Evaluation of Randomized Controlled Trials

Kenneth Stanley, PhD

Although randomized controlled trials (RCTs) play a key role in modern medicine, considerable variability exists in their quality and in the reliability (reproducibility) of their results. The present article discusses features that characterize high-quality trials, such as intent-to-treat (ITT) analysis, the avoidance of patient exclusions, and a procedure for evaluating the reliability of study results. A single serious problem can sometimes invalidate a study. However, most often one must weigh the impact of various strengths and weaknesses of aspects of study design, conduct, and analysis. A companion article[1] on the design of RCTs appeared in a previous issue of this journal.

Study Conduct and Analysis

Exclusion of Cases

A major source of weakness in the reliability of results of a published RCT is the extent to which patients have been excluded from the analysis. Many reasons have been cited by clinicians for their exclusion of cases from analysis, but perhaps the most common reason is a desire to ensure that each patient is "adequately treated." Many investigators find it awkward to retain in an analysis patients who have died within the first few weeks or who were unable or unwilling to adhere to the treatment specified by the protocol document. The general design of many RCTs is given in Figure 1.

During the analysis phase of these studies, some investigators adhere to the concept that patients should be adequately treated; this is sometimes alternatively stated as the exclusion of cases with major protocol violations. The problem with such exclusions is that they are more likely to happen on the more aggressive therapy arm, and they are more likely to happen to patients in a higher-risk group. After such patients are excluded from the analysis, the average result for the remaining patients is thus artificially improved. This is a bias often seen in RCTs; unfortunately, it is easily accepted by the investigators because they expect the more aggressive treatment to do better than the standard therapy.

This issue can also be viewed conceptually. Ideally, the perspective of a data analysis should be looking forward from the point of randomization, including all the good and bad events that occurred to the patients once they started down the treatment path. These are the data (warts and all) a clinician needs to know to appropriately assess what is likely to happen to a patient if the clinician decides a patient should start that treatment. Some patients will experience side effects and will need to stop therapy at various time points, although some will be able to stay on the protocol treatment for the full desired time period. The perspective of a data analysis when cases of inadequate treatment have been excluded is the perspective of looking backward from the end of the trial through rose-tinted glasses—excluding the bad events and focusing only on the good. Results of cases remaining after some poorer-risk cases have been excluded may look impressive in the literature, but such results are not of practical value to clinicians who need to make prospective therapy decisions for their patients.

One of the first points a reader should check in the published report of an RCT is whether the sizes of the analyzed groups are similar. If a publication states that the treatments were randomly allocated in equal proportions but then reports, for example, that 120 patients were analyzed in the standard therapy group and that 108 patients were analyzed in the new treatment group, the reader should be on guard that there may have been a high rate of exclusions. If such a discrepancy is discovered, most often fewer cases will have been analyzed in the new treatment group, suggesting the same directional bias as indicated in Figure 1.

Missing Data

The problem of missing data is similar to that of the exclusion of cases, except that only a proportion of the data from some patients are unavailable for analysis. When data are missing at random, the power of the study is weakened, but it is not a serious concern. However, when data are missing because of aspects of treatment or disease, major problems with bias can arise. Patients with missing outcome observations are more likely to be patients with poor outcomes.

As an example, consider Figure 2, a display of the average (or median) values for a quantitative variable over time. The values plotted in this common type of figure are averages for the available data. This figure, for example, could be reporting left ventricular ejection fraction over time. Does this figure indicate that the average left ventricular ejection fraction is increasing?

Although the line connecting the averages is increasing, it should be noted that the average is based on 200 patients at baseline but only 50 patients at 2 years. The key question to be asked is, why are three quarters of the data missing at 2

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;115:1819-1822.)

Circulation is available at http://www.circulationaha.org DOI: 10.1161/CIRCULATIONAHA.106.618603

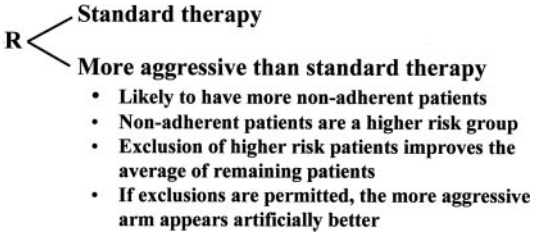

Figure 1. Effect of exclusion of major protocol violations. R indicates randomization.

years? It could be that higher-risk patients were unable to tolerate the therapy for the full 2 years and that the removal of these higher-risk patients from the group of patients forming the 2-year average left behind a group of low-risk patients who, naturally, would have higher left ventricular ejection fraction values. According to Figure 2, the true trend in averages could be stable, decreasing, or increasing; it is just not possible to tell. Alternatively, if the legend below the figure indicated that the average was based on 52 patients at baseline and 50 patients at 2 years, we would know we were looking at nearly all the data and could reliably conclude that left ventricular ejection fraction values were rising. No reliable interpretation of this figure is possible without understanding the pattern of the missing data. However, even if the pattern of missing data is known, there is no guarantee that the results of the figure will be clear if the percentage of missing information is large. But perhaps the most unsettling feature of this type of figure is the realization that most such figures in the medical literature do not include the number of patients below the figure, so it is impossible to form a reliable opinion of what the data actually show.

Every effort should be made to follow all patients and to obtain data values at the key time points. If it is important to have data values at every time point, missing values can be "imputed" by carrying the previous measure forward, by inserting a conservative value, by averaging adjacent values, or by computerized methods that take into account data from similar patients with complete information. A sensitivity analysis also may be useful to determine the extent of the impact of the missing data and the imputation method used. A sensitivity analysis, for example, may calculate the results of an analysis that replaces the missing values for 1 treatment group with a conservative imputation and replaces missing values for the other treatment group with an anticonservative imputation (or no imputation; available case analysis) and compares those results with an analysis that uses the opposite

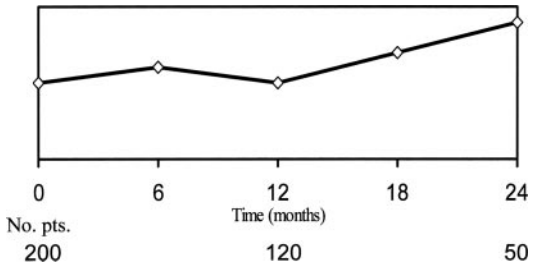

Figure 2. Average of values over time: the problem of missing data.

type of imputation strategy for the 2 treatment groups. If the results are qualitatively similar, one can deduce that the basic study conclusion does not depend on the type of imputation used (or the use of imputation).

Some journals use a rule of thumb for missing data when screening submitted RCT manuscripts. If the proportion of cases excluded or with missing data is similar to or larger than the size of the treatment difference being reported, then the study results are probably unreliable, and the manuscript is rejected before even going out for peer review.

The importance of excluded cases and missing data is underscored by the requirement of some journals that all submitted RCT manuscripts include the Consolidated Standards of Reporting Trials (CONSORT) patient-flow diagram.[2] The CONSORT diagram gives a clear account of all patients entered and their status at every major point in the conduct and analysis of the study. Although some journals do not routinely require the CONSORT patient-flow diagram, it is often requested if suspicion exists of patient exclusions or missing-data problems.

Intention to Treat

Most journals and all regulatory agencies require that an RCT be reported using an ITT study conduct and analysis philosophy. This philosophy can be briefly stated as:

- Comparisons between treatment groups are based on how subjects were randomized (intended to be treated) rather than how they actually were treated, and
- Everyone randomized to the study is analyzed, and all of a subject's follow-up information is included in the analysis, regardless of treatment discontinuation.

The major practical implication of the ITT philosophy is the inclusion of each patient's data in the analysis, regardless of patient withdrawal from treatment or deviations from the protocol. This implies that data should continue to be collected on patients after they enter the study, regardless of their outcome. Unfortunately, some clinical trials stop collecting patient information once the patient stops receiving protocol treatment, possibly biasing the study results. As discussed previously, subjects omitted are often those who are doing poorly or who could not tolerate the treatment. The ITT analysis almost always yields a conservative analysis—dampening the results of new or more aggressive regimens. Although it is easy to insert the phrase "this analysis was conducted by intention to treat" into a manuscript, no guarantee exists that the appearance of that phrase is consistent with the actual analysis being reported.

The ITT philosophy has a number of implications. First, once a patient is entered in the study, every effort must be made to maintain contact and to follow that patient, regardless of treatment cessation or changes. It is often very difficult for patients and clinicians to make this level of commitment when considering consent for an RCT. Second, it is not necessary to collect and analyze extensive information on drug dosing and adjustments, because such data will have no effect on the end-point analysis. The end-point tables and figures will be exactly the same whether or not these data are

TABLE 1. Characteristics of a High-Quality Randomized Controlled Trial

Clear research objective
Use of randomization and stratification
Use of placebos and double-blind designs
Clear primary end point defined a priori
Unbiased assessment of end points
Data and safety monitoring
The power statement
Clear accounting of all patients entered
Report of extent of follow-up
Statement of both idealized therapy and actual therapy
Report of the intent-to-treat analysis
Report of significant concomitant medications
Lack of conflict of interest
Conservative statement of conclusions

collected. Third, the primary analysis is a regimented analysis of the primary end point for all patients in groups as originally assigned, with no room for manipulation of study results. The ITT analysis is often viewed as an inexpensive analysis for a multimillion-dollar study. Fourth, although missing values are not desirable in an ITT analysis, if missing values do occur, they may need to be replaced by imputed values, and the potential impact of the use of imputed values may need to be assessed and discussed.

Evaluating a Published Trial

Characteristics of a high-quality RCT are summarized in Table 1. Few RCTs are optimal on all these points. Deviations from an optimal approach are often the result of specific patient or disease situations outside the control of the investigators. For example, sometimes it may not be feasible to blind the patients or investigators to a treatment intervention, and it may not be possible to obtain 100% patient follow-up in certain clinical settings. Although a single serious problem can invalidate a study, most often one finds varying degrees of deviations from optimal study design, conduct, and analysis in the published literature. One should not undertake an assessment of a published clinical trial using these points as rigid guidelines; rather, one should carefully weigh the strengths and weaknesses of each of the problems to obtain a sense of the reliability of the conclusions.

The published report of a well-designed clinical trial is easy to read and clearly reflects the study objective and design of the investigators. If a manuscript seems complex, this is often a sign of conflicting priorities of the investigators and of an effort to produce a study consistent with their prior beliefs rather than to provide an objective statement of the observed results.

The use of randomization, stratification, placebos, and double-blind designs are signs of a high-quality study. Every RCT should use a stratified randomization to provide balanced treatment groups for comparison. Although it is not feasible to use double-blind designs in all RCTs, they should be used more frequently because they provide objective data on all assessment measures, the importance of some of which

TABLE 2. Coronary Drug Project

	Clofibrate		Placebo	
Adherence, % of capsules taken	5-year Mortality, %	No. of Patients	5-year Mortality, %	No. of Patients
Poor, <80	25*	357	28†	882
Good, >80	15*	708	15†	1813

*P=0.0001.
†P<0.0001.
Adapted from the Coronary Drug Project Research Group,[3] courtesy of the *New England Journal of Medicine*.

may not be recognized until after the study has been completed.

The benefit of placebos in a broad context can be seen in the example from the Coronary Drug Project Research Group[3] given in Table 2. Patients who took 80% or more of their protocol dose of clofibrate had significantly lower 5-year mortality (P=0.0001). One might be inclined to conclude from this that clofibrate was beneficial. However, this study was placebo controlled. When a similar analysis was conducted on the placebo group, it was shown that patients who compliantly took their placebos also had significantly lower 5-year mortality. In this example, the use of placebos clarified the interpretation of the study results. Their use also helped expose a faulty retrospective analysis technique that had attempted to compare results in groups of patients on the basis of patient characteristics (eg, treatment adherence) observed after the time of randomization. Such retrospective evaluations of treatment adherence are misleading.

Identification before the study begins of a single primary end-point measure is key in the eventual production of reliable data from the study. Identification of multiple primary end points permits the investigators to explore the results and then selectively report the data that confirm their biases, misleading the medical community. The establishment of a blinded committee or investigator to provide unbiased assessments of end points and the use of an independent data and safety monitoring board or data monitoring committee to conduct blinded reviews of interim analyses will send a clear message to the readers that the investigators are doing their utmost to conduct the study at a high scientific level with concern for patient safety and the generation of unbiased results.

Prior planning of an RCT using a decision-making structure involving level, power, and the medically important difference can improve the likelihood that the study result will be definitive, whether it is positive or negative (see Stanley[1] for additional information). Assessment of the power statement from the publication of an RCT (often found in the statistical methods paragraph of the Methods section) can provide key insights into the initial primary end point and quality of the study planning. A surprising number of investigators report a power statement on the basis of 1 primary end point and then present the results of an analysis focusing on a different primary end point, demonstrating that they have probably decided to let their biases influence what they report after having reviewed the initial data analysis and not having liked the results. But, perhaps most importantly, if one does not find a power statement in the published report of an RCT, one must conclude that the

investigators are reporting a study that was not carefully planned or that was modified while underway.

Starting in 2004, a number of journals have begun to require the registration of a clinical trial in a public trials registry as a condition for consideration of publication.[4] This clinical trials registration system will markedly improve the ability of investigators to evaluate the published report of a clinical trial by permitting a direct comparison of the initially planned primary end point, study objective, and target sample size, for example, with those reported in the published manuscript.

When a well-designed clinical trial has been completed, the reliability of the results, be they positive or negative, can be assessed. If the study is positive (ie, it concluded a difference between the treatment groups), then the reliability of that statement is given by the *P* value for the primary end-point comparison between the 2 treatment groups. An observed *P* value less than 0.05 but close to the traditional 0.05 cutoff is weak evidence of a difference between the treatments, whereas an observed *P* value of 0.0001 is strong evidence of such a difference. Alternatively, if the *P* value is greater than 0.05, we say that the null hypothesis cannot be rejected and that we have insufficient evidence to conclude that a difference exists between the treatments. The level of confidence we have in that conclusion is given to us by referring to our power and medically important difference in the study-planning stage.

Perhaps the greatest contributions to the generation of misleading results from RCTs are from the exclusion of cases and missing data. Avoidance of biases in reporting RCTs depends on minimal, if any, exclusion of cases and on aggressive follow-up of patients to ensure the most complete follow-up that is possible. One of the first checks a reader should conduct on the published report of an RCT is to compare the number of patients analyzed in each treatment group. If a difference is found, especially in the direction indicating more frequent exclusion of cases on the more aggressive treatment arm, the reader should attempt to gauge the impact of such exclusions by comparing the proportion of excluded cases and missing data with the reported difference in treatment groups. Similarly, the reader should look for clear statements on the extent of follow-up. If data are missing for reasons related to treatment or disease, then the results are suspect and, again, should be assessed by comparing the proportion of missing data with the reported difference in treatment groups. Only rarely will all the patients on a clinical trial be able to stay on the randomized therapy as planned. Inclusion in the publication of a statement on the extent to which the idealized therapy can be followed is important for physicians who may prescribe such therapy for their patients. As mentioned previously, an ideal tool for clarifying the number of patients entered and the number analyzed is the CONSORT patient-flow diagram.[2] The CONSORT statement also provides an extensive checklist of items that can be used to assist in the evaluation of an RCT.

Although it is standard among journals that RCTs be reported using an ITT philosophy, the reader should check to see that sufficient data are given in the article to confirm that such an analysis has indeed been conducted. All too often, data are missing from patients who have stopped protocol therapy, thus artificially improving the average result of the patients with available data. A higher proportion of missing data on the more toxic or complex treatment arm will create a bias that will give that arm artificially inflated results.

The reliability of results from some studies can be harmed by the availability of over-the-counter medications that can affect their end points. Unless investigators carefully track the use of such nonprotocol concomitant medications and adjust the analysis for such use, the study results can be affected. The clearest example of this would be an RCT studying the reduction of pain. The availability of effective over-the-counter pain relievers could invalidate the main study results if use of these pain relievers is not properly tracked and taken into account.

Not all journals are created equal. Some set higher standards and, thus, publish studies with more reliable conclusions. Information obtained from proceedings and abstracts typically undergo no critical scientific review and, thus, should probably not be used for patient-management decisions.

Conflicts of interest may arise at various levels. Study results may be affected by a study design/analysis bias or a chauvinistic bias. As discussed previously, case exclusions, change of the primary end point, and the handling of missing data can all affect the results of a study. From the list of masthead authors and their institutional affiliations, one can often determine where the study's database was located and whether such critical decisions were made by individuals from an institution with a potential conflict of interest. If a study reporting the benefits of a specific therapy is authored by individuals from that same specialization, such a study should not be considered as reliable as one reporting the lack of benefits of that specific therapy. Authorship of a study that has no clear connection with the author's personal goals should be viewed as a sign of greater reliability.

Lastly, when reading the published report of an RCT, one should get the feeling that it has been written by objective scientists. The statement of the author's conclusions should be conservative, and there should be no statement of inferences in the absence of statistical significance. The reader should not see phrases such as "the data suggest that...," "it appears that...," and "although not significant, the data show that...." It is best for medical science overall for authors to be their own best critics and for readers of the medical literature to adopt the perspective of skeptics.

Disclosures

None.

References

1. Stanley K. Design of randomized controlled trials. *Circulation*. 2007;115:9:1164–1169.
2. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*. 2001;357:1191–1194.
3. The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med*. 1980;303:1038–1041.
4. 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. *N Engl J Med*. 2004;351:1250–1251.