Beate Hanson
Branko Kopjar

Clinical studies in spinal surgery

Received: 13 September 2004
Revised: 13 January 2005
Accepted: 24 February 2005
Published online: 7 June 2005
© Springer-Verlag 2005

B. Hanson (✉)
AO-Clinical Investigation
and Documentation, Clavadelerstrasse,
7270 Davos, Switzerland
E-mail: beate.hanson@aofoundation.org
Tel.: +41-81-4142500

B. Kopjar
Department of Health Services H682
Health Sciences Center,
University of Washington,
Box 357660, Seattle,
Washington 98195-7660
Tel.: +1-206-2213349

Abstract There is a growing interest in applying evidence-based approaches also in orthopedic surgery. Despite many challenges to the validity of clinical trials in orthopedic surgery, it is possible to conduct well-designed trials in this field and to produce clinically important findings and reasonably valid conclusions about effectiveness, prognosis and diagnosis in orthopedic surgery. We describe the main principles for conducting clinical trials in this field as well as some of the most common errors and ways to avoid them.

Keywords Surgical treatment · Spine surgery · Clinical trial design · Bias · Confounding

Introduction

Evidence-based medicine (EBM) is the integration of the best research evidence with clinical expertise and patient values [2]. There is a growing interest in the application of the principles of evidence-based approaches to orthopedic surgery. For example, the *Journal of Joint and Bone Surgery* has recently introduced grading of the strength of the evidence for every published article. Courses are organized for orthopedic surgeons to introduce concepts of strengths and weaknesses of different types of research designs.

There are some inherent challenges in conducting high-quality clinical research studies in orthopedic surgery, which contribute to the fact that the strength of scientific evidence in the field of clinical orthopedic research is lower than in many other medical disciplines. Some conditions are rare and it is difficult to collect

sufficient number of cases for running a trial. Surgeons may be more skilful in one particular technique making it difficult to compare the results across the surgical techniques in a randomized fashion. The blinding of patients and evaluators is often not possible. Patients may decide to change their treatment assignment after enrollment crossing over to a different treatment arm.

On the other hand, many studies could have been improved if the researchers had better tools and better mastering of clinical research methods. Many deficiencies occur early on in the planning phase of the trial and cannot be corrected later during the course of the study. The aim of this paper is to offer basic research concepts for orthopedic surgeons who are already conducting or are planning to do clinical research.

Two main disciplines required for conducting good clinical research are clinical epidemiology and biostatistics. Clinical epidemiology focuses on the research

study design and trial conduct issues while biostatistics focuses on data analysis issues. This paper focuses on clinical epidemiology topics, although some basics of the statistical approaches used in clinical orthopedic research will be provided.

Types of clinical epidemiology studies in spine surgery

Spine surgery literature contains a number of different studies with different objectives. A significant number of these studies fall into the category of clinical epidemiological studies. They investigate the topics of treatment effectiveness which aim to establish absolute and relative treatment success; studies of diagnostic approaches which aim to establish the validity of specific diagnostic approaches; studies of harmful effects of treatments and clinical maneuvers or of the risk factors for diseases; and prognostic studies.

Study design

Different study designs and their main advantages and disadvantages are shown in Table 1 and described in more details in the following text.

Observational studies

Studies in orthopedic surgery often use data that are obtained from medical records or are collected prospectively from patients. In these studies, the decision concerning treatment choice is left to the standard clinical approach consisting of the interaction between the patient, physician and other factors rather than being enforced by a study protocol as in experimental studies. The advantage with this approach is that the study results are closer to a real-life situation in regular practice. At the same time these studies have significant limitations in the extent to which they may reach valid conclusions.

There are several types of observational study designs including case reports, case series, cohort studies and case-control studies.

In a case report, data from a limited number of patients are presented without comparison to other treatments. Case reports serve several important purposes. They are used to describe rare clinical events and may give possible explanations and hypotheses. Case reports generate hypotheses rather than test them.

For example, in a report of two patients, Waelchli et al. [7] describe the development of acute osteoporotic vertebral compression fractures of the instrumented vertebral body adjacent to the fractured vertebra. He explains it as a consequence of removing the pedicle screws in those two patients who were previously treated for vertebral lumbar burst fractures.

A case series is the documentation of a larger group of patients (usually ten or more) with a common feature (e.g. the same disease or the same treatment approach). In terms of treatment effectiveness, case series are used to describe experiences with new or complex treatment approaches. These studies provide a bridge between developmental approaches and clinical research and latter, clinical practice.

For example, in a case series of 36 patients, Suetsuna et al. [3] describe the clinical and radiographic results of anterior cervical fusion using porous hydroxyapatite ceramics. The outcomes were reviewed retrospectively from available clinical notes and the average follow-up time was 4.5 years, ranging from 2 years to 7 years.

Sometimes, researchers compare their case series with historical controls and controls from the literature. Such a comparison improves the value of a case series by bringing it into perspective with other treatment options.

For example, none of the 28 patients with degenerative disc disease who were treated with autogenous

Table 1 Types of study design, their advantages and disadvantages

Type of study	Type of design	Advantages	Disadvantages
Observational studies	Case report Case series Cohort studies	Used for rare clinical events Experiences with new or complex treatments Compare two treatments Resemble "real life" clinical situations	No comparison group No comparison group Prone to confounding
	Case-control studies	Small sample size Short duration	Prone to confounding and bias
Experimental studies	Randomized controlled studies	Avoidance of confounding	Expensive Limited generalization Difficulties in study recruitment and conduct

bone graft fusion and pedicle screw/plate fixation had pseudoarthrosis [9]. Data from the literature suggested that the rate of pseudoarthrosis in controls who have not received pedicle screw/plate fixation was about 32%.

Case reports and case series should include all or an unbiased selection of patients. A biased selection of cases may result in the presentation of cases that are not representative of the average clinical population. This may produce misleading conclusions about the effectiveness of a particular treatment approach. Whilst case reports and case series are important in order to generate hypotheses they have substantial disadvantages, the main one being the absence of comparison groups. Any observed outcome could therefore be due to other factors (e.g. improvement in other techniques) and would occur independently of the application of a new treatment approach.

Cohort studies (incidence studies, longitudinal studies, prospective studies) follow-up groups of patients over time in order to observe and compare the occurrence (incidence) of defined outcomes. Cohort studies can be used to elucidate the outcomes and therefore the effectiveness of treatments in groups of patients treated by different treatment approaches

For example, in a prospective study involving 87 patients Suk et al. compared the outcomes of unilateral (47 patients) and bilateral (40 patients) pedicle screw fixation [4]. In this study, there were no significant differences between the two groups in blood loss, clinically satisfactory results, fusion rate, and complication rate. There were however, significant differences in the duration of operating time, duration of hospital stay, and medical expenses which were higher in the bilateral compared with the unilateral group.

Cohort studies can be prospective (also called concurrent) or retrospective (historical). The latter is based upon a retroactive review of data (e.g. medical records). Data sources for cohort studies can be the existing standard data (e.g. clinical records, administrative data) or, better, the controlled data collected for the purpose of the study. The strength of the cohort study depends upon the availability of a direct comparison of two treatments. The major limitation of cohort studies is that the compared patient groups may differ in important baseline characteristics. These characteristics, which may influence the outcome, are usually referred to as *confounders*. This may create a situation in which outcomes are attributed to one factor (e.g. treatment), when they are actually the result of another factor. In well-designed and thought-through cohort studies, researchers try to obtain information on these confounding variables and use different design and statistical approaches in order to minimize their effect.

Case-control studies compare patients with some definitive outcomes to a control group of patients without such outcomes. Information about exposures (e.g. treatment) is then collected retrospectively. Case-control studies are used for the assessment of treatment failures. They are rarely used to evaluate treatment effectiveness.

For example, Vogelsang et al. [6] studied the association between peridural scarring and recurrent pain after lumbar discectomy. They studied 53 patients divided into two groups, those with and those without radicular pain. After examining MRI findings there was no difference in the amount of fibrous tissue between the two groups of patients [6] suggesting that scarring is not associated with radicular pain.

Case-control studies need a small sample size and can therefore be used to study rare events such as treatment failures or complications. Because of their retrospective nature they are of short duration. The disadvantage of case-control studies is that they are prone to different types of biases which will be presented later.

Randomized controlled studies

In randomized controlled studies patients are allocated to treatments and the outcome measurements are part of the study protocol. The standard approach to treatment allocation is randomization which makes sure that compared groups are similar in important characteristics. The problem of confounding is therefore avoided. Randomized controlled trials are therefore the gold standard for evaluating the effectiveness of a clinical treatment. RCT(s) are widely used in order to evaluate pharmaceutical therapies, screening programs or less invasive treatments. The Cochrane Collaboration maintains a database which contains more than 400,000 bibliographic references to RCTs. However, only a small fraction of these are spine surgery trials. It is difficult to recruit elective surgical patients into randomized trials, particularly those that compare surgical and conservative options. Blinding is often not possible in surgical trials, thus creating an evaluation bias. Patients may crossover treatment arms when switching from conservative to surgical treatment. Patients selected for randomized trials need to comply with detailed and specific inclusion and exclusion criteria, and may not be representative of general patient population. Furthermore, it is not always feasible and ethical to let chance decide about treatment allocation. However, there are still areas where RCTs are not only feasible but also necessary, i.e., in comparing different surgical approaches and different types of implants.

For example, in a multi-center randomized study comparing lumbar fusion versus non-surgical treat-

ment for low back pain, 294 patients were randomized into three types of lumbar fusion and a control group of 72 non-surgically treated patients [1]. At a 2-year follow-up, patients who were treated surgically fared better than the conservative group.

Bias

Bias is any systematic error contributing to the difference between the statistical values in a population and a sample drawn from it. Bias is any trend in data collection, analysis, interpretation, publication or review of the findings what may result in systematic departure from truth. Bias can occur because of an error in the design or conduct of a study. It is not necessarily a subjective desire of a particular outcome. Any type of study, including a randomized controlled trial, is susceptible to bias, but some epidemiological studies are particularly prone to certain types of design-associated biases (e.g. recall bias in a case-control study). *Selection bias* is common in clinical studies and is due to the preferential inclusion of subjects in the study arms. This may result in observing an association between the treatment and the outcome when such an association does not exist. Thus, selection bias may invalidate conclusions. A common selection bias is the inclusion of only those patients who have answered a follow-up questionnaire or for whom the follow-up data are available. Non-responders often differ in important characteristics from responders and the loss of non-responders may affect the validity of the results. Another potential selection bias is the inclusion of patients from one physician only or for whom records are available.

For example, in a study of patients with lumbar laminectomy with or without fusion, of 158 potential cases, 6 patients died and 89 patients could not be located or did not return the questionnaires. The analysis was based on remaining 69 patients [8].

Information bias occurs when information about the study subjects is obtained in an inadequate way (mistakes in abstracting medical records, bias in interviewing, bias in accuracy of remembering past exposures or outcomes (recall bias), bias in adequate reporting of treatment outcomes). Subjective outcomes (e.g. pain, functional status) are more open to information evaluation bias than the objective outcomes (e.g. mortality, revision surgery, infection).

Confounding

Confounding is the main source of bias and a major concern in observational studies. It occurs when observing the true relationship between two variables

(e.g. treatment A is associated with a higher proportion of good outcomes) and jumping to conclusions that the relationship is causal (i.e. the outcomes are better in treatment group A because of treatment A). Confounding is a major issue in observational clinical studies because it may result in erroneous inferences about treatment effectiveness and about risk factors for good or poor outcomes. It is also important to emphasize that confounding may occur because of known as well as unknown factors.

The best way to effectively minimize or eliminate confounding is by conducting a randomized controlled trial. The randomization of subjects assures that potential confounding variables are distributed equally to the treatment arms and thus do not distort the relationship between the exposures and the outcomes. However, the majority of studies in orthopedic spine surgery cannot be conducted as randomized controlled trials.

Other ways to minimize the effect of confounding in observational studies is through the design and the analysis phase of the study. In the design phase, we can match subjects from two groups by potential confounders. For example, we can match patients from treatment group A and treatment group B by severity. After matching by severity, we would know that any association between the treatment and the outcome could not be attributed to differences in severity. In the analysis phase, we can stratify analysis by potential confounder and then analyze effects separately for each stratum. For example, we could stratify patients by severity and then analyze outcomes in each patient group. Within each group, severity would not pose a confounding factor. However, often there are multiple confounders and stratifying is not an option for addressing confounding in the study. In such cases researchers apply advanced multivariate statistical techniques such as logistic regression or multivariate regression to make groups comparable. Statistical multivariate adjustment is the most common way of addressing confounding in observational studies. In orthopedic spine surgery studies, multivariate statistical adjustment for confounders is rarely applied.

Statistical analysis

Statistical analysis has several purposes in a research study. Firstly, it is used to describe results by presenting, for example, frequencies, averages, percentages, standard deviations and medians. These simple measurements provide group descriptions of the patients' features and their outcomes.

Analytic studies (analyzing effects of treatment and risk factors for specific outcomes) rely on testing statistical hypotheses. Testing statistical hypothesis (sometimes called testing statistical significance) is a main application of statistical analysis in clinical studies.

While testing can be technically complicated and is best left to professional statisticians, the role of testing is relatively simple—the aim is to distinguish true differences (associations) from chance. As all research is performed on subject samples, there is always the theoretical possibility that the observed results are due to chance only and that no true difference exists between the compared treatments. Statistical tests help to sort out how *likely* it is that the observed difference is due to chance only. Usually an arbitrary test threshold value (e.g. $\alpha=0.05$) is used to distinguish results that are assumed to be due to chance from those results that are due to other factors. If the probability that the results are due to chance only is less than the threshold value ($P<0.05$) then one assumes that the differences are due to other factors (e.g. difference is due to true differences in treatment effects). If the probability obtained by statistical testing is greater than or equal to the threshold value ($P\geq 0.05$), then it can be concluded that it cannot be ruled out that the differences observed in the study are due to chance, and it is usually assumed that there are “no statistically significant differences between the treatments”. However, the absence of statistically significant differences does not necessarily imply that differences do not exist. It only means that differences *could be* due to chance only.

The situation may occur that one treatment is truly superior to another and yet the study shows statistically non-significant results. This will occur if the *sample size* is too small to detect differences in treatment effectiveness. The ability to detect differences is the statistical power of the study. The power of the study depends on the characteristics of the studied variables and on the sample size. As the characteristics of the variable cannot be changed, the only way to influence the power is to change the sample size. A larger sample size will bring more power to the study.

In addition to statistical testing, more complex statistical analysis can be used to adjust for confounding variables or to perform other adjustments (e.g. Cox regression for unequal follow-up time) prior to testing the hypothesis.

Most studies in orthopedic spine surgery use descriptive statistics and simple un-adjusted statistical tests. While these are appropriate in many cases, there is a need to use more complex designs more frequently in order to better address the many open questions in orthopedic spine surgery. Such studies will require the use of more advanced multivariate statistical techniques.

Conclusions about clinical effectiveness

The previous sections have described many aspects of clinical research in orthopedic spine surgery. It is evident that, as in any other research field, the effectiveness of different surgical treatments has to be examined statistically and then viewed carefully in the light of possible biases and effects of confounding. Inferences about causal relationships should be made sparingly, and only after a careful consideration of the strengths and weaknesses of the design and conduct of the study. Due to many reasons such as ethical considerations, surgeon treatment preferences, challenges in patient enrolment, randomized controlled trials continue to be rarely used in the field of orthopedic spine surgery. The next best alternative is a well-designed, prospective cohort study. There are but few examples of such studies in the literature and majority of studies continue to be case series. The orthopedic spine research community should extend efforts to improve quality of the clinical research thus improving outcomes for large patient groups and make better use of scarce health care resources.

References

1. Fritzell P, Hagg O, Wessberg P, Nordwall A (2001) 2001 Volvo Award Winner in clinical studies: lumbar fusion versus nonsurgical treatment for chronic low back pain: a multicenter randomized controlled trial from the Swedish Lumbar Spine Study Group. *Spine* 23:2521–32 discussion 2532–4
2. Sackett D, Straus S, Richardson W, Rosenberg W, Haynes R (2000) Evidence-Based Medicine: how to practice EBM. Churchill Livingstone, Edinburgh
3. Suetsuna F, Yokoyama T, Kenuka E, Harata S (2001) Anterior cervical fusion using porous hydroxyapatite ceramics for cervical disc herniation: a two-year follow-up. *Spine* J 5:348–357
4. Suk KS, Lee HM, Kim NH, Ha JW (2000) Unilateral versus bilateral pedicle screw fixation in lumbar spinal fusion. *Spine* 14:1843–1847
5. Swiontkowski MF, Chapman JR (1995) Cost and effectiveness issues in care of injured patients. *Clin Orthop* 318:17–24
6. Vogelsang JP, Finkenstaedt M, Vogelsang M, Markakis E (1999) Recurrent pain after lumbar discectomy: the diagnostic value of peridural scar on MRI. *Eur Spine J* 6:475–479
7. Waelchli B, Min K, Cathrein P, Boos N (2002) Vertebral body compression fracture after removal of pedicle screws: a report of two cases. *Eur Spine J* 5:504–506
8. White AH, von Rogov P, Zucherman J, Heiden D (1987) Lumbar laminectomy for herniated disc: a prospective controlled comparison with internal fixation fusion. *Spine* 3:305–307
9. Wood GW II, Boyd RJ, Carothers TA, Mansfield FL, Rehtine GR, Rozen MJ, Sutterlin CE III (1995) The effect of pedicle screw/plate fixation on lumbar/lumbosacral autogenous bone graft fusions in patients with degenerative disc disease. *Spine* 7:819–830