Spine
Issue: Volume 21(3), 1 February 1996, pp 259-263

Publication Type: [Analysis - Classics in Spine]
ISSN: 0362-2436
Accession: 00007632-199602010-00001
Keywords: clinical trial, study designs, surgery, intervertebral disk displacement

[Analysis - Classics in Spine]

Classics in *Spine*: Surgery Literature Revisited

Bessette, Louis MD[*]; Liang, Matthew H. MD, MPH[*]; Lew, Robert A. PhD[*]; Weinstein, James N. DO, MS[†]

Author Information

From the [*]Departments of Medicine and Division of Rheumatology and Immunology, Harvard Medical School, Brigham and Women's Hospital, Robert B. Brigham Multipurpose Arthritis and Musculo-skeletal Disease Center, Boston, Massachusetts, and the [†]Department of Orthopaedic Surgery, the University of Iowa Hospitals and Clinics, Iowa City, Iowa.

Supported by NIH grant AR36308. Dr. Bessette is a recipient of the Fonds de la Recherche en Santé du Québec.
Acknowledgment date: January 30, 1995.
First revision date: March 27, 1995.
Acceptance date: May 17, 1995.
Device status category: 1.
Address reprint requests to: Louis Bessette, MD; *Division of Rheumatology; Brigham and Women's Hospital; 75 Francis Street; Boston, MA 02115*

Abstract

Study Design: This article reviews the criteria for evaluating the quality of clinical trials.

Objectives: To outline the current methodologic standards by which the validity of controlled trials need to be evaluated.

Summary of Background Data: Weber's study, published in 1983 in *Spine*, is the only randomized trial comparing surgery and conservative management of sciatica in herniated lumbar discs.

Methods: Weber's article is revisited to illustrate basic principles in the design of clinical trials.

Results: Weber's study is a classic in spine surgery and has changed thinking regarding the benefit of surgery in sciatica related to herniated lumbar discs. However, the authors found potentially critical flaws in this study: a large number of crossovers, inadequate sample size, and insensitive outcome measurements.

Conclusions: A randomized, controlled trial is the most rigorous way to evaluate health intervention. Despite the difficulties of performing such studies, investigators should use the most appropriate scientific methodology.

Management of many spine conditions is often based on opinion or on studies with methodologic limitations. A recent synthesis of the literature on the efficacy of surgery for herniated discs was limited by the quality of the studies.9 Many clinical trials used nonstandardized outcome measures or focused on outcome measures of “objective disease,” such as radiographic appearance, rather than on the patient's function.

The randomized, controlled clinical trial represents the most rigorous way to evaluate therapeutic intervention. In an “efficacy” clinical trial, an intervention is tested on well defined subjects under optimal conditions. The analysis of efficacy is used in an “intent-to-treat” analysis, in which all patients randomized are analyzed. In an “effectiveness” trial, one reports on the subjects who complete the intervention. Both kinds of trials define explicit criteria for including and also excluding certain subjects, and randomize all eligible patients into a treatment group and a control (comparison) group. For subjective endpoints, such as pain and function, a placebo treatment arm is necessary because pain may improve up to 35% with placebo.6 To minimize bias, subjects, health care providers, and evaluators should be blinded to the treatment. Finally, appropriate statistical analyses should be performed to determine whether the results occurred by chance or are statistically significant.

Henrik Weber, a generalist who developed strong interests in spine manipulation and retrained in neurology, published his initial finding in 1978 12 and updated it in 1983 in *Spine*.13 This landmark study of spine surgery in the treatment of leg pain related to herniated lumbar discs has profoundly changed thinking about the benefits of surgery. It is a key study for the Agency for Health Care Policy and Research practice guidelines on the management of acute low back pain.2 It is important to judge a study not only on methodologic standards, but also in the context of what was known and what was acceptable at the time of its conduct. Weber's trial was the first controlled evaluation of spine surgery. As the first randomized trial in back surgery, it rendered obsolete

first controlled evaluation of spine surgery. As the first randomized trial in back surgery, it rendered obsolete previous publications based on case series or case reports, and today it remains the only randomized trial comparing surgery to conservative management in herniated discs.9

Weber's study is an unblinded, randomized, controlled trial with a 10-year follow-up comparing the results of surgery and conservative treatment of patients with leg pain who had a herniated disc. The study showed a better result in the surgically treated group at 1 and 4 years' follow-up, but the difference was statistically significant only at 1 year. No difference was seen at the 10-year follow-up. The critical reappraisal of this study illustrates the criteria by which clinical trials can be judged (Table 1) for their methodologic rigor (internal validity) and their relevance to patient care (external validity).

Sources of patients described (including inclusion and exclusion criteria)
Randomization properly done
Baseline comparability reported (including confounding variables)
Same data collection for all arms of the trial
Subjects, caregivers, and assessors blinded to treatment assignment
Blind assessment of outcome
Interventions and performance of the procedure clearly described
Cointerventions monitored
Compliance, drop out, and cross over assessed and monitored
Side effects assessed
Outcomes defined, measurable, valid, and clinically relevant
Appropriateness of statistical analysis

Table 1. Criteria for Evaluating the Quality of a Clinical Trial

[black small square] Source of Subjects

A study should describe the source of patients, the selection process, and other “filters” by which patients come to participate. Do they come from the general population, with a wide spectrum of severity or complexity, or do they come from a referral center that specializes in treating more complicated or severe conditions? Identifying the source of the patients and specifying exclusions helps readers generalize the results and apply them to the kinds of patients they see.3 In general, the results of a well done trial are so much more valid than anecdote that the reader needs to ask if his or her patients are so different that the results of the study can be ignored. Weber studied consecutive patients admitted to the Department of Neurology at Ullevaal Hospital, a referral center. Whether every patient with a symptomatic herniated disc was admitted to the hospital and evaluated for eligibility, and who decided on admission, cannot be deduced. Admission criteria vary among physicians.

Clear, exhaustive, and mutually exclusive inclusion and exclusion criteria should be stated, leaving no doubt as to how to classify every potential patient into the eligible or excluded group. Weber studied only consecutive patients admitted to his hospital with “sciatica” and a myelogram demonstrating a herniated disc, without specifying how diagnoses were made. He did not state how many patients met inclusion criteria, but were not admitted to the hospital, and how many with symptoms other than sciatica had herniated lumbar discs. An extensive literature on interrater reliability or reproducibility of physical signs and assessment of radiographs shows that experts often disagree, so that readers need some evidence that assessments are reproducible.4,5 Exclusion criteria were spondylolisthesis or previous back surgery, but the author did not report how many patients were included or excluded based on these criteria. Documentation should also be provided on all ineligible patients, and the reasons for their exclusion should be given. This information helps clinicians to

ineligible patients, and the reasons for their exclusion should be given. This information helps clinicians to understand how the results might be applied to their practice.

[black small square] Randomization

Randomization, a critical feature of a rigorous clinical trial, tries to balance known and unknown factors that may affect the results of the intervention.8 If randomization is properly done, potential biased allocation to the study group is reduced, and observed differences are less likely to arise from a selection bias.

In Weber's study, three groups of patients with sciatica with objective disc herniation by myelography are described. Only Group 1 is randomized to surgery or no surgery. The criteria for randomization are defined in laudable detail. Group 1 consisted of patients with sciatica not improved after 2 weeks of conservative hospital management who still had radicular pain “provoked by moderate exercise, by sitting position, or by increased abdominal pressure (coughing, sneezing, or defecation), restricted mobility of the spine, defined scoliosis (tilt), positive straight-leg raising and/or persistent weakness of muscle groups.” Why Groups 2 and 3 were excluded from the randomization is not clear.

Group 2 had 67 patients (24% of the total group) with “definite indications for surgery”: “severe and immobile scoliosis, intolerable pain, sudden onset and/or progressive muscle weakness and/or bladder/rectum paresis.” These indications are not mutually exclusive. Certainly, surgery is indicated in patients with bladder/rectum paresis, but scoliosis, intolerable pain, and progressive muscle weakness are subjective. We are not told how many patients had these indications. If many of these patients had the more subjective indications, we might question whether they in fact should have been included in the randomization.

Group 3 (87 patients or 31% of the total group) is even more problematic from the point of view of excluding potentially eligible patients from the randomization. These patients showed “continuous improvement” with conservative hospital management. Unfortunately, “improvement” is not defined. From the 1978 paper, we learn that Group 3 also included patients who refused operation before randomization and had contraindications to surgery. Therefore, we do not know how many patients were excluded before randomization for these reasons. Most of these concerns would have been answered if more baseline characteristics had been detailed and Groups 2 and 3 followed up. The possibility and importance of a selection bias therefore cannot be assessed.

[black small square] Baseline Patient Characteristics Reported, Including Confounding Variables

The description of subject baseline characteristics after randomization is important because it shows whether the randomization is successful and has balanced the factors associated with outcomes between the groups.7,11 Random allocation usually minimizes the potential for noncomparable groups, but the investigator should test statistically that characteristics are equally distributed between groups.7

The baseline variables of the randomized group are described only in the 1978 article (4-year follow-up), but important data, such as duration of sciatica, baseline physical findings, incapacity, and percentage of subjects working, are not given and no statistical testing is done. Using these data, however, we found no statistical difference between the two randomized groups.

[black small square] Equivalent Data Collection for All Arms of the Trial

A well conducted clinical trial requires that subjects in each arm of the trial be followed with the same frequency and that all endpoints be ascertained with the same intensity.7,8 The length of follow-up should permit complete assessment of all clinically relevant outcomes. Failure to detect a potentially important difference may be the result of a short study duration. The recommended minimum observation period for orthopedic surgical studies in 2 years. The 10-year follow-up in Weber's study is remarkable in this regard.

[black small square] Blinded to Treatment Assignment and Blinded Assessment of Outcome

Ideally, clinical trials should be double blinded so that neither the subject nor the assessor knows which intervention was used.8 Double blinding is clearly the best protection against observation bias. In clinical trials of medication with subjective endpoints such as pain or reported function, the drug to be tested might be compared to another drug or a placebo that tastes and looks identical to the drug being evaluated, so as to blind the subject and the assessor. Blinding the patient is rarely possible in surgical trials, for obvious reasons, and Weber's study is not unusual in that regard.

Although it might be difficult to blind patients in a surgical trial, every effort should be made to ensure that the person who assesses outcomes is blinded to the treatment assignments. The rationale behind this is that caregivers may be biased toward one or another intervention, and modify what they do for the patients or how they treat them. Blinded assessment of outcome is critical to minimize observer bias.8 Weber was present at most of the operations and thus was not blinded. He also evaluated the patients at 1, 4, and 10 years. Although interobserver reliability is not an issue, intraobserver reliability and biased ascertainment of the endpoints are of concern. Blinding of the caregivers is possible with surgical interventions. The surgical wound could have been managed independently, and outcome assessment could have been achieved by an independent observer blinded to whether surgery was done.

[black small square] Interventions and Performance of the Procedure Clearly Described

A detailed description of the intervention or the surgical technique is important for replicating the therapy and conducting confirmatory studies.11 The conservative management and the surgical treatment were well described in Weber's study.

Surgical studies should also describe the skill levels of the surgeons performing the operations, or the success of the technical goals. The results of a surgical intervention may be biased by the technique used and the skill of those performing the procedure. The author does not specify how many surgeons participated in the study, and he gives no description of their level of training and expertise.

[black small square] Cointervention Described

Cointerventions or cotherapies used with the tested intervention should be reported. It is important to know whether cointerventions are comparable among study groups. Therefore, they should be standardized and documented.8 For instance, in Weber's study we are not told whether analgesics or nonsteroidal anti-inflammatory agents or physical therapy were used. Cointerventions could affect the primary endpoints or the incidence of side affects or surgical complications.

[black small square] Compliance, Dropout, and Crossover Assessed and Monitored

The interpretation of any clinical trial result must take into account the extent to which the study participants adhere to their particular treatment assignment. In an intervention trial, three features are important to document: compliance, dropout, and crossover. A high rate in any of these three makes it more difficult to demonstrate a difference between treatment groups.8 In surgical trials, compliance with the surgery is seldom an issue, whereas compliance with cointerventions, loss at follow-up, and crossover may be. Crossover is a major issue in surgical clinical trials because some patients in a nonsurgical group may have surgery.

In Weber's study, 17 people originally assigned to the conservative group (25% of this group) had surgery within the first year of follow-up. Despite this high proportion, the analysis done with these 17 patients included showed a statistically significant better result for the surgical arm at the 1-year follow-up. However, the high rate of crossover in this study decrease the statistical power to detect differences between the conservative and surgical groups at 4 and 10 years of follow-up. This study is nevertheless remarkable for its very low rate of dropout (only two patients lost at follow-up over 10 years)

In any case, noncompliant, dropout, and crossover patients should be included in the primary analysis of every randomized trial. This is called an intent-to-treat analysis, in which all randomized patients are analyzed in their primary assigned group. Techniques for analyzing the results of a study in which patients crossover to the surgical arm must account for when the change occurs, and build on life table methods.1 Therefore, it is essential to obtain as much data as possible on these subjects, including information on their clinical characteristics, the reasons for noncompliance, dropout, or crossover, and their outcome. The Weber articles give a good description of the outcome for these 17 patients, but does not describe in detail why they had surgery (except for “aggravated pain”); nor is it clear who decided on the surgery or where it was done.

[black small square] Side Effects or Complications Assessed

Controlled clinical trials of medical and surgical interventions should systematically assess side effects, toxicity, and surgical complications. In the last analysis, whether an intervention is effective must be balanced against the risks that the patients take with potentially serious or minor side affects. Whether these were

assessed and whether they were done blindly is not clear in Weber's study. In the current climate of health reimbursement, one should also add an economic evaluation to assess the cost-benefit ratio or cost-effectiveness of a new or expensive technology.

[black small square] Outcomes Defined, Measurable, Valid, and Relevant

Outcome events or endpoints should be well defined, reliable (reproducible), valid (measure what they are supposed to measure), and relevant (sensitive to clinically meaningful changes). Because death or complete cure are rare events in spine care, relevant outcomes should be chosen based on detailed, explicit criteria that can be reproduced by other investigators.

The endpoints measured in this study included a spine and neurologic examination, psychosocial evaluation, work capacity, pain, analgesics used, and ability to participate in leisure activity. However, the principal outcome is the author's evaluation of the patient assessment based on an "subjective statement". (Table 2). The patients were asked to endorse a variety of statements, which unfortunately are neither mutually exclusive nor collectively exhaustive, and to choose among the alternatives of completely satisfied, satisfied, or not satisfied, which are not well defined. Furthermore, the reliability and validity of the scale are not known.

Author's Evaluation	Patient's Statement
Good	Completely satisfied
Fair	Satisfied, lesser complaints
Poor	Not satisfied, partly incapacitated
Bad	Completely incapacitated for work due to chronic back pain or sciatica

Adapted from Weber.[13]

Table 2. Principal Outcome Assessment in Weber's Study Based on Patient's Subjective Statement

The published results are vague as to how the patient statements were obtained. Was the patient's assessment obtained by questionnaire, by an open-ended interview, or both? These factors influence the response obtained. If the assessment is done by interview, the interviewer should be blinded to the hypothesis of the intervention. In this study, blinding was not achieved. However, the information was collected by someone other than the operating surgeon, which reduced bias. For any kind of interview, the prompts or elicitations of responses need to be carefully scripted and standardized to avoid interpretation or biased presentation (e.g., How much pain are you having? vs. You're not having pain, are you?).

Most important, one needs to ask whether the outcomes or endpoints are clinically meaningful to the patient. Are they surrogate measures of a desired effect, such as a radiograph showing restoration of anatomic relationships, or are they more relevant to the patients' success, symptoms, function, or quality of life?

[black small square] Appropriateness of Statistical Analysis

Power

Ideally, the number of subjects studied should be large enough to make it likely that the study will have a statistically significant result. A typical study compares two group averages. Averages that differ by little suggest no difference, whereas averages that differ substantially suggest a significant difference. Assuming realistic expected differences in outcome between the proposed intervention and the alternative treatment or placebo, a typically adequate sample size makes the probability of rejecting the null hypothesis of no difference 80%. The typical test criterion for rejecting the null hypothesis is a *P* value less than 5%. More broadly, the statistical test pits those who expect no difference in outcome against those who expect a substantial difference. If there is no

difference, then the test will be significant only 5% of the time. If there is a substantial difference, then the test will be significant 80% of the time; in other words, the test has a power of 80% to detect such a substantial difference. In this contest between hypotheses, a large enough sample size provides decisive evidence in evaluating an intervention.8 In studies showing no effect of a treatment, this may result from inadequate sample size, insensitive measures of endpoints, or a true no effect.

Weber's study, viewed in retrospect, did not have adequate statistical power. With 126 patients randomly assigned to two treatment arms, the difference in the proportion of favorable outcomes between those treated with and (initially) without surgery would have had to exceed 25% for there to be an 80% chance of finding this difference. With 800 patients, the study would have had 80% power to detect a statistically significant result based on a difference of 10% or larger.

Establishing a Small Set of Main Comparisons

Ideally, each study should have only one main hypothesis, with a specific protocol for testing that hypothesis. Naturally, given the extraordinary effort that goes into accumulating data and adhering to rigid protocols, investigators prefer to use the data to address many research questions. This undermines the usefulness of the *P* values and the power calculations, both of which assume the study has only one main question. Performing many tests of statistical significance with a *P* value of 5% for each test increases the likelihood of finding a statistically significant result by chance alone. Lowering the level of significance is a standard technique for avoiding this pitfall. Current practice is to require that significant *P* values be 1% or less, depending on the number of extra tests, as determined by methods such as the Bonferroni correction.10

The lumbar disc study made many comparisons and valiantly tried to deal with the third group of 17 patients who eventually had surgery. Nevertheless, the extra and special statistical comparisons were treated in the statistical analysis as if they had been among the main planned comparisons. None of these was controlled or corrected for the chance that, by a fluke, one or two comparisons among many had erroneously attained statistical significance.

[black small square] Conclusion

The randomized clinical trial represents the gold standard for intervention studies. However, a critical appraisal of a study is necessary to evaluate its validity. Weber's study is one the first randomized clinical trials in spinal surgery, and a classic study. It changed the thinking regarding a surgical role for the herniated disc in a time when the “evidence” was based on case reports, uncontrolled series of cases, and personal experience.

How should one interpret this study? The important and potentially critical defects include the large number of crossovers, the inadequate sample size, and the insensitive outcome measurements. Nevertheless, the author did not stretch the data, and tested obvious hypotheses and found almost no significant differences.

Most physicians consider surgery to be the best option for some patients with herniated lumbar discs with intolerable pain not responsive to conservative treatment. However, based on the existing literature, we believe that surgery is probably not any better than conservative treatment in the long term. No single small trial can be definitive and, as with other questions, only replication can solve the issue. It is interesting to speculate whether Weber's manuscript would have been accepted for publication in *Spine*, for instance, by current standards. We believe it would because it was randomized and controlled, and because objective data, although perhaps with caveats, is far preferable to anecdote.

In this review, we have discussed a paper from the perspective of an ideal study done in an ideal setting. Experimental investigators know well the difficulties of executing perfect studies in human studies. Weber's study is the best of its genre in this field. However, one should not disregard better scientific methods merely because of their difficulties. The payoffs are much greater than in the countless studies done without appropriate consideration of scientific principles.

Acknowledgment

The authors thank Dr. Henrik Weber for his help and support.

References

1. Armitage P. Survivorship tables. In: Wiley J, ed. Statistical Methods in Medical Research, 2nd ed. New York: Halsted Press, 1971:408-14. [Context Link]

2. Bigos S, Bowyer O, Braen G, et al. Acute Low Back Problems in Adults. Clinical Practical Guideline No. 14. AHCRP Publication No. 95-0642. Rockville, MD: Agency for Health Care Policy and Research, Public Health Services, U.S. Department of Health and Human Services, December, 1994. [Context Link]

3. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre. How to read clinical journals: V. To distinguish useful from useless or even harmful therapy. Can Med Assoc J 1981;124:1156-62. ExternalResolverBasic | Bibliographic Links | [Context Link]

4. Deyo RA, McNiesh LM, Cone RO. Observer variability in the interpretation of lumbar spine radiographs. Arthritis Rheum 1985;28:1066-70. ExternalResolverBasic | Bibliographic Links | [Context Link]

5. Feinstein AR. A bibliography of publications on observer variability. J Chronic Dis 1985;38:619-32. ExternalResolverBasic | Bibliographic Links | [Context Link]

6. Gaupp LA, Flinn DE, Weddige RL. Adjunctive treatment techniques. In: Tollison CD, ed. Handbook of Pain Management. Baltimore: Williams & Wilkins, 1994:108-35. [Context Link]

7. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature: II. How to use an article about therapy or prevention. A: Are the results of the study valid? JAMA 1993;270:2598-601. ExternalResolverBasic | Bibliographic Links | [Context Link]

8. Hennekens CH, Buring JE. Intervention studies. In: Mayrent SL, ed. Epidemiology in Medicine. Boston: Little, Brown and Co., 1987:178-212. [Context Link]

9. Hoffman RM, Wheeler KJ, Deyo RA. Surgery for herniated discs: A literature synthesis. J Gen Intern Med 1993;8:487-96. ExternalResolverBasic | Bibliographic Links | [Context Link]

10. Neter J, Wasserman W, Kutner MH. Analysis of factor effects. In: Irwin RD, ed. Applied Linear Statistical Models, 2nd ed. Irwin City: Homewood, 1985:566-601. [Context Link]

11. Rudicel S, Esdaile J. The randomized clinical trial in orthopaedics: Obligation or option. J Bone Joint Surg [Am] 1985;67:1284-93. ExternalResolverBasic | Bibliographic Links | [Context Link]

12. Weber H: Lumbar disc herniation. J Oslo City Hosp 1978;28:33-64. [Context Link]

13. Weber H. Lumbar disc herniation: A controlled, prospective study with ten years of observation. Spine 1983;8:131-40. Ovid Full Text | ExternalResolverBasic | Bibliographic Links | [Context Link]

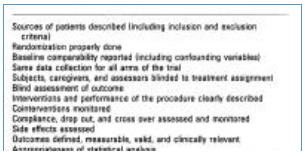
Key words: clinical trial; study designs; surgery; intervertebral disk displacement

---

IMAGE GALLERY

Select All

Export Selected to PowerPoint



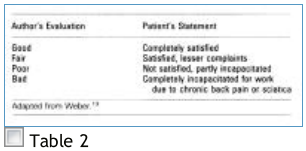


Table 2

Table 1

Back to Top