Common Measures and Analytic Techniques Provide Flawed Assessments of Pain: Modeled Data, and Hip Replacement Study

HUGH A. SMYTHE and EARL R. BOGOCH

ABSTRACT. ***Objective.*** To examine commonly used measures and analytic techniques of pain outcomes, using (1) a synthetic model, and (2) a cohort of patients who underwent total hip replacement.

Methods. (1) A synthetic data set was constructed with 110 visual analog scale (VAS) values, 10 for each integer from zero to 10. Random noise was added to simulate measurement variations. Drift through time and a therapeutic trial were simulated. (2) Eighty-six patients were studied before and a mean of 17 months after total hip replacement. Assessments included a VAS pain scale, the Western Ontario and McMaster Universities Osteoarthritis Index, Harris Hip Score, and SF-36 scores.

Results. The clinical study mirrored the model. Correlation coefficients among treatment differences measured by the pain subscales of 4 instruments varied from 0.53 to 0.22. Floor effects obscured benefit. "Percentage improvement" created a directional bias, and had a hyperbolic distribution that invalidated means, variances, and related statistics. The best outcomes were undervalued when postoperative pain measures approached zero. Standardized means enabled pooling of data from the different instruments and facilitated measurement of the variations due to treatment, methods, and subjects, and other factors.

Conclusion. Outcome measures and analytic techniques are often flawed because of floor and ceiling effects, non-normal distributions, and other problems. Outcomes expressed as "percentage improvement" are inappropriate; changes should be reported in the observed units. Revisions of standard outcome measures to relate pain with activity can better document outcomes, especially favorable results. (First Release Nov 15 2008; J Rheumatol 2008;35:2400–5; doi:10.3899/jrheum.080526)

Key Indexing Terms:
PAIN
OUTCOME MEASURES
PERCENTAGE CHANGE
BIAS
TOTAL HIP ARTHROPLASTY

Two decades ago, I was 60 and my son 30, half my age. Ten years later he was 40, having aged 33.3%. At 50, he has aged a further 25%. This is 58.3% (33.3% + 25%); or is it 66.7% (20/30) in 20 years. I have only aged 31% (16.7% + 14.3%), or perhaps 33.3% (20/60). Clearly, he is aging 200% faster than I. A grandson, now 24, has aged 600% during that same interval. In time I shall stop aging altogether, so they can catch up.

Clinical measures should be clearly understood by users and their audiences, should be designed for efficient statistical analysis, and should fully and fairly present the information sought. These needs are sometimes in conflict. We are in possession of a data set that allows exploration of these issues. The study involves 87 subjects with total hip arthroplasty (THA), who had a battery of standard tests before and following surgery. As a group, they had pain, loss of function, and poor quality of life before therapy, and a major treatment benefit, easily demonstrable with any of the outcome measures chosen. This analysis focuses on the performance of standard measures of pain and on statistical strategies that may be employed in the analysis of pain severity. The general outcomes of the hip surgery study are reported elsewhere[1].

Responses to this manuscript from experts are revealing in that they are contradictory. A statistician stated, "the floor effect of some scales, the difficulty with visual analog scales [VAS], and the issues involved with percent change scores...have been known for the past 50 years or so." On the other hand, a clinician argued that, "Problems with percentage change are severe when there is bidirectional change, but this generally does not happen from pre to post-treatment. At any rate, in rheumatoid arthritis, ...the 20%, 50%, 70% work because these subjects are in trials where a

From the Division of Rheumatology, The Toronto Western Hospital, University of Toronto; and Department of Surgery, University of Toronto, Mobility Program, St. Michael's Hospital, Toronto, Ontario, Canada.

Supported by an unrestricted research grant from Osteonics, Inc., Allendale, New Jersey, USA.

H.A. Smythe, MD, FRCPC, Division of Rheumatology, The Toronto Western Hospital, University of Toronto; E.R. Bogoch, MD, FRCSC, Department of Surgery, University of Toronto, Mobility Program, St. Michael's Hospital.

Address reprint requests to Dr H.A. Smythe, University of Toronto, 2 Heathbridge Park, Toronto, Ontario M4G 2Y6.
E-mail: hasmythe@rogers.com

Accepted for publication August 8, 2008.

certain level of active disease is required for entry. Thus, the example of 1 to 0 does not occur." A third suggested approach was to examine these issues with modeled data, which is the approach presented here.

MATERIALS AND METHODS

Modeling. We constructed a synthetic rectangular data set with 110 "VAS" values, 10 for each of the integers from zero to 10. The set had a mean and median of 5, and a standard deviation of 3.18. Uniform random noise was added to simulate measurement error, giving a "Base" set with a mean of 5.0 and standard deviation (SD) of 3.19. Equal change with further random noise was added to yield "High" and "Low" groups to mimic variation through time without treatment change. When "High" and "Low" values were added in a "followup" group, there was no net change, with a mean of 5.07, an increase in SD to 3.39, and a symmetrical rectangular distribution. Within the VAS strata from 0 to 10 the mean SD was 2.26, with no significant trend.

A treatment study was modeled, with the criterion for "entry" a VAS ≥ 4. Eighty-seven of our 110 "High" subjects met this standard, and mean was 8.94 (SD 2.80). Regression of the selected 87 to the "Base" set produced a mean of 6.07, a change of –2.88. A "Much Better" result was also modeled, with a mean outcome of 2.39 (SD 3.74), and change from entry of –0.6.55. The effects were additive, with uniform mean changes and SD within the strata. The "treatment" changes resulted in "VAS" values of less than zero in 11 "subjects." Rounding to zero led to important distortions; mean outcome 3.04, change –5.30. Though derived from the common synthetic group, the "spectra" of each of the subsets had differing parameter values.

When these "results" were analyzed as percentage change, the mean changes within the strata were not uniform. The standard deviations were even more affected, and both closely fitted a hyperbolic curve. Standard deviations rose to over 1000% in the lowest strata of the "Much Better" group (Figure 1).

The clinical study. Ninety-six patients (of 4 orthopedic surgeons, at 3 large urban university-affiliated hospitals) who had total hip replacements between September 1998 and September 1999 were enrolled. The study's nurse coordinator screened patients who visited the orthopedic outpatient clinics for preadmission evaluation for primary THA. For the study, exclusion criteria were any hip surgery prior to arthroplasty, rheumatoid arthritis (RA) or other inflammatory conditions, longterm steroid therapy, or neurological disease. Ten patients were excluded or lost, and 86 (90%) completed both initial and final assessments, with followup a mean of 17 months after surgery (range 10–29 mo, SD 6.3). Sixty-eight percent were female, and ages varied between 30 and 88 years (mean 61.8, SD 12.8 yrs).

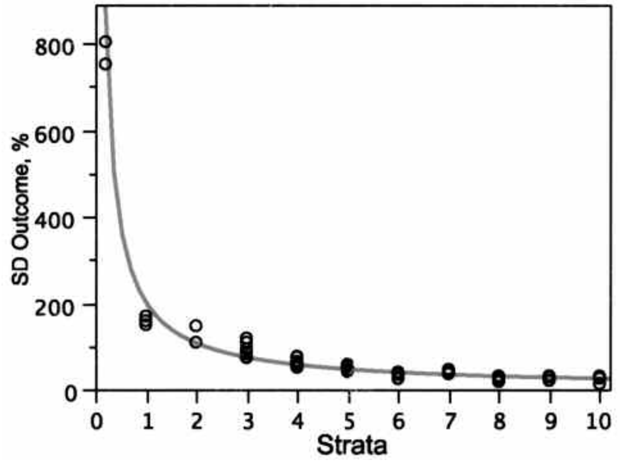

Figure 1. Standard deviations within VAS strata comparing "Much Better" as a percentage change from baseline. The curve is hyperbolic. The departures from homogeneity become severe below 3. Modeled data.

Pain was measured on an analog scale (VAS), recorded as integers from 0 to 10. Other measures were the pain scales within the SF-36[2,3], the Harris Hip Score[4], and the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[5].

The study was approved by the Hospital Research Ethics Board.

Analyses. The pain data were evaluated using several analytical tools and techniques, including analysis of variance and covariance (ANOVA), normal quantile plots, percentage change, standardization and pooling of the outcome measures, and T transformation (described below). The results were subsequently compared to determine correlations. Data were analyzed using JMP (version 6.0.3), a program developed by SAS Institute Inc. (Cary, NC, USA), incorporating powerful analytic and graphic capabilities.

RESULTS

The VAS pain scale. Pre- and postoperative data are presented in Figure 2. Thirty-two of 86 patients (37%) reported zero pain postoperatively, which resulted in a major floor effect. There were even 5 patients who reported zero pain preoperatively on this scale. Six subjects were at the scale maximum prior to surgery.

Most statistical analyses assume a normal distribution, which is not a valid assumption when there are large clumps of subjects at zero or 10. This potential problem was investigated by plotting the same data in a normal quantile plot (Figure 2). The departures from linearity reveal distributions that are not Gaussian. Floor effects are severe. VAS values of zero after surgery are ties, so that there is no measured difference among the lower 37% of the values. The preoperative values of zero before surgery are anomalous and cannot be improved. The slope poorly fits the VAS data between 2 and 9, the heart of the scale to a clinician. These problems are not solved with logarithmic or other data transformations, or with nonparametric ranking statistics.

A common and efficient way of examining change is to calculate the treatment difference (preoperative minus postoperative) as a single value for each participant (mean –2.79, SD 3.39). Each symbol represents the treatment difference measured in one patient. The scale is different, and changes now extend from –8 to +10, from worsening to maximum possible improvement, reflecting the variation of outcome among patients. The floor and ceiling effects have vanished, and the linear fit to the normal probability assumption is excellent. No assumptions about the original data can be made from distributions following additions or subtraction of data drawn from subgroups. The information lost in the ceiling and floor effects in the original preoperative and postoperative pain VAS data remains lost.

Percentage change. Percentage change is frequently calculated to indicate a clinically meaningful outcome. However, when change values are close to zero, as seen with 37% of the patients in this study, it is difficult to interpret outcomes using percentage change. A normal quantile plot of VAS pain change, calculated as a percentage, shows that the data

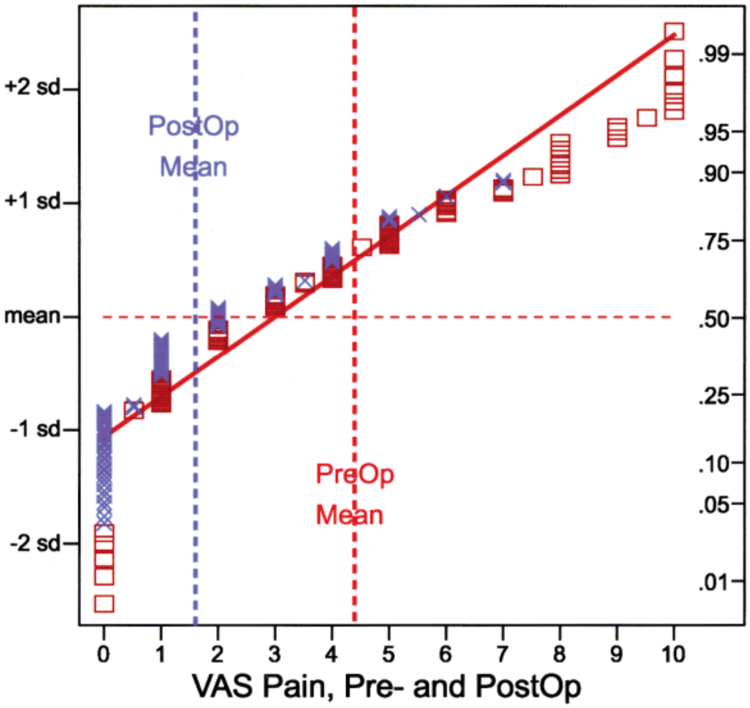

Figure 2. Normal quantile plot of the VAS data. Preoperative (PreOp) measures (mean 4.39) are shown as red squares, postoperative (PostOp) values (mean 1.61) are blue symbols. The vertical axis on the left indicates distance from the mean in standard deviation (SD) units, and on the right the quantiles. Solid red line is the least-squares regression fit to the total data set. It assumes a normal distribution, and passes through the overall (combined preoperative and postoperative) mean (2.96, SD 2.82), represented by the broken horizontal line. The regression line does not match the slope of the data between VAS 2 and 9.

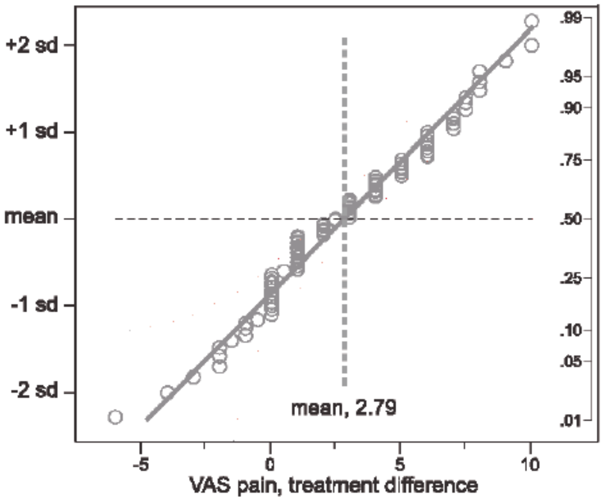

Figure 3. Normal quantile plot of the effect of treatment using VAS pain, calculated as preoperative minus postoperative values. The fit is excellent, in spite of the floor effects shown in Figure 2. From the clinical study.

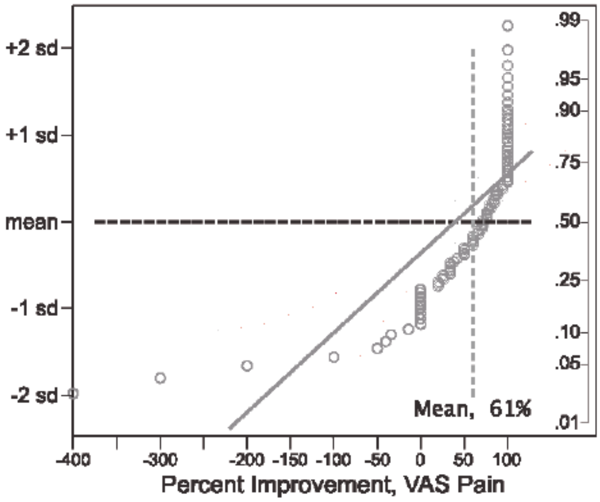

Figure 4. Normal quantile plot of VAS pain change, calculated as percentage change (treatment differences divided by the preoperative values, times 100%). Clinical study data.

from this study are clearly nonlinear and asymmetrical (Figure 4). Percentage changes cannot be meaningfully added, subtracted, averaged, or treated with any standard statistical techniques. If the VAS falls from 10 to 5, that is an improvement of 50%; but if the score worsens from 5 to 10, that is a 100% change. If these results occurred in the same individual (as a result of withdrawal of an effective remedy) an average would be clearly meaningless. It remains meaningless if they were different patients in separate groups of a controlled trial.

There was a large number of patients (27 of 81, or 33%) whose change was the same as the initial value. The same value was given for those whose VAS pain improved from 1 to zero as for those with improvement from 10 to zero — a 100% change. But if they started at 2 and ended at 8 the change is 400%. Changes of 20%, 50%, and 70% are urged or legally required in the reporting of clinical trials[6,7]. These targets are much more easily achieved if the pain scale used has a low value when pain is severe. When there is a zero in the denominator (our 5 patients with VAS of zero preoperatively), it becomes impossible to interpret percentage change. There are possible remedies for some of these problems, but the approach must be prespecified.

Comparing the 4 measures of pain. The pain scales in the 3 other instruments were examined. The analyses were restricted to the 74 patients with complete data for all measures. All scales showed major floor effects (data not shown), but the ceiling effects in Figure 2 were not a problem in the other measures.

The correlations among treatment differences measured by the 4 instruments are shown in Table 1. If the correlation coefficients approached 1.00, any single measure would do. The rest would be redundant, as they provided no new information. If any correlation coefficient approached zero, it should be dropped. Each measure contained information not present in the other 3, and none were perfect.

Table 1. Pearson correlations among measures of treatment difference.

Measures	WOMAC[5]	SF-36 Bodily Pain[3]	VAS Pain
SF-36 Bodily Pain[3]	0.53		
VAS pain	0.43	0.47	
Harris Hip Score[4]	0.28	0.28	0.22

WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index; VAS: visual analog scale.

Pooling the 4 measures of pain. To enable pooling of the data and further analyses, the measures obtained using the 4 different pain scales (VAS, SF-36 pain subscales[2], Harris Hip Score[4], WOMAC[5]) were standardized, converted into common standardized units of unit 1 by dividing by the SD of all values, preoperative and postoperative, for that measure[8,9]. The means were adjusted to give a common mean value (of zero) for the treatment effect (pre- and postoperative differences), and signs were adjusted so that higher values were uniformly more severe. Preoperative values were then (mostly) positive, and postoperative values negative, and were expressed in standard deviation units, as shown

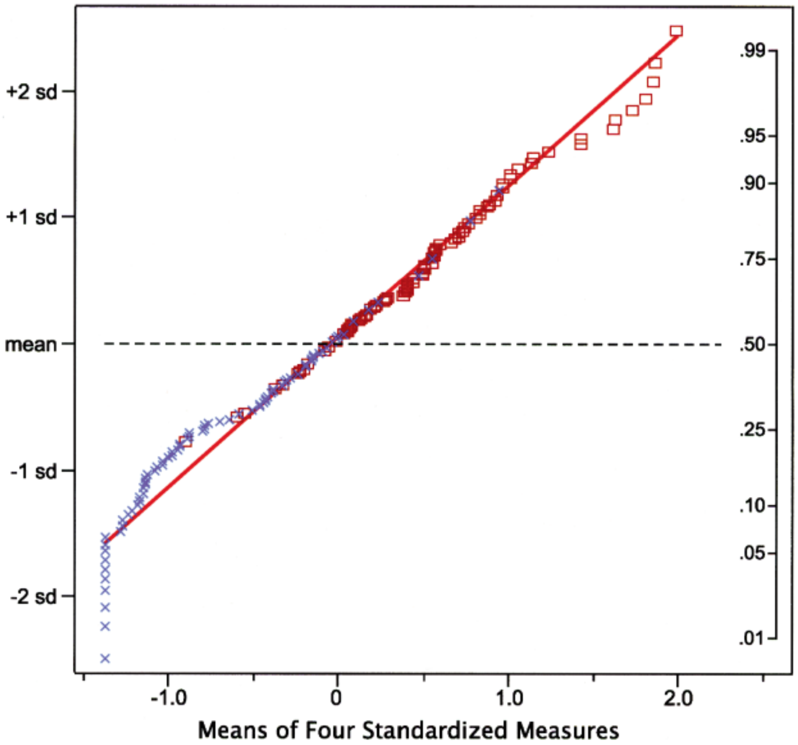

Figure 5. Normal quantile plot of means of standardized data from all 4 outcome measures. The fit to the linear regression line explained 62% of the total variance, but still contained a floor effect within the postoperative group. The mean preoperatively was 0.60 (red squares), postoperatively –0.60 standard deviation units (blue symbols). There is no useful posttreatment information beyond 1 standard deviation from the mean. Clinical study data.

visually in the normal quantile plots (Figure 2). Overall means for the pooled pain measures of treatment effect for each patient were then calculated. The distribution of the resulting values is shown in Figure 5.

Using all the information available, this produced means derived from the 4 separate measures of pain severity within each patient preoperatively, and 4 postoperatively, allowing a measure of "within-patient" variance, separate from the "among measures" term. The preoperative data fit the regression line up to 2 (but not 3) standard deviations from the mean. The remaining severe floor effect means there is no useful postoperative information beyond 1 standard deviation below the overall mean. Twelve subjects are tied at the minimum value. The F-ratio for the pooled standardized mean was 165.2, and the R^2 value was 0.59, slightly better than the WOMAC. Postoperatively, the variation among the measures was significant ($p = 0.030$), and the WOMAC was superior to VAS ($p = 0.027$).

What does zero pain mean? The purpose of hip replacement is to allow the person to be pain-free. Hence, if the procedure is successful, patients will not experience any pain and, if they complete a pain questionnaire, there should be floor effects. Unfortunately, the question on the VAS form given did not differentiate pain at rest from pain with activity. Of 27 subjects who reported zero VAS pain postoperatively, 10 had no pain on the other scales, but 17 reported pain using one or more of the other measures. Seventeen also reported values < 40 on the Physical Function (PF) scale of the SF-36, more than 1 standard deviation below the population-based norm of 50, and these had significantly more pain on the WOMAC pain score, the Harris pain score, and the SF-36 Bodily Pain scores, all with p values < 0.02. Age and sex were not different. The patients were quite satisfied with their outcome. The 17 others had "normal" function, with mean and median just below 50. The Harris and WOMAC scales have several questions relating pain to activities. The SF-36 also has a Physical Component scale, combining aspects of pain and disability.

There were the 5 participants who reported zero VAS pain preoperatively. When the 3 other outcome measures (WOMAC, Harris Hip Score, SF-36 Bodily Pain) were combined and the pooled data evaluated, the initial severity and postoperative improvement in the anomalous VAS group were not significantly different from the others, clinically or statistically.

Table 2. Outcome grouped by percentage change, using T-transformed pooled data. Worse: change negative. Slight: 0 to 20% improvement. Significant: 20.1% to 50%. Major: > 50%.

Subgroup	Count	% of 86
Worse	6	7
Slight improvement	17	20
Significant improvement	41	48
Major improvement	22	25

Percentage change revisited. Percentage change is a ratio, calculated as change divided by initial value plus a constant, which was set at zero in the pooled, standardized data. This is the formula for a hyperbola and is expressed diagrammatically in Figure 6. As the initial value, or preoperative pain measure, approaches zero, the percentage change rises steeply. For the 5 subjects whose initial pain VAS was zero, the value is simultaneously plus and minus infinity, which presents a nonsensical result when evaluating treatment outcomes and effects.

One effective strategy is the T transformation, used in the later evolution of the SF-36[3]. Our standardized data pooled from the 4 outcome measures had the population mean fixed at zero, and a measured standard deviation of 0.84. In the T transformation, the SD is increased to 10 by multiplying our data by 11.94 (10 times the inverse of 0.84). The addition of 50 moves the result from the "hyperbolic" to the "linear" zone in Figure 6. With a population mean of 50 and SD of 10, zero and 100 are a safe 5 standard deviations away. The treatment difference now closely approximates a normal distribution, as does the percentage change. We can now examine the data for 20%, 50%, and 70% change (Table 2). When outcome is grouped by percentage change, 73% of patients had significant or major pain improvement.

Nonparametric approaches. With severe ceiling or floor effects or other major departures from normality, the median is a much more robust measure of central tendency than the mean. However, with these distributions, reliable measures of dispersion may not be available, whether expressed as a standard error of the median or as 0.025 and 0.975 quantiles. We have investigated these questions extensively using jackknife and bootstrap techniques. Varying the proportion of the data excluded during resampling has major effects on the distribution of the derived standard error; the relationship is hyperbolic. A full presentation of these find-

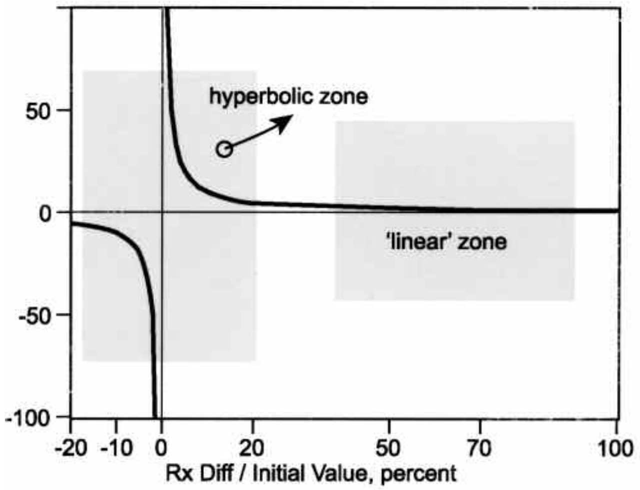

Figure 6. Values on the vertical axis are percentage change. Adding a constant to the change ratio (as in the T transformation) can move a statistic out of the hyperbolic zone into the asymptotic "linear" zone, not clinically or significantly different from a straight line.

ings is beyond the scope of this report — to allow analyses of the effects of multiple factors on multiple measures, analyses of variance techniques were prespecified.

DISCUSSION

In an unrelated controlled trial, treatment "A" resulted in a change of 3.2%, treatment "B" a change of 1.6%. Was "A" 100% different from "B," 50% different, 1.6% different, or really none of the above? (The sponsoring company chose 50%, very prominently claimed in its current advertisements.)

Percentages are easily understood and useful, when used to define proportions. But they can be misleading when comparing outcomes. They are a distinct form of ratios; classically y = 100*n/x. The numerator "n" in our studies is the treatment change, postoperative minus preoperative values. Figure 3 shows an excellent fit of treatment change to a normal distribution, with none of the floor and ceiling effects seen in the data from which it is derived. The range of values was from –8 to +10, not limited by 0 and 10. This set is very suitable for further statistical analyses.

The devil is in the denominator. When it is a variable (a range of values), a curvilinear transformation results, specifically a hyperbola. When the range of values includes or approaches zero, the curves shown in Figures 1 and 4 are the result. After percentage transformation, the modified data are ordinal, but the intervals and deviations are not uniform and not additive. Means, standard deviations, confidence limits, etc. cannot be calculated without error — usually not obvious, and usually not intended. Data from 2 separate studies (expressed as percentage improvement) cannot or should not be compared. When treatment is very effective (joint replacement, anti-tumor necrosis factor therapy), we are most interested in the measures at greatest distance from the pretreatment mean, and most distorted by the calculation.

Directional bias in percentage change. Pain severity typically varies with time. Our modeled data included "High" and "Low" sets, arithmetically at equal distance from "Base" values. The sum of high and low was zero (apart from 0.14 VAS units of added noise). In percentage terms, the "High" values were 94.0% higher than the "Base" values; the "Low" values only 2.9% lower. This directional bias was recognized by Boers, *et al*[10], but their remedy was to choose or modify measures that decreased with improvement, to avoid overestimation of benefit. Bias persists, and overestimation of benefit will occur if preservation of joint space or bone density is the desired outcome. We echo Leslie[11]: "precision and follow-up assessment (should be) based upon absolute measurements, (rather than) relative change."

The VAS is the most familiar and widely used pain scale, but it was worst in terms of clinical and statistical performance. In addition to the problems listed above, the VAS pain scale used did not clearly differentiate pain at rest from pain with activity. A recent study of factors influencing pain after knee arthroplasty noted that 23% had no pain at rest preoperatively (a favorable prognostic factor), but all had pain with movement[12]. Pain with vigorous activity was not specifically queried in any of the scales used.

The problems come into focus at this time, because much more effective treatment strategies have become available. Reconstructive surgery has come of age, and biologic therapies have proved remarkably effective for medical therapy for some patients with rheumatoid arthritis and other autoimmune diseases. With modern therapies, a possible outcome is a return to tennis, golf, skiing, climbing, or heavy lifting. This is not identified by scales or by scoring methods developed decades ago.

ACKNOWLEDGMENT

The authors thank Drs. J. Gollish, E. Schemitsch, and J.P. Waddell for contributing patients to the study. Marjorie Henke, RN, collected the data. Dagmar Gross, MSc, assisted with the preparation of the manuscript.

REFERENCES

1. Bogoch ER, Olschewski E, Zangger P, Henke M, Smythe HA. Increased tender point counts before and after total hip arthroplasty are associated with poorer outcomes, but are not predictive in individual patients. J Arthroplasty 2008 (submitted).
2. Ware JE, Kosinski M, Keller SD. SF-36 Physical and Mental Health Summary Scales; a user's manual. Boston: Health Assessment Lab, New England Medical Center; 1994.
3. Ware JE, Kosinski M. SF-36 Physical and Mental Health Summary Scales; a manual for users of version 1, second edition. Lincoln, RI: QualityMetric Inc.; 2001.
4. Harris WH, McCarthy JC, O'Neill DA. Femoral component loosening using contemporary techniques of femoral cement fixation. J Bone Joint Surg 1982;64A:1063-7.
5. Bellamy N, Buchanan W, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or the knee. J Rheumatol 1988;15:1833–40.
6. Felson DT. Whither the ACR20? [editorial]. J Rheumatol 2004;31:835-7.
7. 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.
8. Dwosh IL, Stein HB, Urowitz MB, Smythe HA, Ogryzlo MA. Azathioprine in early rheumatoid arthritis: comparison with gold and chloroquine. Arthritis Rheum 1977;20:685-92.
9. Smythe HA, Helewa A, Goldsmith CH. "Independent assessor" and "pooled index" as techniques for measuring treatment effects in rheumatoid arthritis. J Rheumatol 1977;4:144-52.
10. Boers M, Verhoeven AC, van der Linden S. American College of Rheumatology criteria for improvement in rheumatoid arthritis should only be calculated from scores that decrease on improvement. Arthritis Rheum 2001;44:1052-5.
11. Leslie W. The importance of spectrum bias on bone density monitoring in clinical practice. Bone 2006;39:361-8.
12. Lundblad H, Kreicbergs A, Jansson KA. Prediction of persistent pain after total knee replacement for osteoarthritis. J Bone Joint Surg Br 2008;90:166-71.