Environmental Health Perspectives
Vol. 32, pp. 31-48, 1979

Clinical Trials in Cancer Research

by Edmund A. Gehan*

This is a review paper which gives a discussion of various aspects of clinical trials in cancer research. Since the conduct of the first randomized controlled clinical trial in cancer patients in the mid-1950's, substantial progress has been made in the utilization of the clinical trial technique for the evaluation of therapeutic efficiacy. The important elements of a protocol are given with some discussion of items to be considered in designing a protocol. The types of clinical trial (phase I, II, III) are defined, and the place of each phase of study in the context of the search for new treatments is delineated.

A comprehensive discussion is given of the elements in the comparative clinical trial (phase III), including objectives, considerations in planning (comparability of treatment groups, stratification of patients, feasibility and size of study, and prospective versus retrospective studies). Brief descriptions are given of designs for comparative clinical trials and a trial in oat cell lung carcinoma is discussed in some detail. Finally, some comments and references are given concerning the analysis of clinical trials.

Introduction

The randomized clinical trial was first used for the evaluation of cancer treatments in the mid-1950's (*1*). In only two decades, it has proven to be a useful way of evaluating the relative effectiveness of treatments and a substantial body of knowledge has been developed which provides objective data about cancer treatments. The organization of comparative clinical trials through the cooperative groups program sponsored by the National Cancer Institute in the U.S. and clinical trials organized in other countries have led to more uniform criteria for diagnosis, objective definitions of response, protocols for the evaluation of treatments, and statistical evaluation of results.

The design of clinical cancer research studies represents a cooperative effort between clinicians and statisticians in the design and analysis of studies and between clinicians and patients in the conduct of studies. The clinical trial is a device for obtaining objective evaluations of the effectiveness of treatments. All would agree that the treatment of patients with cancer is not optimal and that better clinical studies are needed to define better forms of treatment.

In this paper, a description will be given of protocols in clinical studies, and the types of clinical trials being used in cancer research, with special emphasis on the comparative clinical trial. For the comparative clinical trial, discussion will be given to objectives, considerations in planning, designs for comparative studies, and analysis of clinical trials. The overall objective will be to give an introduction to the field with appropriate references to work that has been accomplished.

Some recent articles giving a discussion of general issues relating to cancer clinical trials are Gehan and Schneiderman (*2*), Livingston and Gehan (*3*), and Peto et al. (*4*).

Protocols in Clinical Trials

Since any scientific clinical study requires a well-developed experimental plan, all clinical trials should have a protocol which outlines the design and method of conduct of the clinical trial. A protocol is a written document which gives the purpose of the clinical study including the rationale, the method of administering the treatment(s), and details concerning the plan of investigation. The usual elements included in a protocol are (1) introduction and scientific background for the study; (2) objectives of study; (3) selection of patients; (4) design of study (including schematic diagram) (5) treatment programs; (6) procedures in event of response, no response, or toxicity; (7) required clinical and laboratory data; (8) criteria for evaluating the effect of treatment; (9) statistical considerations; (10) in-

*Department of Biomathematics, The University of Texas System Cancer Center, 6723 Bertner Boulevard, Houston, Texas, 77030.

formed consent; (11) record forms; (12) references; (13) study chairman or responsible investigator and telephone number.

The first requirement of a good study is that it be addressed to an important clinical question. If the design, conduct, and analysis are perfectly carried out but the question is trivial, then the study will be trivial. The protocol should include a statement of the objectives of the clinical study which are specific and well-defined. These may be divided into primary and secondary objectives, and the design of the study should ensure that at least all the primary objectives are achieved. For example, in a study of adult patients with acute leukemia, a clinical trial might be designed to compare treatments for inducing remissions, treatments for maintaining complete remissions, and whether different types of immunotherapy treatment would be useful after a patient has been in remission for one year. If the latter were considered a primary objective, then sufficient patients would have to be entered into study so that the required numbers of patients were on each immunotherapy treatment among patients having remissions lasting at least one year. Since only 35-40% of patients will achieve this status, it is evident that rather large numbers of patients may have to be entered into study to achieve a precise comparison of the immunotherapy treatments. However, if comparison of the immunotherapy treatments was considered a secondary objective, sufficient patients could be entered into study to achieve precise comparisons of remission induction and remission maintenance treatments and simply analyze whatever numbers of patients receive the different immunotherapy treatments.

It is usually not a good idea to plan a study that involves a large number of questions. Such studies are more difficult for investigators to carry out, require more monitoring of patients and completion of forms, are more subject to changes in study plan, and are often difficult to interpret clearly. It is usually better to have a simple design of study which is addressed to a small number of important questions.

The protocol should contain the preclinical and clinical rationale for doing the study in the context of a review of pertinent recent data in the tumors to be studied. The protocol should give the primary reasons for undertaking the study and the scientific basis for the choice of treatments.

The section on selection of patients should define which patients are eligible for study and which will be excluded. Ideally, the patients entering study should be considered as a representative sample from the population of possible patients who meet the selection criteria. Pathological confirmation of diagnosis is a requirement in nearly all studies and should be indicated in this section. Further details that might be given are: tumor site(s) or type(s) to be studied; type (or amount) of prior therapy permitted; age restrictions; requirements about renal and/or liver function and bone marrow status; and requirements concerning the measurability of lesions.

It is desirable to keep a log book of all patients seen during the study period who have a confirmed diagnosis of the disease and are possible candidates for study. The clinical investigator should record data on characteristics related to prognosis, such as age and stage of disease. If a patient is not entered into study, a reason should be given. A log book permits the clinical investigator to make statements about the characteristics of the population of patients who are candidates for study and the proportion of this population that has been entered into study. Changes in the selection criteria might be made if too many patients are being excluded.

The section on design of study should give a detailed statement of the course of a patient's treatment on the protocol. If the course is divided into parts, such as induction, consolidation, and maintenance, the procedures to be followed at each stage should be described clearly. It is helpful to include a schematic diagram giving the general design of the study with the treatments to be administered on specific days following the start of study.

The section on treatment programs is that most frequently referred to by physicians caring for patients and should give precise statements concerning the treatments to be administered. Treatment programs may be defined in terms of the achievement of a given endpoint for the patient, for example a dose of drug may be given according to a certain route and schedule until the white blood count performed at stated intervals reaches a level of $2000/mm^3$ or less. The important point is to have a well defined treatment program that can be followed by the physician.

The section on procedures in event of response, no response, or toxicity should indicate what is to be done to the patient in each circumstance. The general steps should be indicated by the design of the study, however the specifics in terms of alternative treatments, methods of handling toxicity, and requirements for removal from study should be given here.

The section on required clinical and laboratory data gives the tests required for all patients prior to entry into study and at designated time points during the study. It is convenient if a table summarizing these requirements is given in the protocol.

An important section of a protocol is that giving the endpoints for evaluating the effectiveness of treatment. In cancer clinical trials, typical endpoints are response (complete remission, partial remission,

no response), length of complete remission, and survival time. In some types of cancer, for example, advanced breast or lung cancer, the "no response" category is sometimes subdivided into stable disease, mixed response, and disease progression. For patients with advanced disease, achieving a state of "stable disease" may occur more frequently with one treatment than with another. For patients with limited disease, such as certain patients with breast cancer, treatment by surgery alone can remove all evidence of disease. Endpoints for analysis in this circumstance are "time to relapse" and "relapse." Relapse may be defined as local recurrence of disease or occurrence of metastatic disease.

The choice of the primary response variable has an important effect on the number of patients and length of study. If comparison is to be made between the percentages of patients disease-free two years after the start of the study, then the study must last at least two years plus the length of time required to enter sufficient patients into trial. Length of survival is clearly an important endpoint; however, patients sometimes die of causes not related to the disease under study, may receive other treatments after relapse on the study, or may have too long an expected survival for a clinical trial of reasonable length. In each clinical trial, various definitions of response should be considered for their meaningfulness and type of trial implied.

The section on statistical considerations should summarize the major objectives to be achieved in the study and give an estimate of the number of patients required or of how the study will be analyzed sequentially as it proceeds. Some designs for clinical trials comparing two or more treatments are the simple randomized design, the stratified randomized design, and the factorial design. The feasibility of the study should also be considered in this section; if estimates of the number of patients to be entered into study per month can be given based upon previous data, the estimates should be given. An estimate should be made of the total length of time required to accrue and follow sufficient patients to observe the relevant endpoints and achieve major objectives. Failure to give adequate consideration to the feasibility of study has resulted in the starting of clinical trials which had little or no hope of being completed with a definite statement about the effectiveness of treatments.

In the United States, there is a requirement for informed consent either from the patient or an individual responsible for the patient when the patient is a child. As far as possible, the patient should be informed of the design of the study and the rationale for its conduct so that proper consent can be obtained.

Copies of the record forms should either be part of the protocol or should be supplied in conjunction with it. The record forms should contain the required pre-treatment data and the data to be evaluated during the course of study for each patient. Sufficient information should be on the form so that an evaluation can be made of whether or not the patient followed the protocol properly and whether there was sufficient objective evidence for response or relapse. It is usually helpful to have forms with boxes for the recording of information that are self-coding, i.e., data can be entered directly into the computer by keypunching from the forms or entry of data via a computer terminal.

A protocol should be a self-contained document and not have references to other protocols for items such as definition of response or method of administering treatment. When a protocol has been well written, the final report of the study will be easier to write.

Types of Clinical Trials

Figure 1 gives a sequence of clinical trials for a new agent. The diagram is given in terms of a chemotherapeutic agent, but similar steps are involved in the development of new radiotherapy or surgery treatments. New agents may reach the clinical trial stage for various reasons: an indication of effectiveness against cancer in animal systems (mice, rats, dogs, and monkeys), as analogs of agents that have already proven effective, or as combinations of single agents known to be of some effectiveness in previous clinical studies.

Phase I Studies

The major objective of this trial is to determine the maximum safely tolerated dosage regimen for a given schedule of an agent in man. The regimen should be one that can be used in looking for therapeutic effects in later phases of study. A major aspect of the phase I trial is the elucidation of the nature of the agent's side effects, both qualitatively and quantitatively. The basic steps in the phase I study are: selection of a starting dose, selection of a method for dose escalation, and selection of a sample of patients to receive the agent.

Studies in animal systems provide a useful guide to the selection of the starting dosage in man. Freireich et al. (*5*) have shown that the maximum tolerated dose in man was comparable to that in five animal species (mouse, rat, hamster, dog, and monkey) when dosage was expressed per unit of surface area in square meters. Hence, phase I studies in man might be started at dosage levels of ⅓ or less of the

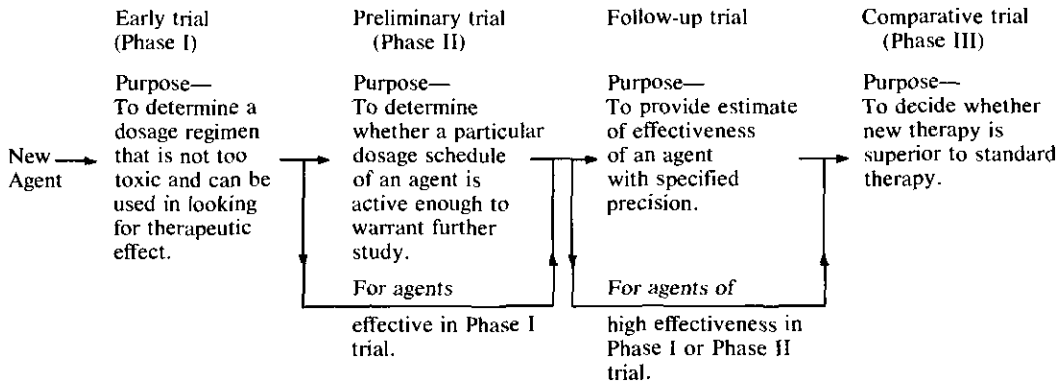

Fig. 1. The sequence of clinical trials for a new agent.

average maximum tolerated dose per unit surface area in the five animal species or a more conservative procedure would be to use ⅓ the dosage in the most sensitive animal species. These can only be taken as general guidelines, since as Freireich et al. (5) state: "It is emphasized and should be clearly understood that it is dangerous to attempt to extrapolate directly from animal toxicity data to maximum tolerated doses in man."

Selecting a method for escalating doses is not well standardized. On assuming that no toxic side effects have been observed at the starting dose level, schemes suggested for increasing dosage may be based upon a Fibonacci search technique (6) in which the dose is raised by decreasing increments, or a simple percentage change in dose scheme based upon toxicity observed. It is clearly desirable to minimize the number of steps required to reach some evidence of biological effect, since neither useful information nor benefit to the patient is obtained by administering doses that are too low; on the other hand, when some undesirable side effects have been observed at a particular dosage, one must be extremely cautious in escalating dosage. As a general guideline, it is advisable to have three patients on a study at each dose level and each should be observed until the period of anticipated risk of acute side effects has passed.

The patient population chosen for a phase I study can influence conclusions in a major manner, depending upon the sensitivity of the patients to the agent being studied. If only patients with very advanced disease are put on a phase I study, then the dosage regimen recommended might be too low for patients who were in an earlier stage of disease.

Every phase I study is at least in part a phase II study. That is, there should be a definite searching for evidence of therapeutic response in the phase I study.

Phase II Studies

The objective of the phase II trial is to determine whether the particular dosage regimen chosen in the phase I study is effective enough to warrant further study. An agent would certainly be of interest if it appeared substantially better than the existing best treatment; alternatively, it could be of interest if it had a lesser degree of activity but represented a different type of agent or if the agent had lower effectiveness but less toxicity.

Figure 1 indicates a category for a follow-up trial which has the objective of providing an estimate of effectiveness of the treatment regimen that has a certain precision. An estimate of the approximate effectiveness of an agent is desirable to have either from the phase II or the follow-up trial prior to the conduct of a phase III study.

A clinical trial is never a test of a drug or radiotherapy or surgery. It is a test of a treatment program when administered at a certain dosage, according to a certain schedule, and in a particular type of patient.

One of the following conclusions should be reached at the end of a phase II trial: (1) agent could be effective in X% of patients or more, or (2) agent is unlikely to be effective in X% of patients or more. If the objective can be stated in that manner, study of only a relatively small number of patients might be needed.

One approach to determining the minimum size of sample for phase II studies has been given by Gehan (7). Table 1 gives the number of patients necessary to have in a phase II trial to decide whether an agent is worthy of further study or is unlikely to be effective in X% of patients or more at given levels of rejection error. Rejection error β is the chance of failing to send an agent on to further study, when it should have been. Hence, if one is interested in an agent of

Table 1. Number of patients required for phase II trial of an agent for given levels of therapeutic effectiveness and rejection error

Rejection error β, %	Therapeutic effectiveness, %									
	5	10	15	20	25	30	35	40	45	50
5	59	29	19	14	11	9	7	6	6	5
10	45	22	15	11	9	7	6	5	4	4

20% effectiveness and is willing to accept a 5% rejection error, a sample of 14 patients is needed. This number has been derived by assuming that the true effectiveness of the agent is 20% and calculating that the chance of 14 consecutive failures is less than 5%. Consequently, if all 14 patients fail to respond, further study of the agent could be stopped because a sequence of 14 failures would occur less than 5% of the time if the true therapeutic effectiveness was 20%.

The numbers in Table 1 should be taken as guidelines because several simplifying assumptions have been made. First, response is assumed to be an all-or-none phenomemon. If complete remission is the endpoint, then nonresponse patients would be those who showed no biological effect at all and also patients who showed some degree of response that was not sufficient to qualify for complete remission. Response should be defined so that when one or more such responses have been observed, it is meaningful to study the agent further. Secondly, the numbers of patients in the table assume that the chance of response is the same for each patient. Based upon previous studies, it is known that the chance of response differs among individual patients according to prognostic characteristics. Consequently, the numbers in the table are those assuming an "average level" of effectiveness for each patient. If there is not a large amount of variation around the average level, the results in the table will be approximately correct and not likely to lead to difficulties in particular studies. Thirdly, no prior information concerning the agent is used in the planning of the study. If the agent has previously been included in a phase I study, it may be desirable to combine the estimates of effectiveness from the phase I and II studies, however no allowance for this is made in Table 1. Fourthly, if the agent has already been studied in other tumor types, there is presumably some correlation between chance of response in various tumor types; however no allowance has been made for this.

The decision rule for a phase II study using the number of patients in Table 1 is as follows: if all patients in the study do not respond, further study of the agent is halted; if one or more patients respond, additional patients are added to study to determine an approximate estimate of effectiveness. Agents with effectiveness substantially lower than that of interest have a substantial chance of being passed to further study. For example, if one enters 14 patients in search of an agent of 20% effectiveness (with 5% rejection error), there is more than a 51% chance that one or more responses will be observed even if the agent has only 5% true therapeutic effectiveness. The numbers of patients given in Table 1 represent the minimum needed to reject an agent for further study, given that no responses have been observed. Passing an agent on for further study when the true response rate is low might be acceptable when no agent of any real effectiveness exists for the given disease, but it may not be satisfactory for diseases for which agents of moderate to high order of effectiveness already exist.

There are alternative methods of specifying a number of patients for phase II studies. A method incorporating decision theory criteria has been given by Sylvester and Staquet (*8*).

When a dosage regimen has passed a phase II trial or has been of moderate effectiveness in a phase I trial, follow-up studies will be instituted to obtain a more precise estimate of effectiveness. Gehan (*7*) has given the number of additional patients to be studied corresponding to number of responses observed in the phase II study.

Some clinical investigators have suggested that two dosage regimens or two types of chemotherapy should be randomized in a phase II trial. In essence, the proposal is for the conduct of two concurrent phase II trials, since there is no objective of determining which of the two therapies is superior. Randomizing patients between therapies provides some control over the investigator's selection of patients for study. If this is likely to be a problem and if sufficient patients are available for conducting two studies at the same time, then randomization of patients in phase II studies might be worthwhile.

As all phase I studies have some elements of phase II, many phase II studies are concerned with phase I problems. In studies of a single agent, cumulative side effects may be observed in the phase II study that were not observed in the phase I study, e.g., studies of adriamycin and daunorubicin cardiac tox-

icity. Further discussion is given by Livingston et al. (*3*) of some of the clinical aspects of phase II studies.

The Comparative Clinical Trial (Phase III)

A comparative clinical trial is a planned experiment in human patients involving two or more treatments when the primary purpose is to determine the relative effectiveness of the treatments. A phase III trial might be begun when there is a preliminary estimate of effectiveness of a new therapy from prior phase I or II studies. A common circumstance for the phase III trial is to compare a proposed new therapy with the best standard treatment for the disease being studied. A phase III trial should begin before opinions have become fixed concerning the relative value of the new treatment; this would make it difficult or impossible to do a controlled trial. However, a phase III trial should not be undertaken when the treatment programs are likely to be modified frequently. In this circumstance, further patients should be added to the phase II study.

These general considerations suggest that the place for the phase III study is when the proposed new therapy has a potential for a small to moderate sized advance in effectiveness over the best standard treatment. If there is no possibility of an advantage for the proposed new therapy, then there is no real rationale for a phase III study. If the new therapy appears very significantly superior to the standard therapy based on phase I or II studies, then it may not be ethical to do a study in which patients would be randomized between the new and standard treatments.

Objectives

The type of comparative trial to undertake differs according to the precise meaning of "determining the relative effectiveness of the treatments." If the primary aim is to select the better of the two treatments for use in future patients and estimation of effectiveness is secondary, then some type of sequential study should be conducted in which the decision to continue study at any stage would be determined by the results available at that stage. As soon as it could be concluded that one treatment was superior to the other, the study would be stopped, even though at that point it might be that only imprecise estimates of the effectiveness of each treatment could be made. Alternatively, the objective might be to select the better treatment and have an estimate of the effect of treatment with some precision. With this combined objective of selecting the better treatment and estimating effectiveness, a sufficient number of patients should be entered on each treatment so that the effectiveness of each could be estimated with some precision. Additional patients might then be needed to satisfy the selection requirement.

As an example, a study by Freireich et al. (*9*) involved a randomized, double-blind comparison of 6-MP versus placebo in the maintenance of remissions in acute leukemia. The trial was a sequential one in which patients were paired at each institution according to complete or partial remission and patients were then randomized within pairs to 6-MP or placebo. A preference was recorded for 6-MP or placebo depending on the length of remission following each treatment, and there was a sequential stopping rule which permitted the study to be stopped after the accumulation of 18 pairs of patients. At that time, the median length of remission for 6-MP was 27 weeks compared with 9 weeks for placebo treatment. At the time the study was stopped, it was clear that 6-MP was the superior treatment, however only relatively imprecise estimates of length of remission could be made based upon 18 patients. Yet, the estimates of median length of remission were quoted by many clinical investigators without indication of the precision of the estimates. This is an example of a study in which the aim was to select the better treatment and estimate effectiveness, however the sequential plan was designed to permit only the selection of the better treatment.

For each particular clinical trial, the primary objectives should be considered carefully and the study design chosen to permit the achievement of major objectives in a reasonable period of time.

Considerations in Planning

Factors to be considered in planning clinical trials are: comparability of groups of patients, stratifications of patients, feasibility and size of study, and prospective versus retrospective studies.

Comparability of Groups of Patients. Having comparable groups of patients is a *sine qua non* of a controlled clinical trial (*10*). A comparative clinical trial should be so planned that the only explanation of a difference observed between treatment groups is a result of the treatment and not differences in types of patient on each treatment. This requires comparability of patients when entered into study, during the conduct of the study, and at the time of analysis when the study is completed.

Randomization of patients is a technique designed to achieve comparability of patients between treatment groups at time of entry into study. Randomization assures that patients will be comparable on the average with respect to factors influencing prog-

nosis. When patients are not comparable, procedures adjusting for differences in prognosis can be utilized so that appropriate tests of difference between treatments can be carried out (*11*). Procedures for randomizing patients are described by Gehan and Schneiderman (2).

In a retrospective study, patients on a proposed new therapy would be compared with patients treated in the past, a historical control group. When patient characteristics related to prognosis are known, patients from the historical control series can be compared with the current series of patients using regression techniques to adjust for differences in prognosis. Procedures for applying adjustment techniques in clinical trials are described in Gehan (*12*), and an application is made to studies of breast cancer. An assumption is needed in this analysis that the differences between treatment groups cannot be explained by factors associated with chronological time; hence, the use of retrospective studies may be expected to have more validity when the interval between groups in chronological time is short.

To achieve comparability, it is also necessary to manage patients on each treatment regimen in the same way. Decisions concerning the removal of patients from study or to stop treatment because of side effects to the patient should be made utilizing the same criteria. If some investigators have a preference for one of the treatments so that patients are maintained on it longer, or are classified as toxic only when the toxicity is very severe, results could be biased in the direction of the preferred treatment. One solution is to do a double-blind study in which neither the patient nor the physician is aware of the treatment being administered. In cancer clinical studies, double-blind studies tend not to be effective in reducing bias because different types of treatments generally lead to different types of toxicity and the types of toxicity are known to the physicians. Double-blind studies are most useful when the measure of response is subjective. The more objective the criteria for toxicity and response, the less the need for a double-blind study. In some circumstances, double-blind studies cannot be carried out, such as studies of some surgical or radiotherapeutic procedures.

Patients in each treatment group should be comparable with respect to the criteria applied in the analysis of the study. Hence, if response is defined as a 50% reduction in the sum of the products of the measured diameters of the tumors, then the same criteria should be applied in both groups. In addition to the investigator treating the patients, it is important to have either a study chairman or reviewing investigator evaluate the patient to assure the application of uniform criteria. Evaluation of response of the patient can be made with the evaluator being unaware of which treatment the patient received. In cooperative groups engaged in clinical trials, often each patient is evaluated by a committee and objective criteria are utilized. This tends to diminish the possibility of bias of a single investigator.

Stratification of Patients. Patients with any given type of cancer may differ in prognosis according to such characteristics as age, stage of disease, bone marrow status, and prior therapy. In adult acute leukemia, for example, patients can be categorized by age, infection status (yes or no), type of leukemia, platelet count, and hemoglobin value. Should important prognostic characteristics be utilized to define stratifications of patients within which patients would be expected to be comparable in prognosis? In a randomized study, patients could be randomized to the possible treatment groups within each stratum. Alternatively, a randomized clinical trial could be carried out by randomizing patients to treatments without regard to strata. A recent paper giving arguments for and against stratification in clinical studies is that of Brown (*13*).

Those arguing in favor of stratification state that the treatment groups will be more comparable than when prognostic factors are ignored; lack of comparability of groups can still be adjusted for in the analysis; not stratifying patients leads to tests of difference between treatment with less sensitivity and power; and the balancing of patients on each treatment within all subgroups makes it possible to estimate and test for interaction effects. Peto et al. (*14*) have argued that even when prognostic factors are ignored in making treatment assignments, randomization will tend to balance the treatment allocations within each stratification group. It can be shown that when there are N patients in a given cell, the amount of information achieved by complete balance, that is by assigning exactly one-half of the N patients to each treatment, can be achieved on the average by randomly assigning $N + 1$ patients to the treatment arms without regard to balance. Secondly, they argue that having different stratifications makes the design and conduct of a clinical trial more complex, possibly discouraging physicians from entering patients. Thirdly, it is pointed out that use of prognostic factors for balancing treatment groups does not eliminate the necessity for utilizing these variables in the statistical analysis. Hence, when multiple strata are part of the design of the study, these stratifications must be used in the analysis. Brown (*13*) finds the arguments against the use of prognostic factors in randomization of treatment assignments "logically valid but not persuasive." He argues that one should not randomize without stratification, a procedure that would perform well on the average,

when one can achieve balance by stratification. Though the loss of information per cell might be small, a study with a large number of cells would have a large loss of power. Assignment of patients to strata is a simple procedure and not likely to discourage investigators from entering patients. Brown (*13*) argues for stratification because of the increased power and precision of the study.

Pocock and Simon (*15*) give methods for achieving a balance of patients on each treatment with respect to prognostic factors. If the assignment of patients is being made at multiple institutions, it is nearly always desirable to balance patients at each institution; failure to do this will result in some investigators criticizing the experimental plan because of the divergences in number of patients on each treatment.

Feasibility and Size of Study. Each clinical trial should be considered for feasibility, number of patients required, and length of study. An important consideration in planning a clinical trial is the number of patients per year that might be expected to enter study. Having an estimate of this number and knowing the approximate follow-up period for observation of an endpoint in the average patient will make it easier to determine the number of patients to enter into study. Clinical trials should not be planned to last for too long a period, since interest and motivation of the clinical investigators tend to decrease with time. Also, alternative treatments are likely to be discovered and become candidates for comparative studies.

The size of a comparative clinical trial is usually determined by considering the clinical trial as a test of a null hypothesis versus an alternative. Sufficient patients are entered into study to provide an adequate test of that hypothesis in terms of significance level and power. For example, in a clinical trial of A versus B, the null hypothesis might be that there is no real difference in the response rate to A or B, while the alternative hypothesis might specify that there is some real difference in response rates. The one-sided alternative specifies the direction of the difference (response rate to A being higher than that to B, say), whereas a two-sided alternative specifies that either A or B may have the higher response rate.

Tables 2 and 3 give the number of patients required in each of two treatment groups to test for given differences in response rate at a certain significance level and power of test. Table 2 gives the sizes of study needed for the one-sided test, i.e., when the clinical trial is designed to test whether a new treatment is better than a standard. Table 3 gives the number of patients required for a two-sided test, that is which of the two treatments is better. The clinical investigator should specify the difference in response rates to be determined, an estimate of the response rate for one of the two treatments, the level of statistical significance (α), the value of the desired power ($1 - \beta$), and whether the test should be one- or two-sided. The level of significance α is the chance of false positive error, that is that the trial results in a statement that there is a real difference between treatments when in fact there is none. The chance of a false negative error is β, that is, the trial results in a statement that there is no difference between treatments when, in fact, there is. The power of the test is defined as $(1 - \beta)$.

As an example using Table 2, suppose that 40% of patients are expected to respond to standard treatment, the clinical trial is to be conducted to determine whether a proposed new treatment results in a response rate that is 20% higher and it is desired to use a statistical significance level of 5% and a power of test of 80%. From Table 2, the number of patients needed in each group to meet these requirements is 76. If the desired power is 90% and the other requirements are the same, then 105 patients are needed in each group. The higher the power needed, the larger the number of patients. Also, the higher the significance level (i.e., the smaller the probability that the observed difference is due to chance) used, the larger the number of patients needed. For example, if the clinical trial is supposed to be conducted using a significance level of 1% and a power of 95%, 195 patients would be needed in each group.

If the set-up for the clinical study was the same, except that it was desired to conduct a two-sided test, then reference to Table 3 shows that 97 patients would be needed in each treatment group (significance level 5%, power of test 80%). Clinical trials for testing two-sided alternative hypotheses always require larger sample sizes than equivalent trials for one-sided alternatives.

If the clinical trial were conducted strictly as a hypothesis-testing procedure, then the given numbers of patients would be entered into each treatment group, response rates determined, and a statistical test carried out to determine whether the difference in outcomes was statistically significant at the 5% level or not. Clinical trials are rarely, if ever, conducted with this degree of rigidity so that the sample numbers of patients should be considered as guidelines in the planning stages of the study. Sometimes, in fact, the procedure for determining sample size is applied in reverse. For example, if 100 patients are expected to enter study per year in a one-sided comparison of A versus B with a response rate of 40% for B, at about one year when 96 patients have been entered, a 25% advantage in response rate for treatment A could be detected (statistical significance level 5%, power of test 80%). At about 1½ years,

Table 2. Number of patients needed in an experimental and a control group for a given probability of obtaining a significant result (one-sided test).

Smaller proportion of success (P_1)	Number of patients at various larger minus smaller proportion of success ($P_2 - P_1$)[a]													
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
0.05	330	105	55	40	33	24	20	17	13	12	10	9	9	8
	460	145	76	48	39	31	25	20	19	15	13	11	10	9
	850	270	140	89	63	37	41	34	21	25	22	18	16	14
0.10	540	155	76	47	37	30	23	19	16	13	11	11	9	8
	740	210	105	64	41	38	30	24	20	17	15	12	11	10
	1370	390	195	120	81	60	46	41	35	28	24	20	17	16
0.15	710	200	94	56	43	32	26	22	17	15	11	10	9	8
	990	270	130	77	52	43	34	26	23	19	16	12	11	10
	1820	500	240	145	96	69	52	41	37	30	24	22	18	16
0.20	860	230	110	63	42	36	27	23	17	15	12	10	9	8
	1190	320	150	88	58	46	36	29	23	18	16	12	11	10
	2190	590	280	160	105	76	57	44	39	30	27	22	18	16
0.25	980	260	120	69	45	37	31	23	17	15	12	10	9	—
	1360	360	165	96	63	46	38	30	23	18	16	12	11	—
	2510	660	300	175	115	81	60	46	40	33	27	22	17	—
0.30	1080	280	130	73	47	37	31	23	17	15	11	10	—	—
	1500	390	175	100	65	46	38	30	23	18	16	12	—	—
	2760	720	330	185	120	85	61	47	39	32	24	20	—	—
0.35	1160	300	135	75	48	37	31	23	17	15	11	—	—	—
	1600	410	185	105	67	46	38	30	23	18	15	—	—	—
	2960	750	340	190	125	85	61	46	39	30	24	—	—	—
0.40	1210	310	135	76	48	37	30	23	17	13	—	—	—	—
	1670	420	190	105	67	46	38	30	23	17	—	—	—	—
	3080	780	350	195	125	84	60	44	37	28	—	—	—	—
0.45	1230	310	135	75	47	36	26	22	16	—	—	—	—	—
	1710	430	190	105	65	44	36	26	20	—	—	—	—	—
	3140	790	350	190	120	81	57	41	34	—	—	—	—	—
0.50	1230	310	135	73	45	36	26	19	—	—	—	—	—	—
	1710	420	185	100	63	41	35	24	—	—	—	—	—	—
	3140	780	340	185	115	76	52	39	—	—	—	—	—	—

[a]Modified from Cochran and Cox. (36). Upper figure: test of significance at 0.05 for α, power equals 0.8 for $(1 - \beta)$; middle figure: test of significance at 0.05 for α, power equals 0.9 for $(1 - \beta)$; lower figure: test of significance at 0.01 for α, power equals 0.95 for $(1 - \beta)$.

after 152 patients had been entered, it would be possible to detect a 20% advantage for treatment A over B (statistical significance level 5%, power of test 80%). At about 2¾ years when about 270 patients have been entered, it would be possible to detect a 15% advantage for treatment A with the same significance level and power requirements. The choice of sample size and length of study could be determined by the clinical investigators to be consistent with the response rate that might be expected for treatment A.

Another way of using Tables 2 and 3 is to obtain a rough estimate of the difference in response rates that had a reasonable chance of being detected in a given clinical study. For example, suppose a clinical trial has been conducted with 25 patients in each treatment group, the response rate of the standard treatment is about 45%, the response rate for the proposed new treatment is 50% and the research paper reporting the results of the study states "there is no evidence in this study of a real difference in response rates between A and B at the 5% level of statistical significance." Reference to Table 3 in the row for smaller proportion of success equal 0.45 and the column for larger minus smaller proportion of success equal 0.40 shows that 25 patients in each group would have had reasonable power (80%) for detecting a 40% difference in response rates. The column for larger minus smaller proportion equal 0.45 shows that 24 patients in each group would have had a 90% power for detecting a 45% difference in response rates. Hence, a study with only 25 patients in each group might have missed a 10-25% advantage in response rates for treatment A. Rough interpola-

tion of the tables could be used when the actual number of patients in the study does not correspond to numbers given in the table.

George and Desu (*16*) give the number of patients needed in clinical trials for comparing two survival distributions when survival is assumed exponentially distributed. The numbers of patients specified are the numbers of failures that must be observed in each treatment group. Also, they give estimates of the length of study required, assuming certain entry rates of patients per year. Table 2 or 3 could be used for this problem also assuming that the clinical investigators wish to determine whether or not a given difference in proportion surviving a given period of time exists between the two treatment groups.

Armitage (*17*) gives some closed sequential plans for comparing treatment groups. These plans are applicable when it is desired to have the option of continuing study at any stage based upon the analysis of results at that stage. The major reason for conducting a sequential trial should be to enable the clinical investigators to stop the study as soon as the observed difference is large enough to reject the hypothesis that no real difference exists. Sequential plans for both quantitative variables (such as degree of response) and qualitative variables (such as complete remission or not) are given. Sequential plans are "closed," meaning that there is a fixed upper limit for the number of patients. This is especially desirable in clinical trials.

Prospective versus Retrospective Studies. An important issue in planning a comparative trial is whether two or more treatments should be compared in a prospective, concurrently controlled, randomized study or whether the new treatment could be compared with a historical control group. There

Table 3. Number of patients needed in an experimental and a control group for a given probability of obtaining a significant result (two-sided test).

Smaller proportion of success (P_1)	Number of patients for various larger minus smaller proportion of success ($P_2 - P_1$)[a]													
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
0.05	420	130	69	44	36	31	23	20	17	14	13	11	10	8
	570	175	93	59	42	37	31	24	21	18	16	13	12	11
	960	300	155	100	72	54	42	38	33	27	24	20	18	16
0.10	680	195	96	59	41	35	29	23	19	17	13	12	11	8
	910	260	130	79	54	40	36	29	24	20	17	16	13	11
	1550	440	220	135	92	68	52	41	38	32	26	23	19	17
0.15	910	250	120	71	48	39	31	25	20	17	15	12	11	9
	1220	330	160	95	64	46	40	31	26	22	18	16	13	11
	2060	560	270	160	110	78	59	47	41	35	29	24	21	18
0.20	1090	290	135	80	53	42	33	26	22	18	16	12	11	9
	1460	390	185	105	71	51	43	33	28	23	18	16	13	11
	2470	660	310	180	120	86	64	50	44	36	27	24	21	17
0.25	1250	330	150	88	57	44	35	28	22	18	16	12	11	—
	1680	440	200	115	77	56	45	36	29	23	18	16	12	—
	2840	740	340	200	130	95	68	52	45	36	29	24	19	—
0.30	1380	360	160	93	60	44	36	29	22	18	15	12	—	—
	1840	480	220	125	80	56	46	36	29	23	18	16	—	—
	3120	810	370	210	135	95	69	53	45	36	29	23	—	—
0.35	1470	380	170	96	61	44	36	28	22	17	13	—	—	—
	1970	500	225	130	82	57	46	36	28	22	17	—	—	—
	3340	850	380	215	140	96	69	52	44	35	26	—	—	—
0.40	1530	390	175	97	61	44	35	26	20	17	—	—	—	—
	2050	520	230	130	82	56	45	32	26	20	—	—	—	—
	3480	880	390	220	140	95	68	50	41	32	—	—	—	—
0.45	1560	390	175	96	60	42	33	25	19	—	—	—	—	—
	2100	520	230	130	80	54	43	32	24	—	—	—	—	—
	3550	890	390	215	135	92	64	47	38	—	—	—	—	—
0.50	1560	390	170	93	57	40	31	23	—	—	—	—	—	—
	2100	520	225	125	77	51	40	29	—	—	—	—	—	—
	3550	880	380	210	130	86	59	45	—	—	—	—	—	—

[a]Modified from Cochran and Cox (*35*).

Upper figure: test of significance at 0.05 for α, power equals 0.8 for $(1 - \beta)$; middle figure: test of significance at 0.05 for α, power equals 0.9 for $(1 - \beta)$; lower figure: test of significance at 0.01 for α, power equals 0.95 for $(1 - \beta)$.

are vocal adherents of both viewpoints and the arguments for and against such studies will be summarized here. The inherent difficulty is that the human patient is the experimental unit in a clinical trial and the design of the study should take due account of this in contrast with studies in experimental animals.

Byar et al. (*18*) state that "randomized clinical trials remain the most reliable method for evaluating the efficacy of therapy." The major advantages of randomization are first that "bias is eliminated in the assignment of treatments (which) means that treatment comparisons will not be invalidated by selection of a patient of a particular kind, whether consciously or not, to receive a particular form of treatment." Secondly, "randomization tends to balance treatment groups in covariates (prognostic factors), whether or not these variables are known. This balance means that the treatment groups being compared will in effect tend to be truly comparable." Thirdly, "randomization guarantees the validity of the tests of significance that are used to compare the treatments." With respect to ethical considerations, Byar et al. (*18*) dismiss ethical objections to randomizing patients because physicians may disagree about what the best treatment for a patient is in certain circumstances. A physician in a randomized clinical trial has made an intellectually honest admission that he does not know the best form of therapy. Chalmers and Shaw (*19*) argued that "random allocation of patients in a scientific trial is more ethical than the customary procedure, that of trying out a new therapy in an unscientific manner by relying on clinical impression and comparison with past experience." Byar et al. (*18*) do not make a suggestion about ethical implications when during the conduct of a study results are accumulating in favor of one treatment. In this circumstance, a physician may wish to not randomize his patients, if one treatment is substantially superior to another. Reference is made to the suggestion of Shaw and Chalmers (*20*) that this problem is minimized if the results of the trial are not made known to participating physicians until a decision has been reached by an advisory committee responsible for stopping the trial.

Cogent arguments for conducting nonrandomized clinical trials in some circumstances have been given by Gehan and Freireich (*21*), Gehan (*22*), and the limitations of the randomized clinical trial have been discussed (*23*). Proponents of the nonrandomized study have argued that all knowledge is dependent on historical data so there is good reason to consider use of historical control patients. Even proponents of randomized controlled studies must accept some historical data, in particular, their own; otherwise, their completed studies would have no predictive value for the future. Nonrandomized studies designed to achieve the same objectives as randomized studies require substantially fewer patients. If an investigator is studying A (new treatment) versus B (standard treatment) in a nonrandomized study and is willing to assume that sufficient data are available so that the response rate to the standard treatment can be taken as a fixed quantity, say P, then the number of patients required to compare response rates of A versus B at a given significance level and power of test is only one-fourth that for an investigator who randomizes patients equally to A and B (*21*). Thus, if patients are entered consecutively on treatment A over a given period of time rather than being randomized to A versus B in a study with the same objectives, then either the nonrandomized study will be completed more quickly or a test of difference of response rates will be accomplished that is more sensitive to smaller differences in response rates.

The major difficulty in conducting the nonrandomized study has been pointed out by Byar et al. (*18*) when they state that "using a control group chosen by any method other than randomization requires the assumption either that the control and treatment groups are identical in all important variables except the treatment under study or that one can correct for all relevant differences. In the latter case, one must assume that all factors affecting prognosis are known." Regression models can be used to determine factors related to prognosis in clinical studies (*11*), and either logistic regression or Cox's regression model (*24*) can be used to test for treatment differences adjusting for differences between groups with respect to prognostic factors. When the time interval between the historical control and the current study is short, the assumption that other factors (such as change in diagnostic procedure, supportive therapy, investigators) are not likely to explain treatment differences is reasonable.

Since a comparative trial is usually not started unless there is preliminary evidence in phase I or II studies suggesting that the new therapy is at least as good or possibly better than the old, clinical investigators are usually concerned with one-sided statistical tests, that is whether the new therapy is better than the standard or not. If this is so, then the question should be raised whether it is ethical to enter patients on the standard treatment when there is little or no chance that it could be better than a new therapy.

Atkins (*25*) succinctly stated the ethical responsibility of the statistician involved in planning a clinical trial. Simply stated, it is if the statistician was willing to have himself or a member of his family as a patient in a randomized clinical trial, then it is ethical.

Otherwise, it is not. Each clinical trial being planned should be considered from this viewpoint before starting; trials should not be begun unless one would be willing to be a participant in the study in appropriate circumstances.

In nonrandomized studies, there is no ethical dilemma for the clinical investigators either in having to decide whether to randomize patients or whether to stop a study early when results are favoring the new treatment. In the nonrandomized study, the clinical investigator should always be administering the treatment program that is considered most likely to succeed for his patients, comparisons of results being made with previous studies. It will be easier to recruit patients for such studies, since all patients presenting for the trial can be offered the proposed new treatment, rather than only a random chance of receiving it. During the conduct of the study, accumulating results on the new treatment, whether favorable or unfavorable to the proposed new treatment, cause no special difficulties. If favorable, the trial is simply continued until precise estimates of effectiveness are available for the new treatment. If unfavorable, standard statistical procedures for comparing a new versus a standard therapy can be utilized for comparing treatments; the trial can be stopped when the proposed new therapy has little chance of being superior to the standard treatment.

Those favoring randomized studies have argued that such studies are inherently more convincing to other clinical investigators. It is argued here that no single clinical study, whether randomized or not, will be completely convincing until results have been confirmed either by the same or other clinical investigators. Confirmation of results will be easier to seek when some promising results have been obtained on the proposed new therapy.

The main advantages for nonrandomized comparative studies are: there is no need for the use of a relatively inactive or poor standard therapy as a concurrent control; studies of a new therapy in consecutive patients with comparison to a historical control group are accomplished with fewer patients and more quickly; there is assurance for the patient and physician that the patient is being offered the treatment that the physician considers best; and the reasonableness of this approach for tumor types with low accrual rates. To validly compare results in a nonrandomized study with patients from a historical control group requires knowing the factors that influence prognosis and using regression methods to adjust for possible differences in prognostic factors. Finally, it must be assumed that differences which do exist between groups (such as chronological time, institution, availability of supportive therapy, etc.) are not sufficiently strongly related to prognosis to explain the differences between treatments observed.

Randomized studies may be preferred to the nonrandomized when: the clinical trial is the first one in a particular tumor type, so there is no basis for choosing a historical control group; there is a long time interval, say, five years or more, between the current and past studies so that substantial differences related to chronological time may exist; or when only small differences are expected between the new and old treatments and a large number of patients are available so that the ethical dilemma does not apply.

There is certainly a valid role for both prospective and retrospective comparisons in phase III studies. Neither the prospective nor the retrospective technique is better in all circumstances. Conducting objective, scientific clinical trials in which observations are recorded which differ significantly from those made in the past forms the basis for new knowledge and therapeutic advances in the treatment of cancer patients.

In circumstances when no firm decision can be reached regarding whether a study should be randomized or not, a compromise solution is sometimes chosen in which patients are allocated to the proposed new versus standard therapy on a 2:1, a 3:1 or some other basis. Sufficient patients should be studied on the control group so that a meaningful comparison can be made with previous results for the control group to determine whether any real changes have taken place. If there have been no significant changes, then results for the control group can be combined between the current and previous studies for comparison with the proposed new therapy. Pocock (*26*) has discussed the issues in combining randomized and historical control groups in clinical studies.

Designs for Comparative Studies

The earliest comparative clinical trials came about by chance, and Bull (*27*) gives some examples. Designs for comparative clinical trials are generally rather simple from the statistical viewpoint. The complexity in clinical trials tends to arise from the multiple observations made on each patient, the difficulties encountered in following the protocol in some patients, and in deciding whether to terminate a clinical study depending upon observed results. Complex designs have not generally been utilized because the delicate balancing features of such designs are more likely to be upset by missing observations or complications created in patient management.

Simple Randomized Design. The simplest situation is one in which patients are randomized to two or more treatments without any attempt to group patients by prognostic characteristics. As patients become available for study, they are assigned to one of the treatments by a formal random procedure. Noncomparability between groups must be accounted for in the analysis. This is the design of choice when patient characteristics related to prognosis are not known or when patients in only a single category are being studied. Usually, the randomization is restricted so that after a certain number of patients have been entered, an equal number of patients have been assigned to each treatment.

Stratified Random Design. If patients can be grouped into prognostic categories or strata such that differences in outcome may be expected among strata, then stratifications may be defined and patients assigned to one of the treatments within each stratum by random allocation. The simplest situation is that in which the strata are pairs of patients and each patient in a pair receives treatment A or B by a random allocation (paired comparison design).

In this type of design, too many stratifications of patients should not be defined. For example, in a study of acute leukemia where age, infection status, type of leukemia, platelet count, and hemoglobin value have been identified as related to prognosis, too many categories of patients would be defined if each of these characteristics were used in defining stratifications. In the limit, as the number of stratifications grows large, there would only be one patient in each stratum and the advantages of stratification would be lost. As a rule of thumb, no more than eight stratification categories should be defined, and preferably four, since differences among up to four prognostic categories would account for major sources of variation and sufficient patients are usually available to have a substantial number of patients within each stratum.

The advantages and disadvantages of stratification have been discussed previously.

Cross-Over Designs. A cross-over design is a combination of the simple randomized and paired comparison design in which each patient is used as his own control. A common way of utilizing this design is to administer the sequence of treatments A followed by B to half the patients and the sequence B followed by A to the other half. A patient is assigned to one of the two sequences by a random allocation, each treatment being administered when the patient's disease is in a comparable state.

In cancer clinical trials, there are practical difficulties with the cross-over design, since some patients may not survive long enough to receive both treatments; alternatively, the patient may have an excellent response to the first treatment so that a long period of time would be required before the second treatment could be given. Because of this, it is nearly always true that sufficient patients will be accumulated in the first phase of study to obtain a sensitive comparison of A versus B in different groups of patients before substantial numbers of patients are available who have received both treatments in the sequence. Of course, there is also the possibility of carryover effects in which the response to the first treatment in the sequence is related to the chance of response in the second phase. Cross-over designs of A versus B in the same patient in cancer clinical studies have been used to confirm the results of comparing A to B in different patients. In studies of acute leukemia by Frei et al. (*28*) and Freireich et al. (*9*), comparisons of treatment effectiveness within patients confirmed the results of comparisons between groups of patients.

Factorial Designs. A common situation in clinical studies occurs when a number of factors are to be studied for their relationship to response. In the 2×2 factorial design, for example, two treatments (or factors) each administered at two levels, say high and low dosage, may be studied for their relationship to response. The treatment-dosage combinations define four treatment combinations to which patients would be randomized. Such trials can be used for testing possible interaction effects, that is whether the difference in effectiveness between treatments was maintained at both levels of dosage. Similarly, differences between high and low dosage could be examined for each treatment.

An example is given of a factorial design in the next section.

Example of Clinical Trial

It is useful to discuss the planning of an actual clinical study to be aware of issues that should be considered in the planning of studies. This example relates to a proposed clinical trial by Dicke et al. (personal communication), though this study has not yet begun. The example is concerned with a study of patients with oat cell lung carcinoma; factors to be studied are: protected environment status (PE) (yes or no), marrow transplant (MT) (yes or no), and chemotherapy treatment (two types). Based upon the accrual of patients in previous studies, it was estimated that 60 patients per year would be eligible for study; however approximately one-third of the patients would not be eligible for marrow transplant because of involvement of the bone marrow or clinical condition. The major endpoints for the evaluation of patients were to be complete remission rate and survival time after the start of study.

Table 4. Design of study of oat cell lung carcinoma.[a]

		Protected environment (PE)		
		No	Yes	
Marrow transplant (MT)	No	30	30	60
	Yes	30	30	60
		60	60	120

[a]Stratifications of patients: limited or extensive disease ambulatory or nonambulatory; treatment groups and number of patients entered in 3 years: All patients receive chemotherapy.

Based upon previous studies of chemotherapy in oat cell lung carcinoma at M.D. Anderson Hospital, a complete remission rate of 45% had been observed with 50% of the patients surviving one year. It was expected that patients in a protected environment and receiving a marrow transplant with the same type of chemotherapy might be expected to have a superior complete remission rate and percentage of patients surviving one year. Because of various factors including expected period of support for the study, it was desirable to have some results within three years after the start of the study. In this period of time, it was expected that 180 patients would be available for study, 120 of whom would be eligible for the marrow transplant program. Patients with oat cell carcinoma can be stratified according to limited versus extensive disease and ambulatory status or non-ambulatory status.

A proposed factorial design for this study is given in Table 4 with the expected numbers of patients on each treatment combination over the three year period. Because of the limited number of patients available and the number of treatment combinations to be studied, it was decided that all patients would receive the same chemotherapy program. The objective of the study was to determine whether any of the treatment combinations yielded a significant improvement over the 45% complete remission rate and/or 50% of patients surviving one year.

For the purpose of planning the study, it was assumed that the effect of the marrow transplant would be similar in patients in or out of the protected environment so that a total of 60 patients receiving a marrow transplant over the three year period could be compared with 60 patients not receiving a marrow transplant. The same type of assumption was needed for protected environment patients so that it was reasonable to compare the 60 patients treated in the environment with 60 patients treated outside the environment.

Table 5 gives the statistical considerations for the study. The simplifying assumption was made in determining these figures that results could be combined from the four stratifications of patients to achieve an overall complete response rate and percentage of patients surviving a given period of time. It was anticipated that the percentage of patients in the four stratification categories would be similar to that in previous studies.

In Table 5, the first comparison is between the total of 180 patients with oat cell carcinoma in the three year period compared with the baseline figures. In making comparisons with previous results, patients not eligible for the marrow transplant must be included. Since it might be expected that patients not eligible for a marrow transplant would be less favorable than other patients, comparing results in only the 120 patients eligible for marrow transplant with previous results could be subject to severe criticism. It could rightly be stated that a more favorable group received the PE and MT and hence, overall comparisons were not meaningful. Comparing results with the 180 patients with baseline data shows that the power would be 80% for detecting an 8% improvement in either CR rate or percentage surviving one year (statistical significance level 5%) and a 90% power for detecting a 15% improvement.

Table 5. Oat cell lung carcinoma: statistical considerations for study.[a]

Comparisons (one-sided)	Total patients	Difference to be detected, %	Significance level (α), %	Power ($1 - \beta$), %
180 patients vs. baseline	180	8	5	80
		15	5	90
PE vs. No PE or MT vs. No MT	120	22	5	80
		28	5	90
PE-MT vs. PE-No MT or other subgroups	60	35	5	80
		40	5	90

[a]Endpoints: complete remission rate, survival; baseline data: 45% CR rate; 50% survive 1 year.

Comparison of PE versus no PE or MT versus no MT would be made based upon 60 patients in each group or a total of 120 patients. This number of patients would be sufficient to detect whether a PE or the MT resulted in a 22% improvement in prognosis compared with the absence of either of these factors (statistical significance level 5%, power of test 80%).

With the proposed design, there would be 30 patients within each subgroup, i.e., receiving both PE and MT, etc. Comparisons of results among subgroups of patients would give an indication of whether or not there were interaction effects. Comparisons between any two subgroups of 30 patients would give an 80% chance of detecting a 35% advantage in prognosis in one group and a 90% chance of detecting a 40% difference (statistical significance level 5%). Though the design of the study is not very sensitive to differences among subgroups, it might be sufficient to determine whether the directions of MT and PE effects were different depending upon the presence or absence of the other factor.

Though this study has not yet been conducted, discussion of the various factors involved should give an appreciation of the issues to be considered in the planning of an actual clinical trial.

Analysis of Clinical Trials

In the analysis of any comparative clinical trial, there are two aspects: estimation of treatment effects and toxicity, and a test of whether or not there is a real difference between treatments with respect to the major endpoints being studied. The relative emphasis on estimation or testing depends upon the objectives of the trial.

In any moderately large clinical trial, some patients will be entered into study who were not eligible and other patients will not follow the treatment program outlined in the protocol for some reason. As a preliminary step in the analysis, patients should be put in appropriate categories depending upon their status. Some patients may be excluded from all results, while others might be included in some calculations. It is recommended that patients first be classified according to whether or not they were eligible for study according to the selection criteria and patients that are not eligible should be excluded from all analyses.

If the analysis being performed is an interim analysis during the conduct of the study, the eligible patients can be classified into those who are too early to evaluate and those who have been evaluated for response and toxicity. Patients who are too early to evaluate must be excluded from all analyses. Patients who have been evaluated may be classified as: evaluable, partially evaluable, or not evaluable. Patients are generally put in the not evaluable category because of a major protocol violation (e.g., treatment not given, treatment program deviated from in a major way, etc.), inadequate data or other reasons. Patients may be classified as partially evaluable because of early death, lost to follow-up shortly after the study was started, refused further treatment, inadequate trial due to toxicity, or other reasons. It is recommended here that patients be identified who have problems associated with their treatment so that analyses can be performed including or excluding patients in certain categories. At our institution, analyses are generally performed first on those patients who are fully or partially evaluable, that is excluding patients that are not evaluable, not eligible, or too early to evaluate. In comparative studies, differences in response rate between treatments may depend upon the categories of patients excluded from the analysis. A recommended procedure is to calculate the differences in endpoints between treatments in more than one way. If all ways of comparing treatments lead to the same conclusion, then there can be reasonable confidence that the conclusion is correct. If there is a difference between treatments that is statistically significant using one denominator (say all evaluable patients) but is not statistically significant for another denominator (say all eligible patients), then the conclusion regarding the study depends upon the patients that were partially or not evaluable. In such a circumstance, the study should be recorded as inconclusive and further patients entered until a conclusive statement can be made. Peto et al. (*14*) argue that "rigorous entry criteria are not necessary for a randomized trial, but rigorous follow-up is. Even patients who do not get the proper treatment must not be withdrawn from the analysis." While this may be an appropriate procedure for determining whether it is better to be randomized to one treatment or another, an objective of a clinical study is to estimate the effectiveness of a given treatment. By including patients who may not have received the treatment at all or who received it in a markedly different manner from that specified in the protocol, one is likely to obtain a biased estimate of the effectiveness of a given treatment. Including all patients randomized is a better general procedure than only considering "evaluable patients," however the recommendation here is to calculate differences in endpoints in multiple ways.

In all comparative trials, it is important to check that patients entered on each treatment are comparable with respect to factors that might influence prognosis. This is important both for randomized and nonrandomized studies, and an analysis should

be carried out by use of adjustment procedures, usually regression procedures, when patients are not comparable with respect to important characteristics. The use of regression procedures to adjust for prognostic characteristics is described for cancer clinical trials in Gehan (*12*).

In estimating treatment effects, the most meaningful statistics for the physicians are the percentages of patients having objective responses to each treatment, disease-free survival curves, survival curves, and other appropriate measures of endpoint in clinical studies. An estimate of variability of appropriate statistics should be given, such as standard error of the percentage responding or standard error of the proportion disease-free at two years. Reporting average values plus or minus two standard error limits give approximate 95% confidence limits for a true value and some measure of how precisely effects have been estimated. Peto et al. (*4*) give procedures for estimating survival curves and tests for differences between them. Gehan (*29*) has also reviewed statistical methodology appropriate for survival studies.

In the frequentist method of testing for differences between treatment groups, a hypothesis is set up which indicates no real difference in the effectiveness of two treatments (null hypothesis) and the assumption is made that the null hypothesis is true. The significance level is the probability of obtaining a sample difference as great or greater than that observed, assuming that the null hypothesis was really true. When the probability is low, say under 0.05, most clinical investigators would be willing to reject the hypothesis of no real difference and declare that some real difference exists in the measure of response. Unfortunately, some clinical investigators adhere too closely to 0.05 as the statistical significance level. If the observed difference has a statistical significance of 0.08, it is declared "not significant" whereas if it is 0.04, it is declared as "statistically significant." Since the difference in observed evidence against the null hypothesis between a significance level of 0.08 and 0.04 must be small, it is better to interpret significance levels, or *P* values, as a measure of the strength of the evidence against the null hypothesis.

An overemphasis on significance level should be avoided. Zelen, in an article by Cutler et al. (*30*), pointed out that "the number of scientific papers that use statistical methods for window dressing is increasing. It appears that the *P* value next to a contingency table is beginning to mean what the 'Seal of Good Housekeeping' means to the housewife."

With the advent of the use of the computer for analyzing clinical studies, it has become possible to not only compare overall differences between treatments, but also to consider differences between treatments within many subgroups of patients. Often, it will be desirable to compare treatments within age groups or patients grouped by stage of disease, but there is a danger, as pointed out by Tukey (*31*), that multiple tests of difference will be carried out between groups, reporting only those statistically significant at the .05 level. Tukey points out that some adjustment should be made for the number of classes of patients, k, that would have been looked at seriously if the results for them had seemed favorable. It does not suffice for k to be only as large as the number of classes actually looked at; what is needed is the larger number of classes, each of which would have been looked at seriously if the results for them had turned out favorably. As a rough rule of thumb, Tukey suggests working with a statistical significance level of $0.05/k$ when multiple tests have been carried out.

Most clinicians understand the statement "Statistical significance is not necessarily equivalent to biological importance." The corollary of this statement which is quoted less often is: "Not statistically significant does not necessarily mean not biologically important." In randomized clinical trials involving a small number of patients in each group, there may be substantial differences between treatment groups that are "not statistically significant." Suppose, for example, that a randomized clinical trial has been conducted with 25 patients randomized to each of two groups. Suppose further that 11 patients or 44% responded in one group compared with 5 patients or 20% in the other group. These results lead to a chi-square value of 2.30 and a statistical significance level between 0.05 and 0.10. Results of this trial could be reported as "not statistically significant," while there was a 24% difference in response rate between the two treatments. To properly interpret the results of such a trial, the power of the clinical trial for detecting differences between treatments should be given. In this particular trial, having 25 patients in each group would have given a reasonable chance of detecting 40% differences in response rates between treatments (statistical significance level 5%, power of test 80%). Freiman et al. (*32*) did a survey of 71 negative randomized clinical trials. Though the conclusion was negative concerning treatment differences, the authors pointed out that the power of the tests for detecting differences between treatments was very low. In 68% of the 71 trials, the confidence limit for the difference in survival included a 50% reduction in mortality and in 85% of the trials, a 25% reduction in mortality was in the 95% confidence interval. Hence, this suggests that many "not statistically significant" clinical trials are being reported in the medical literature that may be

of medical importance, but the sample size was not sufficient to determine the degree of importance.

There are many statistical techniques for the analysis and interpretation of clinical trials data. A description of many appropriate procedures can be found in the excellent textbook of statstical methods of Snedecor and Cochran (*33*); statistical methods especially useful in clinical trials have been given by Armitage (*34*), Burdette and Gehan (*35*), and a recent joint paper by Peto et al. (*4*).

Summary

This is a review paper which gives a discussion of various aspects of clinical trials in cancer research. Since the conduct of the first randomized controlled clinical trial in cancer patients in the mid-1950's, substantial progress has been made in the utilization of the clinical trial technique for the evaluation of therapeutic efficacy. The important elements of a protocol are given with some discussion of items to be considered in designing a protocol. The types of clinical trial (phase I, II, III) are defined, and the place of each phase of study in the context of the search for new treatments is delineated.

A comprehensive discussion is given of the elements in the comparative clinical trial (phase III) including: objectives, considerations in planning (comparability of treatment groups, stratification of patients, feasibility and size of study, and prospective versus retrospective studies). Brief descriptions are given of designs for comparative clinical trials and a trial in oat cell lung carcinoma is discussed in some detail. Finally, some comments and references are given concerning the analysis of clinical trials.

I would like to acknowledge the cooperation of Ms. Betty Greene for an excellent and timely job in typing the manuscript for the conference and publication.

I would also like to acknowledge financial support from grants CA-12014 and CA-11430 from the National Cancer Institute.

REFERENCES

1. Frei, E., III, Holland, J. F., Schneiderman, M. A., Pinkel, D., Selkirk, G., Freireich, E. J., Silver, R. T., Gold, G. L., and Regelson, W. A comparative study of two regimens of combination chemotherapy in acute leukemia. Blood 13: 1126 (1958).
2. Gehan , E. A., and Schneiderman, M. A. Experimental design of clinical trials. In: Cancer Medicine. J. F. Holland and E. Frei, Eds., Lea and Febiger, Philadelphia, 1973, pp. 499-519.
3. Livingston, R. B., Gehan, E. A., and Freireich, E. J. Design and conduct of clinical trials. In: Cancer Patient Care at M. D. Anderson Hospital and Tumor Institute. R. L. Clark and C. D. Howe, Eds., Yearbook Medical Publishers, Inc., Chicago, 1976.
4. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Brit. J. Cancer 35: 1 (1977).
5. Freireich, E. J., Gehan, E. A., Rall, D. P., Schmidt, L. H., and Skipper, H. E. Quantitative comparison of toxicity of anticancer agents in mouse, rat, hamster, dog, monkey and man. Cancer Chemotherapy Repts. 50: 4 (1966).
6. Carter, S. K. Study design principles for the clinical evaluation of new drugs as developed by the chemotherapy program of the National Cancer Institute. In: The Design of Clinical Trials in Cancer Therapy, M. Staquet, Ed., Editions Scientifique, Brussels, Belgium, 1972.
7. Gehan, E. A. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J. Chronic Dis. 13: 1 (1961).
8. Sylvester, R. and Staquet, M. J. A logical approach to the design of phase II trials using decision theory. In: Proceedings of Symposium on Designs for Clinical Cancer Research, Cancer Treatment Repts., in press.
9. Freireich, E. J., Gehan, E. A., Frei, E. III, Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L., Storrs, R., Haurani, F., Hoogstraten, B., and Lee S. The effect of 6-mercaptopurine on the duration of steroid-induced remission in acute leukemia: a model for the evaluation of other potentially useful therapy. Blood 21: 6, 669 (1963).
10. Hill, A. B. Statistical Methods in Clinical and Preventive Medicine, Oxford University Press, 1962.
11. Armitage, P., and Gehan, E. A. Statistical methods for the identification and use of prognostic factors. Intern. J. Cancer 13: 16 (1974).
12. Gehan, E. A. Adjustment for prognostic factors in the analysis of clinical studies. UICC Technical Report Series 36: 35 (1978).
13. Brown, B. W. Designing for cancer clinical trials: selection of prognostic factors. Cancer Treatment Repts. in press.
14. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. Design and analysis of randomized clinical trials requiring observation of each patient. I. Introduction and design. Brit. J. Cancer 34: 585 (1976).
15. Pocock, S. J., and Simon, R. M. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31: 103 (1975).
16. George, S. L., and Desu, M. M. Planning the size and duration of a clinical trial designed to study the time of some critical event. J. Chronic Dis. 27: 15 (1974).
17. Armitage, P. Sequential Medical Trials, 1975.
18. Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H., and Ware, J. H. Randomized clinical trials. Perspectives on some recent ideas. New Engl. J. Med. 295: 74 (1976).
19. Chalmers, T. C., Block, J. B., and Lee, S. Controlled studies in clinical cancer research. New Engl. J. Med. 287: 75, (1972).
20. Shaw, L. W., and Chalmers, T. C. Ethics in cooperative clinical trials. Ann. N. Y. Acad. Sci. 169: 487 (1970).
21. Gehan, E. A., and Freireich, E. J. Non-randomized controls in cancer clinical trials. New Engl. J. Med. 290: 198 (1974).
22. Gehan, E. A. Comparative clinical trials with historical controls: a statistician's view. Biomedicine, 28: 13 (1978).
23. Freireich, E. J., and Gehan, E. A. The limitations of the randomized clinical trial. In: Methods in Cancer Research, Vol. 17, H. Busch and V. de Vita, Eds., Academic Press, New York, 1979, pp. 277-310.
24. Cox, D. R. Regression models and life tables. J. Roy. Statist. Soc. B 34: 187 (1972).

25. Atkins, H. Conduct of a controlled clinical trial. Brit. Med. Journal, 2:377 (1966).
26. Pocock, S. J. A combination of randomized and historical controls in clinical trials. J. Chronic Dis. 29: 175 (1976).
27. Bull, J. P. The historical development of clinical therapeutic trials. J. Chronic Dis. 10: 218 (1959).
28. Frei, E. III, Freireich, E. J., Gehan, E. A., Pinkel, D., Holland, J. F., Selawry, O., Haurani, F., Spurr, C. L., Hayes, D. M., James, G. W., Rothberg, H., Sodee, D. B., Rundles, R. W., Schroeder, L. R., Hoogstraten, B., Wolman, I. J., Traggis, D. G., Cooper, T., Gendel, B. R., Ebaugh, F., and Taylor, R. Studies of sequential and combination antimetabolite therapy in acute leukemia: 6-mercaptopurine and Methotrexate. Blood 18: 4 (1961).
29. Gehan, E. A. Statistical methods for survival time studies. In: Cancer Therapy: Prognostic Factors and Criteria of Response. M. J. Staquet, Ed., Raven Press, New York, (1975), pp. 17-35.
30. Cutler, S. J., Greenhouse, S. W., Cornfield, J., and Schneiderman, M. A. The role of hypothesis testing in clinical trials. J. Chronic Dis. 19: 857 (1966).
31. Tukey, J. W. Some thoughts on clinical trials, especially problems of multiplicity. Science, 198: 679 (1977).
32. Freiman, J. A., Chalmers, T. C., and Smith, H. The role of β and the type II error in the design and interpretation of randomized clinical trials: Survey of 71 negative trials. New Engl. J. Med. in press.
33. Snedecor, G., and Cochran, W. G. Statistical Methods, 6th ed., (1967).
34. Armitage, P. Statistical Methods in Medical Research, 1971.
35. Burdette, W. J., and Gehan, E. A. Planning and Analysis of Clinical Studies, Charles C. Thomas, Springfield, Ill., 1970.
36. Cochran, W. G., and Cox, G. M. Experimental Design, 2nd ed., 1957.