An experimental study of therapeutic efficacy is one in which (a) patients are assigned to one of two or more groups to be offered different therapeutic measures, (b) chance alone dictates whether a particular patient will be assigned to a particular group, and (c) patients in each group are monitored for the abatement of their illness or for the occurrence of the event(s) that the therapy seeks to prevent. Commonly used synonyms for this type of study are "clinical trial" and "randomized controlled trial."

Experimental designs provide results that can be interpreted relatively easily, for the common concern in nonexperimental studies—that the various treatment groups had inherently unequal probabilities of doing well—is much less pressing when it is only chance that determines the membership of the groups. For this reason, the popularity of experimental studies has increased during the last several decades. The approach has been applied to virtually every class of therapy, from pharmaceutical agents to surgical techniques to dietary and other "life-style" interventions.

The concept of an experimental study is straightforward. However, a number of issues that are not so straightforward have to be considered when planning the design and analysis of a particular study. The ways in which these issues are dealt with often have a substantial bearing on the validity and interpretation of the results.
CHOOSING THE SUBJECTS FOR STUDY

Generalizing Beyond the Study Population

The rationale for research in clinical epidemiology (and in all other health research in humans) is that by observing the illness experience of some persons we may come away with lessons that can be applied elsewhere. So, by determining that persons randomly assigned to receive drug A fare better than those assigned to receive drug B, we can conclude that persons similar to those in the study will do better on the average by taking drug A rather than drug B.

To what extent must these two groups—study subjects and persons to whom we would like to refer the findings (reference population)—be "similar" so that valid generalizations can be made from one to the other? To answer this question, it is necessary to take into account the answers to two other questions: (a) Do both the study and reference populations suffer untoward consequences from the condition for which the therapy is being given? (b) If the therapy is effective in the study population, would the means by which it is believed to act (i.e., its biological effect) be present in members of the reference population as well?

To illustrate how the matter of generalizing is approached in practice, let's place ourselves back in the late 1960s, immediately following the publication of the results of the first, large, randomized controlled trial of antihypertensive therapy for the reduction of mortality from cardiovascular disease (Veterans Administration Cooperative Study Group on Antihypertensive Agents, 1967). The study documented a substantial mortality reduction in actively treated versus placebo-treated male veterans with a diastolic blood pressure of 115 to 129 mm Hg who were free of clinical cardiac or cerebrovascular disease and in whom there was no advanced retinal or renal pathology. Now, let's say we had a patient with that level of blood pressure but who was neither a veteran nor male. Should we presume that the findings apply to her? Is an experimental study of the efficacy of antihypertensive drugs needed in non-war veteran women? We would address the issue by considering, first, whether high blood pressure in nonveterans and women predisposes them to an increased risk of mortality from cardiovascular diseases. In 1967, data were available indicating that such persons indeed were at increased risk. [Of course, there are other situations in which the available data suggest no relationship. For example, in hypertensive persons over the age of 75 years there may be little increased mor-
tality over that in normotensive persons (Mitchell, 1983). Also, it may be that data just are not available for the population or disease subgroup to which a patient belongs that link the condition being treated with the outcome."

Second, we would ask whether the postulated mechanism through which antihypertensive therapy exerted its beneficial effect on mortality in male veterans is present in nonveteran women. Unfortunately, it is rarely possible to arrive at an unequivocal answer to a question of this sort, for our knowledge of the pathogenetic mechanism(s) is rarely definitive. While there would be no basis for believing that nonveterans differ from veterans in this respect, it is not out of the question that hormonal and other differences between the sexes could make the extrapolation from men to women inexact. The uncertainty would almost certainly be greater still when trying to extrapolate the study results to persons with blood pressure levels below 115 mm Hg. The benefit of antihypertensive therapy would be expected to be smaller in them (for their excess risk is smaller), but by how much? This is no minor matter, as there is a far greater number of people with modest elevations of blood pressure than there is with large elevations, and at some point a blood pressure threshold must be set below which therapy will not be instituted."

Maximizing the Study’s Ability to Identify Therapeutic Efficacy

Some leap of faith is going to have to be made in applying the results obtained in the study subjects to the reference population. Therefore, the choice of the particular group of subjects for study usually depends less on the degree to which the group represents the reference population than on the group having characteristics that will produce a study with high “power,” that is, characteristics that will successfully identify a difference between treatments, if one is truly present. There are two such important characteristics.

"By the end of the 1970s, experimental studies had demonstrated the efficacy of antihypertensive therapy in nonveterans, women, and persons with less extreme blood pressure elevations. The Hypertension Detection and Follow-up Group (1979a,b) contrasted treatment by regular medical care with that by specialized centers that provided close blood pressure scrutiny and control. Persons randomly assigned to the specialized centers experienced a greater average reduction in blood pressure than did persons receiving regular care, and they had a lower mortality, irrespective of sex or initial diastolic blood pressure (≥90 mm Hg)."
Therapeutic Efficacy

The lower the cost per subject, the larger will be the number of subjects available at a given budget level (it is necessary to get these studies funded) and the more statistically powerful will be the study. An early investigation of the effect of a low-cholesterol, low-saturated fat diet on the incidence of cardiovascular disease was conducted in a Veterans Administration psychiatric hospital (Dayton et al., 1969). Patients receiving meals from one kitchen had a modified diet, whereas the diet of other patients remained as before. The cost of performing this intervention at an institution clearly was less than that of trying to modify dietary cholesterol and saturated fat on an individual basis in individual kitchens.

Once a group of investigators has been established to conduct a randomized controlled trial, there is an economy in having these same investigators conduct additional studies. The initial administrative costs can be avoided (and possibly also the costs that can be associated with the group’s early inexperience) when another therapy for the disease first treated is to be evaluated, or when the investigators begin an evaluation of a therapy for another disease that they encounter. In some instances, the same patients enrolled and participating in the first study can be enrolled in the second study. For example, to determine if aspirin use could prevent myocardial infarction, participants in some treatment groups of the Coronary Drug Project that had been disbanded (e.g., those assigned to receive estrogen or thyroid hormones), but who were still under surveillance, were randomly assigned to receive aspirin or placebo (Coronary Drug Project Research Group, 1976).

High Expected Compliance with the Therapeutic Regimen(s) and Likelihood of Successful Follow-Up

Many clinical trials begin with a “run-in” phase in which all potential subjects are given a placebo or a “control” therapy. Their compliance with the regimen is assessed, and only those in whom compliance is good are entered into the randomized portion of the study (also see p. 58). This is yet another assault on the “representativeness” of the study subjects vis-à-vis the reference population, for the latter invariably would include volunteers and nonvolunteers, compliers and noncompliers, to whom one would like to extrapolate the results. Yet, the resulting study population offers a “clean” separation of subjects exposed to the various treatments,
and thus enhances the ability to find any true between-treatment difference that might exist.

[Occasionally, a study has as its goal the measurement of efficacy in the population to whom the intervention is offered, that is, the efficacy in those who receive it combined with the dilution that results from whatever noncompliance exists in that population. For example, the study of mammography and clinical breast examination conducted within the Health Insurance Plan of New York randomly assigned some of the female members to the intervention group and only then notified them of the study (Shapiro et al., 1971). A sizable fraction of these women, 35%, failed to attend even a single screening examination. Nonetheless, since the aim of the study was to determine if screening of this type would reduce mortality from breast cancer in a population to whom it was offered, the aims of the study were served fully by comparing mortality in the study group, at whatever level of compliance, with that of other women in the Plan.]

NATURE OF THE INTERVENTION

Generalizing Beyond the Therapeutic Measure under Study

Often there are a number of interventions that can accomplish the same biochemical or physiologic change, for example, lowering serum cholesterol, lowering arterial blood pressure, raising gastric pH. If only one of the therapies has been evaluated (say, against a placebo) and a beneficial effect has been found in preventing or controlling a clinical manifestation that results from the original biochemical/physiological derangement, what can be said of the benefits expected from the use of another therapy that is believed to act in the same manner?

Here, just as with the question of generalizing from the study to the reference population, it is not possible to provide a definitive answer that covers all situations. To the extent that the means by which the evaluated therapy exerted its beneficial effect is (a) known and (b) shared by the not-yet-evaluated therapy, there will be confidence that it is also effective. Since this is a subjective assessment, not all who review the evidence will arrive at the same conclusion. For example, a decline in mortality from coronary disease was noted in persons assigned at random to take cholestyramine, an agent that lowers the concentration of certain serum lipids (Lipid Research
true between-treatment differences in the measurement of efficacy is offered, that is, the dilution that exists in that population. For clinical breast examination Plan of New York numbers to the intervention phase of the study (Shapiro et al., 1984), failed to attend even less, since the aim of the s type would reduce mor- phological changes. In the intervention phase, with that of other

re under Study

s that can accomplish the e, for example, lowering pressure, raising gastric a evaluated (say, against a found in preventing or results from the original what can be said of the r therapy that is believed

eralizing from the study to able to provide a definitive extent that the means by beneficial effect is (a) known therapy, there will be cons- a subjective assessment, e at the same conclusion. bronary disease was noted estyrarzine, an agent that in lipids (Lipid Research

Clinics Program, 1984). Is it reasonable to assume that a diet that can accomplish a similar reduction in lipid levels will accomplish a similar reduction in mortality? Experts are not unanimous in their answers to this question, since they are not unanimous in believing that the effect of cholestyramine that is relevant to coronary disease is shared by a low-saturated fat, low-cholesterol diet.

Maximizing the Study's Ability to Identify Therapeutic Efficacy

The finding of no difference between treatments employed in a clinical trial often occurs because the true difference between the treatments is too small to have been reliably detected in the trial. At the planning stage of a clinical trial, there are two ways to try to make a false-negative result less likely: The number of subjects to be enrolled can be made large (see p. 51), or, the intervention chosen can be made as different as possible from that to be offered to the comparison group(s). For example, suppose you are evaluating the effect of dietary modification on disease (e.g., low saturated fat intake in relation to the occurrence of myocardial infarction, or, in older children with phenylketonuria, abridgement of low phenylalanine intake in relation to IQ). It would be important to make the modified diet as different as possible from that of the controls, within the range of what you believe will be acceptable to patients. There will always be critics of a "negative" study who say that, had the intervention only been somewhat more extreme, a positive result would have emerged. You want to avoid giving them any extra ammunition!

NATURE OF THE THERAPY TO BE ADMINISTERED TO THE COMPARISON GROUP

Typically, the comparison group is prescribed "conventional" therapy, which may range from no therapy at all to a complex array of interventions that is believed, at that time, to be the best that can be offered. To justify the conduct of the study, of course, there should be a "reasonable" probability that the comparison therapy, even if it is no therapy, will prove to be as good as or superior to the therapy under study. A judgment as to what is reasonable will be based on existing data, usually from nonexperimental studies. Since a judgment is involved, there often will be disagreement among investigators as to the adequacy of these data in determining
the efficacy of the therapy being considered for study and thus there often will be disagreement as to whether it is ethical to subject some patients to any other therapy.

If some form of placebo is to be used it should be, if possible, totally innocuous. However, there are some situations in which the manipulation necessary to administer the therapy is such that, in the absence of a similar manipulation in the comparison group, the internal validity of the study could be undermined due to a placebo effect. One such situation arose in a clinical trial that evaluated the effect of an intradiskal injection of chymopapain to relieve sciatica due to a herniated lumbar disc (Javid et al., 1983). A needle was placed into the intervertebral disc of each study subject, who was then given either chymopapain or sterile saline on a random basis.

Since the administration of an intradiskal injection of an inert substance could never be considered a legitimate therapeutic alternative—it poses at least some risk and discomfort to the patient with no possibility of benefit—how can it be justified ethically? The only way is to have the prospective subjects informed as to the nature of the study and of the rationale for the use of a placebo that is not free of risk and/or discomfort. They can then decide whether they are willing to subject themselves to the risks so that others may benefit.

Experimental evaluations of coronary artery bypass surgery have eschewed the use of blinding by means of a “sham” surgical procedure, and have instead assigned comparison patients to “medical” therapy (Principal Investigators of CASS and Associates, 1981). When planning the studies the investigators believed that, in terms of some combination of pain relief, functional ability, and mortality, the medical therapy group might well have a better outcome than their surgical therapy counterparts. However, this choice of comparison therapy meant that certain important research questions could no longer be unambiguously addressed. Specifically, it would be impossible to determine if any reduced level of pain in the patients undergoing bypass surgery, relative to that in the medical therapy group, was a result of the bypass per se or simply of the more general effect of a surgical procedure performed within the chest (Cobb et al., 1959). Nonetheless, it would have been inappropriate to do anything other than place the patients’ welfare first, even though that forced some of the goals of the studies to be compromised.

Clearly, this lack of blinding of subjects and investigators will
prove more important in the interpretation of some outcomes (e.g., chest pain) than of others (e.g., mortality). If possible, of course, it is better to keep both subjects and investigators ignorant of the treatment status, as this will minimize the possibility of actions on the part of either group that could bias the results. When complications of disease or therapy arise that necessitate knowledge of the specific therapy to which the patient has been assigned, this information usually can be given to one or more physicians external to the study who can decide on the proper course of action. If the therapy under study is a drug, the blinding is generally done simply by preparing a placebo identical in appearance to the active agent. However, one study in which the identical appearance of drug and placebo was achieved, but blinding was not, is instructive to review here:

Example. In the early 1970s, healthy adults were enrolled in an experimental study in which they were asked to take either vitamin C (3 g/day) or a lactose placebo for 9 months, during which time the incidence of colds was monitored (Karłowski et al., 1975). Because during the follow-up period some subjects indicated that they were biting into and tasting the preparation that they had been given, the investigators asked all subjects at the conclusion of the study to guess the group to which they had been assigned. Of the 102 who attempted a guess, 79 were correct (77%). The following table summarizes the incidence of colds in persons assigned to each of the two treatment groups, as well as in the subgroup of subjects who guessed incorrectly:

<table>
<thead>
<tr>
<th>Treatment guessed</th>
<th>Treatment received</th>
<th>No. of subjects</th>
<th>No. with ≥2 colds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vitamin C</td>
<td>Placebo</td>
<td>11</td>
<td>2 (18%)</td>
</tr>
<tr>
<td>Placebo</td>
<td>Vitamin C</td>
<td>101</td>
<td>36 (36%)</td>
</tr>
<tr>
<td></td>
<td>Vitamin C</td>
<td>12</td>
<td>8 (67%)</td>
</tr>
<tr>
<td></td>
<td>Placebo</td>
<td>89</td>
<td>42 (47%)</td>
</tr>
</tbody>
</table>

In the group assigned to receive placebo, there was an overall excess (47% vs. 36%) in the percentage of subjects with two or more colds. However, a larger difference was associated with a subject’s believing he or she was assigned to a particular group: 36% of subjects assigned to receive vitamin C had two or more colds, twice the incidence in persons who, though they actually were taking placebo, thought they were taking vitamin C. A similar difference was found for persons receiving the vitamin but believing it was a placebo—their incidence was higher than persons receiving placebo (67% vs. 47%). Since a subject’s suspicion of the group to which he or she had been assigned so strongly influenced the results, and since a subject’s suspicion was much more often right than wrong, the validity of the vitamin C-placebo comparison was seriously compromised.
ASSIGNMENT OF SUBJECTS TO TREATMENT GROUPS

How Should the Assignment Be Made?

Any method of patient assignment is acceptable if it guarantees that chance alone dictates the assignment of a particular patient to a particular treatment group. Thus, the more rigid is the process (e.g., the use of a table of random numbers), the better. The provider of care, having made the decision to ask the patient to participate in the study, should have no role in the assignment process. It is necessary to safeguard against letting the provider's judgment as to which patient needs or does not need the therapy under study influence the treatment assignment. Thus, schemes that call simply for alternate assignment of patients to different groups have the potential for bias, at least if it can happen that more than one patient is to be assigned to the groups at the same time.

When Should the Random Assignment Take Place?

In most experimental trials, subjects are assigned to the treatment groups only after they have been informed of the nature of the study and of the therapies they may be offered. For example, in the trial designed to evaluate the efficacy of coronary artery bypass surgery (Principal Investigators of CASS and Associates, 1981), patients who were deemed eligible for surgery, but not those in whom surgery was deemed essential, were asked to allow themselves to be assigned at random to receive the surgery or the medical therapy. The study included only those patients who agreed to submit to the luck of the draw. See Figure 4-1 for a schematic representation of this approach.

There are some circumstances, however, in which it might be advisable to make the random assignment prior to requesting participation ("prerandomization").

1. The investigator and/or clinical collaborators may be uncomfortable with the idea of presenting the possibility of a choice of therapy to a patient. In the case of a life-threatening illness, for example, cancer, perhaps the investigator is concerned that patients not assigned to the newest, most radical measures may withdraw from the trial, even after consent has been given, in order to actively seek these treatments. In such a circumstance, the investigator can randomly assign patients to
TABLE GROUPS

It is not acceptable if it guarantees that a particular patient to a particular group is the process (e.g., group A or B). The provider of care to the patient who is a potential patient to participate in the study is not rigid is the process for the study. It is necessary for the provider's judgment as to whether the patient is eligible for therapy under study influences that call for the randomization. Some patients have the potential to be more than one patient is eligible at any one time.

Place?

are assigned to the treatment group of the nature of the study. For example, in the study, the coronary artery bypass surgery (Ellenberg and Associates, 1981), surgery, but not those in which it might be unusual to be asked to allow themselves to undergo surgery or the medical therapist who agreed to submit to a more radical approach.

However, in which it might be better to request participating patients in a different way than the study of patients was conducted. The challenge of allocation can be understood by the example of a study of two treatments. In such a study, one might randomly assign patients to two treatment groups and then evaluate the outcomes. In this way, the potential for bias due to differences in baseline characteristics is minimized.

Alternative therapies and either (a) seek consent to participate in the study from all patients, whether they have been assigned the standard or the new therapy (Ellenberg, 1984) or (b) seek consent only from patients assigned the new therapy (Zelen, 1979).

2. Sometimes the efficacy of an intervention is to be studied within a large, defined group of individuals (e.g., members of a prepaid health care plan) who have not actively sought care from the investigative team. This situation might arise in an evaluation of vaccine efficacy in healthy individuals or of a screening technique for cancer. In such situations, it could prove logistically difficult and unnecessarily costly to contact and explain the study to all persons, rather than to just the fraction to whom the intervention measure will be offered.
In order for prerandomization to succeed in evaluating efficacy, a high proportion of subjects to whom the treatment is offered will have to accept it. This is because patients who refuse the treatment must, for analytic purposes, be retained in the same group as those who comply. Therefore, a low level of compliance will obscure any true treatment-related benefit.

In the “conventional” type of experimental study (i.e., agreement to participate prior to randomization), it may be desirable to postpone randomization of a patient until his or her compliance can be evaluated (assuming noncompliance is a possibility, e.g., in a drug or lifestyle intervention study). The measurement of compliance can take many forms (pill counts, biochemical tests, etc.), but the goal is the same, that is, to eliminate before the start of the study patients who have a high likelihood of not adhering to the regimen offered and who are likely to be a source of misclassification within the study. With this in mind, the study of the value of lowering blood pressure conducted within the Veterans Administration (Veterans Administration Cooperative Study Group on Antihypertensive Agents, 1967) administered to all eligible hypertensive subjects a placebo tablet that contained 5 mg of riboflavin for 2 to 4 months. Because the urine of persons who take riboflavin is fluorescent yellow under ultraviolet light, the investigators had an objective measure of compliance available only to them. Only those subjects whose compliance achieved a designated level—a bare majority of the potential subjects—were enrolled in that randomized trial of antihypertensive agents.

Under What Circumstances Can a Subject Serve as His or Her Own Control?

There are many conditions that affect more than one part of the body, and for some of these conditions it is possible to administer therapy locally that has no effect on untreated lesions elsewhere. In such a situation, it is appropriate to design a study so that one or more lesions are randomly selected for treatment, with others serving as control sites. Such a design has been used to evaluate measures intended to control diabetic retinopathy (Diabetic Retinopathy Study Research Group, 1981) and a number of dermatologic lesions (Gilchrest et al., 1979).

Studies of the efficacy of therapies intended to reduce the frequency or severity of chronic, recurrent problems, such as seizures, arthritic pain, or menopausal hot flashes, can be more precise in a
ceed in evaluating efficacy, the treatment is offered wills who refuse the treatment in the same group as those compliance will obscure any

mental study (i.e., agreement it may be desirable to post- or her compliance can be possibility, e.g., in a drug measurement of compliance (hemical tests, etc.), but the fore the start of the study pot adhering to the regimen e of misclassification within of the value of lowering orans Administration (Vety Group on Antihypertensive hypertensive subjects iboflavin for 2 to 4 months. iboflavin is fluorescent yellowors had an objective mea- hem. Only those subjects d level—a bare majority of n that randomized trial of

Serve

more than one part of the it is possible to administer treated lesions elsewhere. In sign a study so that one or treatment, with others serv- been used to evaluate mea-opathy (Diabetic Retinopa- a number of dermatologic

intended to reduce the fre- problems, such as seizures, s, can be more precise in a

statistical sense if the subject can serve as his or her own control. By evaluating the same subject at different times, in the presence and the absence of the therapy under study, the variability among subjects in the frequency or severity of the problem will not blur true differences in efficacy. This can be achieved experimentally in a “crossover” design, in which the study subjects are divided into two groups and dealt with as shown in Figure 4-2.

At the end of the study, the primary comparison is between, for each subject, the frequency and/or severity of the condition at time-points 1 and 3 (Hills and Armitage, 1979). Events occurring during the period between time-points 1 and 2, the length of which is determined largely by the amount of time needed for the effects of the measures initially administered to dissipate, are not included in the analysis. (Crossover studies are not appropriate for evaluating the efficacy of therapeutic measures whose effects following discontinuation do not dissipate relatively quickly, that is, within several days.)

In crossover studies it is important to include both sequences, therapy–control and control–therapy, for the effect of the therapy can be confounded by the sequence in which it is given (Louis et al., 1984). An example of what could happen if only one sequence is

![Diagram](image-url)

**Figure 4-2.** Schematic representation of a trial using a crossover design.
used can be seen in a crossover study of estrogen versus placebo for the relief of symptoms associated with the menopause (Coope et al., 1975). Women entering the study averaged 50 to 60 episodes of hot flashes per week. In the group assigned to receive placebo first, the rate fell to 20 per week after 3 months, and fell further, to less than five per week, at the end of a subsequent 3-month period on estrogen therapy. So, in women who started with placebo and crossed over to estrogens, there was evidence of efficacy of estrogen use, although not of an overly great magnitude relative to the “efficacy” of a placebo. It was in the group of women assigned to the other sequence—estrogen first, placebo second—that a large difference occurred. After 3 months on estrogens the frequency of hot flashes fell to less than five per week, but the switch to placebo resulted in a return to the original 50 to 60 per week. While the reasons for the difference in measured efficacy between the two sequences are not well understood, the fact that such a difference can occur reinforces the idea that both possible sequences in a crossover study must be examined.

ASSESSMENT OF ENDPOINTS IN STUDY SUBJECTS

It goes without saying that the vigor with which members of the treatment groups are followed must be equivalent, so that the detection of any study endpoints that do occur is comparable among groups. There should be standard criteria by which the endpoints can be assessed, preferably stipulated at the start of the trial. And, if possible, the person (or persons) who applies the criteria to a particular case should be ignorant of the treatment group to which that case has been assigned.

COMPARISON OF THE TREATMENT GROUPS FOR THE OCCURRENCE OF ENDPOINTS

Which Endpoints to Count

In planning most experimental studies, the choice of endpoint is clear. A study of an analgesic would measure the patient’s perception of pain, a study of an antibiotic would measure the disappearance of infection and its clinical manifestations. However, some therapies have such a high potential for serious adverse effects that...
of estrogen versus placebo for the menopause (Coope et al., aged 50 to 60 episodes of hot
and to receive placebo first, the
s, and fell further, to less than 3-month period on estro-
ed with placebo and crossed
of efficacy of estrogen use, tude relative to the “efficacy” women assigned to the other
and—what a large difference
the frequency of hot flashes switch to placebo resulted in week. While the reasons for
between the two sequences are
a difference can occur rein-
ences in a crossover study

SUBJECTS
with which members of the equivalent, so that the detect-
occurs is comparable among
teria by which the endpoints at the start of the trial. And,
applying the criteria to a par-
treatment group to which that

GROUPS FOR THE

es, the choice of endpoint is measure the patient’s perce-
ould measure the disappear-
manifestations. However, some
serious adverse effects that

a broad range of endpoints must be considered. In an experimental
study of the efficacy of transplantation in prolonging life, it would
not make sense to look at mortality rates in transplanted and non-
transplanted patients only for the cause of death at which the ther-
apy was directed (e.g., renal failure, leukemia). Rather, it would be
ecessary to count causes of death that are related to rejection of
the transplant as well. In practice, one probably would tabulate all
causes of death combined.

In studies with mortality as an endpoint, a problem arises when
the cause of death against which the therapy is directed does not
constitute an overwhelming majority of the total deaths, and yet
there is no good information available a priori as to which other
causes might be affected. The problem exists because neither of the
possible comparisons that can be made, rate of therapy-directed
causes or rate of all causes, is wholly satisfactory. With the latter
comparison, by incorporating many extraneous endpoints, one runs
the risk of diluting a true effect of the therapy and making it harder
to identify (Sackett and Gent, 1979). With the former comparison,
only an incomplete picture of the therapy’s effects may be seen.
Proper interpretation of such a study requires that both compar-
isons be made:

Example. A randomized controlled trial was conducted in which clofibrate or a
placebo was given for an average of 5.3 years to hypercholesterolemic men in
order to influence the incidence and mortality of arteriosclerotic vascular dis-
ease (Committee of Principal Investigators, 1978). The rate of nonfatal myocar-
dial infarction was decreased in the clofibrate-treated men, although the rate of
fatal heart attack was similar in the two groups. However, other causes of death
were more common in the treated group, particularly deaths from digestive dis-
eses and cancer, so that the overall mortality in clofibrate and placebo groups,
respectively, was 2.2 and 1.7 per 1,000 per year.

Some conditions that one is attempting to treat or prevent from
occurring have measurable antecedents. For example, following
treatment for cancer, a death from that cancer is usually preceded
by tumor recurrence or metastasis. Death from ventricular arrhyth-
mia following myocardial infarction is often preceded by nonfatal
episodes of ventricular arrhythmia. Since one of the factors limiting
the power of an experimental study is the number of endpoints, and
since these antecedent conditions often occur more commonly than
do the endpoints themselves, the monitoring and analysis of these
antecedents should increase the statistical power per subject
enrolled.
As an investigator, you would be pleased to accept this increased in power as long as the analysis of the occurrence of such an antecedent condition is giving you, qualitatively, the same “answer” as would a study of a larger number of subjects in which the endpoint itself was measured. You will get the same answer to the extent that the occurrence of the antecedent condition is highly predictive of the endpoint:

*Example.* De Silva et al. (1981) pooled data from six experimental studies of lidocaine prophylaxis in patients with acute myocardial infarction. The incidence of in-hospital ventricular fibrillation (VF) in the group that received lidocaine (16 of 517) was only about one-half the incidence in the group that did not (29 of 505). Deaths due to VF were too uncommon for a meaningful analysis. Nonetheless, since the occurrence of VF following myocardial infarction is a strong risk factor for cardiac death (Ribner et al., 1979), an interim conclusion as to the efficacy of lidocaine prophylaxis in preventing death from VF seems justified.

The following is an example of an intervention that was successful in modifying a suspected antecedent of an endpoint but that turned out to have no influence on the endpoint itself. It should serve as a caution in interpreting experimental studies that measure a therapy’s effect on an antecedent alone.

*Example.* To determine whether the incidence of hepatitis B infection could be reduced in persons undergoing long-term hemodialysis, a large, randomized, controlled trial (n = 1,311) of hepatitis B vaccine was initiated (Stevens et al., 1984). Active production of antibody to hepatitis B surface antigen (anti-HBs) occurred in about 50% of the vaccinated group and in only 2% of those given a placebo. Nonetheless, during the 25-month follow-up period the incidence of hepatitis B infection in the two groups was nearly identical. The authors concluded that “although anti-HBs is traditionally used as an index of immunity to [hepatitis B] infections, it may not be the crucial protective factor. . . . The vaccine we used may have failed to induce such protective responses in immunocompromised patients, resulting in our inability to demonstrate its efficacy, even when there appeared to be an appropriate anti-HBs response.”

Control of the Potentially Distorting Influence of Other Variables

The University Group Diabetes Project (UGDP) was an experimental trial of the ability of hypoglycemic agents to reduce the occurrence of complications of diabetes. One group of patients was assigned, at random, to receive the drug tolbutamide. These patients happened to differ from those who received no active agent in several respects (e.g., age) so that, apart from any influence of
Therapeutic Efficacy: Experimental Studies

The drug, the tolbutamide therapy group would have been expected to have a somewhat higher rate of complications (Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents, 1975).

This sort of imbalance can occur even in an experimental study. It is important to remember that, with respect to other relevant characteristics, randomization merely assures that on the average there will be equality of the groups being compared. In any one trial, between-group differences in characteristics related to the study outcome can and do occur (e.g., in the UGDP study), although rarely are they of any great magnitude.

Two strategies are commonly used to prevent the true measure of efficacy of a therapy from being distorted (confounded) in this way:

1. It is possible to form subgroups ("blocks") of subjects who are homogeneous for the presence or level of risk factors for the study outcome, and then to allocate a fixed proportion of the subjects within each block to each of the various treatments. For example, in the study of coronary artery bypass surgery (Principal Investigators of CASS and Associates, 1981) patients were put into groups based on their symptoms, ventricular function, number of diseased vessels, and the institution in which the treatment was being administered. Within each group, equal numbers were assigned to receive medical therapy and surgery, the order of the assignment being selected at random by the study's statisticians.

This procedure guarantees that treatment groups will be comparable with respect to the factor(s) that define the subgroups. The drawback of the strategy is the added complexity at the time of assignment to the treatment groups, and so it is usually reserved for only those characteristics that are expected to have a strong bearing on the study outcome.

2. When the study has been concluded, the treatment groups can be compared for all characteristics believed to have an influence on the outcome. For those characteristics that differ among the groups, it will be possible to control analytically for their interfering effect, either through adjustment (see Appendix) or through other statistical means. In the UGDP study, for example, the difference in mortality from cardiovascular disease between subjects assigned to receive tolbutamide and those assigned receive placebo, an excess of 13.2 per 1,000 per year in the tolbutamide treatment group, was due in part to the higher mean age of the tolbutamide-treated group.

The incidence of hepatitis B infection could be reduced by hemodialysis, a large, randomized vaccine was initiated (Stevens et al., hepatitis B surface antigen (anti-HBs) group and in only 2% of those given a 3-month follow-up period the incidence of anti-HBs was nearly identical. The authors concluded used as an index of immunity to crucial protective factor. . . . The vacci-
group. The difference fell to 12.4 per 1,000 per year once the age distributions of the two groups were “forced” to be the same by an adjustment procedure.

Handling Those Subjects Who Were Assigned to One Mode of Therapy But Who Did Not Receive It

Studies vary regarding the frequency with which their subjects, once randomized, fail to receive fully the treatment to which they have been assigned. It is the rare study in which the original assignments are adhered to perfectly. The reason for a change in therapy may be physician-initiated: Perhaps the condition of a patient assigned to the medical arm of a study of coronary artery surgery worsens and the physician feels that surgery is necessary. Or, perhaps a patient develops an adverse reaction to a study drug and the physician must discontinue it. Alternatively, failure to adhere to the prescribed regimen can be due simply to patient noncompliance.

In theory, when analyzing the results of an experimental study there are two possible ways of dealing with patients who did not receive or complete the intended course of therapy: (a) keep them in the groups to which they were originally assigned, thus retaining the experimental nature of the design, and accept the resulting misclassification of patients in terms of treatment actually received; and (b) allow them to “transfer” to the therapy they actually received (or to withdraw from the study altogether), losing the experimental nature of the design but reducing the amount of misclassification that would occur in option (a).

In practice, the first approach (keeping the originally assigned groups) is the one to rely on. Categorizing patients on the basis of therapy received allows for the possibility of “selection” bias to appear, that is, are there characteristics of persons who did not receive or complete the originally assigned therapy that also are correlated with the outcome being measured? Based on what we know about the reasons for changes in therapy following randomization, and on the characteristics and outcomes of patients who do change therapies, a considerable amount of selection bias would be expected to result in most studies. For instance, there are therapies that, if effective, are expected to produce demonstrable improvement in the patient’s status before the full course of therapy is finished. It would be expected that patients who do not complete the full course will be disproportionately numerous in the group assigned to the therapy that is truly less effective. In such situations,
the elimination of patients with an incomplete course of therapy from the analysis will diminish the measured efficacy of the superior therapy.

Even when the outcome does not influence compliance in such a direct way, the failure to maintain originally assigned groups can lead to a biased result. An instructive example comes from an experimental study in which, in an effort to reduce cardiac mortality, the drug clofibrate (which lowers the concentration of serum cholesterol) or a placebo was prescribed to patients who had sustained a myocardial infarction (Coronary Drug Project Research Group, 1980). During the 5-year period of the study, adherence to the prescribed regimen was monitored in both treatment and placebo groups. The cumulative mortality was found not to differ between the clofibrate and placebo treatment groups. However, no matter which regimen had been prescribed, those who adhered to it (80% or more of the time) experienced a mortality of about 15%, whereas the mortality in persons who were less compliant was about 27%. If, in the analysis, the investigators had placed the noncompliant clofibrate patients into the placebo group, they would have found a spurious benefit associated with use of the drug.

Statistical Analysis of Experimental Studies

The topic of statistical analysis lies outside the scope of this book. An excellent introduction to the subject, one that emphasizes practical aspects and provides examples, can be found in an article by Peto et al. (1977).

LIMITATIONS OF EXPERIMENTAL STUDIES OF THERAPEUTIC EFFICACY

Randomized controlled trials generally are not cheap to conduct. A considerable expenditure of resources is required to assemble a group of subjects and to monitor them over time. There are also substantial administrative costs associated with the multiinstitutional collaboration that is often required. Since the magnitude of the cost is in part related to the number of subjects studied, there are usually financial restrictions on the size of most experimental studies.

Of course, the smaller the number of subjects, the smaller will be the power of the trial to reliably determine the difference between
treatment groups (see Appendix). The following illustrates the ambiguities in interpretation that can arise when a seemingly large study is not quite large enough:

*Example.* During the mid- to late 1970s, a collaborative investigative group wished to determine whether patients with cutaneous melanoma would benefit from chemotherapy, immunotherapy, or both, after resection of tumor (Vernesi et al., 1982). They randomly assigned 761 patients at high risk of tumor recurrence, based on depth of skin invasion or the presence of regional node involvement, to one of four treatment groups: (1) surgery alone; (2) surgery followed by intravenous dacarbazine administration; (3) surgery followed by immunotherapy with bacille Calmette-Guérin (BCG) vaccine; and (4) surgery followed by both dacarbazine and BCG administration. The survival at 3 years among members of each group was as follows:

<table>
<thead>
<tr>
<th>Treatment group</th>
<th>3-year survival</th>
<th>Difference from surgery alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery alone</td>
<td>41.6%</td>
<td>—</td>
</tr>
<tr>
<td>Surgery + dacarbazine</td>
<td>46.5%</td>
<td>4.9%</td>
</tr>
<tr>
<td>Surgery + BCG</td>
<td>48.7%</td>
<td>7.1%</td>
</tr>
<tr>
<td>Surgery + dacarbazine + BCG</td>
<td>50.0%</td>
<td>8.4%</td>
</tr>
</tbody>
</table>

The difference in survival between the group that received surgery alone and each of the others was quite compatible with there being no true difference, that is, each of the three *p* values was large. The authors concluded that "an advantage of adjuvant treatment was not demonstrated. Either a difference did not exist, or it was of limited clinical importance."

So, what are you, as a physician, to do for the patient whose melanoma and positive regional lymph nodes have recently been resected? A difference in 3-year survival of 5 to 8% may be "of limited clinical importance" relative to the initial hopes of these investigators, but you do not want to deprive your patients of an extra chance of survival of that magnitude: For most of them, that extra chance is probably worth the costs and adverse effects of the chemotherapy and/or immunotherapy. And yet, despite the fact that the results are based on the study of 761 patients, a larger number than in most clinical trials, there is a reasonable probability that the improved survival in the groups that received one of the supplementary therapies was simply the result of chance. Differences in efficacy between treatment regimens that are of clinical importance often cannot be resolved in clinical trials because of the trial's inability to enroll enough patients.

Figure 4-3 depicts the number of subjects (*n*) needed in each of
The following illustrates the arise when a seemingly large

a collaborative investigative group shistemaneous melanoma would benefit th, after resection of tumor (Vero-761 patients at high risk of tumor n or the presence of regional node Lips: (1) surgery alone; (2) surgery nistration; (3) surgery followed by n (BCG) vaccine; and (4) surgery ministration. The survival at 3 years vs:

<table>
<thead>
<tr>
<th>Year survival</th>
<th>Difference from surgery alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.6%</td>
<td>—</td>
</tr>
<tr>
<td>46.5%</td>
<td>4.9%</td>
</tr>
<tr>
<td>48.7%</td>
<td>7.1%</td>
</tr>
<tr>
<td>50.0%</td>
<td>8.4%</td>
</tr>
</tbody>
</table>

up that received surgery alone and ith there being no true difference, e. The authors concluded that “an demonstrated. Either a difference did ance.”

do for the patient whose mel-

odes have recently been al of 5 to 8% may be “of lim-
e initial hopes of these inves-
ve your patients of an extra For most of them, that extra d adverse effects of the che-
and yet, despite the fact that 761 patients, a larger number easonable probability that the cceived one of the supple-
it of chance. Differences in effi-
of clinical importance often can-
of the trial’s inability to enroll ubjects (n) needed in each of

two groups of equal size to reliably (80% of the time) detect a signif- significant difference among them (at $P = 0.05$). The number is related to (a) the frequency with which the study endpoint occurs in one of the groups ($P_1$, ranging from 0 to 1.0) and (b) the true difference between the frequency of the study endpoint in the two groups ($P_2 - P_1$, ranging from −0.5 to 0.6). So, for example, if 30% of patients with a given condition show symptomatic improvement after treatment with a placebo, and if in truth 50% treated with drug

Figure 4-3. Sample sizes required for testing two independent proportions, $P_1$ and $P_2$, with 80% probability of obtaining a significant result at the 5% (two-
sided) level. $n =$ Number of observations per group. (Adapted from P. Feigl: A graphical aid for determining sample size when comparing two independent proportions. Department of Biostatistics, Technical Report No. 6, University of Washington, Seattle, July 1977.)
A improve \((P_2 - P_1 = 0.50 - 0.30 = 0.20)\), then it will require slightly fewer than 200 patients (slightly fewer than 100 per group) to reliably identify this difference. Note that for between-group differences smaller than 0.10, several hundred or more patients per treatment group are needed. The smaller the benefit of the therapy being evaluated, the larger is the study required to determine it.

Because of the limited power of most clinical trials to determine clinically important differences between two therapeutic regimens, it is only under unusual circumstances (e.g., a very large number of subjects relative to that needed to detect the differences expected) that the patient population should be split into three or more groups. There will be strong temptations to form these additional groups. For example, there may be several promising therapies, for example, dacarbazine and BCG for melanoma. Or, there may be some uncertainty as to the proper dose/duration of a single therapy, and it might seem advantageous to try several. Indeed, after a recent clinical trial in which aspirin was found not to lower the risk of recurrence in persons with a previous myocardial infarction (Aspirin Myocardial Infarction Study Research Group, 1980), some evidence has been accumulated to indicate that a beneficial effect might have been present had only a different dose of aspirin been used (Lorenz et al., 1984). Nonetheless, obtaining an adequate number of subjects to provide a valid test of at least one hypothesis should remain the first priority and, as indicated above, that number is often large.

QUESTIONS, CHAPTER 4

4-1. The following is paraphrased from an article published several years ago on the evaluation of the efficacy of a group of drugs:

We do not discuss randomized controlled trials of these drugs because our concern has been to evaluate the effects of the drugs on populations rather than on individuals. Thus, although randomized controlled trials provide unique information on the effects of drugs, they would concern us only if the study groups were representative of defined subgroups of the general population. Such representation is rarely possible in these trials, which generally involve volunteers.

You disagree with the opinion expressed above. Why?

4-2. After obtaining informed consent, 100 patients with cancer X were randomly assigned to one of two therapeutic regimens, surgery or chemotherapy. It turned out that while all 50 patients assigned to chemotherapy received their treatment, 14 of the 50 assigned to receive surgery did not. Five became too ill between the
0.20), then it will require fewer than 100 per group) that for between-group differences or more patients per group the benefit of the therapy required to determine it.

Clinical trials to determine two therapeutic regimens, e.g., a very large number of patients (the differences expected) are split into three or more groups to form these additional promising therapies, for example. Or, there may be a duration of a single therapy several. Indeed, after a sound not to lower the risk of myocardial infarction (Research Group, 1980), some patients that a beneficial effect different dose of aspirin been obtained, obtaining an adequate number of at least one hypothesis indicated above, that number published several years ago on the number of these drugs because our control populations rather than one. Is trials provide unique information us only if the study groups the general population. Such research generally involve volunteers.

If patients with cancer X were randomly assigned to therapy or chemotherapy. It turned out that therapy received their treatment, too became too ill between the two times of randomization and the scheduled date of surgery, and nine had second thoughts and declined to have the operation. However, all of these 14 were offered, and they accepted, chemotherapy. The survival of all patients was monitored for 1 year.

In comparing the relative efficacy of the two therapeutic regimens, which patient groups should be compared? Why?

ANSWERS

4-1. Randomized controlled trials always involve volunteers, so there will always be a need to generalize to a reference population that is not entirely similar to the study population; the former will contain potential nonvolunteers as well. The alternative approaches to evaluating drug efficacy, however—one or other of the nonexperimental designs—have a potentially more important drawback: Patients who receive and do not receive the drug are likely to have inherently different risks of the outcome that the therapy seeks to prevent.

So, in our search for internally valid studies we pay particular attention to the results of experimental studies, and we certainly do not eschew them. The process of generalizing to a reference population may still be tricky, but it cannot even be begun until there is confidence that the comparison among the study subjects themselves has some validity.

4-2. The survival of the 50 patients originally assigned to receive chemotherapy should be compared with that of the 50 patients originally assigned to undergo surgery. If these groups are not preserved in the analysis, and the analysis is based instead on therapy received (or on an analysis that deletes the “crossover” patients altogether), a biased comparison could easily emerge. The patients assigned to undergo surgery but who did not do so may be quite atypical with respect to survival (note that five such patients were too ill). Failure to keep them in the surgery group would probably make surgical therapy appear efficacious in the treatment of cancer X in the absence of any true difference between results for surgical therapy and chemotherapy.

REFERENCES


Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents: JAMA 1975; 231:583–608.


Hypertension Detection and Follow-up Program Cooperative Group: Five-year findings of the Hypertension Detection and Follow-up Program: I. Reduction in mortality of persons with high blood pressure, including mild hypertension. JAMA 1979a; 242:2562–2571.

Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the Hypertension Detection and Follow-up Program: II. Mortality by race-sex and age. JAMA 1979b; 242:2572–2577.


