On bias

Leif Ryd and Leif Dahlberg

We performed a search in Medline to assess the quality of clinical journals in orthopedics from the point of view of study design. 3 levels of quality were chosen: prospective studies, random allocation or double-blind methods and randomized controlled trials; all entries were Medical Subject Headings (MeSH).

Out of 25,538 articles indexed in Medline since 1966–1993 in the 8 most cited general orthopedic journals, 994 were indexed as prospective studies, while 138 were indexed as randomized, controlled trials. In recent years the number of well-designed articles has increased, as has the percentage. In a random check of 208 articles, approximately half were of a clinical type where these issues can be addressed. The agreement between the manual check and Medline indexing was good, but not perfect.

It was concluded that the retrospective study representing clinical production control accounts for the vast majority of all published clinical articles in orthopedics. In recent years, a sharp increase in controlled trials had, however, occurred.

Department of Orthopedics, University Hospital, Lund, Sweden. Tel +46 46-17 15 10. Fax -13 07 32
Submitted 94-02-13. Accepted 94-05-01

Recently, a randomized study on the morbidity of radiographic arthrosis (OA) of the knee was reported (René et al. 1992). The treatment given in the study was restricted to monthly telephone calls by trained lay personnel, while the control group received routine medical attention. Significant effects from treatment were recorded on the Arthritis Impact Measurement Scale (AIMS) regarding pain and to a lesser extent physical function. The improvement was on a par with that observed in studies on a variety of systemic pharmaceutical agents. During the last year a clinical trial of an agent aimed at reducing pain by intra-articular injections in joints suffering from preradiographic OA was conducted (Dahlberg et al. 1994). The results of this double-blind, randomized trial appeared promising, with about 35 percent of the patients experiencing a distinctly positive effect. However, when the code was broken, a majority in the group that experienced improvement had received placebo. The injections as such thus proved to be a potent means of treatment, but the injected substance was of little importance. This observation confirms that of Miller at al. (1958).

It is overwhelmingly agreed that accurate interpretations of the results of treatment in the clinical context can only be reached by controlled trials (Chalmers et al. 1992). In recent years, a number of well-designed investigations have refuted procedures of treatment which often have received acceptance on the strength of history or for theoretical reasons (van der Linden and Larsson 1979, Bretlau et al. 1984, Clay et al. 1991, Chang et al. 1993).

In order to assess the frequency of well-designed studies in the orthopedic literature, a search study in Medline was conducted.

Methods

On October 26, 1993 (first search) and April 2nd, 1994 (second search) Medline searches were conducted via Datastar (Berne, Switzerland). All published articles fulfilling a defined set of criteria regarding the quality of the underlying research from a design point of view were searched. The hits were filed according to year of publication, country of origin and journal of publication.

The search was made primarily on the DESCRIPTOR fields in the Medline entries in order to avoid hits by uncritical title or abstract formulations. Since the interest of the study was clinical research, the descriptor HUMAN was a primary criterion. The controlled thesaurus terms used for indexing articles in Medline are labeled Medical Subject Headings (MeSH) and are listed in the descriptor field of each reference. Some of these descriptors pertain to the design of the underlying study. In this study, the descriptor RANDOMIZED CONTROLLED TRIALS was considered to represent the highest (3rd) level of quality. Articles indexed by the descriptors RAN-

Table 1. Medline search criteria and number of hits based on second search

Level	Algorithm	Description (descriptor field)	Number of hits
		All articles in 8 orthopedic journals	28,745
	and	human	25,538
1	and	prospective studies	994
2	and	random allocation or double-blind methods	323
3	and	randomized controlled trials*	138

* as searched in the publication type and descriptor fields.

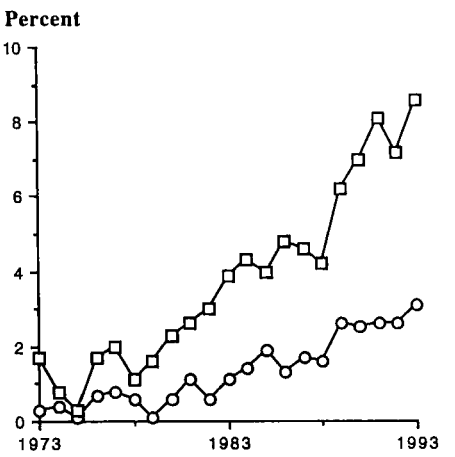

Figure 1. Yearly distribution of well-designed articles as a percentage of all articles published in the eight most cited journals in general orthopedics. □ represents level 1 and ▪ represents levels 2 and 3 based on second search.

DOM ALLOCATION or DOUBLE-BLIND METHOD were considered as representing an intermediate (2nd) level, while the descriptor PROSPECTIVE STUDIES was considered to represent a baseline (1st) level of clinical research (Table 1). Additionally, the publication type field was searched. In this field RANDOMIZED CONTROLLED TRIAL is entered when appropriate.

The orthopedic journals studied were selected according to the Science Citation Index (SCI) reports. The 8 general journals with the highest impact factor and 3 specialized international journals were selected for the study. The SCI report also includes a notation according to the average age of the cited articles as an additional quality parameter (Table 2).

Additionally, the 3 most cited journals in internal medicine (*New England Journal of Medicine, Lancet* and *Annals of Internal Medicine*), surgery (*Annals of Surgery, British Journal of Surgery* and *Surgery*), urology (*Journal of Urology, British Journal of Urology* and *Urology*) and infectious diseases (*Journal of Infectious Diseases, Journal of Infection* and *Infection*) were searched using the same criteria.

For validation purposes, 208 randomly chosen articles were indexed by the authors independently and blindly. Of these we judged that 110 addressed the issues discussed in this article. There was a good but not perfect correlation between the 2 assessments made by us and this also correlated well with the actual indexing of these articles in the *Index Medicus* (r^2 0.55–0.65, $P < 0.0001$). In roughly 50 percent of the articles there was 3-way agreement. In about 25 percent, the 2 authors agreed versus Medline, while in the remaining 25 percent Medline and 1 of the authors agreed against the other author. On 2 occasions there was a complete disagreement.

Results

A total of 28,745 articles had been published in the general orthopedic journals. Of these, 25,538 concerned a clinical study as defined by the descriptor HUMAN and, of these, 994 fulfilled baseline quality criteria, while 323 fulfilled the more stringent criteria of levels 2 and 3. 138 were indexed as RANDOMIZED CONTROLLED TRIAL in the publication type field or in the descriptor field. The percentage of well-designed studies in general orthopedic journals, including all 3 levels, ranged from 1.7 to 6.6. Including only articles of levels 2 and 3 yielded a percentage ranging from 0.4 to 2.4 (Table 2). The percentage of well-designed studies in the specialized journals was somewhat higher, including the baseline level, while there was little difference regarding the more stringent criteria of levels 2 and 3. Of the general orthopedic journals, *Acta Orthopaedica Scandinavica* had published the highest percentage of well-designed studies using the baseline criteria while some of the specialized journals had higher percentages using a higher threshhold.

The trend in time was positive, with an increasing number of well-designed studies being published each year. The percent increase was also favorable despite the increasing over-all number of published articles. In 1973 the percentage of level 1 articles or better was 1.7, increasing to 8.6 in the year 1993 (Figure 1).

The mean percentage of well-designed studies was somewhat lower in orthopedics than in the other

Table 2. SCI Impact factor and percentage of well-designed studies in journals in 1989–1990 based on first search

Journal	SCI Impact factor*	Cited T 1/2	Level 2–3	All levels percent
General orthopedic journals				
J Bone and Joint Surgery (Am+Br)	≈0.802	7.0	1.1	3.7
Clinical Orthopedics	0.684	>10.0	0.7	2.8
Acta Orthopaedica Scandinavica	0.617	>10.0	2.4	6.6
Arch. of Orthopaedic and Trauma Surg	0.287	6.8	1.2	3.1
International Orthopaedics	0.194	8.6	1.5	4.7
Orthopedics	0.190		1.3	2.4
Orthopäde	0.098		0.4	1.7
Revue Chirurgie et Orthopedie	0.072	8.6	1.2	2.4
Mean	0.364		1.2	3.4
Mean 3 most cited	0.701		1.4	4.4
Specialized orthopedic journals				
J of Arthroplasty			1.6	9.1
Spine	0.541	6.1	1.7	5.6
Foot and Ankle	0.177	7.1	0.1	2.9
Other specialities (mean 3 most cited)				
Internal Medicine	16.102		3.8	5.6
Surgery	2.224		2.7	5.8
Urology	0.992		2.3	4.0
Infectious Diseases	2.423		4.0	6.6

* Science Citation Index, 1991 edition. The impact factor = the number of citations to a journal during the years 1989–1990 divided by the total number of articles published by the journal during that time.

areas of medicine (Table 2), the difference, however, was not significant (ANOVA).

Non-parametric regression analysis showed no correlation at all between the SCI Impact Factor and the quality percentage of the different journals as defined here.

The 323 hits at level 2 or better were checked regarding the country of origin. The Scandinavian countries were greatly over-represented, one and a half times the number produced by all of North America. Most of the remaining articles originated in Europe (Table 3).

Table 3. Distribution of articles representing research including a randomization and/or blinded situation reported in the 8 most cited journals in general orthopedics according to country of origin based on second search

Region	Number of hits (level 2–3)
Scandinavian countries	135
North America	95
Great Britain	40
Germany	14
Remaining EU	24
Miscellaneous	15
Total	323

Discussion

Index Medicus has been available in its printed form since 1879. The data base includes articles published 1966 or later and forms the basis of this study. The study performed in this report is certainly open to criticism simply because quality of research cannot be judged as generally as has been done here; our quality indicators are operational for the context of this article only. Even though the indexing of a sample of published articles correlated quite well with our views, obvious mistakes in indexing in *Index Medicus* were identified. Moreover, it was often impossible to judge whether a study was conducted prospectively or represented retrospective production control. However, such errors in indexing should be equally distributed and it is improbable that particular journals or specialities are misrepresented. Hence, the indexing of Medline was accepted as the truth, not by virtue of accuracy but by virtue of objectivity, i.e., blindness of the indexers. For comparison purposes this appears adequate. Further, the Medline database is not constant. Particularly, MeSH headings change and of the terms used in this search only PROSPECTIVE STUDIES has been used since the beginning. RANDOM ALLOCATION and DOUBLE-BLIND METHODS were added in 1978 and RAN-

DOMIZED CONTROLLED TRIAL(S) in 1989. Before 1989 CLINICAL TRIALS was used, but a check revealed that this term was used less stringently and it was not included in this study.

The results of our study indicate that surprisingly little quality research, as defined in this study, is actually published. The percentage found here is far below 10 percent of all articles, and the field of orthopedics does not appear to fare worse than other fields of medicine. The major reason for the large number of well-designed articles as defined here found in *Acta Orthopaedica Scandinavica* is most certainly that researchers in the Scandinavian countries are able to perform prospective and controlled studies to a larger extent. This probably reflects the organization of health care in these countries, with a mostly economically non-competitive situation, where collaboration and the formation of homogeneous materials is more easily organized.

We wished to investigate how the outcome of a treatment has been studied and reported. Much research is performed, which, by the nature of the particular question at hand, is not addressable by controlled designs. These may be laboratory studies, diagnostic reports, cross-sectional designs, case reports, etc. The assessment made here indicates that in about half of all published articles these issues are not applicable. This, however, still leaves a large number of articles to represent the uncontrolled, retrospective report. When the perception of reality in our field is so volatile that it is contaminated by mere telephone calls from lay personnel, the demand for meticulously controlled studies seems imperative. The strong placebo response of surgery as such was demonstrated more than 30 years ago (Cobb 1959, Beecher 1961). Hence, in orthopedics, where a surgical procedure is often studied, this demand becomes even stronger.

The most absolute and objective of the natural sciences, physics, has reached an insight that man can never learn the true state of matter because as soon as measuring equipment is brought to bear on an experiment and an observer–observed pair is established, they start to interact (Heisenberg 1971). When fundamental pieces of matter like electrons and photons can change their inner nature when they are merely looked upon, one must accept that biases on the part of the patient and the investigators can play havoc with the objective truth in the infinitesimally more contaminated clinical investigation; the doctor–patient experimental set-up is certainly a situation where interactions exist. This interaction, of which placebo response is but one example, is termed bias and acts, in both directions, to obscure the true state of affairs.

Bias refers to any systematic error arising from the design and performance of a study and it does not imply willful manipulation of data, fraud or the like. Rather bias lies in the subconscious of the patient and the examiner. It comes in many guises (Rudicel and Esdaile 1985).

One form is usually termed *susceptibility bias* and concerns the fact that when the results of a study group are compared to those of other groups, historically or in the same study, the prognosis of the different groups may vary. Patient selection may vary, additional medical treatments may differ and so on. Randomization is the method of choice to avoid this pit-fall and additional strength is added by the fact that also factors of importance for the prognosis unknown at the time of the study are liable to be evenly distributed between the groups. If specifically important prognostic factors are identified, stratification of the material accordingly can be done.

Another form of bias is termed *performance bias*. When, for example, 2 surgical procedures are compared, they must both be performed with the same skill. This is rather self-evident. Not so evident, however, may be the fact that the follow-up should, for example, include the same number of visits and the patients in both groups should be met with the same enthusiasm. Comparisons between surgical and conservative modes of treatment pose special problems here and the difficulty with, for example, historical controls becomes easily discernible.

Detection bias alludes to the fact that the outcome must be measured by the same criteria. Even if these criteria are established beforehand, they are often not measured by objective means but rather by soft data such as pain, function, etc. Bias of this sort acts on both the patient and the surgeon. The nature of patient bias is, at least superficially, easily appreciable. Expectations, the authority of the doctor, overcoming the psychologic resistance to participate in lengthy, sometimes painful—and expensive—investigations and treatments act in such a way as to make the patient expect improvement (placebo response). This improvement of symptoms must not be misinterpreted as if the patient is being cheated. In the early 1960s, ligation of the internal mammary artery was used to treat coronary angina. The method was found to be effective symptomatically and exercise-ECG revealed normalization of inverted T-waves (Beecher 1961). Not until a blinded test was performed was it revealed that patients, who were sham-operated fared just as well (Cobb 1959). Even their inverted T-wave disappeared! *Bias of the surgeons* is perhaps less often discussed. Surgeons may have a

vested interest in their own procedure or device (Rudicel and Esdaile 1985), making them subconsciously partial. The close collaboration between orthopedic surgeons and commercial parties today is a particular point.

The best method to avoid both detection and (to an extent) performance bias is double-blinding, where neither the patient nor the assessor know what treatment was given. This is often difficult in orthopedics. Blinding of the surgeon during an operative procedure is not an option, but an independent, impartial assessor can be used instead. When comparing surgical and conservative treatments, a double-dummy technique can be envisaged, but is rarely used. Here, all patients receive 2 treatments, 1 surgical and 1 non-surgical, where one randomly is of sham character. Whether this is ethically permissible must be determined in the individual case.

Transfer bias, finally, occurs when there is an incomplete follow-up. If there is a 15 percent loss, this can occur because the patients were so happy with the results that they did not need to go to the doctor anymore, or because they were so displeased that they were reoperated at another center. The difference concerning the results of the study is obvious.

Possibly, one can add the recent phenomenon of *computer bias*. With powerful statistical packages available for uncritical use, correlations and influences between factors, which were not a priori defined as outcome parameters before the study started can often be found. Such posthoc analyses can only serve to generate hypotheses which should be tested in new trials before being accepted. Indeed, with an α (*P*, i.e., probability) of 0.05, 20 posthoc analyses will, by definition, identify a difference where none in reality may exist. In statistical nomenclature this is called a Type I error.

The opposite problem occurs when a false null-hypothesis is accepted as correct and is called a Type II error. The chances for this to occur in a study is termed the ß of the study and 1-ß, i.e., the chances avoiding the Type II error are termed the power of the study. Consider treatment of unstable trochanteric hip fractures and a new device, which promises to cut the complication rate by one half, from, say, 10 percent to 5 percent. With a chosen level of significance, i.e., the level at which the null-hypothesis is rejected, of 0.05 (α) and a power, i.e., probability to find a difference which is indeed present, of 0.8 (1-ß), this would require about 500 patients in each of the study and control groups. This material would take about 15 years to amass at the Department of Orthopedics in Lund although we treat about 400 hip fracture patients yearly. This problem is seldom commented on in published articles and the power is often low, indeed to the extent that also powerful effects of treatment are not identified (Freiman et al. 1978). This dilemma has previously been termed "the squeeze" in this journal, i.e., the squeeze between documented success and the azymtot of Utopia (Bauer 1992). For the scientist there are 2 alternative, diametrically opposite, solutions; multicenter studies (Knutson et al. 1994, Malchau et al. 1993) to provide a sufficient number of patients or to go elsewhere in medicine and find more virgin pastures for research endeavors.

While the effect of bias on any study should be avoided, the placebo response should not be regarded with suspicion, as something false or unethical. It is "the single most potent and versatile tool for relieving the sufferings that man is heir to" (Moertel et al. 1976), and represents the kind of comfort that every doctor should try to give (Ernst et al. 1991). Indeed, in the field of sports medicine, this placebo response may have directly therapeutical or preventive effects. In top athletics, the effects of psychologic training is realized. Mental conditioning gives the athlete a feeling of invincibility which will make him yield less, and became more daring in direct physical contact. This will affect how prone he is to sustain an injury (Jackson et al. 1978). It is, however, important to realize that this placebo effect, given possible cultural differences, is individual and absolute, in time and space, and cannot promote the science of orthopedics. Indeed, the quantity of placebo which we can administer today is probably about the same as Hippocrates was able to do in his time. In other words, we will not be able to improve any treatment by sequentially adding placebo effects to one another. This can only be done by identifying modalities distinctly different from the placebo response in studies that are protected against the effects of bias. In this perspective, the publication explosion in orthopedics today brings some hope; along with quantity comes higher quality.

In our daily practice we owe it to our patients to use a placebo. This will make the individual patient feel better and to achieve this is the primary goal of any doctor. While doing so, we must, however, realize that we are merely applying an inconstant potion of faith and we owe it to orthopedics and to ourselves as scientists to be critical.

References

Bauer G C. What price progress? Failed innovations of the knee prosthesis (editorial). Acta Orthop Scand 1992; 63 (3): 245–6.

Beecher H K. Surgery as placebo. JAMA 1961; 176 (13): 1102–7.

Bretlau P, Thomsen J, Tos M, Johnsen N J. Placebo effect in surgery for Meniere's disease: a three-year follow- up study of patients in a double-blind placebo-controlled study on endolymphatic sac shunt surgery. Am J Otol 1984; 5 (6): 558–61.

Chalmers I G, Collins R E, Dickersin K. Controlled trials and meta-analyses can help resolve disagreements among orthopaedic surgeons (editorial). J Bone Joint Surg (Br) 1992; 74 (5): 641–3.

Chang R W, Falconer J, Stulberg S D, Arnold W J, Manheim L M, Dyer A R. A randomized, controlled trial of arthroscopic surgery versus closed- needle joint lavage for patients with osteoarthritis of the knee. Arthritis Rheum 1993; 36 (3): 289–96.

Clay N R, Dias J J, Costigan P S, Gregg P J, Barton N J. Need the thumb be immobilised in scaphoid fractures? A randomised prospective trial. J Bone Joint Surg (Br) 1991; 73 (6): 828–32.

Cobb L A. Evaluation of internal-mammary-artery ligation by double-blind technique. N Engl J Med 1959; 260: 1115–8.

Dahlberg L, Ryd L, Lohmander S. Effects of intra-articular injections of hyaluronan in the human early-arthritic knee. A prospective, randomized study. Arthritis Rheumat 1994, In press.

Ernst E, Saradeth T, Resch K L. The powerful placebo (letter). Lancet 1991; 337 (8741): 611.

Freiman J A, Chalmers T C, Smith H, Kuebler R R. The importance of Beta, the Type II error and sample size in the design and interpretation of the randomized control trial. N Engl J Med 1978; 299 (13): 690–4.

Heisenberg W. Physics and beyond. Harper & Row, New York 1971.

Jackson D W, Jarrett H, Bailey D, Kausek J, Swanson J, Powell J W. Injury prediction in the young athlete: a preliminary report. Am J Sports Med 1978; 6 (1): 6–14.

Knutson K, Lewold S, Lidgren L, Robertsson O. The Swedish knee arthroplasty project. A nation-wide multicenter study of 30,130 knees 1975–1992. Acta Orthop Scand 1994; 65 (4): 375–86.

Malchau H, Herberts P, Ahnfelt L. Prognosis of total hip replacement in Sweden. Follow-up of 92,675 operations performed 1978-1990. Acta Orthop Scand 1993; 64 (5): 497–506.

Miller J H, White J, Norton T H. The value of intra-articular injections in osteoarthritis of the knee. J Bone Joint Surg (Br) 1958; 40 (4): 636–43.

Moertel C G, Taylor W F, Roth A, Tyce F A. Who responds to sugar pills? Mayo Clin Proc 1976; 51 (2): 96–100.

René J, Weinberger M, Mazzuca S A, Brandt K D, Katz B P. Reduction of joint pain in patients with knee osteoarthritis who have received monthly telephone calls from lay personnel and whose medical treatment regimens have remained stable. Arthritis Rheum 1992; 35 (5): 511–5.

Rudicel S, Esdaile J. The randomized clinical trial in orthopedics: obligation or option? J Bone Joint Surg (Am) 1985; 67 (8): 1284–93.

van der Linden W, Larsson K. Plate fixation versus conservative treatment of tibial shaft fractures. A randomized trial. J Bone Joint Surg (Am) 1979; 61 (6A): 873–8.