Comparison of Internal and External Responsiveness of the Generic Medical Outcome Study Short Form-36 (SF-36) with Disease-specific Measures in Rheumatoid Arthritis

MARTINE M. VEEHOF, PETER M. ten KLOOSTER, ERIK TAAL, PIET L.C.M. van RIEL, and MART A.F.J. van de LAAR

ABSTRACT. ***Objective***. To examine the comparative internal and external responsiveness of the generic Medical Outcome Study Short Form-36 Health Survey (SF-36) and disease-specific measures in patients with rheumatoid arthritis (RA).

Methods. Data were collected from 280 RA patients starting anti-tumor necrosis factor treatment. A total of 168 patients completed a questionnaire including the SF-36, the Arthritis Impact Measurement Scales 2 (AIMS2), the Health Assessment Questionnaire (HAQ), a visual analog scale for general health (VAS-GH), and an 11-point numerical rating scale for pain (NRS pain) at baseline and after 12 months. Internal responsiveness was evaluated with paired samples t-tests and standardized response means (SRM). External responsiveness was investigated with receiver-operating characteristic statistics and Spearman rank-order correlation coefficients. A health transition item was used as the external indicator of change.

Results. No significant differences in internal and external responsiveness were found between the SF-36 and disease-specific measures within the domains physical function, pain, and psychological function. In the domain social function, the SF-36 was more responsive than the AIMS2. In the domain general health, the SF-36 was less responsive (only internal) than the AIMS2 and VAS-GH.

Conclusion. Our study showed comparable internal and external responsiveness of the SF-36 compared with disease-specific measures (AIMS2, HAQ, NRS pain) in all health domains, except social function and general health domains. The assumption that disease-specific measures are more responsive to detect intervention-related changes over time is not confirmed by our data. (First Release Mar 1 2008; J Rheumatol 2008;35:610–7)

Key Indexing Terms:
RHEUMATOID ARTHRITIS
OUTCOME MEASURES
HEALTH-RELATED QUALITY OF LIFE
RESPONSIVENESS

The impact of disease on human life encompasses more than the clinical manifestations of the disease or the pathophysiological process. Therefore, in the 1980s the concept of health-related quality of life (HRQOL) was introduced. HRQOL describes the influence of a disease on all dimensions of health, such as signs and symptoms, function, and psychological and social well-being. To date, the concept of HRQOL has been measured by self-administered questionnaires that provide information from the perspective of the patient. Measurement of HRQOL is warranted, on one hand, to better understand the effects of a disease and, on the other hand, to personalize treatment, assess a patient's progress, and evaluate the effects of treatment.

Several generic and disease-specific measures have been developed to assess HRQOL. Generic instruments focus on general issues of health and are developed for any population irrespective of disease or condition[1,2]. A commonly used generic measure is the Medical Outcome Study Short Form-36 Health Survey (SF-36)[3]. Disease-specific instruments, on the other hand, are developed for a specific disease or condition and thus contain items of particular relevance to the disease or condition[1,2]. Disease-specific measures used frequently in rheumatology are the Arthritis Impact Measurement Scales 2 (AIMS2)[4] and the Health Assessment Questionnaire (HAQ)[5]. Both generic and disease-specific measures have their own advantages and disadvantages. Where generic measures allow comparisons across different diseases and with the normal population,

From the Institute for Behavioral Research, University of Twente, Enschede; Department of Rheumatology, Radboud University Nijmegen Medical Centre, Nijmegen; and Department of Rheumatology, Medisch Spectrum Twente, Enschede, The Netherlands.

Funded by an unrestricted educational grant by Schering-Plough and CVZ (Health Care Insurance Board).

M.M. Veehof, MSc, OT, University of Twente; P.M. ten Klooster, MSc, University of Twente; E. Taal, PhD, University of Twente; P.L.C.M. van Riel, MD, PhD, Professor, Radboud University Nijmegen Medical Centre; M.A.F.J. van de Laar, MD, PhD, Professor, University of Twente and Medisch Spectrum Twente.

Address reprint requests to M.M. Veehof, Institute for Behavioural Research, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: m.m.veehof@utwente.nl

Accepted for publication November 14, 2007.

disease-specific measures have the potential to be more responsive to (intervention-related) changes over time[1,2,6].

The responsiveness of a measure is an important factor to consider when deciding to use a generic or disease-specific measure in research or daily clinical care, particularly when the aim is to measure changes over time[7,8]. Presently, consensus on a definition of responsiveness and the best study design and analysis strategy to assess it is still lacking[9-11]. Husted, *et al*'s review concluded that 2 major types of responsiveness exist: internal responsiveness and external responsiveness. Internal responsiveness describes the ability of a measure to change over a prespecified timeframe, whereas external responsiveness describes the relationship between change in a measurement and change in a reference measurement of health status (external criterion)[9]. Studies on the responsiveness of the SF-36 compared with disease-specific measures (AIMS2, HAQ) in patients with rheumatoid arthritis (RA) are scarce. The few studies that address this subject showed conflicting results and/or used different study designs and analysis strategies for responsiveness[6,12-14]. The aim of this study was to assess the internal and external responsiveness of the SF-36 in comparison with disease-specific instruments in patients with RA.

MATERIALS AND METHODS

Patients and study design. The data for this study were collected as part of the ongoing Dutch Rheumatoid Arthritis Anti-TNF Monitoring (DREAM) study, a register that started in April 2003 to prospectively monitor and evaluate the use of anti-tumor necrosis factor (TNF) in patients with RA in 12 hospitals in The Netherlands. Inclusion criteria for the DREAM study are: diagnosis of RA, active disease [Disease Activity Score 28 (DAS28) > 3.2][15], previous treatment with at least 2 antirheumatic drugs including methotrexate (MTX) at an optimal dose or intolerance for MTX, and no previous treatment with anti-TNF agents.

In the DREAM study, all RA patients starting anti-TNF treatment are seen every 3 months by independent trained research nurses, who collect data on patients' demographics (age, gender, disease duration), clinical condition (DAS28, functional class according to Steinbrocker), health status [SF-36, visual analog scale for general health (VAS-GH)], and functional status (HAQ). For this study, we used data from centers that additionally performed the AIMS2 and an 11-point numerical rating scale for pain (NRS pain) at baseline and at 3 and 12 months.

Measures. *SF-36.* The SF-36 is a generic health status questionnaire containing 36 items, 35 of which are combined into 8 scales: physical function, bodily pain, social function, mental health, general health, vitality, role physical, and role emotional[3,16,17]. Scale scores were calculated according to published scoring procedures[18] and range from 0 (poor health) to 100 (optimal health). Only scales that are identified by disease-specific measures were included for analysis: physical function, bodily pain, social function, mental health, and general health. The SF-36 has been shown to be a reliable, valid, and responsive questionnaire in patients with RA[19-24]. The responsiveness of the Dutch version of the SF-36 has never been investigated in RA.

SF-36: health transition item. A single item of the SF-36, the health transition item, gives an indication of perceived change in general health over the past 12 months. This item is scored on a 5-point scale ranging from "much better" to "much worse"[3,17]. Fitzpatrick, *et al* provided evidence on the validity of the use of a transition item to assess change in health status in RA[25].

AIMS2. The AIMS2 is a disease-specific measure developed for patients with arthritis[4,26]. This 57-item questionnaire contains 12 scales to assess 5 dimensions of health: physical function, symptom, affect, and social interaction and role. One additional item is included to assess general health perception. Component scores were calculated, ranging from 0 (good health) to 10 (poor health). The responsiveness of the Dutch AIMS2 has been investigated by Taal, *et al* and was shown to be satisfactory[27].

HAQ. The HAQ is a disease-specific questionnaire developed to assess functional limitations in patients with rheumatic diseases[5,28-30]. The instrument contains 20 items on 8 domains of life (dressing, arising, eating, walking, hygiene, reach, grip, and common activities). The HAQ standard disability index (HAQ-DI) was calculated, which takes into account the use of aids and devices. The HAQ-DI yields a score from 0 to 3, with higher scores indicating more disability. The Dutch version of the HAQ has been shown to be a responsive measure[31].

NRS pain. Arthritis pain was measured on an 11-point numerical rating scale with verbal anchors from "no pain" (0) to "extreme pain" (10). This scale is part of the Rheumatoid Arthritis Disease Activity Index (RADAI)[32].

VAS-GH. The VAS-GH is a 100 mm line with verbal anchors from "very good health status, could not be better" (0) to "very bad health status" (100). Patients were asked to rate their current general health.

Data analysis. Demographic and clinical characteristics and scores on outcome measures were described. Continuous data were presented as means with standard deviations (SD). Categorical data were presented as proportions. The Kolmogorov-Smirnov test was used to test the normality of the distribution of the scores on the outcome measures. In accordance with Husted, *et al*, we assessed the internal and external responsiveness of the SF-36 and corresponding disease-specific measures[9]. Since high scores on SF-36 indicate good health, while high scores on AIMS2, HAQ, NRS pain, and VAS-GH indicate poor health, we multiplied the change scores of SF-36 by –1, to facilitate comparison among the instruments. Analyses were performed using the statistical packages SPSS 12.0, S-PLUS 6.1, and MedCalc 8.1.

Internal responsiveness. The paired samples t-test (for the normally distributed measures) and Wilcoxon signed-rank test (for non-normally distributed measures) were used to assess the ability of the measures to assess changes between baseline and 12-month followup assessments. Change was considered significant when $p \leq 0.05$. Further, standardized response means (SRM) were calculated. The SRM is calculated as the mean change score divided by the standard deviation of that change score and is seen as an indicator of the ability to distinguish "signal" from "noise"[33,34]. In accord with the criteria of Cohen[35], a SRM between 0.20 and 0.49 can be interpreted as a small effect, a SRM between 0.50 and 0.79 as a moderate effect, and a SRM equal to or greater than 0.80 as a large effect[9]. We applied a bootstrap procedure to obtain 95% confidence intervals (95% CI) for the SRM [36]. Bootstrapping consists of resampling with replacement. We selected 1000 samples (each of 168 observations) with replacement and calculated the SRM for each sample. The SRM of the bootstrap samples were ordered from lowest to highest and the 95% CI for the SRM were obtained by reading the 25th and 975th observations. The comparative responsiveness of the SF-36 and the disease-specific measures was determined by comparing the SRM and calculating a 95% CI for the difference in SRM, using the 1000 bootstrap samples. SRM were considered significantly different if the interval did not contain the value zero[37].

External responsiveness. Receiver-operating characteristic (ROC) curves and Spearman rank-order correlation coefficients with 95% CI were computed to describe the relationship between changes in the measure and an external indicator of change. We used the health transition item of the SF-36 as external indicator. For the ROC curves this item was coded as a binary variable. Patients who judged their health after 12 months of anti-TNF treatment as "much better" or "somewhat better" were classified into the "improved health" group. Patients who judged their health as "about the same", "somewhat worse," or "much worse" were classified into the "non-improved health" group. The areas under the ROC curves (AUC) were cal-

culated to quantify the probability of the measures to correctly classify patients as improved or non-improved. The areas range from 0.5 (no accuracy in distinguishing improvers from non-improvers) to 1.0 (perfect accuracy). The comparative accuracy of the SF-36 and the disease-specific measures was determined by comparing the AUC using the Wilcoxon signed-rank test[38]. A 95% CI was computed for the difference in AUC. The areas were considered significantly different if the interval did not contain the value zero.

RESULTS

Patient characteristics. Two hundred eighty patients were included in this study. Of them, 168 (60%) completed all the questionnaires at baseline and after 12 months of followup. There were no significant differences in demographic (age, gender) and baseline clinical characteristics (disease duration, DAS28, Steinbrocker functional class) between patients who did and who did not complete all questionnaires at both measurement times (data not shown). Data from patients who did not complete all questionnaires at baseline and after 12 months of followup were excluded from further analyses.

At baseline, 71% of the 168 patients were female and mean age and mean disease duration were 54.2 (SD 12.6) and 10.2 (SD 9.2) years, respectively. Mean DAS28 was 5.5 (SD 1.2), indicating high disease activity at study entry. The majority of the patients (81%) had mild disability and were classified into Steinbrocker functional class II.

Internal responsiveness. In Table 1 mean scores at baseline and 12-month changes are described. Results are shown for each domain of health separately. All measures showed significantly improved scores after 12 months of TNF-blocking treatment.

Table 1. Mean scores at baseline and 12-month changes for SF-36 and disease-specific measures. Values are means (SD).

Health Domain	Baseline	12-month Changes
Physical function, n = 151		
SF-36 physical function	37.12 (22.06)	14.56 (19.49)
AIMS2 physical function	3.11 (1.63)	–0.75 (1.24)
HAQ-DI	1.43 (0.57)	–0.28 (0.48)
Pain, n = 167		
SF-36 bodily pain	37.91 (18.21)	18.53 (21.38)
AIMS2 symptom	6.62 (2.19)	–2.17 (2.29)
NRS pain	5.74 (2.59)	–2.30 (3.05)
Social function, n = 161		
SF-36 social function	65.02 (22.62)	11.26 (23.02)
AIMS2 social interaction	3.85 (1.37)	–0.22 (1.11)
Psychological function, n = 158		
SF-36 mental function	71.25 (17.07)	6.80 (14.72)
AIMS2 affect	3.50 (1.60)	–0.67 (1.32)
General health, n = 164		
SF-36 general health	44.42 (18.83)	4.03 (16.89)
AIMS2 general health	6.67 (2.31)	–1.68 (2.51)
VAS-GH	58.94 (21.99)	–18.84 (27.79)

All scores are significantly improved at 12-month followup assessments ($p \leq 0.05$). SF-36: Medical Outcome Study Short Form-36; AIMS2: Arthritis Impact Measurement Scales 2; HAQ-DI: Health Assessment Questionnaire Disability Index; NRS: numeric rating scale; VAS-GH: visual analog scale for general health.

In Table 2 SRM and 95% CI are presented. Within the domains physical function, pain, and psychological function the SRM were quite similar and no significant differences were found between the SF-36 and the disease-specific measures (AIMS2, HAQ, NRS pain). A significant difference was found only between the AIMS2 pain scale and the NRS for pain. The AIMS2 was more responsive to detect improvement in pain than the NRS (difference in SRM = 0.20, 95% CI 0.02–0.38). Within the domains social function and general health the SRM were quite different, and significant differences were found between the SF-36, the AIMS2, and the VAS-GH. In the domain social function, the SF-36 was more responsive than the AIMS2 (difference in SRM = 0.29, 95% CI 0.07–0.54). In the domain general health, the SF-36 was less responsive than the AIMS2 (difference in SRM = 0.43, 95% CI 0.21–0.62) and the VAS-GH (difference in SRM = 0.44, 95% CI 0.22–0.62).

Replication of the analyses with 3-month change scores confirmed these findings (data not shown).

External responsiveness. The health transition item indicated that the majority of the patients judged their health somewhat (30.2%) or much (30.8%) improved after 12 months of anti-TNF treatment. The remainder judged their health about the same (21.9%), somewhat worse (14.8%), or much worse (2.4%).

Results of the ROC analyses are shown in Table 2 and Figure 1. The AUC were quite similar in the dimensions physical function, pain, psychological function, and general health, and no significant differences were found between the SF-36 and disease-specific measures. Differences were more pronounced in the social function dimension. Comparison of the AUC of the SF-36 and the AIMS2 showed significant differences. The SF-36 had higher accuracy than the AIMS2 to distinguish improvers from non-improvers (difference in AUC = 0.15, 95% CI 0.04–0.27, p = 0.01). Results of the correlation analyses (Table 2) confirmed these differences between the SF-36 and the AIMS2. Only the AIMS2 social interaction scale was not significantly correlated with the health transition item.

DISCUSSION

This longitudinal observational study among patients with RA who were starting anti-TNF treatment showed comparable internal and external responsiveness of the generic SF-36 compared with the disease-specific AIMS2 and HAQ within the domains physical function, pain, and psychological function. In the social function domain the SF-36 was more responsive than the AIMS2. In the general health domain the SF-36 was less responsive (just internal) than the AIMS2 and the VAS-GH.

We followed the suggestion of Husted, *et al* and differentiated between internal and external responsiveness[9].

Table 2. Responsiveness statistics for SF-36 and disease-specific meaures.

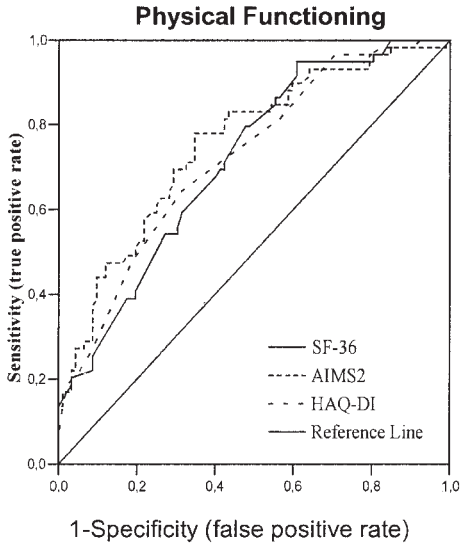
Health Domain	Internal Responsiveness SRM (95% CI)	External Responsiveness AUC (95% CI)	External Responsiveness Spearman's rho
Physical function, n = 151			
SF-36 physical function	0.75 (0.59–0.94)	0.72 (0.64–0.81)	0.48 (0.34–0.59)*
AIMS2 physical function	0.61 (0.45–0.77)	0.75 (0.68–0.83)	0.51 (0.38–0.62)*
HAQ-DI	0.59 (0.37–0.75)	0.72 (0.64–0.81)	0.52 (0.40–0.63)*
Pain, n = 167			
SF-36 bodily pain	0.87 (0.68–1.04)	0.75 (0.67–0.81)	0.45 (0.32–0.56)*
AIMS2 symptom	0.95 (0.76–1.16)†	0.77 (0.70–0.83)	0.50 (0.38–0.61)*
NRS pain	0.75 (0.57–0.93)†	0.71 (0.64–0.78)	0.38 (0.24–0.50)*
Social function, n = 161			
SF-36 social function	0.49 (0.32–0.69)†	0.69 (0.61–0.76)†	0.33 (0.18–0.46)*
AIMS2 social interaction	0.20 (0.02–0.36)†	0.54 (0.46–0.62)†	0.07 (–0.09–0.22)
Psychological function, n = 158			
SF-36 mental function	0.46 (0.28–0.61)	0.68 (0.60–0.75)	0.33 (0.18–0.46)*
AIMS2 affect	0.50 (0.35–0.65)	0.71 (0.63–0.78)	0.36 (0.21–0.48)*
General health, n = 164			
SF-36 general health	0.24 (0.06–0.39)†‡	0.69 (0.61–0.76)	0.39 (0.25–0.51)*
AIMS2 general health	0.67 (0.51–0.85)†	0.75 (0.68–0.81)	0.43 (0.29–0.55)*
VAS-GH	0.68 (0.50–0.84)‡	0.75 (0.67–0.81)	0.45 (0.32–0.56)*

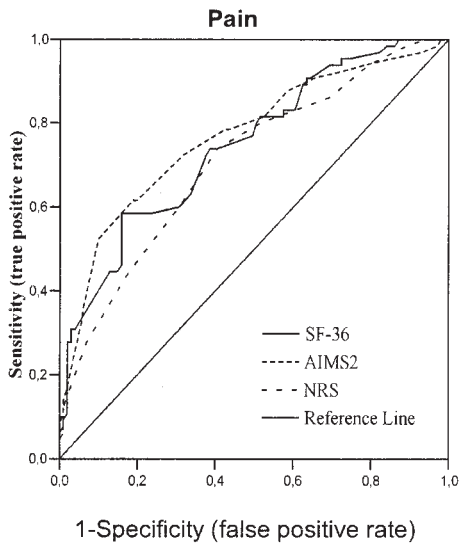
SRM: standardized response mean; AUC: area under the curve. †,‡ significant difference between measures; * $p \leq 0.01$. For definitions of measures, see legend to Table 1.

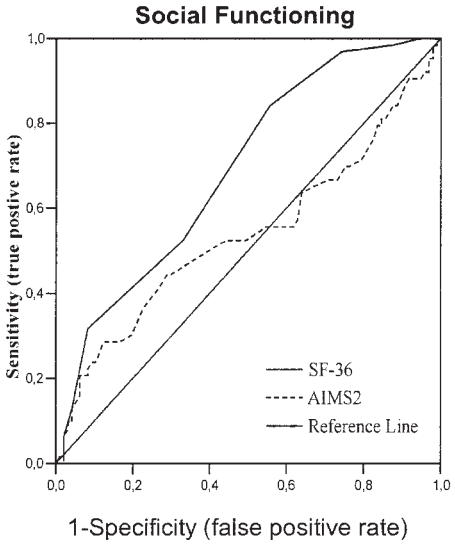
Internal responsiveness, evaluated with the SRM, describes the ability of the measures to detect improvement in HRQOL after 12 months of anti-TNF treatment. The absolute value of the SRM is sample-dependent. This means that the SRM is dependent on the effectiveness of treatment and the variation in change scores. The lowest SRM scores were found in the dimensions social function and psychological function. This may suggest lack of responsiveness of these scales to detect changes in psychological and social function. On the other hand, anti-TNF treatment may have less influence on psychosocial function than on physical function and pain. The responsiveness of these scales needs to be investigated in more detail. External responsiveness describes the relationship between change in the measures and change in an external standard. In contrast to internal responsiveness, external responsiveness is not sample-dependent, but is dependent on the external criterion for judging clinical change. In the absence of a gold standard, we used a self-reported health transition item as external criterion of change, as suggested by Fortin, *et al*[39]. A health transition question describes the magnitude and direction of change in health status over a given time period. The use of self-reported change in health status as indicator of clinical change limits the value of our results. The judgment of change is difficult for the patient and may be determined by psychological factors (e.g., mood, expectations) and current health state[40,41]. On the other hand, self-reported change in health status is a widely accepted external criterion in the evaluation of the responsiveness of HRQOL measures. It has been used in a number of studies and conditions, including rheumatologic conditions[9,42-51]. Self-reported change in health status takes into account the patients' perspective, which is a main focus of HRQOL measures, and is more likely to correlate with HRQOL measures compared with clinical variables[52,53].

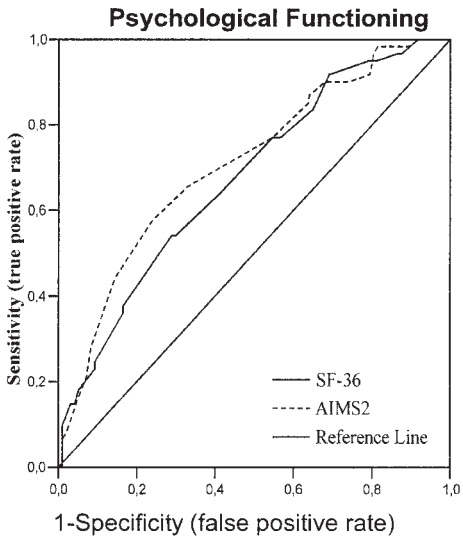
To assess internal and external responsiveness, we used different indices of responsiveness. All methods produced a consistent ranking of the comparative responsiveness of the measures within each domain of health, except for the physical function domain. This means that all methods indicated the same measure as most or least responsive. We found differences, however, in the magnitude of the differences between the measures within a health domain across the indices of responsiveness. In the general health domain, significant differences were found in internal responsiveness between the SF-36, the AIMS2, and the VAS-GH. These differences did not appear using external indices of responsiveness. The same applied to significant differences found in internal responsiveness between the AIMS2 pain scale and the NRS for pain. These results support the conclusion of previous studies that the magnitude of responsiveness is highly dependent on methodological issues such as the definition of responsiveness (e.g., internal versus external responsiveness or general change versus clinically important change), the method to assess responsiveness, the external criterion of change, the study sample, and the effectiveness of the treatment[9,10,54]. Therefore, the absolute values of responsiveness indices cannot be easily compared across studies and should be interpreted with caution.

Our study is one of the first to investigate the comparative responsiveness of the SF-36, the AIMS2, and the HAQ in a cohort of patients receiving a treatment of proven effi-









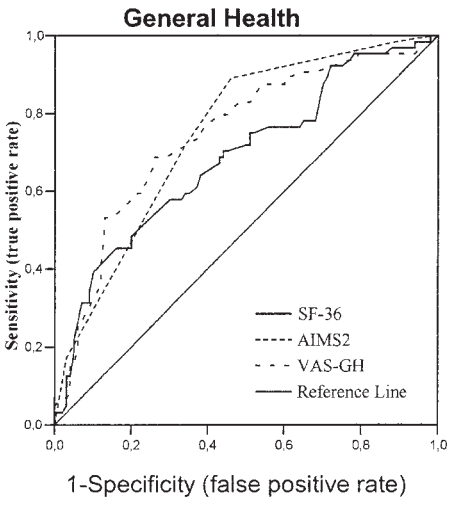


Figure 1. ROC curves for SF-36 and disease-specific measures using the health transition item of the SF-36 as the external criterion. According to this criterion, patients were classified as improved or non-improved after 12-months of anti-TNF treatment.

cacy. Anti-TNF agents have been shown to improve HRQOL in RA patients[55-59]. Most previous studies did not specifically aim at changes after an intervention of known efficacy but followed a group of patients over time[6,13,14]. Changes in HRQOL were less pronounced in these studies, which used disease activity (mostly self-reported) as the external criterion to distinguish patients whose health situation did not change from patients whose situation did improve or deteriorate. Results, which were presented for each subgroup separately, corresponded to our findings with regard to the comparative responsiveness of the measures. However, information on the dimensions social function[6,13] and general health[6,13,14] was not included in these previous studies. Because of differences in methodology, the absolute responsiveness values in the studies cannot be compared with our values. One previous study also focused on the responsiveness of the SF-36 versus a disease-specific instrument following an intervention of proven efficacy[12]. Wells, *et al* investigated the responsiveness of the SF-36 and the HAQ in patients starting methotrexate therapy[12]. In contrast to our findings, they reported a moderate SRM for the SF-36 and a large SRM for the HAQ. These findings, however, were based on a small sample size, and the statistical significance of the difference was not reported. Moreover, they reported on the physical component summary score of the SF-36 only, and not on the physical scale score.

In our study, neither the generic nor the disease-specific instrument was consistently the most responsive measure within the 5 dimensions of health. The assumption that disease-specific measures are more responsive to detect improvements due to RA-specific interventions is not confirmed by our data. The choice for the generic SF-36 or the disease-specific AIMS2 and HAQ depends among other things on the health domain one is interested in. For most purposes the SF-36 is a suitable evaluation instrument. However, if general health is the primary domain of interest, the AIMS2 and VAS-GH are preferred above the SF-36. Moreover, if a specific aspect of physical function, such as arm and hand function, is the primary domain of interest, the AIMS2 and the HAQ are recommended. A disadvantage of the SF-36 is that the physical scale may not reveal all aspects of physical health relevant to arthritis patients. For instance, only few activities related to upper extremity function are included[20]. So, besides the health domain of interest, the specific concepts measured within a health domain should be considered when choosing between the generic SF-36 and the disease-specific AIMS2 and HAQ.

Our study showed comparable internal and external responsiveness of the generic SF-36 in comparison with the disease-specific AIMS2 and HAQ within the physical function, pain, and psychological function domains. In the social function and general health domains the SF-36 was, respectively, more and less responsive than the disease-specific measures. The hypothesis that disease-specific measures are more responsive to detect interventions-related changes over time is not confirmed by our study.

ACKNOWLEDGMENT

We thank T. van Gaalen, W. Kievit, and P. Welsing for their contribution to the organization of the study and data management. We thank the following rheumatologists and research nurses for their assistance in patient recruitment and data collection: J. Alberts, C. Allaart, A. ter Avest, P. Barrera Rico, T. Berends, H. Bernelot Moens, K. Bevers, C. Bijkerk, A. van der Bijl, J. de Boer, A. Boonen, E. ter Borg, E. Bos, B. Botha, A. Branten, F. Breedveld, H. van den Brink, J. Bürer, G. Bruyn, H. Cats, M. Creemers, J. Deenen, C. De Gendt, K. Drossaers-Bakker, A. van Ede, A. Eijsbouts, S. Erasmus, M. Franssen, I. Geerdink, M. Geurts, E. Griep, E. de Groot, C. Haagsma, H. Haanen, J. Harbers, A. Hartkamp, J. Haverman, H. van Heereveld, van de Helm-van Mil, I. Henkes, S. Herfkens, M. Hoekstra, K. van de Hoeven, D.M. Hofman, M. Horbeek, F. van den Hoogen, P.M. Houtman, T. Huizinga, H. Hulsmans, P. Jacobs, T. Jansen, M. Janssen, M. Jeurissen, A. de Jong, Z. de Jong, M. Kleine Schaar, G. Kloppenburg, H. Knaapen, P. Koelmans, M. Kortekaas, B. Kraft, A. Krol, M. Kruijssen, D. Kuiper-Geertsma, I. Kuper, R. Laan, J. van de Laan, J. van Laar, P. Lanting, H. Lim, S. van der Linden, A. Mooij, J. Moolenburgh, N. Olsthoorn, P. van Oijen, M. van Oosterhout, J. Oostveen, P. van 't Pad Bosch, K. Rasing, K. Ronday, D. de Rooij, L. Schalkwijk, P. Seys, P. de Sonnaville, A. Spoorenberg, A. Stenger, G. Steup, W. Swen, J. Terwiel, M. van der Veen, M. Veerkamp, C. Versteegden, H. Visser, C. Vogel, M. Vonk, H. Vonkeman, A. Westgeest, H. van Wijk, N. Wouters.

REFERENCES

1. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med Care 1989;27 Suppl:S217-32.
2. Guyatt GH. A taxonomy of health status instruments. J Rheumatol 1995;22:1188-90.
3. Ware JE, Sherbourne CD. The MOS 36-item short-form health status survey (SF-36). 1. Conceptual framework and item selection. Med Care 1992;30:473-83.
4. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. Arthritis Rheum 1992;35:1-10.
5. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
6. Hagen KB, Smedstad LM, Uhlig T, Kvien TK. The responsiveness of health status measures in patients with rheumatoid arthritis: Comparison of disease-specific and generic instruments. J Rheumatol 1999;26:1474-80.
7. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171-8.
8. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992;45:1341-5.
9. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459-68.
10. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Qual Life Res 2003;12:349-62.
11. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes 2005;3:23.
12. Wells G, Boers M, Shea B, et al. Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis: preliminary findings in the generic health OMERACT study.

J Rheumatol 1999;26:217-21.
13. Haavardsholm EA, Kvien TK, Uhlig T, Smedstad LM, Guillemin F. A comparison of agreement and sensitivity to change between AIMS2 and a short form of AIMS2 (AIMS2-SF) in more than 1,000 rheumatoid arthritis patients. J Rheumatol 2000;27:2810-6.
14. Salaffi F, Stancati A, Carotti M. Responsiveness of health status measures and utility-based methods in patients with rheumatoid arthritis. Clin Rheumatol 2002;21:478-87.
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. Aaronson NK, Muller M, Cohen PD, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol 1998;51:1055-68.
17. Ware JE Jr, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. J Clin Epidemiol 1998;51:903-12.
18. Ware JE, Kosinski M, Keller SD. SF-36 physical and mental summary scales: a user's manual. Boston: The Health Institute, New England Medical Center; 1994.
19. Koh ET, Leong KP, Tsou IYY, Lim VH, Pong LY. The reliability, validity and sensitivity to change of the Chinese version of SF-36 in oriental patients with rheumatoid arthritis. Rheumatology Oxford 2006;45:1023-8.
20. Kvien TK, Kaasa S, Smedstad LM. Performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. II. A comparison of the SF-36 with disease-specific measures. J Clin Epidemiol 1998;51:1077-86.
21. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). Br J Rheumatol 1998;37:425-36.
22. Loge JH, Kaasa S, Hjermstad MJ, Kvien TK. Translation and performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. I. Data quality, scaling assumptions, reliability, and construct validity. J Clin Epidemiol 1998; 51:1069-76.
23. Talamo J, Frater A, Gallivan S, Young A. Use of the short form 36 (SF36) for health status measurement in rheumatoid arthritis. Br J Rheumatol 1997;36:463-9.
24. Tuttleman M, Pillemer SR, Tilley BC, et al. A cross-sectional assessment of health status instruments in patients with rheumatoid arthritis participating in a clinical trial. J Rheumatol 1997; 24:1910-5.
25. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Transition questions to assess outcomes in rheumatoid arthritis. Br J Rheumatol 1993;32:807-11.
26. Riemsma RP, Taal E, Rasker JJ, Houtman PM, Van Paassen HC, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. Br J Rheumatol 1996;35:755-60.
27. Taal E, Rasker JJ, Riemsma RP. Sensitivity to change of AIMS2 and AIMS2-SF components in comparison to M-HAQ and VAS-pain. Ann Rheum Dis 2004;63:1655-8.
28. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167-78.
29. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: dimensions and practical applications. Health Qual Life Outcomes 2003;1:20.
30. Siegert CE, Vleming LJ, Vandenbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305-9.
31. van der Heijde DM, van Riel PL, van de Putte LB. Sensitivity of a Dutch Health Assessment Questionnaire in a trial comparing hydroxychloroquine vs. sulphasalazine. Scand J Rheumatol 1990;19:407-12.
32. Stucki G, Liang MH, Stucki S, Bruhlmann P, Michel BA. A self-administered rheumatoid arthritis disease activity index (RADAI) for epidemiologic research. Psychometric properties and correlation with parameters of disease activity. Arthritis Rheum 1995;38:795-8.
33. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992;30:917-25.
34. Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995;22:1191-2.
35. Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
36. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Statistician 1983;37:36-48.
37. Stratford PW, Kennedy DM. Does parallel item content on WOMAC's pain and function subscales limit its ability to detect change in functional status? BMC Musculoskelet Disord 2004;5:17.
38. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839-43.
39. Fortin PR, Stucki G, Katz JN. Measuring relevant change: an emerging challenge in rheumatologic clinical trials. Arthritis Rheum 1995;38:1027-30.
40. 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.
41. Guyatt G, Norman G, Juniper E, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002;55:900-8.
42. Brazier JE, Harper R, Munro J, Walters SJ, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. Rheumatology Oxford 1999;38:870-7.
43. Garratt A, Ruta DA, Abdalla M, Russell I. SF-36 health survey questionnaire: II. Responsiveness to changes in health status in four common clinical conditions. Qual Health Care 1994;3:186-92.
44. Garratt AM, Hutchinson A, Russell I. The UK version of the Seattle Angina Questionnaire (SAQ-UK): reliability, validity and responsiveness. J Clin Epidemiol 2001;54:907-15.
45. Haywood KL, Garratt AM, Jordan K, Dziedzic K, Dawes PT. Disease-specific, patient-assessed measures of health outcome in ankylosing spondylitis: reliability, validity and responsiveness. Rheumatology Oxford 2002;41:1295-302.
46. Garratt AM, Ruta DA, Abdalla MI, Russell IT. Responsiveness of the SF-36 and a condition-specific measure of health for patients with varicose veins. Qual Life Res 1996;5:223-34.
47. Locker D, Jokovic A, Clarke M. Assessing the responsiveness of measures of oral health-related quality of life. Community Dent Oral Epidemiol 2004;32:10-8.
48. Escobar A, Quintana JM, Bilbao A, Aróstegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. Osteoarthritis Cartilage 2007;15:273-80.
49. Beaton DE, 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.
50. Stavem K, Frøland SS, Hellum KB. Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS. Qual Life Res 2005;14:971-80.
51. Hofer S, Benzer W, Schüssler G, von Steinbüchel N, Oldridge NB. Health-related quality of life in patients with coronary artery disease treated for angina: Validity and reliability of German translations of two specific questionnaires. Qual Life Res

2003;12:199-212.
52. Eichler HG, Mavros P, Geling O, Hunsche E, Kong S. Association between health-related quality of life and clinical efficacy endpoints in rheumatoid arthritis patients after four weeks treatment with anti-inflammatory agents. Int J Clin Pharmacol Ther 2005;43:209-16.
53. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 2000;43:1478-87.
54. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239-46.
55. Lipsky PE, van der Heijde DM, St. Clair EW, et al. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. N Engl J Med 2000;343:1594-602.
56. Keystone EC, Kavanaugh AF, Sharp JT, et al. Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy: a randomized, placebo-controlled, 52-week trial. Arthritis Rheum 2004;50:1400-11.
57. Klareskog L, van der Heijde D, de Jager JP, et al. Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. Lancet 2004;363:675-81.
58. Maini RN, Breedveld FC, Kalden JR, et al. Sustained improvement over two years in physical function, structural damage, and signs and symptoms among patients with rheumatoid arthritis treated with infliximab and methotrexate. Arthritis Rheum 2004; 50:1051-65.
59. Heiberg MS, Nordvag BY, Mikkelsen K, et al. The comparative effectiveness of tumor necrosis factor-blocking agents in patients with rheumatoid arthritis and patients with ankylosing spondylitis: a six-month, longitudinal, observational, multicenter study. Arthritis Rheum 2005;52:2506-12.