Responsiveness of Endpoints in Osteoporosis Clinical Trials

GEORGE WELLS, ANN CRANNEY, BEVERLEY SHEA, and PETER TUGWELL

ABSTRACT. The usefulness of an endpoint depends in part on its responsiveness to clinically important change. From existing randomized controlled trials, the responsiveness of endpoints currently employed in osteoporosis clinical trials were examined. The responsiveness is presented as the sample size per group needed to show a statistically significant difference. The large variation found means that careful attention needs to be given to the responsiveness of the population studied when estimating the sample size. (*J Rheumatol 1997;24:1230–3*)

Key Indexing Terms:
OSTEOPOROSIS CLINICAL TRIALS OUTCOME MEASURES

Unlike other rheumatic diseases, osteoporosis is often symptomatic and, as a result, clinical measures have been difficult to study. A wide range of clinical and nonclinical endpoints can be identified in osteoporosis clinical trials, namely: fracture, bone density, biochemical markers, pain, health status, and height.

In selecting an endpoint for a clinical trial, characteristics such as validity, feasibility, reproducibility, and responsiveness must be considered. In particular, the usefulness of an endpoint depends on its ability to detect meaningful change, which is often termed "responsiveness"[1,2].

For OMERACT III, we examined the responsiveness of various endpoints commonly employed in osteoporosis clinical trials. In this article we present responsiveness as the sample size per group needed to show a statistically significant difference based on the magnitude of benefit seen in clinical trials that have been conducted with various therapeutic agents.

MATERIALS AND METHODS

Using the Cochrane Collaboration's search strategy, a literature search of MEDLINE from 1966 to 1995 was conducted for randomized placebo-controlled trials for the treatment of osteoporosis that assessed the various endpoints of interest. Experts in the field of osteoporosis were also contacted to identify other potential trials. From this compilation of clinical trials, at least one randomized trial for each major treatment for osteoporosis [e.g., alendronate (Al), etidronate (Et), hormone replacement therapy (HRT), vitamin D_3 (D3), calcitonin (Clt), and fluoride (Fl)] were identified. Two independent reviewers extracted the necessary data (GW, AC).

For each endpoint, an effect size was calculated. The effect size for continuous endpoints (e.g., bone densitometry, biochemical markers, pain, health status, height) was calculated as the difference of the means of treatment and placebo groups divided by the standard deviation (SD) of the placebo. For dichotomous endpoints (e.g., fracture), effect size was taken as twice the difference between the treatment and placebo groups of the inverse arc sine of the square root of the rates. A large effect size indicated a high sensitivity to change. Sample size was calculated based on the effect size derived for each endpoint, and for the probability of Type I and II errors of 0.05 and 0.20, respectively. Effect sizes were classified as small (0.2), medium (0.5), or large (0.8) according to Cohen[3]. Using this classification as a guideline, sample sizes were classified as good (< 50 evaluable patients per group), fair (51–150), poor (150–900), or very poor (> 900).

The sites considered for bone densitometry were the lumbar spine, femoral neck, trochanter, and distal forearm. Due to inconsistency in reporting of the SD and the mean change in bone density, percentage of baseline was used and, where necessary, a pooled SD was calculated for both treatment and placebo group using the first order terms of Taylor's series expansion for the ratio of 2 random variables.

RESULTS

Of the articles found in the literature search, 10 were selected for further evaluation based on the treatment under investigation, endpoints used, and the quality and size of the study. The studies, selected and identified by the first author, were Lufkin[4], Overgaard[5], Storm[6], Munk-Jensen[7], Reginster[8], Pak[9], Chapuy[10], Lyritis[11], Liberman[12], and Black[13]. Of the clinical and nonclinical endpoints of interest, radiographic evidence of fracture and bone densitometry were most often used.

Clinical fracture is probably the best endpoint reflecting the clinical outcome of interest; however, little information is available for this endpoint from studies conducted to date. The Black study[13] (Fracture Intervention Trial), with alendronate as the intervention, is the only randomized controlled trial providing this information. There is a distinction between clinical vertebral and clinical nonvertebral fractures, with effect sizes of 0.152 and 0.084, respectively. The resulting responsiveness based on sample size is considered poor for vertebral (703) and very poor for nonvertebral (2232) fractures.

For radiographic evidence of fracture, the distinction between vertebral and nonvertebral fractures is similar to

From the Departments of Medicine and Epidemiology, University of Ottawa, Ottawa, Canada.

Cranney is the recipient of an Arthritis Society fellowship.

G. Wells, PhD, Professor, Department of Medicine, and Associate Director and Senior Biostatistician, Clinical Epidemiology Unit, University of Ottawa and Ottawa Civic Hospital; A. Cranney, MD, Rheumatologist; B. Shea, BScN, Research Associate; P. Tugwell, MD, Professor and Chairman, Department of Medicine, University of Ottawa and Ottawa General Hospital.

Address reprint requests to Dr. G. Wells, Associate Director, Clinical Epidemiology Unit, F6, Ottawa Civic Hospital, 1053 Carling Avenue, Ottawa, Ontario K1Y 4E9, Canada.

the clinical fracture evidence, with very poor responsiveness corresponding to the nonvertebral fractures (Table 1). For vertebral fractures, small to medium effect sizes were found. The smallest effect size was 0.144 in the Liberman study[12] for alendronate. The sample sizes were considered good to fair for etidronate and fluoride (60; 45, respectively) and poor for alendronate, HRT, and calcitonin (845; 182; 156, respectively). For the nonvertebral fractures, 2 studies were available for etidronate and alendronate and extremely small effect sizes were found, with corresponding sample sizes exceeding 2400.

In Table 2, effect size and sample size for bone densitometry outcomes are displayed for various sites. In general, effect sizes were largest for the lumbar spine site, ranging from 0.619 to 1.917 and with corresponding sample sizes of less than 9 to 43. These effect sizes were large and the results were quite consistent across treatments, with the greatest discrepancy for calcitonin in the Overgaard study[5]. In summary, the sample sizes by treatment were: calcitonin 43,9; HRT 14,9; etidronate 12, fluoride 10, and alendronate 9. For the femoral neck, the effect sizes were inconsistent, ranging from a large effect for alendronate, to medium for HRT and vitamin D_3, to small for fluoride. The sample sizes are considered good for alendronate (22), fair for HRT and vitamin D_3 (89; 64, respectively), and poor for fluoride (223). The trochanter was similar to the femoral neck with medium and large effect sizes and sample sizes considered good for alendronate and HRT (14; 22, respectively) and fair for vitamin D_3 (85). The results for the distal forearm were quite variable, with an extremely small effect size for calcitonin in the Overgaard study leading to a very poor responsiveness rating.

Table 1. Responsiveness of fractures (radiography).

	Study	Treatment	Effect Size	Sample Size
Vertebral Fractures	Overgaard	Calcitonin	0.316	182
	Storm	Etidronate	0.554	60
	Pak	Fluoride	0.648	45
	Liberman	Alendronate	0.144	845
	Lufkin	HRT	0.330	156
Nonvertebral Fractures	Storm	Etidronate	0.081	2492
	Liberman	Alendronate	0.075	2877

Table 2. Responsiveness of bone densitometry.

	Study	Treatment	Effect Size	Sample Size
Lumbar spine	Overgaard	Calcitonin	0.619	43
	Reginster	Calcitonin	1.433	< 9
	Munk-Jensen	HRT	1.917	< 9
	Storm	Etidronate	1.194	12
	Pak	Fluoride	1.346	10
	Liberman	Alendronate	1.862	< 9
	Lufkin	HRT	1.096	14
Femoral neck	Pak	Fluoride	0.278	223
	Liberman	Alendronate	0.901	22
	Lufkin	HRT	0.43	89
	Chapuy	Vitamin D_3	0.5	64
Trochanter	Liberman	Alendronate	1.10	14
	Lufkin	HRT	0.898	22
	Chapuy	Vitamin D_3	0.439	85
Distal forearm	Overgaard	Calcitonin	0.103	1536
	Munk-Jensen	HRT	1.416	< 9
	Storm	Etidronate	0.300	175

Relatively little information is available for other endpoints. The responsiveness of the height endpoint was rated as fair, with sample sizes of 98 and 133 required for etidronate and fluoride, respectively. For the pain outcome, little information was available; sample size, based on one trial of calcitonin after incident fractures, was considered good (45)[11]. As for health status, no performance and disease specific or generic quality of life data were available. Data were available on various biochemical markers. In particular, for osteocalcin (bone formation) the sample size was considered good (HRT 9; vitamin D_3 16)[4,10], and for alkaline phosphatase, good to fair (vitamin D_3: 9; calcitonin 63)[5,10]. For the biochemical marker hydroxyproline (resorption), a wide range was found, with sample sizes considered good and poor for HRT and calcitonin, respectively. In Figure 1, a summary of the responsiveness of all the endpoints is shown.

DISCUSSION

A wide range of clinical and nonclinical endpoints have been used in osteoporosis clinical trials. In terms of face validity or relevance, clinical fracture is probably the strongest endpoint, reflecting the primary clinical outcome of interest. Although the feasibility of this endpoint is high, few properly designed randomized controlled trials have been conducted since large sample sizes and a considerable commitment of time and resources is required. Based on scant evidence, a poor responsiveness for vertebral fractures exists and responsiveness of nonvertebral fractures is extremely poor. Using the latter as the specific endpoint of interest would require an extremely large study involving thousands of patients.

Radiographic evidence of fractures is often used as a surrogate measure of clinical fractures. Although necessary, it is only one component of the clinical fracture event. Missing are clinical impact of the event from the distress of pain and effects of the resulting disability, among others. The feasibility of obtaining this evidence is high and several studies have investigated this endpoint, in particular vertebral fractures. Similar to clinical fractures, the responsiveness of vertebral fractures is rated fair to poor and for nonvertebral fractures very poor.

Most information available is for bone densitometry. However, this measure does not assess quality of bone, and different treatments appear to have different relationships

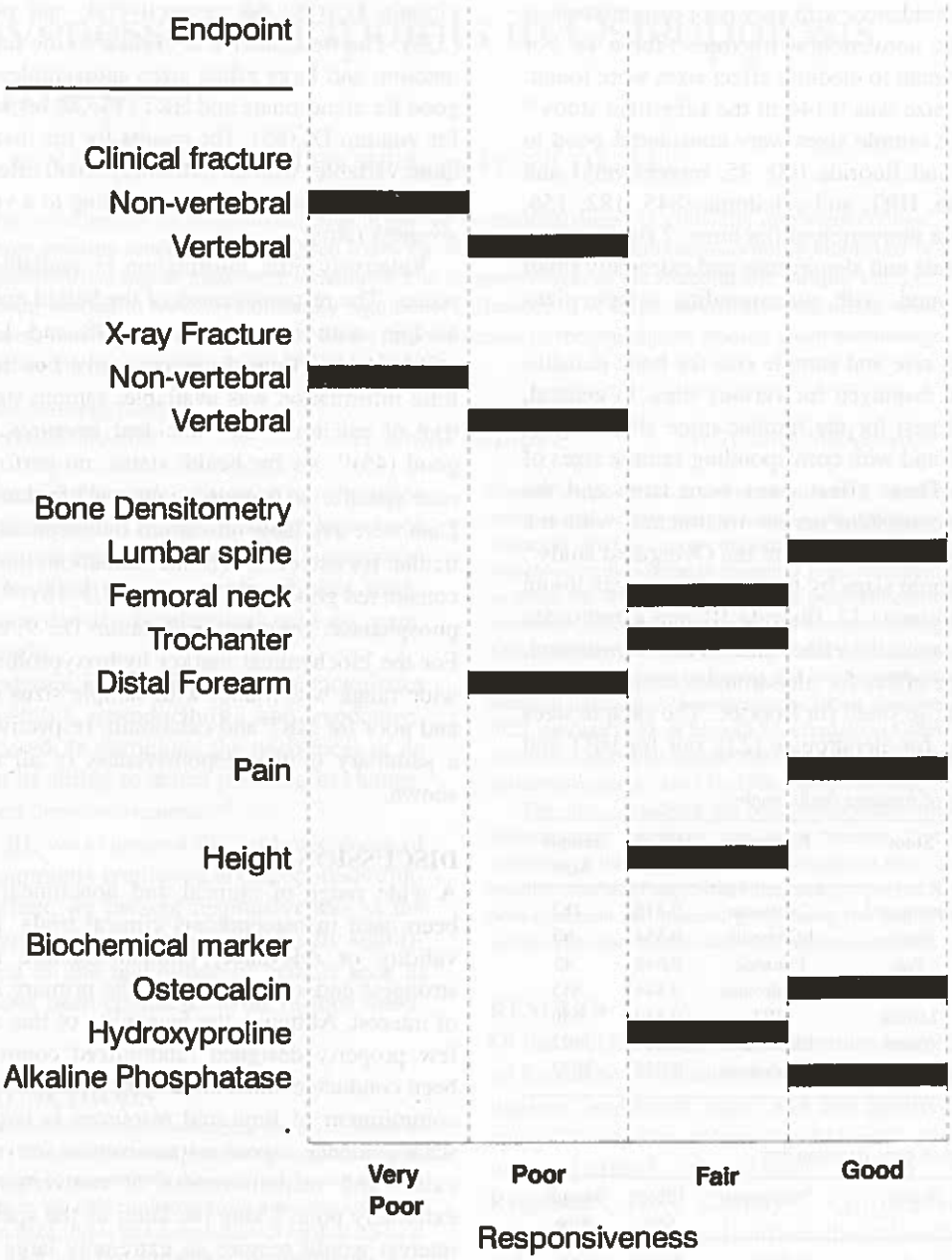

Figure 1. Responsiveness of endpoints in osteoporosis clinical trials.

with bone density and fractures. The feasibility of this endpoint is high for dual energy x-ray absorptiometry, dual photon absorptiometry, and single photon absorptiometry; it is low for quantitative computed tomography. More information is required for ultrasound. The responsiveness of this endpoint is good for the lumbar spine, femoral neck, trochanter, and distal forearm sites, although the distal forearm results are quite variable.

Health status is important for quantitating the effect of fractures on health status, disability, and quality of life. Although a number of disease specific quality of life measures (e.g., OPAQ, Osteoporosis Assessment Questionnaire for spinal osteoporosis) have been developed and are being studied in clinical trials, we were unable to locate published trials with data on this endpoint. Pain is important for assessing clinical fracture[11]. The feasibility of this endpoint is high and, although little information is available, it appears to be responsive. While height is also a feasible endpoint, what it adds to vertebral fractures is unclear. In general its responsiveness is fair. Although biochemical markers are important for understanding the biology, at this time there is not enough evidence for their use as a surrogate for

bone density. However, a recent study by Garnero[14] suggests that urinary markers of resorption such as C telopeptide and free deoxypyridinoline may be useful in assessing risk of hip fracture in the elderly. The responsiveness of these markers is generally good.

In clinical trials, the choice of endpoint is critical and in many instances the feasibility of implementing the trial depends on the responsiveness of the endpoint to the treatment regimen. Often endpoints are relevant but have such poor responsiveness that they require a very large sample size, making a trial impractical. As a result, more responsive surrogate markers may be used, but these may be of questionable clinical relevance. The most relevant endpoint for osteoporosis is clinical or radiographic nonvertebral fractures, which has very poor responsiveness. Vertebral fractures are a reasonable alternative having clinical relevance and better responsiveness. Bone densitometry is also a responsive endpoint, and its usefulness will increase as its exact relationship to clinical features is established.

REFERENCES

1. Fortin PR, Stucki Katz JN: Measuring relevant change: An emerging challenge in rheumatologic clinical trials (editorial). *Arthritis Rheum 1995;*38:1027–30.
2. Kazis LE, Anderson JJ, Meenan RF: Effect sizes for interpreting changes in health status. *Med Care 1989;*27:S178–89.
3. Cohen J: *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. Hillsdale: Lawrence Erlbaum, 1988.
4. Lufkin EG, Wahner HW, O'Fallon WM, *et al:* Treatment of postmenopausal osteoporosis with transdermal estrogen. *Ann Intern Med 1992;*117:1–9.
5. Overgaard K, Hansen MA, Jensen SB, Christiansen C: Effect of calcitonin given intranasally on bone mass and fracture rates in established osteoporosis: A dose response study. *BMJ 1992;*305:556–61.
6. Storm T, Thamsborg G, Steiniche T, *et al:* Effect of intermittent cyclical etidronate therapy on bone mass and fracture rate in women with postmenopausal osteoporosis. *N Engl J Med 1990;*322:1265–71.
7. Munk-Jensen E, Nuelsen Pors S, Obel EB, Bonne Eriksen P: Reversal of postmenopausal vertebral bone loss by estrogen and progestogen: A double blind placebo-controlled study. *BMJ 1988;*296:1150–2.
8. Reginster JY, Denis D, Deroisy R, *et al:* A double-blind placebo-controlled, dose finding trial of intermittent nasal salmon calcitonin for prevention of postmenopausal lumbar spine loss. *Am J Med 1995;*98:452–8.
9. Pak CYC, Sakhaee K, Adams-Huet B, *et al:* Treatment of postmenopausal osteoporosis with slow-release sodium fluoride. *Ann Intern Med 1995;*123:401–8.
10. Chapuy MC, Arlot ME, Duboeuf F, *et al:* Vitamin D_3 and calcium to prevent hip fractures in elderly women. *N Engl J Med 1992;*327:1637–42.
11. Lyritis GP, Tsakalakos N, Magiasis B, *et al:* Analgesic effect of salmon calcitonin in osteoporotic vertebral fractures: A double-blind placebo-controlled clinical study. *Calcif Tissue Int 1991;*49:369–72.
12. Liberman UA, Weiss SR, Broll J, *et al:* Effect of alendronate on bone mineral density and the incidence of fractures in postmenopausal osteoporosis. *N Engl J Med 1995;*333:1437–43.
13. Black DM, Cummings SR, Karpf DB, *et al:* Effect of alendronate on risk of fracture in women with existing vertebral fractures: Results of the Fracture Intervention Trial. *Lancet 1996;*348:1535–41.
14. Garnero P, Hausherr E, Chapuy MC, *et al:* Markers of bone resorption predict hip fractures in elderly women: The EPIDOS prospective study. *J Bone Miner Res 1996;*11:1531–8.