

# The Journal of Clinical Pharmacology

<http://www.jclinpharm.org>

---

## Clinical Trial Design Issues: At Least 10 Things You Should Look For in Clinical Trials

Stephen P. Glasser and George Howard

*J. Clin. Pharmacol.* 2006; 46; 1106

DOI: 10.1177/0091270006290336

The online version of this article can be found at:  
<http://www.jclinpharm.org/cgi/content/abstract/46/10/1106>

---

Published by:

 SAGE Publications

<http://www.sagepublications.com>

On behalf of:

American College of Clinical Pharmacology

Additional services and information for *The Journal of Clinical Pharmacology* can be found at:

**Email Alerts:** <http://www.jclinpharm.org/cgi/alerts>

**Subscriptions:** <http://www.jclinpharm.org/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>

**Permissions:** <http://www.sagepub.com/journalsPermissions.nav>

**Citations** (this article cites 32 articles hosted on the SAGE Journals Online and HighWire Press platforms):  
<http://www.jclinpharm.org/cgi/content/abstract/46/10/1106#BIBL>

# Clinical Trial Design Issues: At Least 10 Things You Should Look For in Clinical Trials

Stephen P. Glasser, MD, FCP, and George Howard, DPH

*Randomized controlled trials remain the gold standard study design and yield the highest level of scientific credence. However, recognition of the limitations of the randomized controlled trial is important. This review highlights 10 potentially problematic areas one should carefully assess when performing or reading an article reporting the results of a randomized controlled trial, problematic areas that can affect the outcome of the trial and therefore mislead the reader. These areas include ethical issues, eligibility criteria,*

*masking (blinding), randomization, analytic methods, the selection of subjects for the interventional and comparison groups, selection of end points, and the interpretation of the results. Each of these is discussed, and examples of published articles are used to highlight the main points.*

**Keywords:** Clinical trials; randomized; limitations; review  
*Journal of Clinical Pharmacology, 2006;46:1106-1115*  
 ©2006 the American College of Clinical Pharmacology

The spectrum of evidence imparted by the different clinical research designs ranges from ecologic studies through observational epidemiologic studies to randomized controlled trials (RCTs). The RCT remains the gold standard study design, and its results are appropriately credited as yielding the highest level of scientific credence. However, recognition of the limitations of the RCT is important. As Grimes and Schultz point out, in this era of increasing demands on a clinician's time it is "difficult to stay abreast of the literature, much less read it critically. In our view, this has led to the somewhat uncritical acceptance of the results of a randomized clinical trial" (p57).<sup>1</sup> Also, Loscalzo, has pointed out that "errors in clinical trial design and statistical assessment are, unfortunately, more common than a careful student of the art should accept" (p3027).<sup>2</sup> To this end, we offer a review of RCTs with an emphasis on some of their pitfalls and highlight some particularly problematic areas that should be considered when performing or reading the results of such trials. In addition, we shall address questions such as what it is that leads the RCT to the highest level of evidence and what are

the features of the RCT that render it so useful. In this article, we will discuss a number of principals (eg, confounding, randomization, and why one needs to monitor the placebo group) that answers these questions.

Let us begin with the example of the postmenopausal hormone replacement therapy (HRT) controversy. Multiple observational epidemiologic studies had shown that HRT was strongly associated with the reduction of atherosclerosis, myocardial infarction risk, and stroke risk.<sup>3-5</sup> Subsequently, 3 clinical trials suggested that HRT was not beneficial and might even be harmful.<sup>6-8</sup> Why can this paradox occur? What can contribute to this disagreement? Why do we believe these 3 RCTs more than so many well-done observational trials?

There are at least 10 problematic areas one should carefully assess regarding clinical trials, problematic areas that can affect the outcome of a trial:

1. Ethical issues (protection of human subjects)
2. Implications of eligibility criteria (sampling)
3. Degree of masking
4. Randomization
5. Intention to treat analysis (the analytic method used)
6. Selection of interventional and comparison groups
7. Selection of end points

From the Division of Preventive Medicine (Dr Glasser) and the Department of Biostatistics (Dr Howard), University of Alabama at Birmingham, Birmingham, Alabama. Submitted for publication April 3, 2006; revised version accepted April 25, 2006.

DOI: 10.1177/0091270006290336

8. Interpretation of results
9. Trial duration
10. Selection of traditional versus equivalence testing

## ETHICAL ISSUES

Consideration of ethical issues is key to the selection of the study design chosen for a given research question or hypothesis. A full discussion of the ethics of clinical research is beyond the scope of this article, particularly as it pertains to using a placebo control. For further discussion see the references noted here.<sup>9-11</sup> (There is also further discussion of this issue under the Selection of Traditional Versus Equivalence Testing section.) The opinions about when it is ethical to use placebo controls is quite broad. For example, Rothman and Michaels are of the opinion that the use of placebo is in direct violation of the Nuremberg Code and the Declaration of Helsinki,<sup>10</sup> whereas others would argue that placebo controls are ethical as long as withholding effective treatment leads to no serious harm and patients are fully informed. Most would agree that placebo is unethical if effective life-saving or life-prolonging therapy is available or if it is likely that the placebo group could suffer serious harm. For ailments that are not likely to be of harm or cause severe discomfort, placebo is justifiable.<sup>11</sup> However, in the majority of scenarios, the use of a placebo control is not a clear-cut issue, and decisions need to be made on a case-by-case basis. One prevailing standard that provides a guideline for when to study an intervention against placebo with a RCT is when one has enough confidence in the intervention that one is comfortable that the additional risk of exposing a subject to the intervention is low relative to no therapy or the standard treatment but that there is sufficient doubt about the intervention that use of a placebo or active control (standard treatment) is justified. This balance, commonly referred to as  *equipoise* , can be difficult to come by and is likewise always controversial. Of importance,  *equipoise*  not only needs to be present for the field of study (ie, there is agreement that there is not sufficient evidence of the superiority of alternative treatments) but also has to be present for individual investigators (permitting individual investigators to ethically assign their patients to treatment at random).

Another development in the continued efforts to protect patient safety is the Data Safety and Monitoring Board (DSMB). The DSMB is now almost universally used in any long-term intervention trial. First a data and safety monitoring plan becomes part of the protocol, and then the DSMB meets at regular and

as-needed intervals during the study to address whether the study requires premature discontinuation. As part of the data and safety monitoring plan, stopping rules for the RCT will have been delineated. Thus, if during the study, either the intervention or control group demonstrates a worsening outcome, the intervention group shows a clear benefit, or adverse events are greater in one group or another (as defined within the data and safety monitoring plan), the DSMB can recommend that the study be concluded. But the early stopping of studies can also be a problem. For example, in a recent systematic review by Montori et al, the question was posed about what was known regarding the epidemiology and reporting quality of RCTs involving interventions stopped for early benefit.<sup>12</sup> Their conclusions were that prematurely stopped RCTs often fail to adequately report relevant information about the decision to stop early and that one should view the results of trials that are stopped early with skepticism.

## IMPLICATIONS OF ELIGIBILITY CRITERIA

In every study, there are substantial gains in statistical power by focusing the intervention in a homogeneous patient population likely to respond to treatment and to exclude patients who could introduce "noise" by their inconsistent responses to treatment. Conversely, at the end of a trial, there is a need to generalize the findings to a broad spectrum of patients who could potentially benefit from the superior treatment. These conflicting demands introduce an issue of balancing the inclusion/exclusion (eligibility criteria) such that the enrolled patients are as much alike as possible but, on the other hand, to be able to apply the results to the more general population (ie, generalizability). Figure 1 outlines this balance. What is the correct way of achieving this balance? There really is no correct answer. There is always a trade-off between homogeneity and generalizability, and each study has to address this, given the availability of subjects, along with other considerations. This process of sampling represents one of the reasons that scientific inquiry requires reproducibility of results; that is, one study generally cannot be relied on to portray "truth."

For example, most of the major studies assessing the efficacy of the treatment of extracranial atherosclerosis with endarterectomy have excluded octogenarians on the basis that this patient population may have a response to the challenges of surgery that is different from that of their younger counterparts.<sup>13-15</sup> Exclusion of these patients may have contributed to

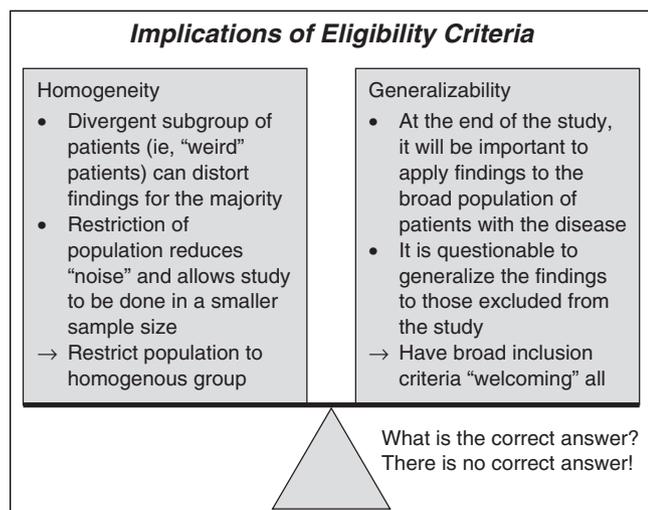


Figure 1. The balance of conflicting issues involved with patient selection.

the successful completion of these positive trials (finding a benefit for the new treatment—endarterectomy). However, now that the trials are complete, there is not level 5 (data from RCTs) evidence to guide the management of octogenarians with extracranial atherosclerosis, one of the subpopulations in which the need for this information is important. In the absence of this information, thousands of endarterectomies are performed in this patient population each year under the assumption that the findings from a younger cohort are generalizable to those at older ages.

Another issue is that RCTs in general are designed to test safety and efficacy (ie, does the drug work under optimal circumstances) and not to answer questions about the effectiveness of a drug, the more relevant question for economic analysis (ie, does the drug work in usual care). Thus, the use of effectiveness trials has been suggested to more closely reflect routine clinical practice. Effectiveness trials use a more flexible dosage regimen and a usual care comparator instead of a placebo comparator. Two approaches to this more real-world trial are the phase 4 trial or the prospective, randomized, open-label, blinded end point (PROBE) trial. The PROBE trial is further discussed in the Degree of Masking section. As to phase 4 trials, they are surrounded by some controversy. Perhaps a more formalized approach with a major emphasis on effectiveness trials would be appropriate for phase 4 trials.

### DEGREE OF MASKING

Although the basic concept of clinical trials is to be at equipoise, this does not change the often preconceived

suspicion that there is a differential benefit (eg, active drug even when investigational is better than placebo). Thus, if study personnel know the treatment assignment, there may be differential vigilance where the supposed inferior group is more intensively monitored (eg, “Are you certain you have not had a problem?”). In this case, unequal evaluations can provide unequal opportunities to differentially discover events. This is why the concept of double-blinding (masking) is an important component of RCTs. But one cannot always have a double-blind trial, and some would argue that double-blinding distances the trial from a real-world approach. An example in which blinding is difficult to achieve might be a surgical versus medical intervention study whereby postoperative patients may require additional follow-up visits and each visit imparts an additional opportunity to elicit events. That is, it is said that the patient cannot have a fever if the temperature is not taken, and for RCTs, events cannot be detected without patient contact to assess outcomes.

To address this more real-world principal, the PROBE design was developed. By using open-label therapy, the drug intervention and its comparator can be clinically titrated as would occur in a doctor’s office. Of course, blinding is lost here but only as to the therapy. Blinding is maintained as to the outcome. To test whether the allowance of open-label versus double-blind therapy affected outcomes differentially, a meta-analysis of PROBE trials and double-blind trials in hypertension was reported by Smith et al.<sup>16</sup> They found that changes in mean ambulatory blood pressure from double-blind controlled studies and PROBE trials were statistically equivalent.

### RANDOMIZATION

Inherent in all clinical research is the issue of confounders of relationships. A confounder is a factor that is associated to both the risk factor and the outcome and leads to a false apparent association between the risk factor and outcome (Figure 2). There are 2 alternative approaches to remove the effect of confounders in observational studies.

- Most commonly used in case-control studies, one can match the case and control populations on the levels of potential confounders. Through this matching, the investigator is assured that both those with a positive outcome (cases) and a negative outcome (controls) have similar levels of the confounder (by design). Because a confounder has to be associated with both the risk factor and the outcome and because through matching the suspected confounder is not associated with the outcome, the factor cannot

act as an outcome (it is not associated with both the risk factor and the outcome). For example, in a study of stroke, one may match age and race for stroke cases and community controls with the result that both those with and without strokes will have similar distributions for these variables, and differences in associations with other potential predictors are not likely to be confounded, for example, by higher rates in older or African American populations.

- In all types of observational epidemiologic studies, one can statistically or mathematically adjust for the confounders. Such an adjustment allows for the comparison between those with and without the risk factor at a fixed level of the confounding factor. That is, the association between the risk factor and the potential confounding factor is removed (those with and without the risk factor are assessed at a common level of the confounder), and as such, the potential confounder cannot bias the association between the risk factor and the outcome. For example, in a longitudinal study assessing the potential impact of hypertension on stroke risk, the analysis can adjust for race and other factors. This adjustment implies that those with and without the risk factor (hypertension) are assessed as though race were not associated with both the risk factor and the outcome.

The major shortcoming with either of these approaches is that one must know what the potential confounders are to match or adjust for them; it is the *unknown confounders* that are the problem. Another issue is that even if one suspects a confounder, one must be able to appropriately measure it. For example, a commonly addressed confounder is socioeconomic status (usually a combination of education and income), but clearly this is a factor in which it is difficult to agree on which measure or cut point is appropriate. The bottom line is that one can never perfectly measure all known confounders, and certainly one cannot measure or match for unknown confounders. To assure that both known and unknown confounders are equally distributed in the investigational and control groups, randomization is necessary.

The introduction of randomization to clinical trials in the modern era can probably be credited to the 1948 trial of streptomycin for the treatment of tuberculosis.<sup>17</sup> In this trial, 55 patients were randomized to either treatment with streptomycin and bed rest or treatment with bed rest alone (the standard treatment at that time). To quote from that article,

determination of whether a patient would be treated by streptomycin and bed rest (S case) or bed rest alone

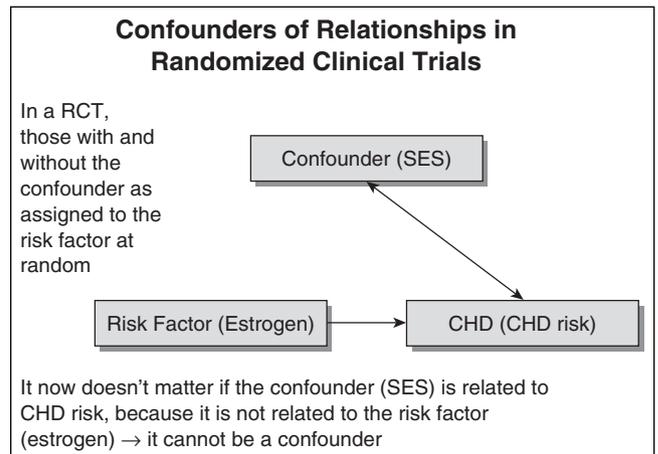


Figure 2. The relationship of confounders to outcome and how they are eliminated in a randomized controlled trial (RCT). SES, socioeconomic status; CHD, coronary heart disease.

(C case), was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each center by Professor Bradford Hill; the details of the series were unknown to any of the investigators or to the co-coordinator and were contained in a set of sealed envelopes each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel and before admission to the streptomycin centre, the appropriate numbered envelope was opened at the central office; the card inside told if the patient was to be an S or C cases, and this information was then given to the medical officer at the centre (p769).<sup>17</sup>

Clearly, there were shortcomings of this trial such as the lack of blinding and lack of informed consent. Bradford Hill was later knighted for his contributions to science, including the contribution of randomization.

### INTENTION TO TREAT ANALYSIS

There are 3 general analytic approaches in clinical trials: analysis as randomized (referred to as intention to treat analysis, or ITT), compliers-only analysis (in which only those patients randomized to a treatment arm who completed the trial and complied with treatment are analyzed), and as-treated analysis (in which only those who received a given treatment are counted, whether or not the patient was initially assigned to that treatment). Intuitively, it makes little sense to include in an analysis patients who did not receive or comply with the treatment (intervention or control)

throughout the period of observation. Intention to treat goes against this intuition because, simply stated, it says once randomized, always analyzed. Thus, a patient in a 5-year clinical trial, who is randomized to an intervention group and drops out of the study during the first week of observation, is analyzed as though he or she received the intervention throughout the study. Why is it then that ITT is the gold standard analytic method? The answer: it is the only analysis that preserves randomization. We have already discussed the critical role that randomization plays in terms of reducing (eliminating) confounding. If a randomized patient is not counted (analyzed), the integrity of randomization is compromised. Thus, there might be (and probably usually is) something different about a patient in the investigational arm who drops out of a study versus one in the control arm who drops out of a study, and this factor could confound the results. Recall that at the beginning of the study, randomization balances potential confounding factors; however, if a confounder is associated with continued participation, then omitting those who fail to participate will reintroduce imbalance on this confounding factor. Thus, withdrawals during the course of a trial could jeopardize the scientific integrity of the trial if ITT were not the primary analysis scheme. This would hold true even if withdrawals from both treatment arms can be shown to be comparable, because this would not account for the comparability of unknown factors. It is also true that ITT will generally dilute the real difference between the investigational and comparator groups, meaning that ITT renders the most conservative result. On the other hand, if upon using the ITT analysis, there is an observed difference between the intervention and control group, one can be more certain that that difference is real.

Some examples of the above principals follow. The Coronary Drug Project compared clofibrate with placebo in patients with a previous myocardial infarction, and it was initially reported that clofibrate was a favorable intervention.<sup>18</sup> However, results were presented according to complier-only analysis and analysis by treatment received (there was a 15% 5-year mortality in the good-compliers group compared to 19.4% mortality in the placebo group;  $P < .01$ ). This analysis was further supported by the fact that there was 24.6% mortality in the poor-compliers group compared to the active-therapy group. When ITT was used as the analysis, the clofibrate group mortality was 18.2% compared to the placebo mortality of 19.4% ( $P < .25$ ). Another interesting finding in this study was that the good-compliers group compared to the placebo group had a 15.1% mortality and the poor-compliers

group compared to the placebo group had a 28.2% mortality, supporting the concept that differences (confounding) might well exist in poor versus good compliers, irrespective of the treatment received. In the Anturane Reinfarction Trial, a favorable outcome was reported for anturane versus placebo in the reduction of myocardial infarction (actually the  $P$  value was .07).<sup>19</sup> In the 1629 patients, 816 were randomized to placebo and 812 to anturane. It was subsequently determined that 71 of the randomized patients did not actually meet the eligibility criteria and had been excluded from the initial analysis. To maintain the scientific integrity of the study (and thereby to maintain randomization), ITT was used for all randomized patients, and the  $P$  value was .20. The above 2 therapies were never approved for their respective indications. To further support the use of ITT, it was later determined that the mortality in the anturane-ineligible patients was 26%, whereas it was only 9% in those who were eligible (again supporting the concept of differential confounding).

#### SELECTION OF INTERVENTIONAL AND COMPARISON GROUPS

Although sometimes studies assess a new active (investigational) treatment versus an approved (standard) active treatment (ie, to assess if the old, standard treatment should be replaced with the new treatment), in other cases, studies are assessing if a new treatment should be added (not replacing, but rather supplementing, current treatment). In this latter case, the comparison of interest is the outcome of patients with and without the new treatment. In this instance, masking can only be accomplished by the use of a double-blind technique. Traditionally, placebo treatment has been used as the comparator to active treatment and has been one of the standards of clinical trials.

The use of the placebo has more and more been the subject of ethical concerns. In addition to ethical issues involved with the use of placebos, there are other considerations raised by the use of placebo controls. For example, an important lesson was learned from the Multiple Risk Factor Intervention Trial (MRFIT) regarding the use and analysis of the placebo control group, which might best be summed up as why it is important to watch the placebo group.<sup>20</sup> MRFIT screened 361 662 patients to randomize high-risk participants (using the Framingham criteria existent at that time) to special intervention ( $n = 6428$ ) and usual care ( $n = 6438$ ) with coronary heart disease mortality as the end point. The design of this well-conducted study assumed that the risk factor profile of

those receiving special treatment interventions would improve, whereas those patients in the usual care group would continue their current treatments and remain largely unaffected. The special intervention approaches in MRFIT were quite successful, and all risk factor levels were reduced. However, there were also substantial and significant reductions observed in the control group. That both treatment groups experienced substantial improvements in their risk factor profile translated to almost identical coronary heart disease deaths during the course of the study. Why did the control group fare so well? Several phenomena may have contributed to the improvement in the placebo control group. First is the Hawthorne effect, which suggests that just participating in a study is associated with increased health awareness and changes in risk factor profile, irrespective of the intervention. In addition, for the longer term trials, there are changes in the general population that might alter events. For example, randomization in MRFIT was conducted during the 1980s, a period when health awareness was becoming more widely accepted in the United States and likely beneficially affected the control group.

Although the ethics of placebo controls is under scrutiny, another principal regarding the placebo control group is that sometimes being in the placebo group is not all that bad. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study was launched in 1994.<sup>21</sup> By the early 1990s, there was mounting clinical epidemiologic evidence of reduced cancer risk associated with higher intake of antioxidants. Treatment with vitamin E and beta-carotene were considered unlikely to be harmful and likely to be helpful, and the question was asked whether antioxidants could reduce lung cancer even in smokers. A double-blind, placebo-controlled RCT was launched with a 2 × 2 factorial design and more than 7000 patients in each cell. No benefit was seen with either therapy, but compared to placebo, a disturbing worsening trend was seen in the beta-carotene-treated group.

## SELECTION OF END POINTS

The choice of which end point(s) to select is critical to any study design. Two additional areas require particular attention: the use of surrogate measures and the use of composite end points.

It has been said that death is a fact, the rest is inference. The ability to definitively determine study end points ranges across a spectrum from definitive end points (such as death or myocardial infarction) to end points that are much more subjective (such as

angina frequency or quality of life). However, the use of surrogate measures (measures that stand in for the true outcome of interest) is common in clinical research, for a number of reasons. A surrogate measure is an outcome (end point) that is used as a substitute for the real end point you would like to choose but cannot. Examples include the use of blood pressure reduction as a surrogate for hypertensive stroke or fasting blood sugar (or HA1c) as a surrogate for diabetic complications. Surrogate measures generally allow for a reduced sample size and a shorter follow-up period (and thereby cost) of a trial. The problem is that one must be assured that the effect of the intervention on the surrogate end point completely (or at least nearly completely) reflects changes in the true end point of interest. A classic example of the choice of a surrogate end point being problematic was the Cardiac Arrhythmia Suppression Trial (CAST).<sup>22</sup> Preceding this study, there was evidence that premature ventricular contractions were a marker for ventricular arrhythmias and sudden cardiac death. The evidence further supported the fact that therapy (antiarrhythmic therapy) was available to reduce premature ventricular contractions and thereby was likely to reduce sudden cardiac death. Therefore, CAST was designed to test the effect of antiarrhythmic therapy on mortality. Amid great ethical debate, a placebo comparison group was ultimately chosen. The results of CAST demonstrated that despite a reduction in premature ventricular contractions, mortality was actually worse in the active compared to the placebo treatment group. In general, the main limitation in the use of a surrogate outcome is the dependency on an unproven assumption that there is a 1:1 (or nearly so) linkage that connects a change in the surrogate outcome and the accepted clinical outcome. As Psaty et al have said, "to use only a surrogate end point is to accept as empirical evidence for clinical practice a hypothesis about health benefits that has never been tested" (p788).<sup>23</sup>

The use of composite end points is also increasingly used in clinical trials. This is the result of the overall reduction in morbidity and mortality related to modern-day therapy, thereby reducing the occurrence of clinical events. Because events drive the sample size of a study, the use of a composite of events as the primary outcome is popular. The problem occurs when one event in a composite is beneficial but several other events are either neutral or adverse. The overall result of the study may then be misleading. The assumption made when composite end points are used is that all components of the composite have equal importance, with similar risk reductions and similar frequency. For example, let us say

that a study is using a composite end point of death, myocardial infarction, and urgent revascularization, and the composite end point demonstrates a statistically significant benefit. However, further analysis suggests that although there is a statistically significant decrease in the need for urgent revascularization (and this was the most common outcome in the composite cluster), there is a nonstatistically significant increase in myocardial infarction (but this is a much less common occurrence), and death is unaffected. What conclusions can be drawn from such a study?

### INTERPRETATION OF RESULTS

Interesting to consider and important to reemphasize is why intelligent people can look at the same data and render differing interpretations. MRFIT is again exemplary of this principal and demonstrates how misinterpretation can have far-reaching effects. One of the conclusions from MRFIT was that reduction in cigarette smoking and cholesterol was effective, but “possibly an unfavorable response to antihypertensive drug therapy in certain but not all hypertensive subjects” (p1465) led to mixed benefits.<sup>20</sup> This possible unfavorable response ultimately has been at least questioned if not proven to be false.

The above principal was also seen in the interpretation of the alpha-tocopherol, beta-carotene cancer study.<sup>21</sup> To explain the lack of benefit and potential worsening of cancer risk in the treated patients, the authors opined that perhaps the wrong dose was used or the intervention period was too short, because “no known or described mechanisms and no evidence of serious toxic effects of this substance (beta carotene) in humans” (p1036) has been observed.<sup>21</sup> This points out how one’s personal bias regarding the intervention can influence one’s shaping of the interpretation of a trial’s results. Finally, there are many examples of trials in which an interpretation of the results is presented, but after publication, differing interpretations are rendered. Just consider the recent controversy over the interpretation of the ALLHAT results.<sup>24</sup>

### TRIAL DURATION

An always critical decision in performing or reading about a RCT (or any study for that matter) is the specified duration of follow-up to come to a meaningful outcome. Many examples and potential problems exist in the literature, but basically in interpreting the results of any study (positive or negative), the question, What would have happened had a longer follow-up period been chosen? should be asked. A recent

example is the Canadian Implantable Defibrillator Study (CIDS),<sup>25</sup> which was a RCT comparing the effects of defibrillator implantation to amiodarone in preventing recurrent sudden cardiac death in 659 patients. At the end of the study (a mean of 5 months), a significant difference was evident in all-cause mortality when comparing the 2 treatment regimens. At one center, it was decided to continue the follow-up in 120 patients who remained on their originally assigned intervention for an additional 5.6 years.<sup>26</sup> All-cause mortality was then found to be increased in the amiodarone group. The Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) trial is an example of a potential problem in which study duration could have been problematic.<sup>27</sup> The central hypothesis of MIRACL was that early rapid and profound cholesterol lowering therapy with atorvastatin could reduce early recurrent ischemic events in patients with unstable angina or no-Q wave acute infarction. Often with acute intervention studies, the primary outcome is assessed at 30 days after the sentinel event. In the MIRACL trial, there was no difference in the primary outcome at 30 days. Fortunately, the study specified a 16-week follow-up, and a significant difference was seen: cumulative incidence of death (any cause), nonfatal myocardial infarction, resuscitated cardiac arrest, or worsening angina with new objective evidence requiring urgent rehospitalization at 16 weeks was 17.4% with placebo and 14.8% with atorvastatin. Finally, an example from the often cited controversial ALLHAT study, which demonstrated a greater incidence in new diabetes in the diuretic arm as assessed at the study end of 5 years.<sup>24</sup> The investigators, however, pointed out that this did not result in a difference in outcomes in the diuretic versus other treatment arms. Many experts have opined that the trial duration was too short to assess adverse outcomes from diabetes, and had the study gone on longer, it is likely that a significant difference would occur negating the study interpretation that diuretics appeared as safe and effective as newer antihypertensive modalities.

### SELECTION OF TRADITIONAL VERSUS EQUIVALENCE TESTING (TABLE I)

Most clinical trials have been designed to assess whether there is a difference in the efficacy to 2 (or more) alternative treatment approaches (against the null hypothesis of no differences between treatments with the comparator treatment to the new treatment traditionally being placebo). There are reasons placebo controls are preferable to active controls, not the least

**Table I** The Types of Randomized Controlled Trials (RCTs) and Their Relationship to Hypothesis Testing

RCT Type	Null Hypothesis	Alternative Hypothesis
Traditional	New = old	New $\neq$ old (ie, new < old or new > old)
Equivalence	New < old + $\delta$ (where $\delta$ is a “cushion”; that is, that the new is at least $\delta$ worse than the old)	New $\geq$ old + $\delta$
Noninferiority	New $\not<$ old	New = old

of which is the ability to distinguish an effective treatment from a less-effective treatment. However, if a new treatment is considered to be equally effective but perhaps less expensive and/or invasive or a placebo control is considered unethical, then the new treatment needs to be compared to an established therapy and would be considered preferable to an older established therapy, even if it is just as good (not necessarily better) as the old. The ethical issues surrounding the use of a placebo control and the need to show a new treatment to be only as good as (rather than better than) has given rise to a recent interest in equivalence testing. With traditional hypothesis testing (ie, superiority trials), the null hypothesis states that there is no difference between treatment groups (ie, new = old or placebo or standard therapy). Rejecting the null then allows one to definitively state whether one treatment is better than another (ie, new > old or new < old). The disadvantage is if at the conclusion of an RCT there is not evidence of a difference, one cannot state that the treatments are the same or as good as one to the other. That is, when the null hypothesis is not accepted, it is simply the case whereby it cannot be rejected. The appropriate statement when the null hypothesis is not rejected (accepted) is there is not sufficient evidence in these data to establish if a difference exists.

Equivalence testing in essence flips the traditional null and alternative hypotheses. Using this approach, the null hypothesis is that the new treatment is worse than the old treatment (ie, new < old); that is, rather than assuming that there is no difference, the null hypothesis is that a difference exists and the new treatment is inferior. Just as in traditional testing, the 2 available actions resulting from the statistical test are (1) reject the null hypothesis or

(2) failure to reject the null hypothesis. However, with equivalence testing, rejecting the null hypothesis makes the statement that the new treatment is not worse than old treatment, implying that the alternative is that the new treatment is *as good as* or better than the old (ie, new  $\geq$  old). Hence, this approach allows a definitive conclusion that the new treatment is as good as the old.

One caveat is the definition of *as good as*, which is defined as being in the neighborhood or having a difference that is so small as to be considered clinically unimportant (generally, event rates within  $\pm 2\%$ ; this is known as the equivalence or noninferiority margin usually indicated by the symbol  $\delta$ ). The need for this “neighborhood” that is considered as good as exposes the first shortcoming of equivalence testing—having to make a statement that “I reject the null hypothesis that the new treatment is worse than the old and accept the alternative hypothesis that it is as good or better – *and by that I mean that it is within at least 2% of the old*” (the wording in italics are rarely included in the conclusions of articles). A second disadvantage of equivalence testing is that no definitive statement that there is evidence that the new treatment is worse can be made. Just as in traditional testing, one never accepts the null hypothesis; one only fails to reject it. Hence, all one can really say is that there is no evidence in these data that the new treatment is as good as or better than the old treatment. Another problem with equivalence testing is that one has to rely on the effectiveness of the active control obtained in previous trials and on the assumption that the active control would be equally effective under the conditions of the present trial.

An example of an equivalence trial is the Controlled Onset Verapamil Investigation of Cardiovascular Endpoints (CONVINCE) trial, a trial that also raised some ethical issues that are different from those usually involved in RCTs.<sup>28</sup> CONVINCE was a large, double-blind clinical trial intended to assess the equivalence of verapamil and standard therapy in preventing cardiovascular disease-related events in hypertensive patients. The results of the study indicated that the verapamil preparation studied was not equivalent to standard therapy because the upper bound of the 95% confidence limit (1.18) slightly exceeded the prespecified boundary of 1.16 for equivalence. However, the study was stopped prematurely for commercial reasons. This factor not only hobbled the findings in terms of inadequate power, but it also could be interpreted to mean that participants who had been in the trial for years were subjected to a breach in contract. That is, they had subjected themselves to

the risk of a RCT with no benefit. There was a good deal of criticism borne by the pharmaceutical company involved in the decision to discontinue the study early. Parenthetically, the company involved no longer exists.

Another approach is noninferiority testing. Here the question is again slightly different in that one asks whether the new intervention is simply not inferior to the comparator (ie, new  $\nless$  old). One advantage is that statistical significance would be only 1-tailed because there is no implication that the analysis addresses whether the new treatment is better, only that it is not inferior. Weir et al used this approach in evaluating a comparison of valsartan/hydrochlorothiazide (VAL/HCTZ) with amlodipine in the reduction of mean 24-hour diastolic blood pressure.<sup>29</sup> Noninferiority of the VAL/HCTZ combination to amlodipine was demonstrated, and fewer adverse events were noted with the combination. The null hypothesis for this analysis was that the reduction in mean 24-hour diastolic blood pressure from baseline to the end of the study with VAL/HCTZ was  $\geq 3$  mm Hg less (the noninferiority margin) than that with amlodipine. Again, a caveat has been recently raised by Le Henanff et al<sup>30</sup> and Kaul et al.<sup>31</sup> Le Henanff et al<sup>30</sup> reviewed published studies listed as equivalence or noninferiority trials between 2003 and 2004 and noted a number of deficiencies, key among them being the absence of the equivalence margin.<sup>30</sup>

## CONCLUSIONS

Although RCTs remain the gold standard proof of efficacy, there are many aspects of trial design that must be appropriately incorporated to ensure the value of the study. The inappropriate use of any tool (including RCTs) compromises the ability to meaningfully interpret the resulting information. We have presented several aspects that a user of the information should consider when establishing the credence to attach to the information from a RCT.

## REFERENCES

1. Grimes DA, Schultz KF. An overview of clinical research: the lay of the land. *Lancet*. 2002;339:57-61.
2. Lascanzo J. Clinical trials in cardiovascular medicine in an era of marginal benefit, bias, and hyperbole. *Circulation*. 2005;112:3026-3029.
3. Hulley SB, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA*. 1998;280:605-613.
4. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA*. 2002;288:321-333.
5. Grady D, Herrington D, Bittner V. Cardiovascular disease outcomes during 6-8 years of hormone therapy. *JAMA*. 2002;288:49-57.
6. Grady D, Rubin SM, Petitti DB, et al. Hormone replacement therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med*. 1992;117:1016-1037.
7. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease. *Prev Med*. 1991;20:47-63.
8. Sullivan JM, VanderZwag R, Hughes JP, et al. Estrogen replacement and coronary heart disease in postmenopausal women. *Arch Intern Med*. 1990;150:847-850.
9. Bienenfeld L, Frishman W, Glasser S. The placebo effect in cardiovascular disease. *Am Heart J*. 1996;132:1207-1221.
10. Rothman KJ, Michels KB. The continuing unethical use of placebo controls (sounding board). *N Engl J Med*. 1994;331:394-398.
11. Clark PI, Leaverton PE. Scientific and ethical issues in the use of placebo controls in clinical trials. *Annu Rev Public Health*. 1994;15:19-38.
12. Montori VM, Devereaux PJ, Adhikari NKJ, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005;294:2203-2209.
13. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *N Engl J Med*. 1991;325:445-453.
14. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Endarterectomy for asymptomatic carotid artery stenosis. *JAMA*. 1995;273:1421-1428.
15. Mayberg MR, Wilson SE, Yatsu F, et al. Carotid endarterectomy and prevention of cerebral ischemia in symptomatic carotid stenosis: Veterans Affairs Cooperative Study Program 309 Trialist Group. *JAMA*. 1991;266:3289-3294.
16. Smith DHG, Neutel JM, Lacourciere Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM measurements. *J Hypertens*. 2003;21:1291-1298.
17. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J*. 1948;2:769-782.
18. The Coronary Drug Project Research Group. Influence of adherence to treatment and response to cholesterol on mortality. *N Engl J Med*. 1980;303:1038-1041.
19. Anturane Reinfarction Trial Research Group. Sulfapyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med*. 1980;302:250-256.
20. Multiple Risk Factor Intervention Trial Research Group. Multiple Risk Factor Intervention Trial: Risk factor changes and mortality results. *JAMA*. 1982;248:1465-1477.
21. Alpha-tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin e and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med*. 1994;330:1029-1038.
22. Greene HL, Roden DM, Katz RJ, et al. The cardiac arrhythmia suppression trial. *J Am Coll Cardiol*. 1992;19:894-898.
23. Psaty BM, Weiss NS, Furberg CD, et al. Surrogate end points, health outcomes, and the drug-approval process for the

treatment of risk factors for cardiovascular disease. *JAMA*. 1999; 282: 786-790.

24. The ALLHAT Collaborative Research Group. Major outcomes in high risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs. diuretic. *JAMA*. 2002;288:2981-2997.

25. Connolly SJ, Gent M, Roberts RS, et al. Canadian Implantable Defibrillator Study (CIDS): a randomized trial of the implantable cardioverter defibrillator against amiodarone. *Circulation*. 2000; 101:1297-1302.

26. Bokhari F, Newman D, Greene M, et al. Long-term comparison of the implantable cardioverter defibrillator versus amiodarone: eleven-year follow-up of a subset of patients in the Canadian Implantable Defibrillator Study (CIDS). *Circulation*. 2004;110: 112-116.

27. Schwartz GG, Olsson AG, Ezekowitz MD, et al. Effects of atorvastatin on early recurrent ischemic events in acute coronary syndromes: the MIRACL study: a randomized controlled trial. *JAMA*. 2001;285:1711-1718.

28. Black H, Elliott WJ, Grandits MS, et al. Principal Results of the controlled onset verapamil investigation of cardiovascular endpoints (CONVINCE) Trial. *JAMA*. 2003;289:2073.

29. Weir MR, Ferdinand KC, Flack JM, et al. A noninferiority comparison of valsartan/hydrochlorothiazide combination versus amlodipine in black hypertensives. *Hypertension*. 2005;46:508-513.

30. Le Henanff A, Giraudeau B, Baron G, Ravuad P. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA*. 2006;295:1147-1151.

31. Kaul S, Diamond GA, Weintraub WS. Trials and tribulations of non-inferiority. *JAMA*. 2005;46:1986-1995.