Design of Randomized Controlled Trials

Kenneth Stanley, PhD

A major factor in the rapid advance of medical science over the past 50 years has been the development and refinement of the clinical research method known as the randomized controlled trial (RCT). A clinical trial is defined as a prospective scientific experiment that involves human subjects in whom treatment is initiated for the evaluation of a therapeutic intervention. In an RCT, each patient is assigned to receive a specific treatment intervention by a chance mechanism.

Nothing more clearly indicates the key role of an RCT in modern clinical research than the placement of this specific research method at the top of the list of levels of evidence in evidence-based medicine.[1] According to this classification, significant results of an RCT are more definitive than any other type of clinical research information.

The purpose of this article is to present an overview of the design of RCTs. Some of the principles of a high-quality study, such as the use of randomization, placebos, and double-blind designs are well known. Other principles such as stratification, use of a decision-making structure, and statistical power are known by many investigators but are not universally recognized or fully understood. These features plus others that indicate the design of a high-quality RCT are discussed. A companion article on the conduct and evaluation of RCTs will appear in a future issue of this journal.

Clarity of Study Objective

One of the most easily recognized aspects of a well-designed and conducted clinical trial is the apparent clarity of the research mechanism evident in its published report. Alternatively, one of the more common problems with a published clinical trial is the apparent "design by committee" in which different members of a protocol team have different goals. Because a clinical trial is a resource-intensive undertaking, many investigators feel that the study should attempt to satisfy a large number of objectives. The end result of this perspective is that some studies come to conclusion without convincing data on any specific question. Ideally, investigators should focus their study on a single major objective, such as the comparison of a new therapy versus the standard therapy with respect to a specific primary end point measure. Development of an explicit statement of the study objective will lead the investigators to the identification of a clear study design.

Classification by Study Design

Overall, clinical trials serve a multitude of functions that include the determination of a maximum tolerated dose, formulation of the basis for drug approval by the FDA, and definition of standard therapeutic management. They can be classified by either design or phase. The 3 most common designs are uncontrolled clinical trials, nonrandomized controlled trials, and RCTs. For uncontrolled trials, no concurrent comparison group exists and controls are implicit. This design is usually considered to provide the weakest level of clinical evidence. In nonrandomized controlled trials, a concurrent comparison group does exist, but patients are allocated to this group by a nonrandom process. Data from such studies are usually only considered reliable if confirmed by a randomized study or by a number of similarly designed nonrandomized studies in a meta-analysis. In an RCT, individuals are randomly allocated to 2 or more treatment groups, which usually include a standard treatment group and one or more experimental groups.

Classification by Study Objective and Phase

The system that classifies clinical trials by phase is given in the Table. Under this system, a new drug or intervention begins testing in phase I trials and then proceeds to phase II and III trials in a sequential manner that culminates in the establishment of the intervention as the new standard or in its licensing. After licensing, a phase IV trial may be undertaken to explore the long-term morbidity and effects that would be too uncommon to be detected in previous studies.

Treatment assignment for phase III trials nearly always uses a randomization mechanism. Although nearly all phase III trials are RCTs, not all randomized trials are phase III trials. The frequency with which randomization is used decreases for phase I and II trials. In addition to ensuring that groups are alike as much as possible, randomization in phase I and II studies is sometimes seen as a fair mechanism to provide patient access to a promising new drug of limited supply.

Although the concept of progression of a drug/intervention through phase I, II, and III trials has served its purpose well for many years, often the progression is not clearly demarcated. For example, phase I/II and phase II/III studies are quite common and may fit clinical needs better than strict adherence to the phase I, II, III progression. Furthermore, with a typical clinical trial gestation period of ≈1 year,

From the Department of Biostatistics, Harvard School of Public Health, Boston, Mass.
Correspondence to Kenneth Stanley, PhD, Department of Biostatistics, Harvard School of Public Health, 651 Huntington Ave, Boston, MA 02115. E-mail kstanley@sdac.harvard.edu

(*Circulation.* 2007;115:1164-1169.)

Circulation is available at http://www.circulationaha.org DOI: 10.1161/CIRCULATIONAHA.105.594945

Phases of Clinical Trials

	Objective	Typical No. of Patients
Phase I	To explore possible toxic effects and determine tolerance of the intervention (and tolerated dose, if a drug study).	10 to 30
Phase II	To determine if treatment has a therapeutic effect or if there is any hope for benefits to outweigh the risks.	20 to 50
Phase III	To compare new treatment to the standard therapy or a control or placebo (if no standard therapy exists).	100 to 1000
Phase IV	To obtain long-term, large-scale information on morbidity and late effects (postmarketing study).	Hundreds or thousands

investigators often adopt a multiphase study design to speed the pace of research.

Equipoise

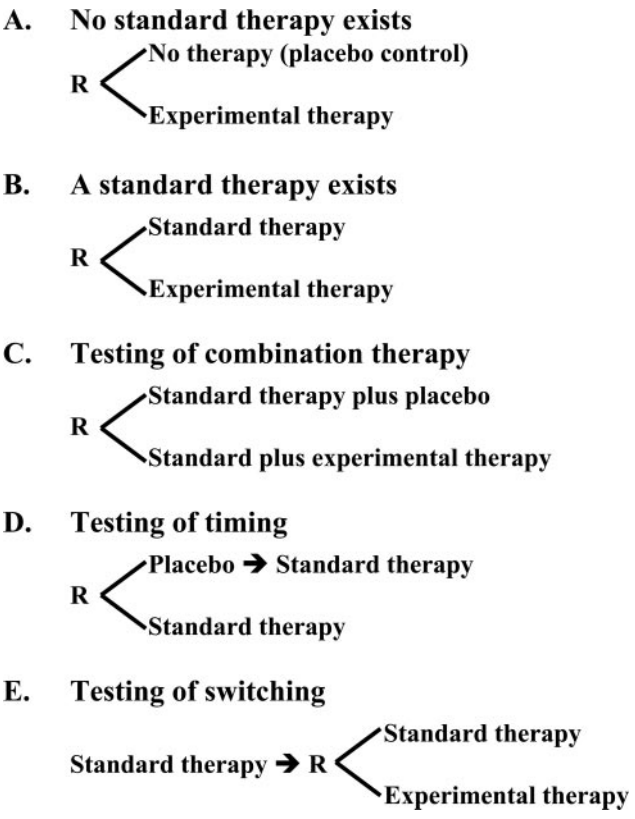
Equipoise is an ethical concept in the design and conduct of clinical trials. This concept states that, ethically speaking, we can only conduct clinical trials in areas of uncertainty and can only continue as long as the uncertainty remains. Thus, for an RCT it is unethical to initiate a clinical trial that does not include the "standard treatment" as 1 of the therapy arms, if such a standard exists, and it is unethical to include a therapy arm that is known to be inferior to any other treatment. This concept obligates investigators who plan a study to perform a comprehensive review of the medical literature during the protocol development phase and to establish a mechanism by which to keep informed of the latest released results from any related trials. Two practical problems are encountered with the concept of equipoise. First, there can be differences of opinion as to the level of evidence associated with "uncertainty." Some investigators may adopt the position of uncertainty unless clear information from an RCT exists, whereas others may use their clinical judgment to make such a determination. Second, it is unclear whether standard therapy is an individual, local, national, or international concept. Most often it is felt to be a local concept in which there may differences in personal preferences among some clinicians but a consensus among most practicing clinicians in that local geographic area.

Common Phase III Designs

A general progression exists for phase III study designs relative to a specific disease (Figure). The appropriate study design is largely dictated by the maturity of the therapeutic knowledge in that disease setting and design issues associated with equipoise.

In the clinical setting in which no prior drug (or intervention) has been established as the standard therapy, the study design for the initial phase III studies would compare a new experimental therapy group to a "no therapy" (eg, placebo control) group (design A). After a drug was found to be effective and identified as the "standard," subsequent phase III study designs would either compare a "new drug" to the standard (design B) or would compare the standard to combination therapy that involves the standard plus the "new drug" (design C). Often the decision to design the study as a head-on-head comparison of the "new drug" (design B) depends on how promising the new drug appeared to be at the phase II level. New drugs that looked promising but are not as potent as the current standard often end up being added to the standard in a combination therapy arm (design C). A sequence of promising but not spectacular drugs that enter a particular disease setting over a period of time often leads to a sequence of 2-, then 3-, then 4-drug combination regimen RCTs.

Other common phase III designs consider issues of timing and switching. The "testing of timing" study design depicts a situation in which the optimal time to initiate therapy is unknown (design D). The study team has selected 2 points in the clinical course of the disease to investigate. Patient entry and randomization is set at the earlier of these points and patients are randomized to the standard therapy or a "delay" arm. The subsequent trigger point (most often a clinical or laboratory event) on the delay arm would determine the initiation of the standard therapy for that group of patients. A comparison of results for these 2 groups would clarify the advantages of a delay in therapy initiation, if any.



Common phase III designs. R indicates randomization.

Similarly, the "testing of switching" study design gives a strategy for evaluation of a switch from the standard therapy to a new experimental therapy (design E). All patients would be treated on standard therapy up to a specific point, often a chronological time or a clinical or laboratory event, at which point the patients would enter the study and be randomly assigned to continue the current standard therapy or switch to the new experimental therapy. The value of the switch could then be evaluated by a direct comparison of the 2 randomized groups.

Randomization

A scientifically valid comparison between 2 treatment groups depends on the groups being alike as much as possible, with the only exception being the specific treatments under investigation. Without such an assurance, healthier patients may be given one treatment and sicker patients another treatment, and the observed result would be biased in favor of the healthier patients rather than serve as a valid comparison of the treatments. The best way to achieve such a balance is by the use of randomization in which a chance mechanism determines the treatment assignment. Randomization will ensure that a specific treatment assignment is not known in advance to either the clinician or the patient. The primary benefit of randomization is that it will eliminate both conscious bias and unconscious bias associated with the selection of a treatment for a given patient.

Although the majority of clinical investigators today are convinced of the benefits of randomization, some disadvantages exist. Many investigators feel that the action of randomization interferes with the doctor-patient relationship. In order to participate in an RCT, clinicians must admit to a patient that it is not known which of the therapies would be best for the patient, which thereby potentially erodes their relationship with that patient. Furthermore, from an ethical perspective, a clinician should believe that these therapies are equivalent with respect to potential patient benefit, a situation many clinicians find uncomfortable.

Stratified Randomization

Although randomization does not guarantee balanced treatment groups, it will tend to produce treatment groups that are alike on average. Additional protection against a possible imbalance, however, is preferred. In common clinical research settings the difference in effect between treatments is small relative to the effect of prognostic factors, such as extent of disease and a patient's performance status. A concept useful for clinical trial design and analysis is that we are trying to detect a soft signal in a noisy environment. The soft signal is the effect of treatment and the noisy environment is patient variability caused by prognostic factors, referral patterns, adherence to therapy, and other factors. Careful study design can help improve the signal to noise ratio, which thereby more readily exposes any true difference in treatments.

Additional protection against a possible imbalance is easily obtained by the use of a stratified randomization. In stratification, patients are formed into risk groups (strata) based on 1 or more prognostic factors, and a separate randomization is conducted for each strata. When the treatment assignment groups are then summed over the various strata, the end result is a forced balance of these overall treatment groups according to the factors used to form the strata. Use of stratified randomization should be viewed as an insurance policy against a potential imbalance, and, because it has virtually no cost (ie, no increase in number of patients needed or additional administrative complexity), it should be routinely used in RCTs.

Selecting the Treatments to be Compared

Because of the ethical principle of equipoise, 1 of the treatment arms in an RCT should be the standard therapy arm. The selection of other new treatment arms for the randomized comparison can often be viewed as an attempt to find a balance between an aggressive step forward and a cautious step forward. The new intervention should have the potential for a meaningful medical advance, capable of producing a benefit that is strong enough to be detected with a moderate-size clinical trial. However, a step forward that is too aggressive may produce a clinical trial in which patients or their physicians are unwilling to participate. On the other hand, a new therapy that represents a cautious step forward may be very appealing to potential patients and their physicians but runs the risk of not being able to produce a benefit that is large enough to be detected by a clinical trial that uses available resources. The smaller the anticipated benefit, the larger the study needed to detect it.

Selection of the Patient Population

Selection of the patient population for a clinical trial is a process of making decisions about contrasts. Restriction of the study to a specific group of relatively homogeneous patients can nearly always minimize the number of patients studied in a clinical trial. The more alike and sensitive a group of patients is to the intervention under investigation, the less other factors can affect the results and the easier it is for the trial to detect the effect of the therapeutic intervention. On the other hand, the population of all patients in the general population that will eventually receive the treatment regimen should theoretically be the population under investigation. However, when the patient population includes a broader range, the number of patients needed increases, the cost of the study increases, and a greater risk exists that the true treatment effect may go undetected because of the noise added by the heterogeneity of the patient population.

A related contrast is the investigator's option to carefully select a set of patients that are motivated and more likely to adhere to the treatment regimen. Some patient groups are not able to adhere to even a moderately complex treatment program, which thereby dilutes the study. One of the best ways to ensure an efficient clinical trial is to establish a run-in period, and then restrict subsequent patient entry onto the main study to only those who demonstrated that they could adhere to the run-in regimen. This strategy is also effective in identification of patients who will be the least likely to be lost to follow-up.

Nearly all patient populations are a blend of different risk groups. When the primary end point is a time-to-failure type

end point (eg, survival), the statistical power is directly proportional to the number of observed failures. For example, consider a completed study of patients with congestive heart failure that used patient survival as its primary end point and had a patient population that could be clearly divided into 2 groups with different risks; call these groups A and B. Assume group A (a high-risk group) comprised 100 patients and that this group experienced 50 deaths. Assume group B (a low-risk group) comprised 200 patients and that this group experienced 10 deaths. Because group A experienced 50 deaths, it provided 5 times (50/10) more statistical information on mortality than group B. This means that the study results were mainly driven by group A. Even though group B had twice as many patients, its contribution to the study survival results was minimal. Inclusion of low-risk patients in a clinical trial population may not be a good investment of resources.

Placebos and Double-Blind Designs

If any of the outcome measures of an RCT are subjective, then it is important that the trial be designed as a double-blind, placebo-controlled study. Only when both the patient and caregiver are unaware of the treatment assignment can their desire for a favorable outcome not potentially bias the results of the trial. The value of blinding, however, extends to all clinical trial assessments. Reliability of the results of a trial is strengthened when, for example, investigators use mechanisms such as an independent blinded end points committee. Another benefit of the use of placebos is the objective assessment of toxicities. If an RCT of a new drug is conducted in an unblinded manner, then all unexpected toxicities on the new drug arm are often ascribed to the new drug. If such a study is conducted in a blinded manner, then the difference in the rate of toxicities between the new drug arm and the standard therapy arm is ascribed to the new drug, frequently a lower rate than what would be reported by an unblinded study. It is not always feasible to blind a clinical trial, for example in studies that involve surgery. Nevertheless, the most influential studies are often those that attempt to establish and maintain the highest scientific standards, which include blinding and the use of placebos.

Primary End Point

Selection of the primary end point is a key design element of an RCT. This is the outcome measure used to make the decision on the overall result of the study and serves as the basis to determine the number of patients needed for the study. Each clinical trial should have only 1 primary end point, which should be defined before initiating the study. Use of multiple primary end points in a clinical trial is often a sign that investigators have let their biases influence the results they wish to report to the medical community. If the primary end point is not defined until after investigators have reviewed the data, it is not difficult to sift through the data and select end points that confirm the investigators' bias. Regulatory authorities and most journals insist on the a priori identification of a single primary end point, which thus insures objective reporting of the study's findings.

Use of composite end points is common in high-quality cardiovascular RCTs. Composite end points are necessary because a number of clinical events, such as a nonfatal myocardial infarction or stroke, may indicate a clinical failure, whereas the selection of only 1 type of clinical event as the end point will not present a comprehensive clinical picture. However, care must be taken when a composite end point is defined to ensure that the clinical failure events include the events of interest as well as "anything worse." For example, consider an RCT that compares 2 treatments for patients with congestive heart failure and the composite end point "nonfatal myocardial infarction or stroke." If there were more deaths on 1 of the 2 treatment arms, then deaths may have prevented the observance of either a nonfatal myocardial infarction or stroke and thus artificially made the arm with more deaths appear better. In this example, one can avoid such interpretation difficulties by inclusion of "death" into the definition of the composite end point.

Sample Size and Statistical Power

A clinical trial should be designed to be definitive, whether positive or negative. A study should not be initiated unless a reasonable likelihood exists that it will provide an answer to the clinical question that is posed—either an intervention works or it does not. One of the most unfortunate possible outcomes of a clinical trial is an inconclusive result. In that case, the good will of the patients and the resources of the clinical research mechanism would have all gone for naught. The chance of this type of outcome can be minimized by adherence to the following principles.

A definitive study result is achieved by placement of a mathematical decision-making structure on the clinical trial in the study development phase. For RCTs, the basic mathematical structure involves (1) identification of the primary end point and the main objective of the trial, (2) formulation of the trial objective as an hypothesis to be tested, (3) specification of the medically important difference the study is designed to detect, (4) identification of the magnitude of the errors that are acceptable (ie, the desired precision of the trial), and (5) calculation of the sample size necessary to achieve this desired precision. As noted in the Table, the typical size of a phase III RCT is 100 to 1000 patients. The sample size determination method outlined below is appropriate for RCTs. Different approaches are used for phase I and II trials.

As an example that uses the most common type of end point seen in RCTs, consider a trial that compares 2 treatments, A and B, with respect to the proportion of successes observed in each treatment group, denoted P_A and P_B. In a randomized trial that compares these 2 treatments, we test the null hypothesis (H_O: $P_A - P_B = 0$) that the 2 treatments yield equivalent results versus the alternative hypothesis (H_A: $P_A - P_B \neq 0$) that the treatments yield different results.

The study is conducted in an attempt to gather sufficient evidence to show that the null hypothesis is incorrect. Samples of patients are selected and the estimated difference in proportions is calculated. The key question is: How far from 0 does this estimate of $P_A - P_B$ need to be before we have sufficient evidence to say the treatments are different? To

answer this question, we formulate the problem in statistical terms, with α as the probability of a conclusion that the treatments are different when in fact they are really equivalent (type I error), and with β as the probability of a conclusion that the treatments are not different when in fact they are different (type II error). For RCTs, traditionally the α level is set to be 0.05. The β level is most often set to 0.20 or 0.10 and is often stated as the power level $(1-\beta)$ for the study.

Let Δ be the difference in the primary end point between the 2 treatment groups that the study is designed to detect—the medically important difference. Therefore, Δ is the difference, P_A-P_B, considered to be both medically significant and biologically plausible. Any smaller difference is considered to be too small to be worth detection and not medically important. Any larger difference is considered to be biologically implausible; it is quite unlikely that there will be such a large difference between these 2 treatments. With α, β, and Δ specified, statistical methods can be used to calculate the sample size necessary to provide the desired precision. Numerous Web sites are available for these calculations, depending on the type of primary end point.[2–4]

Although the premature loss of cases to follow-up weakens the quality of a clinical trial, it is a fact of life for nearly all long-term studies. Sample size for clinical trials should be adjusted to take into account the anticipated proportion of cases lost to follow-up.

It is useful to review the "power statement" in the published report of an RCT. This statement, most often found in the statistical methods paragraph of the methods section, will specify (1) the original primary end point, (2) the medically important difference Δ the study was designed to detect, (3) the size of type I error α (usually 0.05), (4) the power (usually 0.80 or 0.90) or β, and (5) the sample size necessary to achieve this desired precision.

For example, consider the RCT by Dawkins et al that appeared in a recent issue of *Circulation*.[5] On page 3307 one finds that these authors have identified the rate of ischemia-driven target-vessel revascularization at 9 months to be their primary end point, their Δ to be the change from a 20% control rate to a 10% rate in the treatment group, their α to be 0.05, their power to be 80%, and their sample size to be $N=448$. Comparison of the power statement with the observed results from this article allows one to see that the prior planning for this study was well done. The abstract reports an observed control rate of 19.4% and an observed treatment group rate of 9.1%.

Need for Rapid Enrollment

The cooperation of a number of clinical centers is often needed to enter a sufficient number of patients on a clinical trial in a reasonable time frame. If study entry continues beyond 2 years, the investigators open their study up to the risk that the emergence of new advances from a different study may cause their clinical trial to be obsolete or to be stopped prematurely with no results because it may be unethical to continue. Investigators with limited access to patients who wish to participate in RCTs are well advised to join large multicenter efforts rather than attempt to strike out on their own.

Difference Versus Equivalence Trials

RCTs can be classified by their goals. Difference (superiority) trials aim to determine if sufficient evidence exists that 1 treatment arm is different from another. These trials are by far the most frequent. Equivalence (noninferiority) trials aim to determine that 2 treatment arms are equivalent (or nearly so) and are conducted less often than difference trials.

With the common difference trial, the investigators conclude a difference has been demonstrated if they observe a P value <0.05. A series of successful difference trials will thus move medical science forward with a series of improvements in the standard therapy.

An equivalence trial tries to demonstrate similarity between a new treatment and standard therapy. This is most often done to show that a less expensive or less toxic new treatment has clinical benefit very similar to that of the standard therapy. Equivalence trials are sometimes used by a pharmaceutical manufacturer when attempts are made to license a drug in a disease setting that already has 1 or more licensed drugs.

Many researchers who have planned a noninferiority trial, however, do not correctly present their results. The noninferiority design concept is a 1-directional concept. Either the new treatment is inferior to the standard therapy or it is not—a yes versus no type of decision. Statistical procedures for an equivalence trial should focus on that unidirectional decision with 1-sided tests, P values, and confidence intervals. Readers are referred to the COBALT (Continuous Infusion vs Double-Bolus Administration of Alteplase) study[6] and the accompanying editorial[7] for an example of how an equivalence trial should be reported.

Study Monitoring

Safety of the patients who participate in a clinical trial is of paramount importance. For an RCT this is achieved by 2 main mechanisms. One mechanism is the monitoring of each adverse event report as it occurs by a qualified clinician, often on the protocol team, and the resulting assessment as to whether the adverse event was expected or whether additional investigation or a modification of planned protocol treatment may be indicated.

A second mechanism is a Data and Safety Monitoring Board (DSMB), also known as a Data Monitoring Committee (DMC). This is an independent committee established to assess at regularly scheduled intervals the progress of an RCT, regarding enrollment, safety data, data quality, and the critical efficacy end points, as well as the continuing validity and scientific merit of the trial.[8] Because the DSMB/DMC is entirely independent of the clinicians who are participating in the study, it can ensure patient safety and study validity without compromise or bias of the study. Good study design and periodic monitoring also help the investigation maintain appropriate ethical standards. The ability of investigators to monitor and evaluate ongoing clinical trials has improved markedly with the recent initiative by many medical journals to require the registration of a clinical trial in a public trials registry as a condition for consideration of publication.[9]

Disclosures

None.

References

1. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM*. 2nd ed. Edinburgh, United Kingdom: Churchill Livingstone, 2000:173–177.
2. Brant R. Inference for Proportions: Comparing Two Independent Samples. Available at: http://newton.stat.ubc.ca/~rollin/stats/ssize/b2.html. Accessed January 2, 2006.
3. Brant R. Inference for Means: Comparing Two Independent Samples. Available at: http://newton.stat.ubc.ca/~rollin/stats/ssize/n2.html. Accessed January 2, 2006.
4. Schoenfeld D. Find Statistical Considerations for a Study Where the Outcome Is a Time to Failure. Available at: http://hedwig.mgh.harvard.edu/sample_size/quan_measur/para_time.html. Accessed January 2, 2006.
5. Dawkins KD, Grube E, Guagliumi G, Banning AP, Zmudka K, Colombo A, Thuesen L, Hauptman K, Marco J, Wijns W, Popma JJ, Koglin J, Russell ME; TAXUS VI Investigators. Clinical efficacy of polymer-based paclitaxel-eluting stents in the treatment of complex, long coronary artery lesions from a multicenter, randomized trial: support for the use of drug-eluting stents in contemporary clinical practice. *Circulation*. 2005; 112:3306–3313.
6. The Continuous Infusion versus Double-Bolus Administration of Alteplase (COBALT) Investigators. A comparison of continuous infusion of alteplase with double-bolus administration for acute myocardial infarction. *N Engl J Med.* 1997;337:1124–1130.
7. Ware JH, Antman EM. Equivalence trials [editorial]. *N Engl J Med*. 1997;337:1159–1161.
8. Food and Drug Administration. Guidance for Clinical Trial Sponsors on the Establishment and Operation of Clinical Trial Data Monitoring Committees. Available at: http://www.fda.gov/cber/gdlns/clindatmon.htm. Accessed January 2, 2006.
9. DeAngelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJPM, Schroeder TV, Sox HC, Van Der Weyden MB. Clinical trial registration: a statement from the International Committee of Medical Journal Editors [editorial]. *N Engl J Med*. 2004;351:1250–1251.

KEY WORDS: randomized controlled trials ■ statistics ■ trials