To Find the Truth

In Plato's dialogues, Socrates walks around in ancient Athens talking to people in an attempt to find the truth. One characteristic of the dialogues is that the participants try to find unambiguous definitions of important words as a basis for their discussions.

Research of today has the same goal to find the truth in medicine about the cause and cure of disease. Therefore the starting point must be the same as in Plato's dialogues: one must start with solid definitions and a solid framework within which to work.

The article by Dr. Kirkley was intended to be included in issue 10:3, "Measurements and Outcomes of Scientific Studies" but was delayed because of her untimely death immediately before publication. It has taken more than a year to prepare her initial manuscript for publication but we found it very important to let you share her thoughts. The article by Kirkley et al is a presentation of how to accomplish the best framework possible for a study. It outlines different designs and their weak and strong points. If the advice given is followed then orthopedic research will improve tremendously from presenting personal impressions without general application to true answers that are generally useful. This is an important step to take.

You may find it difficult, cumbersome, and may be impossible. It is mostly a question of habit and forethought. If you become used to one of the proposed designs (e.g., the matched control design) you will find that it gives you more reliable information, but sometimes it is not what you believe it would be.

A good surgeon with a certain media back up has a strong placebo effect. His patients certainly benefit from it, but no other surgeon can duplicate his results. Therefore his achievements are only a family matter of no interest to any other surgeon. A true finding from a well-performed study has an impact on the whole orthopedic community and possibly all patients.

Jan Gillquist
Professor emeritus
Linkoping University
Linkoping, Sweden

Clinical Trials in Orthopedic Surgery

A. Kirkley, MD, FRCSC,† R. G. Marx, MD, FRCSC, S. Griffin, CSS,‡ and W. R. Dunn, MD, MPH§*

Abstract: A clinical trial may be defined as a "planned experiment" that involves patients, and it is designed to elucidate the most appropriate treatment of future patients with a given medical condition. The first step in designing a clinical trial is to select the best study type. The major types are the case report, the case series, case control studies, cohort study, and the most rigorous of all study designs; the randomized controlled trial (RCT). The basic design features of an RCT are randomization including stratification, permuted blocks, timing and blinding, measurement of outcome, sample size, analysis, and feasibility.

Key Words: study design, randomized clinical trials

(*Sports Med Arthrosc Rev* 2005;13:61–68)

A clinical trial may be defined as a "planned experiment" which involves patients and is designed to elucidate the most appropriate treatment of future patients with a given medical condition. Essentially, the results of a sample of patients will be used to extrapolate to the general population and make recommendations on how a given condition should be treated. When designing and implementing a clinical trial the clinician scientist attempts to come as close to the truth as possible knowing full well that the evil villain bias will attempt to infiltrate the study at every turn. The selection of study design will be a balance of feasibility and methodological rigor.

*From the Hospital for Special Surgery New York, NY; ‡From the Fowler Kennedy Sport Medicine Clinic, 3M Centre, University of Western Ontario, London, Ontario, Canada; §Hospital for Special Surgery New York, NY. †Dr. Kirkley died in September 2002.
Reprints: S. Griffin, Coordinator, Kirkley Research Group, Fowler Kennedy Sport Medicine Clinic, 3M Centre, University of Western Ontario, London, Ontario, Canada N6A 3K7 (e-mail: sharon.griffin@uwo.ca).

■ *"Indeed, the orthopaedic profession has a propensity to act as a stampeding herd, rushing from the latest change in practice to the newest prophet of an even better result." **Gross '99 (personal communication)*

The purpose of any study is to find out the truth. When one embarks on the path of designing and conducting a trial to answer an important clinical research question one needs to be aware of bias. Bias can creep into a study at each level of execution, from the original design stages, to measurement of variables, and even analysis of data. Hence, it is important to be heedful of bias, and the potential reward is a trial that answers an important question with results that are relevant and generalizable to the population at large and has the potential to change the practice of medicine.

OVERALL DESIGN

The first step is to select the best study type to answer the clinical question. This will be based on a balance of methodological rigor and feasibility. Once the study type is selected the goal is to design a trial that makes it difficult for bias to creep in. Here we discuss the major study types.

THE CASE REPORT

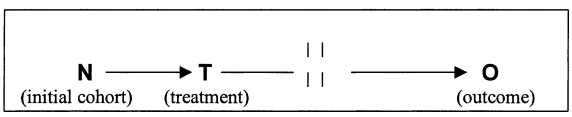
A case report is a description of one or a few patients' condition and response to treatment. It is useful to describe an unusual presentation of a condition or to hypothesize on the cause of the condition or the reaction to treatment. The results are not particularly generalizable to patient care as much is due to chance. Large variations in biologic or psychologic measures and disease course make it impossible to determine, based on clinical observation, if a treatment alone has made a difference on outcome.

A case report in which one patient's condition and response to treatment is described does not constitute a clinical trial. The reason for this is that there are such large variations in biologic or psychologic measures and disease course, that it is impossible to determine, based on clinical observation, if a treatment alone has made a difference on outcome.

THE CASE SERIES

A case series is study of a larger group of patients followed over time for the purposes of describing the natural history of a disease or describing a response to treatment. This type of study suffers from the absence of a control group, which allows for serious potential biases such that it rarely makes a convincing contribution to the evaluation of alternative therapies. Unfortunately, this is the most common study design found in the orthopedic surgery literature. It is clear that this is a simple, cheap, and quick type of study to conduct especially if the data is collected retrospectively. However, it is important for the researcher to know that this study type is particularly susceptible to bias.

The Case Series



This design is most susceptible to the placebo effect that is a well-documented phenomenon. Whereas some have questioned the placebo effect,[1] it is generally estimated that for both objective and subjective outcomes a 35% effect of placebo can be ascribed to the treatment.[2–4] It is likely that the "bigger" the treatment the "bigger" the placebo effect. In orthopedic surgery most of our treatments are perceived by patients as "big" and therefore we can expect substantial placebo effects.

It is well recognized that most new treatments that look effective in case series are found to be ineffective when subjected to a controlled trial. An example of this was laser surgery[5] performed on the endocardium of patients with cardiac dysfunction. The case series showed consistent good results even with objective outcomes such as treadmill testing. Because of these promising results clinicians in the United States carried out over 6000 procedures. However, in a randomized clinical trial in which 300 patients were randomized to receive laser endoscopy or sham endoscopy there were no significant differences between the groups.

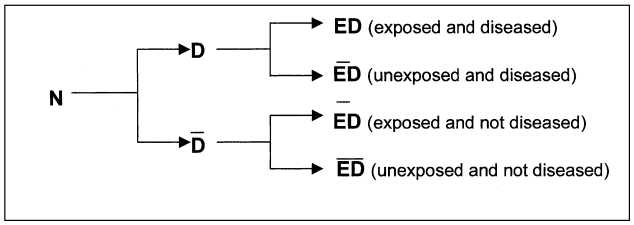
A retrospective case series is an unplanned observational study. This type of study contains serious potential biases such that they can rarely make a convincing contribution to the evaluation of alternative therapies. For instance, it is often the case that sicker patients receive more intensive treatment

In summary, it is rare for a case series to provide definitive information on therapeutic effects and the information should be considered thought provoking, in other words, hypothesis generating or as pilot data for more definitive trials.

CASE CONTROL STUDIES

In this type of design the direction of enquiry is backwards. A group of patients with a condition are matched to a control group who do not have the condition and then the investigator looks back in time to see the proportion of patients in each of the 2 groups who have had the exposure of interest. As an example, a study by Mohtadi[6] evaluating the association between timing of ACL reconstruction and the development of postoperative knee stiffness compared cases (those who underwent ACL reconstruction and developed clinically important stiffness) to controls (those who underwent ACL reconstruction and did not report stiffness). They found that in the case group there was a significantly greater proportion that had undergone surgery within 6 weeks of injury as compared with the control group.

Case Control Studies



This type of design is most appropriate when the condition of interest is rare (as in the example), when there is a very long lag time between the exposure and the development of the condition, or when the exposure of interest cannot be randomized for logistic or ethical reasons (smoking, alcohol consumption during pregnancy etc). It is also prone to many types of bias. For instance, Sackett[7] has catalogued 35 ways that bias can arise in sampling and measurement and reviews the 9 most common in his article of the topic (prevalence-incidence bias, admission rate bias, unmasking bias, non-respondent bias, membership bias, diagnostic suspicion bias, exposure suspicion bias, recall bias, and family information bias).

One strategy to avoid bias in this type of study is to carefully and cautiously select the control subjects. Controls should come from the same joint source population as the cases and represent the exposure experience of this population. There are many strategies to selecting appropriate controls, but the underlying principle is to select controls from a population that contains all of the people who, had they developed the disease of interest, would have become cases in your study. Some common sources include hospital-controls, neighbor-controls, friend-controls, and the general population.[8]

Under matching is the failure to select cases and controls that are sufficiently alike in important characteristics. When under-matched, a case-control study can succumb to confounding. To take a simple example, consider the investigator who wished to evaluate the relationship between cigarette smoking and rotator cuff tears. In his study, patients with rotator cuff tears were compared with age and gender matched controls that did not have rotator cuff tears (patients presenting to the same sports medicine clinic with knee problems). The investigator found a much larger proportion of smoking in the cuff tear patients as compared with the controls. Could confounding have been at play here? The patients with rotator cuff tears were more likely to be manual laborers than those presenting with knee pain. Manual laborers are more likely to be of lower educational level than those in non-laboring occupations and it has been clearly documented that smoking is associated with lower educational level. Therefore, a spurious relationship between smoking and rotator cuff tear may have been set up. What the investigator should have done was match the subjects for occupation, education level and gender, in other words those variables that could have acted as confounders.

Under matching is common but over matching can also occur resulting in the error of selecting controls that excessively resemble the cases. If the investigator happens to match on a factor that is itself related to the exposure under study, there is an increased chance that the matched case and control will have the same history of exposure. When over matched, a case-control study can fail to discover an association that, in fact, is present and real.[9]

Once the investigator has decided on the selection of the cases and controls, the condition under study should be defined in specific, unambiguous terms by operational diagnostic criteria. It is important that similar diagnostic procedures and criteria have been used among cases and controls. In other words cases and controls should have had an equal likelihood of being checked for the occurrence of the condition, and the diagnostic procedures should have been performed and interpreted equally in both groups. This is necessary to ensure that members of the control group are actually free of the condition, and this is especially important when the disease may occur in an asymptomatic form. To return to our rotator cuff example, it is quite conceivable that some of the control subjects had rotator cuff tears that were asymptomatic.

This type of design often relies on recall or health records, both of which are notoriously inaccurate.[10] Bias can easily affect how patients recall exposures. For example, in a recent study evaluating recall bias in rheumatoid arthritis, investigators asked individuals with rheumatoid arthritis and their unaffected siblings whether their parents had rheumatoid arthritis. Interestingly, the subjects with rheumatoid arthritis were twice as likely to report their parents as having the same or similar condition than their siblings who did not have the condition.[11]

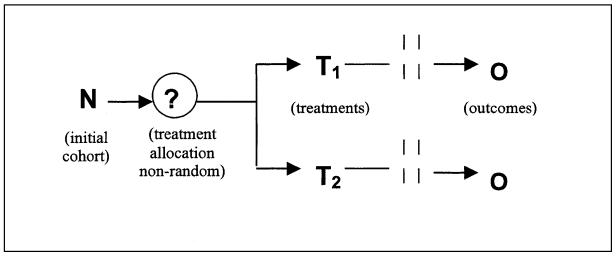
In summary the case control study is classically appealing but the clinician scientist must be aware of the potential biases that can arise using this type of design. However, with the proper selection of controls this can be a powerful design.

COHORT STUDY

In the cohort design a group of people (a cohort) is assembled, none of whom has experienced the outcome of interest, but all of whom could experience it. On entry to the study, people in the cohort are classified according to those characteristics that might be related to outcome. These people are then observed over time to see which of them experience the outcome. From this it is possible to determine how initial characteristics relate to subsequent outcome events. This type of study can be done prospectively where it is possible to collect data specifically for the purposes of the study and with full anticipation of what is needed, therefore avoiding biases that might undermine the accuracy of the data. On the other hand, data for retrospective cohort studies is usually gathered as part of medical records for patient care and these data are usually not of sufficient quality for rigorous research. This type of design is most commonly used as the most rigorous of the feasible means of determining whether exposure to a potential risk factor results in an increased risk for disease.

In the orthopedic literature the cohort design has been modified to compare 2 different treatments of the same condition. When done in a prospective fashion, eligibility criteria, treatments, and outcome measures can be standardized. The subjects can be matched for known prognostic variables. Administratively this type of design is much easier than a randomized clinical trial. The major disadvantage, however, is that there is an implicit but often unstated method by which patients are selected to undergo one treatment versus another. Therefore even if the groups are matched for known prognostic variables they may still be systematically different for important but unknown variables, and therefore the investigator ends up comparing apples to oranges. A good example is the well-known study by Daniels[12] evaluating the long-term effect of anterior cruciate ligament (ACL) injury and reconstruction.

Cohort Study



In this study the group of patients, those who underwent early ACL reconstruction had the worst outcome. However, even if the groups on the surface look similar (age, gender, follow-up time, etc.), it is highly likely that they were systematically different with respect to sport participation, aggressiveness, desire to play, type of sport, and neuromuscular coordination. These factors alone could well have accounted for the differences that were seen.

RANDOMIZED CLINICAL TRIAL

Clinicians may find it useful to know some history of clinical trials to better appreciate the current status of orthopedic clinical trials. Because of the influence of Sir Austin Bradford Hill, the MRC in England conducted the first reported randomized controlled trial (RCT), in 1948. In 1950 a double- blinded study in which a placebo was used, evaluated streptomycin for the treatment of pulmonary tuberculosis.[13] The investigators used sealed envelopes for the randomization process and the evaluators were blinded to group assignment.

Prior to World War II, regulatory bodies had no formal clinical trial requirements for approval of drugs. Rather, these decisions were based on animal studies that demonstrated a lack of toxicity, and some anecdotal clinical experience. In 1960 it became apparent that a large cohort of children born with major deformities of their limbs was the offspring of women who had ingested an anti-nausea drug by the name of Thalidomide during pregnancy. This disaster is credited with the major changes that soon followed in the criteria by which drugs were approved. It was agreed that more stringent criteria demonstrating safety and effectiveness would be ideal. On the other hand, there was much concern that these requirements would dramatically increase the cost of bringing drugs to market, and this would result in innovation being stifled, with the eventual outcome of slowing progress in health care research. Nonetheless, safety was deemed more important and so in 1962 controlled trials were required and by 1969 randomized controlled trials. Currently, a typical drug receiving approval by the Food and Drug Administration in the United States will have been evaluated in over 3,000 patients in phase 1 through phase 3 trials and between 10 and 80 trials will have been conducted.

The same criteria do not exist for surgical procedures or therapeutic devices. One could argue that for procedures that are inherently risky, irreversible, and expensive, the requirements should be at least the same as that of drugs but for some reason this is not the case. Rather, the onus is placed on the clinicians (as opposed to the manufacturers of devices) to rigorously evaluate new techniques and devices to provide evidence of safety and efficacy. Therefore it is important for all orthopedic surgeons to be aware of the appropriate design of randomized clinical trials to critically review the literature or to conduct trials of their own.

THE RANDOMIZED CONTROLLED TRIAL

The randomized controlled clinical trial (RCT) is considered the most rigorous of all study designs, a true experiment in which a group of patients with a specific condition are randomly allocated to treatment groups and then are followed to determine outcome. This type of design, when done well, is likely the least vulnerable to bias. Nonetheless, an investigator who carries out a study of this design without proper attention to detail may find at the end of the trial that bias was there all along slowly but systematically ruining the trial.

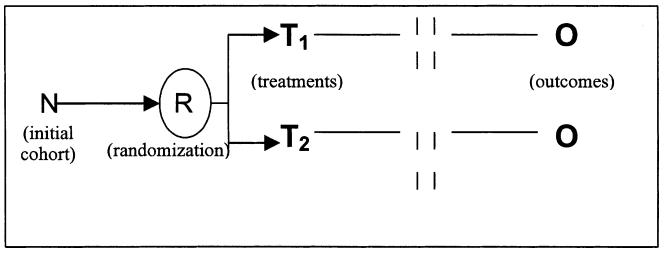
Fundamentally the benefits of this type of design are that it meets the assumption of random allocation for statistical tests; and the groups, if appropriately randomized, are comparable because all confounding variables (both those that are known and those that are unknown) are balanced between the groups. Therefore, patients, staff, and evaluators can usually be blinded.

The major criticisms of this type of design are that it is expensive both in terms of time and money and that the results are often not generalizable to all patients presenting with the condition of interest. This is because it has been shown that volunteer patients are often systematically different from those that choose not to go into a clinical trial.[14–16] These patients are usually of higher educational level, have better general health, and are, on average, more compliant with treatment recommendations.[17] It needs to be emphasized that this issue of generalizability relates to applying the results of the trial to general practice but itself is not a flaw that threatens the validity of the study results. The randomized controlled trial is therefore the most methodologically rigorous method of evaluating and comparing the results of treatments.

In fact Cochrane said:

■ *"It is only by such objective evidence that one can hope to clarify what is the best course of action. Otherwise one is left in a vacuum of uncertainty where the most enthusiastically supported procedures (which may nevertheless be misguided) are likely to be adopted."*

The Randomized Controlled Trial



Ryd et al[18] evaluated all articles in the 8 most cited general orthopedic journals indexed on Medline between 1966–1993 (25,538 clinical studies). There were 994 indexed as prospective studies whereas 138 were indexed as randomized controlled trials. It was concluded that the retrospective study still accounts for most of all published clinical articles in orthopedics. In recent years, a sharp increase in controlled trials had, however, occurred.

BASIC DESIGN FEATURES OF RCTS

The basics of the design of a randomized controlled trial include:

Study question and clearly stated primary and secondary hypotheses
Clear description of the population of interest and how the subjects for the study will be selected to represent this population
Detailed description of the interventions that will be applied and compared
Practical method for allocation to treatment group (randomization)
Detailed description of the primary and secondary outcomes to be used in the study
Justification of the sample size
Detailed description of the planned analysis including interim analyses and stopping rules
Consideration of the ethics of carrying out a clinical trial
Feasibility for carrying out the trial in the setting planned

Randomization

Randomization refers to the simple act of assigning patients to treatment group such that every new subject recruited to the trial has an equal chance of ending up in either treatment group. This eliminates the bias that can creep into trials using pseudo randomization such as when date of birth, date of presentation, or alternate assignment used. Any time the treatment allocation is known by the clinician investigator ahead of time it allows for the opportunity to select the patients for the treatment. Even worse is when assignment is based on judgment.

Randomization is a simple concept but there are some nuances that can make it more likely to be successful, especially in smaller trials.

Stratification

Stratified randomization is a two-stage procedure in which patients who enter a clinical trial are first grouped into strata according to clinical features that may influence outcome. Within each stratum, patients are then assigned to a treatment according to separate randomization schedules.[19] For example, patients who are receiving worker's compensation (WC patients) may be randomized separately from non-WC patients in a trial evaluating treatments for rotator cuff degeneration. Although stratification is very commonly used for clinical trials, investigators and readers are often uncertain about its importance.

Randomization procedures for clinical trials are intended to create groups of patients that are similar with regard to baseline characteristics that influence prognosis (both known and unknown). If the randomization is successful in achieving balanced groups then, the observed differences between groups in outcome may be attributed to the treatments rather than to other prognostic features. Simple randomization can fail if it creates 2 groups of patients that are unbalanced for critical features that are known or suspected to affect prognosis.

Major imbalances are most likely to occur in small trials where chance has a better opportunity to cause one group to be "sicker" than the other. Kernan et al[20] illustrate the chance that simple (unstratified) randomization may lead to treatment groups that are unbalanced with respect to a prognostic factor, by considering a trial of 2 therapies in a disease with an important prognostic factor that is present in 15% of patients. The chance that the 2 treatment groups will differ by more than 10% for the proportion of patients with the prognostic factor is 33% for a trial of 30 patients, 24% for a trial of 50 patients, and 10% for a trial of 100 patients, 3% for a trial of 200 patients and 0.3% for a trial of 400 patients. The chance of an imbalance is greater when prognostic factors are present in 30% of patients than when they are present in 15%. Therefore most methodologists would agree that for small trials (less than 100) patient stratification is advised to assure a valid comparison.

In addition, stratified randomization can increase the power of a study by reducing the variance of the difference between means of the groups.[21,22] Finally, stratification can facilitate subgroup analyses. It forces investigators to identify the subgroups that will be evaluated at study end (prevents data dredging) and helps to ensure that treatment assignments within subgroups are balanced. Each subgroup then becomes a small trial.[23]

The eternal question is how many variables and levels of those variables should one stratify for in a trial. The simple answer is that fewer are better. Although it is beyond the scope of this article, investigators should understand that too many strata can result in incomplete filling of blocks resulting in imbalances. This problem has been termed overstratification.[24] Investigators are encouraged to select only those clinical variables that have a known and important effect on the outcome, for example, individual surgeons in a surgical trial, or individual centers in a multicenter trial.

Permuted Blocks

Permuted block randomization is a modification of simple random allocation in which subjects are allocated in small blocks that usually consist of 2 to 4 times the number of treatment groups, so that at any point in the study the groups are nearly equal. If there are 2 treatment groups, the block size is usually 2 and 4. The subjects in the first block are randomly assigned so that there are equal numbers in each group. The subjects 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 that could occur if the block size is known. For instance, if the block size is 2 in a surgical trial that can not be double-blinded, then the second allocation in each block can be predicted based on the first. For this

reason repeated blocks of 2 are rarely used.[25] This type of randomization ensures that at no point during the study do major imbalances in assignment to group occur.[24,26] Even if the study ends prematurely there will be nearly equal numbers in all groups. The randomization process should be taken into account in the data analysis.[27,28] Hence, a potential disadvantage of this scheme is that data analysis can be more complicated if the randomization is accounted for in the analysis. Because blocking increases the power of the study by creating equal treatment groups, if it is disregarded in the analysis the significance levels are considered conservative estimates.[29]

Randomization Timing

The timing of randomization can be critical to the success of a trial. Ideally it should occur as close to the intervention as possible. This is particularly important in surgical clinical trials where it is often determined intraoperatively whether or not the patient is fully eligible for the study. Consider a trial evaluating the effectiveness of hemiarthroplasty to total shoulder arthroplasty. To allow the nursing staff to have the appropriate equipment in place and to book the appropriate amount of operating room time the clinical trial designers decided to randomize the patients pre-operatively at the time of recruitment in the clinic. However during surgery they find that a number of patients assigned to the total shoulder arthroplasty group cannot have a glenoid implanted (because of severe posterior erosion of the glenoid) and they are dropped from the study. This occurs several times during the study, the net effect being that the hemiarthroplasty group has on average more deficient glenoids than the total shoulder arthroplasty group. This makes the groups systematically different for the most important prognostic variable. In this example it would have been better to randomize intraoperatively after it was determined that a total shoulder or hemiarthroplasty could be carried out on the subject.

It is ideal to use a central randomization process if possible, preferably by someone not involved in the trial. If treatment assignment is contained in envelopes then these should be lined so that trans-illumination cannot reveal the contents.

Blinding

Blinding is defined as a person being unaware of treatment group assignment. This can be applied to the study patient, the caregivers, and/or the evaluator. The purpose of blinding is to minimize bias that is associated with knowledge of treatment. For instance if a patient believes that he or she is in the new promising treatment group that patient may experience more of a placebo effect. Conversely if a patient knows that he or she is in the "not so great" standard of care group that patient may attribute symptoms to the treatment that are unrelated.[4] 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 to the patient his or her personal biases about the treatment.

Total blinding is not always possible especially in surgical clinical trials. If one is comparing the effectiveness of open and arthroscopic treatment of rotator cuff tears, the patient cannot be blinded. The surgeon will not be blind and therefore must be removed from the loop as much as possible to avoid applying cointerventions (such as physiotherapy) unequally and certainly should not be involved in evaluation. The evaluator, however, can be blinded by simply having the patient wear a t-shirt for all follow-up assessments.

Measurement of Outcome

The measurement of outcome in clinical trials is a complex topic of which entire textbooks are devoted. The major issues however are that the instruments used should reflect the hypotheses being tested. For instance if the question is one of efficacy (ie, can a treatment work in an ideal setting) then the outcome will reflect that. If on the other hand the question is one of effectiveness (ie, does a treatment work in the real world) then the outcome will likely be more patient relevant, such as disease specific quality of life.

Sample Size Estimate and Analysis

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).[30] Freedman et al[31] found that only 9% of orthopedic trials in 1997 reported "a priori" sample size calculations, and as a consequence many of these trials were underpowered.

How many patients need to be enrolled in a study so that we can reasonably compare the treatment effect in the 2 groups? This hinges on 4 main variables. First, is the size or magnitude of the difference that we are trying to detect, and how much variation exists between these measurements. How big or small should this difference be? Ideally it should be the smallest difference that is clinically meaningful or relevant. The smaller the treatment affects the larger the number of patients that we will require.

Second, the estimate will be related to our willingness to make a type I error (the error of stating that there is a difference between the groups when this difference 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), most investigators want to minimize this risk and therefore they set their willingness to make this error at 5% ($P = 0.05$).

Third, the estimate will be related to our 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, which will be briefly described later.

Fourth, the number will also be 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. Obviously, if a measurement has a lot of variability it will be more difficult

to show a statistically significant difference between groups, and 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 that needs to be done 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 take place. The number of subjects required will be related to the difference that the investigator wishes to be able to detect between the 2 groups.

The details of how to do the analysis of a trial are beyond the scope of this article. There are however a number of principles that should be followed. The analysis relates directly to the hypotheses that are being tested. The analysis should be clearly described in a detailed fashion prior to starting the study. It is dangerous to decide on the analysis after the study is complete as looking at interesting twists in the data inevitably leads to "data dredging,"[7] which may be a worthwhile endeavor for hypothesis generation but is a dangerous one for hypothesis testing. A statistician should be part of the research team from the outset of designing the study and not an afterthought brought in after the study is finished to try and make something of the mass of data. It is probably true that the best studies have the simplest analysis. For instance, if the primary outcome is a continuous variable such as range of motion then a comparison of the 2 groups at a clinically relevant time point with a Student *t* test is the basic analysis. If the outcome is a rate such as the proportion of people who tear their ACL then the analysis will be a χ^2 analysis.

What if the researcher wishes to compare the groups with respect to multiple variables? First, recall what a statistically significant result means as defined by a *P* value. After all, statistics is all about probabilities, and the *P* value is the probability that the chance collection of patients may imply a difference in outcome that is not real.[32] The ritualistic use of a *P* value of 0.05 for decision rules is problematic to say the least.[33,34] Accordingly, many advocate the use of confidence intervals which are less conceptually rigid than a *P* value and offer more information.[32,35] The confidence interval offers a range of values that we believe, with some predetermined level of confidence, surrounds the true value. Hence, unlike the *P* value that tends to dichotomize results as either "significant" or "not significant," the confidence interval provides an estimate of how great the difference between 2 groups may actually be. Now imagine that the groups are actually equal but one is measuring 100 different variables. If the data are analyzed for any statistically significant associations (data dredging), the chance that the researcher will find a significant association by chance alone increases as the number of tests increase. Therefore, clinical epidemiologists make some recommendations on how to analyze data with multiple outcomes or time points. Most importantly one should declare a primary outcome. If it is positive, the groups will be considered different. The rest of the data can be analyzed in a hypothesis-generating mode only, or a correction can be applied. The most conservative of the multiple corrections that have been described is the Bonferonni Correction.[36] With this correction the *P* value is set at 0.05 divided by the number of variables being analyzed. For instance if one was evaluating 5 variables "*P*" would be set at 0.01. The reason this is overly conservative is that variables are not usually independent of each other. Another option is to combine the multiple end-points into a total score. The clinician scientist will have to decide if this makes clinical sense.

Interim analyses need to be carefully thought out "a priori". They should only be included in the study design if they are required for ethical reasons and then the results should be looked at in a blinded fashion by an independent group that is aware of clear stopping rules.

Feasibility

There is much that goes into assessing the feasibility of a study. Are the investigators experienced enough to deal with the day to day issues that arise? For example, 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 things in place? More importantly, will the treatment be out of vogue before the study is finished?

SUMMARY

The presentation of study designs mentioned earlier follows a hierarchical order from lower levels of evidence to higher levels of evidence. When interpreting the medical literature it is important to conceptually classify studies in such a framework to determine the validity of the conclusions, which ultimately dictate a clinician's decision to change clinical practice in accord with a published study. Although many factors can contribute to bias, the level of evidence of a study correlates with bias and can be used as a general guide when interpreting the scientific rigor of a study. For example, the level of evidence of a case series is very low and is close to that of expert opinion, whereas a randomized blinded trial is the highest level of evidence. However, no study design is perfect, because, as Kuhn pointed out, facts are not neutral—they are theory laden; hence, even the best studies are constrained to some degree by the methodology and by the expectations of the researchers.

REFERENCES

1. Hrobjartsson A, Gotzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *N Engl J Med*. 2001; 344:1594–1602.
2. Kirkley A, Alvarez C, Griffin S. The development and evaluation of a disease-specific quality-of-life questionnaire for disorders of the rotator cuff: The Western Ontario Rotator Cuff Index. *Clin J Sport Med*. 2003;13: 84–92.
3. Benson H, McCallie DP Jr. Angina pectoris and the placebo effect. *N Engl J Med*. 1979;300:1424–1429.
4. Beecher H. The powerful placebo. *JAMA*. 1955;159:1602–1606.
5. Cax J. Surgical treatment of cardiac arrhythmia *Kardiologiia*. 1990;30: 42–43.
6. Mohtadi NG, Webster-Bogaert S, Fowler PJ. Limitation of motion following anterior cruciate ligament reconstruction. A case-control study. *Am J Sports Med*. 1991;19:620–624.
7. Sackett DL. Bias in analytic research. *J Chronic Dis*. 1979;32:51–63.
8. Kelsey J, Whittmore A, Evans A, et al. Case-control studies: planning and execution. Lilienfeld AM, ed. *Methods in Observational Epidemiology*. New York: Oxford University Press; 1996: 188–213.

9. Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology*: *The Essentials*. Baltimore: Williams and Wilkins; 1996.
10. Lingard EA, Wright EA, Sledge CB. Pitfalls of using patient recall to derive preoperative status in outcome studies of total knee arthroplasty. *J Bone Joint Surg Am*. 2001;83:1149.
11. Schull WJ, Cobb S. The intrafamilial transmission of rheumatoid arthritis: the lack of support for a genetic hypothesis. *J Chronic Dis*. 1969;22: 217–222.
12. Daniel DM, Stone ML, Dobson BE, et al. Fate of the ACL-injured patient. A prospective outcome study. *Am J Sports Med*. 1994;22:632–644.
13. Jacob RF. Bias in dental research can lead to inappropriate treatment selection. *Dent Clin North Am*. 2002;46:61–78.
14. Horwitz O, Wilbek E. Effect of tuberculous infection on mortality risk. *Am Rev Respir Dis*. 1971;104:643–655.
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, et al. A comparison between participants and non-participants in a primary preventive trial. *J Chronic Dis*. 1976;29:331–339.
17. American Heart Association. The national diet-heart study: final report. New York: Monograph No. 18; 1980.
18. Ryd L, Dahlberg L. On bias. *Acta Orthop Scand*. 1994;65:499–504.
19. Simon R. Restricted randomization designs in clinical trials. *Biometrics*. 1979;35:503–512.
20. Kernan WN, Viscoli CM, Makuch RW, et al. Stratified randomization for clinical trials. *J Clin Epidemiol*. 1999;52:19–26.
21. Green SB, Byar DP. The effect of stratified randomization on size and power of statistical tests in clinical trials. *J Chronic Dis*. 1978;31: 445–454.
22. Nam JM. Sample size determination in stratified trials to establish the equivalence of two treatments. *Stat Med*. 1995;14:2037–2049.
23. Armitage P, Gehan EA. Statistical methods for the identification and use of prognostic factors. *Int J Cancer*. 1974;13:16–36.
24. Pocock SJ. Allocation of patients to treatment in clinical trials. *Biometrics*. 1979;35:183–197.
25. Friedman L, Furberg C, DeMets D. The randomization process. *Fundamentals of clinical trials*. New York: Springer-Verlag; 1998:61–81.
26. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. *Control Clin Trials*. 1988;9:327–344.
27. Titterington DM. On constrained balance randomization for clinical trials. *Biometrics*. 1983;39:1083–1086.
28. Kalish LA, Begg CB. The impact of treatment allocation procedures on nominal significance levels and bias. *Control Clin Trials*. 1987;8:121–135.
29. Kalish LA, Begg CB. Treatment allocation methods in clinical trials: a review. *Stat Med*. 1985;4:129–144.
30. 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; 83A:1650–1655.
31. 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.
32. Dorey F, Nasser S, Amstutz H. The need for confidence intervals in the presentation of orthopaedic data. *J Bone Joint Surg Am*. 1993;75:1844–1852.
33. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993;137: 485–496.
34. Cohen J. The Earth is Round ($p < .05$). *Am Psychol*. 1994;49:997–1003.
35. Szabo RM. Principles of epidemiology for the orthopaedic surgeon. *J Bone Joint Surg Am*. 1998;80:111–120.
36. Assennato G, Bruzzi P. Bonferroni in biomedical research. *G Ital Nefrol*. 2002;19:178–183.