The Use of Confidence Intervals in Reporting Orthopaedic Research Findings

Patrick Vavken MD, MSc, Klemens M. Heinrich,
Christian Koppelhofer, Stefan Rois, Ronald Dorotka MD

Received: 17 October 2008 / Accepted: 16 March 2009 / Published online: 31 March 2009
© The Association of Bone and Joint Surgeons 2009

Abstract Conflict between clinical importance and statistical significance is an important problem in medical research. Although clinical importance is best described by asking for the effect size or how much, statistical significance can only suggest whether there is any difference. One way to combine statistical significance and effect sizes is to report confidence intervals. We therefore assessed the reporting of confidence intervals in the orthopaedic literature and factors influencing this frequency. In parallel, we tested the predictive value of statistical significance for effect size. In a random sample of predetermined size, we found one in five orthopaedic articles reported confidence intervals. Participation of an individual trained in research methods increased the odds of doing so fivefold. The use of confidence intervals was independent of impact factor, year of publication, and significance of outcomes. The probability of statistically significant results to predict at least a 10% between-group difference was only 69% (95% confidence interval, 55%–83%), suggesting that a high proportion of statistically significant results do not reflect

large treatment effects. Confidence intervals could help avoid such erroneous interpretation by showing the effect size explicitly.

Introduction

The classic research questions in orthopaedics are whether a group of patients improved, ie, had less pain and/or better function, after a particular treatment, or whether a treated group of patients improved compared with an untreated group. To objectively describe the meaning of treatment effects seen in such studies, statistical tests are used. These tests produce a *p* value, which (usually) is the probability that there is no difference between groups; if this probability is (arbitrarily) less than 1 in 20, or 5%, it is typical to consider these results statistically significant and thus as meaningful, and if the probability is larger than 5%, the results are considered not significant or important. However, the *p* value builds on three parameters: size of treatment effect/difference between groups, the variance of this difference, and sample size and ironically, it is not the size of the effect, but the variance and sample size, that are the dominant factors in this equation. Thus a significant result could mean anything from a truly large difference between groups, a small variance of this difference, to a large sample, or any combination of these three. As an example, a reduction in blood loss from 366 to 314 mL in standard versus minimally invasive THA, as reported in a recent study, was significant ($p < 0.001$) [27]. Another recent study regarding care of external fixation pins about the wrist reported no differences in complications among three different treatment protocols ($p = 0.20$), although there was a twofold increase in the number of complications in one group [11]. With these examples we will show

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.

This work was performed at Children's Hospital Boston and Medical University of Vienna.

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

P. Vavken, K. M. Heinrich, C. Koppelhofer, S. Rois, R. Dorotka
Department of Orthopedic Surgery, Medical University of
Vienna, Vienna, Austria

Fig. 1A–D A graph plots statistical significance versus effect size, which we used as a surrogate for clinical importance. The four fields represent (A) clinically false-positive results, (B) clinically true-positive results, (C) clinically true-negative results, and (D) clinically false negative results. Support of an outcome from (A), which is statistically significant but clinically unimportant, is not justified. Consider the example of a statistically significant reduction in blood loss from 366 to 314 mL (see text). At the same time, it would be negligent to ignore an outcome from (D), which is not statistically significant, but has clinical importance. For example, between two groups of 50 patients, there is no significant difference if one and five patients die, despite the clinical importance of a fivefold difference in mortality. However, these distinctions can be made only if measures of effect size and statistical significance are reported.

that a significant p value is not necessarily consistent with a result that would be considered clinically important. Thus, assuming that the size of a treatment effect is most relevant clinically, p values are very susceptible to clinically false positive results [7, 8, 23], and this should be acknowledged when reporting and reading research findings (Fig. 1) [7].

This problem is further aggravated by the fact that orthopaedic surgery is characterized by mostly soft outcomes such as quality of life, function, or pain, which also are context-dependent. For example, a Harris hip score of 80 means something entirely different in a young independent patient of 50 years of age and a patient of 85 years living in an assisted care facility. In light of these facts, we believe the size of a treatment effect, ie, the actual difference between groups, is more relevant to practicing orthopaedic surgeons and patients than its variance or the size of the sample in which it was observed. Confidence intervals (CIs) provide a fairly straightforward and transparent method of describing size and statistical significance [5]. Unlike p values, CIs provide pertinent information to understand the size, significance, and precision of difference, and, by extension, their clinical relevance. An $x\%$ CI is the range of values within which one has $x\%$ confident

that the true effect lies. Therefore $1-x\%$, or, eg, 5% in a 95% CI, is the likelihood of the true effect lying outside this envelope—conveniently that is the same as a p value less than 0.05. Thus, if two 95% CIs do not overlap, or if a certain value is not contained, we can be sure there is a significant difference at the 5% level [3, 18, 24, 25]. At the same time, the width of the CI gives information on precision, meaning how accurately a difference is described, with a narrow CI reflecting a very precise study and a wide CI one less so. Therefore it is not surprising that the use of CIs would be recommended over the use of p values when reporting study findings.

These observations led us to the questions regarding how research findings are presented in orthopaedic publications and whether statistical significance appropriately reflects effect size (Fig. 1). We therefore: (1) assessed the frequency of reporting CIs in a random sample of orthopaedic studies; (2) assessed factors potentially influencing the use of CIs; and (3) studied discrepancies between p values and treatment effect size by asking whether statistically significant or insignificant findings predicted large or small treatment effects, respectively.

Materials and Methods

We included articles from all journals that are listed under the category “Orthopedics” ($n = 48$) by the Institute of Scientific Information (Web of KnowledgeSM Thomson Reuters ISI, Philadelphia, PA). From what was reported in another study on methodology [26] in orthopaedic research, we assumed a prevalence of reporting of CIs of 20% to 40%. We calculated the required sample size to precisely detect this prevalence with an alpha of 5% ($n = 80$), and increased this number by 10% to account for missing information or inaccuracies, resulting in a required number of studies of 88. We used a two-stage sampling procedure using simple randomization. In the first stage, eight journals were selected from the 48 listed orthopaedic journals using a random number generator (*Arthroscopy*; *Clinical Orthopaedics and Related Research*; *European Spine Journal*; *The Journal of Bone and Joint Surgery, American Volume*; *The Journal of Hand Surgery, American Volume*; *Knee Surgery, Sports Traumatology, Arthroscopy*; *Osteoarthritis and Cartilage*; *Spine*). During the second stage, papers reporting on clinical trials, controlled trials, and randomized controlled trials published in 2000, 2003, and 2006 in the eight included journals were identified online using the digital tag terms for publication type in PubMed ($n = 595$), listed chronologically, and 88 papers were chosen for inclusion using a random number sequence.

Data on the following end points were extracted from the included studies: whether CIs were reported, the statistical

significance of study findings, year of publication, impact factor of journal, number of publications per year per journal, and participation of a methodologist. For our purpose, we defined a methodologist as an investigator with a research degree (PhD or MSc) or with an affiliation with a department or institute of (bio)statistics, epidemiology, or an equivalent. All data were obtained independently from all studies by four authors (PV, HMK, CK, SR) and crosschecked for errors. Disagreement was resolved by consensus or with the help of the senior author (RD).

To answer our first question, we calculated the percentages of studies reporting CIs. To address our second question, we assessed the influence of the statistical significance of study findings, year of publication, impact factor of journal, number of publications per year per journal, and participation of a methodologist on the use of CIs using a common, recommended approach in epidemiology [21]. Instead of testing for the significance of associations, we tested for their strength, using odds ratios (ORs) for binary exposures (statistically significant results, participation of a methodologist), Mantel-Haenszel ORs for categorical exposures (year of publication), and regression coefficients for continuous exposures (impact factor, annual number of publications in journal). All variables showing influence were included in a logistic regression model to obtain adjusted ORs for multiple exposures and to adjust for potential confounders [26]. However, there is a possibility that some journals report CIs differently than others, which would mean that reporting of CIs is not uniform across all, but clustered in some journals. Thus we extended the logistic regression model to a generalized linear mixed model, a common method to test and account for such potential clustering.

To answer our third question, we compared statistical significance with effect size. For the purpose of our study, we used effect size as a surrogate of clinical relevance, arguing that generally a large treatment effect is clinically more relevant than a small treatment effect. We used relative differences, ie, the difference between control group and intervention group in percent, as our estimate of effect size. We did not use standardized measures such as Cohen's *d* or Hedge's *g*, which use the same formulae like *p* values but report on a different scale, but aimed at reporting effect sizes based on their actual size instead of based on their statistical variance. We used different cut-off values to establish a gold standard for important and unimportant outcomes and calculated the positive predictive value of statistically significant results and the negative predictive value of statistically not significant results. The predictive value of statistical significance depends on the prevalence of large effects, which is rather intuitive as the higher the prevalence of large outcomes, the higher the chance of the *p* value to find a large effect. To adjust for

this influence of prevalences on predictive values, we calculated the 95% CI for all predictive values.

All calculations were performed using Intercooled Stata[®] 10 (StataCorp LP, College Station, TX). All results are given with 95% CIs, adjusted for potential clustering. The effect of variables is given in ORs, Mantel-Haenszel ORs, or regression coefficients of reporting CIs. All Mantel-Haenszel ORs were tested for homogeneity before pooling. ORs are considered significant if their 95% CIs do not include 1; coefficients are considered significant if their 95% CIs do not include 0.

Results

Among the 88 studies included in our analysis, 19, or 22%, (95% CI, 14%–32%) reported CIs. Eighty-seven, or 99%, (95% CI, 97%–100%) reported on statistical significance, including all studies reporting CIs. Twenty-eight percent (95% CI, 18.7%–37.8%) of the included studies did not report a parameter of variance (standard deviation, standard error of the mean) for the primary end point, thus making it impossible for the reader to judge the probable range of the effect size or to construct CIs (Table 1).

The use of CIs was independent of the statistical significance of the presented findings (OR, 1.93; 95% CI, 0.67–5.58) and the year of publication (OR, 1.01; 95% CI, 0.82–1.24). The average number of articles published per journal per year was 260 (95% CI, 223–297). Articles in journals with larger numbers of annual publications were more likely to report CIs, with a 3% (95% CI, 0.4%–6%) increase in odds of doing so per 10 additional articles per year. There also was evidence for a trend of increasing odds of reporting CIs with larger impact factors, with a 75% (95% CI, 0%–208%) increase in odds per point of impact factor. The mean impact factor was 2.2 (95% CI, 1.99–2.35). As expected, there was an increase in the odds of reporting CIs if a methodologist participated in the study

Table 1. Characteristics of included studies (n = 88)

Variable	Number	%	95% CI
Confidence intervals reported*	19	22%	13.7%–31.98%
P value reported	87	99%	96.6%–100%
Significant result	44	51%	39.6%–61.5%
Parameter of variation (SD/SEM) given	61	72%	62.2%–81.3%
Participating methodologist	45	51%	40%–62%
Average effect size (in % between-group difference)		47%	35.7%–57.5%

* These studies reported *p* values; CI = confidence interval; SD = standard deviation; SEM = standard error of the mean.

Table 2. Risk factors for reporting CIs

Risk factor	Strength of influence*	95% CI
Participation of a methodologist	7.17 (OR)	1.74–96.6
Significant result	1.93 (OR)	0.67–5.58
Year of publication	1.93 (OR)	0.82–1.24
Number of articles published per year in journal	3% (coef)	0.4%–6%
Impact factor	75% (coef)	0%–208%
Adjusted participation of a methodologist [†]	4.71 (OR)	1.02–21.69

* OR = odds ratio; coef = logistic regression coefficient; CI = confidence interval [†]from multivariate model accounting for potential clustering by journal and including participation, impact factor, and number of articles published.

(OR, 7.17; 95% CI, 1.74–96.6). Particular journals, especially those with high impact factors and greater number of articles published, were more likely to publish studies in which a methodologically trained individual participated, and at the same time, were more likely to publish papers reporting CIs, suggesting confounding [26]. After adjusting for confounding, participation of a methodologist was the only variable influencing the odds of reporting of CIs (adjusted OR, 4.71; 95% CI, 1.02–21.69). The same model showed no evidence ($p = 0.21$) for clustering of data in journals (Table 2).

Our last question aimed at comparison of significance of outcomes with effect size, which we used as a surrogate parameter of clinical importance. Studies with statistically significant outcomes had a mean difference in outcomes between study groups of 52.6% (95% CI, 39.8%–65.4%). Studies without a statistically significant result reported a mean difference between groups of 36.1% (95% CI, 15.4%–56.7%); direct comparison of between-study group differences across studies with significant outcomes and not significant outcomes revealed a mean difference of 16.5% (95% CI, –6.0%–39.0%), meaning that the effect size in a study with a statistically significant result could be anything from 39% larger to 6% smaller than the effect in a study without statistically significant results. The positive predictive value of a statistically significant outcome and the negative predictive value of a not significant finding are given in Table 3. Briefly, if a between-group difference of 10% was regarded as clinically important, the positive predictive value of a statistically significant result, ie, the likelihood that a statistically significant difference meant at least a 10% between-group difference, was 69% (95% CI, 55%–83%). However, if a difference of 20% between study groups was assumed clinically important, the positive predictive value of a statistically significant result decreased to 40% (95% CI, 26%–55%). Concurrently, the negative predictive value rose with increasing cut-off values and

Table 3. Predictive values of statistically significant results

Clinical importance cut-off value*	PPV	95% CI	NPV	95% CI
1%	100%	96%–100%	4%	0%–12%
10%	69%	55%–83%	58%	39%–78%
20%	40%	26%–55%	71%	53%–89%
30%	31%	17%–45%	92%	81%–100%
40%	15%	5%–28%	96%	88%–100%
50%	12%	2%–22%	96%	88%–100%
≥ 60%	2%	0%–7%	96%	88%–100%

* In maximal percent difference between groups; PPV = positive predictive value; NPV = negative predictive value; CI = confidence interval.

reached a value greater than 50% at a 10% between-group difference cut-off (Table 3). Our analysis shows that the actual differences between study groups in trials with significant and not significant findings were widely overlapping. Consequently, the predictive power of statistical significance is limited.

Discussion

The use of confidence intervals over p values in reporting research findings has been recommended repeatedly by experts in epidemiology and biostatistics [18, 20, 24]. The main reason for this recommendation is that CIs focus on effect size and statistical significance of results, whereas p values do not reveal all the information needed to interpret study findings [1, 8, 9, 23]. Also, as p values standardize effects for their variance and sample size, clinically not important effects might be inflated to statistical significance by large sample sizes and/or small variances, thus producing clinically false-positive results [10, 20]. A good example of this is meta-analyses, which synthesize data from numerous studies, thus increasing sample sizes. This, in turn, can lead to significant results without changing the effect size. Orthopaedic outcomes often contain rather soft data, such as quality of life or function scores, and in presenting such data, effect sizes are more relevant than sizeless statistical significance. These facts prompted us to formulate the following three purposes: first, to assess the frequency of reporting CIs in a random sample of orthopaedic studies; second, to assess the factors potentially influencing the use of CIs; and third, determine discrepancies between p values and treatment effect size by asking whether statistically significant or insignificant findings predicted large or small treatment effects, respectively.

Our study has some potential weaknesses. We used effect size as a surrogate parameter of clinical importance, arguing that large effects are more important than smaller

effects. However, clinical importance cannot be described by one value, is context-dependent, and therefore must be interpreted in a given situation. Also, it is not possible to define a cut-off value for effect size that will represent clinically important and not important outcomes. Therefore we used a series of cut-off values to describe the association between significance and effect size across a range of values that might be considered clinically relevant depending on context. We did not use standardized differences, such as Cohen's *d*, because they build on the same formula as the *t* test [10]; thus we could not use them for our purpose. We deliberately avoided such standardization of differences because we wanted to focus on the absolute values of group outcomes and their relative difference rather than on fairly abstract manifolds of their pooled variance. We used relative differences (dividing intervention by control) instead of absolute differences (subtracting controls from interventions); thus units of measurement canceled each other out and we could study outcomes on various scales without jeopardizing validity. The use of relative differences, or ratios, is well established for odds, risks, rates, hazards, etc. Finally, our study builds exclusively on reported data, and this information is not necessarily accurate or complete [12]. We included only clinical trials and controlled trials, because these usually have high methodologic quality. CIs are just as valuable in other study designs, but it seems credible that they are even more scarcely used in studies of lower levels of evidence.

Our findings suggest approximately one in five studies reports results with CIs. A frequency we regard as low, despite the fact that most of the commonly used statistical software packages provide CIs with most calculations at default settings. In conjunction with our finding that participation of a methodologist increases the odds of reporting CIs, we interpret this result as an indicator of lacking awareness of the usefulness of CIs, rather than deliberate omission. A similar influence of participation of methodologists was described in a recent study of the management of confounding in orthopaedic research [26]. That study reported a fourfold increase of the odds of controlling for confounding if a methodologist participated [26]. Furthermore, we assumed some journals, or their respective editors, would support the use of CIs more than others. For example, the *British Medical Journal* adopted a policy "expecting all authors to calculate confidence intervals whenever the data warrant this approach" in 1986 [18]. This would have led to clustering in journals; thus we divided our sample size of 88 into 11 papers from eight journals to have sufficiently large numbers to be able to see within and between journal clustering. However, we found no evidence for such an effect. We also included impact factor and number of articles published per year as potential surrogate parameters of overall quality and

activity of a journal, hypothesizing journals with higher impact factors would present more CIs, but we found no effect on the use of CIs. However, for numerous reasons, the impact factor, which is the number of times articles from the last 2 years are cited divided by all citable articles of a journal, is a poor measure of journal importance, especially in slowly moving fields or fields of low general interest [14–16]. Finally, we incorrectly presumed that in parallel with the recent progression of evidence-based medicine in orthopaedic surgery, the use of CIs would have increased since the turn of the century, but we found no evidence for an association with year of publication.

The use of less optimal devices than CIs in reporting research findings, especially the strong focus on standard errors in such methods as *p* values or standardized differences, has been suggested as a source for clinically false-positive results, ie, statistically significant results without clinical relevance [5]. Looking at the 95% CIs for between-study differences, we observed substantial overlap between statistically significant and not significant studies. A likely explanation for such a finding is that many of the statistically significant results occurred owing to small variance or larger sample sizes, rather than large effect sizes. Sample sizes are particularly important in orthopaedic research, because they vary considerably. They become even more important in the increasingly common meta-analyses because these analyses typically have far larger sample sizes than most individual orthopaedic trials so that results that were statistically insignificant in the primary trials eventually will be significant in the meta-analysis owing to inflated sample size and despite the fact that the size of effects might not change. In contrast to *p* values, CIs will not change with alterations in variance or sample size but will only narrow down and still center around the same estimate of the size of effect [6, 13, 19]. However, several authors have suggested that in orthopaedic research there are a fairly high number of type II errors, or statistically false-negative findings, owing to underpowered studies [6, 19]. CIs would help avoid the problem of false-negative findings, as the wide interval would show the low power while still presenting a probable effect size [2, 22]. A couple studies focusing on type II errors in orthopaedics have been published in recent years, and have raised awareness of this problem [4, 17].

Finally, in addition to emphasizing *p* values might not reflect effect size and, by extension, clinical relevance accurately, we wanted to try to quantify this problem. We chose to use predictive values, as they are a well-known method to characterize a test. These analyses showed, even in the best-case scenario, the positive predictive value of a significant result, when assuming a 20% between-group difference as clinically important, reached only 55%, which is hardly beyond random chance. This means differences larger than 20% cannot be predicted by significant

results in our sample, which we consider a rather disappointingly low value. The same is true for the negative predictive values of nonsignificant differences, which do not mean two groups are the same. However, what cut-off value is considered clinically important depends, among other parameters, on a study's aim and setting, personal interpretation, and on patients' and physicians' expectations and cannot be prespecified; it is not our intention to discredit the use of statistical testing based on p values, but rather to suggest its use is limited to comparing groups and should be used only with utmost caution beyond its defined purpose. Reporting of the studied effect together with its 95% CI in addition to or as a substitute for p values would show the size of the effect and its statistical significance explicitly and would allow for identification and elimination of false-positive results.

Our data suggest the frequency of CIs in reporting orthopaedic research findings is low, although it would be a useful instrument in clearly presenting research findings in context and helping to avoid clinically false-positive results, ie, findings with statistical significance but with no or only little clinical relevance or importance. We found participation of a methodologically trained individual increased the odds of reporting CIs, but we do not think this is attributable to special skills, but to higher awareness of this option among these persons. We also found focusing on p values might frequently lead to findings with statistical significance but without clinical relevance. Support of the use of CIs by authors, reviewers, and editors could help to avoid such problems. Such policies already have been adopted by some journals. Readers of any research study should be encouraged to routinely assess effect sizes and significance before drawing conclusions, especially in studies with overly large sample sizes, thus very small standard errors, such as meta-analyses.

References

1. Akobeng AK. Confidence intervals and p-values in clinical decision making. *Acta Paediatr*. 2008;97:1004–1007.
2. Altman D, Bland JM. Confidence intervals illuminate absence of evidence. *BMJ*. 2004;328:1016–1017.
3. Altman DG. Why we need confidence intervals. *World J Surg*. 2005;29:554–556.
4. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ*. 1994;309:102.
5. Altman DG, Machin D, Bryant TN, Gardner S. *Statistics with Confidence*. 2nd Ed. London, England: BMJ Books; 2000.
6. Bailey CS, Fisher CG, Dvorak MF. Type II error in the spine surgical literature. *Spine*. 2004;29:1146–1149.
7. Bhandari M, Montori VM, Schemitsch EH. The undue influence of significant p-values on the perceived importance of study results. *Acta Orthop*. 2005;76:291–295.
8. Bhardwaj SS, Camacho F, Derrow A, Fleischer AB Jr, Feldman SR. Statistical significance and clinical relevance: the importance of power in clinical trials in dermatology. *Arch Dermatol*. 2004;140:1520–1523.
9. Blume J, Peipert JF. What your statistician never told you about p-values. *J Am Assoc Gynecol Laparosc*. 2003;10:439–444.
10. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd Ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
11. Egol KA, Paksima N, Puopolo S, Klugman J, Hiebert R, Koval KJ. Treatment of external fixation pins about the wrist: a prospective, randomized trial. *J Bone Joint Surg Am*. 2006;88:349–354.
12. Emerson J, Burdick E, Hoaglin D, Mosteller F, Chalmers T. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trial*. 1990;11:339–352.
13. 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.
14. Garfield E. How can impact factors be improved? *BMJ*. 1996;313:411–413.
15. Garfield E. Impact factors, and why they won't go away. *Nature*. 2001;411:522.
16. Garfield E. The history and meaning of the journal impact factor. *JAMA*. 2006;295:90–93.
17. Kocher M, Zurakowski D. Clinical Epidemiology and biostatistics: a primer for orthopaedic surgeons. *J Bone Joint Surg Am*. 2004;86:607–620.
18. Langman MJ. Towards estimation and confidence intervals. *BMJ*. 1986;292:716.
19. 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.
20. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc*. 2007;82:591–605.
21. Rothman KJ. Writing for epidemiology. *Epidemiology*. 1998;9:333–337.
22. Sexton SA, Ferguson N, Pearce C, Ricketts DM. The misuse of 'no significant difference' in British orthopaedic literature. *Ann R Coll Surg Engl*. 2008;90:58–61.
23. Sierevelt IN, van Oldenrijk J, Poolman RW. Is statistical significance clinically important? A guide to judge the clinical relevance of study findings. *J Long Term Eff Med Implants*. 2007;17:173–179.
24. Thompson WD. On the comparison of effects. *Am J Public Health*. 1987;77:491–492.
25. Thompson WD. Statistical criteria in the interpretation of epidemiologic data. *Am J Public Health*. 1987;77:191–194.
26. 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.
27. Vavken P, Kotz R, Dorotka R. Minimally invasive hip replacement: a meta-analysis. *Z Orthop Unfall*. 2007;145:152–156.