Fundamentals of Clinical Outcomes Assessment for Spinal Disorders: Study Designs, Methodologies, and Analyses

Patrick Vavken[1,2,3] Anne Kathleen B. Ganal-Antonio[4] Francis H. Shen[5] Jens R. Chapman[6] Dino Samartzis[7,8]

[1] Department of Orthopaedic Surgery, Boston Children's Hospital, Harvard University Medical School, Boston, Massachusetts, United States
[2] Harvard Center for Population and Development Studies, Harvard School of Public Health, Boston, Massachusetts, United States
[3] Department of Orthopedic Surgery, University of Basel, Basel, Switzerland
[4] Department of Orthopedics, Makati Medical Center, Makati City, Philippines
[5] Department of Orthopaedic Surgery, University of Virginia, Charlottesville, Virginia, United States
[6] Swedish Neuroscience Institute, Swedish Medical Center, Seattle, Washington, United States
[7] Department of Orthopaedics and Traumatology, The University of Hong Kong, Pokfulam, Hong Kong, SAR, China
[8] The Laboratory and Clinical Research Institute for Pain, The University of Hong Kong, Pokfulam, Hong Kong, SAR, China

Address for correspondence Dino Samartzis, DSc, Department of Orthopaedics and Traumatology, The University of Hong Kong, Professorial Block, 5th Floor, 102 Pokfulam Road, Pokfulam, Hong Kong, SAR, China (e-mail: dsamartzis@msn.com).

Global Spine J 2015;5:156–164.

Abstract

Study Design A broad narrative review.

Objective Management of spinal disorders is continuously evolving, with new technologies being constantly developed. Regardless, assessment of patient outcomes is key in understanding the safety and efficacy of various therapeutic interventions. As such, evidence-based spine care is an essential component to the armamentarium of the spine specialist in an effort to critically analyze the reported literature and execute studies in an effort to improve patient care and change clinical practice. The following article, part one of a two-part series, is meant to bring attention to the pros and cons of various study designs, their methodological issues, as well as statistical considerations.

Methods An extensive review of the peer-reviewed literature was performed, irrespective of language of publication, addressing study designs and their methodologies as well as statistical concepts.

Results Numerous articles and concepts addressing study designs and their methodological considerations as well as statistical analytical concepts have been reported. Their applications in the context of spine-related conditions and disorders were noted.

Conclusion Understanding the fundamental principles of study designs and their methodological considerations as well as statistical analyses can further advance and improve future spine-related research.

Keywords
- spine
- outcomes
- statistics
- analysis
- study design
- personalized

received
February 7, 2014
accepted after revision
February 9, 2015
published online
March 12, 2015

DOI http://dx.doi.org/10.1055/s-0035-1547525.
ISSN 2192-5682.

Introduction

Outcome assessment is one of the most important features of evidence-based medicine, which has been defined as "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" and therefore depends on accurate description of outcomes.[1] Due to the enormous increases in health care expenditure, numerous health policy makers and regulators have introduced evidence-based medicine and the proof of cost-effectiveness into both the clinical and reimbursement guidelines.

Outcome assessment can be performed for several reasons: to trace the progress of an individual patient, to study the efficacy of a treatment method, to compare the effectiveness of different treatments, or to evaluate the cost-effectiveness of different treatments. In technical terms, outcome assessment aims at establishing four parameters: efficacy (can it work?), effectiveness (does it work?), efficiency (does it produce value?), and safety (do adverse events exist and are they acceptable?). Due to the plethora of innovative spine technologies and the changing face of the management of spinal disorders, understanding outcome assessment is critical to judge the efficacy and safety of a treatment and, in so doing, further advance the field. As such, the following article, part one of a two-part series, describes the rationale of outcome assessment followed by a fundamental description of study designs, methodologies, and analyses while focusing on spinal disorders.

Study Designs

Defining End Points

The first and most important step in any outcome assessment is to clearly define the outcome(s) of interest and the scale of measurement. Preferably, this definition will also include an influential variable or a risk factor that allows formulation of a study question. Unfortunately, more often than not, studies are initiated to research issues such as "We want to look at the clinical advantage of implant X over implant Y"—and, more often than not, such studies fail because they inadequately define what to measure and how to measure it. Instead, a good start might be to ask, "Are there differences in the 5-year rates of radiologic signs of adjacent segment degeneration of implant X versus implant Y in patient population Z?" Such a clearly formulated question will be most helpful in choosing the right type of study design, the appropriate statistical analysis, and the interpretation of the findings in relation to the study question and will even facilitate writing the manuscript.

Study Design Types

Two main study types exist, observational and experimental/interventional, depending on whether the investigator merely observes events unfold or whether he or she intervenes to assess the effectiveness of a treatment in a controlled setting comparing groups that were established at baseline in a certain manner (e.g., randomized controlled trial [RCT]). Further characterization depends on several factors, such as the chronology of outcome occurrence and assessment (prospective/retrospective), the number of interventions/groups (controlled studies/longitudinal cohorts and case series), or patient allocation (randomized/matched/nonrandomized). When choosing from these options, the two most important questions to answer are "Is this study design able to show what I want to see?" or in scientific terminology "Is this design valid?" and—given the constraints in funding and time available, equally important—"Is this design cost-effective to show what I want to see?" It has been recognized that the main antagonist of validity is bias, which is defined as a "systematic deviation from the truth," usually because of a flaw in study design. The probability of biased results in any given study depends on the rigor of its design. The categorization into various "levels of evidence" depends on factors such as study design and objective (e.g., diagnostic, prognostic, therapeutic, among others), with level I studies being regarded as those with the highest degree of evidence and those less likely to be affected by bias (►**Table 1**).

Creation of high-level evidence study designs is desirable for several reasons: they produce the "best evidence" desired for evidence-based medicine, they hold the prospect of the highest return on investment, and they are more likely to get published.[2,3] However, as pointed out previously, cost-effectiveness is another important aspect. A phase III study (which is the postapproval, postmarketing study of Food and Drug Administration–approved drugs and interventions) does not necessarily need to be randomized or controlled to advance insight. For example, it does not matter if a specific drug or device is more or less effective than Vioxx (rofecoxib) or a metal bearing hip like the SR hip, because in both cases the respective manufacturers have withdrawn the drug or recalled the device in question.

In summary, choosing a valid and cost-effective study design will help in obtaining funding for a study, conducting the project, and publishing the findings. Although case reports, case series, and cross-sectional and observational studies are the most common studies published in the spine literature, there has been a steady rise in RCTs and systematic reviews/meta-analyses throughout the past decade.[4] Next, the most common study designs are presented in reverse level-of-evidence hierarchy, and their advantages and disadvantages are further illustrated in ►**Table 2**.

Cross-Sectional Studies, Case Series, and Case Reports—Level IV

Cross-sectional studies, case series, and case reports are observational studies. A cross-sectional study can be considered a survey that is used to gather information about a population at a single point in time. With its propensity to determine prevalence, risk factors, and outcomes, a cross-sectional study can attempt to draw a relationship between factors of interest. An inherent limitation lies in the foundational circumstance that factors of interest and outcomes are both measured at the same time. Subsequently, the interpretation of the results from cross-sectional studies are complicated by not understanding the causal pathway of events, meaning that one cannot be sure which came first, the "risk factor" or the "outcome." Case series and case reports are often used to describe or report unusual

Table 1 Study types and levels of evidence for primary research questions

Level	Types of studies			
	Therapeutic studies: investigating the results of treatment	**Prognostic studies: investigating the effect of a patient characteristic on the outcome of disease**	**Diagnostic studies: investigating a diagnostic test**	**Economic and decision analyses: developing an economic or decision model**
I	• High-quality RCT with statistically significant difference or no statistically significant difference but narrow confidence intervals • Systematic review[a,b] of level I RCTs (studies were homogeneous)	• High-quality prospective study[c] (all patients were enrolled at the same point in their disease with ≥80% follow-up of enrolled patients) • Systematic review[a,b] of level I studies	• Testing of previously developed diagnostic criteria in series of consecutive patients (with universally applied reference gold standard) • Systematic review[a,b] of level I studies	• Sensible costs and alternatives; values obtained from many studies; multiway sensitivity analyses • Systematic review[a,b] of level I studies
II	• Lesser-quality RCT (e.g., <80% follow-up, no blinding, or improper randomization) • Prospective[c] comparative study[d] • Systematic review[a,b] of level II studies or level I studies with inconsistent results	• Retrospective[f] study • Untreated controls from an RCT • Lesser-quality prospective study[c] (e.g., patients enrolled at different points in their disease or <80% follow-up) • Systematic review[a,b] of level II studies	• Development of diagnostic criteria on basis of consecutive patients (with universally applied reference gold standard) • Systematic review[a,b] of level II studies	• Sensible costs and alternatives; values obtained from limited studies; multiway sensitivity analyses • Systematic review[a,b] of level II studies
III	• Case control study[e] • Retrospective[f] comparative study[d] • Systematic review[a,b] of level III studies	• Case control study[e]	• Study of nonconsecutive patients (without consistently applied reference gold standard) • Systematic review[a,b] of Level-III studies	• Analyses based on limited alternatives and costs; poor estimates • Systematic review[a,b] of level III studies
IV	Case series[g]	• Case series[g]	• Case-control study[e] • Poor reference standard	• No sensitivity analyses
V	Expert opinion	• Expert opinion	• Expert opinion	• Expert opinion

Abbreviation: RCT, randomized controlled trial.
Source: Adapted from material published by the Centre for Evidence-Based Medicine. For more information, please see www.cebm.net.
[a]A complete assessment of the quality of individual studies requires critical appraisal of all aspects of the study design.
[b]A combination of results from two or more prior studies.
[c]Study was started before the first patient enrolled.
[d]Patients treated one way compared with patients treated another way.
[e]Patients identified for the study on the basis of their outcome, called *cases*, are compared with those who did not have the outcome, called *controls*.
[f]Study was started after the first patient enrolled.
[g]Patients treated one way with no comparison group of patients treated another way.

events in a small group of patients but offer little information regarding generalizations about the specific treatments, comparisons between groups are not possible, and assessing the outcomes and drawing conclusions based on small sample size are far-stretched. However, such study designs may provide important anecdotal information that lead to future investigations. For example, a recent case series addressing the proof-of-concept principle that magnetically controlled growing rods for the treatment of early onset scoliosis has provided new direction for the treatment of young children with severe deformities by eliminating the need of frequent surgeries and potential complications associated with traditional rod distraction.[5] Observational studies are also useful to document clinically meaningful yet unexpected outcomes, such as adverse effects after spinal treatments.

Case-Control Study—Level III

This study type begins by identifying a group of individuals with an outcome of interest (cases) and a group of individuals without the outcome of interest (controls). However, it is crucial that other than for the outcome being investigated, these groups are as similar as possible (i.e., that controls could

Table 2 Main types of quantitative research designs and their associated advantages and disadvantages

Type of study design	Advantages	Disadvantages
RCT (parallel design)	• In well-constructed study, unbiased distribution of confounds • Blindness of assessment • Can provide temporal sequence of events • Statistical analyses facilitated by randomization	• The ethical issue of who receives a specific treatment and if that treatment can possibly do more harm than good • Designs costly and time-consuming • Compliance issues or participants lost to follow-up may occur, which affects validity • Randomization techniques may be faulty; however, established methods could account for proper randomization of patients • Through time, contamination between groups may occur and should be accounted for • Biases (preallocation, selection, performance, detection, exclusion, publication)
RCT (crossover design)	• All subjects receive treatment and serve as own controls • Error variance reduced and sample size needed is reduced • Blinding may exist	• All subjects receive placebo or alternative treatment at some point • Unknown or lengthy washout period • Not applicable in treatments that are associated with permanent effects
Cohort study	• Determines the incidence of developing the disease in both types of groups • Confounders can be evaluated and their influence upon the outcome can be determine • Can evaluate various outcomes, detect associations, analyze time relationships, monitor changes over time, and assess rare and unique exposures • Can establish the relations between antecedent events and outcomes • Compared with RCT, less expensive and easier administration	• Can be time-consuming (if prospective in nature, but retrospective cohort designs may reduce time and subsequent costs) • Difficult to follow the original sample group through time • Blinding is difficult and randomization not present • Poor sample sizes and short follow-up for rare diseases • Various hidden confounding variables may affect outcome
Case-control study	• Beneficial in studying rare diseases or diseases with long duration to develop outcome • Retrospective and thus inexpensive and quick studies to conduct if time is an issue	• Obtaining an adequate representative control group may be difficult • Sampling bias may exist where defining a homogenous disease group as well as control group could be problematic and contain confounds • Demographics of the groups in question may prevent generalizing results and increasing external validity • How subjects are recruited may be questionable • Obtaining data to determine exposure may be difficult and prove challenging/time-consuming

(*Continued*)

Table 2 (*Continued*)

Type of study design	Advantages	Disadvantages
		• Due to the retrospective nature of the study, establishing a timeline when events occurred that may have contributed to the outcome is difficult, thus, obtaining consistent values of timing of events for both groups may not be probable • The study controls are selected from the investigator and can entail sampling bias, thus are not representative of the population as a whole and risk ratios cannot be analyzed • Recall bias by the participants may be present and dilute the validity of the results • Hidden confounders
Cross-sectional survey	• Inexpensive • Simple • Ethically sound	• Does not establish causality, but possible association • Potential for recall bias • Confounders unequally distributed • Unequal group sizes
Case series and case reports	• May provide insightful information into an area for further investigation • Provides insight for very rare diseases and their management	• Multiple and nonexistence of comparison group

Abbreviation: RCT, randomized controlled trial.
Source: Adapted from Samartzis D, Dominique DA, Perez-Cruet MJ, et al. Clinical outcome analyses. In: Perez-Cruet MJ, Khoo LT, Fessler RG, eds. An Anatomical Approach to Minimally Invasive Spine Surgery. St. Louis, MO: Quality Medical Publishing, Inc.; 2006:103–130.

become cases once they show the relevant end point). Data is collected regarding risk factors or exposures that may have contributed to the outcome of interest, and the resultant differences between the cases and controls are compared. Case control studies are attractive because they are simple and inexpensive to perform. This design is particularly interesting for rare or late-onset outcomes, which would make a prospective study prohibitively complex and expensive, as very large numbers of patients would have to be followed for an extended period of time.

Cohort Study (as a Comparative Prospective Study)—Level II

A cohort study seeks to identify a group of individuals with a risk factor of interest and another group of individuals without this risk factor and follows them over a prespecified period of time while measuring/counting the occurrence of outcomes of interest per group. Such a study can detect associations, analyze time relationships, monitor changes over time, and assess rare and unique exposures. The strength of such a research design lies in its ability to establish associations between antecedent events and subsequent outcomes, and a timeline of events is established. Cost, resource availability, and practicality of sustaining the study goals over time are important considerations preferably made a priori.

Randomized Controlled Study—Level I

RCTs are prospectively conducted and are either short-term (i.e., end points are assessed immediate postintervention) or longitudinal (i.e., end points are assessed at several months or years from intervention). The participants are randomized, which is defined by an unpredictable allocation of individuals to the treatment groups, thereby distributing equally any preexisting factors and biases or confounders that can potentially affect the outcome of interest. Typically, the baseline characteristics of the control and experimental arms should be the same before the trial starts to ensure that differences observed between the two groups at the end of the trial are attributable to the treatment or intervention administered during the study. The two principle subtypes of RCTs are the crossover and parallel designs. The parallel design has two independent experimental arms, usually defined as a control arm and an interventional arm. Two different treatments are administered, the groups are followed for a period of time,

and their outcomes are recorded. The group allocation is maintained for the duration of the study. By comparison, a crossover design represents two paired groups that receive the same treatment at one point or another, but at alternating times. In the crossover design, the main goal is to measure the treatment effect and determine whether a sequential or period effect exists. A crossover RCT is associated with concerns regarding any residual or carryover effects derived from the previous treatment. If these concerns are an issue, a washout period may be incorporated to minimize the effects of the previous treatment. Crossover designs usually do not lend themselves to surgical studies.

Systematic Reviews and Meta-Analyses—Levels I to III

A systematic review entails the systematic collection of data without formal mathematical processing or pooling of data. A meta-analysis is the "quantitative synthesis" or fusion of data to determine the effect size of the intervention across studies, if sufficient data can be found in the systematic review and if the data is homogeneous enough (both mathematically and clinically sensible) to allow such a fusion. The intention is to determine the quality of evidence and the effect size of the treatment.[6] Depending on the origin and quality of the primary data, systematic reviews/meta-analyses are regarded as level I to level III studies. The advantage of systematic reviews/meta-analyses, or a study of studies rather than a study of patients, is that it incorporates many patients often from large and diverse sample sources, different cultures, and countries/facilities in ways that cannot be approximated in a single-site study. Their obvious weakness is that they are only as good as their primary study sources and their reporting as systematic reviews/meta-analyses are data processing and not data generating.

Study Methodologies, Implementation and Analyses

After defining a study question of interest and choosing a valid and cost-effective design, the next problem that presents itself to the researcher is implementation—how to conduct a study. Fortunately, a helpful network of guidelines exist on how to conduct and report a study, such as the CONSORT statement for RCTs,[7] the STROBE panel for case control and cohort studies,[8] and the PRISMA statement for meta-analyses.[9] Also, several highly ranked journals require the use of these tools as a prerequisite for publication. However, a few issues crucial to the quality of a study deserve detailed mentioning.

Power and Sample Size

Statistical power has been one of the most neglected issues in clinical research.[10,11] Understanding statistical power is based on awareness of type I (i.e., false-positive) and type II (i.e., false-negative) errors. Type I error is expressed in form of the *p* value (or the α value), which can be calculated appropriately and without loss of validity at the end of a study. However, type II error has to be considered at the beginning of a study and cannot be addressed during the analysis stage. A power analysis shows how likely a study is to fall into the type II error trap; therefore, the greater the sample size, the higher the power of a study.[12,13]

Power can be thought of as a visibility problem—"If I want to see something very, very small I need a strong microscope, but if I want to see something very big I don't even need my glasses." In this metaphor, the size of the object is the anticipated effect size, the strong microscope is a large study, and the naked eye a small one. Hence, before the beginning of a study, an effect size (i.e., mean difference between fusion rates, risk of developing a spinal complication) has to be chosen as well as the appropriate type of microscope (i.e., study design and sample size) to help understand the power of a study. Choosing an effect size has to strike a balance between scientific rigor, clinical meaning, and cost, leading to a study that is able to show a scientifically important and clinically relevant difference without being prohibitively large. The second crucial parameter in addition to sample size is variance (e.g., standard deviation). An outcome with a high variance may imply that the precision around the effect size may be diminished and not as robust. Power may be achieved by the sample size at the end of the study (i.e., patients available for analysis), not at the beginning of the study (i.e., patients enrolled); as such, it makes sense to measure and account for attrition, usually by increasing the calculated required sample size.[13] However, if ample a priori evidence exists in the literature regarding effect sizes and their variation, one can estimate the sample size needed based on a predetermined chosen power threshold before the onset of a study, in particular if the study design is an RCT.

Intention to Treat and Per Protocol Analysis

For most clinical studies, follow-up is a critical; however, it is often elusive and patients are often lost to follow-up, and still others are not compliant with treatment. Some patients miss follow-up appointments or the documentation is incomplete. At the end of a study, there are always some patients who did not receive the treatment they were supposed to receive or were not followed sufficiently and the question remains whether or not, for all the inconsistencies, to exclude them from the analysis. An example is the often excessive amount of dropout/withdrawal rates seen in RCTs comparing cervical disk arthroplasty to that of anterior cervical diskectomy and fusion cases, whereby such effects render the groups practically noncomparable at follow-up.[14] Analysis after exclusion of such patients is called *per protocol*, because only patients who completed the protocol are considered and will produce an estimate of efficacy (what a treatment can do). Including all patients into the final analysis, regardless of their compliance, is called *intention to treat* analysis and will produce a clinically more realistic estimate, referred to as *effectiveness* (what a treatment actually does), which has usually a smaller effect size and a larger *p* value than a per protocol analysis.

Allocation Concealment and Blinding

A crucial point that unfortunately was not included in the term *randomized controlled trial* is blinding or concealed allocation. Preferably, allocation (i.e., group assignment) is concealed from both patients and outcome assessors because

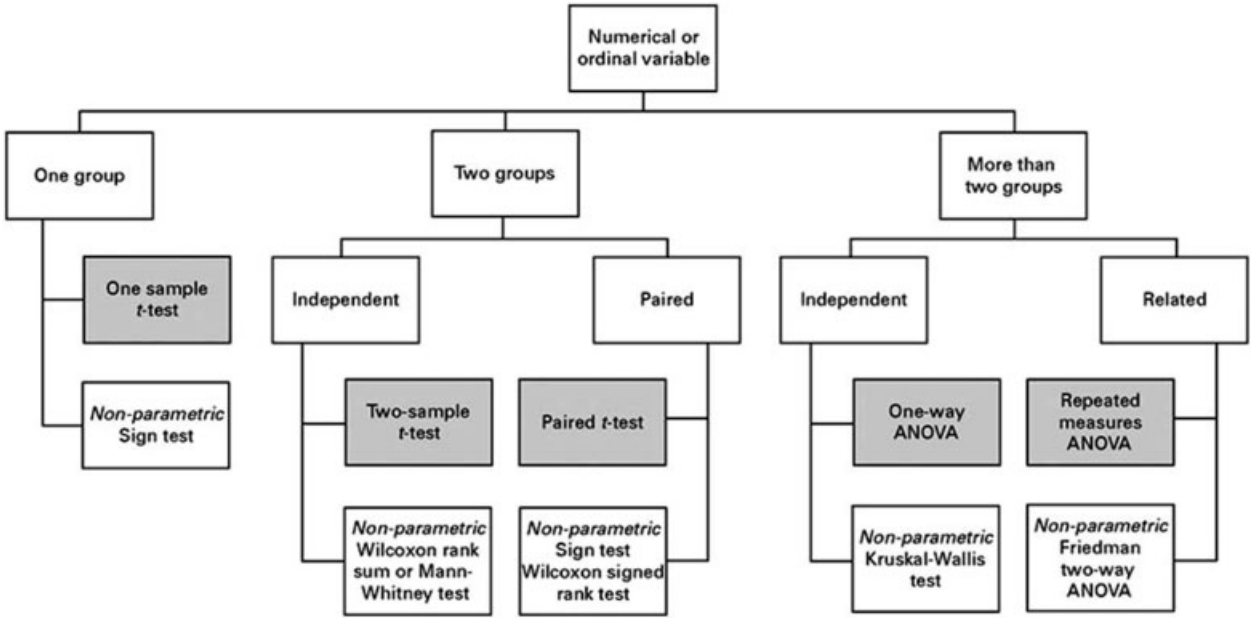

Fig. 1 Flow chart demonstrating appropriate statistical analyses tests when the values are numerical (continuous) or ordinal. (Adapted from Petrie A. Statistics in orthopaedic papers. J Bone Joint Surg Br 2006;88(9):1121–1136.[18]) Abbreviation: ANOVA, analysis of variance.

of the risk of influencing study outcomes (both willingly and subconsciously) based on the allocation. In surgical studies, concealed allocation is not always possible, especially when different surgical treatments or surgical and nonsurgical treatments are compared. However, outcome assessment can be often partially blinded (e.g., radiologic end points but not clinical exam), and sham operations have gained increasingly positive resonance as a method to minimize confounding factors of conventional prospective comparative study protocols (i.e., surgery devoid of a therapeutic step).

Confounding and Bias

Two important problems that arguably do not receive the attention they deserve are bias and confounding. Bias is the systematic deviation from the truth, as mentioned previously, and is usually caused by flawed study design. It can be avoided by diligent study conduct, but it cannot be ruled out, and its impact cannot be assessed. A confounding variable presents when there is a causal association between an exposure and an outcome to a varying degree based upon a third variable (the confounder). Fortunately, confounding can be tested and

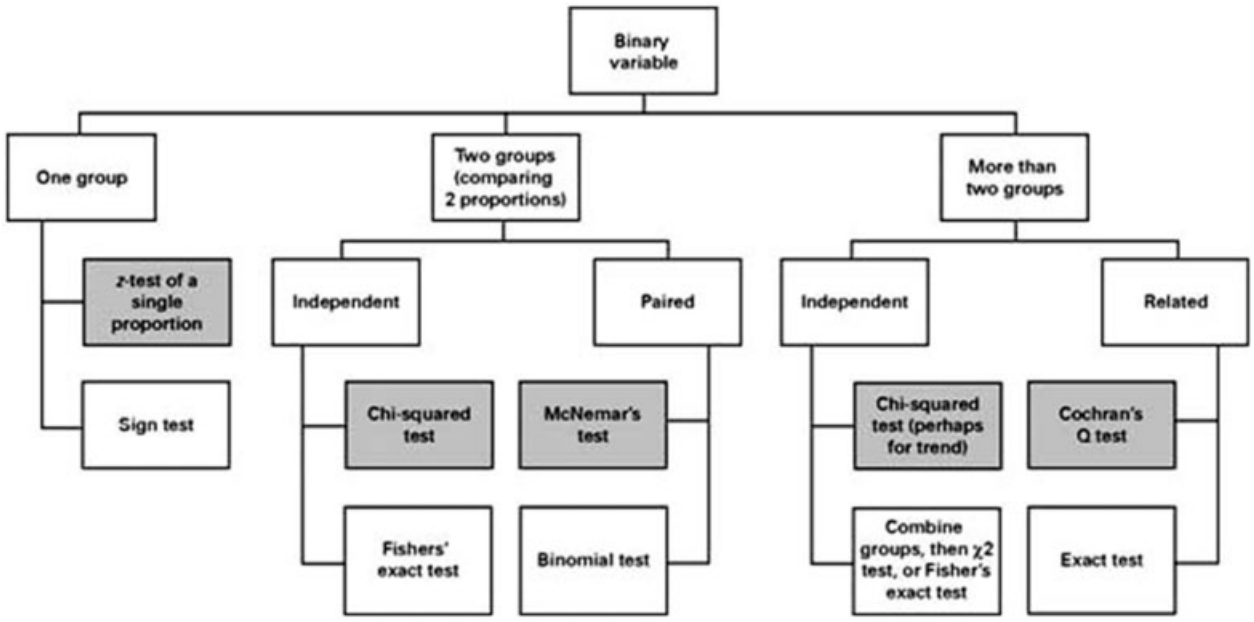

Fig. 2 Flow chart demonstrating appropriate statistical analyses tests when the values are binary. (Adapted from Petrie A. Statistics in orthopaedic papers. J Bone Joint Surg Br 2006;88(9):1121–1136.[18])

accounted for by including the confounding variable into the statistical analysis, such as by stratifying patients by their confounding status. In spine surgery, important examples of confounding include the effect of anesthetic medication or age on the interpretation of motor evoked potentials during pediatric spine surgery,[15,16] or the differential use of steroids in the medical treatment of a spinal cord injury with ganglioside GM1 or growth factors.[17]

Statistical Considerations

Statistical analysis of study results can be a daunting task, and seeking professional help from an individual with training in biostatistics is encouraged for all studies. However, in contrast to the complex, detailed methods, the rationale for statistical analysis is very straightforward. Two types of analysis can be distinguished: descriptive statistics and statistical inference. Descriptive statistics, as the name implies, give a systematic account of what was observed in a study. Statistical inference tries to put findings into relation and searches for associations between variables by formulating a hypothesis and using appropriate tests to prove or refute it.

Choosing the appropriate descriptive statistic and test depends on the type of variable used: binary (pain or no pain), categorical (e.g., disk bulge, extrusion, sequestered), ordinal (e.g., functional disability or lifestyle scores), or continuous/numerical (e.g., estimated blood loss, pain scores; ►**Figs. 1** and **2**).[18,19] The usual rationale behind statistical assessment is testing for superiority, present if one group is better than the other. However, recent developments in scientific methodology have come to emphasize noninferiority and equivalence. Briefly, such studies try to show that two interventions/groups have *substantively the same* outcome (equivalence) or one intervention/group is *no worse within a margin* (noninferiority). There are two good reasons to consider such designs. First, if a current gold standard treatment exist, it is unethical to not use it in a comparative study—a situation where it is advantageous to be able to show *no worse within a margin*. Second, most orthopedic treatments already have very good and excellent primary outcomes (e.g., single-level anterior cervical fusions) and leave only little room for improvement. However, a new technique could be superior in the secondary outcomes, safety, or cost-effectiveness—a situation where demonstrating that primary outcomes are *substantively the same* allows for comparison of such further outcomes.

Conclusion

Spine-related disorders are some of the most disabling conditions worldwide.[20] The prevalence of such disorders continues to rise and the management options continue to evolve. The assessment of the safety and efficacy of various therapeutic interventions or understanding of the natural course of disease is imperative to improve clinical decision making and take measures to enhance patient outcomes. For the field to progress and develop novel technologies to improve patient quality of life or establish preventative models, understanding study designs, their methodologies, and analytical considerations is imperative.

Funding
This work was supported by grants by the Hong Kong Theme-Based Research Scheme (T12–708/12N) and the Hong Kong Research Grants Council (777111).

Disclosures
Patrick Vavken, none
Anne Kathleen B. Ganal-Antonio, none
Francis H. Shen, none
Jens R. Chapman, none
Dino Samartzis, none

References

1 Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996; 312(7023):71–72
2 Poolman RW, Struijs PA, Krips R, Sierevelt IN, Lutz KH, Bhandari M. Does a "level I evidence" rating imply high quality of reporting in orthopaedic randomised controlled trials? BMC Med Res Methodol 2006;6:44
3 Voleti PB, Donegan DJ, Kim TW, Lee GC. Level of evidence: does it change the rate of publication and time to publication of American Academy of Orthopaedic Surgeons presentations? J Bone Joint Surg Am 2013;95(1):e2
4 Gnanalingham KK, Davies BM, Balamurali G, Titoria P, Doyle P, Abou-Zeid A. Improving levels of evidence in studies published in spinal journals from 1983 to 2011. Br J Neurosurg 2013;27(2): 152–155
5 Cheung KM, Cheung JP, Samartzis D, et al. Magnetically controlled growing rods for severe spinal curvature in young children: a prospective case series. Lancet 2012;379(9830): 1967–1974
6 Samartzis D, Perera R. Meta-analysis: statistical methods for binary data pooling. Spine J 2009;9(5):424–425
7 Moher D, Schulz KF, Altman D; CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285(15): 1987–1991
8 Poorolajal J, Cheraghi Z, Irani AD, Rezaeian S. Quality of cohort studies reporting post the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. Epidemiol Health 2011;33:e2011005
9 Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700
10 Altman DG, Moher D, Schulz KF. Peer review of statistics in medical research. Reporting power calculations is important. BMJ 2002; 325(7362):491, author reply 491
11 Freedman KB, Back S, Bernstein J. Sample size and statistical power of randomised, controlled trials in orthopaedics. J Bone Joint Surg Br 2001;83(3):397–402
12 Mayer JD. Sample size calculations. Must do better. BMJ 2009;338: b2323
13 Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ 2009;338:b1732

14 Garrido BJ, Wilhite J, Nakano M, et al. Adjacent-level cervical ossification after Bryan cervical disc arthroplasty compared with anterior cervical discectomy and fusion. J Bone Joint Surg Am 2011;93(13):1185–1189
15 Balvin MJ, Song KM, Slimp JC. Effects of anesthetic regimens and other confounding factors affecting the interpretation of motor evoked potentials during pediatric spine surgery. Am J Electroneurodiagn Technol 2010;50(3):219–244
16 Lieberman JA, Lyon R, Feiner J, Diab M, Gregory GA. The effect of age on motor evoked potentials in children under propofol/isoflurane anesthesia. Anesth Analg 2006;103(2):316–321
17 Sipski ML, Pearse DD. Methylprednisolone and other confounders to spinal cord injury clinical trials. Nat Clin Pract Neurol 2006;2(8):402–403
18 Petrie A. Statistics in orthopaedic papers. J Bone Joint Surg Br 2006;88(9):1121–1136
19 Petrie A, Sabin C. Medical Statistics at a Glance. 2nd ed. Oxford, UK: Blackwell Publishing; 2000
20 Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380(9859):2163–2196