Rationale for and Methods of Superiority, Noninferiority, or Equivalence Designs in Orthopaedic, Controlled Trials

Patrick Vavken MD, MSc, FRSPH

Received: 16 July 2010 / Accepted: 6 January 2011 / Published online: 19 January 2011
© The Association of Bone and Joint Surgeons® 2011

Abstract

Background To provide value-based healthcare in orthopaedics, controlled trials are needed to assess the comparative effectiveness of treatments. Typically comparative trials are based on superiority testing using statistical tests that produce a p value. However, as orthopaedic treatments continue to improve, superiority becomes more difficult to show and, perhaps, less important as margins of improvement shrink to clinically irrelevant levels. Alternative methods to compare groups in controlled trials are noninferiority and equivalence. It is important to equip the reader of the orthopaedic literature with the knowledge to understand and critically evaluate the methods and findings of trials attempting to establish superiority, noninferiority, and equivalence.

Questions/purposes I will discuss supplemental and alternative methods to superiority for assessment of the outcome of controlled trials in the context of diminishing returns on new therapies over old ones.

Methods The three methods—superiority, noninferiority, and equivalence—are presented and compared, with a discussion of implied pitfalls and problems.

Results Noninferiority and equivalence offer alternatives to superiority testing and allow one to judge whether a new treatment is no worse (within a margin) or substantively the

same as an active control. Noninferiority testing also allows for inclusion of superiority testing in the same study without the need for adjustment of the statistical methods. **Conclusions** Noninferiority and equivalence testing might prove most valuable in orthopaedic, controlled trials as they allow for comparative assessment of treatments with similar primary end points but potentially important differences in secondary outcomes, safety profiles, and cost-effectiveness.

Introduction

There has been a trend toward increased use of the principles of evidence-based medicine in orthopaedic surgery [4, 26]. Probably best known and most commonly encountered is the categorization of publications based on levels of evidence. These levels are a measure of the risk of bias in a study, ranking from I to V where Level I evidence is understood to have the least risk of bias, usually stemming from high-quality randomized controlled trials or meta-analyses [9]. It is agreed such trials should satisfy numerous criteria, such as a priori sample-size calculation, randomized and concealed allocation, blinded outcome assessment, and intention-to-treat (ITT) testing, whereby participants are analyzed based on the group to which they were randomized even should they fail to receive the treatment to which they were assigned [5, 6, 19, 30, 56]. Randomized controlled trials (RCT) directly comparing available treatments have become increasingly important in the quest for value-based healthcare where investigators try to establish clinical effectiveness, safety, and cost-effectiveness for different treatments to identify and recommend those with the best overall efficiency.

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

P. Vavken (✉)
Department of Orthopedic Surgery, Children's Hospital Boston,
Harvard Medical School, 300 Longwood Avenue, Enders 1016,
Boston, MA 02115, USA
e-mail: patrick.vavken@childrens.harvard.edu

Current orthopaedic literature contains little discussion regarding how to compare the groups in a controlled trial. This problem begins with the choice of controls, which either can be placebo or active control, ie, a currently used treatment [10, 49]. Most orthopaedic trials use active control designs, in which a new treatment is compared with an old or the current treatment. Usually such studies try to show superiority by testing for evidence of a statistically significant difference between treatment groups with the group having the more favorable absolute outcome being judged as better. Although this approach seems fairly straightforward on paper, there is potential for problems, especially in orthopaedic surgery. Using placebo, sham, or negative control groups often implies that a group of patients will be denied a necessary and helpful treatment and, in such situations presents ethically difficult choices. However, there is no reason to show a new treatment produces better results than no treatment at all if an effective treatment already exists for the same ailment.

A much envied characteristic of our specialty is that there are numerous efficacious treatments where new treatments are likely to offer little improvement. For example, 10-year survivorship after total hip replacement ranges between 90% to 96% [14, 15, 31, 32]. Similarly, the assessment of Short Form-36 quality-of-life subscale outcomes (physical function, bodily pain) after total hip and total knee replacements and ACL replacement showed consistent effect sizes greater than 80% [8]. Several studies comparing 55- to 74-year-old patients undergoing total hip or knee replacement with an age-matched sample of a healthy population, suggest equal or better function compared with those of their healthy peers [33, 57]. The relative success of these procedures creates a ceiling effect. That is, a new treatment can hardly have a better survivorship than 96%, and if so, only by a small margin. A similar situation has become obvious in some studies of minimally invasive or navigated total joint arthroplasties [3, 22, 52, 60]. However, this does not mean a new treatment might not be better in other parameters, such as secondary functional outcomes, safety, or cost-effectiveness. Therefore the task may be to show superiority on secondary parameters, while simultaneously showing comparable efficacy on some primary outcome.

The need to show comparable efficacy of two or more groups leads to two alternative frameworks: noninferiority and equivalence [28, 39]. These methods have been used for a considerable time in pharmaceutical studies, are endorsed and recommended by the US Food and Drug Administration (FDA) and corresponding European institution, the European Medicines Agency (EMA), and are slowly proceeding into the surgical field [24, 25]. However, there are inconsistencies in the literature regarding the use and quality of noninferiority and equivalence trials. One

study produced estimates that 67% of equivalence studies published between 1992 and 1996 were inappropriately named [21]. Another study of 90 randomized controlled trials from surgical specialties revealed only 39% met the criteria for establishing equivalence [11]. Nevertheless, owing to the increasing numbers of such studies and the thus far observed problems with terminology and methodology, the CONSORT group has published guidelines and checklists for reporting noninferiority and equivalence trials as an extension of their earlier publication on reporting randomized trials [35, 39].

The objective of this study is to: (1) describe the fundamental concepts and indices (p values, confidence intervals) of classic superiority statistical testing and highlight the capabilities and limitations of superiority testing, (2) explain how these fundamental concepts and indices lead to alternative designs such as noninferiority and equivalence trials, and, (3) describe the use of these alternative designs in the current literature.

Superiority—Looking for Better

Superiority trial design and analysis, as we most commonly encounter it today, started in the early 1900s, when William S. Gosset encountered the problem of having to find the best barley to brew Guinness beer [50, 51]. He developed a mathematical method, the t-test, to compare small samples and published it using his pen name “Student” because of Guinness’ nondisclosure policy, thus the “Student’s t-test” [50, 51]. This work laid the cornerstone for two related indices pertaining to superiority (and to noninferiority and equivalence): p values and confidence intervals [59].

The p value is the traditional output of null hypothesis testing statistics such as the t-test, analysis of variance (ANOVA), and regression analysis, among others. The p value is the likelihood to see a given result, or a more extreme (even smaller, even larger) one, by random chance [12]. For example, if a difference in mean blood loss of 20 ± 0.3 mL associated with a p value of 0.002 were to be observed in a study comparing surgical procedures A and B, there would be a 0.2% ($0.002 \times 100\%$) chance that this difference, or a larger one, could be seen owing to random chance and a 99.8% ($100\% - 0.2\%$) chance of a true difference in blood loss between procedures A and B. In other words, it is the proportion of times one might expect an effect greater than or of equal size would emerge when the true effect is zero (no difference between A and B), given your sample size. If this is sufficiently unlikely, we infer that the null hypothesis, or no difference between A and B in this example, is not true, and therefore conclude that there must be a difference. Fisher was instrumental in

the creation of the p value. He suggested using 5% as a minimum threshold for concluding that there is evidence against a null hypothesis, but understood the p value as a range of values describing the strength of evidence, rather than a binary cutoff at 5% [12, 17, 48]. Even the value of 5% is not set in stone as minimum threshold. The larger the number of tested hypotheses, the lower the value should be. Also, if prior knowledge of the likelihood exists, the p value should be adjusted, as it should be in tests with intrinsically low power [12, 48]. Finally, it is important to consider that the p value is a compound measure which incorporates central tendency (eg, mean difference), variability, and sample size. The dependence of the p value on sample size and mean difference can result in two problematic situations where statistical results contradict clinical relevance: (1) large, and biologically or clinically meaningful mean differences that are not statistically significant because of very small sample sizes, or (2) marginal or meaningless differences that are statistically significant because of very large sample sizes [59]. The (hypothetical) example above is such a situation where a statistically significant difference of 20 mL blood loss probably bears little clinical relevance. Real life examples often are found in meta-analyses where numerous studies

are combined to achieve very large sample sizes, leading to results such as statistically different ($p < 0.0001$) blood loss between minimally invasive and standard total hip replacements (of 43.4 mL) [58, 60], or statistically different ($p = 0.0002$) joint space narrowing (of 0.13 mm) after treatment with chondroitin sulfate or no treatment in knee osteoarthritis [23, 55, 58]. Another limitation is that the p value is of entirely different scale from the measure of interest, making it difficult to assess the clinical importance of the difference even when statistical significance is observed [59].

In contrast to the p value, the confidence interval is designed to give a sense of the size and significance of the difference in the original units of the measure [13]. It represents the estimated range in which some percentage (eg, 95% for a 95% confidence interval) of the mean differences would fall, given an infinite number of replicates of the same study. Should that range not include 0 for mean differences of continuous variables or 1 for ratios (both of which are consistent with no difference), one may infer statistical significance at a two-tailed, 100% minus width-of-the-confidence-interval alpha level (eg, 95% CI allows two-tailed inference at 100%–95% or 5% alpha) (Fig. 1). Confidence intervals are particularly useful when assessing

Fig. 1 A diagram illustrates the principle of superiority testing in a controlled trial. The effects of an active control and a new treatment are given as 95% CIs of the difference between treatments, measured along the x-axis. The dotted vertical line represents a difference between groups of zero. (1) This 95% CI includes zero implying that there is no difference between groups with an alpha of 5%. (2) The entire 95% CI lies below the zero line on the side of less favorable outcome, implying a significantly worse, inferior result for this

particular comparison. (3) This is a comparison that resulted in a significantly better, superior result, shown by a 95% CI that lies entirely on the more favorable side of the zero line. (4) This is the same mean result as for (3) (depicted by the squares at the center of the CI), but the 95% CI is much wider than the 95% CI in 3, suggesting that the estimate of the difference between treatments is more precise in 3. The reason for this wider 95% CI is either a larger variance or a smaller sample size.

whether a statistically significant effect is of a large enough size to be of clinical importance, something that cannot be accomplished with *p* values alone [59]. This same feature leads to the other two trial designs.

Noninferiority—Being No Worse (within a margin)

Statistical testing, as described above, judges differences between treatments and is capable only of supplying evidence of a difference [1, 28], not a certainty of difference. A nonsignificant *p* value means we cannot conclude the two samples are drawn from different populations, but it is important to understand this does not imply that they are from the same population [1, 29]. Even should two groups in a trial produce the numerically same outcome for a studied end point and a nonsignificant *p* value, one has to consider that might be attributable to variability masking a true difference, ie, that one, or even both groups, have produced outliers that happen to be identical but that would not be reproduced in a new study. However, there are some arguments why it is important to ask that a new treatment be no worse (within a margin), rather than better (Table 1). For instance, a new treatment may be designed to be safer (ie, fewer complications or less severe complications), less expensive, or otherwise more desirable, while providing similar efficacy on a

primary outcome to that of an accepted contemporary treatment. When the new treatment may provide better efficacy, but it was designed based on improving the secondary measures, noninferiority testing is desirable as it also allows the potential to establish that it is better. Noninferiority assessment is one-sided testing, ie, it does not allow the possibility that the new treatment is worse (plus a noninferiority margin) than the active control, but better is also noninferior, and noninferiority testing does not preclude establishing superiority. Confidence intervals supply us with a means to test this hypothesis.

A noninferiority trial starts with defining “no worse (within a margin)” or noninferior before the beginning of the trial (Fig. 2). This requires the definition of an outcome and a threshold in that outcome, below which one would consider a new treatment to be inferior to an older one. One method to define such a margin is to use a value lower than the mean of a current treatment plus a noninferiority margin [10, 20, 24, 45]. The size of this margin should be determined during the design stage of a study before the actual experiment. Choosing the size is difficult, as there are no explicit rules. Usually findings from earlier studies and estimates of clinically relevant differences are combined. For example, the currently used anticoagulant A serves as an active control for a study of the new anticoagulant B. A reduces the incidence of deep venous thrombosis after total hip replacement by 80% and a

Table 1. Reasons for choosing noninferiority over superiority designs

Reason	Explanation
Comparing a new treatment with an active control instead of placebo	It would be unethical to use a placebo group in a controlled study, for example, in a controlled trial of navigated joint arthroplasty. However, testing for superiority of clinical scores after navigated TKA over standard TKA might prove futile because the already extremely high success rates of standard TKA do not leave much room for improvement.
The new treatment is not better in the primary end point but better in secondary end points	Numerous studies have compared minimally invasive hip and knee arthroplasty with the standard procedure without producing statistically significant and clinically meaningful differences for end points such as component alignment, dislocation rates, or scores at 6 months after THA/TKA [3, 22, 52, 60]. However, although minimally invasive arthroplasty probably is not inferior to standard, secondary outcomes such as pain during the first postoperative week and mobilization produce better results, although they cannot outweigh crucial primary end points. Thus, being noninferior in primary end points and superior in secondary end points might be a strong argument for minimally invasive surgery.
The new treatment is not better in the primary end point, but overall efficiency is better	Data from a recent meta-analysis suggest that in the prevention of heterotopic ossification, NSAIDs could be noninferior to radiation in effectiveness and safety, but they are clearly superior in incurred cost, thus producing a greater overall efficiency [55]. Given the strong and strongly growing emphasis on value-based healthcare and cost-effectiveness, these are important findings and might be included in some form in society guidelines and health policy regulations in the near future.
The new treatment can be noninferior and superior	Noninferiority testing is a one-sided test at the “bad end” of the effect of the active control, looking for at least as good. That does not rule out that a new treatment is superior to the active control. Noninferiority testing can be complemented by superiority testing in one study without the need for adjusting for multiple testing or loss of power or validity.

Fig. 2 This diagram illustrates the principle of noninferiority testing in a controlled trial comparing a new treatment with an active control. An a priori defined noninferiority margin (NIM)(dashed vertical line) is added to the line of zero difference between treatments (dotted vertical line). As noninferiority testing is one-sided, this establishes a cutoff and everything toward more favorable outcomes is considered noninferior (gray arrow). Even with the NIM, this cutoff is still well above the effect of a placebo, shown in Treatment 4. Treatment 1a

therefore is noninferior as it is above this cutoff but not significantly different from the active control as the 95% CI contains the zero difference line (dotted vertical). Treatment 1b also is noninferior as it is above this cutoff but also is significantly different from the active control as the 95% CI is entirely above the zero difference line (dotted vertical). Treatment 2 is inconclusive as its 95% CI includes values on both sides of the noninferiority threshold. Treatment 3 is below the cutoff and thus inferior. Treatment 4 is a putative placebo control.

noninferiority margin of 5% of this effect is assumed “no worse”. Therefore B would be required to reduce the incidence of deep venous thrombosis by $80\% \times 0.95$, ie, 76%. To test for this, the 95% CI for the difference of the incidence of deep venous thrombosis with new treatment minus the incidence for the active control is plotted against this threshold and interpreted (Fig. 2). Alternatively, a one-sided test can be used because in noninferiority testing we are interested only in whether B is no worse than 76%, ie, the probability to lie on the left side of the probability distribution curve. “Being no worse than 5% worse than the effect of the active control” also could be interpreted as trying to maintain at least 95% of the effect of the active control. However, there is a risk of so-called “biocreep” [16, 18] (Fig. 3), which refers to the problem of having a new treatment B that is no worse than 5% of the current standard of care treatment A, and then comparing an even newer treatment C again by 5% with treatment B, and then treatment D with treatment C, and so on. Although the noninferiority margin of 5% has never been violated in individual comparisons, treatment Z is far from the effect of the original treatment A. The definition and use of such a margin might seem arbitrary to some, but it actually is more rigorous than a superiority design because it involves

a predefined minimum difference and statistical testing for this specific difference, whereas superiority trials assess only the significance but not the size of a difference.

Another method is to establish a margin that is significantly different from a putative placebo control by using data from pilot studies, earlier studies, or meta-analyses. A benefit of this approach is the inclusion of a placebo or negative control because noninferiority is a closed comparison of two treatments [27]. The problem in closed comparison is that it sometimes can be difficult to differentiate whether a new treatment approximated the effect of the well performing, active control (ie, both treatments worked equally well) or whether the active control approximated the effect of an ineffective, new treatment owing to a problem in the study (ie, both treatments failed clinically to produce a positive outcome). Such problems could be lacking compliance, implant failure, a poorly chosen model, patient crossover or losses to followup, or any other treatment failure. In both scenarios, the new treatment will “be no worse (within a margin)” than the active control, but it is obvious that the latter (no difference between treatments because both failed) should not be interpreted as evidence for the effectiveness of a new treatment. Both methods can be combined to define a

Fig. 3 The reiterative use of noninferior treatments as new active controls for the next study leads to an overall reduction of effectiveness from Treatments 1 to 5 although the noninferiority threshold was never violated in individual studies. The effect of a placebo treatment is lower than the effect in the first, second, and third noninferiority studies, but by the time of the fourth study, the noninferiority threshold (using the noninferior Treatment 3 as new active control) has crept to levels consistent with placebo treatment. NIM = noninferiority margin.

margin that is significantly better than placebo and clinically meaningful.

Next the investigator must establish the population to test. For superiority trials, testing the ITT population is recommended [28, 37], which means patients are analyzed by their initial allocation, regardless of attrition, missed followups, lacking compliance, crossovers, etc, to produce a conservative estimate that reflects the real life situation [6]. Currently it is not unequivocally clear whether ITT leads to a more conservative estimate in noninferiority trials, and it has been suggested that other factors related to study design, patient flow, and statistical analysis influence the conservatism of ITT analyses in noninferiority studies [34]. However, recommendations steer toward the use of ITT, mostly for the sake of consistency with superiority testing [6].

The investigator also must determine sample size for noninferiority. This is straightforward. In superiority testing, the sample size depends on the expected difference between treatments, and analogously, in noninferiority the sample size depends on the expected noninferiority margin [44, 47, 53]. Depending on how large or small this margin is chosen, the required sample size in a noninferiority trial can be substantial. Usually, noninferiority and equivalence studies require (much) larger sample sizes than superiority studies because the typical sizes of the anticipated margins in noninferiority and equivalence studies are (much) smaller than what would be considered a clinically meaningful difference per se between groups in superiority studies.

The question emerges whether noninferiority and superiority designs may be combined in one trial without affecting validity or power. As mentioned above, superiority and noninferiority have distinct testing principles, and adding superiority testing to noninferiority testing is not consistent with multiple comparison testing and can be done without loss of power and the need to adjust p values [28, 37, 38, 61]. It also is valid to test for noninferiority for one (group of) end point(s) and for superiority in other end points in one study [28, 36]. Nevertheless, as with all methodologic details of a study, the analysis plan, use of these approaches individually or in combination, directions of the hypothesis test(s), and required sample size have to be determined a priori.

Equivalence—They Are Substantively the Same

Much of the methodology of equivalence trials comes from studies of pharmaceutical bioequivalence. Other than clinical scores after surgical treatment, pharmaceuticals have a therapeutic window, meaning that too low (ineffective) and too high (toxic) concentrations are detrimental or even dangerous. To account for insufficiently low blood levels and overdoses, equivalence designs have lower and upper margins to show two treatments are equivalent. Again, a nonsignificant p value does not mean two treatments are the same [29]. Similar principles as outlined for noninferiority are applied in equivalence testing and will not be reiterated here (Fig. 4). An equivalence margin is

Fig. 4 The effects (95% CIs for the difference between two treatments) of an active control and a new treatment are given. Equivalence testing is a two-sided test; thus, the equivalence range (gray area), established by the line (dotted) of no difference between an active control and a new treatment plus an equivalence margin is confined on both sides (dashed lines). Treatment 1 is within this range

and thus equivalent to the active control. Treatment 2 is inconclusive as it includes values inside and outside the equivalence range. Treatment 3 is not equivalent, but superior, as it is completely on the more favorable side of the equivalence range. Contrary to noninferiority, equivalence is two-sided and implies not worse but also no better. Treatment 4 is not equivalent and worse than the active control.

estimated and added to either side of the effect of the active treatment, and the effect of the new treatment is tested against this range. This can be done via CIs or using statistical tests. However, equivalence testing is two-sided, meaning a new treatment is equivalent only if it is no better and no worse (both within a margin) than the active control, but noninferiority and equivalence often are confused [11, 21]. One reason might be the better, more positive ring of being equivalent rather than noninferior, despite the fact that this actually inverts the real situation where a noninferior treatment has potential for superiority, whereas an equivalent treatment, by definition, cannot be better than the active control.

Frequency and Quality of Noninferiority Trials and Equivalence Trials in the Literature

Noninferiority and equivalence trials have become common in the assessment of controlled trials in general medicine [54]. However, assessing their frequency is fairly complicated because the terminology is inconsistent and sometimes incorrect. For example, as mentioned earlier, statistically insignificantly different results often are incorrectly interpreted as equivalent, using equivalent not

as a technical term in study design but synonymous with “the same” [39]. One study reported 0.2% of the approximately 1/2-million trials in the Cochrane Central Register of Controlled Trials in 2004 included the words “equivalence” or “noninferiority” [21]. Given that the overall percentage of controlled trials among all orthopaedic publications ranges between 4% and 8%, the frequency of noninferiority trials and equivalence trials probably is low [41, 56]. However, although noninferiority and equivalence are used infrequently in academic, surgical research, they are preferred by the FDA and EMA, and therefore have a striking effect on orthopaedics as a whole.

Discussion

Noninferiority and equivalence designs add valuable tools for evaluation of findings of orthopaedic, controlled trials; both have been endorsed by the FDA and EMA. Reasons for conducting such studies include comparison of a new treatment with an active control rather than a placebo, establishment of a new treatment with better secondary outcome(s) and noninferior primary outcome, or as the first step in testing superiority of a new treatment. However, there also are potential weaknesses in noninferiority

testing. One is the potential to flood the healthcare market with ‘me too’ procedures and products [40] that are non-inferior to current gold standard treatments but do not add additional value. Another potential pitfall is biocrep, ie, the iterative process of establishing noninferiority to the current gold standard of a slightly less effective, new treatment, followed by the use of this new treatment as a gold standard for an even newer, noninferior but again slightly less effective treatment, and so on. Finally, methodologic rigor is even more important in noninferiority than in superiority trials because of the problem of confusing noninferiority with a Type I error [40]. Briefly, this can be understood as follows: a superiority trial looks for a statistically different/clinically meaningful difference, and all flaws in study design and conduct make it harder to find such a difference, which makes this a conservative design (erring on the save side). Noninferiority looks for similarity, or “no statistically different/clinically meaningful difference” as similarity cannot be tested directly. As in superiority trials, flaws in study design and conduct will make it harder to find a difference between treatments, but as “no difference” is the preferred outcome, noninferiority testing is anticonservative: a poorly designed and conducted noninferiority study has a greater chance of a false positive outcome, a Type I error.

The true value of noninferiority studies should be considered in the big scheme of evidence- and value-based orthopaedics. A good example might be the orthopaedic implant market, especially knee and hip prostheses. Currently, this market is characterized by high and steadily increasing prices [2, 7], physician preferences for particular implants [42], and pricing power in the hands of the supplier. The indiscriminate use of noninferiority studies would lead to the addition of no worse, but not necessarily better products to this market, resulting in a price hike on the supply side (device manufacturers trying to compensate research and development costs and make a profit) and the demand side (hospitals having to buy new implants and equipment). This might open new market segments and produce a competitive advantage over business rivals, but will not necessarily improve the provision of effective and needed products. However, the deliberate use of noninferiority studies in conjunction with assessment of secondary outcomes (eg, infection rates, safety, direct and indirect costs, etc) would introduce important information to this market and could help establish healthy competition, transparent cost-effectiveness, and lower prices with increased supply [42, 43, 46].

The most important advantage of noninferiority and equivalence trials is that both designs allow comparison with currently existing, clinically accepted treatment, even if there is a ceiling effect. Additionally, both designs allow shifting the focus of attention to secondary, but no less

important, outcomes and thus enable investigators to compile a more complete picture of the effectiveness of a new treatment in comparison to a current gold standard.

Acknowledgment I am greatly indebted to Jason T. Machan PhD, who offered substantial advice and crucial guidance during revision of this manuscript.

References

- Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.
- Avisory Board. Reducing orthopedic implant costs. Washington, DC: The Advisory Board; 2004.
- Bauwens K, Matthes G, Wich M, Gebhard F, Hanson B, Ekkernkamp A, Stengel D. Navigated total knee replacement: a meta-analysis. *J Bone Joint Surg Am*. 2007;89:261–269.
- Bhandari M, Giannoudis P. Evidence-based medicine: what it is and what it is not. *Injury*. 2006;37:302–306.
- Bhandari M, Richards RR, Sprague S, Schemitsch EH. The quality of reporting of randomized trials in the Journal of Bone and Joint Surgery from 1988 through 2000. *J Bone Joint Surg Am*. 2002;84:388–396.
- Bubbar VK, Kreder HJ. The intention-to-treat principle: a primer for the orthopaedic surgeon. *J Bone Joint Surg Am*. 2006;88:2097–2099.
- Burns LR, Housman MG, Booth RE Jr, Koenig A. Implant vendors and hospitals: competing influences over product choice by orthopedic surgeons. *Health Care Manage Rev*. 2009;34:2–18.
- Busija L, Osborne RH, Nilsdotter A, Buchbinder R, Roos EM. Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. *Health Qual Life Outcomes*. 2008;6:55.
- Centre for Evidence-Based Medicine, Oxford, UK. Available at: www.cebm.net. Accessed November 2, 2010.
- D’Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues: the encounters of academic consultants in statistics. *Stat Med*. 2003;22:169–186.
- Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: equivalency or error? *Arch Surg*. 2001;136:796–800.
- Dorey FJ. In brief: The p value: what is it and what does it tell you? *Clin Orthop Relat Res*. 2010;468:2297–2298.
- Dorey FJ. In brief: statistics in brief: confidence intervals: what is the real result in the target population? *Clin Orthop Relat Res*. 2010;468:3137–3138.
- Eskelinen A, Paavolainen P, Helenius I, Pulkkinen P, Remes V. Total hip arthroplasty for rheumatoid arthritis in younger patients: 2,557 replacements in the Finnish Arthroplasty Register followed for 0–24 years. *Acta Orthop*. 2006;77:853–865.
- Eskelinen A, Remes V, Helenius I, Pulkkinen P, Nevalainen J, Paavolainen P. Total hip arthroplasty for primary osteoarthritis in younger patients in the Finnish arthroplasty register: 4,661 primary replacements followed for 0–22 years. *Acta Orthop*. 2005;76:28–41.
- Everson-Stewart S, Emerson SS. Bio-creep in non-inferiority clinical trials. *Stat Med*. 2010;29:2769–2780.
- Fisher RA. *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd; 1925.
- Fleming TR. Current issues in non-inferiority trials. *Stat Med*. 2008;27:317–332.
- Freedman KB, Back S, Bernstein J. Sample size and statistical power of randomised, controlled trials in orthopaedics. *J Bone Joint Surg Br*. 2001;83:397–402.

20. Gao P, Ware JH. Assessing non-inferiority: a combination approach. *Stat Med*. 2008;27:392–406.
21. Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence? *Ann Intern Med*. 2000;132:715–722.
22. Hart R, Janecek M, Cizmar I, Stipcak V, Kucera B, Filan P. [Minimally invasive and navigated implantation for total knee arthroplasty: X-ray analysis and early clinical results] [in German]. *Orthopade*. 2006;35:552–557.
23. Hochberg MC. Structure-modifying effects of chondroitin sulfate in knee osteoarthritis: an updated meta-analysis of randomized placebo-controlled trials of 2-year duration. *Osteoarthritis Cartilage*. 2010;18(suppl 1):S28–S31.
24. Hung HM, Wang SJ, O'Neill R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biom J*. 2005;47:28–36; discussion 99–107.
25. Hung HM, Wang SJ, O'Neill R. Challenges and regulatory experiences with non-inferiority trial design without placebo arm. *Biom J*. 2009;51:324–334.
26. Kocher MS, Zurakowski D. Clinical epidemiology and biostatistics: a primer for orthopaedic surgeons. *J Bone Joint Surg Am*. 2004;86:607–620.
27. Lasagna L. Placebos and controlled trials under attack. *Eur J Clin Pharmacol*. 1979;15:373–374.
28. Lesaffre E. Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis*. 2008;66:150–154.
29. Lesaffre E. Use and misuse of the p-value. *Bull NYU Hosp Jt Dis*. 2008;66:146–149.
30. Lochner HV, Bhandari M, Tornetta P 3rd. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. *J Bone Joint Surg Am*. 2001;83:1650–1655.
31. Makela KT, Eskelinen A, Paavolainen P, Pulkkinen P, Remes V. Cementless total hip arthroplasty for primary osteoarthritis in patients aged 55 years and older. *Acta Orthop*. 2010;81:42–52.
32. Makela KT, Eskelinen A, Pulkkinen P, Paavolainen P, Remes V. Total hip arthroplasty for primary osteoarthritis in patients fifty-five years of age or older: an analysis of the Finnish arthroplasty registry. *J Bone Joint Surg Am*. 2008;90:2160–2170.
33. March LM, Cross MJ, Lapsley H, Brnabic AJ, Tribe KL, Bachmeier CJ, Courtenay BG, Brooks PM. Outcomes after hip or knee replacement surgery for osteoarthritis: a prospective cohort study comparing patients' quality of life before and after surgery with age-related population norms. *Med J Aust*. 1999;171:235–238.
34. Matilde Sanchez M, Chen X. Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. *Stat Med*. 2006;25:1169–1181.
35. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*. 2001;357:1191–1194.
36. Morikawa T, Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. *J Biopharm Stat*. 1995;5:297–306.
37. Moye L. *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*. New York, NY: Springer; 2003.
38. Ohn F, Jennison C. Optimal group-sequential designs for simultaneous testing of superiority and non-inferiority. *Stat Med*. 2010;29:743–759.
39. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA*. 2006;295:1152–1160.
40. Pocock SJ. The pros and cons of noninferiority trials. *Fundam Clin Pharmacol*. 2003;17:483–490.
41. 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.
42. Robinson JC. Value-based purchasing for medical devices. *Health Affairs*. 2008;27:1523–1531.
43. Rosenthal MB. What works in market-oriented health policy? *N Engl J Med*. 2009;360:2157–2160.
44. Rothmann M, Li N, Chen G, Chi GY, Temple R, Tsou HH. Design and analysis of non-inferiority mortality trials in oncology. *Stat Med*. 2003;22:239–264.
45. Scott IA. Non-inferiority trials: determining whether alternative treatments are good enough. *Med J Aust*. 2009;190:326–330.
46. Simoons S. Health economics of medical devices: opportunities and challenges. *J Med Econ*. 2008;11:713–717.
47. Steinijans VW, Neuhauser M, Bretz F. Equivalence concepts in clinical trials. *Eur J Drug Metab Pharmacokinet*. 2000;25:38–40.
48. Sterne JA, Davey Smith G. Sifting the evidence: what's wrong with significance tests. *BMJ*. 2001;322:226–231.
49. Streiner DL. Alternatives to placebo-controlled trials. *Can J Neurol Sci*. 2007;34(Suppl 1):S37–S41.
50. Student. The probable error of the mean. *Biometrika*. 1908;6:1–25.
51. Student. On testing varieties of cereal. *Biometrika*. 1923;15:271–293.
52. Tashiro Y, Miura H, Matsuda S, Okazaki K, Iwamoto Y. Minimally invasive versus standard approach in total knee arthroplasty. *Clin Orthop Relat Res*. 2007;463:144–150.
53. Tsong Y, Wang SJ, Hung HM, Cui L. Statistical issues on objective, design, and analysis of noninferiority active-controlled clinical trial. *J Biopharm Stat*. 2003;13:29–41.
54. Tuma RS. Trend toward noninferiority trials may mean more difficult interpretation of trial results. *J Natl Cancer Inst*. 2007;99:1746–1748.
55. Vavken P, Castellani L, Sculco TP. Prophylaxis of heterotopic ossification of the hip: systematic review and meta-analysis. *Clin Orthop Relat Res*. 2009;467:3283–3289.
56. Vavken P, Culen G, Dorotka R. [Clinical applicability of evidence-based orthopedics: a cross-sectional study of the quality of orthopedic evidence] [in German]. *Z Orthop Unfall*. 2008;146:21–25.
57. Vavken P, Culen G, Dorotka R. Management of confounding in controlled orthopaedic trials: a cross-sectional study. *Clin Orthop Relat Res*. 2008;466:985–989.
58. Vavken P, Dorotka R. A systematic review of conflicting meta-analyses in orthopaedic surgery. *Clin Orthop Relat Res*. 2009;467:2723–2735.
59. Vavken P, Heinrich KM, Koppellhuber C, Rois S, Dorotka R. The use of confidence intervals in reporting orthopaedic research findings. *Clin Orthop Relat Res*. 2009;467:3334–3339.
60. Vavken P, Kotz R, Dorotka R. [Minimally invasive hip replacement: a meta-analysis] [in German]. *Z Orthop Unfall*. 2007;145:152–156.
61. Wang SJ, Hung HM, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Stat Med*. 2001;20:1903–1912.