Relative Responsiveness of Physician/Assessor-Derived and Patient-Derived Core Set Measures in Rheumatoid Arthritis Trials

TUHINA NEOGI, HUI XIE, and DAVID T. FELSON

ABSTRACT. ***Objective.*** We assessed whether individual American College of Rheumatology core set measures (CSM), and the CSM grouped as composite patient-derived (CPD) or composite physician/assessor-derived (CMD), performed differently in rheumatoid arthritis (RA) clinical trials.

Methods. We used data from 9 RA trials [anti-tumor necrosis factor-α (TNF-α) and disease modifying antirheumatic drug (DMARD)] in which CSM had been assessed, conducted from the early 1990s to present, with a total of 2969 patients. We grouped the CSM as CPD (pain, patient global assessment, function) and CMD [tender joint count (TJC), swollen joint count (SJC), physician global, inflammatory marker]. Using bootstrap simulation, we estimated the sample size that would be required to distinguish active treatment from placebo with the Wilcoxon rank-sum test in the clinical trials for the outcomes of percentage change of each individual CSM, of the Disease Activity Score (DAS), and average percentage change of the CMD or of the CPD.

Results. Comparing the performance of individual CSM relative to one another, the physician and patient global assessments and TJC would require the lowest sample sizes to distinguish active treatment from placebo, while use of the SJC, inflammatory marker, and function would require the highest. The CMD performed similarly to the DAS, requiring similar sample sizes, while the CPD would require 1.7 times greater sample size to distinguish treatment from placebo. The results were similar across DMARD and anti-TNF-α trials.

Conclusion. Because of their demonstrated sensitivity to change, composite measures assessing RA outcomes in clinical trials should continue to include physician/assessor-derived core set measure assessments. (First Release April 15 2008; J Rheumatol 2008;35:757–62)

Key Indexing Terms:
CLINICAL TRIALS RHEUMATOID ARTHRITIS OUTCOME MEASURES

Outcomes of rheumatoid arthritis (RA) clinical trials are assessed by means of composite indices such as the American College of Rheumatology (ACR) core data set[1] and the Disease Activity Score (DAS)[2]. The ACR core set measures include tender joint count (TJC), swollen joint count (SJC), physician global assessment, patient-reported pain, patient-reported function, patient global assessment, and an inflammatory marker, either erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP). The Disease Activity Score (DAS) includes TJC, SJC, patient global assessment, and an inflammatory marker (ESR or CRP). Thus, these composite measures include physician/assessor-derived assessments, patient-derived assessments, and a laboratory measure.

Of the core set measures, those that are patient-derived have been advocated as better predictors of longterm outcomes like disability and death[3-7], and as a better tool to document the longterm course of RA in the clinic[8]. One study has suggested that single measures assessed by patients, such as severity of pain and global assessment, are at least as sensitive to change in trials as such physician/assessor measures as joint counts[9]. However, composite measures are not the same as individual core set items, and the relative sensitivity to change of such composite measures may not parallel those of their specific building-block items. Composite measures combining patient-derived and physician/assessor-derived elements from the core set have rarely been compared for their sensitivity to change[10,11]. These reports, one limited to 3 trials testing anakinra and the other to one trial of methotrexate and leflunomide, suggested equivalence or superiority of the patient-derived meas-

From the Clinical Epidemiology Research and Training Unit, Boston University School of Medicine, Boston, Massachusetts; and the Department of Biostatistics, University of Illinois at Chicago, School of Public Health, Chicago, Illinois, USA.

Supported by NIH AR47785 and a grant to reevaluate response criteria from the American College of Rheumatology. Dr. Neogi was supported by the Abbott Scholar Award in Rheumatology and the Arthritis Foundation Postdoctoral Fellowship Award. She is currently supported by the ACR-REF ASP Junior Career Development Award in Geriatric Medicine, Arthritis Foundation Arthritis Investigator Award, and NIH 1K23AR055127.

T. Neogi, MD, FRCPC, Clinical Epidemiology Research and Training Unit, Boston University School of Medicine; H. Xie, PhD, Department of Biostatistics, University of Illinois at Chicago, School of Public Health; D.T. Felson, MD, MPH, Clinical Epidemiology Research and Training Unit, Boston University School of Medicine, Boston.

Address reprint requests to Dr. T. Neogi, Boston University School of Medicine, Clinical Epidemiology Unit, Suite X-200, 650 Albany Street, Boston, MA 02118.

Accepted for publication December 27, 2007.

ures compared to physician/assessor-derived measures. The implication of this finding would be to place a greater emphasis on composite patient-derived outcomes not only in clinical practice, but also in clinical trials. However, the relative performance of the composite patient-derived versus composite physician/assessor-derived core set measures has not been evaluated in a larger number of patients from trials, or in a comprehensive study that includes trials of a wide spectrum of agents. Nor have the implications for clinical trial design with respect to required sample size been evaluated with such a strategy.

In a recent reevaluation of response definitions for RA, we assembled a large dataset of recently completed multicenter RA trials; this gave us the opportunity to compare the sensitivity to change of patient-derived versus physician/assessor-derived composite measures in RA. We investigated whether a composite of patient-derived (CPD) core set measures performed similarly to a composite of physician/assessor-derived (CMD) core set measures in its ability to distinguish efficacious RA therapies.

MATERIALS AND METHODS

Participants and dataset. We used data assembled as part of the ACR effort to reevaluate improvement criteria in RA[12]. This dataset included 11 large, multicenter randomized RA trials with core set outcome measures assessed, conducted from the early 1990s to the present. For 2 of the trials used for the ACR effort, data were not provided to us, precluding use of these trials in this study. The trials we included were placebo-controlled trials of disease modifying antirheumatic drugs (DMARD), different anti-tumor necrosis factor-α (TNF-α) inhibitors, and a non-anti-TNF-α-inhibiting biologic agent. There was also one comparative trial of a combination of agents versus a single agent. In total, there were 2969 participants (1976 participants in 7 anti-TNF-α inhibitors vs placebo arms, and 993 in 4 DMARD vs placebo or single-agent arms). Based on agreement with sponsors providing the data, we have not provided identifiable trial information. Sponsors provided data either on all patients or on a randomly selected 80% of participants in the trial. Data included the core set measure scores at each study visit and treatment assignment code. Individual participant information other than treatment assignment and ACR core set measures were not released from these trials in developing this dataset. As these were all randomized trials, any known or potential confounders are therefore expected to be equally distributed among the treatment arms.

ACR core set measures, composite measures, and DAS. We assessed individual ACR core set measures separately as well as composite measures, which were defined as ACR core measures grouped as follows[11]: (1) physician/assessor-derived (CMD), which consisted of TJC, SJC, and physician global assessment; (2) physician/assessor- plus laboratory-derived (CMD+lab), which consisted of CMD plus an inflammatory marker (ESR or CRP); and (3) patient-derived (CPD), which consisted of patient-reported pain, patient-reported function, patient global assessment (Table 1). The DAS with the 28-joint count (DAS28) was used as a comparator[13]. For trial data that used CRP, the DAS28-4(crp) was used[14]. Calculations are provided in the Appendix. We did not use the ACR20 as a comparator since it is not a continuous measure.

Statistical analysis. Our overall purpose was to evaluate the ability of the composite measures (CMD, CMD+lab, CPD), and the DAS to distinguish active treatment from placebo. One approach to determining how well an outcome measure performs in distinguishing active treatment from placebo is to estimate the sample size that would be required for the outcome measure to detect a difference between treatment arms. This would be determined by the standardized magnitude of the difference between treatment and placebo, which, in turn, is influenced by the outcome measure's sensitivity to change and variability in the measure.

To distinguish active treatment from placebo in terms of disease improvement, we first calculated the subject-specific improvements using data from each of the trials. For the 7 ACR core set measures and the DAS, improvement was defined as the percentage change from baseline. For the composite measures (CMD, CMD+lab, CPD), improvement was defined as the average percentage change from baseline[11]. We then applied the Wilcoxon rank-sum test to the resultant data and evaluated whether the null hypothesis of equal efficacy of active treatment and placebo was rejected. The Wilcoxon test, a nonparametric test, has the advantage of being robust to distributional assumptions. To evaluate power and sample size requirements for the Wilcoxon rank-sum test using the improvement definitions described above, we conducted nonparametric bootstrap simulations[15] (random sampling with replacement) as follows: for each clinical trial, we resampled with replacement a "K" number of subjects from the original trial data. This process generates one simulated dataset with "K" subjects drawn from the same population as the original data for this trial. We repeated this process of resampling "n" times, which gives rise to "n" simulated datasets for a particular trial with sample size being varied to "K." For each of these "n" simulated datasets, we then applied the Wilcoxon rank-sum test based on the improvement definitions of each measure and determined whether or not the null hypothesis was rejected. The power for sample size "K" was then calculated as the proportion of "n" simulated datasets in which the null hypothesis was rejected. The value of "K" that had a corresponding power of 80% was the required sample size for the improvement definition for a particular measure. We used this same procedure to obtain the required sample sizes for each definition of improvement of the measures (ACR core set, composite, and DAS). To ensure that the results were robust and were not specific to drug type, we also calculated the sample sizes grouped by type of trial, either DMARD or anti-TNF-α.

To put these sample sizes into context, we also estimated the sample size that would be required for a "gold standard" test procedure using all of the core set measures with optimal weighting based on each measure's variability[16]. O'Brien's test was used for testing whether multiple outcomes in one treatment group had consistently larger values than the outcomes in the other treatment group. In O'Brien's test, subjects in each trial are ranked from smallest to largest percentage change for each core set item separately, and for each subject, these ranks are summed across core set measures and then compared[17]. Therefore, although O'Brien's global test statistic reduces the 7 core set measures into a single outcome, its rank-based definition means that one subject's value is dependent on the values of all other subjects, the weights for the core set items change from one dataset to the next. As a result, O'Brien's test does not produce an appropriate definition of response and the summed rank values are not clinically interpretable, although some effort has been made to overcome these issues[18]. We use it here only because it has excellent power to detect differences and can serve as a benchmark for other improvement definitions.

Estimated sample sizes required for each outcome across trials or subgroup of trials were combined using an average weighted by the size of each trial and are reported as a relative ratio to the gold-standard sample size, which is assigned a value of 1. For example, a sample size reported as having a ratio of 2.0 means that use of that measure as a trial's primary outcome would require double the sample size compared to the gold standard.

RESULTS

In comparing the performance of individual core set measures relative to one another, the physician and patient global assessments and TJC would require the lowest sample sizes to distinguish active treatment from placebo, while use of the SJC, inflammatory marker, and function would require the highest (Table 2). For example, if a trial were to

Table 1. Core set and composite measures.

ACR Core Dataset of 7 Disease Activity Measures for RA	Components of Core Set Measures Used for Physician/Assessor-derived Core Set Measures (CMD)	Physician/Assessor-plus Lab-Derived Core Set Measures (CMD+lab)	Patient-Derived Core Set Measures (CPD)	Disease Activity Score (DAS)
Tender joint count	✓	✓		✓
Swollen joint count	✓	✓		✓
Physician global assessment	✓	✓		
Inflammatory marker		✓		✓
Patient-reported pain			✓	
Patient-reported functional disability (HAQ)			✓	
Patient global assessment			✓	✓

Table 2. Relative sample sizes required for individual core set measures versus the O'Brien-derived sample size.

	Relative Ratio of the Estimated Sample Size Required to Distinguish Active from Placebo		
	All Trials	DMARD Trials	Anti-TNF-α Trials
"Gold standard" from O'Brien test	1.0	1.0	1.0
Physician global assessment	1.6	1.3	1.7
Patient global assessment	1.6	1.4	1.7
Swollen joint count	2.2	2.5	2.0
Tender joint count	1.7	1.7	1.7
Pain	1.8	1.8	1.8
Functional disability	2.1	2.4	2.0
Inflammatory marker (generally CRP)	2.4	3.1	1.9

use physician global assessment as the outcome measure, the estimated required sample size would be 1.6 times that of the gold standard (O'Brien's test), while for patient-reported function, the estimated required sample size would be 2.1 times that of the gold standard.

The results were similar across DMARD and anti-TNF-α trials, with the exception of the inflammatory marker and swollen joint count. In DMARD trials, using an inflammatory marker as the outcome measure would require 3.1 times the sample size of the gold standard, while in anti-TNF-α trials, it would require 1.9 times the sample size (Table 2). The differences in sample sizes required for swollen joint count were smaller. The differences in sample sizes between the trial types may speak to differences between drug mechanisms and to the fact that the group of DMARD trials included a variety of agents with differing efficacy.

In the grouped composite measures, the CMD required a sample size of 1.2 times the gold standard, while the DAS required 1.3 times the sample size of the gold standard. The CMD+lab required only 1.1 times the gold-standard sample size to distinguish active treatment from placebo. Although the patient global assessment on its own performed well, the CPD would require on average 1.7 times greater sample size than the gold standard to distinguish treatment from placebo (Table 3). As a hypothetical example, if a trial used the gold standard (O'Brien's test) as its outcome and needed to enroll 200 participants to have an 80% likelihood of detecting a significant difference between active treatment and placebo, using CMD as the outcome would require 240 participants, using CMD+lab would require 220 participants, and using CPD would require 340 participants to detect the difference.

As in the individual core set measures, results of these composite measures were similar across the 2 types of trials, DMARD and anti-TNF-α (Table 3).

DISCUSSION

Like other investigators, we found that the individual core set measures that were least sensitive to change were SJC and Health Assessment Questionnaire (HAQ) disability, while both physician and patient global assessments performed well. Composite physician/assessor-derived (CMD or CMD+lab) outcome measures performed equivalently or slightly better than patient-derived (CPD) ones in distinguishing active treatment from placebo in the 9 large randomized trials we analyzed. Further, this held true for both types of trials, DMARD and anti-TNF-α, suggesting that both physician/assessor- and patient-derived composite measures perform similarly regardless of therapy tested. Interestingly, the inflammatory marker itself did not appear to add to the ability to distinguish active treatment from placebo, even though all trials had average CRP values at baseline that were greater than 2 mg/dl (ranging from 2.2 to 5.3 mg/dl), indicating elevated levels in these trial participants.

Table 3. Relative sample sizes required for composite measures versus the O'Brien-derived sample size.

	Relative Ratio of the Estimated Sample Size Required to Distinguish Active from Placebo		
	All Trials	DMARD Trials	Anti-TNF-α Trials
"Gold standard" from O'Brien test	1.0	1.0	1.0
Disease Activity Score	1.2	1.3	1.1
CMD (physician/assessor-derived)	1.3	1.3	1.3
CMD+lab (physician/assessor- plus lab-derived)	1.1	1.2	1.0
CPD (patient-derived)	1.7	1.9	1.6

These results based on a number of clinical trials with a large number of persons with RA indicate that patient-derived composite measures are not more efficient as an outcome measure than physician/assessor-derived measures in attempting to distinguish effective therapy from placebo. The comparable performance of both composite measures could be ascribed to less change occurring with some of the individual measures, less precision of the measures, or a combination of factors. Studies have shown that the SJC is not among the core set measures with the greatest sensitivity to change[19-21]. However, it is widely advocated as a centrally important measure of patient status in RA[22].

Despite inclusion of SJC and the variably sensitive laboratory measure, composite physician/assessor-derived measures were as sensitive to change as, if not more sensitive than, patient-derived measures, even though patient global assessment performed well. Why is this? One reason is that another core set measure with less sensitivity to change was the HAQ. The HAQ may be poorly sensitive to change because subjects in these trials had longstanding disease, with average disease duration at study entry > 6 years for all trials, with several being 11–13 years, and had fixed functional loss, reflected by average baseline HAQ values ranging from 1.5 to 1.8, with only one study having baseline HAQ of 0.8–0.9[23-25]. Another reason for our findings has to do with the correlation of measures. An index is more sensitive to change when its precision is better than that of its individual components, which can be achieved in part by diminishing the variability or noise of its change. One way indices accomplish this is by combining measures that correlate with one another modestly[26]. The patient-derived measures — patient global assessment, HAQ, and pain — correlate highly with one another[27,28], suggesting that they are not fully independent measures. Indeed, in one study, all 8 subdimensions of the HAQ were explained by pain, suggesting that the HAQ and pain are measuring similar constructs[29]. Items that are too highly correlated fail to add additional information to an index. On the other hand, while SJC and TJC are highly correlated, physician global assessment does not load on the same factor[1], suggesting that physician global assessment is not strongly correlated with joint counts, and laboratory measures are, at best, modestly correlated with these other 3 factors. Thus, although SJC is not very sensitive to change, the covariate structure of the physician/assessor-derived assessment measure ensures that an index of these measures will, overall, be sensitive to change. The DAS also performed well and this may reflect the higher weights assigned to SJC, TJC, and the inflammatory marker in the DAS (Appendix), with the inflammatory marker not correlating highly with the joint counts. Thus, relying solely on individual core set measures or on parts of composite measures may sacrifice sensitivity to change. This has implications for trial design for the number of patients that need to be recruited into a trial.

In comparison to traditional dichotomous-based threshold outcome measures, such as the ACR20, continuous outcome measures have greater statistical power and sensitivity to change. This is reflected in the current ACR recommendation to use the recently developed ACR Hybrid, a continuous measure, as the preferred outcome measure in RA trials[12]. As expected, when we performed additional analyses to evaluate the relative sensitivity to change of the ACR20 measure, the relative sample size required would have been 2.1, indicating that it has less sensitivity to change than either composite patient-derived or physician/assessor-derived measures (both continuous measures), while the continuous ACR Hybrid measure performed similarly to the composite physician/assessor-derived measure (relative sample size 1.1).

Why have previous studies more readily demonstrated equivalence or even superiority of patient-reported outcomes in other trials? These previous studies generally focused on individual patient-derived measures, and measures such as patient global assessment and pain are among the most sensitive to change of outcome measures in RA[1,19,21]. In the one study that directly compared patient and physician/assessor composite measures using data from 3 anakinra trials, the reported findings were opposite to ours, that composite patient measures were more sensitive to change. This difference may be explained by peculiarities of anakinra's effects, with much less effect on SJC, a physician/assessor-derived measure, than on other core set measures[10]. Our results suggest that the anakinra findings may be treatment-specific and not generalizable. In the study comparing patient-derived measures to physician/assessor-derived measures using data from a methotrexate/lefluno-

mide trial, sensitivity to change was not directly compared statistically[11].

We were unable to assess the sensitivity to change of other potentially important patient-derived outcomes, such as fatigue, as these data were not collected/provided in the trial data analyzed. However, current research standards in RA clinical trials continue to promote the use of the ACR core set measures used in these analyses. Similarly, we were unable to assess the predictive validity of composite physician/assessor- and patient-derived core set measures for radiographic structural outcomes as these data were not available in the trial data analyzed. Finally, these results may not be applicable to trials of persons with early RA, in which patient-derived outcomes may have improved sensitivity to change since functional loss may be more sensitive to change in earlier than in later stages of disease.

Nonetheless, our results are based on a large number of RA trial participants with outcome definitions derived from the ACR core set measures that are currently widely used in RA clinical trials, using a wide variety of drugs, with similar results being noted in both anti-TNF-α inhibitor and DMARD trials, and with robust uniform analytic techniques across all trials. Of note, one of the trials used in the other studies evaluating patient-derived outcomes was also included in our current study. While in selected trials, the composite patient-derived outcome may perform at least as well as or even better than other measures, our results suggest that overall a composite patient-derived outcome is no better than a composite physician/assessor-derived measure in terms of sensitivity to change.

In summary, our study provides insight into the performance of the individual core set measures and the ability of physician/assessor-derived versus patient-derived composite measures to detect a difference between the efficacies of treatments in RA clinical trials. Patient-reported measures that are of importance in determining longterm outcomes in the clinic setting may not be sufficient for use in clinical trials on their own. Because of their demonstrated sensitivity to change, composite measures assessing RA outcomes in clinical trials should therefore continue to include physician/assessor-derived core set measure assessments.

APPENDIX

DAS28 calculation using ESR:

DAS28 = 0.56 * sqrt(TJC28) + 0.28 * sqrt(SJC28) + 0.70 * ln(ESR) + 0.014 * GH

DAS28 calculation using CRP:

DAS28-4(crp) = 0.56 * sqrt(TJC28) + 0.28 * sqrt(SJC28) + 0.36 * ln(CRP+1) + 0.014 * GH + 0.96

where GH = patient's general health measured on 100 mm visual analog scale.

REFERENCES

1. 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.
2. van der Heijde DM, van 't Hof M, van Riel PL, van de Putte LB. Development of a disease activity score based on judgment in clinical practice by rheumatologists. J Rheumatol 1993;20:579-81.
3. Callahan LF, Pincus T, Huston JW 3rd, Brooks RH, Nance EP Jr, Kaye JJ. Measures of activity and damage in rheumatoid arthritis: depiction of changes and prediction of mortality over five years. Arthritis Care Res 1997;10:381-94.
4. Pincus T, Brooks RH, Callahan LF. Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. Ann Intern Med 1994;120:26-34.
5. Pincus T, Callahan LF, Sale WG, Brooks AL, Payne LE, Vaughn WK. Severe functional declines, work disability, and increased mortality in seventy-five rheumatoid arthritis patients studied over nine years. Arthritis Rheum 1984;27:864-72.
6. Soderlin MK, Nieminen P, Hakala M. Functional status predicts mortality in a community based rheumatoid arthritis population. J Rheumatol 1998;25:1895-9.
7. Wolfe F, Zwillich SH. The long-term outcomes of rheumatoid arthritis: a 23-year prospective, longitudinal study of total joint replacement and its predictors in 1,600 patients with rheumatoid arthritis. Arthritis Rheum 1998;41:1072-82.
8. Pincus T. The DAS is the most specific measure, but a patient questionnaire is the most informative measure to assess rheumatoid arthritis. J Rheumatol 2006;33:834-7.
9. Strand V, Cohen S, Crawford B, Smolen JS, Scott DL. Patient-reported outcomes better discriminate active treatment from placebo in randomized controlled trials in rheumatoid arthritis. Rheumatology Oxford 2004;43:640-7.
10. Cohen SB, Strand V, Aguilar D, Ofman JJ. Patient- versus physician-reported outcomes in rheumatoid arthritis patients treated with recombinant interleukin-1 receptor antagonist (anakinra) therapy. Rheumatology Oxford 2004;43:704-11.
11. Pincus T, Strand V, Koch G, et al. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the Disease Activity Score (DAS) in a rheumatoid arthritis clinical trial. Arthritis Rheum 2003;48:625-30.
12. American College of Rheumatology Committee to Reevaluate Improvement Criteria. A proposed revision to the ACR20: the hybrid measure of American College of Rheumatology response. Arthritis Rheum 2007;57:193-202.
13. 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.
14. Department of Rheumatology, University Medical Centre Nijmegen, Nijmegen, The Netherlands. [Internet. Accessed February 11, 2008.] Available from: http://www.das-score.nl/www.das-score.nl/index.html
15. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
16. Anderson JJ, Bolognese JA, Felson DT. Comparison of rheumatoid arthritis clinical trial outcome measures: a simulation study. Arthritis Rheum 2003;48:3031-8.
17. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 5. Comparing two therapies with respect to several endpoints. Mayo Clin Proc 1988;63:1140-3.
18. Tilley BC, Pillemer SR, Heyse SP, Li S, Clegg DO, Alarcon GS. Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials. MIRA Trial Group. Arthritis Rheum

1999;42:1879-88.
19. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. J Rheumatol 1993;20:535-7.
20. Bombardier C, Raboud J. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. The Auranofin Cooperating Group. Control Clin Trials 1991;12 Suppl:243S-56S.
21. Gotzsche PC. Sensitivity of effect variables in rheumatoid arthritis: a meta-analysis of 130 placebo controlled NSAID trials. J Clin Epidemiol 1990;43:1313-8.
22. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. J Rheumatol 1993;20:561-5.
23. Aletaha D, Smolen J, Ward MM. Measuring function in rheumatoid arthritis: Identifying reversible and irreversible components. Arthritis Rheum 2006;54:2784-92.
24. Aletaha D, Strand V, Smolen JS, Ward MM. Treatment-related improvement in physical function varies with duration of rheumatoid arthritis: a pooled analysis of clinical trial results. Ann Rheum Dis 2008;67:238-43.
25. Aletaha D, Ward MM. Duration of rheumatoid arthritis influences the degree of functional improvement in clinical trials. Ann Rheum Dis 2006;65:227-33.
26. Streiner D, Norman G. Selecting the items. In: Health measurement scales: a practical guide to their development and use. New York: Oxford University Press; 1989.
27. Sokka T, Kankainen A, Hannonen P. Scores for functional disability in patients with rheumatoid arthritis are correlated at higher levels with pain scores than with radiographic scores. Arthritis Rheum 2000;43:386-9.
28. Ward MM. Clinical measures in rheumatoid arthritis: which are most useful in assessing patients? J Rheumatol 1994;21:17-27.
29. Hakkinen A, Kautiainen H, Hannonen P, Ylinen J, Arkela-Kautiainen M, Sokka T. Pain and joint mobility explain individual subdimensions of the Health Assessment Questionnaire (HAQ) disability index in patients with rheumatoid arthritis. Ann Rheum Dis 2005;64:59-63.