Users' guide to the surgical literature

How to assess a randomized controlled trial in surgery

Achilleas Thoma, MD, MSc; Forough Farrokhyar, PhD; Mohit Bhandari, MD, MSc; Ved Tandan, MD, MSc; for the Evidence-Based Surgery Working Group

Clinicians are presented with problems that require evidence. Randomized controlled trials (RCTs) are considered the optimal study design for evaluating the effect of a new medical or surgical intervention. Surgical RCTs are carried out less often than RCTs of medical interventions, partly because using RCTs to evaluate surgical or interventional procedures is difficult and demands special consideration of issues such as blinding and the effects of surgeon factor, learning curve and differences in pre- or postprocedural care. Despite these obstacles, RCTs are becoming more common as surgeons in many specialties become acquainted with the proper methodology for health research.

The mere description of a study as "randomized" does not allow clinicians to infer validity — to base inferences on the accuracy of the results drawn from the sample of patients studied (internal validity) or the generalizability of these results to other settings (external validity). The purpose of this article is to review, within a framework of a critical appraisal of surgical RCTs, strategies to interpret the results and the issues to consider when applying these results to a clinical practice setting. A clinical scenario will be used to illustrate how surgeons might retrieve and evaluate evidence.

Scenario

A 67-year-old accountant who you know has osteoarthritis in her hips comes to your office. In the past year, the pain in her right hip has increased dramatically despite the conservative treatment you have prescribed. Pain now prevents her from walking for more than 5 minutes, and enjoying recreational activities such as golf.

After examination and a review of her x-rays you conclude that she may benefit from total replacement of her right hip. You discuss with her the nature of her problem, and the expectations and risks of the surgery. You recommend a "cementless" implant recently advocated at an orthopedic conference you attended.

Your patient recalls 2 neighbours who were exceptionally satisfied with their hip replacements. In their cases, however, cement was used to bond the metal hip to the bone. "Whatever way they did it," she tells you, "I want it done the same way."

You ask her to go home, think about what you told her, and return in a few weeks for another discussion. Realizing that you remain uncertain about the latest evidence on outcomes favouring cemented versus uncemented arthroplasty, you decide to search the literature before your patient's next appointment.

The literature search

The ideal article addressing the question would be one comparing cases of cemented hip arthroplasty with those not cemented. From your home computer you enter Medline, the National Library of Medicine's PubMed database. In the search field you enter "hip AND cement AND randomized" with limits to the last 3 years, English articles, studies using human subjects, and clinical trials.

This search yields 10 hits. As you review the titles of the articles found, "Comparison of total hip arthroplasty performed with and without cement," a 6-page article by Laupacis and colleagues published in 2002,[1] attracts your attention. This recent article appears to address your particular question, and you expect it to give you the answer.

Evaluating the search

This search strategy is unfortunately insufficient to identify all the RCTs and other sources that could give you a more complete review of topics. To learn how to plan an appropriate search, refer to the first article of this series, by Birch and associates.[2]

Although various study designs are used in clinical research, RCTs are generally regarded as the most scientifically rigorous for evaluating the effect of a surgical intervention and protecting against selection bias.[3,4] Results from observational studies may generally be valid, but their limitations are acknowledged.

The fundamental criticism of observational studies is that unrecognized confounding factors may distort the results. Although RCTs will remain a prominent tool in clinical research, the results of a single RCT, like those of a single observational study, should be cautiously interpreted. If an RCT is later determined to have given wrong answers, evidence from other trials and from well-designed cohort or case–control studies can and should be used to

find the right answers. The commonly held belief that only RCTs produce trustworthy results and that observational studies are inherently misleading does disservice to patient care, clinical investigation and the education of health care professionals.[5]

Summary of the appraised article

The Laupacis group's study was designed to compare the fixation of a Mallory–Head total hip prosthesis with and without cement. Recruitment was slow; the intended sample size of 300 patients, which had been planned in order to attain a study power of 80%, was reduced to 250.[1] Of these 250 patients, 124 underwent hip arthroplasties with cement and 126 without cement. The outcomes assessed included mortality, revision arthroplasty, health-related quality of life (HRQoL, measured with 5 different scales) and a 6-minute walk test.

The mean time of follow-up was 6.3 years (minimum 0, maximum 9.4 yr). Complete follow-up data were unavailable for 36 patients (14.4%): 31 refused to continue with the HRQoL follow-up examinations, 4 were lost to follow-up and 1 was mistakenly not followed. Overall 35 of the patients studied died, at a mean postoperative period of 4.2 years; but it is unclear how many were in either group.

More revisions were required in patients who had a prosthesis with cement (10.5%) than those without (4.5%), but the difference was not statistically significant. However, patients in the group with cemented implants had femoral revisions significantly more frequently (9.7%) than those with cementless implants (0.8%; $p = 0.002$). (A femoral revision deals with the stem of the implant that goes inside the length of the bone.)

A comparison of pre- and postoperative HRQoL scores revealed substantial improvements across a wide array of dimensions of health status and HRQoL in both groups,

particularly during the first 3 months after surgery. During the 7 years of follow-up, HRQoL was maintained with minor decline between the first and seventh years. Mean time trade-off scores, in total, improved from 0.29 before to 0.8 after operation, but p values were not provided.

Our article has the same structure as previous items in the evidence-based surgery series.[2,6–9] The purpose of ours is to help readers appraise the strengths and weaknesses of RCTs in the surgical literature. This article will apply the principles in the Users' Guides to the critical appraisal of a surgical RCT, as in Box 1.

Are the results of this RCT valid?

Did the investigators take the learning curve into consideration?

In contrast to drug trials, the surgical RCT, which compares a "novel" to the "usual" intervention, has to deal with the learning curve. The learning curve usually refers to the accumulated experience on a new procedure that allows continuous refinement of patient selection, operative technique, adjunctive medication and postprocedural care. Although some investigators suggested that surgical RCTs should start with the first patient,[11,12] most investigators disagree. Most believe that it is inappropriate to compare a familiar with an unfamiliar surgical intervention, as mistakes and adverse outcomes are more likely to occur with the unaccustomed procedure, which will bias the results against the novel intervention.[13–15]

The report by Laupacis and co-workers[1] has no indication that a learning curve was considered. We do not know if the surgeons took a course on the cementless prosthesis or did a few cases to feel comfortable before embarking on the RCT. Considering that the 2 participating surgeons are experienced orthopedic specialists, let us assume that the critical part of the procedure was performed by them many times before, and the learning

curve was therefore not an issue. Any misleading data resulting from this problem would bias against the non-cemented prosthesis; thus, if anything, the estimate of benefit of a non-cemented prosthesis is conservative. Learning curves can be a major factor when the interventions being compared are quite different in technique, such as open versus laparoscopic cholecystectomy.

Were patients randomized?

Randomization tends to produce comparable groups; that is, known and unknown prognostic factors in participating patients will, with randomization, become evenly balanced between the study groups.[16] Because a patient's, physician's or researcher's preferences may bias allocation, randomization removes that potential.

Since bias in the selection of cases can influence outcomes, it is imperative that experimental and control groups be randomized. The authors of the study under scrutiny tell us that their means of randomization was computer-generated, a method that is appropriate.

Was randomization concealed (blinded or masked)?

It is important that those deciding whether a given patient is eligible for the trial remain unaware of the arm of the study to which the patient will be allocated. If randomization or concealment fails, they may systematically enrol sicker or less sick patients to either treatment group, leading to biased results.[4,17] The randomization method is faulty if investigators assign patients according to even or odd birth year, alternate chart numbers, and so on, as these strategies are prone to selection bias.[18] For example, if even/odd birth year is used, the randomization method is unconcealed and investigators, consciously or subconsciously, may select patients based on knowledge of the group to which they are being allocated. In the Laupacis group's article,[1] opaque envelopes provided the sequence of the procedure, which were opened in the operating room. Provided no one tampered with these envelopes, we can be satisfied to this point that in fact concealment in this study was achieved.

A sealed, opaque envelope cannot be read without spoiling the seal. When such envelopes are used, they should be sequentially numbered to circumvent repeated opening. As the authors did not mention any means that was used to prevent premature or repeated opening of their opaque envelopes, we must place a question mark beside concealment as applied in this study.

Central randomization (offsite randomization by an uninvolved party) is the most effective way of ensuring concealment in surgical trials.

In contrast to pharmaceutical studies where concealment and treatment masking can be continued to the end of the study, in surgical studies concealment can only be maintained until the surgeon opens the envelope. From that time forward, the surgeon knows what procedure the patient will undergo.

Were patients stratified?

Some investigators might want to make sure, prior to randomization, that any prognostic factors known to

---

Box 1. Users' Guides for an article about RCT in surgery

Are the results valid?

Did the 2 surgical groups begin the study with a similar prognosis?

- Did the investigators take into consideration the learning curve?
- Were the patients randomized?
- Was randomization concealed (blinded and masked)?
- Were patients stratified?
- Were patients' data analyzed in the surgical groups to which they were randomized?
- Were patients in the 2 surgical groups similar with respect to known prognostic factors?

Did the surgical groups retain a similar prognosis after the study started?

- Were patients aware of group allocation?
- Were surgeons aware of group allocation?
- Were outcome assessors aware of the surgical group allocation?
- Was follow-up complete?

What are the results?

- How large was the treatment effect?
- How precise was the estimate of the treatment effect?

How can I apply the results to patient care?

- Were the study patients similar to my patients?
- Were all clinically important outcomes considered?
- Are the likely treatment benefits worth the potential harm and costs?

This list was modifed from *Users' guides to the medical literature: a manual for evidence-based clinical practice,* by G. Guyatt and D. Rennie (Chicago: AMA Press; 2002).

---

be strongly associated with outcome events are evenly divided between the study groups. This randomization method is called stratified randomization. Laupacis and associates[1] applied this method by stratifying patients by surgeon (initials RB or CR) and age (under and over 60 years) and by selecting a computer-generated randomization system.

Moreso than in medical studies, individual skills and technique may have a major effect on surgical outcomes. In surgical RCTs, stratification by surgeon should therefore be considered whenever possible.

Laupacis and coauthors[1] inform us that they used a stratified randomization method. Unfortunately, they did not report whether age or surgeon factor accounted for any of the observed differences in outcomes (patient survival or need for revision of the prosthesis).

Were patients analyzed in the group to which they were randomized?

Surgical investigators can corrupt randomization if not all patients receive their assigned surgery. For example, if the anesthetist, after induction of anesthesia, voices concern about a patient's arrhythmia or fears a myocardial infarction, the surgeon might decide to carry out whichever procedure would take less time, even if it is not the one randomized to be performed. Another scenario could involve a decision by the surgeon to use alternate procedure because of unexpected findings during the surgery, e.g., poor quality of bone.

Data for such patients might then be dropped from the analysis. If surgeons include such poorly destined patients in one treatment group and drop them from the other, then even a suboptimal surgical procedure may appear to be effective.

Intention-to-treat analysis avoids this potential bias.[10,19] In essence, the analysis of outcomes is based on the treatment arm to which patients were randomized and not on the surgical treatment received. With this analysis method, all known and unknown prognostic factors remain distributed between the surgical groups as equally as they were at randomization.

Were patients in the treatment and control groups similar with respect to known prognostic factors?

Since bias in case selection can influence the effect of the intervention, it is imperative that the experimental and control (more traditional treatment) groups begin their treatment with same prognosis, i.e., stage of disease, underlying morbid condition and age. In the Laupacis study,[1] the authors informed us that they included patients with osteoarthritis (OA) of the hip who were undergoing a unilateral procedure. They excluded patients who were 75 years of age or older. Note that the more elderly patients were likely excluded because uncemented designs can fail in the presence of poor bone quality and unsuitable canal morphology.

In the article itself, we are not given the information necessary to judge for ourselves if in fact the groups are similar. However, an address was provided for a Web page where this information can be found. Checking this page, we found a list of factors provided that included the ratio of women to men, percentage of patients operated on by one of the authors, and preoperative scores on the HRQoL scales used in the study. The distribution between the 2 treatment groups of all the prognostic factors in Box 1 can therefore be determined. Although no statistical analyses were presented, no significant clinical difference was apparent.

It would have been beneficial to know whether patients had the same level of disease; we only know that "severe OA" patients were omitted. Radiographic staging does not always correlate with clinical limitations; it may not have been included for that reason.

Were patients aware of their group allocation?

It has been recognized in the medical literature for some time that if patients know the treatment and believe it to be efficacious, they tend to feel better than those who do not, even when the treatment is the same. This has been dubbed the placebo effect.[20] Placebo effect, which is usually but not necessarily beneficial, is attributable to an expectation that the regimen will have an effect — the power of suggestion.[21] Knowing (or even thinking that they know) that they are in the experimental group could influence patients to answer QoL questions in a favourable manner, exaggerating the response in the surgical arm receiving (in this case) the newer cementless prosthesis.

In this particular study, the authors made it clear that the patient was unaware of the type of prosthesis he or she received. Furthermore, the eventual scar would be identical, as "all patients were operated on by either surgeon with use of an identical direct lateral approach…." However, in surgical trials where the question of interest compares surgical and nonsurgical interventions, this is frequently impossible.

Were clinicians (surgeons) aware of group allocation?

Differences in patient care other than the surgical intervention under consideration can bias the results. For example, if the surgeon was biased toward the cementless prosthesis, he or she might take a little extra time and pay a bit more attention to details such as hemostasis or placement of the prosthesis, producing an overestimate of its good effects.

Even use of correct randomization does not guarantee that 2 surgical groups will remain prognostically balanced. In contrast with pharmaceutical trials where the clinicians do not know whether the patient takes the experimental drug or the pla-

cebo, in surgery the procedure can not be concealed from the surgeon executing it. Despite the attention[22,23] paid to this problem, it unfortunately has no easy solution.

In the article under study, because no effective blinding of the surgeon was possible, we just don't know how much differential surgical care has been applied, consciously or unconsciously, to the cementless implant group.

Were outcome assessors aware of group allocation?

Even when randomization has successfully controlled for selection bias and surgical groups have been kept prognostically balanced, the study can still introduce bias if the assessors of outcomes have not been blinded. We define as assessors those who collect outcome information. If either surgical arm of the study receives more frequent or thorough measurement of outcome or co-intervention (additional physiotherapy, for example), positive interpretation of marginal findings, or a different degree of encouragement during performance tests, then results can be distorted in favour of the group receiving the additional attention.[24]

The authors of the appraised RCT inform us that neither the patient nor the research assistant who assessed the outcomes was aware of the type of prosthesis that had been inserted. Whether this materialized or not, we just don't know. Physiotherapists doing assessments in hospitals do have access to patients' charts. Conscious or unconscious perusal of the operative record could have revealed the type of implant used. We don't know to what extent the investigators concealed this information on patients' charts.

Was follow-up complete?

A major threat to the validity of an RCT is failure to account for all the patients at the end of the study. The greater the number of patients lost in this way, the greater the harm done to the study. Patients who did not return for a follow-up appointment or test may have died or had a bad outcome. It is also possible they may have had a very satisfactory outcome and didn't bother returning for follow-up. When the loss rate is large (greater than 10% or so of the study patients), it affects the study's validity. There are specific guidelines on this subject.[10,25] A large lost-to-follow-up rate reduces the study power.

The occurrence of the event of interest among patients is uncertain after a specified time when follow-up data collection ends. It is unknown when or whether the event of interest occurred subsequently. Such study subjects are described as "censored" or "lost to follow-up".[26,27] These patients still contribute to the study up to the time at which their outcome status was last known. One way to examine the data for censoring effect requires the use of appropriate analytic methods, such as survival analysis (application of life-table method to data).[25,28] In this study, the authors have not described their statistical methods; but when the results are reviewed, it is clear that they used survival analysis.

The authors had planned originally to recruit 300 patients, to provide 80% study power. Study power is an issue decided early in the planning phase; the calculated sample size is the minimum required to maintain that power throughout the study. With 250 patients randomized into the appraised RCT, this study likely became underpowered. Furthermore, complete follow-up data were unavailable for 36 patients: 31 who refused to continue with all of the HR QoL follow-up examinations, 4 who were lost to follow-up and 1 who was mistakenly not followed.

What are the results?

How large was the treatment effect?

The authors reported that during the follow-up of the study, revision sur-

gery occurred more often with the cemented prostheses than with those without cement (13 patients v. 6, $p = 0.11$). They inform us that this difference was not statistically significant; however, cemented implants had significantly more revisions of the femoral component of the prosthesis (12 v. 1, $p = 0.002$). Unfortunately, the results were presented in a way not easily interpreted by practising surgeons.

From the data the authors provided, we were able to calculate the absolute risk reduction (ARR), relative risk (RR), relative risk reduction (RRR) and numbers needed to treat (NNT), all shown in Table 1 and Table 2. (Explanations of these terms and calculations can be found in the article by Urschel and associates[7] in this series.) Notice the subset of values related to the femoral component of the prosthesis.

The NNT is very useful to the clinician. It tells us that every 20 cementless prostheses installed by

Table 1

Number (and %) of hip replacements needing revision surgery

	Cemented implants	Cementless
Patients randomized	124	126
Revisions required	13 (10.5)	6 (4.8)
Femoral only	12 (9.7)	1 (0.8)

Table 2

Calculations from study data

	Revisions	
Factor & equation	All	Femoral
Absolute risk reduction	0.05	0.1

ARR = x – y
  e.g., for all subjects:
    x = 13/124 = 0.10
    y = 6/126 = 0.05
  x – y = 0.10 – 0.05
      = 0.05		

Relative risk	0.5	0.08
RR = y/x		
Relative risk reduction	50%	92%
RRR = (1 – y/x) × 100		
No. needed to treat	20	10
NNT = 1/ARR		

surgeons prevents 1 revision surgery that would be needed if cemented prostheses had been used. Only 10 cementless prostheses need to be performed to prevent 1 revision of a femoral component.

With regard to the various HR QoL scales that we believe are very relevant to surgical studies, the authors made a cursory mention in the results section of their article that "all outcomes measures had substantially improved by 3 months after surgery." HRQoL, the article goes on to report, "was maintained during the 7 years of follow-up, although it appeared to decrease slightly between the first and seventh years." This statement is insufficient. It can be explained by a failure of the investigators to power the study adequately for HRQoL scales.

How precise was the estimate of the treatment effect?

The true reduction to the revision rate caused by use of a cementless prosthesis can never be known. The best estimate we have is that provided by this RCT, called the point estimate.

The true reduction value lies somewhere in this neighbourhood, and it is unlikely to be precisely correct. The best method to communicate this point estimate is to provide the confidence interval (CI). A CI is a set of values within which one can be confident that the true value lies.[10,29] Although the breadth of the CI chosen is somewhat arbitrary, by convention a 95% interval is used. A 95% CI defines the set within which we expect to find the true effect estimate 95% of the time. It becomes narrower as sample sizes increase.

CIs are important for interpreting studies with negative results as well as positive. In a positive RCT where the authors claim that the novel treatment is effective, all we have to do is to look at the lower boundary of the CI. If we believe that this lower-bound effect change is clinically im-

portant, then the indication to adopt this new surgical intervention is strong. If on the other hand we do not consider this value clinically important, despite the fact that the *p* value was significant, we may want further evidence before we adopt it.

Looking at the upper CI boundary can help us respond to negative studies, i.e. those where the *p* value is too large to be significant. When the upper value, if true, shows a result that would be clinically important, the study does not exclude an important treatment effect.[10] The authors of this RCT did not provide a CI value of the revision rate, either for the whole implant or for the femoral component. They did, however, provide a *p* value for both outcomes. From the *p* value we can calculate CI ourselves.[30]

The formula for calculating the 95% CI for a proportion (p) is CI = $p \pm 1.96 \times SE$, where SE is the standard error. The formula for SE of a proportion is $SE = [p \times (1-p)/n]^{1/2}$ where p is the proportion and *n* is the number of patients.

We have calculated 95% CIs for the rates provided in the article. The revision rate in the cemented group was 10.5%, with a CI of 5.1–15.9. The revision rate in the cementless group was 5%, with a CI of 1.2–8.8. The femoral revision rate in the cemented group was 10%, with a CI of 4.7–15.3. The femoral revision rate in the cementless group was 0.8%, with a CI of 0.04–3.95.

With regard to the various HR QoL scales, the authors did not provide an estimate of the change in the score or its CI. They did provide, on a Web page, the raw data from which we can calculate the pre- to postoperative score change.

How can I apply the results to patient care?

Were the study patients similar to the patient(s) in my practice?

Before we accept the RCT's findings

and apply them to our scenario patient, we need to decide whether she is similar to the patients who participated in the RCT.

The patients in any RCT may be quite different from our own because the inclusion/exclusion criteria used in a trial are usually stringent. For example, the RCT may have included male patients only, whereas our patient is female; our patient may have comorbid conditions not present in RCT participants, and so on. One must also consider that RCT report averaged the treatment effects in their study patients, when some individuals undoubtedly benefited more than others, and some may not have benefited at all.

If we look at the inclusion and exclusion criteria set by the investigators of the study scrutinized in our article, we find that indeed, our patient meets their inclusion criteria. We can therefore be confident that their results apply to our patient as well. Although the average age of their study participants was younger (mean 64 yr), upon closer examination of the article we see that a subgroup analysis of the results was performed by age, for those 60 years or younger and those 60–75 years old. The RCT findings still hold for the older subgroup. Our patient is similar in age to the patients in the study, and hence we are willing to apply the RCT's findings to her.

Were all clinically important outcomes considered?

Surgical interventions are indicated only if they produce clinically important changes to the patient. A small physiological change in some variable is not a justification to submit the patient to surgery. For example, a positive change of 10° in the range of motion (RoM) of the knee or hip with a prosthesis is not justification for surgery, unless this change leads to a substantial change in the patient's quality of life.

In their article, Laupacis and co-

authors mentioned that they assessed 3 different outcomes.[1] First, they used mortality as an important outcome. Unfortunately, they do not state the death rate in each group, but only that 35 of all their randomized patients died.

The second outcome they studied was revisional surgery of the hip. As hip prostheses are associated with complications downstream, this is an important outcome to consider—and indeed, it was. Revision of the implant submits a patient to additional surgery with well-recognized potential for perioperative morbidity. Revisional surgeries also add expenses to the health care system, which matter from a societal perspective. The investigators unfortunately did not "piggyback" an economic analysis onto their RCT report.

The third outcome was the assessment of HRQoL. With 5 different scales they demonstrated that scores improved substantially after implant. Although it was left unclear whether the cementless prosthesis outdid the other in this respect, both groups' rise in HRQoL scores appears to have been similar during the study period.

Are the likely treatment benefits worth the potential harm and costs?

Before we decide to use the study's findings to persuade our patient, it is important to consider whether the probable benefits that we expect the patient to derive are worth the effort. Any surgery carries risks, which can include anesthetic complications, surgical wound infection, pneumonia and deep-vein thrombosis.

There is evidence in the literature going back to the 1960s that introduction of a hip prosthesis improves HRQoL in patients afflicted with OA. The RCT we have been examining suggests that in patients younger than 75 years of age without severe OA, infectious arthritis or previous hip/knee arthroplasty, cementless arthroplasty results less often in revi-

sional surgery while exhibiting equivalent risk of mortality and improvements in HRQoL.

Trading off benefit and risk requires that we examine carefully the risks of surgery, not only the initial one but also any revisions downstream. If we believe the findings of the study that the cementless implant is less likely to require later revision, we may want to consider using this rather than the cemented one.

Costs also need to be considered. Our universal health care system covers the surgery costs; as well, there are indirect costs borne by the patient and her family to think of. If we are convinced that the cementless prosthesis is less likely to require revision in the future, there may be compelling reason to advocate this novel intervention.

Resolution of the scenario

After going carefully over the evidence presented in the published manuscript and the Web page, we are convinced that our patient will benefit from either prosthesis in terms of pain control, ability to walk and general improvement in her quality of life. If she chooses the cementless prosthesis, she is more likely to avert revision of her prosthesis downstream. Although you recognize that the results of this study may not be generalizable to all cemented and uncemented prostheses, the one used in your institution is the same as that used by Laupacis and colleagues.[1] Therefore, you recommend the cementless prosthesis.

Discussion

Although this article discussed the appraisal of a single RCT that addresses a single clinical question, the reader should be aware that to grasp the clinical question fully, one should also review other pertinent literature that may include additional RCTs and systematic reviews.

Before accepting RCT findings, it

is important to comprehend the important elements of the study, in particular its design and conduct, and the analyses and interpretation of the results. Although an RCT is considered the best study design to provide an estimate of the effect of a surgical intervention, a study that is methodologically poorly executed may overestimate the true effect. The RCT report must convey to the reader why the study was undertaken, how it was conducted and how it was analyzed; and it is necessary that these steps be transparent to the reader.

There is evidence in the literature that reporting randomization inadequately can lead to bias in the effect of interventions.[4,17] Because of problems associated with the reporting of RCTs, an international group of clinical trialists, statisticians, epidemiologists and journal editors published the Consolidated Standards of Reporting Trials (CONSORT) statement[31] to help investigators and authors. It includes a checklist and flow diagrams incorporating the crucial elements of an RCT. The CONSORT statement has been revised since the original publication to clarify specific checklist items.[32]

In the past few years, as more surgeons have trained in or been exposed to health research methodology, there has been an increase in published RCTs dealing with surgical problems. Unfortunately, surgical trials, in contrast to medical trials, are plagued by unique methodological problems. Although the issue of randomization is straightforward, that of concealment remains problematic. Although investigators and participants in a medical study cannot tell a treatment pill from a placebo right up to study completion, concealment of treatment allocation group disappears once a surgeon opens an envelope or telephones the randomization centre.

Differential surgical care may bias results while the patient is in the operating room or during the follow-up visits, especially if the surgeon is

biased in favour of one or the other procedure.

Surgical interventions may include additional procedures that may influence the postoperative results and overestimate the effect. Standardization of a surgical intervention is a difficult problem.

As a large number of patients may be required to give the study statistical power, an RCT may require multiple surgeons (with different degrees of experience) and multiple centres (with different cultures). The analysis needs to take these variables into consideration, making large RCT studies complex.

As novel surgical interventions are introduced, learning curves need to be taken into account.[11–13,33] An RCT design should be considered only when the surgeons participating are equally capable of performing the novel and comparator interventions. Failure to control for the learning curve may underestimate the effect size of the novel intervention.

The *sine qua non* for undertaking an RCT is a state of equipoise.[34,35] Surgical equipoise is present when surgeons are truly ambivalent as to which of 2 readily available surgical interventions is superior. Furthermore, clinical equipoise reflects ambivalence within the practice community. This ethical consideration was addressed by the authors of the report under consideration in an indirect way. In the introduction of their article they state that "cement disease" may be obviated by the use of the cementless prosthesis, as initial reports were encouraging.

Assessors of outcomes can be blinded if the surgical interventions under consideration have similar incisions and hence, similar scars. Concealment may fail if assessors such as physiotherapists, who usually perform QoL assessments and RoM measurements, have access to patients' charts and operative records. Prudent surgical investigators describe what steps have been taken to avoid bias in outcome assessment.

Since publication of the CONSORT statement, the quality of reporting in published RCTs has improved considerably. Surgeons should make an effort to implement the steps mentioned in this statement when reporting a surgical RCT. Finally, when appraising a surgical RCT, surgeons are encouraged to be critical and use systematically the Users' Guides listed in Box 1.

Competing interests: None declared.

References

1. Laupacis A, Bourne R, Rorabeck C, Feeny D, Tugwell, P, Wong C. Comparison of total hip arthroplasty performed with and without cement: a randomized trial. *J Bone Joint Surg Am* 2002;84(10):1823-8.

2. Birch DW, Eady A, Robertson D, De Pauw S, Tandan V. Users' guide to the surgical literature: how to perform a literature search. *Can J Surg* 2003;46:136-41.

3. Coditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I. Medical. *Stat Med* 1989;8:411-54. [II. Surgical. p. 455-66.]

4. Schultz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-12.

5. Concato J, Shah N, Horwitz RI. Randomized controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342(25):1887-92.

6. Archibald S, Bhandari M, Thoma A. Users' guide to the surgical literature: how to use an article about a diagnostic test. *Can J Surg* 2001;44:17-23.

7. Urschel JD, Goldsmith CH, Tandan VR, Miller JD. Users' guide to the surgical literature: how to use an article evaluating surgical intervention. *Can J Surg* 2001;44:95-100.

8. Thoma A, Tandan V, Sprague S. Users' guide to the surgical literature: how to use an article on economic analysis. *Can J Surg* 2001;44:347-54.

9. Hong D, Tandan VR, Goldsmith CH, Simunovic M. Users' guide to the surgical literature: how to use an article reporting population-based volume–outcome relationships in surgery. *Can J Surg* 2002;45:109-15.

10. Guyatt G, Rennie D. *Users' guides to the medical literature: a manual for evidence-based clinical practice.* Chicago: AMA Press; 2002.

11. Chalmers TC. Randomization of the first patient. *Med Clin North Am* 1975;59:1035-8.

12. Parikh D, Chagla L, Johnson M, Lowe D, McCulloch P. D2 gastrectomy: lessons from a prospective audit of the learning curve. *Br J Surg* 1996;83:1595-9.

13. Bonenkamp JJ, Songun I, Hermans I, Sasako M, Welsaart K, Plukker JTM, et al. Randomized comparison of morbidity and mortality after D1 and D2 dissection for gastric cancer in Dutch patients. *Lancet* 1995;345:745-8.

14. Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT. Statistical assessment of the learning curves of health technologies. *Health Technol Assess* 2001;5:1-79.

15. Mohammed MA, Cheng KK, Rouse A, Marshall T. Bristol, Shipman and Clinical governance: Shewhart's forgotten lessons. *Lancet* 2001;357:463-7.

16. In: Friedman LM, Furberg CD, DeMets DL, editors. *Fundamentals of clinical trials.* 3rd ed. New York: Springer Verlag; 1998. p. 43-6.

17. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomized trials affect estimates of intervention efficacy reported in meta-analysis? *Lancet* 1998;352:609-13.

18. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358-61.

19. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomized controlled trials. *Br Med J* 1999;319:670-4.

20. Kaptchuk TJ. Powerful placebo: the dark side of the randomized controlled trial. *Lancet* 1998;351:1722-5.

21. Last JM. *A dictionary of epidemiology.* New York: Oxford University Press; 1995.

22. Van Der Linden W. Pitfalls in randomized surgical trials. *Surgery* 1980;7:258-62.

23. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomized trials in surgery: problems and possible solutions. *Br Med J* 2002;324:1448-51.

24. Guyatt GH, Pugsley SO, Sullivan MJ, Thompson PJ, Berman L, Jones NL, et al. Effect of encouragement on walking test performance. *Thorax* 1984;39:818-22.

25. Harrell FE Jr. *Regression modelling strategies, with application to linear models, logistic regression, and survival analysis.* New York: Springer Verlag; 2001.

26. Elwood M. *Critical appraisal of epidemiological studies and clinical trials.* New York: Oxford University Press; 1998.

27. Last JM. *A dictionary of epidemiology.* New York: Oxford University Press; 2001.

28. Armitage P, Berry G. *Statistical methods in medical research.* London: Blackwell Scientific Publications; 1994.

29. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG, editors. *Statistics with confidence: confidence intervals and statistical guidelines.* London, UK: BMJ Books; 1989. p. 83-100.

30. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine.* 2nd ed. Boston: Little Brown; 1991.

31. Begg CB, Cho MK, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276:637-9.

32. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Med Res Methodol* 2001;1:2.

33. Testori M, Bartoformei M, Grana C, Mezzetti M, Chinol M, Mazzarol G, et al. Sentinel node localization in primary melanoma: learning curve and results. *Melanoma Res* 1999;9:587-93.

34. Schafer A. The ethics of the randomized clinical trial. *N Engl J Med* 1982;307:719-24.

35. Pocock SJ. Ethical issues. In: *Clinical trials.* Toronto: John Wiley; 1993. p. 100-9.