Precision of Composite Measures of Osteoarthritis Efficacy in Comparison to That of Individual Endpoints

JAMES A. BOLOGNESE, ELLIOT W. EHRICH, and THOMAS J. SCHNITZER

ABSTRACT. ***Objective.*** Osteoarthritis (OA) clinical studies generally employ various endpoints to evaluate a spectrum of disease manifestations. Compared with individual endpoints, a composite measure might (1) provide a uniform outcome measure of OA efficacy with greater face validity, (2) address multiplicity, and (3) enhance precision. Combinations of endpoints were analyzed to investigate precision of composite measures of OA efficacy.

Methods. We reanalyzed three 6 week, placebo controlled, double blind, parallel group studies (2 by the same protocol) of the cyclooxygenase-2 (COX-2) specific inhibitor rofecoxib. The average change from baseline at study weeks 2, 4, and 6 was assessed for 10 individual response variables, including patient and investigator global assessments, WOMAC 3.0V OA Index pain, stiffness and physical function subscales, graded study joint tenderness, and rescue analgesic use. Relationships among variables were evaluated using pairwise correlations and principal components analysis. The precision of variables to differentiate rofecoxib from placebo was evaluated using effect size (i.e., mean difference between rofecoxib versus placebo divided by pooled SD).

Results. Correlations among all pairs of response variables ranged from 0.5 to 0.9, except those with tenderness (0.4 to 0.6) and those with analgesic use (0.2 to 0.4). The first principal component explained about 70% of the total variability, with weights generally similar (0.17 for rescue analgesic use, 0.25 for tenderness, and 0.30 to 0.37 for the others). These results indicate that nearly all measures are closely related. Based on these results, various linear combinations of the 9 endpoints were formed and their precision to discriminate active treatment from placebo was compared to that of the individual endpoints. Effect sizes of the individual endpoints ranged from 0.6 to 1.1; those of the composites from 0.7 to 0.9. The results were very consistent between study protocols.

Conclusion. In comparison to individual endpoints, composite analyses of OA clinical endpoints do not increase precision to discriminate active treatment from placebo. (J Rheumatol 2001;28:2700–4)

Key Indexing Terms:
COMPOSITE MEASURES OSTEOARTHRITIS EFFECT SIZES
ENDPOINT CORRELATION ENDPOINT PRECISION

Clinical evaluation of osteoarthritis (OA) is made by measuring a variety of generally subjective endpoints, for example, patient and investigator global assessments of response to therapy and disease status, pain, stiffness, physical function, joint tenderness, joint swelling, and rescue analgesic use. Most of these endpoints measure symptoms on a visual analog (VAS) or Likert rating scale. There are no firmly established objective measurements, although joint tenderness and swelling are often graded on Likert or present/absent response scales.

From Merck Research Laboratories, Rahway, New Jersey, and Northwestern University, Evanston, Illinois, USA.

Supported by Merck Research Laboratories.

J.A. Bolognese, MStat, Senior Director, Scientific Staff, Merck Research Laboratories; E.W. Ehrich, MD, Vice President, Medical Affairs, Alkermes, Inc.; T.J. Schnitzer, MD, PhD, Professor of Medicine, Northwestern University.

Address reprint requests to J.A. Bolognese, Merck Research Labs, PO Box 2000, Rahway, NJ 07065. E-mail: james_bolognese@merck.com

Submitted December 21, 2000; revision accepted June 29, 2001.

For the evaluation of rheumatoid arthritis (RA), the American College of Rheumatology responder criteria (ACR20) composite endpoint was developed to incorporate several aspects of disease manifestation into a single descriptive indicator of clinical response[1]. In RA, the ACR20 responder criteria have been shown to be sensitive for comparing the magnitude of response among active treatments and placebo[1-3]. Zeng, *et al*[4] showed the ACR20 to be as precise as the best of the individual components in discriminating active treatment from placebo in a clinical trial setting. They and Tilley, *et al*[3] showed that O'Brien's global statistic[5] made up of components of the ACR20 was more precise than any of its individual components. Thus, in RA, composite endpoints have been shown to perform better than their individual component endpoints.

We examined various combinations of OA endpoints attempting to develop a composite measurement of OA more sensitive than individual components for discriminating active treatment from placebo in a clinical trial setting. The precision of various composite measures of OA

efficacy is compared with that of the various individual endpoints using data from 2 independent study protocols.

Composite endpoints were examined with 3 aims: (1) to provide a uniform measure with greater face validity than the individuals; (2) to avoid the multiplicity issue involved with assessing several individual endpoints; and (3) to enhance the precision of measurement of improvement of OA. This paper concerns primarily evaluation of the third aim, since the first 2 are self-evident motivations. To this end, combinations of endpoints from 2 completed clinical trial protocols were reanalyzed to assess composites.

MATERIALS AND METHODS

Data analyzed. The data assessed come from three 6 week, double blind placebo controlled, parallel group dose-response studies of the cyclooxygenase-2 (COX-2) specific inhibitor rofecoxib, also known as MK-0966[6-8]. The latter 2 trials are replicate studies using the same protocol; therefore, their data were combined for the computations in this paper to maximize information. The 25 mg rofecoxib (n = 135) and placebo (n = 140) treatment groups were chosen as the initial ("training") data set for evaluation in Protocol A. Saag, *et al*[7] and Day, *et al*[8] each showed similar efficacy results for their 12.5 (n = 462) and 25 mg (n = 466) treatments; thus, these 2 doses were combined and assessed in comparison to placebo (n = 143) for the Protocol B ("verification") data set to maximize the amount of information from which to derive the summary statistics in the analysis. Note that each of rofecoxib 12.5 and 25 mg once daily was shown to be clinically effective in OA of the hip or knee[7,8]; the latter is the maximum recommended dose in OA.

The OA efficacy response of each patient for each endpoint that was entered into the analyses was the average change from baseline across the entire 6 week treatment period. Baseline was the value at which flare criteria were satisfied after prior OA therapy washout. The average change from baseline was based on all planned observations at study weeks 2, 4, and 6, and time of discontinuation, if appropriate, for each of these 10 endpoints: patient global assessment of response to therapy (0–4 point Likert scale); investigator global assessment of response to therapy (0–4 point Likert scale); patient global assessment of disease status (100 mm VAS); investigator global assessment of disease status (0–4 point Likert scale); Pain Walking on a Flat Surface [Western Ontario and McMaster University Osteoarthritis Index (WOMAC) 3.0V question 1, 100 mm VAS, included because it was among the 3 primary endpoints of the 2 study protocols (patient global assessment of response to therapy and investigator global assessment of disease status were the others)]; WOMAC 3.0V pain subscale (100 mm VAS); WOMAC 3.0V physical function subscale (100 mm VAS); WOMAC 3.0V stiffness subscale (100 mm VAS); study joint tenderness (of the primary joint of OA involvement; 0–3 Likert scale); overall average analgesic rescue medication use (number of tablets per day).

The introductory instructions to the patients for each subscale of the WOMAC questionnaires included specific reference to the primary joint of OA involvement. The above 4 WOMAC endpoints were included in the analyses as 4 individual endpoints. Except for Pain Walking on a Flat Surface, as mentioned above, responses to the individual WOMAC questions were not included separately in any of the analyses. They were only included for their contribution to their respective WOMAC subscale score.

Composite endpoints were formed via several methods chosen to encompass a wide variety of potential summary measures of the individual endpoints. The following composite endpoints were examined: overall average of the 10 individual standardized endpoints; the first principal component (see Statistical Analyses section) of all individual endpoints; O'Brien's global statistic[5] based on all individual endpoints (this was computed by ranking all individual endpoint responses across all patients, then summing the ranks across endpoints to yield a rank sum score for each patient); overall average of all endpoints; overall median of all endpoints; minimum response value of all endpoints; maximum response value of all endpoints; WOMAC overall average score (average across the 24 questions); average WOMAC subscale score (average of the 3 WOMAC subscale scores).

All individual endpoints' response ranges were normalized to the 0 to 100 range to correspond to the 0 to 100 mm VAS for computation of all the composite endpoints.

Statistical analysis. Pearson correlation coefficients were computed to measure the degree of association between pairs of individual endpoints. Principal components analysis was carried out to assess the relative importance of the individual endpoints. This type of analysis considers each individual endpoint as a separate dimension and generates the particular weighted average of the individual endpoints that best accounts for the total variability in the multidimensional data. Once this first principal component is computed, and its explained variability is removed from the total variability in the multidimensional data, a second principal component can be computed, and a third, and so on, until all the variability is accounted for. The principal component coefficients were computed via standard techniques from the covariance matrix and scaled such that the sum of their squared values equaled 1[9]. Precision was assessed using effect size, i.e., the difference between treatment mean changes from baseline (rofecoxib vs placebo) divided by the pooled standard deviation (SD) of change from baseline. Thus, effect size is a measure of treatment effect in SD units.

RESULTS

Correlation coefficients between pairs of individual endpoints are shown in Table 1. They ranged from 0.4 to 0.9, except that those with acetaminophen use were small, between 0.2 and 0.4. Among the other pairs of endpoints, correlations with study joint tenderness were moderate, ranging from 0.4 to 0.6. Correlations between all pairs not containing acetaminophen use or study joint tenderness were high, ranging from 0.5 to 0.9.

In Protocol A, the first principal component accounted for 70% of the total variability; the second, for 10%. Thus, the first principal component appeared to be the only useful one in terms of summarizing most of the overall variability of the average change from baseline data. The principal component coefficients (Table 2) of the individual endpoints ranged from 0.30 to 0.37, with the exception of tenderness (0.24) and rescue medication (0.18). Similar results were observed in Protocol B — 66% of total variability was accounted for by the first principal component; 11% by the second; coefficients for the endpoints in the first principal component ranged from 0.31 to 0.37, except for 0.23 for tenderness and 0.17 for rescue medication. The magnitudes of the principal component coefficients indicate the relative importance of their respective endpoint to the composite principal component measure[9]. These principal component results indicate a general similarity with regard to relative importance of all the individual endpoints in explaining the overall variability, with tenderness and rescue medication of generally less importance than the others.

Values of correlation coefficients and principal component coefficients were within 10% between studies, indicating close consistency of results in the 2 separate study data sets. Effect sizes of the individual endpoints (Table 3)

Table 1. Correlation coefficients between pairs of individual endpoints (calculated from each patient's average change from baseline for each endpoint.

Protocol A	B	C	D	E	F	G	H	I	J
Endpoint									
A.WOMAC Pain Walking	0.68	0.59	0.72	0.66	0.90	0.80	0.74	0.44	0.31
B. Patient: response to therapy		0.65	0.66	0.86	0.68	0.67	0.65	0.51	0.42
C. Investigator: disease status			0.56	0.76	0.61	0.59	0.54	0.62	0.27
D. Patient: disease status				0.63	0.76	0.79	0.73	0.43	0.27
E. Investigator response to therapy					0.67	0.64	0.63	0.58	0.41
F. WOMAC pain subscale						0.87	0.79	0.45	0.31
G. WOMAC function subscale							0.86	0.43	0.30
H. WOMAC Stiffness subscale								0.38	0.34
I. Joint tenderness									0.24
J. Rescue analgesic use									
Protocal B	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Endpoint									
A. WOMAC Pain walking	0.62	0.52	0.68	0.63	0.84	0.71	0.65	0.35	0.34
B. Patient: response to therapy		0.64	0.61	0.86	0.67	0.64	0.59	0.44	0.43
C. Investigator: disease status			0.57	0.70	0.56	0.54	0.46	0.55	0.30
D. Patient: disease status				0.61	0.74	0.75	0.67	0.39	0.30
E. Investigator response to therapy					0.64	0.63	0.57	0.49	0.44
F. WOMAC pain subscale						0.87	0.77	0.40	0.31
G.WOMAC function subscale							0.79	0.40	0.32
H.WOMAC stiffness subscale								0.33	0.28
I. Joint tenderness									0.20
J. Rescue analgesic use									

Table 2. First principal component coefficients† for individual endpoints (calculated from each patient's average change from baseline for each endpoint).

	Protocol A	Protocol B
Endpoint		
WOMAC Pain Walking	0.35	0.35
Patient: response to therapy	0.34	0.34
Investigator: disease status	0.31	0.30
Patient: disease status	0.34	0.34
Investigator: response to therapy	0.34	0.34
WOMAC pain subscale	0.36	0.37
WOMAC function subscale	0.36	0.37
WOMAC stiffness subscale	0.34	0.34
Joint tenderness	0.25	0.24
Rescue analgesic use	0.18	0.19
Percentage of total variation explained	70*	64*

†The magnitudes of the principal component coefficients indicate the relative importance of their respective endpoint to the composite principal component measure[9]. *Without rescue analgesic use, which was excluded because it did not enhance the precision of the composite endpoints.

Table 3. Effiect sizes of endpoints (mean difference between rofecoxib and placebo divided by pooled standard deviation).

	Protocol A	Protocol B
Individual endpoints		
WOMAC Pain Walking	0.70	0.78
Patient: response to therapy	1.06	1.03
Investigator: disease status	0.60	0.75
Patient: disease status	0.88	0.94
Investigator response to therapy	0.84	0.85
WOMAC pain subscale	0.65	0.72
WOMAC function subscale	0.70	0.71
WOMAC stiffness subscale	0.73	0.67
Joint tenderness	0.65	0.59
Rescue analgesic use	0.55	0.53
Composite endpoints		
Mean of individual endpoints	0.91	0.95
First principal component	0.91	0.95
O'Brien's statistic	0.93	0.98
Median of individual endpoints	0.85	0.88
Minimum of individual endpoints	0.83	0.91
Maximum of individual endpoints	0.88	0.78
WOMAC overall average	0.72	0.74
WOMAC subscale average	0.74	0.75

ranged from 0.6 to 1.1, except for acetaminophen use, which had effect size 0.55. Recall that the unit of measurement for effect size is SD. For example, the difference between rofecoxib 25 mg and placebo in acetaminophen use was 0.55 SD.

Based on the results of the individual endpoints, rescue analgesic use was eliminated from the composites for several reasons. First, it had the smallest principal component weight; that is, it contributed the least toward

explaining variability in the overall individual endpoint data. Second, it had the smallest effect size; that is, it was least precise in measuring treatment effect. Third, the effect sizes of the composites were the same or greater without rescue analgesic use. Thus, rescue analgesic use was eliminated from the composites in the subsequent data summaries. The effect sizes of the composite endpoints ranged from 0.7 to 0.9 (Table 3), in the same range as the best individual endpoints.

Figure 1 shows a plot of effect sizes (with confidence intervals) of the individual endpoints and selected representative composites in Protocols A and B. Note that the larger the effect size, the more precise the endpoint in terms of measuring treatment effect in SD units. The individual endpoints are at the top of the plot, sorted by decreasing effect size, and the composite endpoints are at the bottom. Note that some of the individual endpoints are similarly precise as many of the composites, and the composites are generally similarly precise. In addition, note that similar results are observed for each separate study protocol.

The second set of trials also had an ibuprofen treatment group. The general ordering of ibuprofen effect sizes was similar to that of the rofecoxib effect sizes, except that those of rofecoxib were uniformly larger (data not shown).

DISCUSSION

Measuring the clinical assessment of OA involves potentially many endpoints. This paper evaluated a representative set of endpoints that encompasses all subjective aspects of the disease. These endpoints were shown in 3 studies, by 2 independent protocols, of the COX-2-specific inhibitor rofecoxib to be generally highly correlated (correlation coefficients on the order of 0.6 to 0.9), suggesting that they are generally measuring the same aspect of the disease — that is, symptoms involved in the related aspects of pain, stiffness, and function. Principal component analysis yielded results consistent with this notion, since component coefficients were all of generally the same size.

In light of the high correlation among the endpoints, it is not surprising that a variety of composite measurements of these endpoints yielded none with any enhanced precision for discriminating active treatment from placebo in either of the 2 clinical trial protocols assessed. Thus, although development of a composite measure of OA efficacy may be useful for obtaining a single measurement useful for descriptive purposes, based on the analyses reported in this paper it appears unlikely that such a composite would lead to enhanced precision for discriminating active treatment from placebo in an OA clinical trial. This is unlike the situ-

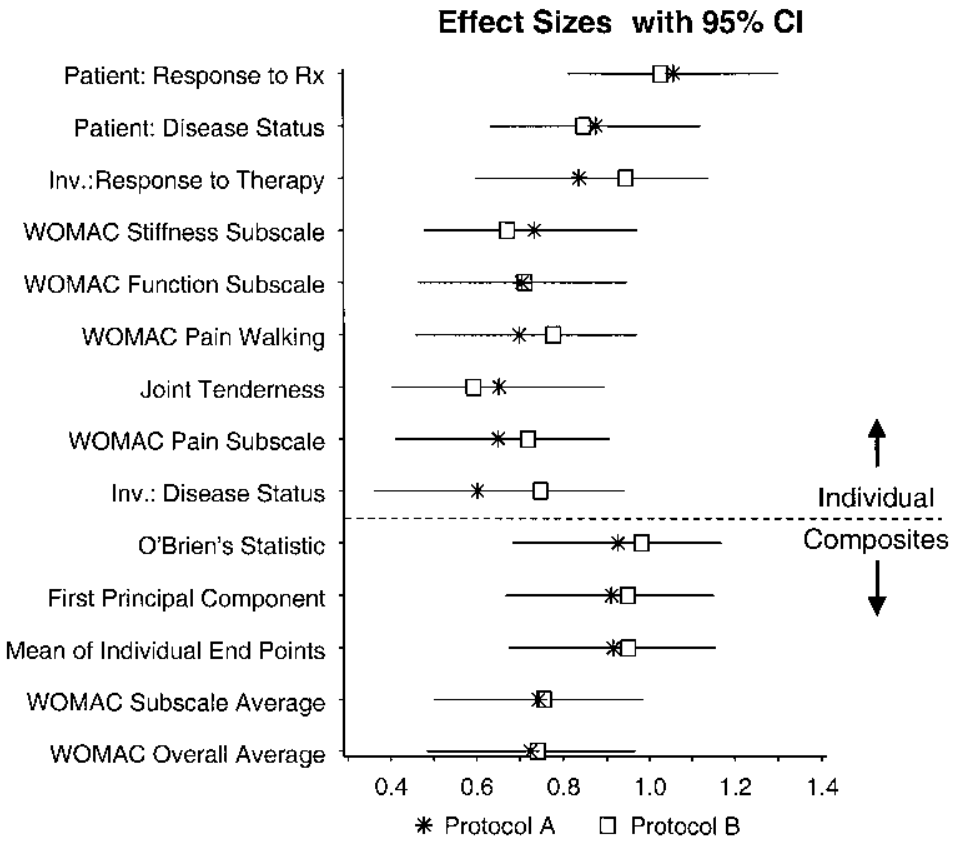

Figure 1. Effect sizes with 95% confidence intervals (CI) of individual and composite endpoints.

ation in RA, where the binary ACR20 endpoint has been shown to be similarly precise to the best from among its components[1], and a continuous composite of the ACR core set of endpoints has been shown to be more precise than the individual components for discriminating active treatment from placebo[4]. Tilley, *et al*[3] have shown that correlations among components of the ACR20, while statistically significantly positive, are of moderate magnitude. Their estimates of correlation coefficients among the RA endpoints range from 0.13 to 0.73, with most in the 0.2 to 0.6 range. These contrast to those found among OA endpoints in this paper. This could be part of the explanation for not having found an OA composite endpoint with substantially better sensitivity than its individual components.

The correlation, principal components, and effect size analyses were also carried out on the change from baseline to the last on-treatment observation, and yielded results (data not shown) similar to those of the average change from baseline. In addition, the correlation and principal component analyses were carried out on the average change from baseline after removal of the treatment effect by assessing residuals from analysis of variance. These analyses also yielded qualitatively similar results, although the magnitudes of the correlation coefficients were slightly less because correlated treatment-induced changes were removed from the data. Thus, these additional analyses demonstrated the robustness of the results reported in this paper.

Several additional potential interpretations can be drawn from careful examination of trends in the results from Tables 1 and 3 and Figure 1. First, not all correlation coefficients in Table 1 are large (> 0.7), suggesting that each of the individual endpoints does provide some degree of information independent of the others. However, that degree appears not to be great enough to have a marked influence on the precision of composite endpoints in relation to the individual endpoints. Second, from Table 3, the 2 individual endpoints that have effect sizes rivaling those of the composite endpoints are both the global evaluations of response to therapy, the use of which tends to be limited to trials of short duration since they depend on patients' recall of comparative response versus baseline. All except one composite endpoint show effect sizes in excess of all individual endpoints except for the response to therapy endpoints; however, as shown in Figure 1, the confidence intervals overlap substantially.

Third, it is surprising that effect sizes of all WOMAC endpoints tend to be lower than patient global evaluations and other composite endpoints. In addition, the individual WOMAC endpoint Pain Walking on a Flat Surface has an effect size similar to those of the individual WOMAC subscales and the Overall WOMAC score; however, this could, at least in part, be accounted for by the requirement of at least a 15 mm worsening at baseline of Pain Walking on a Flat Surface during prior OA therapy washout. These 3 trends should be viewed as hypothesis generating, rather than conclusive, since they are based on subtle trends in the data rather than on marked differences that exceed limits of variability of the estimates. Hence, they require further study to establish or refute.

In conclusion, analyses of data from 2 separate 6 week placebo controlled, flare designed study protocols that showed the efficacy of the COX-2-specific inhibitor rofecoxib in patients with hip or knee OA revealed that individual study endpoints were highly correlated. Consequently, composite endpoints did not increase the precision of efficacy comparisons beyond that of individual endpoints. Although composites were not shown to increase precision, they may be useful in OA to address other issues, for example, multiplicity of endpoints and estimation of overall level of response.

ACKNOWLEDGMENT

The authors thank the reviewers for their careful and thoughtful reviews. Their questions and suggestions led to addition of valuable information to this manuscript.

REFERENCES

1. Felson D, Anderson JJ, Boers M, et al. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
2. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 1995;38:1568-80.
3. Tilley BC, Pillemer SR, Heyse SP, et al. Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials. Arthritis Rheum 1999;42:1879-88.
4. Zeng Q, Bolognese J, Ehrich E, Daniels B. Comparison of several composite endpoints in the assessment of MK-0966 efficacy in rheumatoid arthritis [abstract]. Arthritis Rheum 1998;41 Suppl 9:S207.
5. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984;40:1079-87.
6. Ehrich E, Schnitzer T, Kivitz A, et al. MK-966, a highly selective Cox-2 inhibitor, was effective in the treatment of osteoarthritis of the knee and hip in a 6-week placebo controlled study [abstract]. Arthritis Rheum 1997;40 Suppl 9:S85.
7. Saag K, Fisher C, McKay J, et al. MK-0966, a specific Cox-2 inhibitor, has clinical efficacy comparable to ibuprofen in the treatment of knee and hip osteoarthritis in a 6-week controlled clinical trial [abstract]. Arthritis Rheum 1998;41 Suppl 9:S196.
8. Day R, Morrison BW, Luza A, et al. A randomized trial of the efficacy and tolerability of the COX-2 inhibitor rofecoxib vs ibuprofen in patients with osteoarthritis. Arch Intern Med 2000;160:1781-7.
9. Morrison DF. Multivariate statistical methods. New York: McGraw-Hill; 1967:222-7.