Epidemiologic Reviews

Vol. 24, No. 1
Printed in U.S.A.

Design of Randomized Trials

Sylvan B. Green[1,2]

INTRODUCTION

When considering any medical study, we have to keep in mind two issues related to participant (patient) heterogeneity: the effect of chance and the effect of bias (1). These issues are addressed by having adequate numbers of study participants (i.e., an adequate sample size) and by using randomization for intervention (treatment) assignment. Randomized trials are recommended to achieve a valid determination of the comparative benefit of competing intervention strategies, whether for prevention, screening, treatment, or management.

To make the following discussion generically applicable to all of these settings, the individuals who will be enrolled into the trial will be referred to as "participants" and the experimental conditions as "interventions." Clearly, in the usual clinical trial, these are "patients" and "treatments," respectively. In a prevention or screening trial, the participants may be drawn from the "normal," healthy population, or they may be selected because they are "high risk" based on some known or putative risk factors. There is a continuum from healthy to elevated risk to precursor abnormality or preclinical condition to frank disease (and, of course, disease itself may have a continuum of stages), so distinctions here can be somewhat arbitrary. Thus, many of the aspects of the design of randomized trials apply across the board. The location of the study along the disease continuum is relevant to issues such as the population to which the study applies, the interventions that are appropriate, and the sample size and duration of the study (both of which are usually greater for prevention and screening trials than for treatment trials).

While this article deals with randomized trials, this subject should be considered in the context of the broad range of clinical trials and intervention studies. Although it is important not to overrely on categorization of clinical trials, it is useful to consider the concepts embodied in the description of clinical trials by "phase." This categorization may vary depending on the disease area; it is quite common in cancer studies (2), on which the following discussion focuses as an example. Phase I describes the "formulation" stage of the clinical trial process, whose major objective is to investigate dosage and route of administration of a new intervention and to determine toxicity. In many situations, this stage involves determining the maximally tolerated dose by using a dose escalation scheme in successive patients or groups of patients. Because of the need to progress sequentially to different dosing regimens, phase I trials in diseases such as cancer usually are not randomized. In a number of disease areas, in contrast to cancer, phase I trials may be conducted in healthy volunteers.

The usual objective of phase II trials is to look for evidence of "activity" of an intervention. Such trials may use as an outcome measure some indication that the intervention is having a desired effect (e.g., in cancer trials, shrinkage of a measurable tumor would indicate promise as a treatment; a favorable effect on a biomarker in persons at risk of malignancy might indicate promise as a preventive agent). Customarily, phase II trials in diseases such as cancer do not involve a control group. However, randomization may be of value in some phase II settings. For example, if several new agents are under investigation for a given group of patients, we make a much better decision regarding which agent(s) to advance to subsequent phase III trials if bias in selection of patients is avoided by using randomized assignment.

Phase III trials are designed to investigate the "efficacy" of interventions. They involve randomized comparison of intervention approaches, which may be an active intervention versus nothing, an active intervention versus placebo, or one intervention versus another. These trials should involve a "definitive" endpoint; the choice of the appropriate endpoint for comparing efficacy depends on the disease area in question.

The term "phase IV" is used less commonly but refers to studies that evaluate the "effectiveness" of proven interventions in wide-scale use. Sometimes these are uncontrolled studies, perhaps part of postmarketing surveillance by the industrial provider of the intervention. However, randomization should be encouraged in this setting; such phase IV trials may be conducted in the context of large, simple trials (discussed below).

Designs of nonrandomized phase I and phase II trials are covered well elsewhere (2). This article focuses on comparative randomized trials. The advantages of randomization are well known (3–6). With randomization, bias (whether conscious or unconscious) is avoided. Predictive factors,

Received for publication January 3, 2002, and accepted for publication May 3, 2002.

Abbreviations: DSMB, data and safety monitoring board; PSA, prostate-specific antigen.

[1] Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH.
[2] Present affiliation: Arizona Cancer Center, University of Arizona, Tucson, AZ.
Reprint requests to Dr. Sylvan B. Green, Arizona Cancer Center, 1515 North Campbell Avenue, P.O. Box 245024, Tucson, AZ 85724-5024 (e-mail: sgreen@azcc.arizona.edu).

known and unknown, tend to be balanced between intervention and comparison groups. In addition, randomization provides a valid basis for statistical tests of significance. Having a concurrent comparison group controls for time trends. For all of these reasons, randomized trials are able to achieve two goals: 1) validly determining the comparative benefit of competing intervention strategies and 2) convincing the community of the results. Obviously, randomization is not in itself sufficient reason to be convinced of the study conclusions, but the increased strength of evidence provided by randomization is an important factor when evaluating the results.

Certainly, observational (nonrandomized) studies are useful, as demonstrated by important epidemiologic studies, for example. However, nonrandomized studies are not recommended for comparing interventions. Unlike epidemiologic investigations of possible etiologic factors, in which adjustment for confounding factors often seems quite sensible, bias in intervention assignment is inherently part of the practice of medicine and public health, and to hope that it can be simply adjusted away is no more than wishful thinking (5). Thus, in comparative phase III trials, randomization is essential.

CONCEPTS IN THE DESIGN OF RANDOMIZED TRIALS

The process of designing a randomized trial can be conceptualized as answering five questions. These questions are discussed in the paragraphs that follow.

What?

Conducting a randomized trial implies that we are comparing two or more intervention groups with regard to outcome. The first question is what is being compared with what? Most of this article focuses on a two-group (often called "two-arm") trial, in which participants are randomly allocated to one of two intervention conditions. This simple structure can encompass a variety of different situations, including the following possible types of comparisons (as well as others):

- A new experimental intervention versus nothing (i.e., one of the intervention conditions is "no intervention").
- A new experimental intervention versus placebo.
- One intervention versus another (e.g., comparing a new experimental intervention with a "standard" intervention, or comparing two alternative, commonly used interventions with each other).
- One intervention versus the same intervention plus something else (e.g., testing the effect of adding something new to an existing standard regimen).
- An intervention at a specified dose (or duration or intensity) versus the same intervention at a higher dose (or longer duration or greater intensity).
- lntervention now versus later, perhaps for only those participants who experience a certain event (e.g., adjuvant therapy for cancer patients versus treating only those whose cancer recurs later).

In any of these situations, the term "intervention" may refer to a drug (or drug regimen), a surgical procedure, a medical device, a therapeutic modality (radiation, biologic therapy, etc.), a micronutrient, a diet, a behavioral intervention, or a clinical approach to diagnosis, symptom management, or palliative care. All of these situations involve a choice between two alternative approaches, and we are uncertain as to which one is preferable. This uncertainty involves balancing both potential beneficial and possible adverse outcomes. Thus, in planning a trial, it is important to have an honest assessment of what we do know and do not know about the net benefits of any intervention in a given situation.

When designing a trial, it is often tempting to include more than two intervention groups, leading to a multiarm trial. Sometimes doing so is quite reasonable, but one must proceed cautiously. It is important to justify each intervention group to assure that an important question is truly being assessed. A larger sample size will be required for the trial; is the cost justified, and are adequate numbers of participants available? If the trial does include more than two arms, are all pairwise comparisons of interest? A special form of multiarm trial involves factorial designs, discussed later in this article.

Which?

Obviously, the value of a trial depends to a great extent on the specific intervention(s) being tested. Presumably, a comparative randomized trial would have been preceded by earlier studies that formed the basis for selecting the intervention(s), including deciding on the method of administration and identifying possible beneficial as well as adverse outcomes. It is important that an intervention be well defined; however, this requirement does not eliminate the possibility of flexibility. The extent to which this is reasonable depends on the nature of the intervention and its state of development. With a new experimental intervention (e.g., a new drug), there are advantages to investigating a relatively tightly controlled regimen. When investigating interventions already in general use, one might want to compare alternatives as they are given in practice, where heterogeneity is realistic and increases the generalizability of the result. It is often said that what makes a randomized trial ethical is the presence of uncertainty, that we do not know which, if any, of two (or more) alternatives is more effective; such uncertainties not only justify the trial but also form the basis for its value to the medical and public health communities. As noted above, this uncertainty involves balancing both beneficial and adverse outcomes. Thus, in the appropriate settings, trials of flexible regimens can enable evaluation of a strategy that represents an interventional approach, more than just a specific drug. Think of the trial question as being similar to a fork in the road, where the clinician or public health professional must make a decision. The randomized trial provides an unbiased assessment of the consequences of proceeding down one path compared with another. The answers to such questions enable us to make appropriate choices to achieve better outcomes.

Why?

The objectives of the trial need to be defined clearly. Although this statement may seem obvious, designs of trials may suffer because they lack focus. Contemplating the reasons for conducting the trial will help select outcome measures. It is often said that a trial should have a single primary outcome measure and a limited number of secondary outcome measures. This is generally good advice, although with care it is possible to evaluate more than one question in a single trial. However, it is important to balance the natural desire to get as much as possible from a trial against the risk of including too much, thereby jeopardizing the ability to obtain valid answers to any question.

Although sometimes the primary outcome measure for a trial will be a quantity measured on a continuous scale, many trials have either a dichotomous (binary) outcome or a time-to-event outcome. In many situations, the choice of the dichotomous outcome measure is readily apparent from the context of the trial: nonresponse versus response, recurrence versus no recurrence, incident case of disease versus disease free, dead versus alive in a specified time period subsequent to randomization, and so forth. In situations in which multiple measures of outcome are of interest, it may be useful to develop a composite assessment of "failure" versus "success" by which each participant will be categorized at the completion of the study. If an objective definition of success can be agreed upon in the design stage, such a dichotomous variable can be a very practical outcome measure for the trial. Although there is usually an interest in reporting the separate effects on different measures of outcome, often the ultimate goal is to decide at the completion of the trial which intervention approach to recommend in general, and a composite success-rate measure may suit this purpose.

Sometimes, even when separate outcome measures are of primary interest, a global index may be useful for summarizing outcomes and/or monitoring a trial. For example, consider a prevention trial in which an intervention may reduce the incidence of some diseases but increase the incidence of others; here, a global measure of health may be of interest in addition to results for a specific targeted disease. Such a global index has been investigated with regard to the Women's Health Initiative clinical trial (7). This trial is discussed later in this article.

When the length of time until an event occurs is more important than simply the occurrence of the event, then time-to-event may be the primary endpoint, such as time to failure, disease-free survival, or overall survival. If the effect of an intervention is to delay the occurrence of an event, then a time-to-event endpoint may provide greater power than simply looking at the incidence of the event during a fixed time interval following randomization. Statistical methods for censored survival data are applicable to these endpoints; such methods are discussed in textbooks (e.g., Marubini and Valsecchi (8)) and elsewhere in this issue of *Epidemiologic Reviews* (9). Not only do these methods deal with the situation in which some participants have not experienced the event by the time of data analysis, but these methods also can account for varying follow-up times of participants entered during a lengthy accrual period, and they incorporate the entire "survival" experience of the participants. These methods enable the proportion of participants experiencing the event to be estimated as a function of time.

In other situations, the functioning of each participant is graded at multiple points in time; examples of such measures include performance status, neurologic function, visual function, and quality of life. Various statistics may be of interest in such situations, such as the mean level of functioning over time, the rate of change over time (slope), the percentage of time spent above or below a certain threshold of function, or simply the probability of ever crossing a specified threshold. Statistical methods for longitudinal data analysis are often appropriate here.

Who?

It is essential to define eligibility for the trial. Considerations that enter into defining the study population include identifying persons with the potential to benefit, selecting participants in whom the effect can be detected, excluding those with an unacceptable risk of adverse effects, considering competing risk (e.g., excluding those likely to succumb to some other condition), and enrolling those participants considered likely to adhere to the intervention (10). The nature of the trial may determine how broad or narrow the eligibility criteria are. Possible arguments for restricting eligibility are to enhance statistical power by having a more homogeneous population, providing a higher rate of outcome events, and perhaps producing a larger potential benefit. Toxicity concerns may also increase the list of reasons for exclusion from the trial. Conversely, advantages arise from defining broad eligibility. The ability to recruit larger numbers may actually enhance statistical power while providing greater generalizability. Peto et al. (11) have advocated using the "uncertainty principle," whereby a randomized trial is open to participants for whom the choice of a recommended intervention (treatment) strategy remains uncertain. If a clinician and patient are reasonably certain that one of the options would be definitely inferior or otherwise inappropriate for that particular patient, the patient is not entered into the study; otherwise, the patient is randomly assigned to one of the intervention groups. In addition, broad eligibility criteria allow the benefits of trial participation to be available more widely across the population.

How many?

An essential part of any trial design is to determine the sample size and trial duration. When designing a randomized trial, one must plan for an adequate number of participants (sample size) to ensure the desired power to detect a particular intervention effect. To increase the power of a trial, we increase the sample size. Increasing the sample size increases the precision of the estimate of intervention effect. Expressed another way, increasing the sample size decreases

the variability (decreases the standard error) of the estimate, thus producing narrower confidence intervals. This topic is also discussed in textbooks (e.g., Friedman et al. (10)) and elsewhere in this issue (12, 13).

An important topic is the investigation of intervention effects in subgroups of participants. Inquiries about subgroups may arise when data from trials are being analyzed, but this issue should be considered earlier during the design stage. The first question that should be addressed is, Is it expected that the actual intervention effect may differ in a meaningful way between different subgroups? This question is of concern because apparent differences (in intervention effect across subgroups) can result from chance alone. The risk of spurious results increases with a greater number of subgroup analyses. Statistical tests can be performed to investigate whether differences in intervention effect between subgroups are consistent with chance alone (e.g., by testing for treatment-covariate interactions (14)); however, achieving adequate statistical power to formally test interactions requires a larger sample size. Therefore, when there is prior reason to suspect an important interaction, the trial should be designed to be large enough to investigate subgroups. Otherwise, the focus should be on the primary question(s); explore the data for subgroup interactions, but interpret cautiously, and limit the data-derived subset analyses to suggesting hypotheses for future study (1, 15).

FACTORIAL DESIGNS

The following discussion of factorial trials is taken from my recent summary of this topic (1). Factorial designs in randomized trials are sometimes appropriate and can lead to efficiencies by answering more than one question (addressing more than one comparison of interventions) in a single trial (16). The simplest design is the balanced 2×2 factorial, addressing two intervention comparisons: A versus not-A, and B versus not-B. Conceptually, participants are first randomized to A or not-A and then also to B or not-B. In effect, equal numbers of participants are randomly allocated to one of four intervention conditions: A alone, B alone, both A and B, and neither A nor B. Broadly speaking, the concept of not-A can be no intervention, placebo for A, or some standard intervention being compared with A. The intervention not-B is defined similarly. As an example, the Physicians' Health Trial randomized participants to aspirin versus placebo tablet, as well as to beta-carotene versus placebo capsule, with the goal of preventing coronary heart disease and cancer, respectively (17). Such factorial designs can be generalized to more than two dimensions, for example, a three-dimensional $2 \times 2 \times 2$ design. In addition, a dimension can have more than two options; for example, a 3×2 design could compare, for the "A" question, high-dose A, low-dose A, or not-A (1).

In the analysis of a balanced 2×2 factorial trial, the effect of A can be tested by comparing all participants randomized to A with all participants randomized to not-A. Of course, there are two strata of participants for this comparison, those randomized to B and those randomized to not-B. The effect of B can be tested analogously. Thus, all participants contribute to hypothesis testing for both questions. Factorial trials can be designed so that the process of data monitoring during the trial can lead to one question being answered (and that aspect of the trial stopped) while the rest of the trial proceeds. For example, in the aforementioned Physicians' Health Trial, during the follow-up period, the aspirin component was stopped early because of an observed benefit of aspirin (refer to reference (18) for more details), while the beta-carotene component continued until the planned end of the trial (19).

An interesting example of a three-dimensional factorial design is the ongoing Women's Health Initiative clinical trial (20, 21). This trial is studying approximately 68,000 postmenopausal women, looking at several major outcomes: cancer, cardiovascular disease, and osteoporosis-related fractures. The three dimensions (intervention comparisons) are dietary modification (a group-intervention program led by a nutritionist teaching a "low-fat dietary pattern") versus control (a standard packet of health promotion materials); hormone replacement therapy versus placebo; and calcium plus vitamin D supplementation versus placebo. An interesting feature here is what the designers of the study call a "partial factorial" design; I prefer the label "variable dimension factorial" (1). A participant is randomized to one or more dimensions of the factorial trial, depending on which of the randomized comparisons (dimensions) she is eligible for and also for which she is willing to consent. Thus, a woman with a contraindication to hormonal replacement therapy would not be randomized to the comparison of hormone replacement therapy versus placebo, nor would a women who either insisted on receiving hormone replacement or refused to receive it. However, such women could be randomized to dietary modification versus control and to calcium plus vitamin D versus placebo. The analyses of such a trial would focus on the valid randomized comparisons, with each participant included in the analyses of intervention effects for which she was randomly allocated (1).

Factorial designs are applicable when there is a genuine interest in more than one intervention question. It is important that the interventions can actually be given together (i.e., they are not known to interfere with each other, and the toxicity of the combined interventions does not reach unacceptable levels). Factorial designs make more sense when the mechanisms of action of the interventions in different dimensions are different. Otherwise, the combination of interventions may not produce a greater effect than that of either one alone. Factorial designs may be considered either when serious interactions between interventions are not expected or when information on interactions is of particular interest (1); in the latter case, however, a larger sample size may then be needed to obtain adequate statistical power to formally test the interaction effect. Although factorial designs should be reserved for situations in which they specifically can be justified, more use of factorial designs than is currently the case could be helpful in increasing the efficient use of clinical trial resources (16).

THE ROLE OF LARGE, SIMPLIFIED TRIALS

A compelling argument can be made that more use of randomized trials is needed to address areas of uncertainty in medicine. This argument leads to the call for more large, simple trials (11), when appropriate. Given patient heterogeneity and the play of chance, large numbers of patients are necessary to provide reliable estimates of the effect of treatment, especially when realistic effects are relatively modest in size (but still potentially of great public health importance). Depending on the setting and on the stage of development of new therapies, more or less complexity of trials may be indicated. Some settings require rigorously defined randomized clinical trials with tightly controlled eligibility criteria. However, the methodology of randomized trials can also be applied to wide-scale studies of practical effectiveness by implementing large, simplified trials with broad eligibility criteria. Simplicity of trials permits larger numbers of patients with a lesser expenditure of resources, enhanced reliability of data, and possibly greater generalizability.

As stated by Peto et al., "There is simply no serious scientific alternative to the generation of large-scale randomized evidence. If trials can be vastly simplified, as has already been achieved in a few major diseases, and thereby made vastly larger, then they have a central role to play in the development of rational criteria for the planning of health care throughout the world" (11, p. 39).

DESIGN FEATURES TO MAINTAIN THE INTEGRITY OF RANDOMIZATION

As noted above, there are compelling reasons to use randomized trials to compare alternative approaches to intervention. The goal is to have comparable groups and to avoid bias in the comparisons. Having committed the effort to undertake a randomized trial, it is essential to design the trial to maintain the integrity of the randomization process. Although some of these topics overlap with a discussion of trial conduct and analysis, their initial consideration is an inherent part of proper trial design.

Randomization procedures

The procedures for implementing the randomization should be unbiased, unpredictable (for the participants and for the study personnel who are recruiting and enrolling them), and tamperproof. The timing of randomization is important; it should be done after determining eligibility and as close as possible to the start of intervention. Carefully planning the design of the trial should avoid delays between these events, to minimize the probability of patients becoming noncandidates for starting intervention after randomization. Because all patients randomized will be included in an intention-to-treat analysis (as discussed below), we want to maximize the opportunity for each randomized patient to embark on the randomized assignment. If the trial is viewed as providing evidence on which of two (or more) alternative approaches to choose at some decision point, the allocation should be made as close as possible to the time point at which the competing approaches actually diverge.

In some situations—for example, in prevention trials in which beginning the randomly allocated intervention is not urgent—a relatively brief run-in period may be used. Here, all participants are started on the same intervention (perhaps a placebo, perhaps one of the study interventions whose short-term administration is not thought to have a lasting effect). Those who comply successfully during the run-in period are then randomized to one of the intervention groups in the study. While the use of a run-in period may hypothetically reduce the representativeness of the randomized study participants, it does answer the question pertaining to intervention effect among willing compliers, and having a group in which compliance is better may allow a smaller sample size for the actual randomized trial while maintaining statistical power. A good example is the Physicians" Health Study (17, 22); this topic is also discussed elsewhere in this issue (13).

Masking (blinding)

There is a spectrum of opinions on this subject, and only a brief discussion can be provided here. Certainly, there are advantages to the classic double-masked (double-blind) study, where neither the participants nor the study personnel who interact directly with the participants know the result of the randomized assignment. The feasibility of masking depends heavily on the nature of the intervention. When a drug is being compared with "no drug," creating a placebo will help maintain comparability of the groups. When two drugs are being compared, it may be possible to mask the identity of the drug being given, depending on the route of administration. In some of these situations, a "double dummy" approach can be used; for example, each participant takes two pills, one active and one placebo, where each drug has its own matching placebo.

Sometimes, masking of the participant and/or the clinical team is impossible. One example is a trial comparing a surgical with a medical intervention. Another example is a trial in which the side effects of a particular drug are readily apparent and unlikely to be seen in the comparison group(s); a classic example is a drug that changes the color of the patient's urine. However, it still may be possible, and quite valuable, to design the trial so that the outcome is assessed by a masked observer. The necessity and feasibility of doing so depend on the nature of the outcome measure. For instance, trials with survival as the primary outcome measure may be little affected by an inability to mask interventions. Trials with a more subjective outcome measure may benefit greatly by having each participant assessed by an independent observer masked as to intervention assignment. Even seemingly objective endpoints (e.g., those based on scans, photographs, histopathology slides, or cardiograms) may be evaluated by a central expert panel whose members are masked with regard to intervention. The key point is that the randomization was used to avoid bias by producing groups comparable in expectation (i.e., on average), and we want to prevent any bias from being introduced subsequently in the assessment of outcome.

In some settings, avoiding assessment bias presents special difficulties. Designers of the Prostate Cancer Prevention Trial (a chemoprevention trial of finasteride versus placebo, with a primary endpoint of biopsy-proven prostate cancer) confronted interesting questions on how to use prostate-specific antigen (PSA) monitoring while avoiding differential biopsy rates between treatment groups resulting from an effect of the intervention on PSA. This discussion led to the suggested use of blinded PSA determinations reported simply as "elevated" or "not elevated," categorized by using group-specific threshold values (23).

Some would extend masking to the members of an independent data and safety monitoring board (DSMB). However, many clinical trial experts oppose this idea. Often, accumulating data on side effects will easily reveal the nature of an intervention, thus unmasking an A versus B designation. A DSMB should review the full spectrum of outcome measures and endpoints, weighing possible benefits against side effects or other adverse effects. Therefore, it would be folly for the DSMB not to know which benefits corresponded with which harms. Furthermore, it is somewhat disingenuous to advocate that a DSMB start masked, with the ability to request its own unmasking when differences are seen. Before differences emerge, it would not matter whether members were masked or not, and, once differences emerge, members will want to know which group is which, to evaluate properly the relative benefits and risks of the competing interventions. The following question also arises: If the informed consent document reassures participants that a DSMB is watching over the accumulating data, then can the DSMB fully meet its ethical responsibilities if it too is masked? These are arguments for allowing a DSMB to evaluate unmasked data in closed (confidential) sessions.

Intention-to-treat analysis

An "intention-to-treat" analysis of outcome data from a randomized trial includes all participants randomized, counted in the group to which they were randomized (regardless of what occurs subsequently). All randomized trials should be designed so that an intention-to-treat analysis is the first one performed. This is the analysis supported by the randomization; it maintains the comparability (in expectation) across intervention groups. Analyses that exclude participants after randomization may introduce the bias that randomization was designed to avoid by making the resultant intervention groups inherently different regarding prognosis (1, 24). In the design stage, the sample size should be chosen so that the trial will have adequate power to detect the realistic alternative that will be observed with an intention-to-treat analysis, if in fact the null hypothesis (of no difference) is false.

An intention-to-treat analysis provides a valid answer to a real question. It provides a test of the "policy" ("strategy" or "intention") embarked upon at the time of randomization. If we think of a randomized assignment as a choice we make when we encounter a fork in the road, as noted previously, then the intention-to-treat analysis gives us an estimate of the difference in overall consequences upon deciding to turn right rather than left, regardless of what changes in course we make later (1).

Sometimes trials will need to randomize patients immediately to avoid delay in treatment but will incorporate a subsequent verification of eligibility. This verification may occur either after the result of a specified diagnostic test (based on a prerandomization specimen) becomes available or after an expert review of eligibility has been completed (e.g., by an central panel in a multicenter trial). While I advocate starting with a pure intention-to-treat analysis in this situation as well, it also may be reasonable to analyze just the eligible participants, provided that the assessment of eligibility is not influenced by the randomized assignment. In other words, the determination of eligibility should be masked with regard to intervention assignment and should be based exclusively on information collected before randomization, with equal ascertainment across intervention groups. In such carefully defined situations, it may be argued that the spirit of the intention-to-treat principle has not been violated (1).

Special considerations may apply to equivalence trials, whose goal is to demonstrate that competing interventions have approximately equal efficacy; in such trials, an excessive amount of noncompliance may lead to an apparent equivalence that is misleading. An intention-to-treat analysis is of interest in equivalence trials, but interpretation of the results must consider the possible impact of noncompliance.

Accounting for losses to follow-up

The study design and analysis of clinical trials must account for losses to follow-up. It is important to remember that such losses can both decrease power and introduce bias. It is common practice to deal with the issue of adequate power by planning to have a greater sample size (appropriately inflated to account for expected losses to follow-up, so that outcome data will be available on a sufficient number of patients). However, the potential risk of bias may be a more difficult issue, depending on the reasons for losses to follow-up and missing outcome data. Planning for the trial should include careful consideration of ways to minimize losses to follow-up. Note that, for intervention dropouts, outcome data should not necessarily be missing. Therefore, trials (and informed consent processes) should be designed so that treatment modifications and/or dropout (so-called off-treatment) do not lead to the participant being "off-study"; such participants should still be followed to ascertain outcome. Extra efforts should be incorporated in the study design to locate and assess primary outcome measures from participants who may no longer be receiving the intervention. For trials with a survival outcome, central databases may provide the needed vital status information.

GROUP (CLUSTER) RANDOMIZATION

Trials that randomize groups or clusters of individuals, rather than randomizing the individuals themselves, raise some interesting issues. Units of cluster randomization include communities or villages, workplaces, schools or

classrooms, religious institutions, chapters of social organizations, families, and clinical practices (e.g., in health services research). Randomization by cluster is less efficient statistically than randomization by individual, in that the persons in a cluster-randomized study will contribute less information than if randomized individually (25). The key principle here is that the design and analysis of cluster-randomized trials must account for the correlation of individuals within a cluster (26). In particular, it is essential to have an adequate sample size; often, the number of clusters drives the calculation of sample size. Reasons for randomizing by cluster, in spite of the loss of statistical efficiency resulting from the intracluster correlation, include the feasibility of delivering the intervention, political and administrative considerations, the desire to avoid contamination between those assigned to competing interventions, the basic nature of the intervention (which may be at the group level), the use of site-specific resources to decrease cost, and possibly greater generalizability (27). The following interesting examples can be cited to illustrate the use of cluster randomization:

- In Indonesia, 450 villages were randomly assigned whether to participate in vitamin A supplementation. The intervention consisted of capsules containing 200,000 IU vitamin A distributed to preschool children at baseline and 6 months later. Reported results included lower mortality and xerophthalmia prevalence; other outcomes (growth, morbidity) also were investigated (28, 29).
- Twelve communities (six pairs) were randomized in a community trial of the impact of improved sexually transmitted disease treatment on human immunodeficiency virus (HIV) infection in rural Tanzania (Mwanza). The intervention included establishing a reference clinic, staff training, a regular supply of drugs, supervisory visits to health facilities, and health education about sexually transmitted diseases. A lower rate of HIV seroconversion was reported in the intervention groups (30, 31).
- The Community Intervention Trial for Smoking Cessation (COMMIT) randomized 22 communities (11 matched pairs) in North America to investigate a behavioral intervention at the community level. The basis of the intervention was to use existing channels capable of influencing smoking behavior in large groups of people: community organization and mobilization, media and public education, health care providers, worksites, and smoking cessation resources. Although the study reported nearly identical mean quit rates for intervention and comparison communities among heavy smokers, it also reported that an additional 3 percent of light-to-moderate smokers quit in the intervention communities versus comparison communities; favorable secular trends for smoking prevalence were reported in both intervention and comparison communities (27, 32, 33).
- The Eating Patterns Study randomized 28 physician practices (within six primary care clinics) regarding whether to use a self-help booklet, with physician endorsement thereof, to lower dietary fat intake and raise dietary fiber intake. The reported result was a favorable intervention effect at 1 year in self-reported dietary behavior change (34).

Interventions in communities or other groups have frequently been investigated as nonrandomized studies. However, the use of randomization in such settings is just as important as in individual-level studies.

Increased use of cluster-randomized trials can be anticipated. Likely areas of application include studies of behavioral and lifestyle interventions, infectious disease interventions (including vaccines), studies of screening approaches, and health services research. An additional application hypothetically would involve randomizing "clinics" or "communities" to have available a new drug (or other agent) in short supply (newly licensed or available through an expanded access program). In all of these situations, well-designed randomized trials allow rigorous evaluation of the interventions being investigated.

REFERENCES

1. Green SB. Hypothesis testing in clinical trials. Hematol Oncol Clin North Am 2000;14:785–95.
2. Simon R. Clinical trials in cancer. In: DeVita VT, Hellman S, Rosenberg SA, eds. Cancer: principles and practice of oncology. 6th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2001:521–45.
3. Byar DP, Simon RM, Friedewald WT, et al. Randomized clinical trials: perspectives on some recent ideas. N Engl J Med 1976;295:74–80.
4. Green SB. Patient heterogeneity and the need for randomized clinical trials. Control Clin Trials 1982;3:189–98.
5. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. Stat Med 1984;3:361–70.
6. Yusuf S. Randomised controlled trials in cardiovascular medicine: past achievements, future challenges. BMJ 1999; 319:564–8.
7. Freedman L, Anderson G, Kipnis V, et al. Approaches to monitoring the results of long term disease prevention trials: examples from the Women's Health Initiative. Control Clin Trials 1996;17:509–25.
8. Marubini E, Valsecchi MG. Analyzing survival data from clinical trials and observational studies. West Sussex, England: John Wiley & Sons, 1995.
9. Peduzzi P, Henderson W, Hartigan P, et al. Analysis of randomized controlled trials. Epidemiol Rev 2002;24:26–38.
10. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 3rd ed. New York, NY: Springer-Verlag, 1998.
11. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 1995;48:23–40.
12. Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002;24:39–53.
13. Buring JE. Special issues related to randomized trials of primary prevention. Epidemiol Rev 2002;24:67–71.
14. Byar DP. Assessing apparent treatment-covariate interactions in randomized clinical trials. Stat Med 1985;4:255–63.
15. Yusuf S, Wittes J, Probstfield J, et al. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93–8.
16. Freedman LS, Green SB. Statistical designs for investigating several interventions in the same study: methods for cancer

prevention trials. J Natl Cancer Inst 1990;82:910–14.
17. Buring JE, Hennekens CH. Cost and efficiency in clinical trials: the US Physicians' Health Study. Stat Med 1990;9: 29–33.
18. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. N Engl J Med 1989;321: 129–35.
19. Hennekens CH, Buring JE, Manson JE, et al. Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. N Engl J Med 1996;334:1145–9.
20. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials 1998;19:61–109.
21. Rossouw JE, Hurd S. The Women's Health Initiative: recruitment complete—looking back and looking forward. J Womens Health 1999;8:3–5.
22. Lang JM, Buring JE, Rosner B, et al. Estimating the effect of the run-in on the power of the Physicians' Health Study. Stat Med 1991;10:1585–93.
23. Feigl P, Blumenstein B, Thompson I, et al. Design of the Prostate Cancer Prevention Trial (PCPT). Control Clin Trials 1995;16:150–63.
24. Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000;21:167–89.
25. Cornfield J. Randomization by group: a formal analysis. Am J Epidemiol 1978;108:100– 2.
26. Donner A, Klar N. Design and analysis of cluster randomized trials in health research. London, England: Arnold, 2000.
27. Green SB, Corle DK, Gail MH, et al. Interplay between design and analysis for behavioral intervention trials with community as the unit of randomization. Am J Epidemiol 1995;142:587–93.
28. Sommer A, Tarwotjo I, Djunaedi E, et al. Impact of vitamin A supplementation on childhood mortality: a randomized controlled community trial. Lancet 1986;1:1169–73.
29. Djunaedi E, Sommer A, Pandji A, et al. Impact of vitamin A supplementation on xerophthalmia: a randomized controlled community trial. Arch Ophthalmol 1988;106:218–22.
30. Grosskurth H, Mosha F, Todd J, et al. Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomized controlled trial. Lancet 1995;346:530–6.
31. Hayes R, Mosha F, Nicoll A, et al. A community trial of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania: 1. Design. AIDS 1995;9:919–26.
32. COMMIT Research Group. Community Intervention Trial for Smoking Cessation (COMMIT): 1. Cohort results from a four-year community intervention. Am J Public Health 1995;85:183–92.
33. COMMIT Research Group. Community Intervention Trial for Smoking Cessation (COMMIT): 2. Changes in adult cigarette smoking prevalence. Am J Public Health 1995;85: 193–200.
34. Beresford SAA, Curry SJ, Kristal AR, et al. A dietary intervention in primary care practice: the Eating Patterns Study. Am J Public Health 1997;87:610–16.