Randomized Controlled Trials in Knee Surgery

Robert G. Marx, MD, MSc, FRCSC
Warren R. Dunn, MD, MPH

INTRODUCTION

Randomized controlled trials (RCTs) are the most rigorous clinical scientific investigations. This study methodology is the closest to a true experiment in which a group of patients with a specific condition are randomly allocated to treatment groups and are then followed to determine outcomes. This type of design, when done well, is the least likely to include bias. It is powerful because if the patients are appropriately randomized, the two groups are comparable because all confounding variables (known and unknown) are balanced between the two groups, and therefore the only difference between them is the intervention. Any difference in outcome can only be attributed to the treatment.

The difficulty with this study design is that it requires a significant amount of research time and it also is very expensive. As well, the results may not be generalizable to all patients who have the specific condition because patients who volunteer for this type of study are often systematically different from those who don't volunteer.[8,15,16] Patients who participate in RCTs[5] are more educated, have better general health, and on average are more compliant with treatments.[2] Therefore, the results of RCTs need to be interpreted in this light; however, the validity of the results of RCTs is greater than with any other study design.

Over the past decade, the number of RCTs has increased significantly in orthopedics in general and in subspecialty journals as well.[4] It is important for the orthopedic surgeon to understand issues relating to the conduct and evaluation of this type of research. This article focuses on specific issues relating to RCTs and provides examples relating specifically to knee surgery.

Dr Marx is from Foster Center for Clinical Outcome and the Sports Medicine and Shoulder Service, Hospital for Special Surgery, New York, NY; and Dr Dunn is from Vanderbilt Sports Medicine Center and Vanderbilt Health Services Research Center, Nashville, Tenn.

Reprint requests: Robert G. Marx, MD, MSc, FRCSC, Sports Medicine and Shoulder Service, Hospital for Special Surgery, 535 E 70th St, New York, NY 10021.

BASIC DESIGN FEATURES OF RANDOMIZED CONTROLLED TRIALS

Several basic principles must be adhered to before conducting an RCT. The study question and primary and secondary hypotheses, must clearly be stated. The population of interest should be described clearly and the study participants who represent this population should be specified as well. For example, a randomized trial of anterior cruciate ligament (ACL) reconstruction may define the age of patients included and whether acute or chronic reconstructions will be studied, or both. Additionally, the investigators may include patients with meniscal or chondral pathology, however this should be described prior to the study.

The intervention should be described in detail and there should be a practical method for randomized allocation to treatment group. The primary and secondary outcomes should be described in detail and the sample size justified prior to undertaking the study. The analysis as well as potential interim analyses and stopping rules should be determined prior to starting. Lastly, the feasibility of completing the study and the ethics of randomizing patients with the specific condition should be clarified in advance.

RANDOMIZATION

Randomization refers to the act of assigning patients to treatment groups such that every new patient recruited to the trial has an equal chance of ending up in either of

the two groups. This eliminates the potential bias that can result from using inadequate randomization methods such as date of birth, date of presentation, or alternate assignment. Whenever the treatment allocation is known by the clinician investigator prior to the patient being allocated to a treatment group, the opportunity to select a patient's treatment exists. For example, if nonoperative treatment for ACL insufficiency is compared to surgery and the treatment allocation is determined by date of birth, the investigator may end up with one group that is different from another for an important prognostic variable, such as activity level.[1]

Stratified randomization is a two-stage procedure where patients initially are grouped into strata according to clinical features that have an important influence on outcome. Patients then are assigned to treatment according to separate randomization schedules within each stratum.[14] For example, patients who are being treated for a worker's compensation claim may be randomized separately when a trial is evaluating treatments for knee arthritis. This is because the patients being treated for a worker's compensation claim generally will have an inferior outcome; and if one group has more patients who are being treated for worker's compensation claims, the effect of this variable may outweigh the effect of the intervention being studied to treat their arthritis.

Randomization can fail if the two groups created by the randomization are unbalanced for critical features that may affect outcome. When the randomization is successful in achieving balanced groups, then the observed differences in outcome between the two groups can be attributed to the treatment rather than other prognostic features. In surgical trials, it is critical to stratify for surgeon because the individual performing the procedure has an affect on the outcome. For example, a randomized trial may compare two graft sources for ACL reconstruction using five surgeons. If randomization was not stratified for surgeon, then two surgeons could perform most of the procedures using one graft source and three surgeons could perform most of the procedures using the other graft source. In this case, if the graft sources perform differently, it would not be clear whether the difference in outcome is due to the individual performing the surgery or due to the graft itself. Therefore, all surgical randomized trials must be stratified for surgeon to avoid this problem. Imbalances between the groups are more likely to occur in small trials where chance alone has a better opportunity to create a situation where one group has a worse prognosis. As the numbers of patients increase, the groups are more likely to be similar for known and unknown prognostic variables, by random chance alone.

The important question is how many variables can be stratified for a given trial. In general, the fewer the better. If there are too many strata, they may not be completely filled due to small numbers, and there may be overstratification.[12] Therefore, investigators must select only those clinical variables that have a known and important affect on outcome, such as surgeon or another important clinical factor. For example, in a RCT of ACL graft type, if the investigator stratifies for surgeon, meniscal repair, patient activity level, patient age, and chondral injury, unless a large number of patients are recruited for the trial, there will be so many different strata that many will have such small numbers that analysis will be impossible.

Permuted block randomization is a modification of simple random allocation in which patients are allocated in small blocks that usually consist of two to four times the number of treatment groups so that at any point in the study the groups are nearly equal. If there are two treatment groups, the block size is usually two and four. The patients in the first block are randomly assigned so that there are equal numbers in each group. The patients in the succeeding blocks are then randomized in turn until the final sample size is achieved. The size of the blocks is randomly laid out and not disclosed to the investigators to prevent potential selection bias, which could occur if the block size is known. For instance, if the block size is two in a surgical trial that cannot be double-blinded, then the second allocation in each block can be predicted based on the first. For this reason, repeated blocks of two are rarely used.[7] This type of randomization ensures that no major imbalances in assignment to group occur.[11,12] Even if the study ends prematurely there will be nearly equal numbers in all groups. Without block randomization, by chance alone the first 20 patients randomized may lead to 18 in one group and 2 in the other. While this is unlikely, it is an undesirable possibility that may occur without block randomization.

THE TIMING OF RANDOMIZATION

The timing of randomization can be critical to the success of the trial. Ideally it should occur as close to the intervention as possible. This is particularly important in surgical trials where it may be determined intraoperatively whether or not the patient is eligible for the study. For example, in a study evaluating the effectiveness of an all-inside device for meniscal repair versus a traditional inside-out technique, the investigator may be more likely to repair patients who are randomized to the all-inside technique. This is because it is less invasive and a second incision is not required. They therefore may tend to use the device and repair smaller tears. The surgeon may only perform the inside out technique on large tears due to the increased morbidity with the second incision. In other words, the indications for meniscal repair may expand with the all-inside device due to the decreased morbidity. This may lead to more meniscal repairs with the all-inside

device including smaller tears, which have a better prognosis. Ultimately, in this scenario, the two groups have different prognoses, which is not ideal. In this hypothetical study, the ideal method to avoid this problem is to have the patients randomized after the meniscal tear is identified and the decision to repair the tear is made. If patients are randomized after the final decision is made to repair the meniscus, the surgeon is then "locked in" to the treatment option.

In addition to the timing of the randomization, a central randomization process is ideal. Preferably, the randomization is performed by an individual who is not involved in the trial. If treatment assignment is contained in envelopes, they should be lined so that transillumination cannot reveal the contents. Randomization by envelope is not ideal and this technique is more susceptible to problems with the randomization. Computer generated randomization techniques are preferred.

BLINDING

Blinding is defined as a person being unaware of treatment group assignment. This can be applied to the study patient, caregivers, or the evaluator. The purpose of blinding is to minimize bias that is associated with knowledge of treatment. For example, if a patient thinks they are in the new promising treatment group, they may experience more of a placebo effect. Patients who know they are in the "not so great" standard of care group may attribute symptoms to the treatment that are unrelated.[3] For example, if patients with chondral defects are randomized to either microfracture or debridement, the patients who receive the debridement may feel that their treatment is inferior because it is less involved. Similarly the physician who is aware of treatment assignment may look harder for a known complication in the experimental group than he or she might otherwise and apply a cointervention or relate personal biases to the patient about the treatment.

In surgical trials it is sometimes impossible to blind the patient. When comparing arthroscopic debridement to nonoperative treatment for osteoarthritis of the knee, the patients can not be blinded. The surgeon also will not be blinded, and it is important to try to decrease the number of cointerventions. The evaluator can be blinded by possibly having the patient wear tights or long pants for all follow-up assessments.

MEASUREMENT OF OUTCOME

The measurement of outcome in clinical trials is complex.[10] The major issue is that instruments used to evaluate outcome should reflect the hypothesis being tested. For example, if the question is one of efficacy (ie, if a treatment can work in an ideal setting) then the outcome should reflect that. If the question is one of effectiveness (ie, does the treatment work in the real world) then the outcome should be more patient-relevant, such as a disease-specific quality-of-life measure. Previous work by Marx[10] presents information regarding knee-rating scales

SAMPLE-SIZE CALCULATION

Sample-size calculation should occur early in the planning of a study to ensure that the trial has adequate statistical power to identify differences between treatment groups. This fundamental step is often skipped in orthopedic trials, which can lead to sample sizes too small to detect a difference between groups (type II error).[9] Freedman et al[6] found that only 9% of orthopedic trials in 1997 reported a priori sample size calculation, and as a consequence many of these trials were underpowered. An underpowered study means that the investigators found no statistically significant difference between the two groups but it is not clear whether the lack of statistical significance is due to an insufficient number of patients.

Four main variables affect the number of patients that must be enrolled in a study to allow a reasonable comparison. The first variable is the size or magnitude of the difference the investigator is trying to detect and how much variability exists between the measurements. The difference should be the smallest that is clinically meaningful or relevant. The smaller the treatment effect, the larger the number of patients that will be required.

Second, the estimate is related to willingness to make a type I error (the error of stating that there is a difference between the groups when one does not really exist). Because the implications of this type of error can be enormous and far reaching (ie, adopting a new expensive treatment when it doesn't really work), to minimize this risk most investigators set their willingness to make this error at 5% ($P=.05$).

Third, the estimate is related to willingness to make a type II error (the error of concluding that there is no significant difference between the groups when one does exist). For most trials the implications of this type of error are not as grievous as the type I error and therefore most investigators are willing to take a greater risk than for a type I error. The risk of making a type II error is usually set at 20% ($\beta=0.2$). The exception to this would be in the equivalency study.

Fourth, the number also is related to the standard deviation of the measure that will be used as the primary outcome. This number reflects the average distance that an individual measure will differ from the mean. If a measurement has a lot of variability, it will be more difficult to show a statistically significant difference between

groups and therefore a larger sample size will be required.

The type of outcome being measured (proportions of events or continuous variables) and the comparison group (between or within patients) in the study will indicate the exact equation required to calculate the sample size. The sample size lets the investigator know if the study is feasible and, if feasible, whether it should be done in one center or multiple centers and over what period of time the recruitment will occur. The number of study participants required will be related to the difference that the investigator wishes to be able to detect between the two groups.

The investigator should plan the analysis and the comparisons prior to conducting the trial. If this is not done, the investigator may risk "data dredging,"[13] which may be worthwhile for hypothesis generation but is dangerous for hypothesis testing. To avoid this, a statistician should be part of the research team from the outset to design the study and plan for the subsequent analysis. Interim analyses must carefully be determined prior to the beginning of the study as well. Data should not be analyzed during the study unless an interim analysis is required for safety or ethical reasons. If this is the case, the results should be studied in a blinded fashion by an independent panel who is aware of predetermined stopping rules should one group have a clear advantage over the other.

FEASIBILITY

Much goes into assessing the feasibility of a study. Are the investigators experienced enough to deal with the day-to-day issues that arise? Are there suitable numbers of patients? Will the patients be willing to be recruited to the trial as designed? Is the treatment feasible and can it be done in a reproducible way? Are there research staff in place to evaluate patients? Are the appropriate precautionary measures in place? More importantly, will the treatment be out of vogue before the study is finished? If these questions are not answered, the investigator risks wasting time attempting to conduct a study that cannot be done appropriately.

SUMMARY

Randomized controlled trials are complex, require great effort and attention to detail by the investigator, and are expensive. This research methodology is important to allow us to base our decision-making for patients on solid evidence to ultimately improve patient care.

ACKNOWLEDGEMENT

The authors acknowledge the late Sandy Kirkley, MD who was a leader in randomized controlled trials in orthopedic surgery. At the time of her death in September 2002, she was a pioneer in randomized controlled trials in sports medicine and shoulder and knee disorders. She made a substantial contribution directly and indirectly to this work, and she is greatly missed by colleagues, collaborators, friends, and patients.

REFERENCES

1. Aglietti P, Buzzi R, Zaccherotti G, De Biase P. Patellar tendon versus doubled semitendinosus and gracilis tendons for anterior cruciate ligament reconstruction. *Am J Sports Med.* 1994;22:211-218.
2. American Heart Association. *American Heart Association: The National Diet-Heart Study: Final Report.* New York, NY: Monograph No. 18; 1980.
3. Beecher HK. The powerful placebo. *J Am Med Assoc.* 1955;159:1602-1606.
4. Brophy RH, Marx RG. An assessment of the methodological quality of research. Presented at: the 35th Annual Eastern Orthopaedic Association Meeting; October 13-16, 2004; Rio Grande, Puerto Rico.
5. Dunn WR, Lyman S, Marx R; ISAKOS Scientific Committee. Research methodology. *Arthroscopy* 2003;19:870-873.
6. 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.
7. Friedman L, Ferburg C, DeMets D. The randomization process. In: *Fundamentals of Clinical Trials*. New York, NY: Springer-Verlag; 1998:61-81.
8. Horwitz O, Wilbek E. Effect of tuberculous infection on mortality risk. *Am Rev Respir Dis.* 1971;104:643-655.
9. Lochner HV, Bhandari M, Tornetta P III. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. *J Bone Joint Surg Am.* 2001;83:1650-1655.
10. Marx RG. Knee rating scales. *Arthroscopy.* 2003;19:1103-1108.
11. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. *Control Clin Trials.* 1988;9:327-344.
12. Pocock SJ. Allocation of patients to treatment in clinical trials. *Biometrics.* 1979;35:183-197.
13. Sackett DL. Bias in analytic research. *J Chronic Dis.* 1979;32:51-63.
14. Simon R. Restricted randomization designs in clinical trials. *Biometrics.* 1979;35:503-512.
15. Smith P, Arnesen H. Mortality in non-consenters in a post-myocardial infarction trial. *J Intern Med.* 1990;228:253-256.
16. Wilhelmsen L, Ljungberg S, Wedel H, Werko L. A comparison between participants and non-participants in a primary preventive trial. *J Chronic Dis.* 1976;29:331-339.