Responsiveness of the core set, response criteria, and utilities in early rheumatoid arthritis

A C Verhoeven, M Boers, S van der Linden

Abstract

***Objective*—Validation of responsiveness and discriminative power of the World Health Organisation/International League of Associations for Rheumatology (WHO/ILAR) core set, the American College of Rheumatology (ACR), and European League for Rheumatology (EULAR) criteria for improvement/response, and other single and combined measures (indices) in a trial in patients with early rheumatoid arthritis (RA).**

***Methods*—Ranking of measures by response (standardised response means and effect sizes) and between-group discrimination (unpaired *t* test and χ^2 values) at two time points in the COBRA study. This study included 155 patients with early RA randomly allocated to two treatment groups with distinct levels of expected response: combined treatment, high response; sulfasalazine treatment, moderate response.**

***Results*—At week 16, standardised response means of core set measures ranged between 0.8 and 3.5 for combined treatment and between 0.4 and 1.2 for sulfasalazine treatment (95% confidence interval ±0.25). Performance of patient oriented measures (for example, pain, global assessment) was best when the questions were focused on the disease. The most responsive single measure was the patient's assessment of change in disease activity, at 3.5. Patient utility, a generic health status measure, was moderately (rating scale) to poorly (standard gamble) responsive. Response means of most indices (combined measures) exceeded 2.0, the simple count of core set measures improved by 20% was most responsive at 4.1. Discrimination performance yielded similar but not identical results: best discrimination between treatment groups was achieved by the EULAR response and ACR improvement criteria (at 20% and other percentage levels), the pooled index, and the disease activity score (DAS), but also by the Health Assessment Questionnaire (HAQ) and grip strength.**

***Conclusions*—Responsiveness and discrimination between levels of response are not identical concepts, and need separate study. The WHO/ILAR core set comprises responsive measures that discriminate well between different levels of response in early RA. However, the performance of patient oriented measures is highly dependent on their format. The excellent performance of indices such as the ACR improvement and EULAR response criteria confirms that they are the preferred primary end point in RA clinical trials.**

(*Ann Rheum Dis* 2000;**59**:966–974)

Department of Rheumatology/Internal Medicine, University Hospital Maastricht, The Netherlands
A C Verhoeven
S van der Linden

Department of Clinical Epidemiology, VU University Hospital, Amsterdam, The Netherlands
M Boers

Correspondence to:
Professor M Boers, Department of Clinical Epidemiology VE9–78, VU University Hospital, PO Box 7057, 1007 MB Amsterdam, The Netherlands
m.boers@azvu.nl

Accepted for publication
26 April 2000

Many end point measures are available to assess treatment efficacy in rheumatoid arthritis (RA). The OMERACT consensus on end point measures in RA facilitates comparison of results from different trials in treatment evaluation.[1] The OMERACT initiative has called for further validation of the measures included in the World Health Organisation/International League of Associations for Rheumatology (WHO/ILAR) core set (also known as American College of Rheumatology (ACR) core set[2]) and combined measures (indices) such as improvement and response criteria.[3 4] To determine the applicability of a measure in a certain setting, the OMERACT filter has been proposed, containing three elements: truth, discrimination, and feasibility.[5] The first two elements capture classic validity concepts. The topic of this study is discrimination. To be discriminative in trials, a measure has to detect clinically relevant change; moreover, it has to detect clinically relevant differences in change between treatment groups. Highly responsive measures are preferred because they allow clinical trials to be done with fewer patients, and also because they facilitate detection of small—but potentially important—differences in treatment effect.[6] For the clinician, applying responsive measures in patient care will allow better tailoring of individual treatment. However, individual patient care may require another selection of measures than those used in trials: for example, assessment of morning stiffness and disease activity in the feet joints remains useful in the clinic despite their exclusion from the core set.[7]

As every measure is bound to pick up some noise together with the intended signal, its responsiveness is determined by the ratio of treatment effect to its variability (signal-to-noise ratio). Two classes of responsiveness statistics can be distinguished: the first is based on measurement of change in the course of a therapeutic intervention with known efficacy (external criterion, gold standard); the second is directed at the correlation of change in the tested measure with change in a "criterion measure". However, as this last class of responsiveness statistics is based on the variability between subjects, regardless of whether group changes occur, they yield little information about the ability of a measure to detect treatment effects.[8]

The purpose of this study was to validate further the responsiveness of the core set, of the

ACR improvement[3] and European League for Rheumatology (EULAR) response criteria,[4] and of other single measures and indices with data from a recent trial. The COBRA study[9] (Dutch acronym: COmbinatietherapie Bij Reumatoïde Artritis) was a randomised controlled trial in patients with early RA that showed excellent clinical response, low toxicity, and less progression of radiographic joint damage with treatment of combined step-down prednisolone, methotrexate, and sulfasalazine, compared with treatment with sulfasalazine alone. The trial allowed us to create one "gold" and one "silver" standard for relevant response against which to validate performance of end point measures: we proposed the hypothesis a priori that patients in the combination group would on average show large and, certainly, relevant improvements at week 16 owing to the corticosteroid pulse (gold standard). Also, we assumed on the basis of the well known effects of sulfasalazine on disease activity that patients in the sulfasalazine-only group would also show relevant improvements, but to a lesser degree (silver standard).

We then ranked the end point measures and indices used in the COBRA study by their relative responsiveness, and also by their ability to discriminate between the (changes in the) treatment groups. Ultimately, this discriminatory ranking yields the most relevant results. However, one must realise that this ranking is post hoc: as the difference in response between the groups was the primary study question of the trial, the presence and extent of such a difference was unknown before the start of the trial.

Patients and methods

THE COBRA STUDY

The COBRA study[9] was a 56 week clinical trial that randomly assigned 155 patients with RA (ACR criteria[10]), aged 23–70, to one of two treatments. All patients had early, active disease (diagnosis <2 years). No prior treatment with disease modifying antirheumatic drugs, apart from antimalarial drugs, was allowed. One group was treated with a combination of sulfasalazine, methotrexate, and, initially, high dose oral prednisolone, the other group with sulfasalazine and double placebo. The prednisolone dose was 60 mg daily in the first week, tapered in weekly steps to the maintenance dose of 7.5 mg in week 7. Prednisolone and methotrexate (or the placebos) were tapered and stopped after weeks 28 and 40, respectively, while sulfasalazine was continued.

CORE SET MEASURES

A broad variety of end points was assessed, including all disease activity measures of the WHO/ILAR core set.[1] This comprises tender and swollen joint count (68 and 48 joints, respectively[11]), pain, assessor's and patient's global assessment (on a 10 cm visual analogue scales (VAS)), acute phase reactant (that is, erythrocyte sedimentation rate, Westergren method (ESR) or C reactive protein (CRP)), and physical function (by Health Assessment Questionnaire; Dutch HAQ[12 13]).

NON-CORE SET MEASURES

Non-core set measures included other joint counts and scores such as the Ritchie index, grip strength (by vigorimetry; Martin, Tottlingen, Germany, mean of medians of three measurements in both hands[14]), Arthritis Impact Measurement Scale (AIMS)[15]—a modified and validated Dutch version with scales for mobility, pain, and self efficacy, and the McMaster Toronto Arthritis patient preference questionnaire (MACTAR).[16] The MACTAR is an instrument that follows improvement in five impaired activities, elicited and ranked in priority by the patient at baseline, together with changes in quality of life, psychological, social, and emotional wellbeing. Its scores increase as functional ability improves and vary from 11 (maximum deterioration) to 47 (maximum improvement). In its original format the baseline scores differ from the follow up scores because items inquiring about change are not included. To make these scores directly comparable mock change items were added at baseline and scored as "unchanged". To compare the responsiveness and discriminatory power of different formats of patient global assessment (see Appendix), two items from the MACTAR interview (change in disease activity by seven point Likert scale, and physical function by two questions with a six point scale), and a question on the actual state of disease activity (from a monitoring questionnaire; five point Likert scale) were evaluated together with the patient's global assessment of health indicated on a 10 cm VAS.

GENERIC MEASURES

Whereas these disease-specific measures are sensitive to clinical change in RA, other—generic—measures yield a broader picture of patients' health status and allow comparison across a range of conditions.[17] Utility served as the central concept of generic measures in the COBRA study.[18] Utility is a single value or preference that patients assign to a particular health state. This value is expressed on a scale ranging from 1 (perfect health) to 0 (death) and takes into account both the positive effects of treatment and negative side effects. The rating scale and standard gamble methods assessed utility; the rating scale method derives utilities directly by asking the patients to place health states on a thermometer scale (that is, vertical VAS), the standard gamble method derives utilities from the patients' responses to decision situations under risk.[19 20] Utility scores were assessed at baseline, and weeks 28 and 56.

INDICES

Various indices (that is, composites from several measures) were assessed in the COBRA study (table 1). In fact, a pooled index of five measures (composite measure to reflect each patient's standardised improvement) was the assigned primary outcome. Pooling is a validated method to increase responsiveness of separate measures.[21] To obtain a patient's pooled index score, the standardised change score was calculated by dividing change in one measure by its pooled standard deviation of

Table 1 Definition of indices

Index (ref)	*Calculation*
Pooled index (measures A, B, C...)[21]	Mean of (change in measure A/$SD_{change\ A}$; change in measure B/$SD_{change\ B}$; change in measure C/$SD_{change\ C}$; etc)
Disease activity score (DAS)[23]	0.54(Ritchie index of painful joints) + 0.065(swollen joint count) + 0.33ln (ESR) + 0.072(patient global)
ACR remission[24]	5 out of 6, for a period of 2 months: morning stiffness 15 min, no (joint pain), no joint pain on examination, no joint swelling, ESR <30 mm/1st h (men, <20 mm/1st h), no fatigue
COBRA, "probable" ACR remission[9]	4 out of 5, absence of fatigue assumed (not assessed in trial)
DAS remission[25]	DAS ⩽1.6
ACR improvement (20%)[3]	Improvement by at least 20% in tender joint count *and* swollen joint count *plus* in at least 3 out of the remaining 5 core set measures: acute phase reactant, physical function, doctor global assessment, patient global assessment, pain
Modified ACR improvement (##%)*	Improvement by at least ##% in tender joint count and swollen joint count *plus* in at least 3 out of the remaining 5 core set measures
No of core measures improved (##%)[26]*	Count of core set measures improved by at least ##%
EULAR response[4]	*"Good":* improvement in DAS >1.2 *and* final DAS ⩽2.4 *"Moderate":* not meeting criteria for "good", but improvement in DAS >0.6 *and* final DAS ⩽3.7 *"None":* remaining

*##% = thresholds varying from 0 to 70%.

change for each treatment group at week 28. This procedure was repeated for five measures; the pooled index is the mean of standardised scores. Finally, a constant was added so that all index values started with a zero value at baseline. To obtain pooled index values for time points other than week 28, change scores at that point were divided by the same factor (the SD of change of the measure at week 28). The trial was designed before the conception of the WHO/ILAR core set.[1] Recommendations at that time were to select five measures for maximum sensitivity to change[22]: tender joint count, global assessment by an independent assessor, ESR, grip strength, and MACTAR. The original disease activity score (DAS)[23] was also calculated. This index contains the Ritchie tender joint index, swollen joint count, ESR, and patient's global assessment on a 10 cm VAS (calculation: 0.54(Ritchie) + 0.065(swollen joint count) + 0.33ln (ESR) + 0.072(patient's global).

REMISSION, IMPROVEMENT, AND RESPONSE

Improvement in individual patients was also assessed in several ways: the ACR[24] and DAS remission criteria,[25] ACR improvement[3] and EULAR response[4] criteria, and count of improved core set measures[26] (table 1). Because fatigue was not measured in the trial "probable remission" described instances in which a patient would be in remission when absence of fatigue was assumed. Modified ACR improvement criteria and counts of improved core measures were also calculated with improvement thresholds varying from 0 to 70% (table 1). To calculate percentage improvement, values were recoded where necessary to ensure that all scales decreased on improvement.

FOLLOW UP

Initially, grip strength, ESR, and patient's assessment of disease activity (five point Likert scale) were registered weekly by research nurses, later at least every four weeks. All other reported assessments—with the exception of utilities—were made at baseline and at weeks 16, 28, (40, and 56) by trained independent assessors who contacted the patients only at these times. In this way, the assessors were unaware of the effects of high dose prednisolone during the first six weeks of the protocol. Utility scores were assessed biannually; thus for these measures only 28 week follow up measures are reported.

STATISTICAL ANALYSIS

All analyses were based on intention to treat: only five patients (3%, all in the sulfasalazine group) were lost to follow up before week 56 of the trial. The primary statistic of responsiveness was the standardised response mean (SRM): mean observed change from baseline divided by the standard deviation of this change.[27] The effect size (ES)[28]: the mean change from baseline divided by the standard deviation (SD) of baseline scores was also calculated. Confidence intervals of SRMs were calculated with the assumption that its distribution is approximately Gaussian with mean zero and SD of one over the square root of the sample size.[29] From the confidence intervals, statistical difference between SRMs could be evaluated. With no correction made for multiple comparison these findings are solely informative. Most evaluated variables are on an ordinal rather than interval or ratio scale level. However, as the underlying phenomenon (disease activity) is on an interval scale, these measures can be analysed parametrically if the sample size is large enough, as in the COBRA study database (central limit theorem).

Ceiling and floor effects may impair responsiveness when baseline values are found on the upper and lower end of the scale. We arbitrarily defined these extremes at the upper and lower one sixth of the scale (comparable with baseline HAQ scores <0.5 or >2.5) and analysed the variables in the core set.

As stated in the introduction, statistics based on change from therapeutic interventions need a priori confidence that the treatment is effective—that is, that the mean improvement in the treated group is relevant. The a priori criterion was a large change from baseline in

the combined-treatment group at 16 weeks ("gold standard"); less change from baseline was expected in the group treated with sulfasalazine ("silver standard"). This proved to be true, though the change reached at 28 weeks was slightly larger, especially in the sulfasalazine-only group. Thus the combined-treatment group SRM at week 16 was the primary statistic of responsiveness to form a league table for responsiveness. The SRMs in the sulfasalazine group can be used to assess the ability of a measure to detect smaller—but still meaningful—changes, or changes occurring in a smaller proportion of the treatment group.

To indicate the discriminative power between groups unpaired Student's *t* test values are reported. χ^2 Values reflect between-group contrast (that is, discriminative power) in the nominal variables: improvement and remission criteria. Because the primary study question of the COBRA trial concerned contrast between treatment groups, this contrast could not be an a priori criterion such as improvement in the combined-treatment group as outlined above.[30] Consequently, the ranking based on discrimination must be interpreted with caution. The 28 week data are included to allow further exploration of trends in responsiveness and discriminatory power.

Results

The combined-treatment group included 76 and the sulfasalazine group 79 patients. The groups were balanced in important demographic and prognostic variables.[9 18] At week 16 the mean improvement based on the pooled index was 1.4 for the combined-treatment group and 0.7 for the sulfasalazine group ($p<0.0001$). At week 28 these values were 1.5 *v* 0.8 ($p<0.0001$). In the combined-treatment group rates and rapidity of ACR 20, 50, and 70 improvement were similar to those reported in recent trials on anti-tumour necrosis factor (anti-TNF) treatment (see below).

At week 16, most measures indicated large improvement from baseline in both groups (SRM 0.4–4.1, ES 0.3–3.2), and all measures except patient's global assessment of health (VAS) significantly distinguished between combined treatment and sulfasalazine (table 2, fig 1). Statistics of responsiveness were larger in the combined-treatment group than in the sulfasalazine group, confirming a priori assumptions of greater improvement in this group. The relative responsiveness ranking of measures was similar in both groups, suggesting the ranking is stable over a broad range of relevant response. However, in the sulfasalazine-only group the absolute differences in responsiveness between measures were less, in proportion to the overall decreased response. All indices (that is, pooled index, MACTAR, DAS, and count of improved core set measures) were—in both treatment groups—considerably more responsive than single core set measures such as tender joint count. The only exception to this was the highly responsive single item patient's assessment of change in disease activity on a seven point Likert scale. The responsiveness of most single measures was satisfactory but not equal (for example, high responsiveness for pain and ESR, lower for tender joint count and CRP). A confidence interval smaller than 0.5 around the SRM estimates indicates that a difference between SRMs of 0.35 or greater would be significant when tested at the two sided 0.05 level. The results at week 28 were generally similar (table 3, fig 1).

The format of patient assessment of disease activity, physical impairment, and global well-being strongly influenced responsiveness (for a description of the formats see Appendix). The item in the MACTAR interview that asked for change of disease activity (seven point Likert scale) proved to be most responsive, patient's

Table 2 Indices of responsiveness at week 16 of follow up, for each treatment group, ordered by standardised response mean (SRM) of the combination group

	Combined treatment (n=75)				*Sulfasalazine (n=79)*				
	Mean change	SE_Δ	**SRM**	*ES*	*Mean change*	SE_Δ	*SRM*	*ES*	*t Value*
Count core set measures ≥20% improved	5.9	0.2	**4.1**	—	4.1	0.2	2.2	—	5.7
Change in disease activity by patient, 7 point Likert	2.6	0.1	**3.5**	—	1.7	0.2	1.2	—	5.3
Pooled index composite measure	1.4	0.1	**2.4**	—	0.7	0.1	1.1	—	6.5
MACTAR interview	12	0.6	**2.2**	3.2	8.5	0.8	1.3	2.1	3.7
Count core set measures ≥50% improved	4.6	0.2	**2.2**	—	2.4	0.3	1.0	—	6.0
Disease activity score (DAS) composite measure	−2.1	0.1	**1.9**	1.9	−1.2	0.1	1.0	1.2	4.9
AIMS pain scale	−8.4	0.5	**1.8**	2.3	−5.7	0.5	1.2	1.6	3.6
Health Assessment Questionnaire	−1.1	0.1	**1.5**	1.5	−0.4	0.1	0.8	0.6	6.2
Pain assessment by patient, VAS	−3.3	0.3	**1.5**	1.6	−1.8	0.3	0.7	0.8	4.0
Disease activity by patient, 5 point Likert	−1.7	0.1	**1.5**	1.6	−1.1	0.1	0.9	1.1	2.7
ESR (mm/1st h)	−41	3	**1.4**	1.2	−23	3	0.9	0.7	4.1
Swollen joint count ARA 48	−9	1	**1.4**	1.0	−5	1	0.7	0.6	4.2
Grip strength (kPa)	22	2	**1.3**	1.5	8	1	0.6	0.4	5.8
Count core set measures ≥70% improved	3.0	0.3	**1.3**	—	1.4	0.2	0.7	—	4.8
Global assessment by observer, VAS	−30	3	**1.3**	1.3	−14	3	0.6	0.7	4.3
Physical function by patient, 6 point Likert	1.8	0.2	**1.3**	1.7	1.3	0.2	0.9	1.2	2.2
Tender joint count ARA 68	−14	1	**1.2**	1.0	−6	1	0.5	0.4	4.1
Ritchie Arthritis Index tender joints	−10	1	**1.2**	1.0	−7	1	0.8	0.6	2.5
CRP (mg/l)	−33	4	**0.9**	0.8	−18	5	0.5	0.5	2.2
AIMS mobility scale	5.3	0.7	**0.9**	1.0	2.1	0.5	0.5	0.3	3.9
Global assessment by patient, VAS	−2.2	0.3	**0.8**	1.1	−1.7	0.3	0.6	0.9	1.1
AIMS self efficacy scale	5.1	0.8	**0.8**	0.9	2.4	0.6	0.4	0.4	2.7

SE_Δ = standard error of change; SRM = standardised response mean; ES = effect size; *t* value, unpaired (i.e. between group). SRM 95% confidence intervals have a width of <0.5 around the listed value, between-SRMs differences ≥0.35 are significant (two sided $p<0.05$, no correction for multiple comparison).

global assessment of health indicated on a VAS, least responsive. MACTAR, HAQ, some AIMS subscales, and single item patient global assessment of physical function were not equally responsive. The utility rating scale showed responsiveness close to the patient global assessment of disease activity, whereas utility measured by standard gamble was the least responsive of all measures.

Analyses on floor and ceiling effects showed that, of the core set variables, ESR and tender and swollen joint count were vulnerable to a certain degree of floor effect, with respectively 17, 15, and 15% of the patients in the lowest one sixth segment of the scale. Global and pain assessments, and also the HAQ had fewer patients that scored at the extremes of the scale.

The ranking for between-group discrimination showed interesting trends. This is best seen in fig 1: highest *t* values (that is, most discriminative power) were found for pooled

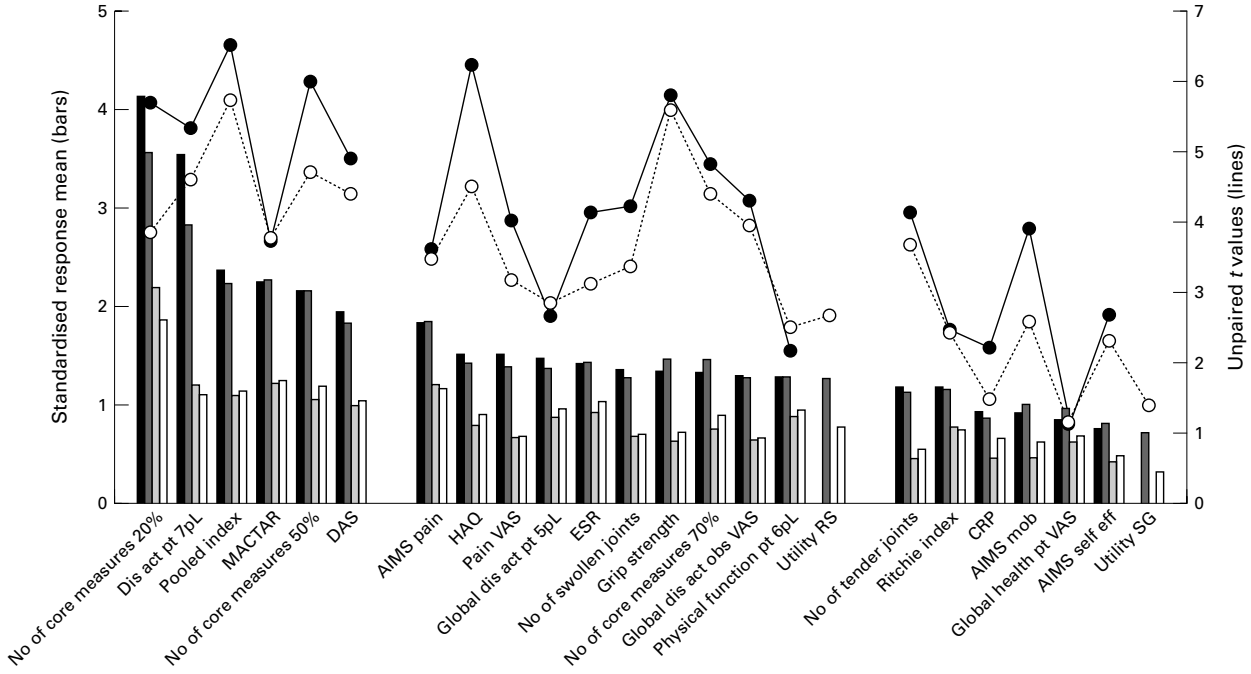

Figure 1 Comparison of responsiveness and discrimination performance of end points. The order of the measures evaluated corresponds with that of table 3. Dark bars: standardised response means of the combination group at 16 (black) and 28 weeks (dark grey). Light bars: standardised response means of the sulfasalazine group at 16 (light grey) and 28 weeks (white). Dots: unpaired t values of the between-group comparison at 16 (black) and 28 weeks (white). pL = points Likert; MACTAR = McMaster Toronto Arthritis patient preference questionnaire; DAS = disease activity score; AIMS = Arthritis Impact Measurement Scale; HAQ = Health Assessment Questionnaire; VAS = visual analogue scale; ESR = erythrocyte sedimentation rate; RS = rating scale; CRP = C reactive protein; SG = standard gamble.

*Table 3 Indices of responsiveness at week 28 of follow up, for each treatment group**

	Combined treatment (n=75)				*Sulfasalazine (n=79)*				
	Mean change	*SE_Δ*	**SRM**	*ES*	*Mean change*	*SE_Δ*	*SRM*	*ES*	*t Value*
Count core set measures ≥20% improved	5.8	0.2	**3.6**	—	4.5	0.3	1.9	—	3.9
Change in disease activity by patient, 7 point Likert	2.5	0.1	**2.8**	—	1.6	0.2	1.1	—	4.6
Pooled index composite measure	1.5	0.1	**2.4**	—	0.8	0.1	1.1	—	5.7
MACTAR interview	12	0.6	**2.3**	3.3	8.6	0.8	1.2	2.2	3.8
Count core set measures ≥50% improved	4.8	0.3	**2.2**	—	3.0	0.3	1.2	—	4.7
Disease activity score (DAS) composite measure	−2.3	0.1	**1.8**	2.2	−1.4	0.1	1.0	1.4	4.4
AIMS pain scale	−8.5	0.5	**1.8**	2.4	−5.8	0.6	1.2	1.4	3.5
Health Assessment Questionnaire	−1.1	0.1	**1.4**	1.5	−0.6	0.1	0.9	0.8	4.5
Pain assessment by patient, VAS	−3.4	0.3	**1.4**	1.7	−2.0	0.3	0.7	0.9	3.2
Disease activity by patient, 5 point Likert	−1.8	0.2	**1.4**	1.7	−1.2	0.1	1.0	1.2	2.8
ESR (mm/1st h)	−40	3	**1.4**	1.2	−27	3	1.0	0.8	3.1
Swollen joint count ARA 48	−10	1	**1.3**	1.1	−5	1	0.7	0.7	3.4
Grip strength (kPa)	25	2	**1.5**	1.6	11	2	0.7	0.5	5.6
Count core set measures ≥70% improved	3.6	0.3	**1.5**	—	1.9	0.2	0.9	—	4.4
Global assessment by observer, VAS	−3.3	0.3	**1.3**	1.4	−1.7	0.3	0.7	0.8	3.9
Physical function by patient, 6 point Likert	2.0	0.2	**1.3**	1.7	1.4	0.2	0.7	1.2	2.5
Utility† rating scale technique	0.24	0.02	**1.3**	1.3	0.15	0.02	0.8	0.8	2.7
Tender joint count ARA 68	−16	2	**1.1**	1.2	−8	2	0.5	0.5	3.7
Ritchie Arthritis Index tender joints	−11	1	**1.2**	1.1	−7	1	0.8	0.7	2.4
CRP (mg/l)	−32	5	**0.9**	0.8	−22	4	0.7	0.7	1.5
AIMS mobility scale	5.3	0.6	**1.0**	1.0	3.2	0.6	0.6	0.5	2.6
Global assessment by patient, VAS	−2.4	0.3	**1.0**	1.2	−1.9	0.3	0.7	1.0	1.1
AIMS self efficacy scale	5.2	0.7	**0.8**	0.9	2.9	0.7	0.4	0.4	2.3
Utility† standard gamble technique	0.10	0.02	**0.7**	0.7	0.06	0.02	0.3	0.3	1.4

*Abbreviations, see table 2. SRM 95% confidence intervals have a width of <0.5 around the listed value, between-SRMs differences ≥0.35 are significant (two sided p<0.05, no correction for multiple comparison).
†Utilities were only assessed at week 28.

Table 4 Discriminative ability of the preliminary ACR improvement criteria at different thresholds, EULAR response criterion, ACR and DAS remission criteria to distinguish between two treatment groups at two points of follow up

	Week 16			*Week 28*		
Criterion	*Improved**	χ^2	*p Value*	*Improved**	χ^2	*p Value*
ACR improvement						
0% Threshold	88 *v* 63%	12.9	0.0003	82 *v* 62%	7.3	0.007
10% Threshold	78 *v* 44%	18.0	<0.0001	74 *v* 49%	9.7	0.002
20% Threshold	**72 *v* 32%**	**25.7**	**<0.0001**	**72 *v* 49%**	**8.6**	**0.003**
30% Threshold	64 *v* 24%	25.7	<0.0001	67 *v* 39%	12.1	0.0005
40% Threshold	58 *v* 16%	28.6	<0.0001	55 *v* 30%	9.8	0.002
50% Threshold	**43 *v* 14%**	**16.6**	**<0.0001**	**49 *v* 27%**	**8.1**	**0.004**
60% Threshold	26 *v* 8%	9.7	0.002	37 *v* 22%	4.4	0.04
70% Threshold	**16 *v* 6%**	**3.6**	**0.06**	**29 *v* 10%**	**8.8**	**0.003**
EULAR response		20.2	<0.0001		13.4	0.001
Good + moderate	86 *v* 56%			86 *v* 63%		
Good	37 *v* 15%			47 *v* 24%		
ACR remission						
'Probable'	12 *v* 6%	1.4	0.23	21 *v* 11%	2.7	0.10
DAS remission						
(DAS ≤1.6)	12 *v* 9%	0.4	0.54	17 *v* 10%	1.6	0.20

*Combined treatment (n=76) *v* sulfasalazine treatment (n=79) group.

index, count of core set measures improved by 50%, DAS, HAQ but, also, grip strength. Between the two assessments a catch-up effect is seen in the sulfasalazine group: whereas improvements in the combination group were already maximum at week 16, the sulfasalazine group improved further between week 16 and week 28, resulting in a smaller between-group difference (and thus a smaller *t* value).

At 16 weeks, ACR 20% improvement and EULAR response criteria showed large χ^2 values, consistent with significant differences in response between treatment groups (table 4). The discriminatory performance of these criteria ranks high among all the measures tested (table 4): based on p values a χ^2 value of 8 roughly corresponds with a *t* value of 3; similarly, a χ^2 value of 12 corresponds with a *t* value of 4, and a χ^2 value of 25 with a *t* value of 5. At week 28 the differences between the groups were smaller. At week 16, modification of the percentage value in the ACR improvement criterion between 0% (no improvement, no worsening) and 50% did not change its discriminatory capacity; at week 28, this was also true for the 70% cut off point (table 4). The ACR and DAS remission criteria did not show a significant between-group difference.

Discussion

This study is the first independent confirmation of the responsiveness of the WHO/ILAR core set measures and response criteria in a trial of patients with early RA. In addition, it lends strong support to the use of other indices in such a trial. The conclusions are strong because they are based on the findings in two groups of patients with a high and moderate level of expected response. They extend the validity of both the core set and the ACR response criteria, because these had initially been selected, designed, and tested mainly in placebo controlled studies.

The fact that indices are more responsive than most single measures is not surprising, as combining measures (or items in a questionnaire) reduces scatter. Even a simple count of improved core set outcome variables proved to be a very responsive index, especially at the 20% threshold. Directly asking for change can also reduce scatter, even though the answer may be biased towards the current condition. Evidence for this is shown by the high responsiveness of the patient change question and the MACTAR (that incorporates many change items). The responsiveness of functional scales may be partly explained by the fact that they generally comprise several items in a multi-item questionnaire. Nevertheless, a set of two physical function questions on a six point Likert scale was also responsive.

The strong influence of format and content of the patient's global assessment questions on responsiveness is worrying. Similarly, the responsiveness of pain as a measure depends on the format. It is likely that the focus of doctors (or other assessors) is on the patients' disease, but this seems not always be the case for the patients themselves. Although not specified in great detail in the original formulation of the WHO/ILAR core set, we advocate focusing the format of patient oriented instruments on the disease, and paying close attention to the exact wording of the question(s).

Utility scores are advocated as a generic measure of treatment benefits. The two methods to derive utilities proved to have quite different levels of responsiveness. The rating scale (which is a patient preference rather than a true utility) performed adequately (comparable with observer's global assessment on VAS). However, the standard gamble method (a true utility because choices are made in a situation of uncertainty) showed low responsiveness. Economists prefer standard gamble because it conforms better to theoretical principles, but in practice its application was hindered by limited comprehension of the method by our patients and their risk aversive attitude. This phenomenon has been seen before in patient groups with a non-fatal or chronic disease.[31 32]

The data on between-group discrimination must be interpreted with caution. The extent of differences between the groups was not known before the trial, and might have been large in comparison with expected differences in current and future head-to-head trials. Nevertheless, the results are unique and extremely interesting as they indicate that responsiveness, the ability to detect change, may not parallel the ability to discriminate between different levels of response. Both the ACR improvement criteria (at various percentage levels) and the EULAR response criterion showed excellent discriminatory ability. This is at odds with the other trials in the review of Felson *et al*, who concluded that 20% remained the best cut off point for the ACR criteria.[33] A possible explanation is the relatively large contrast between treatment groups in the COBRA study.[34] Other indices were also better discriminators than most single measures, with the exception of grip strength. In contrast, the discriminatory capacity of the MACTAR, though good, was less than expected based on its excellent responsiveness. This difference in performance between the HAQ and the MACTAR is hard to explain, and will need replication in other studies. Grip strength was

included in the design of the trial based on the work of Anderson *et al.*[22] Despite its good performance in trials up to 1989, grip strength was eventually excluded in the core set for reasons of redundancy. Nevertheless, in early RA the fact that it is a composite measure of hand function with pain, swelling, stiffness, and muscle strength may contribute to its excellent performance. Muscle strength, particularly, may be a physical function variable with potency in early and established RA.[35]

From published reports we know that different responsiveness statistics—also those that are solely based on change from therapeutic intervention—may[36] or may not[37] yield different rank orders. In general, rankings based on paired Student's *t* test values and SRM will only be discrepant when different sample sizes are used for different measures. SRM is least influenced by sample size as it avoids the use of standard error of the mean in the denominator. Sample size was not an issue in this report as few values were missing. ES and SRM generally yield similar ranks, though discrepancies occur when the within-group SD at baseline (the denominator in the ES calculation) differs much from the SD of within-group change (the denominator in the SRM calculation). Obviously, ES cannot be calculated for measures that directly evaluate change, because they do not have a baseline variance. It may be a typical feature of these transitional measures—and indices that include change questions, such as the MACTAR—to pair a large SRM with a relatively small unpaired *t* value.

Despite their strong evidence, the data represent only one study in one subgroup—that is, early RA. The generalisability of our findings may be slightly limited, as the effects of treatment in the combined-treatment arm were large compared with many other trials in RA, but similar to those seen in recent anti-TNF trials. A meta-analysis of effectiveness of low dose corticosteroids in RA reported somewhat smaller ES in measurements of grip strength, swollen and joint tender count, and ESR (0.4–1.0) than we did; in particular, the ES values found in the corticosteroid treatment arms were smaller.[38] However, the effect in the sulfasalazine group resembles that found in trials of methotrexate and intramuscular gold.[39–42] Thus for studies of such moderately effective drugs, the ranking of the sulfasalazine group might be more appropriate.

Analyses on floor and ceiling effects showed that ESR and tender and swollen joint count were vulnerable to a certain degree of floor effect. The study's inclusion criteria towards disease duration and disease activity, with evaluation based on ESR and joint counts, probably prevented a serious floor effect in the study group. With global and pain assessments on a visual analogue scale, people tend to put their mark somewhere at the middle of the scale.

Buchbinder *et al* studied the ability of end points to discriminate between treatment effects in a placebo controlled trial of cyclosporin in RA.[43] As the difference between cyclosporin and placebo was the primary study question of that trial, their approach is similar to the post hoc discrimination tests between treatment groups in this report. Compared with the COBRA study, differences between treatment groups were smaller for ESR and swollen joint counts but similar in other measures. They found doctor's and patient's global assessments (measured as a change question), as well as the AIMS pain subscale to be most discriminatory, and ESR and pain (five point scale) to be least discriminatory, with all other core set measures, including another doctor's global question, the HAQ, and a modification of the MACTAR (that is, PET), falling in between. These results agree with our observation on the importance of the exact format of the questions. The discrepancy found in the ESR is expected: lack of responsiveness of ESR is well known during treatment with cyclosporin. More surprising is the relatively poor performance of the physical function questionnaires. It may be that the cyclosporin trial included patients with longstanding disease and more fixed disability that was less likely to respond to treatment. Patients in the COBRA study had a median disease duration of only four months.

Differences in responsiveness, and especially discrimination, have important implications for trial design. The use of responsive and discriminative measures allows reduced patient numbers or detection of smaller—yet relevant—differences between groups. This is important especially in trials of early RA. Simpler trial design through use of a limited number of measures saves costs and effort, and facilitates interpretation of the results. In routine patient care, use of a limited number of highly responsive measures facilitates the collection and interpretation of long term follow up data. Obviously, additional measures should be applied according to the characteristics of the individual patient.

In summary, this study convincingly shows that responsiveness and the ability to discriminate between different levels of response are not identical concepts. The data provide strong evidence for the responsiveness and discriminatory capacity of the WHO/ILAR core set as well as the ACR and EULAR response criteria in the study of moderately and strongly effective drugs in early RA. However, where information is elicited from the patient, researchers should select and focus their instruments on the disease, as performance is strongly dependent on the exact format of questions.

AC Verhoeven is research-fellow in the COBRA trial supported by grant of the "Ziekenfondsraad, fonds Ontwikkelingsgeneeskunde" (92–045), The Netherlands.

Appendix: Formats of items with patient assessment

Assessment of change in disease activity (7 point Likert scale)

Q. *When you think of your arthritis during the two weeks before the first interview, how much better or worse overall has your arthritis become ?*

- ❑ 7 *A great deal better*
- ❑ 6 *Moderately better*
- ❑ 5 *Slightly better*
- ❑ 4 *No change*
- ❑ 3 *Slightly worse*
- ❑ 2 *Moderately worse*
- ❑ 1 *A great deal worse*

Assessment of disease activity (5 point Likert scale)

Q. *When you think of your arthritis, how would you say your condition has been over the past week ?*

- ❑ 5 *Good*
- ❑ 4 *Reasonably good*
- ❑ 3 *Moderate*
- ❑ 2 *Poor*
- ❑ 1 *Very poor*

Assessment of physical function (two questions, 6 point Likert scale)

Qa. *How would you say your overall physical functioning has been ? Over the past week you think of it as...*

- ❑ 5 *Good*
- ❑ 4 *Good to fair*
- ❑ 3 *Fair*
- ❑ 2 *Fair to poor*
- ❑ 1 *Poor*

Qb. *Is your physical function not as good as it might be because of your arthritis ?*

❑ 0 *Yes* ❑ 1 *No*

Global assessment (10 cm visual analogue scale)

Q. *How has your general health been during the past week ?*

Pain assessment (10 cm visual analogue scale)

Q. *How much pain did you have during the past week ?*

1 Boers M, Tugwell P, Felson DT, van Riel PL, Kirwan JR, Edmonds JP, *et al*. World Health Organisation and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol 1994;41(suppl):86–9.
2 Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, *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.
3 Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, *et al.* American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
4 van Gestel AM, Prevoo MLL, van 't Hof MA, van Rijswijk MH, van de Putte LBA, van Riel PLCM. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Arthritis Rheum 1996;39:34–40.
5 Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for outcome measures. J Rheumatol 1998;25:198–9.
6 Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995;22:1191–2.
7 Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50: 869–79.
8 Pincus T, Stein CM. What is the best source of useful data on the treatment of rheumatoid arthritis: clinical trials, clinical observations, or clinical protocols? J Rheumatol 1995;22:1611–17.
9 Boers M, Verhoeven AC, Markusse HM, van de Laar MAFJ, Westhovens R, van Denderen JC, *et al*. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309–18.
10 Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, *et al*. The American Rheumatism Association revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
11 The cooperating clinics committee of the American Rheumatism Association. A seven-day variability study of 499 patients with peripheral rheumatoid arthritis. Arthritis Rheum 1965;8:302–34.
12 Fries JF, Spitz PW, Young DY. The dimensions of health outcome: the Health Assessment Questionnaire, disability and pain scales. J Rheumatol 1982;9:789–93.

13 Siegert CEH, Vleming LJ, Vanderbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305–9.
14 Jones E, Hanly JG, Mooney R, Rand LL, Spurway PM, Eastwood BJ, *et al*. Strength and function in the normal and rheumatoid hand. J Rheumatol 1991;18:1313–18.
15 Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis; the Arthritis Impact Measurement Scale. Arthritis Rheum 1980;23:146–52.
16 Tugwell P, Bombardier C, Buchanan WW, Goldsmith C, Grace E, Hanna B. The MACTAR patient preference disability questionnaire: an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987; 14:446–51.
17 Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. J Rheumatol 1993;20:429–36.
18 Verhoeven AC, Bibo JC, Boers M, Engel GL, van der Linden S. Cost-effectiveness and cost-utility of combination therapy in early rheumatoid arthritis: randomized comparison of combined step-down prednisolone, methotrexate, and sulphasalazine with sulphasalazine alone. Br J Rheumatol 1998;37:1102–9.
19 Bennett K, Torrance GR, Tugwell P. Methodological challenges in the development of utility measure of health-related quality of life in rheumatoid arthritis. Control Clin Trials 1991;12(suppl):118–28.
20 Bakker CH, Rutten-van Mölken MPMH, van Doorslaer EKA, Bennet K, van der Linden S. Health related utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. Patient Education Counsel 1993;20:145–52.
21 Smythe HA, Helewa A, Goldsmith CH. 'Independent assessor' and 'pooled index' as techniques for measuring treatment effects in rheumatoid arthritis. J Rheumatol 1977;4:144–52.
22 Anderson JJ, Felson DT, Meenan RF, Williams HJ. Which traditional measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 1989;32:1093–9.
23 Van der Heijde DMFM, van 't Hof MA, van Riel PLCM, Theunisse LM, Lubberts EW, van Leeuwen MA, *et al*. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 1990;49:916–20.
24 Pinals RS, Masi AT, Larsen RA. Preliminary criteria for clinical remission in rheumatoid arthritis. Arthritis Rheum 1981;24:1308–15.
25 Prevoo MLL, van Gestel AM, van 't Hof MA, van Rijswijk MH, van de Putte LBA, van Riel PLCM. Remission in a prospective study of patients with rheumatoid arthritis. American Rheumatism Association preliminary remission criteria in relation to the disease activity score. Br J Rheumatol 1996;35:1101–5.
26 Tugwell P, Pincus T, Yocum D, Stein M, Gluck O, Kraag G, *et al*. Combination therapy with cyclosporine and methotrexate in severe rheumatoid arthritis. N Engl J Med 1995; 333:137–41.
27 Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632–42.
28 Kazis L, Anderson J, Meenan R. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178–9.
29 Beaton BE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79–93.
30 Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use*. 2nd ed. Oxford: Oxford University Press, 1995:164–6.
31 Bakker CH, Rutten-van Mölken MPMH, van Doorslaer EKA, Bennet K, van der Linden S. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. J Rheumatol 1994;21:269–74.
32 Goossens MEJB, Rutten-van Mölken MPMH, Leidl RMJ, Bos SGPM, Vlaeyen JWS, Teeken-Gruben NJG. Cognitive-educational treatment of fibromyalgia: randomized clinical trial. II. Economic evaluation. J Rheumatol 1996;23:1246–54.
33 Felson DT, Anderson JJ, Lange MLM, Wells G, LaValley MP. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998;41:1564–70.
34 Anderson JJ, Wells G, Verhoeven AC, Felson DT. Factors predicting response to treatment in rheumatoid arthritis: the importance of disease duration. Arthritis Rheum 2000; 43:22–9.
35 Stucki G, Bruhlmann P, Stucki S, Michel BA. Isometric muscle strength is an indicator of self-reported physical functional disability in patients with rheumatoid arthritis. Br J Rheumatol 1998;37:643–8.
36 Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997:50:239–46.
37 Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spine stenosis. J Clin Epidemiol 1995;48:1369–78.
38 Saag KG, Lindsey AC, Sems K, Nettleman MD, Kolluri S. Low-dose corticosteroids in rheumatoid arthritis. A meta-analysis of their moderate-term effectiveness. Arthritis Rheum 1996;39:1818–25.
39 Weinblatt ME, Kaplan H, Germain BF, Merriman RC, Solomon SD, Wall B, *et al*. Low-dose methotrexate compared with auranofin in adult rheumatoid arthritis; a thirty-six-week, double-blind trial. Arthritis Rheum 1990; 33:330–8.
40 Ward JR, Williams HJ, Egger MJ, Reading JC, Boyce E, Altz-Smith M, *et al*. Comparison of auranofin, gold sodium thiomalate, and placebo in the treatment of rheumatoid arthritis; a controlled clinical trial. Arthritis Rheum 1983;26:1303–15.
41 Rau R, Herborn G, Karger T, Menninger H, Elhardt D, Schmitt J. A double blind randomized parallel trial of intramuscular methotrexate and gold sodium thiomalate in early erosive rheumatoid arthritis. J Rheumatol 1991;18: 328–33.
42 Jeurissen MEC, Boerbooms AMT, van de Putte LBA, Doesburg WH, Mulder J, Rasker JJ, *et al*. Methotrexate versus azathioprine in the treatment of rheumatoid arthritis. A forty-eight-week randomized double-blind parallel trial. Arthritis Rheum 1991;34:961–72.
43 Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial. Arthritis Rheum 1995;38:1568–80.