The Randomized Clinical Trial: Bias in Analysis

GRAHAM S. MAY, M.B., B. CHIR., DAVID L. DEMETS, PH.D.,
LAWRENCE M. FRIEDMAN, M.D., CURT FURBERG, M.D.,
AND EUGENE PASSAMANI, M.D.

SUMMARY The realization that bias in patient selection may influence the results of clinical studies has helped to establish the randomized controlled clinical trial in medical research. However, bias can be equally important at other stages of a trial, especially at the time of analysis. Withdrawing patients from consideration in the analysis because of ineligibility on account of study entry criteria, lack of compliance to the protocol, or data of poor quality may be a source of systematic error. Examples to illustrate the possible consequences are taken from trials in the cardiovascular field. We recommend that reported study results should include outcome data from all subjects randomized in the group to which they were originally assigned.

CLINICAL INFORMATION varies in reliability. The need for unbiased assessment has produced a variety of methods to evaluate new or established interventions, from the individual case report to the randomized controlled clinical trial. The randomized trial is often regarded as the standard against which other methods should be measured and has become an integral part of medical research. Our confidence in this tool, however, is appropriate only if we ensure that certain methodologic requirements are satisfied.

In this report we examine ways in which bias might be introduced during the analysis of trial results even if the design and conduct up to that time meet desired standards. Most of the examples come from multicenter trials, but the principles are applicable to the smaller, single-center endeavor. Our aim is not to formulate hard-and-fast rules, but to enhance the awareness of journal readers to the dangers of data omission. We wish to encourage critical appraisal of trial design and analysis rather than acceptance of results at face value.

Randomization is the key to proper trial design because it removes the potential for bias in the allocation of patients to different interventions, it tends to ensure that study groups are comparable with respect to both known and unknown prognostic factors, and it guarantees the validity of statistical tests of significance used to interpret the results. Randomization does not ensure the validity of the entire trial; it is only one of several measures that attempt to minimize systematic error and the associated benefits can easily be lost.

The aim in all clinical trials is to recruit patients who conform to specific inclusion criteria. This ideal is difficult to achieve, and a few inappropriate subjects are usually randomized. Furthermore, during the course of the study, some subjects stray from the requirements of the protocol and the result is a study sample that differs from what was originally intended. Two classes of subjects are often confused, namely exclusions and withdrawals. To distinguish them, we advocate general acceptance of the terminology used by Peto and colleagues. Exclusions are subjects who are initially considered as candidates for the study, but are found not to meet all the entry criteria; hence, they are never randomized. Withdrawals are subjects who were once deemed eligible for the study on all accounts and were randomized to one or another of the intervention groups; later, however, their inclusion in the study is felt to be inappropriate and they are deliberately omitted from analysis. Therefore, the primary distinction between the two classes is related to the time of randomization, exclusions occurring before and withdrawals afterwards. Prerandomization exclusions are not a problem because, as allocation to intervention never occurs, the internal validity of the study is unaffected. The more selected population may well restrict the inferences that can be made from the study results. Our concern is with postrandomization withdrawals as a source of bias.

During the course of a trial or at the time of analysis, enrolled patients may be withdrawn for various reasons. It may be decided that entry criteria were violated, compliance with intervention was less than expected, or the collected data were of inadequate quality. Whatever the justification, the three cardinal features of prior randomization are in jeopardy if withdrawals are permitted. Even if equal numbers of subjects are withdrawn from the intervention groups, there is no assurance that group comparisons remain valid, because the reasons for withdrawal may differ between groups. Subjects are not withdrawn randomly, although it may be shown that they are well-matched on a number of baseline characteristics. Therefore, in strict terms, the remaining subjects in the intervention groups can no longer be assumed comparable. Three common reasons for withdrawal are ineligibility of subjects, poor compliance and data of inferior quality.
Ineligible Subjects

Ineligible subjects are randomized but are later considered to have violated the entry criteria because of protocol misinterpretation, clerical error or because a definitive diagnosis was not available at the time of randomization. The decision to label a subject as ineligible may be based on data that are beyond question (e.g., age or sex) or, more often, on evidence that requires some degree of judgment (e.g., x-ray, ECG, or severity of symptoms). Withdrawals that are based on subjective assessment open the possibility for bias, especially if treatment assignment or outcome is known beforehand. Whether outcome is known for the trial as a whole or only for the subject in question, it should be assumed that bias weighed on the decision to withdraw.

The Anturane Reinfarction Trial (ART) was designed to compare the platelet-active drug sulfinpyrazone (Anturane) with a placebo control in patients who had suffered a myocardial infarction. Subjects were randomized into the study at several clinical centers. The data were collected and monitored centrally, the usual practice in large-scale studies. At subsequent review, it was decided that a number of patients should never have been randomized because they did not meet all the entry criteria. Based on the final report, 1,629 patients were randomized, 816 in the placebo group and 813 in the sulfinpyrazone group (table 1). Seventy-one patients who were judged not to fulfill the entry criteria were declared ineligible and disregarded in the published analysis. The reported difference in total mortality between sulfinpyrazone and control groups was not statistically significant (8.3% vs 10.9%; $p = 0.07$). If the more guarded approach of including deaths in all subjects (eligible + ineligible) had been adopted, total mortality would have been 74 of 813 (9.1%) for sulfinpyrazone and 89 of 816 (10.9%) for placebo ($p = 0.20$).

At first glance, the published results of ART might indicate that sulfinpyrazone was of some clinical benefit even though conventional statistical significance levels were not met. If we focus on patients withdrawn from the analysis (a matter of subtracting those patients analyzed from all those randomized), four of 33 omitted placebo patients died and 10 of 38 omitted sulfinpyrazone patients died ($p = 0.13$). In the placebo group there was little difference in the mortality of patients who were withdrawn and those who were analyzed (12.1% vs 10.9%), in contrast to the sulfinpyrazone group, in which mortality was almost three times greater in patients who were withdrawn than in those who were retained (26.3% vs 9.1%). This decision to withdraw subjects led to criticism of both study reports. The ART exemplifies how differences between groups can be changed if patients are withdrawn from a study after randomization, albeit in approximately equal numbers from each treatment group. It is the outcome of these subjects that is important. The 38 withdrawals from the sulfinpyrazone group have only a small effect on the denominator of any mortality equation because they are buffered by the large number of study participants. However, the 10 deaths in these withdrawn patients greatly affect the small numerator, and hence the level of significance.

Additional difficulties arise if the study intervention itself can possibly alter the diagnostic signs of the qualifying event. In the Multi-Institutional Study to Limit Infarct Size (MILIS), propranolol, hyaluronidase or placebo is given shortly after hospital admission for acute myocardial infarction. If, as supposed, myocardial damage starts soon after the onset of chest pain and becomes irreversible over 12–24 hours, intervention must be started early to have an effect. However, early treatment requires randomization before the diagnosis is confirmed; inevitably, some patients who later appear not to have had an infarction are enrolled. Moreover, if the interventions do succeed in limiting infarct size or preventing threatened infarction, they may affect subsequent ECG tracings and serum enzyme levels on which the conventional definitive diagnosis of infarction depends. Subjects with a small myocardial infarction in the intervention group may have reduced serial enzyme levels and therefore appear not to have had an infarction. This situation will not arise for those in the placebo group. If, at some later stage, the subjects are withdrawn from the study because they did not meet the criteria for myocardial infarction, more subjects would be withdrawn from the intervention groups (those with no infarction + those with small infarction), assuming that intervention is efficacious, than from the control group (those with no infarction only). Thus, MILIS allows no postrandomization withdrawals on diagnostic grounds.

There are three acceptable policies for dealing with withdrawals on the grounds of entry criteria violation. The first is to delay enrollment until all the diagnostic tests have been confirmed and the entry criteria have been carefully checked. No withdrawals are subsequently allowed. The second is to enroll presumptive or unconfirmed cases and later withdraw

---

**Table 1. Anturane Reinfarction Trial: Mortality**

<table>
<thead>
<tr>
<th>Patients</th>
<th>Sulfinpyrazone</th>
<th>Placebo</th>
<th>Significance of the difference ($p$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzed</td>
<td>64/775 (8.3%)</td>
<td>85/783 (10.9%)</td>
<td>0.07</td>
</tr>
<tr>
<td>Randomized</td>
<td>74/813 (9.1%)</td>
<td>89/816 (10.9%)</td>
<td>0.20</td>
</tr>
<tr>
<td>Ineligible*</td>
<td>10/38 (26.3%)</td>
<td>4/33 (12.1%)</td>
<td>0.13</td>
</tr>
</tbody>
</table>

*Ineligible patients excluded.
subjects who are shown to be misdiagnosed. The decision to withdraw should be made as early as possible after patient entry and be based on objective data collected before randomization. To further reduce bias, these data should be assessed without knowledge of study group or outcome. The third policy is to recognize that some subjects with an unconfirmed diagnosis will be enrolled, but to subsequently withdraw none of them even if the initial decision proves incorrect. This is always valid because two randomized groups are being compared, but somewhat conservative because each group includes subjects who may respond differently to the interventions.

We advocate a combined policy. Every effort should, of course, be made to establish the eligibility of subjects for entry into the study and randomization should be delayed as long as possible within the limits of the design. After randomization, no withdrawals should be allowed and the primary analyses should include all subjects enrolled. This does not preclude subgroup analyses on the basis of baseline data, including eligibility status. If the conclusions from analyses using all enrolled subjects and the subgroup analyses agree, then interpretation of the latter is easier. Should the results differ, one must be cautious in drawing firm conclusions from subgroups. In general, the analysis with all the enrolled subjects should be emphasized.

Compliance

Full patient compliance with study protocol, especially any prescribed treatment regimen, is essential if the maximum effects of any intervention are to be seen. Patient cooperation with the best-planned study is unpredictable and impossible to estimate in advance. In drug trials, medication may be taken irregularly, either because patients simply forget or because of unpleasant taste, the inconvenience of the treatment schedule, or the development of unwanted effects that are thought to be associated with the treatment. The investigator or the private physician may encourage default by an unwillingness to implement full protocol dose because of the occurrence or expectation of adverse effects. Whatever the reason, any action that an intervention might possibly have will be reduced. If compliance falls below given levels, in the worst case medication being avoided altogether, benefit certainly cannot be anticipated. If the objective of a study is to determine the actual effectiveness of a particular treatment, it is appealing not to count poorly compliant subjects in any final evaluation. Nevertheless, this distillation of the data is not without hazard.

The Coronary Drug Project (CDP) compared the effects on total mortality of five lipid-lowering drug regimens with a placebo control group in subjects who had suffered a myocardial infarction. In patients who received clofibrate, total mortality for the 5-year period was 20%, compared with 20.9% in the placebo group.**14** Compliance was determined by pill count; the patients who took 80% or more of the protocol dose were considered good compliers. As reported by the CDP Research Group,**12** overall mortality for those in whom compliance could be assessed was 18.2% in the clofibrate group and 19.4% in the placebo group. In the treatment group the mortality of the good compliers (15.0%) was considerably less than mortality in the poor compliers (24.6%). This might seem to suggest that clofibrate was of some benefit to those who actually took it, until it is realized that a similar difference is also seen in those receiving placebo; placebo adherers fared considerably better (mortality 15.1%) than placebo nonadherers (28.2%). In such a case it is tempting to compare the good adherers between treatment groups, but this can lead to misinterpretation.

When β-blocking drugs are given at the time of myocardial infarction, their cardiodepressant action may cause potentially serious unwanted effects, such as hypotension and bradycardia. Subjects who develop these effects would appear to be at higher risk of dying than those who do not. In a recent double-blind trial of propranolol and atenolol,**18** more patients in the β-blocker groups than in the placebo control group developed these particular adverse effects, as expected, and had their medication stopped. Despite this, patients discontinued treatment almost equally in all three groups because a disproportionate number were withdrawn from the placebo group for nonspecific reasons. The results of this study illustrate the problem of withdrawing patients, even if the number withdrawn in each group is the same. Overall mortality was analyzed at 6 weeks and included every patient randomized; there was no statistically significant difference between the three groups (propranolol 7.6%; atenolol 8.7%; placebo 11.6%), but mortality in those patients who stopped taking either of the β blockers (propranolol 15.9%; atenolol 17.6%) was over four times greater than in those who continued (propranolol 3.4%; atenolol 2.6%). The development of serious adverse effects in patients on β-blocker treatment appears to have identified those already at high risk. There was no difference in those who stopped (12.5%) or continued (11.2%) to take the placebo tablets. This reinforces the view that comparison of good compliers between groups is inappropriate. A less cautious approach to analysis would have omitted the noncompliant patients and thereby biased the results in favor of the β-blocking group by excluding those at high risk.

Conflicting results can arise even if patients are rejected on the basis of compliance measures that are more objective than pill counts. The ultimate would be to determine if the intervention actually had the physiologic effect that is believed to confer benefit, although this is the result of two factors that are not necessarily related, namely the quantity of medication actually taken and the response of the individual to it. If anticoagulants are beneficial in patients who have had a myocardial infarction, this is presumably due to their antithrombotic properties. Wasserman et
al.14 conducted one of the randomized trials of anticoagulant therapy and found that although mortality was lower in those receiving warfarin (16%) compared with the control group (21%), the difference was not statistically significant. Results of subgroup analysis, based on the prothrombin time to indicate the degree of anticoagulation, are hard to explain in the context of the proposed mechanism of action. The mortality rate among patients who were inadequately anticoagulated was found to be lower (13%) than that among patients who were receiving, supposedly, therapeutic levels of warfarin (17%). In the CDP Research Group study,12 similar unexpected results were obtained when analyses were performed using cholesterol reduction as a measure of treatment efficacy.

In the extreme case a subject may not actually start medication before being withdrawn for one reason or another. In an open (unblinded) trial of the β-blocker alprenolol given to patients after a myocardial infarction, Ahlmark and colleagues16 randomized patients at the time of hospital admission to either the β-blocker or usual treatment. The start of medication, however, was delayed until 2 weeks later and during the intervening period 60% of the originally randomized 393 patients were withdrawn from the study, the major reasons being lack of confirmation of the infarction, death or contraindication to β blockers. One hundred sixty-two patients actually started therapy, 69 on alprenolol and 93 in the control group. This imbalance is statistically significant (p < 0.05), but even if it were not, doubts are raised about the comparability of the two groups at entry. The managing physicians were aware of future medication assignment. Therefore, if the condition of a patient was thought to be endangered by a β-blocking drug, he was only permitted to continue in the study if not assigned to alprenolol. To avoid this complication, it is wise to delay randomization, if possible, until treatment can be initiated immediately.

When discussing the issue of withdrawing poorly compliant subjects the distinction has been made between the "management" and "explanatory" trial.16,17 The former reflects the situation in clinical practice, focusing on the observed effect of intervention when this is prescribed within the bounds of normal doctor-patient relations, that is to say with a policy of "intention to treat." All patients randomized and all events, without exception, occurring during the observation period are counted and analyzed. Conversely, the explanatory trial concentrates on patients who complied with the protocol as planned, with the aim of elucidating the mechanism of action of an intervention. It is argued18 that in an explanatory trial, withdrawal of subjects is permissible, provided the rules for doing so are stated in advance, especially if the treatment group assignment is not known at the time of withdrawal. This does not take into account the possible interdependence of compliance and outcome. The potential for bias is still present, even if the intention to withdraw subjects is stated ahead of time.18 In the propranolol/atenolol study,19 subjects were not withdrawn, but if rules such as withdrawal for poor compliance had been specified in advance, this would in no way have lessened the important difference in observed outcome between groups. The explanatory policy may be appropriate under highly controlled conditions with a small number of hospitalized patients and with the sole intent to generate new hypotheses. We are concerned that the policy may be misapplied and we emphasize that inclusion of both good and poor compliers is the only assurance of valid group comparisons.

The policy of analyzing all randomized subjects in the groups to which they were originally assigned, so that the conclusions are not biased, does have an impact on the study design. The potential effect of the therapy may be diluted by noncompliant subjects, so the sensitivity, or ability to detect a real difference due to treatment, will be reduced. To compensate, the number of study subjects must be increased, the amount in relation to the degree of dilution. Methods have been proposed20-21 that attempt to model the effect of lack of compliance on the sample size, and invariably this results in increased sample size and cost. However, the alternative of withdrawing subjects in order to enhance treatment effects, and thereby to possibly introduce substantial bias, carries a more serious penalty — trial results that claim benefit when none exists.

Quality of Data

The conclusions of a trial can go no further than the quality of the data collected. These may fall short of desired standards on account of quantity, quality or both. Data may be missing entirely, incomplete, or, at worst, falsified. Throughout the trial monitoring should be of the highest standard so that missing or erroneous data can be spotted early and, when appropriate, corrected. Time-dependent data are irretreivable and must remain classified as missing. The extent of incomplete data may be used as one of the measures of clinic performance in a trial. In the analysis of any variable, the amount of information not available for consideration should be clearly stated.

In large trials, some subjects are certain to be lost to follow-up, and these may be regarded as enforced withdrawals. Every effort should be made to at least ascertain the essential data, e.g., vital status, on these subjects at some common point at the end of the study. The problem remains of what to do if even this limited measure is unsuccessful. Survival analysis is not restricted to mortality data and has the advantage that the information that is available, namely, the status of the patient at the last time of contact, can at least be used even if subsequent data are missing, although this assumes that the loss to follow-up is independent of treatment. For other parameters, more contrived analyses may be tried by replacing missing data with values that are only an approximation of the truth. For example, the value that was last observed, the mean of preceding values or a value extrapolated from existing data, might be substituted for informa-
tion that is missing. Whatever method is chosen, if the number of subjects lost to follow-up represents a large proportion of the total, doubt remains as to whether the overall result might have been altered if the missing information had been available. For this reason, extreme case analyses should be done, assigning both the most optimistic and pessimistic values for the missing data. If the conclusions are the same in each case, then despite the incomplete data set, firm conclusions are warranted. However, it is unusual for an analysis of this kind not to have some impact on the conclusions. Therefore, every effort must be made to keep loss to follow-up and missing data to a minimum.

In some multicenter trials, a special variation of this problem has occurred: One or more of the clinics involved has been found to be performing inadequately in terms of data quality.22,23 Patient follow-up at the involved center was stopped in each instance and the data ignored in the final analysis. As outcome was unknown and randomization of patients had been stratified by clinic, an advisable precaution in any multicenter trial, these omissions probably did not affect the overall conclusions. Nevertheless, it does indicate that, at least in the involved centers, the study requirements had been poorly understood.

Data that are present but of suspected validity, either because they are inconsistent with known information or because they lie well outside prespecified boundary values, are a special problem. Laboratory measures, where a repetition of the test at some later date is often inappropriate, fall into this category. Two rules should always be observed: No material alterations to the recorded data should be allowed without documented justification and no editing or suppression of the data should be permitted unless exact details are recorded and reported in any analysis. An outlying value should never be ignored because it might be an accurate finding and therefore important. To account for this, the analysis of any variable should be done with and without wild observations; a similar result with both alternatives makes interpretation from the restricted data set justified. If the results differ, the analysis including all the observations should take precedence.

We advocate that all the randomized patients in a clinical trial be accounted for as fully as possible and included in the analysis in at least part of the presentation. Bias can easily be introduced and should always be suspected if subjects are withdrawn, especially if the decision relies on subjective assessment of data obtained after randomization. Subgroup analyses using prerandomization characteristics are justified if they are limited in number, preferably specified in advance, and are presented as being secondary to the overall study conclusions. Subgroup analyses based on post-randomization variables, subject compliance being the most common example, are ill-advised. Readers of journals must be cautious and satisfy themselves that at no stage of a trial has randomization, and thus the comparability of study groups, been compromised.

References

10. Multi-Institutional Study to Limit Infarct Size: Protocol. Cardiovascular Diseases Branch, Division of Heart and Vascular Diseases, National Heart, Lung, and Blood Institute, Bethesda, Maryland, 1978