Don't be misled by the orthopedic literature

Tips for critical appraisal

Rudolf W Poolman[1], Gino M Kerkhoffs[2], Peter A A Struijs[3] and Mohit Bhandari[4]

On behalf of the International Evidence-Based Orthopedic Surgery Working Group

[1]Department of Orthopedic Surgery, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands, [2]Department of Orthopedic Surgery and Traumatology, Kantonsspital, St. Gallen, Switzerland, [3]Department of Orthopedic Surgery, Orthopedic Research Center, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands, [4]Division of Orthopedic Surgery, McMaster University, Ontario, Canada
Correspondence RWP: Poolman@trauma.nl
Submitted 07-01-31. Accepted 07-02-01

Surgeons are constantly inundated with medical information. There are over 4,000 different journals in PubMed and over 10,000 new citations are added to the medical literature every week. Unfortunately, not all studies are equally valid and nowadays patients have free access to the medical literature, although most are not trained to judge the validity of medical information on the internet. Thus, surgeons, despite their busy schedules, have no choice but to become familiar with the principles and practice of critical appraisal to ensure well-informed decisions based upon evidence of the highest quality.

In this second article in our series, we provide tips for appraisal of the validity of a published orthopedic study. We provide guidelines to avoid being misled by the orthopedic literature, and also checklists for assessment of the reporting of a published study.

Evidence-based cycle

To perform daily patient care based on evidence-based practice requires a systematic approach. A series of steps known as the "Evidence Cycle" provides a guideline to approaching clinical problems in a systematic manner using evidence-based medicine (Sackett et al. 2000, Guyatt et al. 2002, Hayward 2007). The Evidence Cycle can be conceptualized to consist of the five A's:

1. *Assess.* The physician needs to understand the patient's problem and to determine the full context of patient characteristics, demographics, and differential diagnosis. Once the patient and his/her problem are clear, the clinician can formulate a clear research question.

2. *Ask.* Compose a clear research question for the patient's problem. The patient-oriented clinical question (PICO) is a very useful tool to achieve this (Table 1).

3. *Acquire.* The third step is to retrieve the evidence from literature databases. A search strategy with appropriate MeSH terms/subheadings is essential. Librarians can help you if the PICO is well designed; they can also help you focus the question or even develop it further to assist in finding the most appropriate articles.

4. *Appraise.* Determine whether the evidence you retrieved is useful. Assess the quality of randomized controlled trials (RCTs) using an RCT validation score, for example the CLEAR NPT (Table 2). Another option is to use the key validity questions as advocated by Guyatt et al. (Guyatt et al. 2002). Although originally designed to evaluate RCTs, we consider that parts of these guidelines are applicable to other study designs, as the guidelines evaluate possible methodological safeguards. For example, even in case series, outcome assessment could be done by an independent outcome assessor.

DOI 10.1080/17453670710013636

Table 1. Clinical patient-oriented questions

PICO clinical patient-oriented questions		
P	Patient Population Problem	Describe your patient group.
I	Intervention	Which main intervention, prognostic factor, or exposure is being considered?
C	Comparison	What is the main alternative treatment to compare to I (intervention)?
O	Outcome	What result can be expected from the intervention? What outcome is relevant to you and your patient? What type of question are you asking? What would be the best study design/methodology?

Primary guides:

- Was the assignment of patients to treatments randomized?
- Were all patients who entered the trial properly accounted for and attributed at its conclusion?
- Was the follow-up complete?
- Were patients analyzed in the groups to which they were randomized?
- Were the patients in the population analyzed comparable to your patient?

Secondary guides:

- Were the patients, health workers, and study personnel "blind" to the treatment?
- Were the groups similar at the start of the trial (in non-RCTs)?
- Aside from the experimental intervention, were the groups treated equally?

We recommend that the original articles should be read using this guide (Bhandari et al. 2001, Guyatt et al. 2002). In addition, we suggest a guide to prevent readers from being misled by articles reporting the results of intervention studies as described in detail below (Montori et al. 2004).

5. *Apply.* Apply the evidence you have retrieved to your particular patient. The evidence you found must be fit into the patient's context of complaints, characteristics, demographics, and desires. Treatment of a patient can never be based on evidence alone. The available evidence must always be put into the context of the specific clinical circumstances and the patient's values and preferences.

Patient-Oriented Clinical Questions (PICO) (Richardson et al. 1995)

To be able to apply evidence to your daily clinical practice, a useful tool is to use a so-called PICO (which stands for patient, intervention, comparison, outcome) (Table 1). Using a PICO will result in a good clinical question, and this will save you time when researching and help you to concentrate directly on your patient's requirements. A clinical situation or problem may raise more than one clinical research question. It is important that multiple topics should be fitted into an equal number of research questions. Try to keep your PICO as simple as possible. An example of a clear PICO would be the following: a 48-year-old female patient has complaints of a trochanteric bursitis. You are not sure whether to treat your patient with a corticosteroid injection or with NSAIDs.

Question: in the case of a 48-year-old woman with trochanteric bursitis (Patient), is injection with a corticosteroid injection (Intervention) as effective as NSAIDs (Comparison) for complete resolution (Outcome) of trochanteric bursitis? From this PICO, your relevant and clear research question allows you to perform a directed search of the literature databases (Richardson et al. 1995).

Key methodological safeguards in surgical trials

Optimal study design is based on methods to prevent bias. For interventions, the most optimal study design is a randomized controlled trial (RCT). However, an RCT is not always feasible in orthopedic research, and study questions may be answered with comparative case series or observational studies. In these "other" forms of evidence, methodological safeguards can still be used. Currently available checklists, like the CheckList to Evaluate A Report of a Non-Pharmacological Trial

Table 2. Checklist of items to assess quality of randomized controlled trials of non-pharmacological treatment (CLEAR NPT) (Boutron et al. 2005)

1. Was the generation of allocation sequences adequate?	Yes / No / Unclear
2. Was the treatment allocation concealed?	Yes / No / Unclear
3. Were details of the intervention administered to each group made available? [a]	Yes / No / Unclear
4. Was the experience or skill [b] of care providers for each arm appropriate? [c]	Yes / No / Unclear
5. Was participant (i.e. patient) adherence assessed quantitatively? [d]	Yes / No / Unclear
6. Were the participants adequately blinded?	Yes / No, because blinding was not feasible / No, although blinding was feasible / Unclear
6.1. If participants were not adequately blinded:	
6.1.1. Were all other treatments and care (i.e. co-interventions) the same in each randomized group?	Yes / No / Unclear
6.1.2. Were numbers of withdrawals and of individuals lost to follow-up the same in each randomized group?	Yes / No / Unclear
7. Were care providers or those caring for the participants adequately blinded?	Yes / No, because blinding was not feasible / No, although blinding was feasible / Unclear
7.1. If care providers were not adequately blinded:	
7.1.1. Were all other treatments and care (i.e. co-interventions) the same in each randomized group?	Yes / No / Unclear
7.1.2. Were numbers of withdrawals and of individuals lost to follow-up the same in each randomized group?	Yes / No / Unclear
8. Were outcome assessors adequately blinded to assess the primary outcomes?	Yes / No, because blinding was not feasible / No, although blinding was feasible / Unclear
8.1. If outcome assessors were not adequately blinded, were specific methods used to avoid ascertainment bias (systematic differences in outcome assessment)? [e]	Yes / No / Unclear
9. Was the follow-up schedule the same in each group? [f]	Yes / No / Unclear
10. Were the main outcomes analyzed according to the intention-to-treat principle?	Yes / No / Unclear

[a] The answer should be "yes" for this item if these data were either described in the report or made available for each arm (reference to a preliminary report, online addendum etc.) .

[b] The experience or skill of care providers will be assessed only for therapist-dependent interventions (i.e. interventions where the success of the treatment is directly linked to the technical skill of care providers). For other forms of treatment, this item is not relevant and should be removed from the checklist, or answered "unclear".

[c] Appropriate experience or skill should be determined according to published data, preliminary studies, guidelines, run-in period, or a group of experts and pre-specified in the protocol for each study arm before the start of the survey.

[d] Treatment adherence will be assessed only for treatments necessitating iterative interventions (e.g. physiotherapy that supposes several sessions, in contrast to a "one-shot" treatment such as surgery). For one-shot treatment, this item is not relevant and should be removed from the checklist, or answered "unclear"

[e] The answer should be "yes" for this item if (1) the main outcome is objective or "hard", or (2) if outcomes were assessed by a blinded or at least an independent endpoint review committee, or (3) outcomes were assessed by an independent outcome assessor trained to perform the measurements in a standardized manner, or (4) the outcome assessor was blinded to the study purpose and hypothesis.

[f] This item is not relevant in trials in which follow-up is part of the question. For example, this item is not relevant in the case of a trial assessing frequent versus less frequent follow-up for cancer recurrence. In these situations, this item should be removed from the checklist or answered "unclear".

(CLEAR NPT) (Table 2) are suitable for helping to identify utilization of key methodological safeguards in surgical trials (Boutron et al. 2005).

Methodological safeguards in clinical research

Was the study randomized?

The reason for randomizing study participants is to ensure that baseline characteristics are equally distributed among groups. These characteristics include prognostic factors that are both known and unknown. By randomizing study participants properly, investigators can be more confident that any effect on outcomes is attributable to the intervention under study—and known and unknown prognostic factors that are not under investigation are spread equally among the control and treatment groups (Altman and Bland 1999b, Juni et al. 2001). Studies that allocate patients to two groups, or cohorts, without randomization run the risk of having important prognostic imbalances between the study groups. The authors must rely upon matching patients for known factors in each group or on using complex statistical techniques to 'adjust' for differences between the two groups. Under both sets of circumstances, the situation is never ideal and introduces bias.

Did the authors conceal treatment allocation?

Papers reporting an adequate method of randomization may still be at risk of unequal distribution of prognostic factors between treatment groups if randomization is not concealed. Individuals responsible for randomizing patients should be blinded to the process of randomization, in order to prevent systematic influence on the randomization process (Busse and Heetveld 2006). For example, the system of odd and even days is prone to abuse: a surgeon who is more comfortable with one of two elective surgical procedures under investigation might include a patient for randomization either on Thursday or on Friday depending on the date, in order to be able to perform the preferred operation. In this way, the randomization process can be systematically influenced and bias introduced. A solution to the problem would be to employ a protocol of remote randomization. Using a remote randomization system, investigators have to visit a password-secured website or telephone number to include a patient in a study and the patient is assigned randomly to the treatment or control group. The "remote" aspect prevents "convenient" selection of treatment- or control-group patients by investigators, which can skew the purpose of randomization (Altman and Bland 1999a, Juni et al. 2001). After confirming the eligibility of a patient, a randomization center allocates the patient to experimental or control therapies. Closed envelopes can be held up to the light and may thus be readable to the investigators (Guyatt et al. 2002).

The intention-to-treat principle

Excluding from the final analysis those subjects who fail to complete treatment (dropouts), or who elect to switch to a treatment group other than the one to which they were originally assigned (referred to as contamination) upsets the balance that randomization strives to achieve (Fergusson et al. 2002). When excluding patients from the analysis, investigators may well be removing a group of patients with a worse prognosis. The remaining patients will be destined to have a better outcome, thereby inflating the true effect of the intervention under study. Analysis of participants in the groups to which they were originally assigned is referred to as the intention-to-treat principle, and is meant to preserve the value of randomization. Prognostic factors that are known and those that are not known will be, on average, distributed equally in the two groups, and the observed effect will only be that which is due to the assigned treatment. This analysis can be combined with a per-protocol analysis, in which patients are analyzed according to the treatment protocol they actually received. Exclusion of randomized patients from the primary analysis may be legitimate when study personnel have made errors in the implementation of eligibility criteria, or patients have never received the intervention. In such cases, exclusion of patients does not introduce bias and may lead to a more informative analysis if an independent, blinded adjudication committee makes this determination after evaluating all randomized patients without being informed about the final outcome of the patients (Fergusson et al. 2002).

Blinding of patients, clinicians, or outcome assessors

Another potential threat to correct interpretation of the results of a trial is the awareness of participants, clinicians, or outcome evaluators of the allocation groups. Whether done consciously or

unconsciously, everyone has the capacity to introduce bias. Patients enter studies with preconceived ideas about what they believe, and in almost all cases they would prefer to receive an active treatment rather than a placebo. Clinicians will have expectations concerning the value of certain therapies relative to placebo based on their clinical experience, and outcome evaluators can also be expected to have similar biases. It is of utmost importance to try to blind patients, treatment providers and outcome assessors as far as possible (Poolman et al. 2007). In surgical trials, blinding of treatment providers is impossible; still, patients can often be blinded if different procedures are compared using the same surgical approach (Moseley et al. 2002). Next, outcome assessors can nearly always be blinded. If blinding is still not feasible, investigators can assign independent outcome assessors to minimize bias (Devereaux et al. 2001, 2004, Boutron et al. 2004). The terms single-, double-, or triple-blinded are perceived to be confusing by most clinicians (Devereaux et al. 2001). We suggest evaluating a manuscript on reported blinding of: 1) treatment providers, 2) patients, 3) outcome assessors, and 4) data-analysts; describing who was blinded—and not using the confusing terms double-or triple-blinded. Whether blinding was really performed during the trial can only be revealed after contacting the authors, which is something the average reader will not do (Devereaux et al. 2004).

Was the follow-up sufficiently long and complete?

Adequate follow-up would be a period of time that is mandatory for a treatment effect to be measured and for a difference between two interventions or treatments to be analyzed. Ideally, the perfect follow-up period would permit definite conclusions to be drawn regarding the primary study hypothesis. For example, in one particular trial there was a trend that after 5 years of alendronate treatment there were less hip fractures to be reported in the treatment group than in the placebo control group (Black et al. 1996). The question remained whether this trend could be extrapolated to ten-year use of alendronate as well. Thus, there was a need to lengthen the period of follow-up to ten years. This extension of long-term follow-up (FLEX) in the Fracture Intervention Trial (FIT) revealed that there was a limited difference in clinical outcomes between treatment of women with post-menopausal osteoporosis with alendronate for 5 years and for 10 years (Black et al. 2006). The authors concluded that women who stopped alendronate after 5 years showed a modest increase in bone loss, and in clinical vertebral fractures, but no increase in other fractures (Black et al. 2006). They therefore suggested that for many women, stopping alendronate after 5 years does not significantly increase their fracture risk. Those at very high risk of clinical vertebral fractures may, however, benefit from continued treatment (Black et al. 2006, Colon-Emeric 2006).

Another important issue is that at the conclusion of a trial, the status of each patient with respect to the target outcome or study endpoint will be known to the investigators. Patients whose status is unknown are often referred to as being lost to follow-up. As the number of patients who are lost to follow-up increases, so does the potential compromise to the validity of a given study. This is because patients who are lost to follow-up often have different prognoses from those who are not lost; the former group may be lost because they had an adverse outcome or because they were doing well and so did not return to the clinic to be assessed (Joshi et al. 2003). Evidently, if the percentage of patients lost to follow-up is high, the study is generally considered to be inadequate. With an insufficient number of patients reporting their outcome, no valid conclusions on the outcome of the interventions can be drawn without introducing another form of bias. For long-term follow-up studies, generally a maximum of 20% of the patients being lost to follow-up is acceptable when the reasons for their absence are adequately listed. Journals such as "Evidence-Based Medicine" only include trials where there has been at least 80% follow-up (Guyatt et al. 2002).

How not to be misled by claims made in an article

Montori suggested a guide to help readers avoid being misled by articles reporting the results of intervention studies (Table 3). Although originally described to analyze pharmaceutical trial reports, many items may also apply to surgical trials—and

Table 3. Guide to avoiding being misled by biased presentation and interpretation of data [a]

1. Read only the Methods and Results sections. Bypass the Discussion section.
2. Read the Abstract reported in evidence-based secondary publications.
3. Beware of composite endpoints.
4. Beware of small treatment effects.
5. Beware of subgroup analyses.
6. Beware of limbs as opposed to patients.

[a]Adapted and modified (focusing on surgical trials) from the guide to avoiding being misled by biased presentation and interpretation of data by Montori and coworkers (Montori et al. 2004).

in fact can be used to evaluate any interventional study (Montori et al. 2004).

1. Read only the Methods and Results sections. Bypass the Discussion section

This is probably the best way not to be misled by the orthopedic literature. It is better to make your own inferences based on the methods and results presented. The authors may interpret their results differently than you would after carefully going over the Methods section. Also, you will be directly informed about the study design and you can make your own judgments about the level of evidence. One uses the key methodological safeguards (as discussed above) in these sections and the reader can judge the likelihood of bias without having to read the entire manuscript.

Furthermore, the Methods section will reveal important aspects of outcome measurement. It is important to evaluate the outcome instruments used (Poolman et al. 2007). Were these instruments validated, and were they used as the primary endpoint (Pynsent 2001)? Next, the outcome should be of importance to the patient. For many orthopedic conditions, pain is the most bothersome factor. Range of motion may be recorded easily by the investigators, but may be less important to the patient. This is why "patient-important" outcome instruments measure changes in healthcare status better than traditional outcome evaluation using physical examination.

2. Read the Abstract reported in evidence-based secondary publications

If lack of time is an issue, you may consider reading pre-appraised articles. These articles are published in "secondary" journals such as Evidence-Based Medicine, Clinical Evidence, evidence-based abstracts of Journal of Bone and Joint Surgery, and the Evidence-Based Orthopaedic Trauma section of Journal of Orthopaedic Trauma (Wright and Swiontkowski 2000, Bhandari and Sanders 2003, Poolman et al. 2006, Struijs and Kerkhoffs 2006,).

3. Beware of composite endpoints

Composite endpoints (CEPs) are outcome measures build into one single "sum of events" (Montori et al. 2005a, b). For example, reoperation can be used as a composite endpoint, being a combination of reoperations for nonunion, infection, implant failure, malunion, or compartment syndrome. As long as all the separate items of the composite endpoint are equally important to the patient, the use of that endpoint can be justified. However, if the composite endpoint is build on outcomes ranging from death to a relatively simple endpoint such as a complication, one should be careful. An example of a composite endpoint with large gradients in importance to the patient comes from the "DVT prevention literature" (Turpie et al. 2005). In a study by Turpie et al. on the prevention of venous thromboembolism in patients after total knee replacement, the primary efficacy endpoint was "the composite of the incidence of proximal and/or distal DVT (at screening or confirmed symptomatic events), confirmed non-fatal PE, and all-cause mortality during the treatment period" (Turpie et al. 2005). When a large gradient in patient-importance is present in endpoints, one should evaluate the results with the following questions:

- How much did endpoints of least importance to the patient contribute the majority of events to CEPs? In the study by Turpie et al. (2005), the majority of events were non-symptomatic DVTs,

which is perhaps an outcome of lesser importance to the patient. However, this was seen in both control and treatment groups.

- Was there a large gradient in the effect of therapy among component endpoints, particularly with the largest effects seen on least patient-important outcomes? In our example the treatment effect was similar in both groups. The results should be interpreted with caution if the patient-important outcomes are unbalanced.
- Was complete component data reported for endpoints that comprise CEPs? Turpie reported all individual endpoints. Clear reporting helps interpretation of balance of patient-important outcomes as a part of the composite outcome.

4. Beware of small treatment effects

Clinically relevant treatment effect is different from statistically significant treatment effect (Bhandari et al. 2005). We first need to know how large the treatment effect was (Bhandari and Haynes 2005).

How is a treatment effect reported?

Investigators conducting randomized clinical trials will often report the proportion of patients who experience adverse events or outcomes. Examples of these dichotomous outcomes (yes-or-no outcomes that either happen or do not happen) include presence of pain, incident of complications, and death. Patients either do or do not experience an event, and the investigators report the proportion of patients for whom such events have occurred. Consider, for example, a study in which 20% of the control group but only 5% of the treatment group developed complications. How might these results be expressed? One way would be to express them as the absolute difference (known as the absolute risk reduction, or risk difference) between the proportion who had complications in the control group (χ) and the proportion who had complications in the treatment group (γ), or $\chi - \gamma = 0.20 - 0.05 = 0.15$. Another way of expressing the effect of treatment would be as a relative risk: the risk of events among patients receiving the new treatment compared to that among controls, or $\gamma/\chi = 0.05/0.20 = 0.25$. The most commonly reported measure of dichotomous treatment effects is the complement of this relative risk, known as relative risk reduction. This measure is expressed as a percentage: $(1 - (\gamma/\chi)) \times 100 = (1 - 0.25) \times 100 = 75\%$. A relative risk reduction of 75% means that the new treatment has reduced the risk of developing complications by 75% compared to the risk among control patients; the greater the relative risk reduction, the more effective the therapy. Investigators may calculate the relative risk over a period of time, as in a survival analysis; this is called a hazard ratio.

Investigators may use other methods to describe the size of a treatment effect, such as an odds ratio. However, regardless of what measure is used, point estimates of effect can be misleading unless accompanied by a measure of precision, such as confidence intervals. The reporting of a significant p-value provides limited information compared to a confidence interval. The determination of a p-value is calculated on the assumption that the finding is due to chance—the null hypothesis. If the likelihood that the observed differences between groups are due to chance is less than what the investigators decide is acceptable, the result is reported as being statistically significant. Usually this threshold is set at 5%, or $\alpha = 0.05$, so a p-value of ≤ 0.05 is considered "significant". This threshold is somewhat arbitrary, and if investigators wish to be more confident that chance cannot explain a particular difference, they can choose a lower threshold p-value.

What is the treatment effect-size?

The true treatment effect of any given intervention could only be calculated if all individuals on Earth were randomized to receive either the intervention or a placebo. Investigators usually use the 95% confidence interval, which can be considered as defining the range that includes the true treatment effect 95% of the time. The true treatment effect will lie beyond these extremes only 5% of the time, a property of the confidence interval that is closely related to the conventional level of statistical significance of $p < 0.05$. The treatment effect can either be the improvement noted for a given outcome measure (pain, range of motion), or the decrease in an adverse event. The larger the sample size of a trial, the larger the number of outcome events and the greater the confidence that the true treatment effect is close to what we have observed.

In a study comparing open and percutaneous techniques in the surgical treatment of tennis elbow, the American Academy of Orthopaedic

Surgeons Disability of Arm, Shoulder and Hand (DASH) score was used (Dunkow et al. 2004). A DASH score of 150 signifies maximal disability and 30 points minimal disability. This score was "normalized" to a score ranging from 0 to 100. The authors presented the following results. The preoperative median basic normalized DASH score was 70 (64–75) in the open group and 70 (64–80) in the percutaneous group. The median postoperative basic normalized DASH score was 53 (48–57) in the open group and 49 (46–51) in the percutaneous group. The change in the median basic DASH score for the open group was 17 (11–19) as compared to 20 (18–26) for the percutaneous group. The authors reported that this finding was highly significant ($p = 0.0011$, Mann-Whitney U Test). One can question whether a difference of 4 points between the open and percutaneous group on a scale ranging from 0 to 100 points is clinically relevant, even though statistically significant.

Although the authors used a "patient-important" outcome instrument, they should have decided a priori on the clinical relevance of the range of the scale. Are patients with a 5-point lower score really doing better, or is true change in health status reflected in a difference of 20 points (a property of the scale called the minimal clinically important difference) on a 100-point scale? This raises an important issue, especially since sample size calculations are often performed on the best estimates of "important treatment effects". Not only is the choice of a primary endpoint important; the relevance of change on a scale is also based on the clinical expertise of the investigators and can be arbitrary. Small trials risk the chance of false negative conclusions (type II errors); thus, calculation of sample size before the start of the trial is of paramount importance (Zlowodzki et al. 2004).

In an attempt to reduce the emphasis on p-values, journals such as the Journal of Bone and Joint Surgery (Am), Acta Orthopaedica, and the Canadian Journal of Surgery now require the presentation of confidence intervals around a point estimate of effect (Bhandari et al. 2005).

5. Beware of subgroup analyses

Subgroup analysis may be the result of "data-dredging" (Bhandari et al. 2006). Study power is calculated on the primary study endpoint. Unfortunately, it can happen that the trial does not reach significance; this may reduce the chance of acceptance of the manuscript due to publication bias. Thus, authors perhaps look for significant results in secondary outcomes or subgroups after the trial has been completed. If the subgroups were not described in the protocol and are performed post hoc, this may result in false positive findings. Specifically, if multiple endpoints are evaluated this might result in positive results due to chance alone: alpha error or type I error (Bhandari et al. 2003, Zlowodzki et al. 2006). To prevent these false positive results, the investigators must perform a correction for multiple endpoints or multiple comparison like the Bonferroni method (Bland and Altman 1995). When validated outcome instruments are used, they can be reported as a single measure or they can be reported as the individual items. One should be careful about the reporting of individual items when the investigators place emphasis on the few significant subitems. These instruments are designed to be analyzed as a whole, and not as individual subgroups.

An alpha error is the erroneous conclusion that there is a difference between groups, when in fact there is none. Alpha errors can be avoided by clearly stating primary and secondary outcome parameters before conducting the trial, and adjusting the significance level of secondary outcome measures to the number of calculated secondary outcome parameters, e.g. by using the Bonferroni method: alpha = 0.05/N, where N = Number of secondary outcome parameters.

An example illustrating multiple endpoints comes from the study of Chimento et al. (2005) reporting minimally invasive hip replacement. This study had significant results and reported twelve different outcome parameters. The intraoperative blood loss ($p = 0.003$), total blood loss ($p = 0.009$), and proportion of patents with a limp at six weeks ($p = 0.04$) were significant with a p-value set at < 0.05. Using the Bonferroni method, the real significance level should be $0.05/12 = 0.004$, leaving only intraoperative blood loss as a true positive result. No sample size calculation was performed in that study and the no endpoint was clearly identified as the primary one.

Only one-third of RCTs in orthopedics journals have clearly described the primary endpoint in the

manuscript (Bhandari et al. 2003). Interestingly, this lack of reporting of primary endpoint was associated with a larger total number of endpoints. This, again, was associated with an increased risk of false positive results (type I errors) (Bhandari et al. 2003) as described above. Readers should be careful if the authors of an article do not clearly describe the primary endpoint of interest, and they should be even more careful if many endpoints are described. In such cases it is quite likely that the results obtained were due to chance alone.

6. Beware of limbs as opposed to patients

Musculoskeletal injuries or diseases often affect only one limb or joint, but at times both the left and right sides are affected. This "limb-specific" feature in orthopedics can complicate interpretation of research findings, particularly if multiple observations from single individuals have been used inappropriately. As surgeons, we deal with patients, not with limbs. Unfortunately, however, scientific articles dealing with orthopedics regularly report limbs, not patients (Bryant et al. 2006). Articles reporting numbers of limbs and using limbs—not patients—for statistical calculations should be interpreted with caution. A patient with two joints affected may score differently on a patient-reported outcome instrument than a patient with only one joint affected. If the first patient is used twice in the calculation, this could clearly skew the results. Thus, investigators should plan ways of incorporating patients who have bilateral involvement into trials before the start of the trial. Bryant et al. (2006) suggest that: "A biostatistician can help to formulate a plan of analysis and to determine the degree of within-patient correlation so that the estimate of variability within each group is closer to the true underlying variance and not an overestimation or underestimation that has been influenced by the direction and magnitude of the association between limbs or joints in individual patients. Other options include excluding the second limb or joint from the study, randomly choosing which limb or joint to include in the analysis, or analyzing bilateral patients as a distinct subgroup". If these precautions are not met and the article only reports limbs, the results should be regarded with caution.

Conclusion

Irrespective of the study design, investigators can use methodological safeguards to prevent bias. The guidelines given here may help readers to avoid being misled by articles reporting the results of intervention studies.

We are grateful to Jason W Busse for his useful tips for critically appraising composite endpoints.

Altman D G, Bland J M. How to randomise. BMJ 1999a; 319: 703-4.

Altman D G, Bland J M. Statistics notes. Treatment allocation in controlled trials: why randomise? BMJ 1999b; 318: 1209.

Bhandari M, Haynes R B. How to appraise the effectiveness of treatment. World J Surg 2005; 29: 570-5.

Bhandari M, Sanders R W. Wher's the evidence? Evidence-based orthopaedic trauma: a new section in the Journal. J Orthop Trauma 2003; 17: 87.

Bhandari M, Guyatt G H, Swiontkowski M F. User's guide to the orthopaedic literature: how to use an article about a surgical therapy. J Bone Joint Surg (Am) 2001; 83: 916-26.

Bhandari M, Whang W, Kuo J C, Devereaux P J, Sprague S, Tornetta P, III. The risk of false-positive results in orthopaedic surgical trials. Clin Orthop 2003; (413): 63-9.

Bhandari M, Montori V M, Schemitsch E H. The undue influence of significant p-values on the perceived importance of study results. Acta Orthop 2005; 76: 291-5.

Bhandari M, Devereaux P J, Li P, Mah D, Lim K, Schunemann H J, Tornetta P, III. Misuse of baseline comparison tests and subgroup analyses in surgical trials. Clin Orthop 2006; (447): 247-51.

Black D M, Cummings S R, Karpf D B, Cauley J A, Thompson D E, Nevitt M C, Bauer D C, Genant H K, Haskell W L, Marcus R, Ott S M, Torner J C, Quandt S A, Reiss T F, Ensrud K E. Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Fracture Intervention Trial Research Group. Lancet 1996; 348: 1535-41.

Black D M, Schwartz A V, Ensrud K E, Cauley J A, Levis S, Quandt S A, Satterfield S, Wallace R B, Bauer D C, Palermo L, Wehren L E, Lombardi A, Santora A C, Cummings S R. Effects of continuing or stopping alendronate after 5 years of treatment: the Fracture Intervention Trial Long-term Extension (FLEX): a randomized trial. JAMA 2006; 296: 2927-38.

Bland J M, Altman D G. Statistics notes: Multiple significance tests: the Bonferroni method. BMJ 1995; 310: 170.

Boutron I, Tubach F, Giraudeau B, Ravaud P. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. J Clin Epidemiol 2004; 57: 543-50.

Boutron I, Moher D, Tugwell P, Giraudeau B, Poiraudeau S, Nizard R, Ravaud P. A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. J Clin Epidemiol 2005; 58: 1233-40.

Bryant D, Havey T C, Roberts R, Guyatt G. How many patients? How many limbs? Analysis of patients or limbs in the orthopaedic literature: A systematic review. J Bone Joint Surg (Am) 2006; 88: 41-5.

Busse J W, Heetveld M J. Critical appraisal of the orthopaedic literature: therapeutic and economic analysis. Injury 2006; 37: 312-20.

Chimento G F, Pavone V, Sharrock N, Kahn B, Cahill J, Sculco T P. Minimally invasive total hip arthroplasty: a prospective randomized study. J Arthroplasty 2005; 20: 139-44.

Colon-Emeric C S. Ten vs five years of bisphosphonate treatment for postmenopausal osteoporosis: enough of a good thing. JAMA 2006; 296: 2968-9.

Devereaux P J, Manns B J, Ghali W A, Quan H, Lacchetti C, Montori V M, Bhandari M, Guyatt G H. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001; 285: 2000-3.

Devereaux P J, Choi P T, El Dika S, Bhandari M, Montori V M, Schunemann H J, Garg A X, Busse J W, Heels-Ansdell D, Ghali W A, Manns B J, Guyatt G H. An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. J Clin Epidemiol 2004; 57: 1232-6.

Dunkow P D, Jatti M, Muddu B N. A comparison of open and percutaneous techniques in the surgical treatment of tennis elbow. J Bone Joint Surg (Br) 2004; 86: 701-4.

Fergusson D, Aaron S D, Guyatt G, Hebert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 2002; 325: 652-4.

Guyatt G H, Rennie D, The Evidence-Based Medicine Working Group. Users' guides to the medical literature. A manual for evidence-based clinical practice. AMA press, Chicago 2002.

Hayward R. Centre for Health Evidence. http://www.cche net/2007.

Joshi A B, Gill G S, Smith P L. Outcome in patients lost to follow-up. J Arthroplasty 2003; 18: 149-53.

Juni P, Altman D G, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001; 323: 42-6.

Montori V M, Jaeschke R, Schunemann H J, Bhandari M, Brozek J L, Devereaux P J, Guyatt G H. Users' guide to detecting misleading claims in clinical research reports. BMJ 2004; 329: 1093-6.

Montori V M, Busse J W, Permanyer-Miralda G, Ferreira I, Guyatt G H. How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: should I dump this lump? ACP J Club 2005a; 143: A8.

Montori V M, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse J W, Pacheco-Huergo V, Bryant D, Alonso J, Akl E A, Domingo-Salvany A, Mills E, Wu P, Schunemann H J, Jaeschke R, Guyatt G H. Validity of composite end points in clinical trials. BMJ 2005b; 330: 594-6.

Moseley J B, O'Malley K, Petersen N J, Menke T J, Brody B A, Kuykendall D H, Hollingsworth J C, Ashton C M, Wray N P. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med 2002; 347: 81-8.

Poolman R W, Kocher M S, Bhandari M. Pediatric femoral fractures: A systematic review of 2422 cases. J Orthop Trauma 2006; 20: 648-54.

Poolman R W, Struijs P A, Krips R, Sierevelt I N, Marti R K, Farrokhyar F, Bhandari M. Reporting of outcomes in orthopaedic randomized trials: does blinding of outcome assessors martter? J Bone Joint Surg (Am) 2007; 89 (3): 550-8.

Pynsent P B. Choosing an outcome measure. J Bone Joint Surg (Br) 2001; 83: 792-4.

Richardson W S, Wilson M C, Nishikawa J, Hayward R S. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995; 123: A12-A13.

Sackett D L, Straus S E, Richardson W S, Rosenberg W, Haynes R B. Evidence-based medicine. Churchill Livingstone, 2000.

Struijs P, Kerkhoffs G. Ankle sprain. Clin Evid 2006; 15: 1493-501.

Turpie A G, Fisher W D, Bauer K A, Kwong L M, Irwin M W, Kalebo P, Misselwitz F, Gent M. BAY 59-7939: an oral, direct factor Xa inhibitor for the prevention of venous thromboembolism in patients after total knee replacement. A phase II dose-ranging study. J Thromb Haemost 2005; 3: 2479-86.

Wright J G, Swiontkowski M F. Introducing a new journal section: Evidence-based orthopaedics. J Bone Joint Surg (Am) 2000; 82: 759.

Zlowodzki M M D, Bhandari M M D, Brown G A M, Cole P A M, Swiontkowski M F M. Planning a randomized trial: Determining the study sample size. Tech Orthop 2004; 19: 72-6.

Zlowodzki M, Jonsson A, Bhandari M. Common pitfalls in the conduct of clinical research. Med Princ Pract 2006; 15: 1-8.