical measures

The VARA registry includes baseline and longitudinal clinical data collected and recorded during the process of routine rheumatologic care. At enrollment, baseline demographic information, serologic studies, radiographs, and clinical data were collected as previously described (11). At baseline and subsequent clinic visits, clinical measures of disease activity were recorded including: tender and swollen joint counts (0-28), erythrocyte sedimentation rate (ESR, mm/hr), C-reactive protein (CRP), the ten-item functional assessment used in the Multidimensional Health Assessment Questionnaire (MDHAQ; range 0-3) (14), patient global health assessment (Patient-Global, 10.0 cm visual analog scale [VAS]), patient pain score (Pain, 0-10 point VAS), and physician global health assessment (Physician-Global, 10.0 cm VAS).

The Disease Activity Score for 28 joints (DAS-28) (15), CDAI, and SDAI (9) were calculated for both the baseline and follow-up visits. These measures were chosen either because they have already been used in clinical trials or because they have been validated and shown to produce results similar to standard outcome measures such as the ACR-20 in re-evaluations of clinical trial data or other clinical datasets.

When comparing disease activity between the index and follow-up visits, several response measures were calculated for each patient, including: ACR-20 (16), EULAR good or moderate (EULAR-GM) (17), SDAI moderate (SDAI-M, DSDAI ≥ 7) (18), CDAI moderate (CDAI-M, DCDAI ≥ 6) (18), Patient Reported Outcomes Index-M 20 (PRO-IM-20) (19) and RADARA (10). Patients were classified as having a RADARA improvement response if all three of the following were observed: a decrease in the tender joint count, a decrease in the swollen joint count, and a decrease in MDHAQ.

Using an inverse calculation for each measure (ACR-20, EULAR-GM, SDAI-M, CDAI-M, PRO-IM-20, and RADARA), a patient was considered "worse" if change corresponding in magnitude to that required for improvement was seen, but in the opposite direction. For example, to meet the criteria for a ACR-20 "worse" response, a patient must show an increase of at least 20% in both the swollen and tender joint counts and to have 3 out of the 5 other sub-measures (ESR, Physician-Global, Patient-Global, MDHAQ, and Pain) worsen by at least 20%. It follows that patients were classified as having a RADARA worse response if they experienced an increase in the tender joint count, an increase in the swollen joint count, and an increase in MDHAQ. If neither an "improvement" nor a "worse" status was detected by means of a change in any of these measures, the patient was classified as having "no change".

Statistical analyses

Demographic variables, including age and disease duration, were calculated at the time of the index clinic visit. Agreement of ACR-20 and EULAR-GM responses with RADARA and the other disease activity response measures was examined by calculating Cohen's kappa coefficient, a measure of inter-rater agreement, using the interpretation criteria proposed by Landis and Koch (kappa < 0 = poor agreement, 0–0.20 = slight agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = substantial agreement, and 0.81–1.00 = near perfect agreement) (20). In secondary analyses, we examined the agreement of these measures in patients with low disease activity/remission (DAS-28 <2.6) and in patients with higher levels of disease activity (DAS-28 ≥2.6) at their index visits. All analyses used two-tailed *p*-values and were performed using Stata release 10 (StataCorp, College Station, TX).

Results

As of September 2008, there were 1146 patients with RA enrolled in VARA, of whom 529 qualified for the present analysis. Of the 617 patients excluded, 248 did not have two clinic visits within a 3-12 month period, and the remainder (n=369) had at least one missing ACR core variable. There were no significant differences in age (66.5 *vs.* 66.3 years, *p*=0.80), disease duration (10.7 *vs.* 11.8 years, *p*=0.09), or tender (6.1 *vs.* 6.4, *p*=0.45) and swollen (5.7 *vs.* 5.5, *p*=0.62) joint counts at the enrollment visit between those enrolled and those excluded from the analysis.

The participants' characteristics are summarized in Table I. The average age was 67 years with 91% being male, most (84%) were seropositive for rheumatoid factor, and almost half (41%) had rheumatoid nodules. The mean disease duration was 13 years. The mean baseline DAS28, CDAI and SDAI scores were 3.5, 13.2 and 14.3, respectively, suggestive of overall moderate disease activity. The mean tender joint count was 3.3 and the swollen joint count was 3.0. The average time between visits was almost 5 months. Baseline RA treatments included methotrexate (60%), prednisone (44%), hydroxychloroquine (33%) and TNF-inhibition (39%).

Only 61 patients (12%) had a RADARA worse response, while 37 (7%) showed a RADARA improvement response. At the index visit, the primary difference between these patients and the 'no change' group was that RADARA improvement patients uniformly showed more severe disease activity and RADARA worse patients had less severe joint counts (Table I).

Comparisons of ACR-20 and EULAR-GM responses with other composite disease activity response measures are summarized in Table II. All kappa coefficient values were statistically significant (*p*<0.001). When compared to ACR-20, RADARA demonstrated concordance in 486 (92%) patients. In a similar comparison to ACR-20, the

Table I. Characteristics (% or mean [SD]) of patients with RA at index visit and by RADARA response.

Variables	All (n=529)	Worse (n=61)	Improvement (n=37)
Demographics			
Age (years)	66.5 (10.7)	65.6 (10.4)	65.3 (12.0)
Male sex (%)	91.3	90.2	83.8
High school education (%)	64.1	62.3	64.9
Caucasian (%)	80.2	85.3	70.3
ACR core measures			
Swollen joints (0-28)	3.0 (4.3)	1.7 (2.2)	8.4 (7.0)
Tender joints (0-28)	3.3 (5.4)	2.2 (3.7)	8.5 (7.4)
ESR (mm/hr)	23.6 (19.8)	24.2 (20.7)	26.5 (22.8)
MDHAQ (0-3)	0.88 (0.59)	0.97 (0.49)	1.16 (0.56)
Pain (0-10)	4.1 (2.8)	4.4 (2.5)	5.9 (2.3)
Patient global (0-10)	3.8 (2.5)	4.2 (2.5)	5.6 (2.9)
Physician global (0-10)	3.2 (2.2)	3.1 (1.8)	4.8 (2.4)
Other clinical measures			
Disease duration (years)	12.9 (11.0)	12.6 (11.1)	11.2 (9.5)
Time between visits (days)	144 (48)	144 (53)	139 (39)
DAS28	3.5 (1.3)	3.4 (1.0)	5.0 (1.1)
CDAI	13.2 (11.0)	11.2 (6.9)	27.3 (15.2)
SDAI	14.3 (11.2)	12.1 (7.1)	29.0 (15.2)
Rheumatoid nodules (%)	41.2	42.6	32.4
RF positive (%)	84.1	86.9	81.1
CRP	1.06 (1.46)	0.87 (0.91)	1.63 (2.40)
Medications			
Methotrexate (%)	60.0	55.0	62.9
Prednisone (%)	43.9	46.7	45.7
Hydroxychloroquine (%)	32.6	41.7	22.9
Leflunomide (%)	14.6	16.7	11.4
Sulfasalazine (%)	13.6	11.7	17.1
Anti-TNF (%)	38.5	33.3	42.9

ESR: erythrocyte sedimentation rate; DAS: Disease Activity Score; MDHAQ: modified Health Assessment Questionnaire; CDAI: Clinical Disease Activity Index; SDAI: Simplified Disease Activity Index; RF: rheumatoid factor; CRP: C-reactive protein; TNF: tumor necrosis factor.

SDAI-M response showed concordance in 395 (75%) patients, the CDAI-M response in 388 (73%) patients, the EULAR-GM response in 321 (61%) patients, and the PRO-IM-20 in 283 (53%) patients. When compared to EULAR-GM, RADARA was concordant in 326 (62%) patients. Similarly, SDAI-M showed concordance with EULAR-GM in 387 (73%) patients, CDAI-M in 379 (72%) patients, and PRO-IM-20 in 237 (45%) patients. Of the measures examined, RADARA responses aligned best with ACR-20 responses, with a corresponding kappa = 0.67 (substantial agreement). In contrast, RADARA responses showed only fair agreement with EULAR-GM responses (kappa = 0.29). Of the measures examined in comparison to EULAR-GM response, the SDAI-M responses showed the strongest agreement (kappa = 0.54, moderate agreement).

The majority of instances where discordance was found between different response measures were cases in which one measure reported ‘no change’ and the other reported either improvement or worsening. However, there were a small number of subjects with improvement by one measure and concomitant worsening by another (Table II). These major discrepancies were rare and only found in comparisons with the EULAR-GM response; no such discrepancies were observed in comparisons with ACR-20.

When we analyzed one case in which EULAR-GM was worse while the CDAI-M and SDAI-M showed improvement, we discovered that the worsening EULAR-GM score was the result of an abrupt rise in the ESR that led to a marked increase in the DAS28 score notwithstanding improvement in the tender and swollen joint counts (which led to an improvement in the CDAI-M and SDAI-M).

In a similar comparison of EULAR-GM with PRO-IM-20, 30 subjects with a marked discordance in response were noted. Of the 19 patients whose PRO-IM-20 score worsened while their EULAR-GM improved, all experienced worsening in two of the three patient measures (HAQ, patient global, and pain score) with simultaneous improvement in the joint counts that accounted for the improvement in EULAR-GM. Only 2 had a worse ESR with improvement in EULAR-GM. Of the 11 patients with PRO-IM-20 improvement and a worsened EULAR-GM, most exhibited worsening joint tender and swelling counts, but simultaneous improvement in 2 out of 3 patient-reported measures. One patient showed no change in the joint tenderness count and a small increase in the joint swelling count from 3 to 4, but the change in ESR from 1 to 8 accounted for the worsening in the EULAR-GM response.

Comparison of the ACR-20 and EULAR-GM responses to RADARA and the other response measures was also performed in sub-sets of the patient cohort stratified by disease activity (Table III, a and b). Compared to ACR-20, RADARA demonstrated concordance in 142 (96%) patients who had been in DAS28 remission and in 338 (89%) patients with active disease. In the comparison with EULAR-GM, RADARA was concordant in 107 (72%) patients in remission and 219 (57%) patients with active disease. Similar rates of major discrepancies were seen in both the remission and active disease subsets when the ACR-20 and EULAR-GM responses were compared to SDAI-M, CDAI-M and PRO-IM-20 (Table III).

Discussion

Composite clinical response measures such as the ACR-20 have been used as the standard outcome measure in clinical trials for over a decade, but they are only rarely applied in the clinic setting. The primary reason for this may be obvious; the standard dichotomous response measure provides relatively little information on the individual patient, especially in cases of non-re-

Table II. Frequency tables comparing ACR 20 and EULAR Good/Moderate responses with other real-time response measures (n=529)*.

		ACR 20				EULAR Good/Moderate		
		Worse	No change	Improvement		Worse	No change	Improvement
RADARA	Worse	42	19	0	Worse	50	11	0
	No change	6	419	6	No change	89	247	95
	Improvement	0	18	19	Improvement	0	8	29
			κ=0.67				κ=0.29	
SDAI Mod.	Worse	39	70	0	Worse	89	20	0
	No change	9	324	2	No change	49	230	56
	Improvement	0	62	23	Improvement	1	16	68
			κ=0.37				κ=0.54	
CDAI Mod.	Worse	39	67	0	Worse	86	20	0
	No change	9	326	2	No change	52	227	58
	Improvement	0	63	23	Improvement	1	19	66
			κ=0.37				κ=0.52	
PRO-IM 20	Worse	43	139	0	Worse	73	90	19
	No change	5	219	4	No change	55	116	57
	Improvement	0	98	21	Improvement	11	60	48
			κ=0.21				κ=0.14	
ACR 20					Worse	42	6	0
					No change	97	257	102
					Improvement	0	3	22
							κ=0.26	

*For all κ, $p<0.001$.
ACR: American College of Rheumatology; EULAR: European League Against Rheumatism; RADARA: Real-time Assessment of Disease Activity in Rheumatoid Arthritis; SDAI: Simplified Disease Activity Index; CDAI: Clinical Disease Activity Index; PRO-IM: Patient Reported Outcomes Index (majority response).

sponse. Responses that are governed by percentage changes have an even lower informative power; without taking into account the baseline value, the change that triggers a response can range from minimal to very large. Clearly, a physician should not make a clinical decision based on these responses alone. However, a robust clinical response measure could offer the physician a useful framework within which to judge the progression of a patient's disease and possibly provide a quality measure to justify the addition or interruption of a specific therapy (21).

One such simple response measure is the RADARA, which we developed and are using in some of our rheumatology clinics, where it has shown moderate agreement with EULAR-GM responses and high agreement with ACR-20 responses. This measure was developed in part because of the long delays may occur when acute phase reactant (ESR and CRP) test results are not routinely available at the time of the clinic visit but are needed to determine the full DAS28 score. Since RADARA is a relatively simple measure that can be calculated immediately, it could potentially provide an indication in 'real-time' of a general worsening or improvement in the patient's condition in comparison to their previous visit. We observed substantial agreement between the RADARA and ACR-20 responses, as would be expected since change in both measures is based on joint counts and both have a HAQ component. There was only fair agreement between RADARA and EULAR-GM (SDAI-M showing the highest – albeit modest – level of agreement with EULAR-GM). The lack of concordance between RADARA and EULAR-GM highlights the different weights assigned to ESR and Physician-Global in the two measures; the similarity in the use made of Physician-Global probably explains why the SDAI-M agrees so well with EULAR-GM.

The criteria for the enrollment of patients in RCTs are strict; Sokka and Pincus estimated that only 5% of their community RA patients would have been eligible for the ATTRACT study (22). Moreover, the 'gold-standard' RA response measures in all recent RCTs are not used nor are they recommended for use in regular clinical practice. We found that very few patients under routine clinical care with, on average, medium disease activity (DAS28: 3.5) experienced changes in disease activity corresponding to an ACR-20 improvement or an ACR-20 worse response (4.7% and 9.1%, respectively) (Table II). Surprisingly only 1 of the 25 patients with improvement was in DAS28-defined remission at the first visit (Table IIIa); this is surprising because patients starting with low levels of disease activity would require only small changes to meet the 20% threshold. In contrast to the ACR-20, there were a large number of patients with changes in disease activity corresponding to a EULAR-GM improvement response (27.2%). The EULAR-GM response requires a change in DAS that falls within one of three pre-defined ranges (≤3.2, >3.2 & ≤5.1, or >5.1), and in our study the mean index DAS28 was 3.5, which is very close to the 3.2 threshold. Thus, on average patients needed to show relatively small changes in disease activity to meet the EULAR-GM

Table III. A. Frequency tables comparing ACR-20 and EULAR Good/Moderate responses with other real-time response measures for patients meeting remission criteria (DAS28 <2.6) at the index study visit (n=148).

		ACR 20				EULAR Good/Moderate		
		Worse	No change	Improvement		Worse	No change	Improvement
RADARA	Worse	12	2	0	Worse	10	4	0
	No change	4	129	0	No change	26	97	10
	Improvement	0	0	1	Improvement	0	1	0
			κ=0.79 p<0.001				κ= -0.04 *p*=0.142	
SDAI Mod.	Worse	12	17	0	Worse	22	7	0
	No change	4	110	0	No change	14	91	9
	Improvement	0	4	1	Improvement	0	4	1
			κ=0.43 p<0.001				κ= -0.04 *p*=0.149	
CDAI Mod.	Worse	12	17	0	Worse	21	8	0
	No change	4	111	0	No change	15	91	9
	Improvement	0	3	1	Improvement	0	3	1
			κ=0.44 p<0.001				κ= -0.04 *p*=0.161	
PRO-IM 20	Worse	14	48	0	Worse	20	38	4
	No change	2	62	0	No change	13	46	5
	Improvement	0	21	1	Improvement	3	18	1
			κ=0.16 p<0.001				κ= -0.01 *p*=0.337	
ACR 20					Worse	13	3	0
					No change	23	98	10
					Improvement	0	1	0
							κ= -0.05 *p*=0.125	

B. Frequency tables comparing ACR-20 and EULAR Good/Moderate responses with other real-time response measures for patients not meeting remission criteria (DAS28 ≥2.6) at the index study visit (n=381).

		ACR 20				EULAR Good/Moderate		
		Worse	No change	Improvement		Worse	No change	Improvement
RADARA	Worse	30	17	0	Worse	40	7	0
	No change	2	290	6	No change	63	150	85
	Improvement	0	18	18	Improvement	0	7	29
			κ=0.64 p<0.001				κ=0.08 *p*=0.007	
SDAI Mod.	Worse	27	53	0	Worse	67	13	0
	No change	5	214	2	No change	35	139	47
	Improvement	0	58	22	Improvement	1	12	67
			κ=0.35 p<0.001				κ=0.18 *p*<0.001	
CDAI Mod.	Worse	27	50	0	Worse	65	12	0
	No change	5	215	2	No change	37	136	49
	Improvement	0	60	22	Improvement	1	16	65
			κ=0.35 p<0.001				κ=0.17 *p*<0.001	
PRO-IM 20	Worse	29	91	0	Worse	53	52	15
	No change	3	157	4	No change	42	70	52
	Improvement	0	77	20	Improvement	8	42	47
			κ=0.22 p<0.001				κ=0.01 *p*=0.363	
ACR 20					Worse	29	3	0
					No change	74	159	92
					Improvement	0	2	22
							κ=0.05 *p*=0.057	

ACR: American College of Rheumatology; EULAR: European League Against Rheumatism; RADARA: Real-time Assessment of Disease Activity in Rheumatoid Arthritis; SDAI: Simplified Disease Activity Index; CDAI: Clinical Disease Activity Index; PRO-IM: Patient Reported Outcomes Index (majority response).

criterion for improvement. The 124 patients who showed EULAR-GM improvement had a DAS28 index of 4.4 and ΔDAS -1.4, whereas 116 other patients also showed an improvement in DAS (mean ΔDAS -0.3), but these did not cross any pre-defined states.

Three other, less well-known response measures were also used for comparison in this study: CDAI-M, SDAI-M and PRO-IM-20. Like the EULAR-GM, these response measures have threshold values for change, but they differ from the EULAR-GM in that their criteria for improvement (and worsening in this study) are not dependent on the index level of disease activity (with the caveat that those with a minimum score cannot improve). The CDAI and SDAI are relatively new measures that have achieved popularity due to their simplicity (summing of values), their reliance on physician-based measures, and their perceived

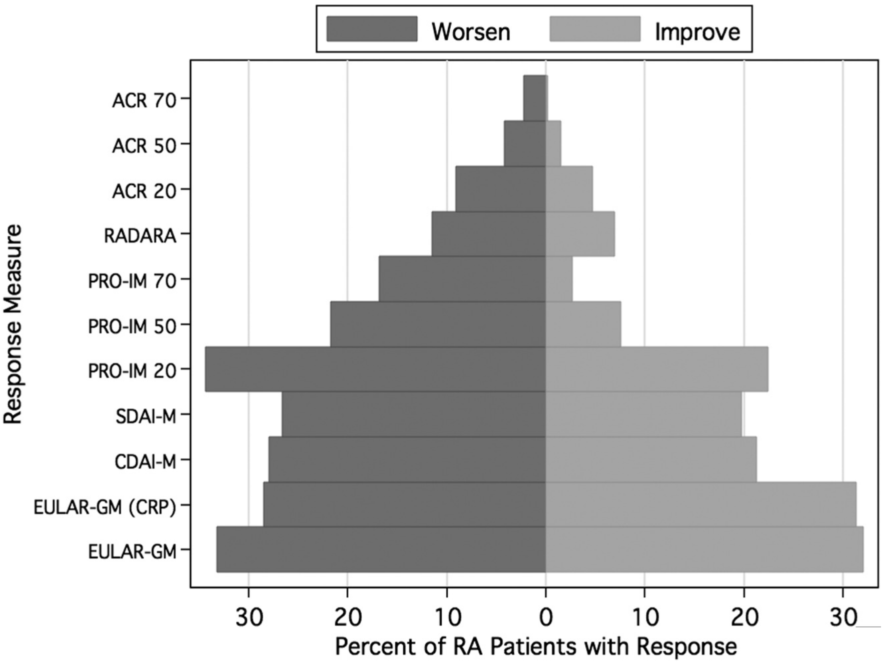

Fig. 1. Horizontal bar graph showing the percentage of the 529 RA patients studied who showed a worse response or improvement in several measures at their second clinic visit.
ACR: American College of Rheumatology; RADARA: Real-time Assessment of Disease Activity in Rheumatoid Arthritis; PRO-IM: Patient Reported Outcomes Index (majority response); SDAI-M: Simplified Disease Activity Index moderate; CDAI-M: Clinical Disease Activity Index moderate; EULAR-GM: European League Against Rheumatism Good/Moderate.

usefulness in the clinic (23). Because they had scores below the minimum response threshold at their index visit, it was impossible for 148 (137) of our patients to demonstrate CDAI-M (SDAI-M) improvement even when 27% (26%) of them showed improvement in their CDAI (SDAI) scores. The PRO-IM-20 is the newest, and the only patient-exclusive response measure analyzed in this study. Since PRO-IM-20 does not include any joint counts or laboratory measures, it is not surprising that it showed the lowest concordance with ACR-20 and EULAR-GM. However, this discordance contrasts with the level of agreement described between the ACR-20 and PRO-IM-20 in previous clinical trials (19).

Another limitation to the use of these response measures is that most (with the exception of ACR-20 and RADARA) allow for an improvement response even when the swollen joint count, tender joint count or HAQ increases (worsens). For the EULAR-GM response, this arises from the fact that the DAS formula has four competing components. Depending on the initial joint count (1 vs. 28), a decrease in one tender joint (keeping swollen joints, ESR and Patient-Global unchanged) could decrease the DAS28 by 0.56 (1 tender joint) vs. 0.10 (28 tender joints). In the first example (1 to 0 tender joints), it would take an increase in at least 4 (0 to 4) and up to 18 (10 to 28) swollen joints to prevent the change in DAS28 from decreasing. While not based on clinical data, this example illustrates how the relative weighting of each DAS component limits how the DAS28 can change. This phenomenon occurred in 46% of our patients; i.e., the EULAR-GM improved, but one or two of the swollen or tender joint counts or the HAQ score worsened. For example, one patient experienced a change in DAS28 from 4.09 to 3.47, giving him a EULAR-GM improvement response, even though his swollen joints increased from 4 to 9 and his MDHAQ increased from 0.9 to 1.1.

One concern relating to RCTs is that they only report the results of response measures where patients show improvement; we do not know of any that have reported the number of patients who actually worsened. It can be argued that knowing how much the condition of non-responders in an RCT improves or worsens could help in deciding which therapy to initiate. For example, suppose the trial for drug A has 70% ACR-20 improvement and 20% ACR-20 worsening and the trial for drug B has 60% ACR-20 improvement and 0% ACR-20 worsening. Without knowing the frequency of worsening, as is current practice, drug A would likely appear to be more efficacious. The inclusion of worsening rates or the weighting of patients' concerns about a decline in functional status with a new therapy (24) might tip the balance in favor of drug B.

While the patient cohort in our study represents a unique group that differs from the majority of RA patients due to its older age and greater percentage of men, the results are directly applicable to the US VA health care system, which has nearly 8 million enrollees and, with 24 million veterans and 37 million dependents, will see dramatic growth in the near future (25). The VA is the largest integrated health system in the US, with an extensive network of databases that has contributed to advances in health outcomes research not only in the VA, but also in the general population. Our results showing the lack of comparability between different clinical measures of RA and our development of RADARA has immediate applications in the VA population and could be applied in future studies to the general RA population where women predominate. The modification of these measures/this measure to include symmetric worsening could make them/it applicable in any RA cohort.

We have proposed a novel and intuitive response measure, RADARA, which is based on the inclusion of worsening responses. While RADARA was created with this in mind, the number of patients with a change in disease activity corresponding to a 'worse response' in the average clinic visit was not expected (Fig. 1); all measures showed a slightly greater percentage of patients worsening rather than improving, which is consistent with the natural course of this disease. On average, patients who worsened were similar to those who improved in terms of index sociodemographic parameters, but they differed based on RA disease characteristics, with those worsening/improving tending to show regression towards the mean with better/worse index values, respectively. The results of these analyses suggest that a proportionate number of patients worsen over the follow-up period whether they begin in a low disease activity state or a higher disease activity state.

While we address the limitations of using a response measure in clinical practice, this should not be interpreted as promoting the alternative use of composite disease activity scores and their pre-defined states (26), as these have many limitations as well (27). Instead, we advocate being aware of the limitations when using any measures to guide clinical practice, and suggest that there is a benefit to allowing for worsening when adopting a response measure.

In summary, our observations from clinical practice demonstrate that changes in disease activity and the associated measures of clinical response can be interpreted differently depending on the composite outcome measure employed. Discrepancies between the conclusions of different response measures are a reflection of the criteria employed. We propose that future studies using these response measures should also report the number of patients that worsen, using the same criteria in the inverse sense. We present RADARA as an example of how to incorporate this worsening in a response measure and as a useful and simple tool for clinical practice. Further studies will be needed to define composite outcome measures that can be employed to direct treatment decisions during clinical practice.

Acknowledgements

The authors would like to acknowledge Dannette Johnson, DO, of the VA Medical Center in Jackson, Mississippi for her contributions to VARA and for proofreading this manuscript.

References

1. SMOLEN JS, VAN DER HEIJDE DM, ST CLAIR EW *et al.*: Predictors of joint damage in patients with early rheumatoid arthritis treated with high-dose methotrexate with or without concomitant infliximab: results from the ASPIRE trial. *Arthritis Rheum* 2006; 54: 702-10.
2. GENOVESE MC, BATHON JM, MARTIN RW *et al.*: Etanercept versus methotrexate in patients with early rheumatoid arthritis: two-year radiographic and clinical outcomes. *Arthritis Rheum* 2002; 46: 1443-50.
3. WEINBLATT ME, KEYSTONE EC, FURST DE *et al.*: Adalimumab, a fully human anti-tumor necrosis factor alpha monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial. *Arthritis Rheum* 2003; 48: 35-45.
4. MAINI R, ST CLAIR EW, BREEDVELD F *et al.*: Infliximab (chimeric anti-tumour necrosis factor alpha monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a randomised phase III trial. ATTRACT Study Group. *Lancet* 1999; 354: 1932-9.
5. MORELAND LW, BAUMGARTNER SW, SCHIFF MH *et al.*: Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein. *N Engl J Med* 1997; 337: 141-7.
6. ALETAHA D, NELL VP, STAMM T *et al.*: Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. *Arthritis Res Ther* 2005; 7: R796-806.
7. SMOLEN JS, BREEDVELD FC, SCHIFF MH *et al.*: A simplified disease activity index for rheumatoid arthritis for use in clinical practice. *Rheumatology* 2003; 42: 244-57.
8. ALETAHA D, SMOLEN J: The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. *Clin Exp Rheumatol* 2005; 23 (Suppl. 39): S100-108.
9. MAKINEN H, KAUTIAINEN H, HANNONEN P *et al.*: Disease activity score 28 as an instrument to measure disease activity in patients with early rheumatoid arthritis. *J Rheumatol* 2007; 34: 1987-91.
10. CALL SE, CIPHER D, HOOKER R *et al.*: Real-time assessment of disease activity in rheumatoid arthritis (RADARA): use of threshold criteria improves specificity of composite clinical outcome measures [abstract]. *Arthritis Rheum* 2007; 56 (Suppl.): S706.
11. MIKULS TR, KAZI S, CIPHER D *et al.*: The association of race and ethnicity with disease expression in male US veterans with rheumatoid arthritis. *J Rheumatol* 2007; 34: 1480-4.
12. 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.
13. FELSON DT, ANDERSON JJ, BOERS M *et al.*: The American College of Rheumatology preliminary core 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.
14. PINCUS T, SWEARINGEN C, WOLFE F: Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. *Arthritis Rheum* 1999; 42: 2220-30.
15. 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.
16. FELSON DT, ANDERSON JJ, BOERS M *et al.*: American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995; 38: 727-35.
17. FRANSEN J, VAN RIEL PL: The Disease Activity Score and the EULAR response criteria. *Clin Exp Rheumatol* 2005; 23 (Suppl. 39): S93-99.
18. ALETAHA D, SMOLEN JS, KVIEN TK: Definition of moderate and major response for disease activity indices in rheumatoid arthritis (RA). *Ann Rheum Dis* 2006; 65 (Suppl. II): 692.
19. PINCUS T, CHUNG C, SEGURADO OG, AMARA I, KOCH GG: An index of patient reported outcomes (PRO-Index) discriminates effectively between active and control treatment in 4 clinical trials of adalimumab in rheumatoid arthritis. *J Rheumatol* 2006; 33: 2146-52.
20. LANDIS JR, KOCH GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-74.
21. SOLOMON DH, GABRIEL SE: Quality Measures 101: what every rheumatologist should know. *Clin Exp Rheumatol* 2007; 25 (Suppl. 47): 18-21.
22. SOKKA T, PINCUS T: Eligibility of patients in routine care for major clinical trials of anti-tumor necrosis factor alpha agents in rheumatoid arthritis. *Arthritis Rheum* 2003; 48: 313-8.
23. ALETAHA D, SMOLEN JS: The Simplified Disease Activity Index (SDAI) and Clinical Disease Activity Index (CDAI) to monitor patients in standard clinical care. *Best Pract Res Clin Rheumatol* 2007; 21: 663-75.
24. WOLFE F, MICHAUD K: Resistance of rheumatoid arthritis patients to changing therapy: Discordance between disease activity and patients' treatment choices. *Arthritis Rheum* 2007; 56: 2135-42.
25. National Center for Veterans Analysis and Statistics (http://www.va.gov/vetdata/ accessed Oct 2, 2008).
26. ALETAHA D, FUNOVITS J, SMOLEN JS: The importance of reporting disease activity states in rheumatoid arthritis clinical trials. *Arthritis Rheum* 2008; 58: 2622-31.
27. WOLFE F, MICHAUD K: The challenges of determining RA disease activity and remission in clinical practice. *Nature Clinical Practice* 2008; 4: 462-3.