Eur Spine J (2006) 15: S4–S16
DOI 10.1007/s00586-005-1057-5

REVIEW

Rudi Hiebert
Margareta Nordin

Methodological aspects of outcomes research

Received: 1 November 2005
Accepted: 2 November 2005
Published online: 30 November 2005

R. Hiebert (✉) · M. Nordin
Department of Orthopaedic Surgery, Occupational and Industrial Orthopaedics Center (OIOC), NYU Hospital for Joint Diseases, New York University Medical Center, 63 Downing Street, New York, NY 10014, USA
E-mail: Rh44@nyu.edu
Tel.: +1-212-2556690
Fax: +1-212-2556754

Abstract A critical evaluation of existing scientific evidence of treatment efficacy can be an important part of communicating risk and benefits of treatment options to patients during the course of clinical practice. A checklist of key methodological issues to examine when reading a research study is presented and discussed. Steps in reading a paper include: identifying the research question; identifying the manner in which subjects get enrolled in the study; identifying the treatments and outcomes used; identifying the study design and the comparisons being made; evaluating the study methods for the possibility of bias and uncontrolled confounding; assessing whether the statistical analysis used is appropriate for the study design; assessing whether the study has sufficient statistical power to demonstrate hypotheses being tested. Finally, procedures for grading and evaluating evidence, as used by systematic review groups and international best evidence synthesis consensus groups is briefly described.

Keywords Spine · Research methods · Literature review · Tutorial

Introduction

Outcomes research provides evidence of health care practices and interventions in terms of, for example, the ability to function, or the level of pain, the costs or the risks involved in undergoing a surgical procedure [18]. Outcomes research ties patient experience to the type of treatment they receive. In this way outcomes research serves as a powerful tool for improving the quality of care [18].

Drawing conclusions about the value of outcome studies can be daunting. Numerous studies are published, but few actually provide strong evidence for claims of treatment efficacy. For example, to formulate concensus findings regarding management of whiplash-associated disorders, the Quebec Task Force identified titles of over 10,000 whiplash-related articles and conference abstracts published over a 10 year period of time. Of these, 1204 studies met criteria for relevance, and only 294 met initial scientific quality criteria for further, in-depth review. Finally, only 62 studies were evaluated as having sufficient methodologic quality for acceptance into the Task Force recommendations. Consequently it is important to become a critical consumer of research information, to effectively judge whether findings from outcomes studies are trustworthy and relevant. This article is intended to help the reader to better understand the quality of an outcomes study, and how the findings of the study support the conclusions made.

Reading a study

Some systematic means for reading is needed to be able to identify those studies that provide useful information

from those that provide information of no value or which are misleading. A list of steps can be a useful aid for efficiently reading research studies, one example is given in (Fig. 1).

Step 1: identify the research question

Evaluating outcome means asking the question whether one kind of treatment is better than another. What do we mean by better? Examples of outcomes research questions are in the spine surgery literature include: is patient satisfaction, quality of life, return to work, improved function and or pain with those getting one type of treatment as compared to another? Is pain reduced and mobility maintained? Is treatment cost-effective? Do the benefits of a particular treatment outweigh the risks?

A central methodological issue in evaluating outcomes research is whether the research question is worth investigating in the first place. Clinical research, particularly randomized controlled trials, involves subjecting patients to risks they would not otherwise normally encounter in seeking health care. For example, patients enrolled in clinical research are frequently asked to give information that does not pertain directly to decisions about their treatment, such as questions about depression, social activity, feelings about work, application for social compensation, and so on. These questions can be sensitive in nature and expose the patients to risks if disclosed to individuals or organizations that are not part of the research study. Second, patients participating in studies of new, experimental technologies undergo treatments and procedures where the risk of a serious adverse event is largely unknown. Finally, in some study designs, the treatment decision is not left up to the clinician and patient, but up to chance.

- Indentify the research question
- Identify how subjects get enrolled in the study
- Identify the treatments being compared
- Identify the outcomes being assessed
- Identify the study design
- Evaluate for the possibility of bias and confounding
  - Blinding
  - Random allocation
  - Differential drop out
  - Appropriate statistical analysis
- Evaluate for statistical power

Fig. 1 Suggested checklist for evaluating methodologic aspects of outcomes research

These aspects of participating in research fundamentally alter the normal patient–physician relationship. In routine clinical care the physician's obligation is to provide scientifically validated treatment that best meets the patient's needs. In clinical research, the obligation of the physician-scientist is to ensure that study participants are not subject to unreasonable risk for unnecessary or trivial reasons [39]. This means that the research question must have genuine medical, social and/or scientific value, that the research methods used give a good chance of obtaining findings that are trustworthy, and that the value and quality of the new knowledge gained offset risks to the subjects of participating in the study [39]. It also means that the research subject fully understands the nature of the changed doctor–patient relationship, the risk and benefits of the study, and voluntarily agrees to serve as research subject in a study.

Consequently, the first step a reader takes is to understand the purpose and merit of a study's research question. A study's literature review frames and justifies the research question being posed. A reader can evaluate the quality of a study's literature review by looking for certain key features. For example, a good literature review should draw from multiple sources, both electronic and printed [34]. The keywords and search strategy used should be reported. The table of contents of journals specializing in the field can be hand searched, and international resources should be consulted when possible [6]. Finally, when possible experts in the field can be consulted to draw on their own personal libraries and experience to identify important, but hard-to-find studies. These techniques are used by international best evidence synthesis review groups [3, 7, 35, 48, 52] and systematic review groups [41, 51, 54].

Step 2: identify how subjects get enrolled in the study

The procedures by which a study enrolls potential study subjects helps the reader understand to what degree the findings in the study population can be generalized to other clinical populations. This is particularly important in studies of spine surgery because there may be disagreement among clinicians about the specific indications for surgical intervention. For example in a recent review, Carragee [15] found that there was little consensus among practioners regarding management of chronic, disabling low back pain. Some clinicians focus on identifying a specific anatomic feature thought to be responsible for the pain. However, in many cases the specific anatomic features thought to be 'pain

generators' are also present in imaging studies of asymptomatic individuals [15]. Some researchers have suggested that the observed geographic variation in rates of spine surgery can be explained, in part, by local differences in what are considered the clinical indications for back surgery and clinical practice patterns [21].

For this reason it is essential that a good quality study report the exact enrollment procedures used by the study investigators. Ideally, a study protocol should not rely solely on the report of a diagnosis for purposes of inclusion or exclusion; instead, the clinical workup and diagnostic criteria need to be described sufficiently well so that a reader would be able to replicate the study's enrollment procedures [40]. This may be difficult to do in retrospective studies because existing clinical records may only document the diagnosis and the specific procedures used for coming to the diagnosis and decision criteria may be incompletely or inconsistently documented. However, for prospective studies, particularly clinical trials, it is possible to employ uniform procedures for screening and enrollment, and these procedures should be described in the study's publication.

In addition to using uniform clinical procedures to enroll subjects, a good quality study will also report results of the enrollment process. For example, the Consolidated Standards of Reporting Trials (CONSORT), an international group of biomedical journal editors and experts in clinical trial methodology, epidemiology and biostatistics, recommend that the following information be reported: the number of potentially eligible subjects from the study population, the number of subjects screened for inclusion/exclusion, the number of individuals eligible, and the number of subjects enrolling in the study [40]. If significant number of subjects drop out during the enrollment process (for example if a large number of subjects are screened as eligible but fail to enroll in the study), the reasons for the drop must be explored and reported in the study [40].

Step 3: identify the treatments being compared

The next step for the reader is to identify the treatments or interventions the study seeks to compare. The choice of comparisons is the central feature that confers clinical and scientific value to a study. What is an appropriate comparison group? There is no simple answer because the study questions reflect the interests, values and motivations of the clinical and research community to provide optimal care for patients. From a methodological standpoint, however, the treatments compared should be (1) distinguishable, (2) medically justifiable, (3) compatible with the needs of the patient, (4) have reasonable doubt regarding relative efficacy, (5) a mode of administration that is compatible with the methodological requirements of the study (for example, concealment is possible when blinding is needed), and (6) a mode of administration that is similar to real-world clinical practice [38].

It is essential that the treatments (both the experimental and control) are fully and completely described. For example, Brox et al. [12] compared fusion with conservative care among patients with chronic, nonspecific low back pain. In this study, conservative care was defined as a cognitive-behavioral, multidisciplinary program. Brox et al. described the components of the program, the frequency of administration and the duration of treatment. On the other hand, Friztell et al. [28] evaluated fusion against conservative care. In this study, however, multiple, different modalities were used in the comparison group. Fritzell et al. did not capture data on the frequency, duration and type of treatment modalities used among those getting usual care. In this case, while the internal validity of the comparison was not substantially weakened, the lack of a clear description of the control group limits the generalizability of the study findings.

The issue of the use of a placebo in spine surgery research is hotly debated [21, 24]. Some researchers argue that placebo is the only control group that permits a study to evaluate whether or not a new treatment is effective as compared to no treatment or treatment known to be ineffective. For example, when a treatment and control show no differences, one of two situations could have occurred. Both the treatment and control are effective, but equally so; or, both the treatment and control are ineffective [39]. Does the control group have to be a placebo or sham intervention? No, but the reader should understand that in those situations where no placebo or sham treatment control group is used, the best that the trial can do is make statements about relative efficacy, that is, whether one treatment does better than another in the outcomes presented, but cannot make statements regarding overall efficacy where no treatment would have been offered.

Step 4: identify the outcomes being assessed

When critically reading a study, the reader needs to ask three questions about the outcome measure reported being used. First, are the outcome measures meaningful? In the spine literature outcome measure fall into two different categories: patient-oriented outcomes and non-patient oriented outcomes. Patient-oriented outcomes reflect outcomes that are of importance to the patient, for example: pain, function, social and family life, ability to take care of oneself, ability to work and so on. Two of the most common and widely used patient-oriented outcome measures in spine literature include the Roland–Morris Disability Scale [44] and the Oswestry Disability Index [25] (see Table 1). Non-patient-oriented

Table 1 Selected outcome measures used in studies of low back pain

	Domain	Examples
Patient oriented outcomes	General health status and quality of life	SF-36, SF-12, Sickness impact profile, EuroQuol
	Back-specific disability	NASS low back pain instrument, Prolo scale Low back outcome score Dallas pain questionnaire Oswestry disability questionnaire Roland and Morris disability scale
	Patient satisfaction	Patient satisfaction index
	Pain level	0–10 analog pain scale
	Medication use	Medication type and doseage measured pre- and post-operatively Change in medication usage following surgery
	Return to work	Prolo economic scale, return to work, work retention, LBP-related work disability recurrence
Non-patient oriented outcomes	Econometric outcomes	Return to work, work retention, LBP-related work disability recurrence, quality-adjusted years of life, direct and indirect costs associated with surgery and rehabilitation, Prolo economic scale
	Biomechanical	Implant durability
	Surgical	Frequency and type of revision, salvage, replacement
	Radiological	Fusion status

Adapted and modified from Blount et al. (Blount KJ, Krompinger WJ, Maljanian R, Browner BD (2002) Moving toward a standard for spinal fusion outcomes assessment. J Spinal Disord Tech 15(1):16–23) and Bombardier et al. (Bombardier C (2000) Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. Spine 25(24):3100–3103)

outcome measures reflect the interests of other stakeholders, such as the surgeon, the patient's employer, social and health insurance organizations. For example, surgeons may be particularly interested in outcomes such as the stability of an implant, or the success of a fusion, or blood loss during surgery. On the other hand, social and health insurance organizations are interested in the cost benefit and cost-effectiveness of treatment. Importantly, there is a growing consensus in the surgical and research community that a single outcome measure is not sufficient. Investigators have called for using multiple outcome measures, such as a visual analogue scale in conjunction with the Oswestry and return to work status simultaneously [9]. Today, four domains of outcome are commonly recommended. They are: patient-oriented outcomes (for example, perceived pain, function and well-being), clinician-oriented outcomes (for example, fusion rate and blood loss in surgery), cost/health care utilization (for example, medical and disability costs) and societal outcomes (for example, return to and retention of work).

Second, are the outcome measures sensitive enough to detect important changes in the condition of the patient? There is a growing literature on establishing the 'Minimally Clinically Important Difference' (MCID), a criterion, which can be thought of as the difference in a specific measure of outcome that reflects a meaningful change in the health status of a patient. For example, for certain patient populations the MCID for the Oswestry Disability Index has been estimated by different observers between 4 and 17 points out of a 100 point scale [50]. What constitutes a MCID, however, is not necessarily fixed, and the MCID can (and should) vary depending on what perspective the investigator feels important [5].

Third, is the duration of follow-up sufficient? Although there is no fixed requirement for the duration of follow-up, investigators are calling for longer-term follow-up following surgical intervention to identify unexpected events related to the implant device. One study found an unexpected increase in the level of pain between 1 and 2 years following fusion [28]. The recently published United States Food and Drug Administration Investigational Device Exemptions (IDE) trials follow patients for 2 years after implant. As part of its recommendation of market approval of the SB Charité III artificial disc (June 2–3, 2004), the Food and Drug Administration's Orthopaedic And Rehabilitation Devices Advisory Panel asked that the manufacturer continue follow-up of the IDE trial's participants up to 5 years [2].

Step 5: identify the study design being used

Once the outcome and exposure measures are identified, the next question the reader needs to evaluate is: how are outcomes going to be compared between the various treatment groups? The study design embodies the procedures and analytic approach used to make the essential comparisons. We want to identify the study design for two reasons: (1) to anticipate what kind of statistical

analysis will be used to analyze the data, and (2) to anticipate areas where bias and confounding may have been inadvertently introduced that would limit the strength of the conclusions being drawn from the findings. There are a finite number of clinical study designs. The major observational and experimental designs are discussed below.

Cross-sectional

A cross-sectional study can be thought of as a survey, where patients having a similar condition are identified and then characterized (for example by demographic features, the type of treatment they received, or by health status or outcome) (Fig. 2). Cross-sectional studies can provide information about prevalence and also information about associations between risk factors and outcome. For example, a recent cross-sectional study investigated the relationship between smoking, global health status and depression among spine patients in the United States. Patients from the National Spine Network ($n = 25.455$) were characterized according to smoking status (smoker or non-smoker), global health status (SF-36) and depression (Zung depression scale). The prevalence of smoking was 16% among the spine patients. The study found that smokers were more likely to report symptoms of depression than non-smokers (54 vs. 37% respectively), were more likely to report severe back symptoms (50 vs. 37% respectively) and scored lower on measures of overall physical and mental health [53].

Because the risk factor and the outcome are measured at the same time, cross-sectional studies cannot provide evidence that the exposure is causally related to the outcome. For example, data from Vogt et al. [53] suggest that smoking is related to depression among spine patients. But because the study is cross-sectional, it is not possible to tell whether individuals in the study began smoking and then became depressed, or whether they started to smoke because they felt depressed.

Case control

A study that can offer evidence for causality is the case control design (Fig. 3). The study design is retrospective in nature. It first identifies all those individuals with the outcome of interest (for example, all those exhibiting poor function documented by a high Oswestry score) called cases. It then collects a comparison group of individuals, called controls, who are similar in many respects to the cases, but who do not have the outcome of interest (for example, all those exhibiting good function documented by a low Oswestry score). Within each group a ratio, called an odds, is formed by comparing the number of cases exposed to a risk factor (for example, spinal surgery) to the number of cases not exposed to the risk factor. A similar odds is calculated for the controls. Estimate of association is calculated by forming the odds ratio, which is the ratio of the odds of exposure among the cases compared to the odds of exposure among the controls. When the odds ratio is at or around the value 1, it means the odds of exposure to the risk factor is about the same among the cases as it is among the controls. When the odds ratio deviates well away from the value of 1, it means that the odds of exposure in the cases is different from the odds of exposure among the controls. This is interpreted to mean that there is a statistical association between having the disorder and the likelihood of having been exposed to the risk factor.

Selecting appropriate controls to serve as a comparison group is a fundamental methodological issue in case control studies. Controls are those individuals that have had a chance of experiencing the exposure risk factor of interest (for example, surgery or conservative care), but are free of the outcome of interest (meaning they express no dysfunction). Controls meeting this definition can be drawn from any population; however, for reasons of generalizability it is desirable that the controls consist of a representative sample of the population from which they a drawn [45]. Matching is sometimes used to limit the influence of covariates on the findings of the study.

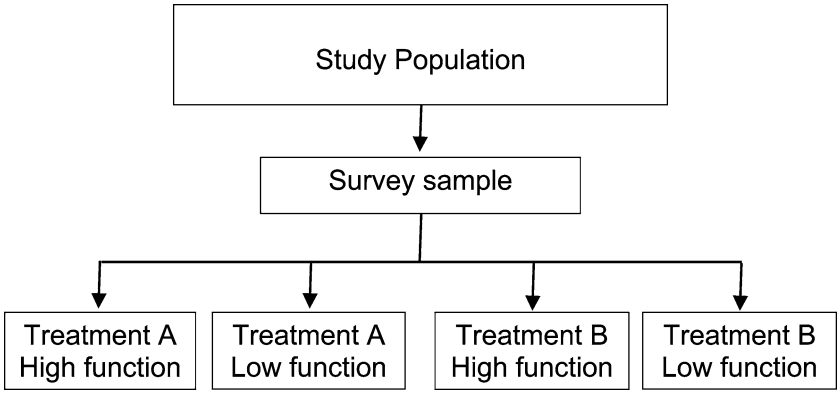

Fig. 2 Cross-sectional study design

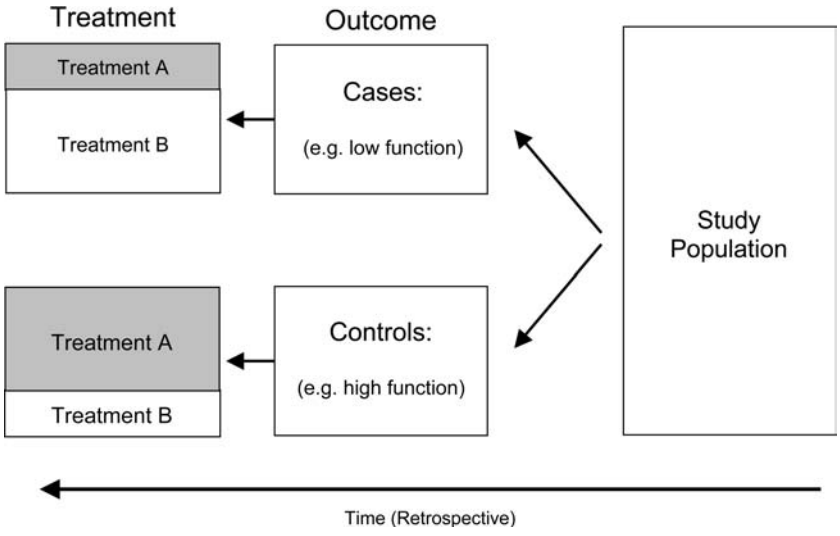

Fig. 3 Case control study design

Matching means pairing a case with a control on a certain characteristic, for example, age, gender or smoking. By doing so, the influence of these factors on the findings of the study are nullified. This makes it easier to detect statistically significant association between outcome and the exposure of interest. Care should be taken, however, not to over-match. Some authors recommend that matching not be done on more than three covariates [45]. Moreover, matching on the exposure variable of interest (for example, the kind of treatment a patient receives) makes it impossible in a case control study to assess the relationship between that treatment and outcome. Finally, a special statistical analysis (called a matched odds ratio) is required when analyzing data from a matched case control study [45].

Comparative and cohort study designs

A frequently used study design in the spine literature is a comparative study of outcome of patients drawn from a clinical practice or hospital. The design of this kind of study involves identifying a group of patients undergoing treatment for a specific disease. These patients are then classified according to the type of treatment they received. Outcome is then compared between the groups (Fig. 2).

Because of the possibility of uncontrolled confounding and lack of generalizability, simple comparative studies (case series) can provide only weak or suggestive evidence of treatment efficacy. However, because they are relatively inexpensive and easy to conduct, comparative studies are often used early in the development of new surgical procedures or devices. For example, early evaluative studies of artificial discs [13, 17, 19, 23, 32, 47, 56] identified clinical populations of individuals with chronic, discogenic low back pain eligible for spinal fusion. These early observational studies found that reduction in pain and improvement in function among those getting the artificial disc was no different from those getting spine fusion, but mobility of the spinal segment was largely maintained.

A form of a comparative study is the cohort study design (Fig. 4). The cohort study design is distinguished from the comparative study in that the number of individuals followed is larger and the catchment area is well described and representative. In epidemiologic studies, a cohort consists of an entire community of individuals sharing some common characteristic, for example living in the same geographic region or all sharing the same year of birth. In surgical studies, a cohort can be identified as all those individuals getting a particular experimental intervention for a specific condition (such as an artificial disc), or all those individuals who belong to a well defined community (such as members of a health maintenance organization, a geographical area or occupational group).

The benefit of following an entire (or statistically representative) population of individuals is that a probability value, called risk, can be calculated for the cohort. Risk is defined as the proportion of individuals experiencing improved outcome among the entire cohort within a specific period of time. To evaluate treatment efficacy, relative risks are calculated by comparing the risk of improved outcome among those in the cohort obtaining one kind of treatment and comparing that to the risk of improved outcome among those getting an alternative treatment. A relative risk of 1 means the outcome experience is the same between groups getting different treatments. Relative risks diverging well away from 1 means the outcome experience between the comparison groups is different. Because an entire population is followed, the findings from a cohort study are

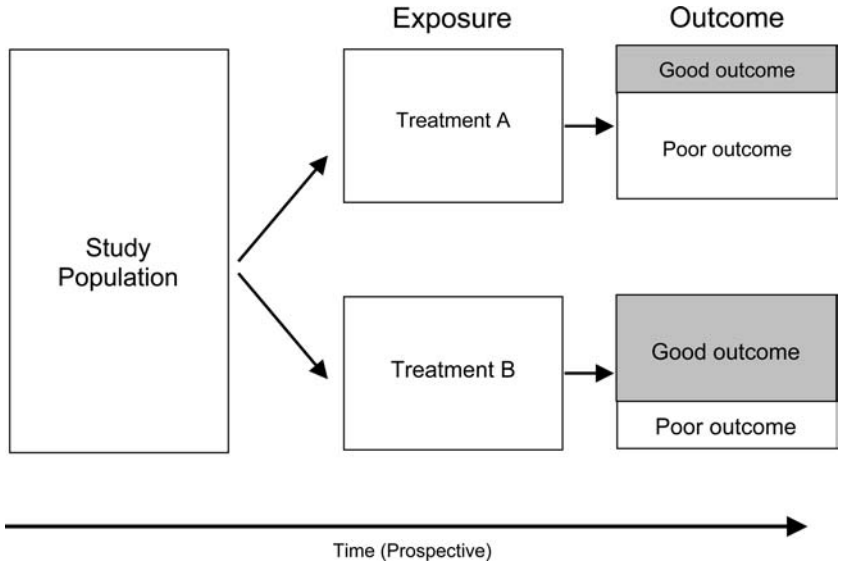

Fig. 4 Cohort study design

far more generalizable than that of a simple comparative design (case series).

Randomized controlled trial

In an idealized experiment, identical samples are selected and exposed to various treatments. The logic behind this experiment is that, when all other factors that could be related to outcome are identical except for the type of treatment, then any difference in outcome is uniquely and solely attributable to the type of treatment used. Of all of the basic clinical research study designs, the randomized controlled trial comes the closest to this experimental ideal.

The key features of a randomized controlled trial are the following: (1) recruitment of potential subjects and screening for eligibility, (2) voluntary enrollment by means of informed consent, (3) screening to exclude those candidates for whom study participation would be contra-indicated, (4) experimental allocation by means of a random process, (5) pre-intervention assessment, (6) intervention, (7) post-intervention assessment, and (8) comparison of outcome between the study arms. To assess treatment efficacy, outcome among those in the intervention arm is compared with outcome among those in the control arm at the end of the study (Fig. 5).

The statistical analysis in a randomized clinical trial computes rates of outcome between the treatment groups. When the outcome is continuous in nature (for example, the Oswestry score, which is a index scale from 0 to 100) difference in outcome can be assessed using analysis of variance statistical method. When the outcome is categorical in nature (for example, the proportion of individuals having 'excellent' outcome following surgery) can be compared by calculating rate ratios. The interpretation of a rate ratio, as used in a clinical trial, is the same as that for a relative risk. A rate ratio of 1 means no difference in outcome between treatment groups; rate ratios that diverge well away from 1 mean the outcome experience between the treatment groups is different.

Step 6: evaluate the methods used to control for bias and confounding

Confounding is a factor that is associated with both with treatment and outcome [45]. Confounding can mask the observed statistical association between treatment and outcome, either inflating the estimate or diminishing it. For example, in a study comparing opioid patient controlled analgesia with non-steroidal analgesia (NSAIDs) following lumbar fusion surgery, Park et al. found that individuals infused with a combination of ketorolac and fentanyl were more likely to have delayed fusion than those using fentanyl alone, and concluded that NSAIDs were responsible for the delayed recovery [43]. However, in Park et al. study, individuals using NSAIDs were far more likely to smoke cigarettes than those taking opioids. Smoking is thought to interfere with osteogenesis. Since smoking in the study was associated both with the treatment (opioid versus NSAID) and also outcome (rate of fusion), smoking confounded the observed association between analgesia and fusion.

Bias is the systematic misrepresentation of rates or associations observed in a study [45]. Bias can be introduced in the way subjects are recruited, data are collected, or by using an incorrect statistical analysis procedure for the study design used [45]. The common methodological tools used to control for bias and

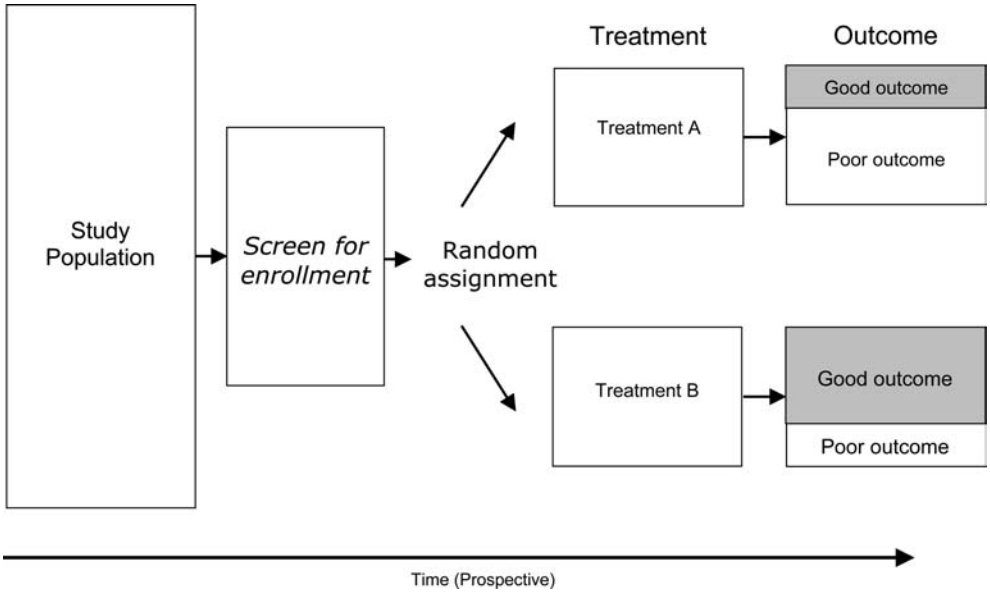

Fig. 5 Randomized controlled trial

confounding include: blinding, control of cross-over, randomization, control of differential follow-up, and correct statistical analysis.

Blinding

Blinding refers to whether or not the allocation status of a study participant is known or not [38]. Blinding takes place at several levels. For example, a research subject is 'blinded' when that person does not know which treatment (the experimental or control) they are receiving. An independent evaluator, used to obtain measures of outcome during follow-up, is 'blinded' if that person does not know what kind of treatment the subject has received. Finally, the treating physician/clinician is 'blinded' if that person does not know the kind of intervention a particular research subject is allocated to receive. The terms 'single', 'double' and 'triple' blinding have been used to describe concealment of treatment allocation in a study. However, a recent survey found that these terms are not interpreted consistently among readers, for example some individuals thought 'double blinded' meant that the research subject and the outcomes evaluator are blinded to treatment status, whereas others interpreted 'double blinded' as meaning the subject and the treating clinician were blinded to treatment status [46]. Consequently, the terms 'single', 'double' or 'triple' blinded in themselves have little meaning; a good quality study will fully and completely describe the means of concealing treatment allocation [46].

Blinding serves two purposes. First, it helps ensure that the nature of the treatments remain distinct and separate between the intervention and control arms of the study [38]. For example, if the subject knows the study arm to which they belong, there can be a tendency to seek out the treatment being administered in the opposing arm. When this happens, the distinction between the treatment and comparison arm blurs, reducing the internal validity of the study. This phenomena is called 'cross-over', which will be discussed shortly.

The second purpose of blinding is to limit prevarication bias and interviewer bias. Prevarication bias occurs when a study subject over- or under-estimates outcome because of knowledge of the kind of treatment they had received. There can be any number of reasons, for example, a subject may subconsciously feel the need to provide answers pleasing to the interviewer or study coordinator, or the subject may feel that there is some secondary gain from consciously mis-reporting their clinical status. Interviewer bias occurs when a clinical outcomes evaluator subconsciously or deliberately influences responses from a subject because of an awareness of which treatment the subject has received. A study can be particularly susceptible to interviewer bias when the same person who provides treatment in a study is also the same person who evaluates outcome at the end of the study. Because clinical personnel have an interest in outcome of a patient, it is likely that person will not be able to fairly evaluate the outcome of the subject. For this reason it is important that outcomes be evaluated by an independent person who has no knowledge of the treatment status of the subject.

Blinding may be very difficult to maintain in studies comparing surgical and non-surgical treatments. For example, it is difficult to conceal a scar from a patient, and some trials, as a policy, reveal the type of implant to the subject [8, 20]. However, the effect on study findings

from lack of blinding is not trivial. One study found that in studies not using full blinding, estimates of treatment effect were exaggerated by about 19% (32) [37]. Yet, few studies report the degree to which blinding was maintained during the course of the study [31].

How can we know whether blinding was maintained? One technique used is to ask the study participant, at the end of the study, which treatment they thought they had received. The response is graded on scale ranging from 'treatment', 'control' and 'do not know'. A statistical test compares what the subject guessed they had received against their actual treatment allocation. If blinding was maintained, the test would show a random distribution among those guessing correctly and those guessing incorrectly, that is, about 50% guessing correctly. If blinding was lost, then the proportion guessing correctly would be expected to be different from 50% [4].

Good outcomes studies will report on the kind of blinding that was employed. At a minimum, an independent outcomes evaluator, who does not know what kind of treatment the patient has received, should be used. The clinician who provides treatment must never also assess outcome. Whenever reasonably possible to do, the patient themselves should be blinded to the treatment that they have received. Finally, a good outcomes study will report in the Discussion section of the paper how blinding (or lack thereof) could have potentially affected the findings of the study.

Control of cross-over

If the subject knows the study arm to which they belong, there can be a tendency to seek out the treatment being administered in the opposing arm. When this happens, the distinction between the treatment and comparison arm blurs, reducing the internal validity of the study. This phenomena is called 'cross-over'.

Cross-over is a modest methodological issue in studies comparing different kinds of implant technologies, for example when comparing instrumented and non-instrumented fusion. This is because the rate of salvage or revision of spinal implants is low, so the opportunity to exchange one kind of spinal implant with another is limited. However, cross-over is a significant methodological issue when comparing surgical with non-surgical management. A good quality trial will report cross-over, study withdrawals and cases with missing data. Useful information can be gleaned from reasons why subjects withdrew from a study or otherwise were unable to complete the study protocols.

Randomization

Randomization is an allocation procedure that theoretically produces study arms where the overall group characteristics are identical in all respects. This property is guaranteed by the Central Limit Theorem, which states in part that a random selection process will always produce a sample whose average estimates the population mean from which the sample is drawn [10]. Where several groups are formed by random allocation from a large pool of eligible subjects, the Central Limit Theorem would lead us to expect that the groups would be representative of the total pool of eligible subjects, and by extension, similar to each other. When comparison arms are equivalent in all respects except for the type of treatment received, then the only conclusion that can be drawn is that differences in outcome are attributable solely to the type of treatment given. This logic is the underpinning of why properly randomized controlled trials are thought of as providing the strongest evidence of treatment efficacy [10].

There are natural barriers to using randomization in spine surgical trials [29]. One of the most important barriers is the natural disinclination by the surgeon to permit treatment to be left up to chance. Leaving treatment decision up to chance can be thought of as unethical, especially where the surgeon or the patient has a particularly strong belief about the benefit of a particular treatment option. However, recent studies have shown that significant bias can be introduced in randomized controlled trials where the allocation schedule is not concealed from the patients or clinicians [46].

Consequently, a good randomized controlled trial will report: the method of randomization used, and demonstrate the equivalence of the control and intervention groups by comparing selected, baseline characteristics.

Control of differential drop out during follow-up

It has been shown that measures of functional outcome can be affected by differences in post-surgical management following lumbar fusion [42]. To avoid biasing comparison between study arms, randomized controlled trials should employ uniform protocols for post-surgical rehabilitation and management [14]. In addition, problems with patient compliance can confound long-term findings, even when rehabilitation protocols are carefully planned and implemented. Research into patient compliance is evolving, but current thinking is that deviations from prescribed treatment reflect a process whereby a patient tries to accommodate the requirements of treatment into the context of their own, competing, life and situational demands [22]. The protocols for post-operative management need to be clearly described in any study of surgical intervention. For example does the post-operative protocol of care include certain medication and physical therapy.

Studies can loose stochastic equivalence between study arms when individuals in the study drop-out from follow-up. Drop out can occur for a host of reasons, for example, study participants may die, or move, or they may decide to withdraw their voluntary consent. Cases with missing or incomplete data can be considered a form of drop-out. Study investigators may choose to exclude participants from further follow-up if those individuals require salvage, revision or replacement surgery, or if medical conditions develop that preclude them from continued participation in the study. In the same way that a good trial reports a flowchart of individuals recruiting and enrolling into the study, a good trial will report the number, characteristics and study arm membership of those individuals dropped or lost to follow-up.

Correct statistical analysis

Each study design has its own associated statistical analysis procedure. The case control design expresses associations in terms of the odds ratio; the cohort design uses the relative risk; the clinical trial expresses associations in terms of rate ratios (for categorical outcomes) or differences in mean values (where the outcome in continuously scaled). The reader needs to review the statistical analysis and ensure that, for the findings to be statistically sound, that the statistical analysis methodology matches the study design. Important bias can result where the analysis methodology does not reflect the study design. For example, the use of a unmatched analysis technique in a case control study where matching is used can strongly bias the findings and invalidate the conclusions drawn [45].

It is beyond the scope of this article to discuss the specific techniques for conducting statistical analyses appropriate for each study design. Numerous texts are available for cross-sectional, [26] cohort, [11] case control [45] and randomized clinical trial study designs, [27, 38] and the reader is referred to these texts for more information.

Step 7: determine if the study was adequately powered

There is a need to have sufficient numbers of subjects enrolled in study to avoid drawing erroneous conclusions from the statistical findings. The first error is when the statistical analysis shows that there is a difference in outcome between study arms when in fact differences in treatment efficacy are negligible. This is referred to as a 'Type I' statistical error. The second statistical error is referred to as a 'Type II' error and occurs when the statistical analysis fails to demonstrate that a difference in treatment efficacy exists when in fact the treatments do produce decidedly different outcomes but the number of subjects were too small. How many subjects need to be enrolled to avoid these two different inferential errors?

A study is sufficiently powered if the differences in outcome the study investigator wishes to demonstrate is larger than the study's standard error [55]. The standard error represents the range of possible 'best guesses' of the average outcome score among those getting a particular treatment. A ratio, called the effect size, consists of the difference in outcome the investigator wishes to demonstrate (numerator) compared with the study's standard error (denominator).

The study's standard error is inversely proportional to the number of subjects enrolled. When the number of subjects enrolled is high, the standard error is low. When the number of study subjects is low, then the standard error is high [55].

Studies have adequate statistical power when the outcome differences are much larger than the standard error. There are two ways to ensure a study has adequate statistical power: define outcome differences that are very large, or enroll large numbers of study subjects [55].

Usually a balance is drawn between the two. The study investigator justifies how big a difference in treatment outcome should be tested; a good outcomes study will rationalize this choice by referring to the literature on the minimally clinically significant difference (described earlier). Then the investigator will identify, by means of a sample size calculation, the number of research subjects required to demonstrate this minimally clinically significant difference as statistically significant.

Consequently, to be of best use, clearly indicate the reasons for sizing a study to detect a difference in outcome between treatment arms of a given magnitude and report the results of the sample size calculation that was done in preparation for the study. We believe the best way to determine a correct sample size is to conduct a pilot study on selected outcomes.

Synthesizing information from multiple studies

Conclusions regarding treatment efficacy come from synthesizing available evidence identified from systematic literature searches. There are multiple approaches to synthesizing evidence, one of which can be labeled as a 'systematic literature review', and another that can be termed a 'best evidence synthesis'. The systematic review is characterized by a thorough literature review which grades the methodology of each of the studies included in the review. Usually some sort of grading criteria for each methodological element is developed for the review effort. These grading criteria are agreed upon by the individuals reading and rating the studies before the systematic review takes place. For example, a typical

way of rating 'Blinding' would be: 'Single blinded', 'Double-blinded', 'No blinding', 'Not described' [16]. Some systematic reviews conduct analyses of inter- and intra-rater consistency in assigning grades, although this practice, while desirable, is still uncommon [36]. Scores from each of the methodological elements are added together to form and overall quality score for the study. The quality score represents, in essence, how confident the reader can be that important methodological elements that limit bias and confounding are addressed by the design of the study. Examples from the low back pain literature include (but not limited to) the various reviews from the Cochrane Back Pain group (such as the review of surgery for degenerative lumbar spondylosis, [31] multidisciplinary biopsychosocial rehabilitation for chronic back pain, [33] and surgery for lumbar disc prolapse [31] to name few), and evidence reviews of common management techniques for low back pain [54].

The best-evidence synthesis, on the other hand, can be thought of as a generalization of the systematic literature review. Whereas a systematic literature review grades studies with regard to methodological rigor, the best evidence synthesis combines evidence to summarize what is currently known regarding risk factors, diagnosis, treatment and prognosis of a medical condition. Some examples of best evidence synthesis for spine pain include the Quebec Task Force on Spinal Disorders, [48] the Quebec Task Force on Whiplash-Associated Disorders, [49] the Paris Task Force on Back Pain, [3] and the current ongoing Neck Pain Task Force [1].

Conclusion

This paper reviews some of the basic methodological considerations when reviewing evidence from outcomes studies, with particular reference to studies of spine surgery. A suggested checklist for drawing attention to key methodological issues is presented. The process of critical evaluation of outcome studies does not necessarily resolve controversy. For example, systematic review, after nine randomized clinical trials, still has not provided definitive evidence whether chronic, non-spe cific low back pain is best treated with surgical or non-surgical approaches, or which patient best benefit from either approach [30]. Where one study has important methodological flaws that limit the strength of the evidence, other later studies can address these methodological limitations to test hypotheses. The process of coming to consensus can be slow and difficult. However, for the clinician, the immediate benefit of critical evaluation of the literature is to provide to patients the strongest available evidence for making treatment choices, but also to inform the patient in those situations where weak (or no) evidence exists of efficacy of an existing or new treatment.

References

1. The Bone and Joint Decade 2000–2010 Task Force on Neck Pain and its Associated Disorders. Available at: http://www.nptf.ualberta.ca/. Accessed November 1, 2005
2. Summary minutes. Orthopaedic and Rehabilitation Devices Advisory Panel, United States Food and Drug Administration. Yaszemski MJ, Chairperson. Gaithersburg, MD, June 2–3 2004
3. Abenhaim L, Rossignol M, Valat JP, Nordin M, Avouac B, Blotman F, Charlot J, Dreiser RL, Legrand E, Rozenberg S, Vautravers P (2000) The role of activity in the therapeutic management of back pain. Report of the International Paris Task Force on Back Pain. Spine 25(4 Suppl):1S–33S
4. Bang H, Ni L, Davis CE (2004) Assessment of blinding in clinical trials. Control Clin Trials 25(2):143–156
5. Beaton DE, Boers M, Wells GA (2002) Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol 14(2):109–114
6. Bereczki D, Gesztelyi G (2000) A Hungarian example for hand searching specialized national healthcare journals of small countries for controlled trials: is it worth the trouble?. Health Libr Rev 17(3):144–147
7. Bigos SJ, Bowyer O, Braen G (1994) Acute low back problems in adults, clinical practice guideline, No. 14. vol AHCPR Pub 95-0642. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, Rockville, MD
8. Blumenthal S, McAfee PC, Guyer RD, Hochschuler SH, Geisler FH, Holt RT, Garcia R Jr, Regan JJ, Ohnmeiss DD (2005) A prospective, randomized, multicenter Food and Drug Administration Investigational Device Exemptions study of lumbar total disc replacement with the Charité artificial disc versus lumbar fusion: Part I: evaluation of clinical outcomes. Spine 30(14):1565–1575; discussion E1387–1591
9. Bombardier C (2000) Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. Spine 25(24):3100–3103
10. Box GEP, Hunter WG, Hunter JS (1978) Statistics for experimenters. Wiley, New York, NY
11. Breslow NE, Day NE (1987) Statistical methods in cancer research: the design and analysis of cohort studies, vol 2. International agency for research on cancer, Lyon, France
12. Brox JI, Sorensen R, Friis A, Nygaard O, Indahl A, Keller A, Ingebrigtsen T, Eriksen HR, Holm I, Koller AK, Riise R, Reikeras O (2003) Randomized clinical trial of lumbar instrumented fusion and cognitive intervention and exercises in patients with chronic low back pain and disc degeneration. Spine 28(17):1913–1921
13. Buttner-Janz K, Schellnack K, Zippel H (1989) Biomechanics of the SB Charité lumbar intervertebral disc endoprosthesis. Int Orthop 13(3):173–176

14. Carey TS (1999) Randomized controlled trials in surgery: an essential component of scientific progress. Spine 24(23):2553–2555
15. Carragee EJ (2005) Clinical practice. Persistent low back pain. N Engl J Med 352(18):1891–1898
16. Chalmers CC, Smith H, Blackburn B, Silverman B, Schroeder B, Reiter D, Ambroz A (1981) A method for assessing the quality of randomized clinical trial. Control Clin Trials 2:31–49
17. Cinotti G, David T, Postacchini F (1996) Results of disc prosthesis after a minimum follow-up period of 2 years. Spine 21(8):995–1000
18. Coyle YM (2000) Developing theoretical constructs for outcomes research. Am J Med Sci 319(4):245–249
19. David T (1993) Lumbar disc prosthesis. Eur Spine J 1:254–259
20. Delamarter RB, Bae HW, Pradhan BB (2005) Clinical results of ProDisc-II lumbar total disc replacement: report from the United States clinical trial. Orthop Clin North Am 36(3):301–313
21. Deyo RA, Nachemson A, Mirza SK (2004) Spinal-fusion surgery—the case for restraint. N Engl J Med 350(7):722–726
22. Donovan JL, Blake DR (1992) Patient non-compliance: deviance or reasoned decision-making? Soc Sci Med 34(5):507–513
23. Enker P, Steffee A, McMillin C, Keppler L, Biscup R, Miller S (1993) Artificial disc replacement. Preliminary report with a 3-year minimum follow-up. Spine 18(8):1061–1070
24. Errico TJ, Gatchel RJ, Schofferman J, Benzel EC, Faciszewski T, Eskay-Auerbach M, Wang JC (2004) A fair and balanced view of spine fusion surgery. Spine J 4(5 Suppl):S129–138
25. Fairbank JC, Pynsent PB (2000) The oswestry disability index. Spine 25(22):2940–2953
26. Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. Wiley, New York
27. Fleiss JL (1986) The design and analysis of clinical experiments. Wiley, New York
28. Fritzell P, Hagg O, Wessberg P, Nordwall A (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 26(23):2521–2532
29. Fu WK (2000) Re: randomized controlled trials in surgery: an essential component of scientific progress (Spine 1999; 24:2553–2555). Spine 25(15):2002–2004
30. Gibson JN, Grant IC, Waddell G (1999) The cochrane review of surgery for lumbar disc prolapse and degenerative lumbar spondylosis. Spine 24(17):1820–1832
31. Gibson JN, Waddell G (2005) Surgery for degenerative lumbar spondylosis. Cochrane Database Syst Rev (2):CD001352
32. Griffith SL, Shelokov AP, Buttner-Janz K, LeMaire JP, Zeegers WS (1994) A multicenter retrospective study of the clinical results of the link SB Charité intervertebral prosthesis. The initial European experience. Spine 19(16):1842–1849
33. Guzman J, Esmail R, Karjalainen K, Malmivaara A, Irvin E, Bombardier C (2001) Multidisciplinary rehabilitation for chronic low back pain: systematic review. BMJ 322(7301):1511–1516
34. Haynes RB, McKibbon KA, Fitzgerald D, Guyatt GH, Walker CJ, Sackett DL (1986) How to keep up with the medical literature: IV. Using the literature to solve clinical problems. Ann Intern Med 105(4):636–640
35. Hildebrandt J, Ursin H, Mannion AF, Airaksinen O, Brox JI, Cedraschi C, Klaber-Moffett J, Kovacs F, Reis S, Staal B, Zanoli G, Broos L, Jensen I, Krismer M, Leboeuf-Yde C, Niebling W, Vlaeyen JW (2005) European guidelines for the management of chronic non-specific low back pain. European Co-Operation in the field of Scientific and Technical Research (COST). Available at: http://www.backpaineurope.org/web/files/WG2_Guidelines.pdf. Accessed August 25, 2005
36. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ (1996) Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 17(1):1–12
37. Juni P, Altman DG, Egger M (2001) Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 323(7303):42–46
38. Meinert CL, Tonascia S (1986) Clinical trials: design, conduct, and analysis, vol 8. Oxford University Press, New York, NY
39. Miller FG, Brody H (2003) A critique of clinical equipoise. Therapeutic misconception in the ethics of clinical trials. Hastings Cent Rep 33(3):19–28
40. Moher D, Schulz KF, Altman DG (2003) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Clin Oral Investig 7(1):2–7
41. Nachemson AL, Jonsson E (eds) (2000) Neck and back pain: the scientific evidence of causes, diagnosis, and treatment. Williams & Wilkins, Lippincott
42. Ostelo RW, de Vet HC, Waddell G, Kerckhoffs MR, Leffers P, van Tulder M (2003) Rehabilitation following first-time lumbar disc surgery: a systematic review within the framework of the cochrane collaboration. Spine 28(3):209–218
43. Park SY, Moon SH, Park MS, Oh KS, Lee HM (2005) The effects of ketorolac injected via patient controlled analgesia postoperatively on spinal fusion. Yonsei Med J 46(2):245–251
44. Roland M, Fairbank J (2000) The Roland-Morris disability questionnaire and the oswestry disability questionnaire. Spine 25(24):3115–3124
45. Schlesselman JJ (1982) Case control studies: design, conduct, analysis, 1st edn. Oxford University Press, New York, NY
46. Schulz K (2001) Assessing allocation concealment and blinding in randomized controlled trials: why bother? Evid Based Nurs 4(1):4
47. Sott A, Harrison D (2000) Increasing age does not affect good outcome after lumbar disc replacement. Int Orthop 24:50–53
48. Spitzer WO, LeBlanc FE, DuPuis M (1987) Scientific approach to the assessment and management of activity-related spinal disorders. A monograph for clinicians. Report of the Quebec Task Force on Spinal Disorders. Spine 12(7 Suppl):S1–59
49. Spitzer WO, Skovron ML, Salmi LR, Cassidy JD, Duranceau J, Suissa S, Zeiss E (1995) Scientific monograph of the Quebec Task Force on whiplash-associated disorders: redefining ''whiplash'' and its management. Spine 20(8 Suppl):1S–73S
50. Taylor SJ, Taylor AE, Foy MA, Fogg AJ (1999) Responsiveness of common outcome measures for patients with low back pain. Spine 24(17):1805–1812
51. van Tulder M, Furlan A, Bombardier C, Bouter L (2003) Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine 28(12):1290–1299
52. van Tulder MW, Becker A, Bekkering T, Breen A, Gil del Real MT, Hutchinson A, Koes BW, Laerum E, Malmivaara A, Nachemson AL, Niehus W, Roux E, Rozenberg S (2005) European guidelines for the management of acute nonspecific low back pain in primary care. European Co-Operation in the field of Scientific and Technical Research (COST). Available at: http://www.backpaineurope.org/web/files/WG1_Guidelines.pdf. Accessed August 25, 2005

53. Vogt MT, Hanscom B, Lauerman WC, Kang JD (2002) Influence of smoking on the health status of spinal patients: the national spine network database. Spine 27(3):313–319
54. Waddell G, McIntosh A, Hutchinson A, Feder G, Lewis M (1999) Low back pain evidence review. Royal College of General Practitioners, London
55. Walters S, Campbell M, Paisley S (2001) Methods for determining sample sizes for studies involving health-related quality of life measures: a tutorial. Health Serv Outcomes Res Methodol 2:83–99
56. Zeegers WS, Bohnen LM, Laaper M, Verhaegen MJ (1999) Artificial disc replacement with the modular type SB Charité III: 2-year results in 50 prospectively studied patients. Eur Spine J 8(3):210–217