The Hawthorne Effect, Sponsored Trials, and the Overestimation of Treatment Effectiveness

FREDERICK WOLFE and KALEB MICHAUD

ABSTRACT. ***Objective.*** To determine if the results of rheumatoid arthritis (RA) clinical trials are upwardly biased by the Hawthorne effect.

Methods. We studied 264 patients with RA who completed a commercially sponsored 3-month, open-label, phase 4 trial of a US Food and Drug Administration approved RA treatment. We evaluated changes in the Health Assessment Questionnaire disability index (HAQ) and visual analog scales for pain, patient global, and fatigue during 3 periods: pretreatment in the trial, on treatment at the close of the trial, and by a trial-unrelated survey 8 months after the close of the trial, but while the patients were receiving the same treatment.

Results. The HAQ score (0–3) improved by 41.3% during the trial, but only by 16.5% when the endpoint was the post-trial result. Similar results for the other variables were patient global (0–10) 51.9% and 34.6%, pain (0–10) 51.7% and 39.7%, fatigue (0–10) 45.6% and 24.6%. Worsening between the trial end and the first survey assessment was HAQ 0.29 units, pain 0.8 units, patient global 0.8 units, and fatigue 1.1 units.

Conclusion. Almost half the improvement noted in the clinical trial HAQ score disappeared on entry to a non-sponsored followup study, and from 23% to 44% of improvements in pain, patient global, and fatigue also disappeared. These changes can be attributed to the Hawthorne effect. Based on these data, we hypothesize that the absolute values of RA outcome variables in clinical trials are upwardly biased, and that the treatment effect is less than observed. (J Rheumatol First Release Sept 15 2010; doi:10.3899/jrheum.100497)

Key Indexing Terms:
HAWTHORNE EFFECT
CLINICAL TRIALS
EFFECTIVENESS
RHEUMATOID ARTHRITIS

Although biologic therapy improves the health status of patients with rheumatoid arthritis (RA), we have observed that RA patients treated with biologics and followed in the large National Data Bank for Rheumatic Diseases (NDB) observational data bank, and patients followed in clinical practice, do not have RA outcomes that are as good as those seen in clinical trials[1,2,3,4]. This observation is somewhat unexpected because RA patients treated in the community have less severe RA than participants in clinical trials and should be expected to have better results. By "outcomes," we mean the actual levels of the clinical variables [for example, a Health Assessment Questionnaire (HAQ) score of 1.1] as distinct from the change or percentage change in the variables. However, the percentage change reported in clinical trials is also important because it can lead readers to expect that the same degree of improvement that is observed in clinical trials will also be found in clinical practice in similar patients.

RA clinical trials may be biased toward better results by several factors. Some patients may be entered into the trial at the time of a transient flare rather than when they are in a steady state. In addition, the enrollment evaluation process in itself can be biased toward the overreporting of baseline abnormalities because patients must have predefined levels of severity to enter the trial. To the extent that these biases are present, they work on both the active treatment and the control group equally, and ordinarily are accounted for by the blinding and randomization process. However, the degree of improvement, which is usually presented as American College of Rheumatology (ACR) percentage of improvement or Disease Activity Score (DAS) improvement, is generally emphasized only for the active treatment group in publications and marketing materials, where one might say, for example, that 68% of treated patients achieved an ACR 20% response. These 2 biases, regression to the mean and examination bias, can lead to exaggerated improvement for the treated and untreated groups. However, they do not lead to final values that are biased toward the less abnormal, as they affect only starting values.

There is a third bias that could result in both increased

From the National Data Bank for Rheumatic Diseases and University of Kansas School of Medicine, Wichita, Kansas; and the University of Nebraska Medical Center, Omaha, Nebraska, USA.

Kaleb Michaud received partial funding for this study from the Arthritis Foundation's New Investigator Award and NIH ARRA grant 1RC1AR058601-01.

F. Wolfe, MD, National Data Bank for Rheumatic Diseases and University of Kansas School of Medicine; K. Michaud, PhD, University of Nebraska Medical Center and National Data Bank for Rheumatic Diseases.

Address correspondence to Dr. F. Wolfe, National Data Bank for Rheumatic Diseases, 1035 N. Emporia, Suite 288, Wichita, KS 67214, USA. E-mail: fwolfe@arthritis-research.org

Accepted for publication July 15, 2010.

improvement compared with baseline and improved final values — the Hawthorne or trial effect. This bias could have important implications about the effectiveness of therapy. The Hawthorne effect was originally defined in an industrial setting as an increase in worker productivity produced by the psychological stimulus of being singled out and made to feel important[5]. Subsequently the definition has been broadened so that in medical settings it refers to treatment response rather than productivity[6]. In the clinical trial setting, the effect may be defined as the additional clinical response that results from increased attention provided by participation in the clinical trial.

The Hawthorne effect, if present, could result in RA improvement in 2 ways. First, it could result in true improvement — as in the Hawthorne effect of observed productivity in the factory — and, second, it could result in reported but not true improvement; for example, a patient might indicate by HAQ score that his function is improved, while in reality it is not. If the Hawthorne effect is present in that manner, we would expect that patients would not do as well in the subsequent clinical setting, where the intensive clinical trial attention is absent. We call the first type of Hawthorne effect Type A and the second type of Hawthorne effect Type B.

Authors who have studied the Hawthorne effect have all noted that it is extremely difficult to obtain reliable results. There is some evidence from non-RA studies that the Hawthorne effect may be present. Braunholtz, *et al* in 2001[7] performed a systematic review of the literature (mostly cancer trials) with respect to a Type A "protocol/Hawthorne effect (benefit from improved routine care within a trial)." They noted that there was "weak evidence to suggest that clinical trials have a positive effect on the outcome of participants [and concluded] that it is more likely that clinical trials have a positive rather than a negative effect on the outcome of patients. In the limited data available, the effect seems to be larger in trials where an effective treatment already exists and is included in the trial protocol."

Peppercorn, *et al* evaluated 24 published articles of outcomes among cancer patients who were enrolled and not enrolled in clinical trials. They noted that "14 comparisons provided some evidence that patients enrolled in trials have improved outcomes. However, strategies to control for potential confounding factors were inconsistent and frequently inadequate," and they concluded that "Despite widespread belief that enrollment in clinical trials leads to improved outcomes in patients with cancer, there are insufficient data to conclude that such a trial effect exists."[8]

By contrast, McCarney, *et al* found that more intensive followup of individuals in a placebo-controlled clinical trial of Ginkgo biloba for treating mild to moderate dementia resulted in a better outcome than minimal followup, as measured by their cognitive functioning[6]. The intensive group had comprehensive assessment visits at baseline and 2, 4, and 6 months postrandomization compared with an abbreviated assessment at baseline and a full assessment at 6 months.

The Hawthorne effect has not been evaluated in RA. We hypothesized that a Type B Hawthorne effect might be present. In this report we studied RA patients who received comprehensive care in a clinical trial, including free treatment, and received ordinary RA care by their non-study physicians after the trial. During followup, patients used the same questionnaires as used in the trial and paid 0–100% of the cost of their own medications, according to their insurance coverage. We compared outcomes at the end of the clinical trial with outcomes noted in the non-sponsored followup assessments to see if improvement noted in the clinical trial was maintained in clinical practice. Sponsored extension studies of RA trials usually show maintenance of improvement. The RA patients in this followup study were volunteers who had a good response to therapy.

MATERIALS AND METHODS

In order to conceal the treatment, as desired by the sponsor of the clinical trial, we have used approximate numbers of patients and dates of treatment in the next paragraph. Patients in this study had been participants in a 1500–2500 person phase 4 clinical trial of a currently marketed and US Food and Drug Administration approved RA treatment. The trial began with a double-blind period, followed by a 3-month open-label treatment period during which all patients received active treatment. Enrollment to the open-label study began after 2005 and the last patient completed the study before 2008. Physicians were compensated for their participation, and patients received the treatment without cost. There was a total of 5 study visits.

We were allowed to invite patients in the above study to participate in the National Data Bank for Rheumatic Diseases (NDB) longitudinal study of RA outcomes[9,10] after completion of the phase 4 study. The NDB is a research data bank that surveys patients by mail and Internet at 6-month intervals (January and July). However, administrative procedures significantly limited our ability to recruit for the followup NDB study, and enrollment did not begin until half of the phase 4 study was completed and easy access to patients was lost. We were not allowed to contact patients directly, but were required to contact their physicians. Physicians were not compensated, but were asked to contact their patients, explain the study, ask for their signed informed consent, and forward details to the NDB. As a result, many physicians did not participate and an unknown number of patients were contacted. From this process 264 patients consented to participate in the study followup. Patients who participated were systematically different from nonparticipants. Compared with nonparticipants, participants had increased rates of ACR20, ACR50, and ACR70 responses[11], and had lower (better) HAQ[12], pain, patient global, and fatigue scores. Patients were not compensated for the participation in the NDB.

Patients received a standard, FDA-approved dose of the study medication throughout the open label study, and this dose was continued during the NDB followup, except for 16 patients (6.1%) who discontinued the treatment. For the purposes of this study we evaluated only variables that were included in both the open-label and followup study. These variables included the HAQ disability index[12], and visual analog scales (VAS) for pain, patient global, and fatigue. The assessment formats used in both studies were the same, except that some NDB patients completed the questions on the Internet.

Statistical methods. We compared the 4 clinical variables at 3 timepoints: the start of active drug administration in the open-label trial, the final observation in the open-label trial, and the first observation in the followup NDB survey. We used t-tests to compare variables at the different timepoints, and we calculated effect sizes using the pooled standard deviation. Data were analyzed using Stata (Stata Corp., College Station, TX, USA) version 10.1.

To compare the final NDB values with values of other NDB RA patients, we identified 5686 RA patients in the NDB receiving the same class of treatment at their first NDB observation as those in the clinical

trial. We excluded patients participating in safety registries because their RA might be more severe than the RA of the average patient. By regression analysis, we adjusted their HAQ, pain, patient global, and fatigue scores to the characteristics of the study population, as noted in Table 1.

RESULTS

At the time of evaluation in the NDB surveys, the median RA duration of the 264 patients was 10.4 years (Table 1). The median duration from the last clinical trial assessment to the first post-trial survey assessment was 0.8 years. All patients received the study drug during the 6-month recall period of the survey questionnaire. However, 16 patients (6.1%) discontinued the treatment during this period. Sensitivity analyses indicated only minimal change in study results if these patients were included or excluded, and therefore we elected to include them in the study. The dose of the study drug was unchanged between the end of the study and the first post-study reporting.

All study measures improved during the trial and worsened thereafter (Table 2 and Figure 1). The HAQ score improved by 41.3% during the trial, but only by 16.5% when the endpoint was the post-trial result. Similar results for the other variables were patient global 51.9% and 34.6%, respectively, pain 51.7% and 39.7%, and fatigue 45.6% and 24.6%. Worsening between the trial end and the first survey assessment was HAQ 0.29 units, pain 0.77 units, patient global 0.77 units, and fatigue 1.1 units.

Table 1. Characteristics of study population during NDB followup (n = 264).

Variable	Median (IQR) or %
Age, yrs	57.2 (50.1–64.4)
Disease duration, yrs	10.4 (4.3–23.5)
Time from clinical trial closure, yrs	0.8 (0.5–1.4)
Male, %	17.0
Study treatment, %	100.0
Prednisone, %	30.7
Methotrexate, %	68.2
Leflunomide, %	3.4

We also calculated the effect sizes for the study variables for the 2 time periods in Table 2. Pain and patient global improved most in the trial (effect sizes 1.29 and 1.22, respectively) and HAQ and fatigue improved the least (effect sizes 0.83 and 0.93). From the end of the trial to the survey evaluation, the effect sizes decreased by 0.30 (pain) to 0.44 (HAQ). The reduction in effect size for HAQ was more than half of the noted improvement.

ACR20%, 50%, and 70% improvement was noted by 63.1%, 43.4%, and 20.2% in the trial, respectively. We examined the effect of ACR improvement on changes in HAQ score between trial end and first survey evaluation in Figure 2. Except for the large difference in the small ACR70 group, changes in HAQ were similar in the nonresponders and ACR20 and ACR50 responders.

DISCUSSION

In our study we found a general loss of improvement in patient-reported outcomes in treated patients after they stopped participating in a clinical trial, but continued their initial study treatment. The study and non-study evaluations used the same questionnaire but differed only in the attention and setting of the evaluations. While this improvement meets the definition of a Hawthorne effect — additional clinical response that results from increased attention provided by participation in the clinical trial — it seems likely that the result could be attributed to a Type B effect in which the improvement is reported but is not true improvement. The alternative interpretation — that true improvement occurred in the trial but is lost at the end of the trial — seems untenable.

What is the nature of this improvement? One possibility is that it represents expectation bias, a form of placebo effect. Epstein noted this bias to be present in unblinded studies, but not in blinded controlled trials[13]. Hróbjartsson and Gøtzsche in a systematic review of placebo effects found placebo had minimal clinical effect compared to usual treatment[14]. However, the pooled standardized mean difference was significant for the trials with subjective outcomes but not for those with objective outcomes. In 27 trials

Table 2. Changes in study variables according to study time and setting.

Variable	Trial Start, mean (SD)	Trial End, mean (SD)	Post-trial Followup		
			Positive ES (95% CI)	Mean (SD)	Negative ES (95% CI)
ACR20 (%)		63.1			
ACR50 (%)		43.4			
ACR70 (%)		20.2			
HAQ, 0–3	1.21 (0.60)	0.71 (0.61)	0.83 (0.72, 0.95)	1.01 (0.71)	0.44 (0.60, 0.26)
Global, 0–10	5.2 (2.3)	2.5 (2.1)	1.22 (1.07, 1.38)	3.4 (2.5)	0.38 (0.49, 0.25)
Pain, 0–10	5.8 (2.3)	2.8 (2.4)	1.29 (1.12, 1.45)	3.5 (2.7)	0.30 (0.41, 0.19)
Fatigue, 0–10	5.7 (2.7)	3.1 (2.7)	0.93 (0.78, 1.09)	4.3 (2.9)	0.41 (0.54, 0.27)

Trial start: pretreatment; trial end: 6 months later; post-trial followup: 0.8 years later. Differences between trial end and trial followup are significant at $p < 0.001$. Positive: clinical improvement; negative: clinical worsening; ES: effect size.

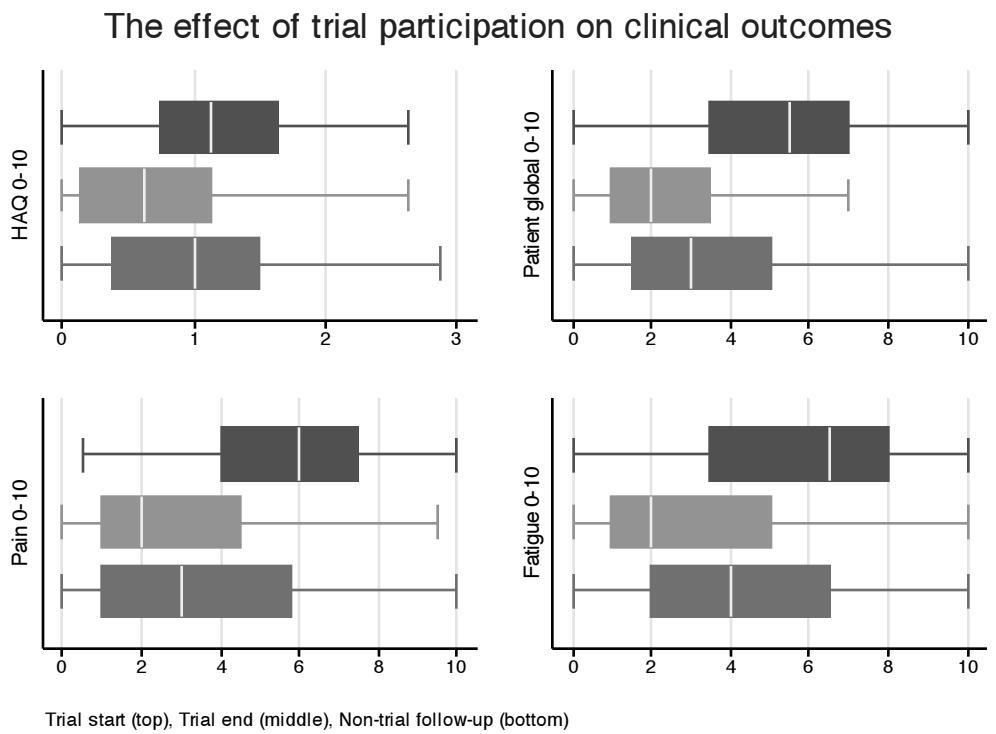

Figure 1. The effect of trial participation by categories on clinical outcomes for 264 patients treated with study drug. The median time from trial end to non-trial followup assessment was 0.8 years.

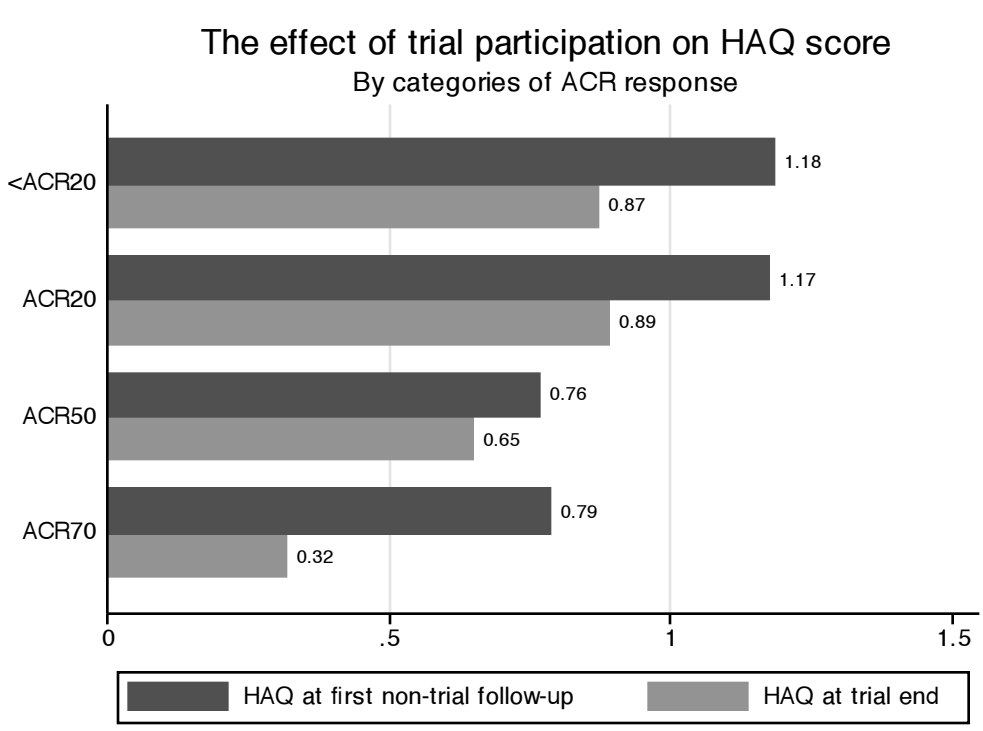

Figure 2. The effect of trial participation by categories of American College of Rheumatology (ACR) response for 264 patients treated with study drug. The median time from trial end was 0.8 years.

involving the treatment of pain, placebo had a beneficial effect, as indicated by a reduction in the intensity of pain of 6.5 mm on a 100-mm VAS.

Of interest, in the current study the difference in pain in the 2 settings was 0.70 units, very similar to Hróbjartsson and Gøtzsche's 6.6 mm. In standardized units of the effect

size, this difference was 0.3. For the HAQ score the effect size was –0.44 and the absolute difference was –0.30 units. A change of 0.22 units is considered to be the minimally important difference[15]. The difference between the clinical trial HAQ and the community (clinic) HAQ is important because HAQ values are commonly used to map to utility scores and then in the calculation of cost effectiveness[16]. In addition, the HAQ is a powerful predictor of mortality[17] and medical costs[18]. If clinical trial results overstate HAQ scores, then the true effectiveness — the real functional status — is overstated. In addition, US Food and Drug Administration indication requires a sustained improvement on the HAQ score. But what if the observed HAQ score is the result of a Type B Hawthorne effect?

Extension studies following clinical trials are common, and usually report sustained improvement. In a review of such extension studies, Landewé and van der Heijde report that "if there are no differences in treatment effects during an RCT, they are unlikely to appear at followup," and conclude that biased selection, dropouts, crossover, and confounding render such studies useless for providing longterm information[19].

Can our findings be extrapolated to other settings? We used the NDB to obtain scores for other patients with RA participating in survey research. We adjusted the results to the characteristics of the patients in the current study. For current study compared with other NDB RA patients, the results were HAQ 1.01 versus 1.12, patient global 3.4 versus 3.5, pain 3.5 versus 3.8, and fatigue 4.3 versus 4.5, respectively.

There are a number of real and potential limitations to this study. Participants were volunteers who continued on the study therapy after the trial. Patients who chose not to participate or discontinued therapy were not included in the study, and such patients had worse outcomes or were unsatisfied with their therapy. However, this exclusion process worked to the advantage in the study becaause it allowed us to observe how patients who are doing well fared when they changed setting — the model for Hawthorne effect observation. Even so, it is possible, although unlikely, that nonparticipants would have responded differently in the followup period. The Hawthorne effect is an important issue, and our hypothesis could be tested following future studies, with appropriate planning.

We also did not have physician data. So we were unable to extend this observation to physical examination and laboratory data. However, because patient global and tender joint count are part of the DAS-28[20], and pain and HAQ are part of the ACR improvement criteria, it seems likely that the Hawthorne effect plays a role here, too. However, the extent of the role cannot be determined.

In summary, almost half of the improvement noted in the clinical trial HAQ score disappeared on entry to a non-sponsored followup study, and from 23% to 44% of improvements in pain, patient global, and fatigue also disappeared. These changes can be attributed to the Hawthorne effect. Based on these data, we hypothesize that the absolute values of patient-reported RA outcome variables in clinical trials are upwardly biased, and that treatment effect is less than observed.

REFERENCES

1. Wolfe F, Michaud K, Dewitt EM. Why results of clinical trials and observational studies of antitumour necrosis factor (anti-TNF) therapy differ: methodological and interpretive issues. Ann Rheum Dis 2004;63 Suppl 2:ii13-ii7.
2. Wolfe F, Rasker JJ, Boers M, Wells GA, Michaud K. Minimal disease activity, remission, and the long-term outcomes of rheumatoid arthritis. Arthritis Rheum 2007;57:935-42.
3. Shaver TS, Anderson JD, Weidensaul DN, Shahouri SS, Busch RE, Mikuls TR, et al. The problem of rheumatoid arthritis disease activity and remission in clinical practice. J Rheumatol 2008;35:1015-22.
4. Wolfe F, Michaud K. The loss of health status in rheumatoid arthritis and the effect of biologic therapy: a longitudinal observational study. Arthritis Res Ther 2010;12:R35.
5. Franke RH, Kaul JD. The Hawthorne experiments: first statistical interpretation. Am Soc Rev 1978;43:623-43.
6. McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P. The Hawthorne effect: a randomised, controlled trial. BMC Med Res Methodol 2007;7:30.
7. Braunholtz DA, Edwards SJ, Lilford RJ. Are randomized clinical trials good for us (in the short term)? Evidence for a "trial effect". J Clin Epidemiol 2001;54:217-24.
8. Peppercorn JM, Weeks JC, Cook EF, Joffe S. Comparison of outcomes in cancer patients treated within and outside clinical trials: conceptual framework and structured review. Lancet 2004;363:263-70.
9. Wolfe F, Michaud K. A brief introduction to the National Data Bank for Rheumatic Diseases. Clin Exp Rheumatol 2005; 23:S168-S71.
10. Wolfe F, Michaud K. Biologic treatment of rheumatoid arthritis and the risk of malignancy: Analyses from a large US observational study. Arthritis Rheum 2007;56:2886-95.
11. 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.
12. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
13. Epstein WV. Expectation bias in rheumatoid arthritis clinical trials: The anti-CD4 monoclonal antibody experience — commentary. Arthritis Rheum 1996;39:1773-80.
14. Hrobjartsson A, Gotzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. N Engl J Med 2001;344:1594-602.
15. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557-60.
16. Kobelt G, Jonsson L, Lindgren P, Young A, Eberhardt K. Modeling the progression of rheumatoid arthritis: A two-country model to estimate costs and consequences of rheumatoid arthritis. Arthritis Rheum 2002;46:2310-9.
17. Wolfe F, Michaud K, Gefeller O, Choi HK. Predicting mortality in patients with rheumatoid arthritis. Arthritis Rheum 2003; 48:1530-42.
18. Michaud K, Messer J, Choi HK, Wolfe F. Direct medical costs and their predictors in patients with rheumatoid arthritis: a three-year study of 7,527 patients. Arthritis Rheum 2003;48:2750-62.
19. Landewe R, van der Heijde D. Follow up studies in rheumatoid arthritis. Ann Rheum Dis 2002;61:479-81.
20. van der Heijde DMFM, van 't Hof M, van Riel PLCM, Theunisse LAM, 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.