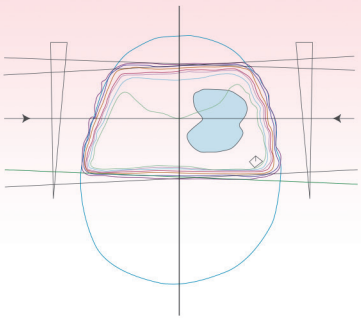
CHAPTER

13 Statistics and Clinical Trials



Qian Shi, Wenting Wu, and Daniel J. Sargent

In this chapter we address some of the statistical issues associated with clinical trials. We begin with a short description of the basic role of statistics and statisticians and illustrate some sample communications between investigators and statisticians. We provide an overview of the statistical issues relevant to each phase of clinical trials (phases I, II, and III) and discuss some of the unique aspects of phase II and III trials. We introduce standard and newly developed phase II experimental designs. We are particularly interested in the essential roles of randomization and stratification in comparing new and standard therapies. We discuss the important *intent-to-treat principle*, which is fundamental to analyzing phase III trials. We emphasize the importance of appropriate monitoring in ongoing trials. In addition, we consider some of the special problems that arise in the analysis of survival data that are caused by the phenomenon of *censoring*, which occurs when a patient's time to death cannot be completely determined either because he or she becomes lost to follow-up before death or is still alive when the data are to be analyzed. Finally, we briefly discuss several relevant topics in modern clinical trials: surrogate endpoints, biomarkers, and adaptive design.

Our intent in this chapter is to provide the reader with insight regarding the use of statistics (and statisticians) in the design and analysis of clinical trials rather than to provide all of the details required for an investigator to perform his or her own analyses. Several excellent texts provide the details required for performing analyses, and we cite several such references.

We define a clinical trial as a designed study involving the treatment of prospectively accrued humans that is specified by a document (protocol) with specific goals and analysis plans. Meinert[1] describes some of the unique aspects of clinical trials that distinguish them from other medical research studies, including, among others, observational studies and case-control studies, and enumerates the requirements of a good protocol document.

WHY ARE STATISTICS AND STATISTICIANS USEFUL?

It has been a long-standing paradigm that clinical research is best conducted through collaborations requiring multidisciplinary expertise. Among these scientists, statisticians are often viewed as technicians who provide statistical services, rather than collaborators who make critical scientific contributions to the research. For example, an improper opinion that occasionally exists is that the involvement of statisticians only starts with analyzing data after the completion of the study. Actually, optional collaboration between statisticians and clinical investigators is involved in every step of a clinical trial and can be extremely fruitful. These collaborations cover the whole cycle of a study—from design, conduct, continuous monitoring, data analysis, and interpretation to decision making and publishing results, as well as generating new hypotheses for subsequent studies. Furthermore, the methodology research centered on clinical trials is a driving force of innovative designs in clinical research.

The ultimate goal of conducting clinical trials is to make beneficial influences in future clinical practice, which requires that the conclusions made at the end of the trial have a high degree of credibility, reproducibility, and external validity.[2] Just like all experiments, realities of environment, conditions, and resources introduce both controllable and uncontrollable errors. This is especially true in oncology studies. Genetic, behavioral, environmental, and sociological heterogeneities introduce great complexities in disease progresses and the impact of the studied treatment. Generally speaking, there are two types of errors. One is random, which is purely the result of chance. Another is systematic, known as bias, which describes errors that are not a consequence of chance alone. A fundamental difference between random errors and bias is that pure random errors have no preferred direction, whereas biases present effects of distortion.

Random errors are commonly caused by sampling variability, subject heterogeneity, measurement error, and other source of noises. In reality, the random errors in any experiment can never be completed eliminated. For instance, two patients who have cancer may have nearly identical clinical (and pathologic) characteristics and receive the same treatment. But one may fail (experience relapse or die) within months whereas the other may be cured. A practical definition of *variance* could be the unexplained differences in outcomes that exist among seemingly similar patients. If one looks at two sets of 15 patients who differ in one characteristic (possibly treatment) and notices that one group has a median survival time that is 2 months longer than the other group, one cannot immediately conclude that the characteristic is truly associated with survival difference or just by chance. Statistical techniques will be required to evaluate the likelihood that a median survival difference of 2 months was quite likely to occur even if these two sets of patients who actually did not differ in the characteristic had been compared (a so-called type I error). Quantifying the magnitude and likelihood of the errors resulting from chance requires knowledge of probability and statistical theory. Reducing the relative impact of random errors can be achieved by averaging over a large number of observations. This is commonly known as the *sample size calculation* in clinical trial design.

Bias, on the other hand, cannot be reduced by averaging after repletion or taking additional observations. In clinical settings, although biases can arise in numerous ways, there are often a set of factors that have been studied and well known to contribute to the possibility of bias. In many cases, these sources of bias can be understood well enough to be controlled. For example, bias can occur in medical research studies if treatments are compared between groups of patients who are not equivalent in terms of characteristics that are associated with prognosis. The fundamental strategy used in clinical trials to avoid biased treatment comparisons is randomization of treatment assignments (with stratification), which guarantees the comparability of patients assigned to different treatments under the comparison(s) in terms of his or her prognosis under the condition of no treatments. In observational studies, such an assurance is simply not possible. Although it may be feasible to adjust for potential sources

of bias at the data analysis stage, based on our present knowledge of factors influencing disease outcomes within individual patients, there is no way to ensure that patients receiving two different treatments are comparable in the absence of randomization. There are other critical considerations to reduce bias and include concurrent controls, objective assessments, consistent active follow-up and endpoint ascertainment, no post-hoc exclusions, and so on.

Reducing and controlling random errors and systematic biases are essential elements of clinical trials to provide valid and generalizable trial inferences. These cannot be achieved without sound statistical design. Sophisticated analyses can almost never salvage design faults and poor implementation of the design.[2,3] Since the recognitions of statisticians' involvements in clinical trials in 1970s, a subgroup of statisticians have been actively contributing their knowledge and creativity in cancer therapeutic research by providing core techniques to transfer the conceptional clinical ideas into sound, efficient, and practical clinical trials throughout the entire lifetime of the study.

INTERACTION BETWEEN AN INVESTIGATOR AND A STATISTICIAN

Inadequate communication between an investigator and a statistician might be the most frequent cause of inappropriate designs or analyses in clinical research. This may occur because the investigator does not explicitly state what he or she is trying to learn from a particular study or because the statistician lacks the insight of what methods that satisfactorily address the study question. Planning and conducting a clinical trial is an interactive and iterative collaboration process between investigators and statisticians. For example, in the designing phase, many critical communications are centered on the following components:

1. What is the targeted disease population? This provides the set of critical disease characteristics and known (potential) confounders that the statistician needs to consider in the design or analyses for controlling biases or identifying effect modifier (e.g., predictive markers).
2. How is study "success" defined? This indicates the choice of study type including decision-making studies with superiority, noninferiority, or equivalence designs, exploratory studies aimed to quickly screen promising agents, or trials conducted to provide estimation of the key parameters of the new agents.
3. What is the existing knowledge of the tested agent? This describes the stage of the study design, for example, phase I trials for dose finding, phase II trials for preliminary evidence of efficacy, or phase III trials for confirming efficacy and regulatory approval for marketing. Commonly, the level of uncertainty tolerance (i.e., significance level and power) varies according to the phase of the design.
4. What is the clinical meaningful treatment effect? This requires an appropriate choice of endpoint, which is a measurement that reflects how patients feel and function and can be reliably and precisely measured. Historical data of this endpoint in the targeted population serves the benchmark of the design. Discussions of the expected clinical meaningful effect size give reasonable derivation of sample size and study operating characteristics.

Other details include whether a control group should be included; should a biomarker be used in the design, and if so how should it be incorporated; and what are the logistical and practical challenges in trial conduct? These components of communications between investigators and statisticians are never isolated from each other. Transparent, informative, comprehensive, and persistent communications are essential for a good trial design.

CLINICAL TRIALS

Clinical trials performed to develop and test new agents or modalities in oncology are often categorized as phase I, phase II, or phase III trials according to their primary aims.

Phase I Trials

Phase I trials are generally the first trials involving human subjects in which a new treatment is tested. The goal of a phase I trial is to test a new regimen's tolerability and toxicity. These trials usually enroll a limited number of patients who have exhausted other treatment options. Phase I trials may enroll patients with a specific tumor type only or be open to patients with different tumors. Different primary endpoints and study designs should be considered for regimens with different cancer therapeutic mechanisms.

Traditional Cohorts-of-3 Phase I Design

Historically, cancer therapies have been designed to act as cytotoxic, or cell-killing, agents. The fundamental assumption regarding the dose-related activity of such agents is that there exists a monotone nondecreasing dose-response curve, meaning that as the level of the dose increases, tumor shrinkage will also increase, and this phenomenon should translate into increasing clinical benefit. Under this assumption, both the toxicity and the clinical benefit of the agent under study will increase as the dose increases; therefore, an appropriate goal of a phase I trial is to find the highest dose with acceptable toxicity. Because the monotone nondecreasing dose-response curve has been observed for most cytotoxic therapies, toxicity has historically been used as the primary endpoint for identifying the dose that has the greatest likelihood of being effective, yet tolerable, in subsequent testing.

In this context, the typical goal for phase I clinical trials has been to determine the maximum tolerated dose (MTD), which is the highest dose level at which the rate of dose-limiting, toxicity (DLT) does not exceed the targeted toxicity level (TTL). Which adverse events will be considered as DLT vary depending on the type of cancer and the specific agent under study. The traditional design is the cohort-of-3 design, which is easy to use and requires no complicated statistical computations. Specifically, three patients are enrolled at the starting dose level. If no DLT is observed, three patients will be enrolled at the next higher dose level. If one DLT is observed, another three patients will be enrolled at the same dose level. Out of the six patients at a given dose level, if one DLT is observed, three patients will be enrolled at the next higher dose level. If two or more DLTs are observed (out of three or six patients at a given dose level), then the MTD will be considered to have been exceeded, and the next lower dose level will be defined as the MTD as long as six patients have been evaluated at that level.

Although the cohort-of-3 phase I design has considerable appeal (it is straightforward to conduct, easy to explain, and has considerable historical precedent), there are several limitations with this design. Particularly with newer agents that will be discussed, use of the cohort-of-3 design poses increasing challenges in modern clinical trials. For instance, the TTL is arbitrarily set to 33%, which may not suitable for every disease-agent setting. High uncertainty is unavoidable because of the inherited small sample size. For example, the exact 95% confidence interval for the observed rate of DLT at MTD is 0.00 to 0.64 when 1 of 6 patients experienced DLT. Also, it is a common misunderstanding that the true DLT rate at MTD

identified by this design is 0.33. It is actually closer to 0.22.[4] For the assumption of monotone nondecreasing dose-response curve for cytotoxic agents, this fact increases the risk of failure in phase II studies because the lower toxicity also indicates lower efficacy. When the true DLT rate presents small variations across dose levels, it may require a long time and many patients treated at suboptimal dose levels, such as in a cohort-of-3 phase I study[5] in which 56 patients were accrued during a period of 38 months to determining the MTD of CPT-11 in the regimen of CPT-11/5-FU/LV for patients with metastatic or locally advanced cancer.

Newer Phase I Designs

Over the last several decades, statistically based improvements have been proposed for phase I study design. One of the statistical theories (i.e., Markov Chain) demonstrates that when assigning patients to the next dose level, up or down, based on observed number of DLTs (0 or ≥1) and prespecified TTL, the dose assignment will converge to MTD when the number of assignments is sufficiently large.[6] A class of up-and-down designs was developed based on this theory. One example is biased-coin design (BCD) with dose assignment based on a single patient's DLT occurrence.[7] Later this method was extended to enroll patients in cohorts to each dose level.[8]

In parallel to up-and-down design, another class of nonfixed dose assessment design, using a model-based approach, was started with the first proposal of continual reassessment method (CRM),[9] with further improvements for practical use.[10,11] One of the key features of the CRM is that the DLT rates at different dose levels are reestimated based on data from treated patients across all dose levels examined to date, whenever there are new data observed. Each newly enrolled patient is assigned to the dose level whose current DLT rate estimate is closest to TTL. At the end of study, the MTD is determined by final DLT rate estimates using all treated patients' data. Garrett-Mayer gave an excellent tutorial on the CRM design.[12]

As the therapeutic development in oncology moves from cytotoxic to cytostatic or targeted agent, additional challenges are present in dose-finding studies. For example, the dose-response relationship for novel agents may take forms other than monotone nondecreasing curve, such as quadratic, or increasing with a plateau. In this case, evaluating toxicity and efficacy simultaneously is necessary to find the biologically optimal dose (BOD), which balances the needs of minimizing toxicity (to an acceptable rate) and maximizing the efficacy. Weighted-outcome[13] or utility-based[14] approaches were proposed to handle the tradeoffs between two endpoints. Another example is dose finding in combination treatment; CRM methods have been extended to generalized CRM[15] or partial order CRM[16] methods aimed to address this situation.

Despite the development of more effective designs, the standard chort-of-3 still remains the most popular phase I study design in practice. Rogatko et al examined more than 1,200 published phase I studies and found only 1.6% adapted more sophisticated design than chort-of-3 design.[17] The major reasons may be statistical complexity and lack of user-friendly software.[18] However, increased awareness and greater communication between investigators and statisticians increases the possibility of applying these newer methods.

Phase II Trials

In phase II trials, the goal is to establish clinical activity and further evaluate the treatment's toxicity. Unlike phase I trials, which commonly accrue patients with a variety of cancers, a phase II trial should restrict accrual to a reasonably well-defined patient population. Historically, the response rate has been the most common endpoint for phase II trials. However, in the last decade, progression-free survival/disease-free survival (PFS/DFS) and overall survival (OS) rates have been used increasingly as the primary endpoint in the phase II setting.

Traditional Single-Arm Phase II Designs: One-Stage Design

The starting point for a discussion of phase II trials is the single-arm, one-stage design. In these trials, patients are accrued to a single arm and are treated at a single dose level that has been suggested by previous phase I trials. If the treatment appears sufficiently active in comparison with historical success rates (typically, tumor response) in patients with disease and prognosis similar to those accrued to the phase II trial, the treatment may be considered a viable candidate for the phase III setting. Toxicities, expense, and other considerations will often influence the decision as to whether the drug or radiation versus chemoradiation regimen has enough promise compared with standard regimens to warrant further interest.

We explain this type of design by means of an example. Suppose that in recent treatment trials for a given disease, the response rate ranged between 15% and 20%. A new trial might then be designed to show that the new treatment will not be worth continuing if the true response rate is 10% or less and that the new treatment should be continued if the true response rate is 25% or greater. Such a design implies, by default, that if the response rate is in the range of 15% to 20%, the probability of either decision being made (i.e., to abandon the new therapy or to bring it to phase III testing) may be nearly equal. One critical consideration is to make sure that patients in the new study are comparable to those enrolled in the previous studies, a factor that in many cases is impossible to verify.

As described, the typical phase II trial is structured to provide the basis for a recommendation either to abandon the therapy or to consider it for further testing. Working together, the investigator and the statistician choose an "unacceptable" response rate, p_0 (10% in our example) and a "promising" response rate, p_1 (25% in our example). The study is designed so that if the response rate is as low as p_0 there will be little chance that the treatment will be recommended, whereas if the response rate is as high as p_1 there is a high probability that the treatment will be recommended. These benchmark response rates should in no case be chosen arbitrarily; considerable thought must be given to their selection, and they should be fully justified by comparison with historical data in patients with comparable disease and prognosis.

The considerations just given might lead to the following study design criteria:

1. If the true response rate is $p_0 = 10\%$, there should be no more than a probability of 0.10 (significance level) of erroneously concluding that the treatment should be carried forward.
2. If the true response rate is 15% to 20%, it will not be harmful to have a fairly high probability of reaching either conclusion.
3. If the true response rate is $p_1 = 25\%$, there should be a probability of at least 0.90 (power = 0.90) of concluding the treatment is effective enough to be carried forward.

Using well-known methods, the statistician could determine that each of these criteria would be satisfied if 40 patients were accrued, and the treatment would be recommended for further consideration if, and only if, at least 7 of the 40 patients responded. Table 13-1 gives the probabilities that the treatment will be carried forward for various hypothetical response rates.

It is worth noting that the formal decision rule is to conclude the treatment is worth carrying forward if the estimated

TABLE 13-1 Operating Characteristics of a One-Stage Phase II Study Design

	True Response Rate					
	5%	10%	15%	20%	25%	30%
Probability of recommending treatment for further consideration	0.003	0.10	0.39	0.71	0.90	0.98

TABLE 13-2 Operating Characteristics of a Two-Stage Phase II Study Design Allowing Early Stopping for Zero Responses in the First 15 Patients

	True Response Rate					
	5%	10%	15%	20%	25%	30%
Probability of recommending treatment for further consideration	0.003	0.10	0.39	0.71	0.90	0.97
Probability of stopping after first stage of accrual	0.46	0.21	0.09	0.04	0.01	0.005

TABLE 13-3 Operating Characteristics of a Two-Stage Phase II Study Design Allowing Early Stopping for Zero Responses or One Response in the First 15 Patients

	True Response Rate					
	5%	10%	15%	20%	25%	30%
Probability of recommending treatment for further consideration	0.003	0.09	0.35	0.66	0.86	0.95
Probability of stopping after first stage of accrual	0.83	0.55	0.32	0.17	0.08	0.04

response rate is at least $7/40 = 17.5\%$ but not if the response rate is less than or equal to $6/40 = 15\%$. As Table 13-1 indicates, the probability of observing a response rate of at least 17.5% in our trial if the true response rate is 10% is only 0.10. In contrast, the probability of observing a response rate of at least 17.5% is 0.90 if the true response rate is 25%, and a positive recommendation is extremely likely (probability = 0.98) if the true response rate is as high as 30%. Other factors may enter into the final decision, particularly if the observed response rate is close to 15% or 20%.

Traditional Single-Arm Phase II Designs: Two-Stage Design

Because many experimental treatments ultimately do not demonstrate efficacy, ethical considerations lead to the usual recommendation of an early stopping rule for lack of efficacy in phase II trials. For example, suppose that the first 15 patients who are accrued do not respond to the therapy. At that point, an investigator might begin to believe that the treatment is no better than standard therapy and might even be worse. This raises the question of whether the trial should be terminated. On the one hand, there is some probability that 7 of the next 25 patients will respond, resulting in a recommendation for continued consideration of the treatment. Statisticians have developed clinical trial designs that allow a prospectively specified examination of the data from the ongoing trial to address this issue.

Generally, the issue of *early stopping rules* is as follows: When designing a trial, it may be appropriate to introduce the possibility of stopping early if there is strong evidence that the treatment is no more and possibly less effective than the standard treatment, as long as doing so would not substantially reduce the probability of detecting a true beneficial effect (power). Of course, one might also want to stop early if the treatment appears to be extremely effective, although there is no ethical constraint against assigning patients to an apparently effective treatment, unless doing so would unduly prolong the further development of a promising therapy.

Consider the following modification of the phase II study design just described. Accrual will proceed in two stages. In the first stage, 15 patients will be accrued, treated, and observed for clinical response. If no patients respond, the trial will be terminated, and the candidate treatment will not be recommended for further consideration; otherwise, an additional 25 patients will be accrued. If in total at least eight responses are observed, the treatment will be recommended for further consideration; otherwise, it will not be so recommended.

This new design has some nice properties. The maximum number of patients required is the same as before, and, as shown in Table 13-2, the new design still satisfies each of the desired criteria 1 to 3 listed previously. However, if the new treatment has a true response rate of 5% or less, the study has a 0.46 chance of stopping after 15 patients have been accrued, therefore sparing 25 patients from treatment with an inactive regimen.

Now consider a design that requires early stopping if only 0 or 1 of the first 15 patients responds; otherwise, 25 more patients are accrued, and the treatment is recommended for further consideration if at least 7 patients respond. This design provides much better protection against the possibility of accruing an excessive number of patients to an ineffective regimen. However, if this design is used, it would be slightly less likely that a truly effective regimen would be recommended for further consideration. Table 13-3 shows the properties of this design.

Notice that with this design, there is a 0.83 probability that the study will terminate after 15 patients are entered if the regimen has only a 5% response rate. However, the probability of recommending the treatment is now only 0.86 if there is a 25% response rate. Numerous authors have discussed optimal strategies for choosing phase II designs, including Gehan,[19] Herson,[20] Lee, Staquet, and Simon,[21] Fleming,[22] Chang et al,[23] Simon,[24] Therneau, Wieand, and Chang,[25] Bryant and Day,[26] and Thall, Simon, and Estey.[27]

Newer Phase II Designs: Randomized Phase II Design

Single-arm phase II designs can demonstrate biologic activity of an agent with relatively small sample size and short study duration. However, the potential for patient selection bias, the evolving methodologies for the assessment of "success" (i.e., changes in computed tomography [CT] scanners, changes in

response criteria), and a general lack of robustness of historical rates of this type of design result in a high possibility that a positive single-arm phase II trial will be followed by a negative phase III study.[28] A potential solution to this fundamental limitation of single-arm phase II designs is the use of a randomized design. Two types of randomized phase II designs, the randomized selection designs and the randomized screening designs, have been proposed and used widely.

In a selection design, all arms are considered experimental arms. These could be different new regimens or different doses or schedules of the same regimen. The sample size is calculated to guarantee that with a small probability, perhaps 10%, an inferior arm (for instance, one in which the response rate is 15% lower) will be selected for the future phase III study. In a standard selection design, the arm with the highest estimated success rate for the primary endpoint will be recommended for a follow-up phase III study, no matter how small the difference of estimates might be among arms.[29] On the other hand, a flexible selection design, or "pick-the-winner" design, will recommend the arm with the best estimated primary endpoint if the difference is larger than a prespecified criterion.[30] Otherwise, toxicities, expense, and other factors will be taken into consideration when making the decision.

If a standard of care is included in the comparison, a screening design should be applied.[31] A screening phase II design is similar to a randomized phase III trial but has larger type I and type II errors. For instance, a range of 10% to 20% for both error rates is acceptable. A screening phase II design provides a head-to-head comparison between the experimental regimen and the standard of care using a relatively smaller sample size and clear guidance as to the likelihood of success if the agent is moved forward into phase III testing.

Overall, randomized phase II designs balance prognostic factors among all arms, avoid patient selection bias, and provide robust arm comparison results. The major concerns for randomized phase II designs are (1) the larger required sample size, (2) the relatively high risk of false-negative and false-positive rates resulting from the small sample size, and (3) the possible desire (or ethical mandate) to skip a confirmatory phase III study given a positive phase II result.

Another type of newly developed randomized phase II design is the phase II/III design. A phase II/III design starts with a phase II component. Patients are randomized among several experimental arms or a standard-of-care arm versus one or more experimental arms. Experimental arms could be compared with the historical control or the standard-of-care control to select arms for entering the phase III component. The main advantage of this type of design is (1) the speed of transition from phase II to phase III and (2) the fact that data obtained from the phase II component could be applied toward the phase III comparison as long as the protocol treatment is not altered between phases.

Phase III Trials

It is generally recognized that judging the value of a new therapy by comparing it with historical data may give an erroneous impression of the therapy's efficacy. Pocock[32] related a number of illuminating examples of this phenomenon. Therefore the phase III trial, in which a new agent, modality, or combined treatment modality is tested against an accepted standard treatment in a randomized comparison, is considered the most satisfactory method of establishing the value of the proposed treatment.

Under most circumstances, the goal of a phase III trial is to determine whether the proposed treatment is superior to the standard treatment (superiority trial). Alternatively, if the alternative treatment is easier to administrate, costs less, or is less toxic than standard of care, showing superiority is not necessary; a definitive demonstration of noninferiority (noninferiority trial[33]) is all that would be required. Because the goal of a phase III trial is to make a definitive conclusion regarding a new treatment's efficacy, enough patients need to be accrued to guarantee a small probability of false-positive results (i.e., declaring that a regimen is effective when in fact it is not) and a large probability of true-positive results (i.e., declaring that a regimen is effective when in fact it is). The probability of false-positive results is also called the size of the study, which usually is limited to 5% or less. The probability of true-positive results is also called the power of the study, which usually is set at 80% to 95%. The primary endpoints for comparison are usually OS and PFS/DFS, and a common secondary endpoint is quality of life (QOL). In this section, we discuss the rationale for requiring randomized treatment assignment and stratification in a phase III trial. We introduce the intent-to-treat principle, and we emphasize the importance of monitoring in ongoing trials.

Rationale for Randomization

A random assignment is defined as any assignment resulting from a mechanism ensuring an equal probability to all possible assignments. The most obvious reason for randomized assignment is to eliminate the possibility of bias, that is, the intentional or inadvertent assignment of treatments by means of procedures that allow patient characteristics to influence the assignments in a way that may promote an imbalance in patient characteristics between the two treatment arms. The basic idea is simple. If chance alone determines treatment assignments and enough patients are accrued, the law of averages will ensure that the characteristics of patients in different treatment arms will be nearly identical.

Treatments must be assigned in such a way that the accruing physician or coordinator will have no reasonable basis for predicting the arm to which a patient will be assigned. For example, if an accruing physician suspects that at some point their next patient may be assigned to a new treatment study, he or she may recruit patients who might have more favorable outcomes, thereby allowing subtle differences in patient characteristics to influence treatment comparisons. For this reason, it is common for the randomization of patients to be permitted only through contact with a centralized randomization coordinator.

Other potential sources of bias exist that cannot be eliminated by randomized treatment assignment. In some diseases, application of a treatment may convey psychological effects or promote expectations that affect some endpoints of interest. In cancer research, the most obvious example is that treatment with chemotherapy or other agents gives rise to expectations of toxicity. Therefore, whenever it is feasible without causing undue hardship to patients, treatment assignments should be masked by use of placebos. Perhaps even more important is the need to keep medical personnel unaware of actual treatment assignments whenever possible to eliminate bias in assessing response, relapse, or attribution of causality to adverse events.

Stratification of Treatment Assignments

For trials with large sample sizes, randomization of treatment assignments—together with the law of averages—will result in nearly balanced patient characteristics across treatment arms. However, if certain patient characteristics are known to be strongly prognostic of outcome, a better balance can be achieved by allowing separate randomization of treatments to patients within subpopulations defined by the various levels of prognostic variables. Such a restricted randomization scheme is referred to as *stratified random assignment*, and the

prognostic variables are referred to as *stratification variables*. For example, in clinical trials of adjuvant therapy for breast cancer, it is known that rates of relapse are strongly related to the number of axillary nodes found to be positive for tumor on histologic examination and are less strongly related to the age of the patient. Therefore, the assignment of treatments to patients might be stratified within eight subpopulations, defined by combinations of nodal status (node negative, 1 to 3 positive nodes, 4 to 9 positive nodes, 10+ positive nodes) and age at randomization (≤49 years, ≥50 years). In this way, stratification (to balance the effects of strongly prognostic characteristics) and randomization (to prohibit bias resulting from less obvious or even unforeseen factors) may be used together to promote an unbiased comparison of treatments.

In some cases, it is difficult or even impossible to achieve completely balanced treatment assignments within every combination of the levels of all potential stratification variables. For example, in multicenter cancer clinical trials, it is generally thought to be good practice to balance treatment assignments within each accruing institution because patient selection may differ from institution to institution. However, it is not unusual for cooperative groups in oncology to conduct phase III clinical trials using several hundred accrual sites. Clearly, it would be difficult to stratify completely in the accruing institutions together with other potentially prognostic variables, because the number of resulting combinations of levels of all the stratification variables would be excessive. To circumvent this problem, cooperative groups commonly use algorithms that dynamically maintain nearly balanced assignment of treatments across the levels of all stratification variables marginally but that do not guarantee balance within all combinations of the levels of all of the stratification variables. This approach also results in nearly balanced treatment assignments at all times, which will be beneficial if, as often happens, the characteristics of patients for whom the trial is considered most appropriate in the judgment of accruing physicians tend to change as the trial matures. Details of several such dynamic balancing algorithms may be found in articles by Taves,[34] Pocock and Simon,[35] Freedman and White,[36] and Begg and Iglewicz.[37]

Intent to Treat

If treatment assignments are randomized as described previously, one can be reasonably confident that important prognostic factors will be balanced across treatment arms. However, if any patients are excluded at the time the data are analyzed, the possibility of bias arises once again. An analysis that includes all randomized patients, regarded as though each had received precisely the treatment to which he or she had been randomized, irrespective of compliance or administrative errors (e.g., errors in eligibility or other protocol deviations), is an intent-to-treat (ITT) analysis. Another type of analysis is called per-protocol (PP) analysis, in which only eligible patients who adhered to the treatment according to the protocol are included. In this section, we show how treatment comparisons may be seriously biased by excluding patients from an analysis, even for apparently valid reasons. We then show how an intent-to-treat analysis may effectively reduce biases that can creep into analyses when not all patients actually receive the treatments to which they have been randomized. We discuss these issues in the context of the following example:

Example 1. A clinical trial was designed to compare the effect of radiation therapy relative to observation in patients with rectal cancer. Three hundred patients were randomized between radiation therapy (arm A) and observation (arm B), 150 to each arm. Suppose that all patients actually received all of their intended therapy and that after 7 years of follow-up,

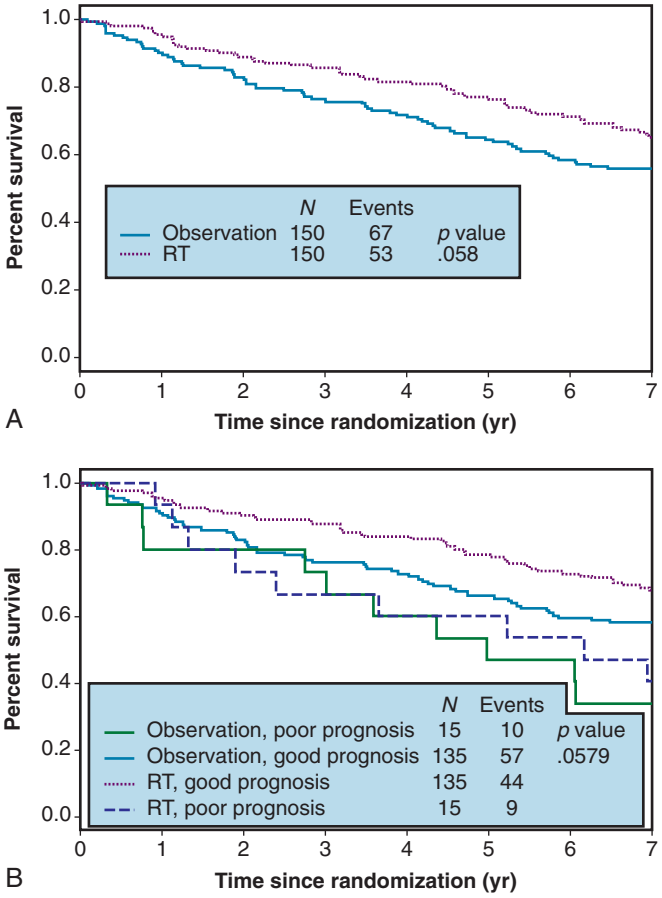


Figure 13-1 **A,** Survival curves that would be observed if all patients receive the prescribed therapy (percent survival versus years from randomization). **B,** Survival curves that would be observed if all patients receive the prescribed therapy, stratified by risk group (percent survival versus years from randomization). *RT*, Radiation therapy.

survival curves were as shown in Figure 13-1, *A*. Although the survival curve for patients receiving radiation therapy lies above that for patients in the observation-only arm, suggestive of a benefit from radiation therapy, the difference is not quite statistically significant ($p = 0.058$).

In any clinical trial, the patients who are accrued exhibit considerable variation in their prognoses. Suppose that we could make use of certain prognostic patient characteristics that were unavailable in the analysis summarized previously. Suppose further that by using this new information, we could identify the 10% of patients with the worst prognoses. Then we might redraw the survival curves after stratifying according to risk, resulting in the plots shown in Figure 13-1, *B*. An analysis of these data that takes this stratification into account again yields a p value of 0.058, which agrees with the p value from the unstratified analysis to three decimal places. This insensitivity of analysis to stratification is a direct consequence of the randomization used in the design of the trial, which tends to balance the poor-prognosis patients equally across the two arms of the trial.

Returning to the original analysis of these data, suppose that not all patients in arm A had actually received the radiation therapy to which they had been randomized. The statistician must decide what to do with these noncompliant cases. There are at least three ways to proceed: (1) it may seem logical to discard data corresponding to the noncompliant patients because they did not receive the radiation therapy that was assigned to them; (2) it might even seem reasonable to analyze

the noncompliant patients as though they had been randomized to arm B rather than arm A because, in fact, they received no adjuvant therapy, similar to those patients who were randomized to observation; (3) the ITT principle states that the noncompliers should be analyzed just as though they had received radiation therapy, even though in actuality they did not.

We now consider the possible consequences associated with these three possibilities. Under the assumption that patients' reasons for refusal of treatment are completely unassociated with prognosis, methods 1 and 2 just presented should lead to estimated treatment differences that are essentially identical to the analysis of Figure 13-1, *A*. If there is a treatment difference, the ITT analysis will tend to underestimate this difference because some of the patients who were not irradiated are counted as having received that therapy. The differences in the three analyses should be rather small under this assumption of no association between refusal of treatment and prognosis.

Alternatively, it is rather likely that reasons for refusal of treatment will be associated with patient prognosis. For example, patients with poorer performance status, or their physicians, may be less liable to accept assignment to a treatment that is known to be associated with significant toxicity. Deteriorating health may lead to difficulties in traveling to receive treatment; advancing disease may correlate with a patient's level of depression or anxiety, which, in turn, may correlate with compliance. Suppose that it is the 10% of patients with the worst prognosis who refuse to accept their radiotherapy. Under these conditions, analysis of the data after deleting the noncompliant patients overstates the effect of radiotherapy, as shown in Figure 13-2, *A*. This is because the patients with the worst prognosis are deleted from consideration in arm A, but no similar deletion of patients with poor prognosis is made in arm B. The result is a spuriously significant p value of 0.021. Treating the noncompliers as though they had been randomized to observation (see Figure 13-2, *B*) results in an even more serious overstatement of the radiotherapy effect, associated with a p value of less than 0.009. In contrast, the ITT analysis yields a slightly attenuated treatment effect ($p = 0.065$) (Figure 13-2, *C*) but no serious misrepresentation of the true effect.

Why did the ITT analysis correspond to the "truth" more closely than the analysis that omitted the subset of patients who did not receive their assigned radiation? This is because the effect of the prognostic factor (risk group) was larger than the treatment effect and the likelihood of patient refusal was associated with both the treatment received and the prognostic group of the patient. Although patient prognoses were balanced as randomized, they are not balanced as treated. In this case, the baseline prognosis of patients who actually received radiation was better than the baseline prognosis of those who were under observation. It is precisely this type of imbalance that the ITT analysis is designed to prevent.

There is no one correct answer as to the best approach to analysis when not all patients receive their assigned therapy. However, there is general agreement among statisticians that it is always appropriate to perform an ITT analysis in which each patient is analyzed according to the treatment assigned at randomization, regardless of what treatment was actually received. Results of any other analysis should be compared with the results of the ITT analysis, and if the results differ substantively, the interpretability of the data must be questioned.

The prior examples illustrated only one type of problem that may be addressed by an ITT analysis. Biases similar to those just described can occur if patients elect to cross over study arms, accept a nonstudy therapy, or receive only a portion of the assigned therapy or if they are determined to be ineligible. Therefore, we advocate attempting to obtain complete follow-up for all patients registered to every phase III trial and performing an ITT analysis. Sometimes this may not be possible because patients who refuse protocol therapy may also refuse to be followed. If only 1% or 2% of patients

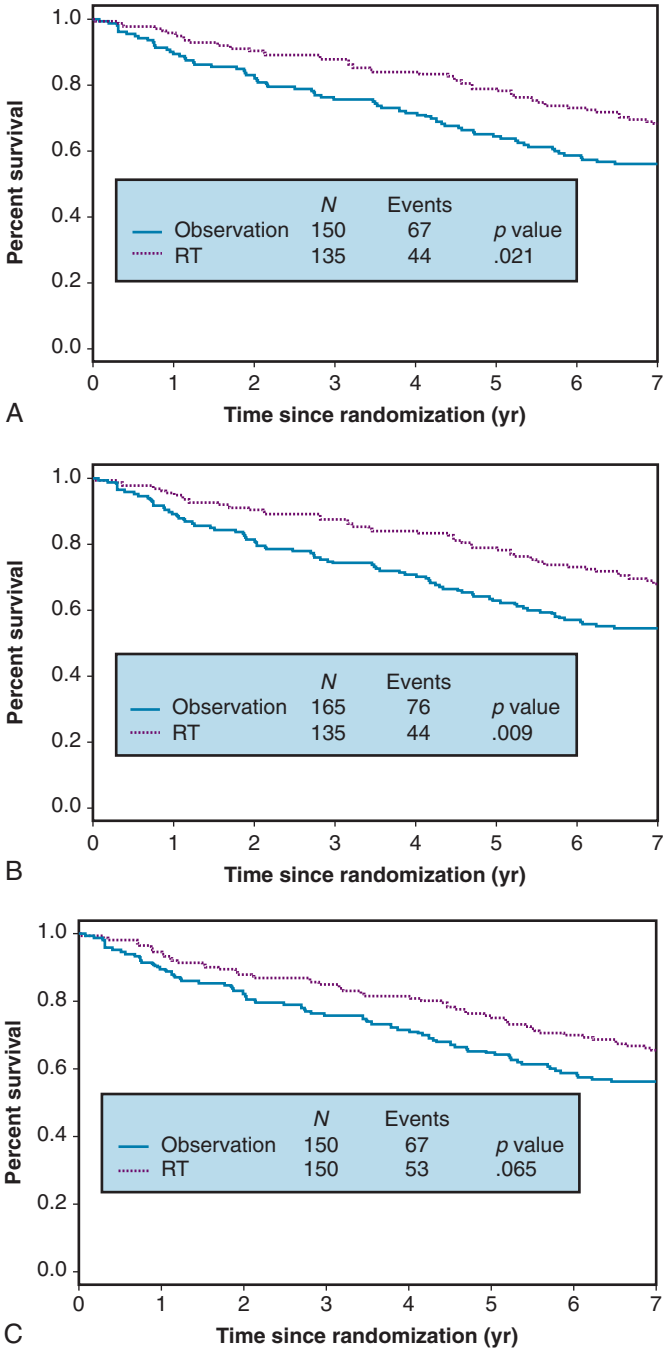


Figure 13-2 **A,** Survival curves that would be observed if the patients of poor risk who refuse radiation therapy *(RT)* are excluded from the analyses (percent survival versus years from randomization). **B,** Survival curves that would be observed if the patients of poor risk who refuse radiation therapy are included as untreated patients (percent survival versus years from randomization). **C,** Survival curves that would be observed if the patients of poor risk who refuse radiation therapy are included as radiation therapy patients (i.e., the results that would be obtained using an intent-to-treat analysis [percent survival versus years from randomization]).

fall into this category, there is little danger of important bias, but if this number is larger, say 5% to 10%, there is a real danger that the study results may be biased. One might then want to perform what are called sensitivity analyses to examine the possible effect of having so many patients without follow-up. In its simplest form, one might perform one analysis assuming that all of the patients without outcome data on arm A failed early on, whereas those on arm B survived; and then perform an analysis assuming that all of the patients without outcome data on arm A survived, whereas those on arm B failed. If one reaches the same conclusion about treatment effect in either analysis, then the exclusions do not represent a serious problem.

There is often disagreement concerning the inclusion of ineligible patients in analyses, particularly in analyses dependent on patient covariates because ineligible patients may not fit any reasonable classification. The decision as to whether it is appropriate to exclude ineligible patients depends on the methods used for determining ineligibility. For example, in some studies, each patient is reviewed for eligibility in a uniform way by a reviewer (or review committee) who is unaware of the assigned treatment and uses only information that was obtained before randomization. If that is the only mechanism for classifying patients as ineligible, one could reasonably exclude such patients from analyses. However, any less stringent approach that allows patients to be classified as ineligible after randomization could introduce subtle biases. For example, in a trial of radiation versus observation, one might routinely perform a pretreatment examination immediately after a patient is assigned to radiation therapy. If, during that examination, it is determined that the patient had not met an eligibility requirement, exclusion would cause a bias because a similar patient not assigned to radiation would not have had the pretreatment examination and would, therefore, not have been determined to be ineligible.

To summarize, we recommend always performing an ITT analysis using all patients as randomized and reporting the results of that analysis in any publication. Further analyses may be performed, especially for noninferiority trials, but one should be aware of potential biases similar to those described previously. Further discussion of these issues can be found in articles by Pocock[32] and Gail.[38]

The data used in Example 1 were generated using the following assumptions. The expected 5-year survival rate for an untreated good-risk patient was 0.73, and for an untreated bad-risk patient, 0.54. The reduction in the death rate associated with radiation treatment was assumed to be 20% in each group. For Figure 13-2, we assumed that all the patients with poor risk had an expected 5-year survival rate of 0.54 because none received treatment.

Monitoring Ongoing Trials

In clinical trials, patients enter a study sequentially over time; therefore, information about the treatment accumulates as the trial progresses. An interim analysis is a planned analysis conducted before the final planned analysis, which allows the study sponsor to evaluate the trial's success probability while controlling the overall statistical error rates. As the trial progresses, interim analyses are usually conducted at prespecified time points, which are quantified by the number of total events. Interim analyses could be used to monitor superiority, futility, or both: If a new regimen is convincingly demonstrated to be superior to the active control, there is an ethical obligation to stop the randomization and provide the new regimen to every patient. If an interim analysis shows that there is little chance for the new regimen to outperform the control, it would be prudent to terminate the trial and save patients for other promising regimens.

There are different approaches in setting interim monitoring boundaries for superiority. For instance, Pocock[39] proposed to spend the type I error (false-positive rate) equally throughout all interim and final analyses. O'Brien and Fleming[40] proposed to reserve most of the type I error for the final analysis. The general approach is to set the interim boundaries conservatively so that the trial will not stop early for efficacy unless convincing evidence is present.

Typically, setting the interim boundaries for futility is not done as conservatively as is setting them for superiority. A nice rule of thumb proposed by Wieand et al[41] is as follows: Assuming that a study has a time-to-event primary endpoint, an interim futility check will be conducted when 50% of events targeted for final analysis have been observed. If the hazard ratio of the experimental treatment versus the control is larger than 1 (i.e., the outcomes on the experimental arm are poorer than those on the control arm at that time), the trial should be stopped for reasons of futility. The rule is simple to follow, and the probability of falsely determining that an experimental regimen is no better than the control situation when, in fact, the experimental regimen actually works is 2% or less.

Interim monitoring mainly focuses on treatment efficacy. To monitor additional aspects of a trial—for instance, to protect the safety of enrolled patients, to ensure the validity of study results, or to identify unacceptably slow accrual rates, unusually high dropout rates, or unacceptable ineligibility rates—a Data Safety Monitoring Board (DSMB) is usually recommended, if not required. The DSMB is composed of an independent panel of experts. Usually, a physician and a statistician are required. Other representatives could include epidemiologists, laboratory scientists, a patient representative, or representatives of different groups. DSMB members have access to all data and report directly to the study sponsor. In current clinical trial practice, the U.S. Food and Drug Administration (FDA) strongly recommends using a DSMB for all phase III trials and many institutional internal review boards now require use of a DSMB.

SURVIVAL ANALYSIS

Survival analysis differs from other types of statistical analysis in that the analysis of time to an event often is complicated by the lack of complete follow-up for every patient (i.e., there is censored data). An example illustrates why this is important. The example uses hypothetical data chosen to illustrate clearly the problem that censoring can cause.

Example 2. Suppose a radiation oncologist decides to review the outcomes of patients with rectal cancer treated with radiation therapy. Of primary interest is the determination of the proportion of patients who remain relapse free for 5 years. Suppose patients have been accrued over the past 7.5 years, and by chance the patients have started treatment in pairs at intervals of 1 year. Suppose, furthermore, that if no more patients were accrued and the oncologist could wait 5 more years to obtain 5-year follow-up for every patient, he or she would find that half of them were relapse free for more than 5 years. This (unobserved) relapse pattern is shown in Figure 13-3.

Twelve years after the first patient was entered, the 5-year relapse-free pattern would be as shown in Figure 13-4 (time 0 is the date of entry for each patient). If the investigator had waited 12 years (i.e., until 5-year data for all 16 patients were complete), he or she would have observed that half of the patients were relapse free 5 years after entry (i.e., the 5-year relapse-free rate is 0.50). Suppose, in fact, that the investigator decided to look at the patients' experience 7.5 years after the first patient had entered. Then the investigator would have observed everything to the left of the vertical line in

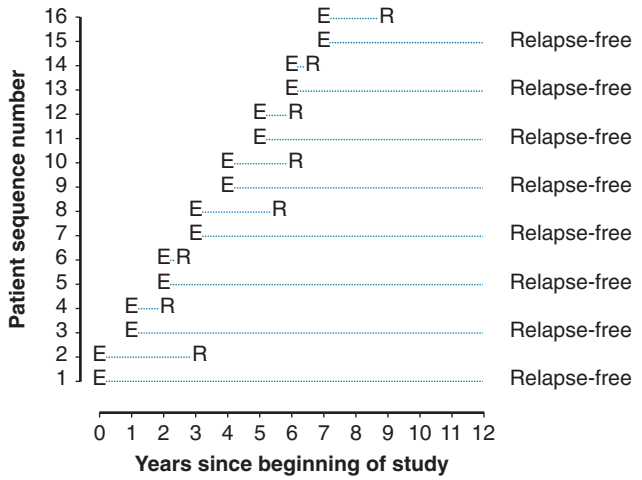


Figure 13-3 Date of entry *(E)* and relapse *(R)* of 16 patients monitored over a 12-year period; 8 did not experience relapse.

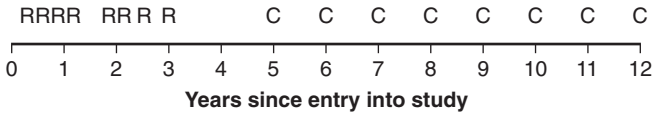


Figure 13-4 Relapse history of 16 patients seen 12 years after the beginning of the study: time from entry into the study until relapse or last follow-up. *C*, Censored; *R*, relapsed.

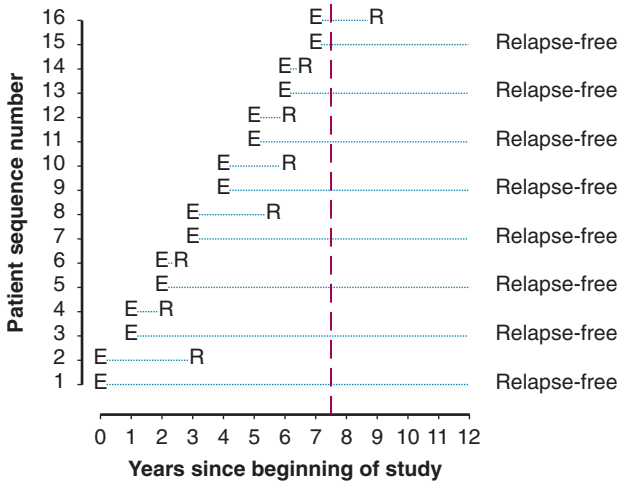


Figure 13-5 Status of follow-up 7.5 years after the first patient was entered in the trial. *E*, Entry; *R*, relapse.

Figure 13-5. Translated into time from entry, this would be represented as shown in Figure 13-6. In that case, the investigator would have complete 5-year data for 7 of the 8 patients who relapsed and for 3 patients who were relapse free at 5 years (those who were entered at year 0, 1, or 2). For the other 6 patients, his or her knowledge would be that they were relapse free for some length of time less than 5 years. The investigator's first instinct might be to exclude these 6 patients from analysis because he or she does not know what their relapse status will be after 5 years of follow-up. Of the remaining 10 patients, 7 are known to have relapsed and 3 are known to be disease free for more than 5 years, so that the estimated relapse-free rate at 5 years is 0.30. This estimate does not seem to reflect the data accurately.

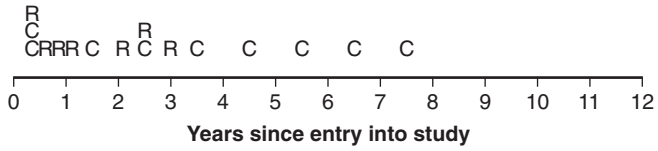


Figure 13-6 Relapse history of 16 patients seen 7.5 years after the beginning of the study: time from entry into the study until relapse or last follow-up. *C*, Censored; *R*, relapsed.

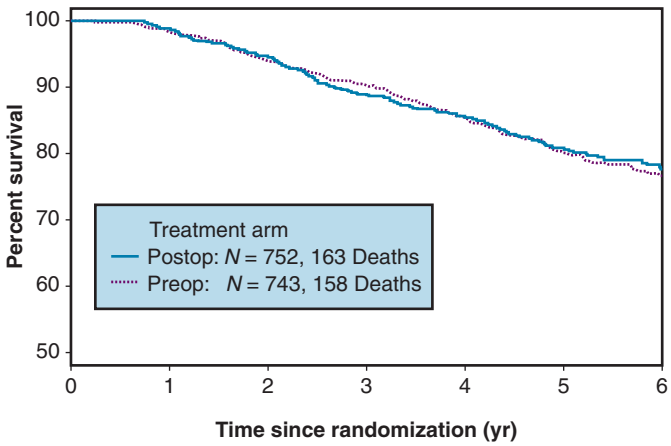


Figure 13-7 Survival of patients with operable breast cancer treated with doxorubicin and cyclophosphamide.

The estimate is so different from 0.50 because the method of calculation is biased. The cause of the bias can be seen by studying the seventh and eighth patients in the series of patients (i.e., the patients entered 3 years after the first patient entered). Notice that the seventh patient entered was still relapse free when the radiation oncologist performed his or her analysis but had only been followed for 4.5 years (i.e., did not have a known status at 5 years) and, hence, was excluded. However, the eighth patient (who was entered on the same day) had relapsed by the time of analysis and, therefore, was included. The bias is that patients who relapse have a better chance of being included in the analysis than those who do not relapse. One way to avoid a biased estimate would be to exclude all patients who did not have the potential to be monitored for 5 years. Therefore the oncologist would only be able to use the information from the first six patients entered. In this artificial example, this would have led to a correct estimate of the 5-year relapse-free rate because three of the first six patients relapsed within 5 years.

This method is unsatisfying because not all of the available information is used. In statistical terms, the disadvantage of our proposed solution is that there is considerably more variance associated with an estimate that uses the data from only six patients than one that uses all of the available data. For example, if patients 6 and 7 had entered the study in the opposite order, the relapse-free estimate would jump from 50% to 67%. A better method is described in the next section.

Kaplan-Meier Method

A common way to summarize survival data is to estimate the "survival curve," using a method proposed by Kaplan and Meier.[42] The Kaplan and Meier curve shows—for each value of time—the proportion of subjects who survive at least that length of time. Figure 13-7 shows survival curves for patients with operable breast cancer who have been treated with either preoperative or postoperative chemotherapy (doxorubicin

and cyclophosphamide) in a large clinical trial. Here, we offer two examples that provide some insight regarding the Kaplan-Meier method. The first example concerns a data set having no censored observations; the second example extends these ideas to accommodate censored observations.

Example 3. Suppose four patients are entered into a study, possibly at different times, and all four die, with death times of 10 months, 20 months, 25 months, and 40 months, respectively, from time of entry. We want to estimate the 3-year survival rate. Three of the four patients died within 3 years from entry, and one patient remained alive more than 3 years from entry (i.e., one fourth of the patients entered lived at least 3 years). Therefore, a reasonable estimate of the 3-year survival rate is $S(3) = 0.25$. This estimate (the proportion of patients who remained alive at 3 years) is referred to as the empirical survival estimate, and the graph of estimates obtained this way at all time points is called the *empirical survival curve.*

Example 4. Suppose that in the previous example three of the patients have died at 10 months, 25 months, and 40 months, respectively. The fourth patient is still alive but has been monitored for only 20 months. We want to estimate the 3-year survival rate. Notice that if we could wait another 16 months to obtain the estimate (so that we would have 3-year data for all patients), we would obtain an estimate of either one fourth (if the fourth patient dies soon) or one half (if the patient remains alive for another 16 months). The rather intuitive approach we used in Example 3 to estimate the 3-year survival rate does not work here because we do not know whether there will be two or three survivors when all the patients have died or have been monitored for 3 years (i.e., we do not know how to handle the patient whose follow-up was censored at 20 months). Kaplan and Meier proposed an approach that updates the survival estimate at the time of each death using only those patients who are at risk of failing at each update.

Because none of the patients died during the first 10 months, the Kaplan-Meier estimate of the probability of surviving to any time point less than 10 months is equal to 1. Because four patients are alive at 10 months but only three quarters of them survive beyond 10 months, the Kaplan-Meier estimate changes to three quarters for time points beyond 10 months but before the next death. Although one patient is censored at 20 months, this gives no information regarding the likelihood of a death, so the Kaplan-Meier estimate remains at three quarters for all time points between 10 months and 25 months. One of the two patients who are still at risk just before 25 months dies at that time. Therefore, the estimate of the probability of a patient surviving beyond 25 months, given that the patient survived at least 25 months, is one half. The Kaplan-Meier estimate of surviving more than 25 months is the product of three quarters (the estimated probability a patient will survive until 25 months) times one half (the estimated probability that a patient will survive more than 25 months given that the patient was alive at 25 months), which is three eighths. There are no other deaths before 3 years, so the Kaplan-Meier estimate of the 3-year survival rate is three eighths. The Kaplan-Meier method uses all the relevant available data from each patient but excludes the patient who was censored at 20 months from all computations beyond that time.

If the Kaplan-Meier approach is applied to the data in Example 3, the estimate $S(3)$ is equal to the (probability of surviving 10 months) × (probability of surviving more than 10 months given survival of 10 months) × (probability of surviving more than 20 months given survival of 20 months) × (probability of surviving more than 25 months given survival of 25 months), or $1 \times \frac{3}{4} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{4}$, which matches the empirical survival estimate. In fact, the Kaplan-Meier estimate and the empirical estimate always match when they are applied to uncensored data.

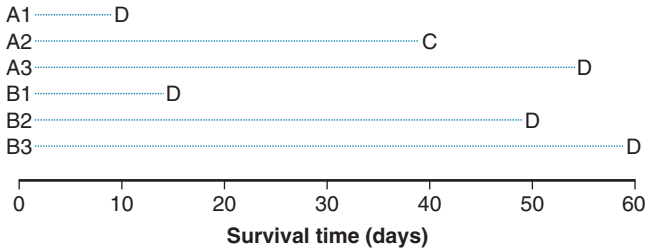


Figure 13-8 Survival times of six patients treated with one of two regimens. *C*, Censored; *D*, death.

TABLE 13-4 Observed Deaths (t=10)

Treatment	Dead	Alive	Total
A	1	2	3
B	0	3	3
Total	1	5	6

The reader may verify that application of the Kaplan-Meier approach to the data in Figure 13-6 will result in an estimate of 0.45 for the probability of remaining relapse free through 5 years.

Log-Rank Statistic

Perhaps the most common application of survival analysis techniques is the comparison of survival times (or other times to event, such as time to disease relapse) between two or more groups of patients that differ in some aspect (e.g., male versus female, or treated versus untreated). The log-rank statistic, the stratified log-rank statistic, and the Cox proportional hazard model are commonly used to compare survival times between groups.

The log-rank statistic deals with the problem of censoring by comparing the groups only when a patient within any of the groups experiences an "event" (if survival times are to be compared across groups, an event would be a death; if times to relapse are to be compared, an event would be a relapse, and so on). This idea is most easily explained in the context of a simple example. Suppose we monitored three patients receiving a standard treatment regimen (this might even be no treatment), which we refer to as treatment A, and three other patients receiving an experimental regimen, which we refer to as treatment B. Suppose, furthermore, that the survival times for the patients receiving treatment A are 10, 40+, and 55 days, respectively, and for the patients receiving treatment B, 15, 50, and 60 days, respectively (a plus sign after a value refers to a censored time; i.e., a patient with a time of 40+ days was last known to be alive at 40 days and no further follow-up is available). These data are represented graphically in Figure 13-8; the three patients receiving treatment A are labeled A1, A2, and A3, and those receiving treatment B are labeled B1, B2, and B3. Deaths are denoted by the letter D, and censored survival times are labeled with the letter C.

When evaluating the log-rank statistic, the first computation occurs at time $t = 10$, the time at which the first death is observed. Just before this point in time, all six patients are known to be alive (and, hence, are "at risk" to die at time $t = 10$). Three of these patients received treatment A and three received treatment B. Exactly one of these six patients is known to have died at time $t = 10$, and he or she received treatment A. Table 13-4 summarizes the status of patients at this time point. Notice that at the time of this computation

TABLE 13-5 Observed Deaths (t=15)

Treatment	Dead	Alive	Total
A	0	2	2
B	1	2	3
Total	1	4	5

TABLE 13-6 Observed Deaths (t=50)

Treatment	Dead	Alive	Total
A	0	1	1
B	1	1	2
Total	1	2	3

TABLE 13-7 Observed Deaths (t=55)

Treatment	Dead	Alive	Total
A	1	0	1
B	0	1	1
Total	1	1	2

there are three patients on each arm. Therefore, if treatment B was equivalent to treatment A and only one death occurred on one arm or the other, there would be a one half chance that the death is on arm A. In fact, the death is on arm A; hence, we observe one death when the probability of observing a death on arm A is one half (i.e., there is one half more death on arm A than is expected).

Observations at time $t = 15$ are listed in Table 13-5. At this point there would be a two fifths probability that the death would occur on treatment A (if the treatments were equivalent), but the death did not occur on treatment A, so there were two fifths fewer deaths on arm A than expected (i.e., the observed deaths minus the expected number is $0 - \frac{2}{5} = -0.4$).

The next computation occurs at time $t = 50$; observations are listed in Table 13-6. Notice that patient A2 is not included in this table, even though she is not known to have died at any time before $t = 50$. Because she was lost to follow-up (censored) at time 40, she is no longer "at risk" at time 50. At this time, there would be a one third probability that the death would occur on treatment A (if the treatments were equivalent), but the death does not occur on treatment A, so there are one third fewer deaths on arm A than expected (i.e., the observed deaths minus the expected number is $0 - \frac{1}{3} = -0.33$).

One may go through the same computations at time $t = 55$ (Table 13-7) and will determine that the number of deaths minus expected deaths on treatment A is $1 - \frac{1}{2} = 0.5$. At time $t = 60$, all remaining patients are on the same arm, so the observed minus the expected number of deaths must be 0. Adding up the observed minus the expected number of deaths at times 10, 15, 50, and 55, one obtains 2 observed deaths minus 1.733 expected deaths, so that there were 0.27 deaths more than expected on arm A, indicating that this treatment might be harmful. However, one's intuition is that this is not a significant difference (i.e., such a small difference could easily be attributed to the play of chance), and in fact this is true. Therefore, there is no strong evidence in this example that treatments A and B differ.

More formally, the log-rank statistic is defined to be the difference in observed and expected numbers of deaths on one of the two treatment arms. The statistic may be standardized by dividing by the square root of its variance, yielding a score that, under the hypothesis of equivalent treatments, is approximately standard normal. In the present example, the variance of the log-rank statistic can be shown to equal 0.9622, so the standardized test statistic is $(2 - 1.733)/0.9622 = 0.27$, corresponding to a two-sided p value of about 0.79.

Cox Proportional Hazard Model

When controlling for the association of a single variable with patient survival, the stratified log-rank test can be used. However, in terms of adjusting for additional explanatory variables, the most popular method is the Cox proportional hazards model.[43] The Cox model explores the relationship between survival experience and prognostic variables or explanatory variables. It estimates and tests the hazard ratio between different groups. In the Cox regression model, explanatory variables could be continuous (for instance, age or weight or systolic blood pressure), categorical (for instance, gender or race or Eastern Cooperative Oncology Group [ECOG] performance score), or an interaction between different categorical variables. The Cox regression model does not make any parametric assumption about the survival probability distribution of each group; however, it does assume a proportional hazard between two groups. If we use $h_i(t)$ and $h_0(t)$ to denote the hazards of death at time t for the ith patient and the baseline patient, that is, the patient with all explanatory variables taking values 0, respectively, the proportional hazard model can be expressed as

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_p x_{pi}) h_0(t),\ t \geq 0$$

where $x_1, x_2, \ldots x_p$ are p different explanatory variables, $h_0(t)$ is called the baseline hazard function, and $h_i(t)/h_0(t)$ is a constant, called the hazard ratio. The βs are the regression coefficients. For a continuous explanatory variable x_p, β_p stands for the log of the hazard ratio between the ith patient and the baseline patient. For a categorical explanatory variable x_1, for instance, assume that $x_1 = 0$ if the patient is randomized to the control group, or $x_1 = 1$ if the patient is randomized to the experimental group. Then β_1 stands for the log of the hazard ratio between patients in the experimental group versus the control group, whereas all other explanatory variables are the same.

Because the Cox regression model assumes the proportional hazard between different groups, it is necessary to check this assumption before concluding the estimations. Most of the model check procedures are based on graphs or plots of model-fitting residuals. If the proportional hazard assumption does not fit the data even after the data transformation, other models should be considered.

CURRENT TOPICS IN PHASE III CLINICAL TRIALS

Surrogate Endpoints

The primary endpoint is one of the most critical elements in formulating the study design, data collection, and statistical analysis plan of a clinical trial. It is a quantitative measurement implied or required by the primary objective of a study. The best endpoint is a clinical measurement reflecting the most relevant potential treatment effect of the new regimen that can be defined with rigorous mathematical and statistical properties. The OS rate has historically been the primary outcome for most phase III oncology trials because of its clear virtues: it is simple to measure, unambiguous, the least susceptible to investigator bias, and of unquestionable clinical relevance. Despite these many advantages, the role of OS as a primary endpoint is challenged in modern phase III clinical trials. Two primary challenges to the OS endpoint

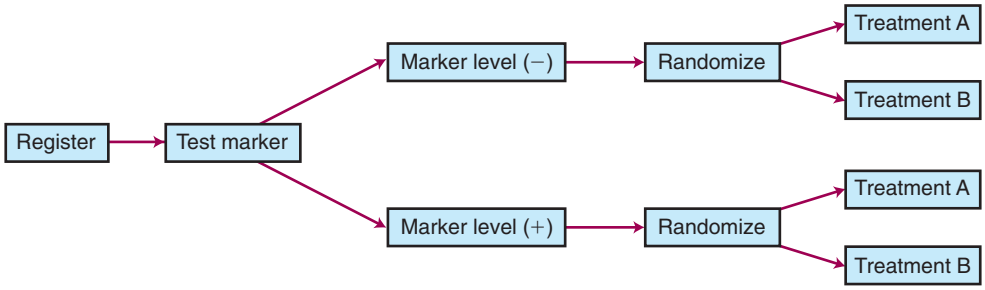


Figure 13-9 Marker by treatment interaction to test a predictive factor question; same treatments in both prognostic groups.

are that in many cases an extensive follow-up period is needed to obtain survival status and that in many diseases, multiple effective therapies are now available. In this case, OS is affected by all therapies given to a patient and, as such, this endpoint is insensitive to the impact of changing a single line of therapy.

One of the alternatives is to use a surrogate endpoint instead of a clinical endpoint in a phase III trial. A surrogate endpoint is an endpoint obtained sooner, at lesser cost, or less invasively than the long-term clinical efficacy endpoint. The Biomarkers Definitions Working Group (BDWG) defines a surrogate endpoint as "a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit (or harm, or lack of benefit or harm)."[44] When using a surrogate endpoint, one would like to make the same inference as if one had observed a true endpoint. Therefore, a critical requirement of a valid surrogate endpoint is that the treatment effect observed on a surrogate endpoint should reliably predict the treatment effect on the clinical endpoint. This implies a stronger requirement for a surrogate endpoint than simply a significant correlation between it and the clinical endpoint. In circumstances where the speed of the study is a major concern, where second-line therapies that affect the OS potentially obscure the assessment of first-line treatment benefit, or where crossover becomes unavoidable, a surrogate endpoint might be considered in place of the long-term clinical endpoint.

From statistical perspectives, there are two branches of surrogate evaluation methodologies: a single-trial or meta-analytic evaluation. In the single-trial approach, the "proportion of treatment effect" (PTE)[45,46] explained by the surrogate, has been the predominant approach. However, a single study is unable to provide reliable information regarding the prediction of the treatment effect on the clinical endpoint based on the observed treatment effect on potential surrogate endpoint. Therefore, to validate an endpoint as a legitimate surrogate endpoint, a meta-analysis is usually required. In addition, heterogeneity between the trials included in the meta-analysis strengthens the robustness of results from individual trials. In conducting a meta-analysis, using individual patient data instead of summary data is highly recommended, and, ideally, both positive and negative trials should be included. Examples of meta-analyses conducted to examine potential surrogate endpoints can be found in articles by Sargent and colleagues[47] and Burzykowski and associates.[48]

Biomarkers

A marker is a single trait, or a group of traits, that differentiates patients with respect to an outcome of interest. If a marker can be used to identify patients with differing risks of a clinical outcome (such as progression or death) in the absence of therapy or when receiving nontargeted standard treatment, the marker is usually called a *prognostic marker.* If a marker predicts differential efficacy of a specific therapy, the marker is called a *predictive marker.*

Validation of a prognostic marker is usually conducted retrospectively in patients treated with placebo or a standard treatment. Validation of a predictive marker could also be conducted retrospectively based on data from a randomized controlled trial.[49] However, a prospective randomized controlled trial would be ideal. Two types of clinical designs can be used to validate a predictive marker: a targeted/selection design or an unselected design. A targeted trial enrolls only patients who are most likely to respond to the experimental therapy based on their molecular expression levels. On the one hand, a targeted trial could result in a large savings of patients for other trials. On the other hand, it could miss efficacy in other patients and miss the opportunity to test the association of the biologic endpoints with clinical outcomes.

Different types of unselected designs have been proposed and discussed. For instance, the marker-by-treatment interaction design (Figure 13-9) stratifies patients according to their marker status and randomly assigns patients in each marker group to two different treatments. Hypothesis setting and sample size estimation could be different in each marker-defined subgroup. Either a formal test for marker-by-treatment interaction or a separate superiority test within each marker group could be conducted. Another popular unselected design is the marker-based strategy design (Figure 13-10), in which the random treatment assignment could be based on the patient's marker status or could be independent of it. Examples for each type of design can be found in the article by Sargent and colleagues.[50]

Adaptive Design

An adaptive design could be defined as a design "that allows adaptations to trial procedures (for instance, eligibility criteria, study dose, or treatment procedure, etc.) and/or statistical procedures (for instance, randomization, study design, study hypothesis, sample size, or analysis methods, etc.) of the trial after its initiation without undermining the validity and integrity of the trial."[51] Many types of adaptation could be applied to an ongoing trial. For instance, a study could be resized based on an interim review of the outcomes in the control group. Assume, for example, that in the design of a study, the progression-free survival (PFS) under the standard of care was assumed to be 12 months. However, on an interim review of accrued data from the control arm, the estimated PFS was 15 months. In this case, more patients need to be accrued for the study or patients need to be followed longer for the same hazard ratio to be detected. No statistical penalty is required for this review, because only data from the control arm have been analyzed.

For another example of an adaptive design, assume that a study started with n experimental arms and one control arm. Within this study, based on a prespecified plan, all

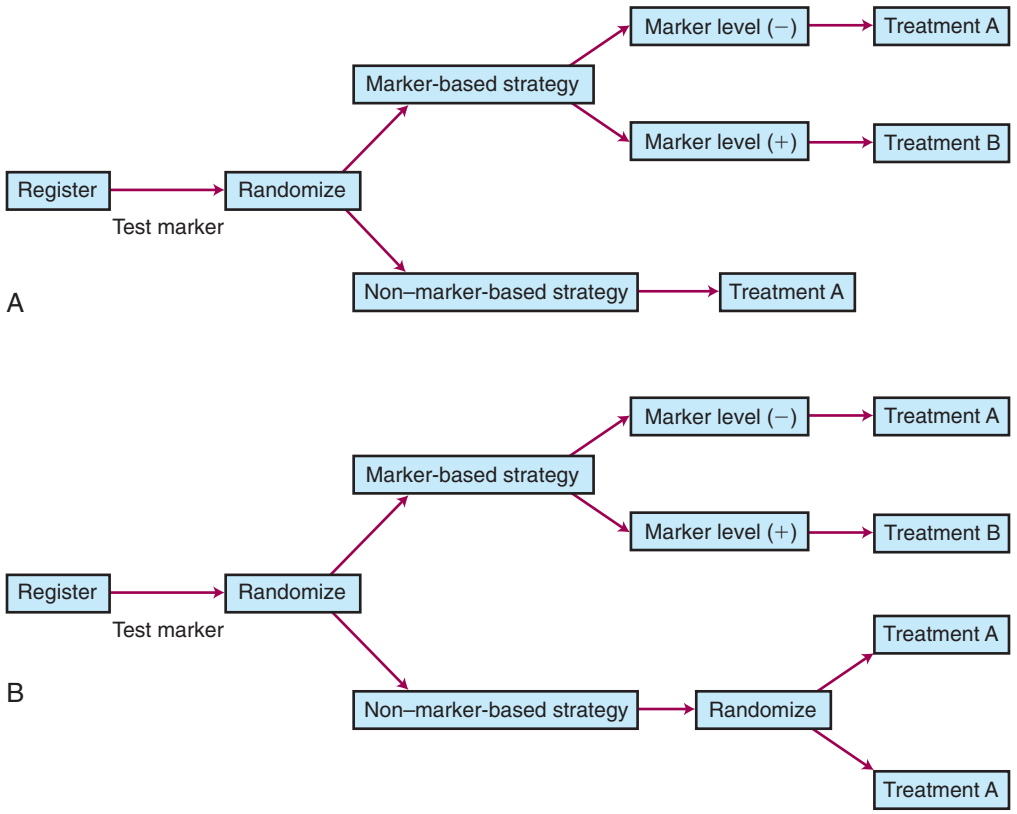


Figure 13-10 **A,** Marker-based strategy design to test predictive factor question; no randomization in non–marker-based arm. **B,** Marker-based strategy design to test predictive factor question; randomization in both arms.

experimental arms are to be compared with the control arm, at the interim analysis, and only the most promising experimental arm and the control arm will continue to accrue when patient accrual is resumed. In other words, the study will have started with $n+1$ arms and ended with 2 arms, and $n-1$ arms will have been dropped after the interim analysis. Because no conclusions regarding superiority were allowed to be made at the interim analysis, no statistical penalty for the interim review is needed. Under other circumstances, if an adaptation relates to (1) altering the randomization ratio between the arms, (2) resizing the trial based on interim comparison between the experimental and control arms, or (3) changing the primary endpoint, then a penalty for looking at the data before the end of the study is definitely needed. In planning an adaptive design, possible adaptations must be specified in the protocol before accrual takes place. In applying an adaptive design, an efficient mechanism for processing and analyzing data is critical.

As drug development moves toward targeted agents with particular treatment mechanisms through a targeted pathway, the disease population for testing the treatment effects may need to be refined and subseted based on a relevant biomarker signature. By subgrouping the disease population, the traditional design becomes challenging. Under this circumstance, Bayesian adaptive designs guided by biomarkers have been explored and applied. For example, the Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE) trial and Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis 2 (ISPY-2) trials are two recent biomarker-based adaptively randomized studies. In the BATTLE trial, four treatments (Erlotinib, Vandetanib, Eroltinib + bexarotene, Sorafenib) and four biomarker profiles (EGFR mutation, KRAS/Braf mutation, VEGF/VEGFR-2 expression, RXRs/Cyclin D1 expression and CCND1 copy number) were considered. The trial started with equal allocation randomization to four treatments, then was switched to adaptive randomization that would increasingly assign patients into treatments with greatest potential for efficacy based on their individual biomarker profile.[52] Similarly, ISPY-2 was aimed to identify regimens which can improve outcomes for patients with particular biomarker signatures in the neoadjuvant chemotherapy setting for patients with breast cancer. A concurrent randomization between experimental and standard care regiments was conducted within the patient populations defined by biomarkers. The adaptation was based on the Bayesian predictive probability of the experimental agent being more effective than standard therapy to allow "graduation" or "dropping out" of the pair of regimen and biomarker.[53] Both studies highlight the feasibilities of applying advanced trial designs in practice and also illustrate powerful collaboration between statisticians and clinical, laboratory, and bioinformatics investigators.

ACKNOWLEDGMENT

The authors wish to acknowledge that parts of this chapter were taken directly from prior versions of the chapter, written by Dr. Sam Wieand and Dr. John Bryant, who are both deceased.

CRITICAL REFERENCES

A full list of cited references is published online at www.expertconsult.com.

1. Meinert CL: Clinical trials: Design, conduct, and analysis, Oxford, 1989, Oxford University Press.
2. Piantadosi S: Clinical trials: A methodological perspective, New York, 1997, Wiley.

6. Giovagnoli A, Pintacuda N: Properties of frequency distributions induced by general "up-and-down" methods for estimating quantiles. J Stat Plan Inference 74:51–63, 1998.
12. Garrett-Mayer E: The continual reassessment method for dose-finding studies: A tutorial. Clin trials 3:57–71, 2006.
14. Thall PF, Cook JD: Dose-finding based on efficacy-toxicity trade-offs. Biometrics 60:684–693, 2004.
15. Braun TM, Jia N: A generalized continual reassessment method for two-agent phase I trials. Stat Biopharm Res 5:105–115, 2013.
18. Braun TM: The current design of oncology phase I clinical trials: Progressing from algorithms to statistical models. Chin Clin Oncol 3:2, 2014.
21. Lee YJ, Staquet M, Simon R, et al: Two-stage plans for patient accrual in phase II cancer clinical trials. Cancer Treat Rep 63:1721–1726, 1979.
22. Fleming TR: One-sample multiple testing procedure for phase II clinical trials. Biometrics 38:143–151, 1982.
23. Chang MN, Therneau TM, Wieand HS, et al: Designs for group sequential phase II clinical trials. Biometrics 43:865–874, 1987.
24. Simon R: Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10:1–10, 1989.
25. Therneau TM, Wieand HS, Chang M: Optimal designs for a grouped sequential binomial trial. Biometrics 46:771–781, 1990.
26. Bryant J, Day R: Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics 51:1372–1383, 1995.
27. Thall PF, Simon RM, Estey EH: Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med 14:357–379, 1995.
28. Tang H, Foster NR, Grothey A, et al: Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. J Clin Oncol 28:1936–1941, 2010.
29. Simon R, Wittes RE, Ellenberg SS: Randomized phase II clinical trials. Cancer Treat Rep 69:1375–1381, 1985.
31. Rubinstein LV, Korn EL, Freidlin B, et al: Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol 23:7199–7206, 2005.
32. Pocock SJ: Clinical trials: A practical approach, New York, 1984, Wiley, pp 182–186.
35. Pocock SJ, Simon R: Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31:103–115, 1975.
38. Gail MH: Eligibility exclusions, losses to follow-up, removal of randomized patients, and uncounted events in cancer clinical trials. Cancer Treat Rep 69:1107–1113, 1985.
40. O'Brien PC, Fleming TR: A multiple testing procedure for clinical trials. Biometrics 35:549–556, 1979.
47. Sargent DJ, Wieand HS, Haller DG, et al: Disease-free survival versus overall survival as a primary endpoint for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol 23:8664–8670, 2005.
48. Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al: Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate endpoints in metastatic breast cancer. J Clin Oncol 26:1987–1992, 2008.
50. Sargent DJ, Conley BA, Allegra C, et al: Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 23:2020–2027, 2005.
51. Chow SC, Chang M: Adaptive design methods in clinical trials—a review. Orphanet J Rare Dis 3:11, 2008.

REFERENCES

1. Meinert CL: Clinical trials: Design, conduct, and analysis, Oxford, 1989, Oxford University Press.
2. Piantadosi S: Clinical trials: A methodological perspective, New York, 1997, Wiley.
3. Steyerberg EW: Clinical prediction models: A practical approach to development, validation, and updating, New York, 2009, Springer.
4. Ivanova A: Escalation, group and A + B designs for dose-finding trials. Stat Med 25:3668–3678, 2006.
5. Goldberg RM, Kaufmann SH, Atherton P, et al: A phase I study of sequential irinotecan and 5-fluorouracil/leucovorin. Ann Oncol 13(10):1674–1680, 2002.
6. Giovagnoli A, Pintacuda N: Properties of frequency distributions induced by general "up-and-down" methods for estimating quantiles. J Stat Plan Inference 74:51–63, 1998.
7. Stylianou M, Flournoy N: Dose finding using the biased coin up-and-down design and isotonic regression. Biometrics 58:171–177, 2002.
8. Ivanova A, Flournoy N, Chung Y: Cumulative cohort design for dose-finding. J Stat Plan Inference 137:2316–2327, 2007.
9. O'Quigley J, Pepe M, Fisher L: Continual reassessment method: A practical design for phase 1 clinical trials in cancer. Biometrics 46:33–48, 1990.
10. Faries D: Practical modifications of the continual reassessment method for phase I cancer clinical trials. J Biopharm Stat 4:147–164, 1994.
11. Goodman SN, Zahurak ML, Piantadosi S: Some practical improvements in the continual reassessment method for phase I studies. Stat Med 14:1149–1161, 1995.
12. Garrett-Mayer E: The continual reassessment method for dose-finding studies: A tutorial. Clin trials 3:57–71, 2006.
13. Braun TM: The bivariate continual reassessment method. extending the CRM to phase I trials of two competing outcomes. Control Clin Trials 23:240–256, 2002.
14. Thall PF, Cook JD: Dose-finding based on efficacy-toxicity trade-offs. Biometrics 60:684–693, 2004.
15. Braun TM, Jia N: A generalized continual reassessment method for two-agent phase I trials. Stat Biopharm Res 5:105–115, 2013.
16. Wages NA, Conaway MR, O'Quigley J: Continual reassessment method for partial ordering. Biometrics 67:1555–1563, 2011.
17. Rogatko A, Schoeneck D, Jonas W, et al: Translation of innovative designs into phase I trials. J Clin Oncol 25:4982–4986, 2007.
18. Braun TM: The current design of oncology phase I clinical trials: Progressing from algorithms to statistical models. Chin Clin Oncol 3:2, 2014.
19. Gehan A: The determination of the number of patients required in a follow-up trial of new chemotherapeutic agent. J Chronic Dis 13:346–353, 1961.
20. Herson J: Predictive probability early termination plans for phase II clinical trials. Biometrics 35:775–783, 1979.
21. Lee YJ, Staquet M, Simon R, et al: Two-stage plans for patient accrual in phase II cancer clinical trials. Cancer Treat Rep 63:1721–1726, 1979.
22. Fleming TR: One-sample multiple testing procedure for phase II clinical trials. Biometrics 38:143–151, 1982.
23. Chang MN, Therneau TM, Wieand HS, et al: Designs for group sequential phase II clinical trials. Biometrics 43:865–874, 1987.
24. Simon R: Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10:1–10, 1989.
25. Therneau TM, Wieand HS, Chang M: Optimal designs for a grouped sequential binomial trial. Biometrics 46:771–781, 1990.
26. Bryant J, Day R: Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics 51:1372–1383, 1995.
27. Thall PF, Simon RM, Estey EH: Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med 14:357–379, 1995.
28. Tang H, Foster NR, Grothey A, et al: Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. J Clin Oncol 28:1936–1941, 2010.
29. Simon R, Wittes RE, Ellenberg SS: Randomized phase II clinical trials. Cancer Treat Rep 69:1375–1381, 1985.
30. Sargent DJ, Goldberg RM: A flexible design for multiple armed screening trials. Stat Med 20:1051–1060, 2001.
31. Rubinstein LV, Korn EL, Freidlin B, et al: Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol 23:7199–7206, 2005.
32. Pocock SJ: Clinical trials: A practical approach, New York, 1984, Wiley, pp 182–186.
33. Blackwelder WC: "Proving the null hypothesis" in clinical trials. Control Clin Trials 3:345–353, 1982.
34. Taves DR: Minimization: A new method of assigning patients to treatment and control groups. Clin Pharmacol Ther 15:443–453, 1974.
35. Pocock SJ, Simon R: Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31:103–115, 1975.
36. Freedman LS, White SJ: On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. Biometrics 32:691–694, 1976.
37. Begg CB, Iglewicz B: A treatment allocation procedure for sequential clinical trials. Biometrics 36:81–90, 1980.
38. Gail MH: Eligibility exclusions, losses to follow-up, removal of randomized patients, and uncounted events in cancer clinical trials. Cancer Treat Rep 69:1107–1113, 1985.
39. Pocock SJ: Group sequential methods in the design and analysis of clinical trials. Biometrika 64:191–199, 1977.
40. O'Brien PC, Fleming TR: A multiple testing procedure for clinical trials. Biometrics 35:549–556, 1979.
41. Wieand S, Schroeder G, O'Fallon JR: Stopping when the experimental regimen does not appear to help. Stat Med 13:1453–1458, 1994.
42. Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. J Am Statist Assoc 53:457–481, 1958.
43. Cox DR: Regression models and life-tables (with discussion). J Roy Statist Soc Ser B 34:187–220, 1972.
44. Biomarkers Definitions Working Group: Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Ther 69(3):89–95, 2001.
45. Freedman LS, Graubard BI, Schatzkin A: Statistical validation of intermediate endpoints for chronic diseases. Stat Med 11:167–178, 1992.
46. Lin DY, Fleming TR, De Gruttola V: Estimating the proportion of treatment effect explained by a surrogate marker. Stat Med 16:1515–1527, 1997.
47. Sargent DJ, Wieand HS, Haller DG, et al: Disease-free survival versus overall survival as a primary endpoint for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol 23:8664–8670, 2005.
48. Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al: Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate endpoints in metastatic breast cancer. J Clin Oncol 26:1987–1992, 2008.
49. Simon RM, Paik S, Hayes DF: Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst 101:1446–1452, 2009.
50. Sargent DJ, Conley BA, Allegra C, et al: Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 23:2020–2027, 2005.
51. Chow SC, Chang M: Adaptive design methods in clinical trials—a review. Orphanet J Rare Dis 3:11, 2008.
52. Kim ES, Herbst RS, Wistuba II, et al: The BATTLE trial: Personalizing therapy for lung cancer. Cancer Discov 1(1):44–53, 2011.
53. Barker AD, Sigman CC, Kelloff GJ, et al: I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin Pharmacol Ther 86:97–100, 2009.