

Randomized Phase II Clinical Trials¹

Richard Simon,* Robert E. Wittes, and Susan S. Ellenberg²

The sources of variability influencing the results of phase II trials are reviewed. Randomized designs for phase II testing are presented and evaluated. Phase II designs with "standard therapy" control groups are not found to be broadly useful. Designs which randomize among new agents or schedules appear to be of value both scientifically and logistically where patient accrual is adequate. The rationale and advantages of this design are described. The concept of ranking and selection of new agents and schedules is presented as an alternative to testing the null hypothesis of therapeutic equivalence. Sample size calculations demonstrate potential advantages of this approach in appropriate situations. The conduct of pilot or "phase II" studies of combinations is also discussed and randomized designs with early stopping rules are proposed. [Cancer Treat Rep 69:1375-1381, 1985]

Better drugs are needed for the treatment of most human malignancies. Consequently, it is important that the clinical evaluation of new agents be carefully performed. Phase I and II studies present special difficulties, because they involve use of agents whose spectrum of toxicity and likelihood of benefit are poorly defined. It is the purpose of this paper to review problems with phase II studies and to explore the potential role of randomization in such trials.

Objectives of Phase II Trials

There are three basic objectives in treating patients on phase II studies (table 1). The first is to benefit the patients. This must be a primary objective of any therapeutic intervention. The second objective is to screen the experimental agent for antitumor activity in a given type of cancer. Agents which are found to have substantial antitumor activity and an appropriate spectrum of toxicity are generally incorporated into combinations to be evaluated for patient benefit in controlled phase III trials. The third objective of the phase II trial is to extend our knowledge of the toxicology and pharmacology of the agent. It is useful to clearly distinguish Objective 1 from Objective 2. Response rate is an appropriate endpoint for evaluating Objective 2. We generally cannot adequately evaluate the extent to which Objective 1 is achieved in

phase II trials. Response rate is only meaningful to the patient if causing tumor shrinkage means extending survival or improving quality of life. This may or may not be the case. Because an untreated control group is generally not available, one cannot properly evaluate whether the new agent influences survival. Comparing survivals of responders to survivals of nonresponders is not a valid way of demonstrating that there has been an impact of treatment on survival. Such comparisons are biased by the fact that responders must live long enough for a response to be documented. Also, responders may have more favorable prognostic factors than nonresponders, leading to a difference in survival regardless of treatment (1-3). It is even possible that treatment may shorten the survival of nonresponders rather than lengthen that of responders. Figure 1 shows an example in which non-small cell lung cancer responders survive longer than nonresponders, but the composite survival of the entire group is no different from an untreated historical control (4).

TABLE 1.—Objectives of phase II trials

- | |
|--|
| 1. Benefit the patients |
| 2. Screen drug for antitumor activity |
| 3. Extend knowledge of toxicology and pharmacology of drug |

¹Received Nov 5, 1984; revised Apr 4, 1985; accepted Apr 26, 1985.

²Biometric Research Branch (R. Simon and S. S. Ellenberg), Cancer Therapy Evaluation Program (R. E. Wittes), Division of Cancer Treatment, National Cancer Institute, Bethesda, MD.

*Reprint requests to: Richard Simon, PhD, Biometric Research Branch, Cancer Therapy Evaluation Program, Division of Cancer Treatment, National Cancer Institute, Landow Bldg, Rm 4B06, National Institutes of Health, Bethesda, MD 20892.

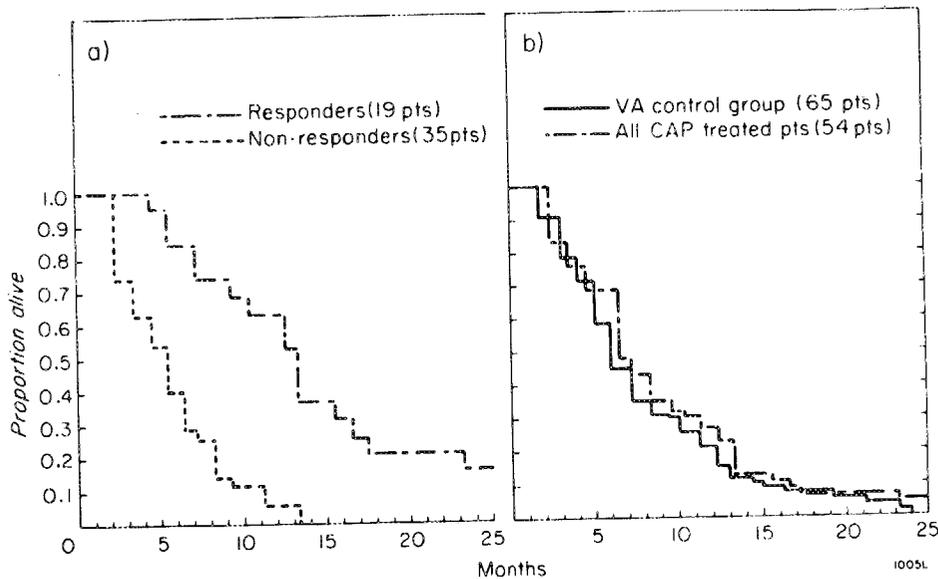


FIGURE 1.—(a) = survival curves (ref 4) for 19 “responding” and 25 “nonresponding” patients with non-small cell lung cancer treated with cyclophosphamide, doxorubicin, and cisplatin; (b) = survival curve derived from the curves in (a) for the whole population of treated patients compared with a historical control group of 65 untreated patients.

Reliability of Phase II Trials

There is frequently great variability in the response rates reported from different phase II studies of the same agent. Some major factors that contribute to this variability are listed in table 2. The first factor is patient selection. Response rates generally decrease as the extent of prior therapy increases. Patients who have failed several prior regimens are more likely to have tumors composed of large numbers of resistant cells, and such patients are also less likely to be able to tolerate full doses of the investigational drug. Whereas it may be appropriate to screen for drugs cytotoxic to tumor cells that are resistant to standard agents, such screening generally cannot be effectively performed in patients with poor bone marrow and other organ system reserves. The steep dose-response relations observed in experimental tumors for some cytotoxic agents suggest that adequate phase II evaluations should be performed at maximally tolerated doses. Probably the most frequent problem with phase II studies is that the patients selected are so debilitated by disease and prior therapy that an adequate evaluation of

antitumor activity is impossible. Such patients are less likely to benefit from the drug, and the trial itself will not contribute meaningfully to evaluation of the drug. Debilitated patients are more likely to die or withdraw early in the course of treatment. Some investigators consider the response of such patients invaluable, and the variable number of such patients contributes to variability in reported response rates.

A second important factor is variability in response criteria among institutions and groups (5). Some reports of phase II trials do not describe or reference the exact criteria used in sufficient detail to adequately interpret the results. A third factor is subjectivity in assessment of response. There have been few formal evaluations of the impact of measurement error on reported rate of tumor response (6-9). In a recent study, Warr et al (9) compared measurements of several physicians on real or simulated malignant lesions, and found that some commonly used criteria of response are subject to large errors. For this reason, they have recommended requiring more extensive and longer lasting evidence of tumor reduction as a basis for the definition of partial response. A fourth source of variability in results includes differences in dosage modification guidelines and differences in protocol compliance. These features reflect differences in aggressiveness of treatment and carefulness in performing clinical trials. They may also reflect differences in patient selection, as protocol compliance may be more difficult in a population of heavily pretreated patients. The variations in policies for handling exclusions in the calculation of response rates and the small sample sizes of

TABLE 2.—Sources of variability in results of phase II trials

1. Patient selection
2. Response criteria
3. Inter-observer variability in response assessment
4. Dosage modification and protocol compliance
5. Reporting procedures
6. Sample size

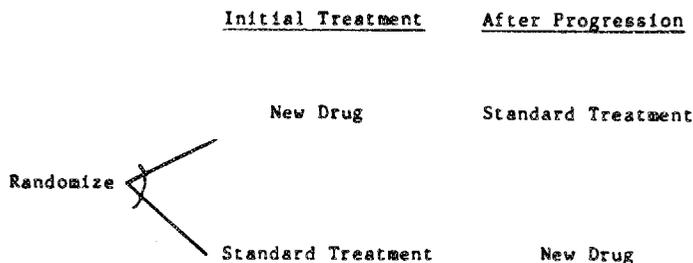


FIGURE 2.—Randomized phase II design of new agent vs positive control treatment.

phase II trials also influence the variability of results (10). Certainly, the outcome of treatment for all patients entered in a trial should be reported regardless of whether the investigator considers some of the tumor responses to be unevaluable.

In the following sections, we shall describe randomized designs for phase II studies and attempt to evaluate whether the problems resulting from the sources of variability listed in table 2 are reduced by such designs.

Randomized Phase II Trials With Control Arms

One type of phase II design which has been employed involves randomization between an investigational agent and an active standard treatment (fig 2). Crossover after progression of disease is shown in the figure. The major objective of the randomization is to help in the interpretation of a poor response rate to the investigational agent. If the new agent has an observed response rate of, say, 25%, then the drug would be identified as having antitumor activity regardless of the magnitude of the response rate to the control treatment. Suppose, however, that the new agent has a low observed response rate (eg, < 10%). In this case, one's conclusion about activity would depend on the response rate observed in the control group. If that response rate is sufficiently large and if the sample size is adequate, then we would conclude that the new agent is inactive for this type of tumor. If the response rates to the new agent and control are both poor, however, then the trial is indeterminate due to inappropriate patient selection. The purpose of randomization here is not to determine whether the new agent is better or worse than the active control.

The design does not seem to offer a broadly useful approach to phase II testing. It is potentially useful in circumstances where an adequate response rate on the active control is not assured, but the identification of an "active control" treatment may be difficult. With the usual phase II sample sizes, it may also be difficult to reliably determine whether the patients are sufficiently responsive to the control treatment for the test of the investigational agent to be meaningful. The expected attrition prior to crossover usually precludes obtaining substantial secondary response rate or cross-resistance infor-

mation. This design may be useful in some studies involving patients with responsive tumor types who have received several previous treatments. Generally, however, a better approach would be to limit eligibility criteria to patients with less prior therapy instead of employing this design.

Randomized Phase II Trials Without Control Arms

The second type of randomized design is shown in figure 3. Two or more treatment arms are possible with this design and the arms are all experimental agents. When such a design is mentioned, investigators are sometimes puzzled at the rationale for conducting a large randomized comparison of two agents which may have no activity in the disease. Such a proposal would indeed be inappropriate, but the fact that randomization is employed does not mean that phase III-type sample sizes are to be employed nor does it mean that comparison, in the usual sense, is the objective. The design can be performed with sample sizes, and early stopping rules, conventionally used for nonrandomized phase II studies.

There are several advantages to the design shown in figure 3. First, randomization helps ensure that patients are centrally registered before treatment starts (11). Such registration is essential for checking patient eligibility, terminating accrual when the target sample size is reached, and establishing a reliable centralized record of treated patients in order to ensure that all such patients are reported upon. Establishment of a reliable mechanism to ensure patient registration prior to treatment is of fundamental importance for all clinical trials. Randomization facilitates this task, since the protocol treatment is unknown until the patient is registered and the randomization performed.

Other advantages of the randomized design shown in figure 3 compared to independent phase II studies are that differences in results obtained for the two agents will more likely represent real differences in toxicity or antitumor effects rather than differences in patient selection, response evaluation, or other factors listed in table 2. With separate studies, the institutions and participating investigators may differ. Even if the eligibility criteria are identical, the protocol priorities may differ.

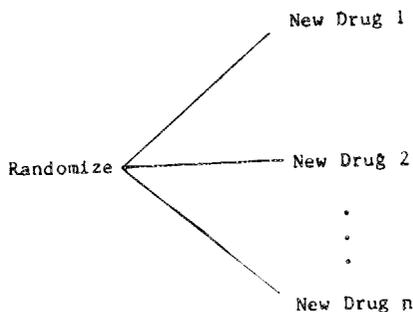


FIGURE 3.—Randomized phase II design of 2 or more new agents.

advantages of the design shown in figure 3 for purposes of ranking and selecting agents and schedules for further study. Whereas traditional procedures for determining sample sizes for nonrandomized phase II trials remain generally applicable (12), there are also circumstances where the direct methods of statistical ranking and selection theory are useful (13). Since these methods have not been described in the oncology literature or applied to cancer clinical trials, they will be presented briefly here. With the statistical selection theory criterion, one always selects for further study the treatment with the greatest response rate (or best value of the other primary endpoint used, eg, lowest incidence of toxicity), regardless how small or "nonsignificant" its advantage over the other treatments seems to be. The sample size is established to ensure that if a treatment is superior to all the others by an amount D , then it will be selected with high probability P . If there are two groups of treatments and the best ones are of equal efficacy with a response rate D greater than those in the other group, then the selection therapy approach ensures that one of the better treatments will be selected with a high probability P . Table 3 shows the required number of patients per treatment with $P = 0.90$ and $D = 15\%$. For example, suppose that three schedules of administration are to be studied and the expected baseline response rate is 20%. With 44 patients per schedule, we have probability 0.90 of selecting the schedule that has a true response rate of $20\% + 15\% = 35\%$. This compares favorably with a sample size of 91 per group for a phase III comparison of two treatments with power 0.90 for detecting differences of 20% with a two-sided 5% significance level (14). Using the selection theory approach, a treatment will still be selected even if the schedules are equivalent or the differences are $< 15\%$, but the probability of correct selection (if there are small true differences) will not be as great as 90%. Conventional statistical designs used in clinical research require much larger numbers of patients, because they select a treatment as superior only when the data are incompatible with the hypothesis that the treatments are equivalent. Selection theory designs always select a treatment as superior, even if they are actually equivalent. With these designs, the probability of selecting a treatment which is actually inferior by amount D (eg, 15%) is limited to be $1-P$; however, the concepts of significance level and power do not have direct analogs in selection theory. More complex selection theory designs involving sequential analysis have also been studied (13).

Although there is substantial statistical literature on methods for determining sample sizes for phase II trials, the issue of patient heterogeneity is rarely discussed. For trials in which both previously treated and non-previously treated patients with the same type of cancer are eligible, the results should generally be analyzed separately for the two groups, since response rates are often so different. For such a study, the number of non-previously

Thus, patients who are more debilitated and less likely to respond may be entered on one protocol. Rates of patient inevaluability also differ substantially among institutions, and the randomized design tends to distribute inevaluable patients evenly among the treatments. Because partial response assessment is somewhat subjective and of limited reliability, there is an advantage to the randomized phase II study in that uniform second-party review of response for both treatments is more readily recognized as being appropriate.

One may ask why comparability of prognostic makeup or response assessment matters if we are not attempting to compare agents. The answer is that a limited form of comparison is often desired. The objective of most phase II studies involves more than just determining whether the agent has any antitumor activity. It also includes estimating the degree of antitumor activity, the extent of tumor shrinkage, the proportion who respond, and the durability of responses. Although conventional sample sizes limit the precision with which degree of activity can be estimated, the estimation is important. Such estimates are regularly used to rank available agents in selecting therapy for patients and in developing plans for introducing the investigational agent with the most promising phase II results into front-line combinations. Even though the sample sizes in randomized phase II studies may not be sufficient to test hypotheses of equality of effect, the inherent comparability of results assures that drugs can be reliably ranked when large differences are obtained. The comparative aspect of phase II trials is even more obvious in cases involving analogs, different schedules of the same agent, different preparations of the same drug, or different agonists and antagonists of a physiologic substance. Some analog studies, in fact, should probably be designed as large-scale phase III trials employing early stopping rules to handle the contingency that an analog could be totally ineffective.

Sample Sizes for Selecting Among Drugs or Schedules

In the previous section, we have tried to indicate the

TABLE 3.—No. of patients/treatment for selection designs*

Smallest response rate (%)	No. of treatments		
	2	3	4
10	21	31	37
20	29	44	52
30	35	52	62
40	37	55	67
50	36	54	65
60	32	49	59
70	26	39	47
80	16	24	29

* Probability of correctly selecting best treatment is 0.90 when it is superior by absolute difference of 15% in response rate.

treated patients should be sufficient for separate analysis, and previously treated patients should probably not be entered on study until results for the former group indicate that the agent has sufficient antitumor activity. There is little potential benefit in administering an agent to previously treated patients which is inactive in non-previously treated patients.

Phase II Studies of Combinations

Though this paper is primarily oriented to evaluating single agents, there are also numerous small "phase II" studies of combinations. For some diseases, there are many active combinations, but it is difficult to know which components contribute to the activity or how to build more effective combinations with tolerable toxicity. Combination studies are clearly oriented towards determining level of activity rather than just presence of activity. Consequently, the problems listed in table 2 are even more severe for pilot studies of combinations. For two-drug combinations, a useful solution, at least for some diseases, would be to conduct randomized pilot phase III trials in which the combination is compared to one of the component single agents as indicated in figure 4. The sample size would be planned as for a usual phase III trial, but accrual would be terminated early if preliminary results indicate that the combination is unlikely to produce a clinically meaningful improvement over the single agent. There are a variety of sequential designs available which effectively accomplish this (15,16). The same approach can be used to assess the contribution of a drug to a multidrug combination (eg, ABC vs AB). Sequential designs are less practical, however, if the endpoint is survival and takes a long time to be observed. Chalmers (17) has emphasized the importance of such randomized pilot studies for avoiding over interpretation of initial results on highly selected patients. Use of this approach, perhaps with a 2 to 1 randomization weighted in favor of the combination, could facilitate the development of effective combinations and reduce the number of patients treated on studies whose results are ambiguous.

The 2 to 1 weighted randomization may be desirable in some trials, as it permits accumulation of greater experience using the new combination with little loss in power compared to an equal randomization (eg, about 13% more patients are required) (18).

DISCUSSION

In this paper, we have attempted to outline problems in phase II evaluations of chemotherapeutic agents and to explore whether randomization is of value in solving these problems.

The randomized design shown in figure 2 may be useful in some circumstances where the appropriateness of the patient population for phase II trials is in question. Although this is often the case, a better solution to the problem would be to utilize patients with less prior therapy. For unresponsive solid tumors, it would also be difficult to identify an active control treatment for use in this design.

The design shown in figure 3 is of much broader usefulness. For settings where patient accrual rate is sufficient, it contributes in several ways to the interpretability of phase II results. The design also facilitates the logistics of conducting phase II trials of common tumors and is used extensively by some cooperative groups. When accrual is adequate, the only substantial disadvantages of this design seem to be the necessity of discussing randomization as part of the informed consent process, and the potential for misinterpreting the results as if they were from a large comparative phase III trial (19). For purposes of ranking and selecting drugs and schedules for further study, however, the randomized design of figure 3 is of value when several drugs or schedules are available. The selection theory approach to planning sample size for such studies is also worthy of consideration.

Phase II trials are problematic in diseases such as prostatic, pancreatic, or brain cancers because of the difficulties in evaluating response. In such circumstances, one may be forced to use survival as the primary endpoint. A new agent could be directly evaluated in a phase III comparison to either a standard treatment or a palliative treatment control; however, an alternative approach is to design a trial to select among several new agents. Surviv-

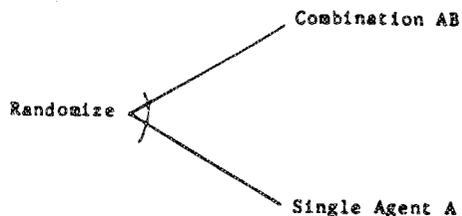


FIGURE 4.—Randomized phase II/III design for evaluating new combinations.

al would be the endpoint, but the trial would require fewer patients per new agent than a phase III trial. The selection trial would then be followed by a phase III comparison of the selected treatment to the standard or palliative control.

The design shown in figure 4 seems particularly useful

for studies of new drug combinations (eg, AB or ABC) in which a single agent (A) or combination (AB) serves as control. Although small pilot studies are required for ensuring tolerability of a new regimen, such designs with appropriate early stopping rules provide more reliable screens for improved combination therapy.

APPENDIX

We present here the formula used for calculation of the numbers in table 3. The treatment selected is that with the largest observed response rate. For selecting among K treatments when the difference in true response rates of the best and next best treatment is D , the probability of correct selection is smallest when there is a single best treatment and the other $K-1$ treatments are of equal but lower efficacy. If we stipulate that the response rate of the worst treatment equals a specified value p , then the probability that the best treatment produces the highest observed response rate is:

$$\sum_{i=0}^n f(i) [1 - B(i; p + D, n)] \quad [1]$$

where $B(; p + D, n)$ is the cumulative distribution function for a binomial distribution with success parameter $p + D$ and n patients and $f(i)$ is the probability that the maximum number of successes observed among the $K-1$ inferior treatment is i . Thus,

$$f(i) = [B(i; p, n)]^{K-1} - [B(i-1; p, n)]^{K-1}$$

If there is a tie among the treatments for the largest observed response rate, we shall assume that one of the tied treatments is randomly selected. Hence, in calculating the probability of correct selection, we add to expression [1] the probability that the best treatment was selected after being tied with 1 or more of the other treatments for the greatest observed response rate. This probability is:

$$\sum_{i=0}^n b(i; p + D, n) \sum_{j=1}^{K-1} g(i, j) / (j + 1) \quad [2]$$

where $g(i, j) = \text{COMB}(K-1, j) [b(i; p, n)]^j [B(i-1; p, n)]^{K-1-j}$, b denotes the binomial probability mass function and $\text{COMB}(K-1, j)$ denotes the number of combinations of $K-1$ objects takes j at a time. The quantity $g(i, j)$ represents the probability that exactly j of the inferior treatments are tied for the largest number of observed responses among the $K-1$ inferior treatments, and this number of responses is i . The factor $1/(j+1)$ in expression [2] is the probability that the tie among the best treatment and the j inferior treatments is randomly broken by selecting the best treatment.

The probability of correct selection is the sum of expressions [1] and [2]. For specified values of p , D , and K , the value of n , number of patients per treatment, is determined to provide a probability of correct selection equal to that desired (P).

REFERENCES

1. WEISS GB, BUNCE H, and HOKANSON JA. Comparing survival of responders and non-responders after treatment: a potential source of confusion in interpreting cancer clinical trials. *Controlled Clin Trials* 4:43-52, 1983.
2. SIMON R, and MAKUCH RW. A non-parametric graphical representation of the relationship between survival and the occurrence of an event: application to responder versus non-responder bias. *Stat Med* 3:35-44, 1984.
3. ANDERSON JR, CAIN KC, and GELBER RD. Analysis of survival by tumor response. *J Clin Oncol* 1:710-719, 1983.
4. TANNOCK I, and MURPHY K. Reflections on medical oncology: an appeal for better clinical trials and improved reporting of their results. *J Clin Oncol* 1:66-70, 1983.
5. DAVIS HL, JR, MULTHAUF P, and KLOTZ J. Comparisons of cooperative group evaluation criteria for multiple-drug therapy for breast cancer. *Cancer Treat Rep* 64:507-517, 1980.
6. GURLAND J, and JOHNSON RO. How reliable are tumor measurements. *JAMA* 194:125-130, 1965.
7. MOERTEL CG, and HANLEY JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer* 38:388-394, 1976.
8. LAVIN PT, and FLOWERDEW G. Studies in variation associated with the measurement of solid tumors. *Cancer* 46:1286-1290, 1980.
9. WARR D, MCKINNEY S, and TANNOCK I. Influence of measurement error on assessment of response to anti-cancer chemotherapy and a proposal for new criteria of tumor response. *J Clin Oncol* 2:1040-1046, 1984.
10. FLEMING TR. One sample multiple testing procedure for phase II clinical trials. *Biometrics* 38:143-151, 1982.
11. HERSON J. Patient registration in a cooperative oncology group. *Controlled Clin Trials* 1:101-110, 1980.
12. HERSON J. Statistical aspects in the design and analysis of phase II clinical trials. *In Cancer Clinical Trials: Methods and Practice* (Buyse ME, Staquet MJ, and Sylvester RJ, eds). Oxford University Press, 1984.
13. GIBBONS JD, OLKIN I, and SOBEL M. *Selecting and Ordering Populations: A New Statistical Methodology*. New York, Wiley, 1977.
14. SIMON RM. Design and conduct of clinical trials. *In Cancer: Principles and Practice of Oncology* (DeVita VT, Jr, Hellman S, and Rosenberg SA, eds). Philadelphia, Lippincott, 1982.
15. LAN KKG, SIMON R, and HALPERIN M. Stochastically curtailed tests in long term clinical trials. *Communications in Statistics - Sequential Analysis* 1:207-214, 1982.
16. DEMETS DL, and WARE JH. Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 69:661-663, 1982.
17. CHALMERS TC. Randomization of the first patient. *Med Clin North Am* 59:1035-1038, 1975.
18. PETO R, PIKE MC, ARMITAGE P, ET AL. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br J Cancer* 34:585-612, 1976.
19. FREIMAN JA, CHALMERS TC, SMITH H, JR, ET AL. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 299:690-694, 1978.