

---

# THERAPEUTIC STUDIES

Lawrence V. Rubinstein, PhD

The investment in therapeutic oncology studies in North America is enormous. Every year, approximately 20,000 potential anticancer agents are screened by the United States National Cancer Institute (NCI) in vitro human tumor cell line assay, and in 1999 approximately 30 new agents were taken to clinical trial.<sup>22, 37</sup> In North America there are currently 10 government-funded cancer cooperative clinical trials groups (American College of Surgeons Oncology Group, Cancer and Leukemia Group B, Children's Oncology Group, Eastern Cooperative Oncology Group, Gynecologic Oncology Group, National Cancer Institute of Canada, National Surgical Adjuvant Breast and Bowel Project, North Central Cancer Treatment Group, Radiation Therapy Oncology Group, and Southwest Oncology Group) and approximately 60 cancer centers, primarily large academic medical centers. Among them, they represent approximately 8000 investigators at 1700 institutions and place more than 20,000 patients on therapeutic studies annually.<sup>22, 37</sup>

By 1960, the current clinical pathway for the development of a potential anticancer agent or regimen had been defined<sup>14</sup>: phase I trials of approximately 20 patients to determine the appropriate dose and schedule for further testing, usually the maximal tolerated dose (MTD); followed by phase II trials of approximately 30 to 50 patients to establish an indication of clinical effect, usually tumor shrinkage; followed by randomized phase III trials involving hundreds or thousands of patients to test clinical efficacy, usually defined as the ability to prolong survival, compared with a control therapy. Since the early 1960s, the cooperative groups have accrued patients according to standardized clinical protocols, with standardized criteria of diagnosis, treatment, and measurement of effect and with a prospective statistical design and collaborative

---

From the Biometric Research Branch, National Cancer Institute, Bethesda, Maryland

---

HEMATOLOGY/ONCOLOGY CLINICS OF NORTH AMERICA

analysis and reporting. The early development of the cooperative group statistical centers was led by NCI statistician Marvin Schneiderman,<sup>14</sup> and since then statisticians from the National Institutes of Health (NIH) and from cooperative group and cancer center statistical centers have developed most of the new statistical methodology for the design, conduct, and analysis of cancer clinical studies.<sup>37, 41</sup>

A significant body of literature reviews clinical trials methodology. For cancer trials in particular, possibly the most inclusive recent work is that of Piantadosi.<sup>34</sup> Two shorter, but still comprehensive, studies are that by Leventhal and Wittes,<sup>26</sup> which gives a clinician's perspective, and that of Green et al,<sup>16</sup> which gives a statistician's perspective. Simon<sup>43</sup> has written a useful chapter in the text by DeVita et al and has also written a review of clinical trials methodology<sup>41</sup> in the active decade of the 1980s. The author has drawn on these sources liberally in this article and has also referred to original methodology papers. This article attempts, as much as possible, to keep the discussion on an intuitive and nontechnical level. It elucidates the problems in basic design, conduct, and interpretation associated with phase I, phase II, and phase III trials and explains how the various statistical approaches have arisen as solutions to these problems. The fundamental problem common to all three trial types is that of achieving a correct and precise answer to the question posed by the trial, to inform future testing and treatment better, while protecting the trial patients from receiving treatment that has demonstrated excessive toxicity or lack of clinical efficacy. This shared problem gives rise to statistical designs with basic similarities across the three trial types.

## PHASE I TRIALS

The objective of a phase I trial is to determine the appropriate dosage of an agent or combination regimen that will be taken to further study and to provide initial pharmacologic and pharmacokinetic studies. At this stage of testing it is generally assumed that increased dose is associated with increased chance of clinical efficacy, so the phase I trial is designed as a dose-escalation study to determine the MTD, that is, the maximum dose associated with an acceptable level of dose-limiting toxicity (DLT). (Dose-limiting toxicity is usually defined as toxicity of grade 3 or higher, except for grade 3 neutropenia unaccompanied by either fever or infection.<sup>45</sup>) This MTD is then taken into further testing. Because efficacy is generally not an important endpoint in phase I trials, it is not necessary to restrict the trial population to patients homogeneous with respect to disease or even to patients with measurable disease (for which tumor response is determinable), but it is important to exclude patients with impaired organ function who might therefore be more prone to serious toxicity. The fundamental conflict in phase I trials is between escalating the dosage too rapidly, exposing patients to

excessive toxicity, and escalating too slowly, depriving patients of treatment at potentially efficacious dose levels.<sup>4</sup>

The first problem is deciding on a safe but not overly conservative initial dosage for the trial. If the agent is new to clinical testing, this dosage must be based on animal studies. It has been determined that the dose (defined in mg/m<sup>2</sup> of body surface area) associated with 10% lethality in mice (MELD<sub>10</sub>) can be predicted to be roughly equivalent to the human MTD.<sup>17</sup> Therefore, the initial dose is taken to be either one tenth of the MELD<sub>10</sub> or, if smaller, one third of the LD<sub>10</sub> associated with the beagle dog.<sup>26</sup> The next problem is defining dose increments for the subsequent dose levels, and it is here that the various phase I trial designs differ.

### Standard Phase I Design

To define dose levels beyond the initial dose, the *standard* phase I design uses a set of decreasing *Fibonacci* increments proposed by Schneiderman,<sup>38</sup> currently taken to be 100%, 67%, 50%, 40%, and 33% thereafter.<sup>4</sup> These increments are added to each dose level to determine the succeeding level. In other words, the second dose level is 100% greater than the first, the third is 67% greater than the second, and so forth. This schedule allows more aggressive dose escalation for the initial levels, which are expected to be sufficiently removed from the MTD for such escalation to be safe.

The *standard* rule governing dose escalation from one level to the next makes no assumptions concerning the shape of the dose-toxicity curve or the potential for cumulative toxicity; the decision to escalate to the next dose level is based solely on first-course toxicity results of the current level. The dose escalation rules proceed as shown in Table 1 escalating in cohorts of three to six patients per dose level.<sup>45</sup> Up to three patients are treated at the current dose level. If at least two patients are observed to have DLT, the prior dose level is defined as the MTD (unless only three patients have been treated at that level, in which case it is the

**Table 1. DOSE ESCALATION RULES FOR THE STANDARD PHASE I TRIAL**

<b>Outcome: First-course DLT/Patients</b>	<b>Action: Escalate, Suspend, or Halt Dose Escalation</b>
0 DLT out of 3 patients	Escalate dose for next cohort of 3 patients
1 DLT out of 3 patients	Treat next cohort of 3 patients at the same dose
≥2 DLT out of 3 patients	Halt dose escalation: treat total of 6 patients at previous dose to determine MTD
1 DLT out of 6 patients	Escalate dose for next cohort of 3 patients
≥2 DLT out of 6 patients	Halt dose escalation: treat total of 6 patients at previous dose to determine MTD

DLT = dose-limiting toxicity; MTD = maximal tolerated dose (the highest dose for which no more than one of the six treated patients exhibits DLT)

tentative MTD). If none of the three patients is observed to have DLT, the dose level is escalated one step for the next cohort of up to three patients, and the process continues as described. If exactly one of the three patients treated shows DLT, up to three additional patients are treated at the current dose level. If none of these additional three patients shows DLT, the dose level is escalated for the next cohort of up to three patients, and the process continues; otherwise, the prior dose level is defined as the MTD (unless only three patients have been treated at that level, in which case it is the tentative MTD). A tentative MTD becomes final when a total of six patients are treated at that level with fewer than two showing DLT.

Table 2 shows the statistical operating characteristics of this approach. If at least two of three patients treated at a particular dose show DLT, there is 90% confidence that the true probability of DLT at that dose is greater than 20%. (In other words, as shown in Table 2, unless the true probability of DLT at that dose is at least 20%, the probability of at least two out of three patients exhibiting DLT is less than 10%.) On the other hand, if none of three patients shows DLT, there is 90% confidence that the true probability of DLT is less than 55%. (Again, as in Table 2, unless the true probability of DLT is less than 55%, the probability of none of three patients exhibiting DLT is less than 10%.) In the interest of efficiency, either of these situations can be accepted as sufficient to decide whether to halt or continue escalation after treating only three patients at the current level. Allowing for expansion to six patients in case one of the initial three patients shows DLT, the dose

**Table 2. PROBABILITIES OF HALTING OR CONTINUING DOSE ESCALATION FOR VARIOUS PROBABILITIES OF DOSE-LIMITING TOXICITY ASSOCIATED WITH THE DOSE LEVEL, FOR THE STANDARD PHASE I DESIGN**

<b>True Probability of DLT for Dose Level</b>	<b>.05</b>	<b>.1</b>	<b>.2</b>	<b>.3</b>	<b>.4</b>	<b>.5</b>	<b>.6</b>	<b>.7</b>
Probability of halting dose escalation after accruing either 3 or 6 patients (≥2 DLT)*	.03	.09	.29	.51	.69	.83	.92	.97
Probability of continuing escalation after only 3 patients (0 DLT)†	.86	.73	.51	.34	.22	.13	.06	.03
Probability of halting escalation after only three patients (≥2 DLT)†	.01	.03	.10	.22	.35	.50	.65	.78

\*This row gives probabilities of halting dose escalation, at a given dose, if the true probability of DLT for that dose level is as indicated.

†These rows give probabilities of continuing or halting dose escalation after accruing only three patients, at a given dose, if the true probability of DLT for that dose level is as indicated. In all cases, the cohort will be limited to three patients with at least 50% probability, and for the more extreme DLT probabilities (.05 or .7), the cohort will be expanded to six patients with less than 20% probability.

escalation rule gives 91% probability that dose escalation will not halt at doses associated with DLT probability less than 10%, and it gives 92% probability that escalation will not proceed beyond doses associated with DLT probability in excess of 60% (Table 2). The process of approaching the MTD from below, in successive steps, further protects against defining an MTD associated with excessive toxicity. Table 2 plus simulations<sup>15, 23</sup> show that, for a wide variety of dose-toxicity curves, the probability is approximately 85% to 90% that the defined MTD will be associated with DLT probability of approximately 10% to 45%.

The primary criticisms of the standard phase I design are<sup>15, 32, 45, 49</sup>:

1. The design does not target a particular probability of DLT to be associated with the MTD; in practice, the DLT rate associated with the defined MTD will depend somewhat on the DLT rates of the various dose levels.
2. The MTD definition is unnecessarily imprecise because it does not make adequate use of all the available toxicity data.
3. The dose escalation is unnecessarily slow, so that excessive numbers of patients are treated at dose levels less likely to be efficacious.

Storer<sup>49</sup> proposed defining the MTD by fitting all the first-course toxicity data to a logistic dose-toxicity curve (a sigmoidal curve that maps dose levels to associated DLT rates, for example, as in Equation [1]) and letting the MTD be the dose level associated with the targeted DLT rate (usually, 20%–30%), thus addressing the first two criticisms of the standard design. To address the third criticism, he suggested escalating the dose in single-patient cohorts until DLT is observed, at which point dose escalation would revert to the standard design.

### Continual Reassessment Method

Others<sup>15, 32</sup> have proposed using a logistic dose-toxicity model to guide the dose escalation and to define the MTD, using a Bayesian approach<sup>27</sup> to define the model initially according to investigator expectations and updating it with toxicity data obtained during the trial. Goodman et al<sup>15</sup> proposed using the following one-parameter logistic model for DLT probabilities  $P_i$  at dose levels  $x_i$ :

$$p_i = \frac{\exp(3 + \alpha x_i)}{1 + \exp(3 + \alpha x_i)} \quad (1)$$

with parameter  $\alpha$  given at trial start the standard exponential distribution with mean and variance equal to one. To represent the investigators' initial expectations of the dose-toxicity relationship, the doses  $x_i$  are recalibrated so that, when substituted into the model, letting  $\alpha$  be equal to one (its mean according to the initially given exponential distribution), they yield for the  $P_i$ s the investigators' initial expectations for the proba-

bilities of DLT at the doses  $x_i$ . In Bayesian terms, the dose-toxicity model is initially based on the *prior* distribution of  $\alpha$ , the distribution used before the collection of data. Through the variability associated with  $\alpha$ , the model represents the substantial uncertainty of the investigators' initial expectations. An indication of this uncertainty is that the dose associated with DLT rate of 20%, according to the investigators' initial expectations, is given by the model, a 33% probability of actually being associated with a DLT rate in excess of 75% and also a 20% probability of being associated with a DLT rate less than 5%. As each successive patient is treated, the distribution of  $\alpha$  is recalculated according to Bayes' theorem<sup>27</sup> to reflect the new data and the greater certainty associated with the dose-toxicity relationship. Equation (1), with  $\alpha$  having this recalculated *posterior* distribution, eventually reflects the dose-toxicity pattern actually observed, with substantially less uncertainty associated with the predicted DLT rates  $p_i$ .

The continual reassessment method (CRM) involves defining the MTD as the dose associated with the target DLT rate (usually 15%–25%), according to Equation (1), letting  $\alpha$  be the mean of its posterior distribution, thus addressing the first and second criticisms of the standard method. The original method of O'Quigley et al<sup>32</sup> involved treating each successive patient at the successively recalculated MTD. Because this approach required awaiting toxicity results for each successive patient and also could result in excessive toxicity,<sup>23</sup> Goodman et al<sup>15</sup> suggested escalating in cohorts of two to three patients, no more than one dose level at a time and argued that this approach, although conservative, resulted in faster escalation than allowed by the standard method.

### Accelerated Titration Design

Simon et al<sup>45</sup> proposed using a much richer stochastic model than that of Equation 1, to model the toxicity  $y_{ij}$  for dose level  $d_{ij}$ , the  $j^{\text{th}}$  dose of the  $i^{\text{th}}$  patient:

$$y_{ij} = \log(d_{ij} + \alpha D_{ij}) + \beta_i + \epsilon_{ij} \quad (2)$$

where the coefficient  $\alpha$  represents the influence of total prior dose  $D_{ij}$ , and the normally distributed variables  $\beta_i$  and  $\epsilon_{ij}$  represent the effects of interpatient variability and inpatient variability, respectively. (In practice, the continuous variable  $y_{ij}$  is observed as a toxicity grade level [none to mild, moderate, dose-limiting, and unacceptable], and each grade level corresponds to a range of values for  $y_{ij}$ .) Simon et al used this model to analyze historic toxicity data from 20 phase I trials involving nine different agents and to perform extensive phase I trial simulations based on the results of these analyses for a wide variety of designs. Using a model that accommodates the possibility of cumulative toxicity makes it possible to include toxicity data from courses subsequent to the first in both the analyses and the simulations.

Simon et al<sup>45</sup> concluded that it is safe to conduct the initial dose

escalations with single-patient cohorts, using 100% increments, reverting to use of the standard dose-escalation design, using 40% increments, with the first instance of DLT or the second instance of moderate toxicity (for patients' first or subsequent courses). They also concluded that it is safe to allow for inpatient dose escalation so long as the patient exhibits no more than mild toxicity on the current dose. Compared with the standard design, phase I trial simulations combining these two strategies substantially reduced the number of patients required and dramatically reduced the number of patients treated without exhibiting at least moderate toxicity (and thus, potentially, receiving no biologic effect from the agent). Finally, Simon et al<sup>45</sup> demonstrated that using a dose-toxicity model that distinguishes between interpatient and inpatient variability allows the investigators to choose a more appropriate phase II starting dose for an agent for which a subgroup of patients is especially prone to serious toxicity.

### Further Considerations

Several other alternative phase I designs have been proposed for special situations. For drugs with variable pharmacokinetic properties across species, Collins et al<sup>4</sup> suggested that the area under the concentration over time curve (AUC), when measurable, may be a more constant indicator of drug effect than drug dosage. They proposed accelerating the initial dose escalations, using the AUC associated with the MELD<sub>10</sub>, rather than the MELD<sub>10</sub> itself, as the target. For drugs with variable dose effect based on a patient baseline characteristic (initial white blood cell count, in particular), Mick and Ratain<sup>30</sup> suggested using a dose-toxicity model incorporating this additional variable to define both the MTD and the dose-escalation schema. For phase I studies of drug combinations, Korn and Simon<sup>24</sup> point out that there may be a wide variety of combined MTDs, involving different drug proportions. They provide guidance in arriving at a favorable combination from a dose-intensity perspective and in designing the combined dose-escalation schema. Phase I studies in children are generally performed after an adult MTD has been established, and dose escalation begins at 80% of the adult dose to minimize undertreatment.<sup>46</sup> Finally, in the so-called *phase IB study* of a biologic agent the objective is to find the optimal biologic dose (OBD) rather than the MTD. These studies often involve randomly allocating 6 to 10 patients to each of two to four dose levels. The concern is that the biologic endpoint should have at least prognostic clinical relevance and that the variability associated with it be sufficiently small so that a cohort of 6 to 10 patients per dose will give a sufficiently precise indication of biologic effect.<sup>43</sup> Outside the phase IB trial, the MTD cohort is sometimes expanded to approximately 10 patients. In these cases, care should be taken not to overinterpret any responses or, more importantly, any lack of response. As discussed later, phase II trials of efficacy are significantly larger than 10 patients and still give only crude

indications of response rate. Moreover, 10 patients without response is generally insufficient evidence on which to reject the potential efficacy of an agent. Most importantly, the patient population of phase I trials is generally less likely than that of phase II trials to exhibit tumor response.

## PHASE II TRIALS

The objective of a phase II study is to determine whether a new agent or combination regimen has sufficiently promising biologic activity to warrant further, more definitive, clinical testing. Because biologic activity may vary by tumor type, phase II studies are restricted to a particular histology or closely related set of histologies (the uncommon exception being a study of loosely related rare histologies). To maximize the likelihood of seeing biologic activity in the initial phase II studies of an agent, the patient population should be restricted to those with maximum performance status and minimum prior chemotherapy.<sup>43</sup> If an effective standard therapy is available for the patients, it is sometimes medically justifiable to postpone that therapy and treat the patients first with one or two test courses of the experimental agent, using a so-called "window of opportunity" design. After an agent proves its activity in a favorable population, it may undergo further testing in a less favorable population.

### Basic Phase II Designs

In the late 1950s, there were few, if any, effective agents against most forms of cancer; therefore, the primary role of the phase II trial was to screen out clinically ineffective agents as quickly as possible.<sup>14</sup> This goal required a short-term endpoint, indicative of clinical benefit and minimally affected by selection bias (the potential for particularly promising patients to be favored in accrual to the trial). Tumor shrinkage was chosen as the endpoint. Until recently, this endpoint restricted enrollment in phase II trials to patients with bidimensionally measurable lesions. Recently, an international committee of investigators has proposed that phase II response evaluation be based on the longest tumor diameter,<sup>53</sup> opening phase II studies to patients with unidimensional lesions. The paucity of active agents in the late 1950s also suggested use of a statistical design that allocated the minimal number of patients to receive totally ineffective agents. It was determined that a tumor response rate less than 20% was not clinically promising, and Gehan<sup>13</sup> suggested that a run of 14 patients with no response was the minimum number necessary to establish with 95% confidence that the true response rate to the agent did not attain this 20% threshold. (In other words, if the true response rate was at least 20%, there would be at least a 95% likelihood of seeing at least 1 response among 14 patients.) If there was no response among 14 patients, the trial would be terminated

and declared negative. The standard form of Gehan's design also dictates that if at least 1 response is observed among the initial 14 patients, an additional 11 to 16 patients will be treated, to allow estimation of the response rate with a 95% two-sided confidence interval spanning approximately the observed response rate plus or minus 0.1. (With a 95% two-sided confidence interval of response rates, there is a 97.5% confidence that the true response rate does not exceed those in the interval; if it did, there would have been more responses than were observed with 0.975 probability. Likewise, there is a 97.5% confidence that the true response rate does not fall below those in the interval; if it did, there would have been fewer responses than were observed with 0.975 probability.) Estimating the response rate from a phase II trial is discussed in more detail later.

As effective anticancer agents were identified in the 1960s and 1970s, it became apparent that a more comprehensive statistical approach was required for phase II trials. The Gehan design gave little guidance about how to designate an observed response rate as promising or unpromising, nor did it allow for limiting the probability of making an error in such a designation. Fleming<sup>10</sup> proposed a two-stage design that involves prospectively defining the minimal response rate (called  $P_1$ ) that is sufficiently promising so that the investigators would want, with high probability, to recommend further testing of the agent or combination, and, likewise, the maximal response rate ( $P_0$ ) that is sufficiently discouraging so that the investigators would want, with high probability, to recommend no further testing. (In statistical terms, a response rate no more than  $P_0$  is said to satisfy the *null hypothesis* of no treatment benefit, whereas a response rate of at least  $P_1$  is said to satisfy the *alternative hypothesis*.) Furthermore, the design allows for limiting both the *type I error* (the error of calling an agent promising if the response rate is no more than  $P_0$ ) and the *type II error* (the error of calling the agent not promising when the response rate is at least  $P_1$ ). The design requires that the total sample size of the two stages ( $n_1 + n_2$ ) be sufficiently large so that when the investigators designate the minimal number of responses ( $r_2$ ) necessary for declaring the agent worthy of further testing, the study will have the following property: the probability of a false positive (that the number of responses will be at least  $r_2$  when the true response rate is no more than  $P_0$ ) and the probability of a false negative (that the number of responses will be less than  $r_2$  when the true response rate is at least  $P_1$ ) satisfy the desired type I and type II error bounds, respectively. Finally, the design provides for early stopping after approximately half the patients ( $n_1$ ) have been accrued, if the results are dramatically positive or negative. This provision involves designating bounds  $r_1$  and  $a_1$  so that if the number of responses among the initial  $n_1$  patients is at least  $r_1$  or at most  $a_1$ , the trial will be terminated early and declared positive or negative, respectively. The positive and negative bounds  $r_1$  and  $a_1$  are chosen to be sufficiently extreme so that the early-stopping option has minimal

effect on the type I and type II errors which would be obtained from a one-stage trial of  $n_1 + n_2$  patients.

Table 3 gives an example of a Fleming two-stage design to distinguish between response rates of  $P_1 = 20\%$  and  $P_0 = 5\%$ , with type I and type II error rates of 5% and 8%, respectively. The total sample size is  $n_1 + n_2 = 40$ , and the final threshold value for declaring the trial positive is  $r_2 = 5$  responses (12.5%). Interim stopping occurs at  $n_1 = 20$  patients if the number of responses is  $a_1 = 0$  or at least  $r_1 = 4$  (20%). These bounds are sufficiently extreme so that the operating characteristics of the two-stage trial are essentially identical to what they would be without the option of early termination. The cohort of three to six patients in the standard phase I dose escalation design can be viewed as a severely reduced two-stage phase II design in which the endpoint is dose-limiting toxicity rather than tumor response, and the much smaller size results in greatly reduced statistical precision.

Simon<sup>39</sup> improved Fleming's two-stage design by suggesting that early stopping not be allowed for dramatically positive results, in the interest of achieving more precise estimates of response rate by accruing to the full sample size in these cases. He also suggested that because most phase II trials are negative, it is appropriate to choose a design that minimizes the average sample number (ASN) when the response rate is equal to  $P_0$ . Table 4 gives designs for four commonly chosen pairs of  $P_0$  and  $P_1$ , with type I and type II error rates set at 0.1 (the standard choice). In these designs early termination occurs for observed response rates less than  $P_0$ , which occurs with approximately 0.55 to 0.65 probability under the null hypothesis, and the trial is declared positive for observed response rates that attain the halfway point between  $P_0$  and  $P_1$ .

In choosing an appropriate two-stage design, investigators must make two sets of decisions. First, they must define an appropriate  $P_1$  and  $P_0$ . In cases in which there are few, or no, effective therapies,  $P_1$  is generally chosen to be 20%, the conventional lower bound for a promising response rate. When a number of effective therapies are available, however,  $P_1$  may be set at 30% to 40%, or higher. In particular, if the

**Table 3. OPERATING CHARACTERISTICS OF EXAMPLE OF FLEMING TWO-STAGE PHASE II DESIGN\***

	True Response Rate					
	2.5%	5%	10%	15%	20%	25%
Probability of positive outcome	.004	.052	.377	.737	.922	.983
Probability of positive outcome after stage I	.002	.016	.133	.352	.589	.775
Probability of negative outcome after stage I	.603	.358	.122	.039	.012	.003
Probability of positive outcome for one-stage trial	.003	.048	.371	.737	.924	.984

\*Design: Declare the agent promising if at least five responses (12.5%) are observed among the total sample of 40 patients. Stop early, after 20 patients, if there are at least four responses (20% response rate: agent is declared promising) or if there are 0 responses (agent is declared not promising).

**Table 4.** EXAMPLES OF SIMON OPTIMAL DESIGNS ( $\alpha = \beta = 0.1$ )\*

$P_0, P_1$ †	$n_1, n_2$ ‡	$a_1, r_2$ §	ASN ( $P_0$ )	PET ( $P_0$ )
5%, 20%	12, 25	0%, 11%	23.5	.54
10%, 30%	12, 23	8%, 17%	19.8	.65
20%, 40%	17, 20	18%, 30%	26	.55
30%, 50%	22, 24	32%, 39%	29.9	.67

\*The probability of falsely declaring the trial positive ( $\alpha$ -type I error rate), given a true response rate equal to  $P_0$ , and the probability of falsely declaring the trial negative ( $\beta$ -type II error rate), given a true response rate equal to  $P_1$ , are both equal to 0.1.

† $P_0$  and  $P_1$  are, respectively, the maximum response rate that is sufficiently discouraging so that the investigators would want, with high probability, to recommend no further testing of the agent or combination, and, likewise, the minimum response rate that is sufficiently promising so that the investigators would want, with high probability, to recommend further testing.

‡ $n_1$  and  $n_2$  are, respectively, the sample sizes of the first and second stages of the trial.

§ $a_1$  is the upper limit on the observed response rate for terminating the trial after stage 1 and declaring it negative (accepting  $H_0$ ).  $r_2$  is the lower limit on the observed response rate for declaring the trial positive (rejecting  $H_0$ ) after continuing through stage 2.

||Average sample number ( $P_0$ ) and PET ( $P_0$ ) are, respectively, the average sample number and the probability of early termination, given a true response rate equal to  $P_0$ .

phase II trial involves a combination regimen,  $P_1$  should be set 10% to 20% higher than would be attainable with the most active component of proven effectiveness. The choice of  $P_0$  is dictated by the practical need to keep the phase II trial relatively small. In general,  $P_0$  is set equal to  $P_1 - 20\%$  (the exception is setting  $P_0$  to 5% when  $P_1$  is set to 20%). The second set of decisions involves setting the desired type I ( $\alpha$ ) and type II ( $\beta$ ) error bounds. Common practice is to set both  $\alpha$  and  $\beta$  equal to 0.1, because it is generally accepted that in phase II trials, false-negative results (which may terminate development of a useful agent) are at least as serious as false-positive results (which result in wasted time and resources at the phase III level).<sup>39</sup> In testing agents against solid tumors, however, because a large percentage of new agents unfortunately prove ineffective, many investigators prefer to use an  $\alpha$  of 0.05, with a  $\beta$  of either 0.1 or 0.2.<sup>39</sup>

### Estimating the Response Rate from a Phase II Trial

The Fleming and Simon two-stage designs described previously include a precise decision rule relating to whether the results of a trial are sufficient to warrant further testing. It is still important, however, to obtain an appropriate measure of tumor response. Response rate should be measured by dividing the number of responders by the total number of patients who received at least one course of therapy. Excluding patients who suffer early treatment failure inflates the response rate.<sup>16</sup> It is also important to calculate a confidence interval for the response rate, to reflect the fact that the relatively small phase II trial results in a relatively imprecise measure of response rate. Care should be taken to calculate this confidence interval correctly. Many investigators make the mistake

of naively treating the response estimator as an asymptotically normal random variable with a two-sided 95% confidence interval equal to the response rate plus or minus 1.96 times the estimated standard deviation:

$$(k/n) \pm 1.96\sqrt{(k/n)(1 - k/n)/(n - 1)} \tag{3}$$

where  $k$  equals the number of responses among  $n$  patients. Others calculate an exact binomial confidence interval but fail to account for the early-stopping option for dramatically negative results and for the sacrifice of statistical precision that accompanies the opportunity to avoid unnecessarily prolonging a negative trial. An appropriate exact binomial two-sided confidence interval for the result of a two-stage Simon optimal design, accounting for the early-stopping option, is easily calculated on a hand calculator, using the approach of Jennison and Turnbull.<sup>19</sup> (A more complex alternative approach by Duffy and Santner<sup>8</sup> is also frequently used.) Table 5 illustrates the significant inaccuracy, for low estimated response rates, that can result from using either of the other two approaches. For example, if the investigators were to observe three responses among the 37 patients, the highest observed response rate that would still result in a negative trial, they would know from the design of the trial that they were only 90% confident that the true response rate falls below  $P_1 = 20\%$ , because the  $\beta$  (false-negative error rate) of this trial design is 10% for a true response rate of 20%. They would know, therefore, that the normal approximation and the one-stage exact binomial confidence intervals, which ascribe 97.5% confidence to the conclusion that the true response rate falls below either 17% or 22%, respectively, are both misleading.

**Table 5. 95% CONFIDENCE INTERVALS ("NAIVE" VERSUS 1-STAGE VERSUS 2-STAGE) FOR A SIMON OPTIMAL DESIGN ( $n_1 = 12, n_2 = 25, a_1 = 0, r_2 = 4, \alpha = \beta = 0.1$ ): NUMBER OF RESPONSES = 2-6 OUT OF 37**

Observed Response Rate	Two-Sided 95% Confidence Intervals		
	Naive (Asymptotic Normal Approximation)*	1-Stage (Ignoring Early Stopping)†	2-Stage (Accounting for Early Stopping)‡
2/37 = 0.054	(0, 0.128)	(0.007, 0.182)	(0.009, 0.266)
3/37 = 0.081	(0, 0.170)	(0.017, 0.219)	(0.019, 0.272)
4/37 = 0.108	(0.007, 0.209)	(0.030, 0.254)	(0.032, 0.285)
5/37 = 0.135	(0.023, 0.247)	(0.045, 0.288)	(0.047, 0.305)
6/37 = 0.162	(0.042, 0.282)	(0.062, 0.320)	(0.063, 0.330)

\*The naive two-sided 95% confidence interval is obtained by treating the observed response rate as an asymptotically normal random variable. It is symmetric except when truncated at zero.

†The one-stage confidence interval is obtained by treating the observed number of responses as the sum of independent identically distributed binomial random variables, but assuming the sample size is fixed at  $n_1 + n_2$ .

‡The two-stage confidence interval is obtained by treating the observed number of responses as the sum of independent identically distributed binomial random variables, and accounting for the possibility that the trial will terminate at  $n_1$  patients if the observed number of responses is less than  $a_1$ , using the approach of Jennison and Turnbull.<sup>20</sup>

## Phase II Trials Versus Historical Controls and Randomized Phase II Trials

In determining the appropriate  $P_1$  and  $P_0$  for a phase II trial, investigators often use historical data on standard treatments applied to the targeted patient population (often restricted to the involved institutions, to help assure patient similarity). In particular, if the phase II trial involves a combined regimen, historical data concerning the component agents are used. Although this procedure implies a comparison between the trial results and historical data, both of which have inherent variability, the common practice is, unfortunately, to treat the response rate derived from the historical data as if it were a constant. Thall and Simon<sup>51</sup> demonstrate that this practice can lead to a serious underestimation of the true type I error rate, particularly when the historical data are limited, for example, to 40 to 100 patients, from a number of separate studies, each with its own separate underlying response rate (reflecting interstudy and intrastudy variation). Table 6 shows, in fact, that the true  $\alpha$  may be as high as 0.10 to 0.15, when the assumed  $\alpha$  equals 0.05. Clearly, this inflated false-positive rate could lead to a lot of wasted time and resources performing phase III trials doomed to negative results. Furthermore, Thall and Simon<sup>51</sup> suggest that in these cases random allocation of up to 35% of the patients to the control treatment may help compensate for the relatively small number of historical controls. In fact, Table 6 shows that placing all the patients on the experimental treatment may result in as low as a 55% power to detect the targeted 20% improvement in response rate, when an 80% power can be attained with the same number of patients if some of them are randomly assigned to the treatment associated with the historical controls.

An alternate situation commonly arises from the simultaneous availability of several experimental agents for testing in phase II trials. These agents could be tested in separate trials at separate institutions, but if more than one agent seems promising, potential differences in patient selection, response assessment, and dosage modification and compliance could make the results difficult to compare for purposes of prioritizing the agents for further testing.<sup>47</sup> In this situation, Simon et al<sup>47</sup> propose using a randomized phase II study to assign patients, usually from several institutions, randomly to the various experimental treatments. Regardless of the difference in rates, the agent demonstrating the highest promising response rate would be given the highest priority for further testing. Table 7 gives the required sample sizes per arm, for trials of two to four arms, to assure, with 90% probability that an agent which has a true response rate at least 15% greater than its competitors will be chosen. It must be remembered that this design gives no assurance of the magnitude of the true difference in response rates. It does not allow a definitive comparison of the agents but only a tentative ranking for purposes of further testing. For example, if two or more agents have equal underlying response probabilities, the probability is high that one will seem superior by chance. Therefore, this approach is not appropriate

Table 6. INCORPORATING HISTORICAL CONTROLS IN PLANNING PHASE II TRIALS ( $\alpha = 0.05$ ,  $\beta = 0.2$ ,  $P_0 = 0.2$ ,  $P_1 = 0.4$ )\*

90% Probab. Interval for the $\theta_i$ †	Number of Historical Controls†	Sample Size of Experimental Trial†	% Allocated to Control on Experimental Trial†	Power Without Controls‡	$\alpha$ for Naive Approach§
0.15-0.25	40	76-80	23-26	0.72-0.75	0.12-0.13
0.15-0.25	60	65-72	9-18	0.76-0.79	0.10-0.11
0.15-0.25	100	48-61	0-3	0.80	0.09-0.10
0.10-0.30	40	87-92	33-37	0.54-0.62	0.15-0.17
0.10-0.30	60	83-90	30-35	0.58-0.68	0.14-0.16
0.10-0.30	100	78-88	25-34	0.61-0.73	0.12-0.15

\*The historical controls are assumed to consist of a number of separate studies with overall mean response rate  $P_0 = 0.2$ . Individual studies are assumed to have their own separate mean response rates  $\theta_i$  ( $i = 1$  to  $k - 1$ ), which vary about  $P_0$ , falling within either (0.15-0.25) or within (0.10-0.30) with 90% probability. The experimental trial is assumed also to have its own separate mean response rate of  $\theta_k$  for the controls and  $\theta_k + 0.2$  for the treated patients, where  $\theta_k$  is assumed to have the same distribution as the  $\theta_i$ , and thus  $\theta_k + 0.2$  varies about  $P_1 = 0.4$ .

†The total number of historical controls is 40 to 100. The required number on the experimental trial is given for the optimal percentage of controls included in the experimental trial.

‡The power is given for the option of not allocating any experimental patients to a control arm.

§The type I error ( $\alpha$ ) is given for not allocating any experimental patients to a control arm, and, in addition, comparing the response rate of the treated patients to the overall response rate of the historical controls, treating this latter rate as a constant rather than as the random variable it is.

**Table 7. RANDOMIZED PHASE II TRIAL: NUMBER OF PATIENTS PER TREATMENT ARM REQUIRED TO GIVE 90% POWER TO CORRECTLY SELECT\* A TREATMENT YIELDING RESPONSE RATE 15% HIGHER THAN THE HIGHEST OF THE OTHER ARMS**

Superior Response Rate†	Number of Treatments to be Randomized		
	2	3	4
25%	21	31	37
35%	29	44	52
45%	35	52	62
55%	37	55	67
65%	36	54	65

\*In this design, the treatment with the highest response rate is assigned the highest priority for further testing, regardless of how small the difference in response rates is, compared with the other treatments.

†The superior response rate is the response rate associated with the best arm.

for comparing an agent against a control treatment or for comparing a combination regimen against one or more of its components. In these cases, even if there is no advantage to the experimental regimen, it has a 50% probability of seeming at least nominally superior.

### Further Considerations

In some situations tumor response may be an inappropriate endpoint. Recently, increasing attention has been given to cytostatic agents, which serve to reduce or halt tumor growth or metastatic spread. If these agents are to be assessed in a phase II trial, time to progression is most often the best endpoint. Also, in some diseases such as lung cancer, tumor response has not proven to predict for a survival advantage; in other diseases such as brain and prostate cancer, response may be difficult to measure, and few patients may have measurable disease. In these diseases, time to progression or survival may be the best phase II endpoint.<sup>43</sup> Because time to progression and survival have been found to be more affected by patient selection factors than is tumor response, these situations require a careful choice of the historical controls or a careful use of randomization. In other situations, toxicity is of such concern that it should be incorporated into the phase II decision process, along with tumor response.<sup>3, 5, 52</sup> Further discussions concerning both standard and new phase II designs may be found in the articles by Thall and Simon<sup>51</sup> and by Mariani and Marubini.<sup>29</sup>

In some situations a phase II study of a new regimen is of little value and one may proceed directly from the phase I determination of MTD to a randomized phase III comparison. This would be the case if the schedule of administration has been altered in a way that is unlikely to affect efficacy dramatically or if an additional agent or modality of known efficacy has been added. In either case, a one-armed phase II

study is unlikely to provide useful information concerning the added benefit. Certainly, if a known regimen has been altered to make it less toxic, a one-armed phase II study is likely to be useless in demonstrating equivalent efficacy. On the other hand, if a modulator has been added to an agent in hopes of significantly boosting its efficacy, and there is some doubt as to whether the modulator will have any effect whatsoever, it may be useful to conduct a phase II study before the phase III comparison, especially if a biologic endpoint can be identified which would suggest the presence or absence of effective modulation.

## PHASE III TRIALS

The objective of a phase III trial is to determine definitively the clinical efficacy of an experimental agent or regimen, compared with a standard treatment, which may be observation. In rare circumstances the standard treatment may be so clearly inadequate or the experimental treatments may be so promising that the investigators feel compelled to compare the experimental treatments against historical controls. The potential for introducing selection bias (in favor of the experimental regimen or against it), which may or may not be identifiable from a comparison of baseline prognostic variables, suggests that use of randomized controls is almost always necessary. This necessity is especially true in oncology trials, in which the anticipated benefits of the experimental treatment, compared with the standard, are generally small.<sup>16, 43</sup>

In statistical terms, the objective of a phase III trial is to test the null hypothesis (of no treatment difference) against the alternative hypothesis. The alternative hypothesis is often one-sided, restricted to the case where the experimental treatment is superior, because often there is no interest in proving that the experimental treatment is actually worse than the control. Otherwise, the alternative hypothesis is two-sided, and investigators look for a statistically significant treatment difference in either direction. Unfortunately, recently the distinction between one-sided and two-sided alternative hypotheses has been blurred; in a desire to be conservative investigators have come to accept a two-sided 0.05 significance level as the appropriate standard of proof, preferring to give up the one-sided designation, even when appropriate, rather than replace the conventional 0.05 significance level with a 0.025 significance level.

### Phase III Trial Design

The primary endpoint of a phase III trial is generally survival, because this endpoint best defines clinical benefit and is totally objective. Occasionally, if there is concern that use of effective salvage therapy will obscure the benefit of the experimental treatment, other endpoints such

as progression-free survival (defined as time from randomization to death or progression, whichever comes first) may be chosen as primary.<sup>43</sup> Use of progression as the primary endpoint may be misleading; the primary endpoint should always include deaths, which may be treatment-related.<sup>43</sup> The patient population of a phase III trial should be as inclusive as possible to allow maximal generalization of the trial results to the potential beneficiaries.<sup>16, 43</sup>

The first consideration in designing and administering a phase III trial is prevention of bias, which is any systematic design flaw that favors one treatment arm over the others. A valid randomization, in which the treatment arm is not known to the investigator at patient registration, protects against the kind of selection bias that may be introduced by using historical controls or by revealing the treatment arm before the patient is entered on trial. The primary analysis should be an *intention to treat* analysis, including all eligible patients randomized.<sup>16, 43</sup> Excluding early deaths could exclude deaths that are treatment-related. Excluding patient withdrawals or patients with serious protocol violations, on the grounds that they did not receive sufficient therapy to benefit, could bias the results, because these patients could have a prognosis that differs by treatment arm. Indeed, eligibility exclusions should be made only on the basis of information that relates to patient status before randomization and that is available independent of treatment arm. Extreme care should be taken that patients are never excluded on the basis of factors that might have been influenced by treatment or treatment-arm assignment. For example, excluding patients who refuse treatment would introduce bias if patients with poor prognosis refused the more toxic treatment more often than the less toxic one. Finally, loss to follow-up not associated with end of study may differ by treatment arm and by current status of patient and therefore may introduce bias. For the final analysis and all interim analyses, all endpoint data should be brought up-to-date as of a uniform cut-off date. Covert loss to follow-up associated with data not being updated uniformly could also differ by treatment arm and by patient status. For the final analysis, in particular, survival data should be acquired from national death indices for patients lost to follow-up.

A second major consideration in designing a phase III trial is restricting the type I error, which is the probability of observing a statistically significant treatment difference when, in fact, there is none. By convention, type I error in phase III trials is almost always restricted to 0.05.<sup>43</sup> Endpoint comparisons should all be planned prospectively and kept to a minimum, with the total type I error controlled by use of the Bonferroni adjustment (allocating to each comparison an appropriate fraction of the total type I error, usually  $1/k$  where there are  $k$  comparisons). In general, it is best to restrict to a single primary endpoint. Otherwise, use of the Bonferroni adjustment, where the multiple endpoints are likely to be positively correlated, reduces the ability of the trial to detect a potential treatment effect. Ad hoc analyses should be avoided, and subset analyses should be kept to a minimum, with the

results treated as exploratory.<sup>43</sup> More than two treatment arms may be used, but it may be best to restrict the treatment arm comparisons prospectively to those of interest (rather than making all possible comparisons), to minimize the cost of the Bonferroni adjustment for each one. Finally, the interim monitoring plan should be defined prospectively so that the multiple interim analyses may be accounted for in defining the total type I error (because these multiple analyses are correlated in a determinable way, it is not necessary to use the Bonferroni adjustment).

A third major consideration in designing a phase III trial is restricting the type II error, which is the probability of failing to detect a medically significant benefit associated with use of an experimental treatment. By convention, type II error in phase III trials is almost always restricted to 0.1 to 0.2.<sup>43</sup> More precisely, the investigators must first identify the minimal difference in the primary endpoint which would be considered medically significant. If the primary endpoint is survival (or some other time-to-event endpoint, such as disease-free survival), the most efficient comparison is generally by means of the log-rank test (as described later), and this minimal difference is best expressed as a percentage increase in median survival associated with use of the experimental treatment. The log-rank test is statistically optimal (and, in particular, more efficient than comparing estimated survival percentages for the treatment arms at a particular time point) if the death rates (or, more generally, the event rates) of the experimental and control arms, over time, maintain a constant proportion, which is the assumption made in using this test.<sup>33, 36</sup> This constant proportion in the death rates can be shown to be the multiplicative inverse of the ratio of median survival times. It is far better to express the treatment difference in terms of the ratio of median survival times (or, equivalently, in terms of the percentage increase in median survival associated with use of the experimental treatment) than in terms of the difference in survival percentages at a given time. For example, a ratio of two in median survival times gives rise to survival percentages of 90% versus 81% (a 9% difference) at an early time point and gives rise to survival percentages of 50% versus 25% (a 25% difference) later on. Determination of the ratio that corresponds to the minimal clinically significant difference depends on the median survival of the control treatment arm. For a control median survival of only 4 months, the minimal clinically significant difference may be determined to be a 100% increase in median survival associated with the experimental treatment. On the other hand, for a control median survival of 4 years, the minimal clinically significant difference may be determined to be a 25% increase in median survival.

Once the minimal clinically significant difference has been defined and the desired type I error rate bound has been set, the investigators may assure the desired power (which is equivalent to restricting the type II error, because power is 1 minus the type II error), for trials in which survival comparisons are made by the log-rank test, by assuring that a sufficient number of deaths (or, more generally, events) are observed. It has been shown<sup>36</sup> that the power of the log-rank test depends

on the number of observed deaths, not on the number of patients, and that the required number of observed deaths can be calculated from the equation:

$$1/D_c + 1/D_e = (\ln\Delta)^2 / (Z_\alpha + Z_\beta)^2 \quad (4)$$

where  $D_c$  and  $D_e$  represent the numbers of deaths observed on the control and experimental arms, respectively,  $\Delta$  represents the median survival ratio targeted as the minimal medically significant difference, and  $Z_\alpha$  and  $Z_\beta$  represent the standard normal  $(1 - \alpha)$  and  $(1 - \beta)$  quantiles, respectively, corresponding to the specified type I ( $\alpha$ ) and II ( $\beta$ ) error bounds. The requirement of Equation 4 can be very closely approximated by the following simplified requirement, in terms of the total number of deaths  $D = D_c + D_e$ :

$$D = 4(Z_\alpha + Z_\beta)^2 / (\ln\Delta)^2 \quad (5)$$

Table 8 gives the required number of observed deaths, for survival comparisons based on the log-rank test, for a range of deltas (1.25–2), for one-sided and two-sided 0.05 type I error rates, and for 0.8 or 0.9 power. If the investigators wish, instead, to base the survival comparison on a percentage survival at a prespecified time point (for example, at 5 years beyond randomization), other methods for calculating sample size to assure desired power are available.<sup>43</sup>

Two types of phase III trials deserve special mention. In some trials, the experimental treatment is conservative and is expected to be less toxic than the standard treatment. For these *equivalency* trials, the objective is to demonstrate that the experimental treatment has efficacy equivalent, but not necessarily superior, to that of the standard. In other words, the experimental treatment will be preferred over the standard unless the null hypothesis of no treatment difference is rejected in favor of the standard treatment. Therefore, in contrast with the usual scenario, a type II error involves mistakenly discarding a proven treatment (the standard) in favor of an unproven one (the experimental), which has, in fact, lesser efficacy. This is a more serious error than the type I error, which, in this case, involves failing to accept an unproven treatment (the experimental), which has equivalent efficacy and lesser toxicity compared with a proven treatment (the standard). Therefore, equivalency studies should be designed to have small type II error (0.05–0.1) and may have relatively larger (0.1–0.2) type I error.<sup>43</sup> Also, as pointed out by Simon,<sup>44</sup> the investigators should make sure that the experimental treatment retains significantly more efficacy than no treatment at all.

A second special form of phase III trial, the *factorial* study, is designed to answer two therapeutic questions, instead of one, at little additional cost.<sup>40</sup> The design involves simultaneously testing two experimental approaches which are thought to have no interaction. For example, a surgical adjuvant trial may involve randomizing patients to chemotherapy versus none and also randomizing patients in each of these two treatment groups to immunotherapy versus none. The four treatment arms are not compared individually. Instead, the two arms with

**Table 8. NUMBER OF OBSERVED DEATHS REQUIRED\* TO DETECT A TARGETED INCREASE IN MEDIAN SURVIVAL (EXPRESSED AS THE RATIO OF MEDIAN SURVIVAL TIMES, EQUAL TO 1.25-2) FOR  $\alpha = 0.05$  (ONE-SIDED OR TWO-SIDED) AND FOR POWER = 0.8 OR 0.9**

Ratio of Median Survival Times ( $\Delta$ )	Decrease in Death Rate for Experimental Arm	$\alpha = 0.05$ One-Sided		$\alpha = 0.05$ Two-Sided	
		Power = 0.8	Power = 0.9	Power = 0.8	Power = 0.9
2	50%	52-59	72-81	66-74	88-99
1.8	44%	72-78	100-109	91-99	122-133
1.65	39%	99-105	137-146	126-134	168-179
1.5	33%	151-157	209-218	191-199	256-267
1.35	26%	275-281	381-390	349-357	467-478
1.25	20%	497-503	689-698	631-639	845-856

\*The number of required observed deaths (or, more generally, observed events) is given as a range. The lower limit of the range is the solution to equation<sup>†</sup> and is taken by most authors to be the actual requirement<sup>‡</sup>; it is sufficiently accurate so long as most patients on trial are observed until death. As the percentage of patients who are censored before death approaches 100%, however, the actual requirement, which is the solution to equation<sup>†</sup> with  $D_c/D_t$  approaching  $\Delta$ , approaches the upper limit.

chemotherapy are compared with the two without, and the two arms with immunotherapy are compared with the two arms without. Each of these two simultaneous comparisons is done at the 0.025 significance level to maintain an overall type I error of 0.05. The sample size need be only 18% larger than that of a two-armed study addressing only one question. The investigators must be certain, a priori, however, that no treatment interaction exists between the two modalities. A negative interaction, in particular, would seriously reduce the power of the trial to answer either treatment question, and any interaction would make it difficult to assess the individual treatment effects with sufficient precision. It must also be stressed that the factorial study lacks adequate power to test for interaction.<sup>40</sup>

### Phase III Trial Analysis

Some phase III trial endpoints, such as tumor response or survival to a particular time point beyond randomization (for example, 5 years), are binomial variables and can be analyzed as such.<sup>18</sup> The success rate (response or survival) for each treatment can be estimated by  $k/n$ , where  $k$  is the number of successes and  $n$  is the number of patients. In this case, the estimated success rates can be treated as asymptotically normal with standard deviations and confidence intervals calculated as in Equation 3, and statistical comparisons can be made in the usual way.<sup>34</sup>

Most phase III trial endpoints, however, are *survival times*, such as time to death or time to progression, and for many patients the study ends (in statistical terms, *the time is censored*) before the endpoint occurs.<sup>1</sup> For these patients, it is known that the survival time exceeds the time to censorship, but the exact survival time is not known. The statistical challenge is to make appropriate use of this partial information, both in calculating survival time distributions and in comparing survival between treatments.

The Kaplan-Meier product-limit estimator is the standard estimator of the probability of *surviving* the event of interest (which, for concreteness, is taken to be death), for every time point, and was developed to make full and accurate use of the censored survival data.<sup>21</sup> It is a *nonparametric empirical* survival estimator, which means that it does not rely on estimating the parameters of a particular survival distribution form, such as the exponential, nor does it involve any a priori assumptions about the shape of the survival distribution curve, but uses only the data themselves. For each patient, time is measured from point of randomization. The estimator assigns a discrete probability of death to each time point for which a death is observed and to only those time points. For the death time  $t_i$  the actual probability is taken to be  $d_i/n_i$ , where  $d_i$  is the number of deaths observed at time  $t_i$ , and  $n_i$  is the number of patients still at risk (those who have not died or been censored yet) at time  $t_i$ . Likewise, the probability of surviving a small interval about  $t_i$ , which contains no other death times, is taken to be

$(n_i - d_i) / n_i$ . The probability estimator of surviving through time  $t$  is taken to be the product of all the individual survival estimates associated with surviving through the individual death times  $t_i$  which fall below  $t$ :

$$\hat{S}(t) = \prod_{i:t_i \leq t} (n_i - d_i) / n_i \quad (6)$$

As the number of observed death times increases, the estimator  $\hat{S}(t)$  approaches the smooth survival function which accurately gives the probability of survival for each time point  $t$ . Hence the name, *product-limit estimator*. Its variance may be conservatively estimated<sup>33</sup> as

$$\text{Var}[\hat{S}(t)] = \hat{S}^2(t) \sqrt{(1 - \hat{S}(t)) / N(t)} \quad (7)$$

where  $N(t)$  is the number of patients still at risk (those who have not died or been censored) at time  $t$ . (An alternative variance estimator<sup>1, 21</sup> is given by the Greenwood formula.)

The log-rank test, proposed by Mantel,<sup>1, 28, 33</sup> is the standard test used to compare survival distributions in the presence of censored data. It is also nonparametric, in that it makes no assumptions concerning the distributions to be compared, except that the ratio of the death rates remains constant over time. It is based on the following reasoning. At each observed death time  $t_i$  (where, as for the Kaplan-Meier estimator, time is measured from point of randomization, for each patient),  $d_{ie}$  deaths are observed from the experimental group (out of  $d_i$  deaths total for the experimental plus control groups). If the hazard of death is equal for the two groups (the null hypothesis), then in each case,  $d_{ie}$  has expectation  $d_i(n_{ie} / n_i)$ , where  $n_{ie}$  is the number of patients at risk at time  $t_i$  in the experimental group and  $n_i$  is the total number of patients at risk at time  $t_i$ . Therefore, under the null hypothesis, the sum of the observed  $d_{ie}$ 's minus the sum of their respective expectations is asymptotically normal with mean 0. Divided by its estimated standard deviation, this difference is a standard normal variable, under the null hypothesis, which can be used to test the equality of the survival distributions.<sup>1</sup>

Linear regression, a standard statistical tool for simultaneously relating an outcome variable to a set of covariates, has been applied to survival data by means of the Cox model.<sup>6</sup> In standard linear regression, the outcome variable is modeled as the sum of a constant term plus the individual covariates, each multiplied by an associated coefficient. For each data point (or patient), the statistician has measures of the outcome variable and the associated covariates. By finding the best fit for the linear model, he estimates the effect of each covariate on the outcome variable. In the Cox model, it is assumed that the death rate  $\lambda$  is a function of time  $t$  and a set of prognostic covariates  $x_i$  and that it is made up of an *underlying* death rate  $\lambda_0(t)$ , over time, which is increased or decreased multiplicatively as a result of the effects of the prognostic covariates. In other words, it is assumed that the logarithm of the death rate is modeled as the sum of the logarithm of the underlying death rate plus the individual covariates, each multiplied by an associated coefficient:

$$\ln \lambda(t, x_1, \dots, x_n) = \ln \lambda_0(t) + a_1 x_1 + \dots + a_n x_n \quad (8)$$

All the usual linear regression methods are available,<sup>1</sup> by which the death rate can be related to the prognostic covariates, and the methods for fitting the model<sup>8</sup> to the death times and their associated prognostic covariate values, across the patients, are given by Cox.<sup>6</sup>

### Phase III Trial Monitoring

When patients are randomly allocated to a particular treatment on a clinical trial, there is an implicit understanding that neither arm has shown itself inferior to the other. It is imperative that the investigators are diligent in thus protecting the interests of the randomly allocated patients. On the other hand, the scientific integrity of the trial must be protected from the potential inflation of type I error by multiple unplanned interim analyses<sup>12</sup> and from the potential inflation of type II error by premature closure, either intentionally or because of a fall-off of accrual from prematurely disenchanted investigators. It has become required practice in phase III trials, particularly in oncology, that specific interim monitoring guidelines for early trial analysis and stopping be written into the protocol and be enforced uniformly by a data monitoring committee.<sup>42, 43</sup> It has also become required practice on NCI-sponsored phase III trials that these committees be independent of the investigators conducting the trial. This independent monitoring avoids any conflict, real or apparent, with investigators' professional interests, and it insulates the investigators randomizing and treating patients from the possible effects of exposure to potentially misleading early *trends* in the data.

Broadly speaking, two situations call for consideration of early trial stopping: either the experimental treatment has already proven itself superior, or it has become apparent that it will not prove superior. To address the first situation, that of early positive stopping, O'Brien and Fleming<sup>31</sup> propose the use of conservative (high) upper bounds on the log-rank test statistic, restricting early positive stopping to the extreme situations in which the log-rank test statistic exceeds these bounds. The result is that the final analysis, in the event that early stopping does not occur, may proceed almost as if early stopping were not an option. Their approach involves defining in advance the number of interim monitoring analyses and their times (in terms of numbers of deaths observed). The significance level of the final analysis, assuming no early stopping, is adjusted to reflect the amount of type I error associated with the possibility of early stopping under the null hypothesis. Because early stopping is restricted to extreme situations by the O'Brien-Fleming approach, and these situations are very rare under the null hypothesis, the amount of type I error used by allowing for early stopping is relatively small.

To address early positive stopping, DeMets and Lan<sup>7</sup> propose a more flexible method which does not require predefining either the number or times of the interim monitoring analyses. First, the investigators define the rate at which they wish to expend type I error across

potential interim monitoring times, with respect to percent of total number of deaths observed. In other words, they define how they will divide the total probability of false rejection of the null hypothesis (0.05), across the time period of the trial. For example, if they wish to adapt the conservative O'Brien-Fleming philosophy, they would allow only approximately 0.005 probability of early stopping, under the null hypothesis, before seeing 50% of the total number of deaths (no matter how many interim analyses occur), out of the total 0.05 probability allowed for falsely rejecting the null hypothesis, across all analyses, interim plus final. Once this *alpha spending* function has been defined, the investigators are free to conduct interim monitoring analyses at convenient times (such as just before annual or semiannual meetings of the data monitoring committee), so long as they set stopping bounds on the log-rank statistic, at each successive interim analysis, which correspond to an expenditure of type I ( $\alpha$ ) error that is in accordance with their predefined time table.

To address the situation in which the interim data analysis indicates that the experimental treatment will almost certainly not prove superior to the control, even if the trial is continued to its planned number of observed deaths, Lan et al<sup>25</sup> propose *stochastic curtailment*. They propose that a trial be terminated early if, given the current data and assuming that the targeted minimal treatment difference, in fact, pertains, the likelihood of rejecting the null hypothesis, if the trial is continued to the required number of observed deaths, is no more than 20%. This approach is conservative because if such a situation pertains, the estimated treatment difference, based on the current data, would surely not favor the experimental treatment by the targeted minimal amount, if at all. Lan et al<sup>25</sup> show that this approach increases type II error by less than 25% relative to the fixed design. (For example, if the fixed design had 90% power to detect the targeted minimal treatment difference, allowing for stochastic curtailment in the event of early nonfavorable results would still yield at least 87.5% power.) Wieand et al<sup>54</sup> propose a simpler approach to negative early stopping. They propose terminating the trial when half the required deaths are observed if the observed treatment difference, no matter how small, favors the control treatment. They argue that this approach leaves the power of the trial to detect the targeted treatment difference virtually unchanged.

Either the approach of Lan et al<sup>25</sup> or that of Wieand et al<sup>54</sup> for negative trial stopping may be used in conjunction with the O'Brien-Fleming bounds<sup>31</sup> for positive trial stopping. The resulting *asymmetric* stopping bounds are entirely appropriate, because the treatment comparison is usually asymmetric. Investigators are interested in disproving the null hypothesis of no treatment difference only in one direction, that is, in that which corresponds to superiority of the experimental treatment. Generally, investigators do not care whether the experimental treatment is definitely inferior to the control; if it is not superior, it is of no further interest. Two-sided O'Brien-Fleming bounds,<sup>31</sup> which result in early stopping only if one treatment appears dramatically superior, should be

used only when two experimental treatments are being compared and the null hypothesis will be rejected only if there is a significant treatment difference in one direction or the other. Unfortunately, as discussed previously, the growing practice of using a two-sided 0.05 significance level, in the interest of being conservative, even when the inherent comparison is one-sided, sometimes leads to the inappropriate use of symmetric early stopping bounds. This practice could result in continuing allocation of patients to a decidedly inferior experimental treatment, well beyond the point at which its lack of superiority to the control has been demonstrated.

Another approach to interim monitoring, which may be used for either one-sided or two-sided comparisons, is by means of the *repeated confidence intervals* of Jennison and Turnbull.<sup>20</sup> In this approach, for example, two-sided 95% confidence intervals are constructed repeatedly over the interim monitoring times, so that the true value stays within the repeated confidence intervals with 95% probability. One may stop early and reject the null hypothesis if it falls outside the repeated confidence interval. Likewise, in the case of one-sided treatment comparisons, one may stop early and accept the null hypothesis if the targeted minimal treatment difference falls outside the repeated confidence interval.

Further statistical issues arising from the use of early stopping guidelines are discussed in the literature.<sup>9, 11, 42</sup> In particular, estimates of treatment difference may be biased after early termination or even when the trial goes to completion, if early termination was an option. An advantage of O'Brien-Fleming bounds<sup>31</sup> and the negative stopping approach of Wieand et al<sup>54</sup> is that, in either case, if the trial proceeds to completion, estimates of treatment difference are left virtually unchanged.

### Phase III Trial Reporting

There is a significant literature on the failings of statistical reporting, and how these failings relate to weaknesses in trial design and analysis.<sup>35</sup> Conversely, there have been attempts to enforce improvements in trial design and analysis by having the medical journals enforce corresponding guidelines for clinical trial reporting. Simon and Wittes<sup>46</sup> and a recent international consortium,<sup>2</sup> in particular, give such guidelines. The synthesis of these guidelines, and those of others,<sup>35</sup> are summarized here. This summary outlines and underscores important statistical issues of design, analysis, and monitoring. These guidelines have some applicability to phase I and II trials, as well.

#### *Statistical Guidelines for Reporting Clinical Trials*

**Introduction.** Prospectively state defined hypotheses and planned subgroup or covariate analyses.

**Methods.** Describe the planned study population and inclusion and exclusion criteria. Give the primary and secondary outcome measures and the minimal important differences and indicate how the target sample size was projected. Describe the rationale and methods for statistical analyses. Give the prospectively defined stopping rules. Describe the method of randomization. Give the number of eligible patients not entered or not randomly allocated and the reasons. Briefly describe the methods used to ensure that the data are complete and accurate, that all patients entered on study are reported, and that the assessment of major endpoints is reliable. The study should not have an inevaluability rate for major endpoints in excess of 15%.

**Results.** For each randomly allocated group, give the timing of follow-up and the number of patients withdrawn or lost to follow-up. Not more than 15% of eligible patients should be lost to follow-up. State the estimated effect of treatment on primary and secondary outcome measures, including point estimate and confidence interval. Significance tests not relating to prespecified hypotheses must be considered exploratory. Present summary data and appropriate descriptive and inferential statistics in sufficient detail to permit alternative analyses and replication. Describe prognostic variables by treatment group and any attempt to adjust for them. Describe protocol deviations, including the number of randomly allocated patients subsequently found ineligible or not treated as assigned, together with the reasons. State interpretation of study findings, including sources of bias and imprecision.

## References

1. Anderson JR, Crowley JJ, Propert KJ: Interpretation of survival data in clinical trials. *Oncology* 5:104, 1991
2. Begg C, Cho M, Eastwood S, et al: Improving the quality of reporting of randomized controlled trials: The CONSORT Statement. *JAMA* 276:637, 1996
3. Bryant J, Day R: Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 51:1372, 1995
4. Collins JM, Zaharko DS, Dedrick RL, et al: Potential roles for preclinical pharmacology in phase I clinical trials. *Cancer Treat Rep* 70:73, 1986
5. Conaway MR, Petroni GR: Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics* 52:1375, 1996
6. Cox DR: Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 34:187, 1972
7. DeMets DL, Lan KKG: Interim analysis: The alpha spending function approach. *Stat Med* 13:1341, 1994
8. Duffy DE, Santner TJ: Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* 43:81, 1987
9. Emerson SS, Fleming TR: Interim analyses in clinical trials. *Oncology* 4:126, 1990
10. Fleming TR: One sample multiple testing procedure for phase II clinical trials. *Biometrics* 38:143, 1982
11. Fleming TR, DeMets DL: Monitoring of clinical trials: Issues and recommendations. *Control Clin Trials* 14:183, 1993
12. Fleming TR, Green SJ, Harrington DP: Considerations of monitoring and evaluating treatment effects in clinical trials. *Control Clin Trials* 5:55, 1984

13. Gehan EA: The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* 13:346, 1961
14. Gehan EA, Schneiderman MA: Historical and methodological developments in clinical trials at the National Cancer Institute. *Stat Med* 9:871, 1990
15. Goodman SN, Zahurak ML, Piantadosi S: Some practical improvements in the continual reassessment method for phase I studies. *Stat Med* 14:149, 1995
16. Green S, Benedetti J, Crowley J: *Clinical Trials in Oncology*. London, Chapman & Hall, 1997
17. Grieshaber CK, Marsoni S: The relation of preclinical toxicology to findings in early clinical trials. *Cancer Treat Rep* 70:65, 1986
18. Harrington DP, Andersen JW: Common methods of analyzing response data in clinical trials. *Oncology* 4:95, 1990
19. Jennison D, Turnbull BW: Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* 25:49, 1983
20. Jennison C, Turnbull BW: Repeated confidence intervals for group sequential clinical trials. *Control Clin Trials* 5:33, 1984
21. Kaplan G, Meier P: Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* 53:457, 1958
22. Klausner RD: *The Nation's Investment in Cancer Research*. Bethesda, MD, National Cancer Institute, 1999, NIH Publication No. 99-4373
23. Korn EL, Midthune D, Chen TT, et al: A comparison of two phase I trial designs. *Stat Med* 13:1799, 1994
24. Korn EL, Simon R: Using the tolerable-dose diagram in the design of phase I combination chemotherapy trials. *J Clin Oncol* 8:374, 1990
25. Lan KKG, Simon R, Halperin M: Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics-Sequential Analysis* 1:207, 1982
26. Leventhal BG, Wittes RE: *Research Methods in Clinical Oncology*. New York, Raven Press, 1988
27. Lindley DV: Bayesian inference. In Kotz S, Johnson NL (eds): *Encyclopedia of Statistical Sciences*, vol 1. New York, John Wiley & Sons, 1982, p 197
28. Mantel N: Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 50:163, 1966
29. Mariani L, Marubini E: Design and analysis of phase II clinical trials: A review of statistical methods and guidelines for medical researchers. *International Statistical Review* 64:61, 1996
30. Mick R, Ratain MJ: Model-guided determination of maximum tolerated dose in phase I clinical trials: Evidence for increased precision. *J Natl Cancer Inst* 85:217, 1993
31. O'Brien PC, Fleming TR: A multiple testing procedure for clinical trials. *Biometrics* 35:549, 1979
32. O'Quigley J, Pepe M, Fisher L: Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* 46:33, 1990
33. Peto R, Pike MC, Armitage P, et al: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 35:1, 1977
34. Piantadosi S: *Clinical Trials: A Methodologic Perspective*. New York, John Wiley & Sons, 1997
35. Rubinstein LV: Statistical review for medical journals, guidelines for authors. In Armitage P, Colton T (eds): *Encyclopedia of Biostatistics*. Chichester, UK, John Wiley & Sons, 1998, p 4275
36. Rubinstein LV, Gail MH, Santner TJ: Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* 34:469, 1981
37. Rubinstein LV, Ungerleider R: Cooperative cancer trials. In Armitage P, Colton T (eds): *Encyclopedia of Biostatistics*. Chichester, UK, John Wiley & Sons, 1998, p 933
38. Schneiderman MA: Mouse to man: Statistical problems in bringing a drug to clinical trial. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1967, p 855

39. Simon R: Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 10:1, 1989
40. Simon R: Designs for efficient clinical trials. *Oncology* 3:43, 1989
41. Simon R: A decade of progress in statistical methodology for clinical trials. *Stat Med* 10:1789, 1991
42. Simon R: Some practical aspects of the interim monitoring of clinical trials. *Stat Med* 13:1401, 1994
43. Simon R: Clinical trials in cancer. In DeVita VT, Hellman S, Rosenberg SA (eds): *Cancer: Principles and Practice of Oncology*, ed 5. Philadelphia, Lippincott-Raven, 1997, p 513
44. Simon R: Bayesian design and analysis of active control clinical trials. *Biometrics* 55:484, 1999
45. Simon RM, Freidlin B, Rubinstein LV, et al: Accelerated titration designs for phase I clinical trials in oncology. *J Natl Cancer Inst* 89:1138, 1997
46. Simon R, Wittes RE: Methodologic guidelines for reports of clinical trials. *Cancer Treat Rep* 69:1, 1985
47. Simon R, Wittes RE, Ellenberg SS: Randomized phase II clinical trials. *Cancer Treat Rep* 69:1375, 1985
48. Smith M, Bernstein M, Bleyer WA, et al: Conduct of phase I trials in children with cancer. *J Clin Oncol* 16:966, 1998
49. Storer BE: Design and analysis of phase I clinical trials. *Biometrics* 45:925, 1989
50. Thall PF, Simon R: Incorporating historical control data in planning phase II clinical trials. *Stat Med* 9:215, 1990
51. Thall PF, Simon R: Recent developments in the design of phase II clinical trials. In Thall PF (ed): *Recent Advances in Clinical Trial Design and Analysis*. Boston, Kluwer Academic Publishers, 1995, p 49
52. Thall PF, Simon RM, Estey EH: New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J Clin Oncol* 14:296, 1996
53. Therasse P, Arbuck SG, Eisenhauer EA, et al: New guidelines to evaluate the response to treatment in solid tumors (RECIST guidelines). *J Natl Cancer Inst* 92:205, 2000
54. Wieand S, Schroeder G, O'Fallon JR: Stopping when the experimental regimen does not appear to help. *Stat Med* 13:1453, 1994

*Address reprint requests to*

Lawrence V. Rubinstein, PhD  
Biometric Research Branch, National Cancer Institute  
EPN 739  
6120 Executive Boulevard MSC 7434  
Bethesda, MD 20892-7434

e-mail: rubinsteinl@ctep.nci.nih.gov