

---

# False Positive Rates of Randomized Phase II Designs

P. Y. Liu, PhD, Michael LeBlanc, PhD,  
and Manisha Desai, MS

*Fred Hutchinson Cancer Research Center, Southwest Oncology Group Statistical Center (P.Y.L., M.L.) and the University of Washington, Department of Biostatistics (M.D.), Seattle, Washington, USA*

---

**ABSTRACT:** The randomized Phase II design for the purpose of selecting a treatment for eventual Phase III testing has recently gained popularity in cancer clinical research. Unfortunately, along with its wider use also come frequent misapplications. The major misuse of the design is the treatment of the Phase II results as ends in themselves without further, definitive evaluation. For binary and censored exponential survival data, we quantify the chance of observing “impressive” between-group differences when underlying distributions are exactly the same in 2-, 3-, and 4-arm selection designs. Depending on one’s view of what is impressive, the “false-positive” rates range from 20% to over 40%. We stress that randomized Phase II results are pilots to Phase III evaluations. One should not regard them as conclusive. We caution especially against the inclusion of control arms in such designs because of the propensity for erroneous inferences. We also discuss the inappropriate practice of performing post-hoc hypothesis testing and presenting  $p$ -values that are less than 0.05. *Control Clin Trials* 1999;20:343–352 © Elsevier Science Inc. 1999

**KEY WORDS:** *Randomized Phase II design, pick-the-winner design, binary data, censored survival data, false positive rate*

## INTRODUCTION

The randomized Phase II selection design was first introduced to cancer clinical trials by Simon et al. [1]. As stated by the authors, the intent of the design was “ranking and selecting agents and schedules for further study.” The execution is exceedingly simple. Patients meeting the same eligibility criteria are randomized to  $K$  experimental treatments. At the conclusion of the trial, the treatment with the best outcome is selected for further testing regardless of the magnitude of its observed advantage over the other treatments. Statistical hypothesis testing is not performed. For binary outcomes such as tumor response, Simon et al. gave the sample sizes required for a 90% correct selection

---

*Address reprint requests to: P. Y. Liu, PhD, Southwest Oncology Group Statistical Center, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP-557, PO Box 19024, Seattle, Washington 98109-1024.*

*Received October 13, 1998; accepted January 4, 1999.*

probability should there be one treatment for which the underlying response rate is superior to all others by an absolute 15% [1]. Liu et al. generalized the design to censored survival data [2].

The attraction of the selection design lies in its simplicity and moderate sample sizes consistent with traditional Phase II studies. Yet the aim of selecting the best treatment for further study when and only when such a treatment exists also results in an inescapable limitation. As intended, the design provides a high probability of taking the superior treatment forward when such a treatment exists. When there is no difference among the treatments in true efficacy, a superior-looking treatment can still appear purely by chance. Unlike the hypothesis-testing methodology, which assesses the likelihood of a false-positive finding, the selection design makes no attempt to distinguish the false positive from the true positive. This task is left to the mandatory follow-on Phase III studies where statistical error rates are properly controlled. Results of selection studies thus do not provide any answers concerning the relative merit between treatments. Owing to the statistical complexity and the large sample size required for a multi-arm ( $>2$ ) Phase III study [3, 4], the selection design serves an important screening function towards a simple, two-arm Phase III study when there are multiple promising experimental regimens. Selection results, however, are mere catalysts; they are not definitive on their own. As Gibbons et al. pointed out, the false-positive or type I error rate of the selection design is 100% [5].

Unfortunately, correct application of the randomized Phase II selection design has proven illusive in our experience. The randomization issues a license to compare and investigators often forget the lack of precision inherent in small samples. "Impressive" observed differences are taken at face value and not followed by Phase III investigations. Control arms are included in the design. Hypothesis tests are conducted and  $p$ -values presented when they are  $< 0.05$ . In short, instead of being regarded as an interim selection step, the randomized Phase II design is construed as an easy substitute for a Phase III comparison without the burden of statistical significance and large sample size. This approach can be more misleading than comparing single-arm Phase II results to historical benchmarks because although investigators commonly recognize that single-arm pilots are only predecessors to Phase III studies, they often view randomized Phase II results as ends in themselves. To caution against such misuse of the methodology, we quantify the high false-positive rates that could result. For the null case where no treatment differences exist, we present tail probabilities of obtaining seemingly impressive differences for binary and exponential survival data. We provide two examples and further discuss such issues as the danger of including a control arm and the confused practice of presenting  $p$ -values in this setting.

## TAIL PROBABILITIES OF THE LARGEST OBSERVED DIFFERENCE

For both binary and censored exponential survival data, one can approximate the distribution of the largest observed treatment difference by that of the range of order statistics from normal random variables. We assume the sample size ( $n$ ) to be equal in all treatment arms throughout this derivation.

For binary data, let  $\hat{p}_{(1)} \leq \dots \leq \hat{p}_{(K)}$  be the ordered response rates of the  $K$  treatments. When the underlying response rates are the same ( $p$ ) for all  $K$  treatments and  $np$  is large, we can approximate the distribution of  $\hat{p}_{(K)} - \hat{p}_{(1)}$  by the range of normal order statistics,  $R = X_{(K)} - X_{(1)}$ , where the  $X$ s are independently and identically distributed normal random variables with mean  $p$  and variance  $p(1 - p)/n$ . The exact distribution function for  $R$  is known as (Kendall & Stuart [6]):

$$P(R > r) = 1 - K \int_{-\infty}^{\infty} (F(x + r) - F(x))^{K-1} dF(x), \quad (1)$$

where  $F(x)$  is the cumulative distribution function of a normal variable with mean  $p$  and variance  $p(1 - p)/n$ .

For selection designs based on censored survival data, Liu et al. [2] proposed fitting the Cox proportional hazards model [7],  $h(t, z) = h_0(t) \exp(\beta'z)$ , where  $t$  is the survival time,  $h(t, z)$  is the hazard function,  $z$  is the  $(K - 1)$ -dimensional covariate vector of treatment group indicators, and  $\beta = (\beta_1, \dots, \beta_{K-1})$  is the vector of log hazard ratio relative to the  $K$ th treatment (the baseline group). We select the treatment with smallest estimated  $\hat{\beta}_i$ ,  $i = 1, \dots, K$ , ( $\hat{\beta}_K \equiv 0$ ) to be the group with the best observed survival. The distribution of the range of  $\hat{\beta}_i$ ,  $i = 1, \dots, K$  cannot be easily specified in general. For exponential data with survival functions,  $S_i(t) = P(T_i > t) = \exp(-\lambda_i t)$ ,  $0 < t < \infty$ ,  $i = 1, \dots, K$ , however, it is known that the distribution of  $\log(\hat{\lambda}_i)$  is approximately normal with mean  $\log(\lambda_i)$  and variance  $1/d_i$  when  $d_i$  is large, where  $d_i$  is the number of deaths in group  $i$ . In the context of the Cox proportional hazards model, the largest observed log hazard ratio is

$$\begin{aligned} \max_i(\hat{\beta}_i) - \min_i(\hat{\beta}_i) &= \max_i[\log(\hat{\lambda}_i/\hat{\lambda}_K)] - \min_i[\log(\hat{\lambda}_i/\hat{\lambda}_K)] \\ &= \max_i[\log(\hat{\lambda}_i)] - \min_i[\log(\hat{\lambda}_i)]. \end{aligned}$$

Therefore, when  $\lambda_1 = \dots = \lambda_K = \lambda$ , the distribution of the largest observed log hazard ratio can also be approximated by the range of normal order statistics with mean  $\log(\lambda)$  and variance  $1/d$ , where  $d$  is the expected number of deaths in each group.

### Tail Probabilities of the Largest Observed Difference for Binary Outcomes

For binary outcomes, we evaluated equation (1) using S-Plus [Mathsoft Inc., 1998] for  $K = 2, 3$ , and  $4$ , and equal response rates in all  $K$  groups, ranging from 10% to 60%. The sample sizes are such that the correct selection probability would be 0.90 if the response rate in one treatment were superior to the others by an absolute 15% [1]. This is the Simon design widely used for selection studies. Figure 1 presents the probabilities for observing an absolute between-treatment difference greater than 10%, 15%, and 20%, when there are no true differences. We also performed simulations using S-Plus (3000 replications). As expected, the results agree well with those from equation (1) except when  $p = 10\%$ . Because the normal approximation to the binomial distribution is less than ideal when  $np$  is small, we present simulation results in Figure 1 for  $p = 10\%$ .

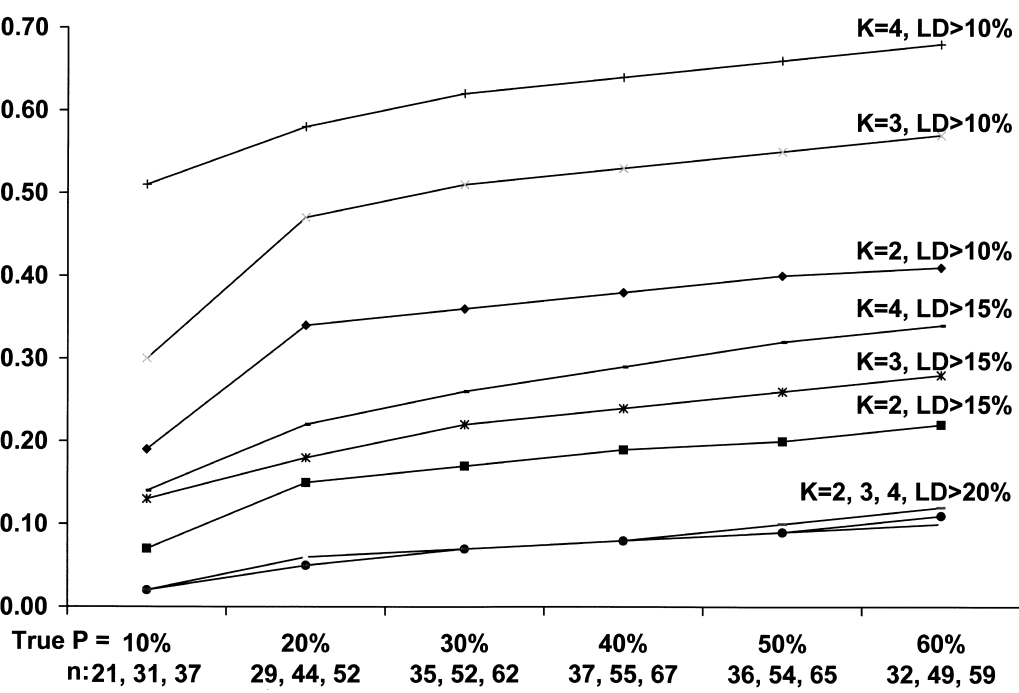


Figure 1 Probability of the largest difference in observed response rates (LD) > 10%, 15%, and 20% when the true response rates (P) are the same for K = 2, 3, and 4 groups. Sample sizes (n) are per arm for K = 2, 3, and 4.

For K = 2, when the underlying response rates are the same and in the 10% to 20% range, the selection design has a 0.20 and 0.35 chance that the observed response rate in one sample is higher than that in the other by an absolute 10% or more. The chance of observing a difference greater than 15% is 0.07 to 0.15. At face value, a difference of 10% to 15% is substantial for rates centered around 10% to 20%. When the underlying response rate lies between 30% and 60%, the chance of obtaining a greater than 10% and 15% difference between the two samples is approximately 0.40 and 0.20, respectively. The chance of observing a difference of at least 20% is close to 0.10 when the underlying response rate is 50% or 60%. Again, without further testing a 15% to 20% difference would appear “very real.”

The false-positive rate increases when the number of arms K increases, despite a corresponding change in the sample size. For example, when K = 4, the chance of observing a 10% or greater difference between the “best” and the “worst” group is 0.50 to 0.70; the same chance for a 15% difference is 0.15 to 0.35. Only the chance for a 20% or greater difference remains unchanged from K = 2 to K = 4.

Tail Probabilities of the Largest Observed Hazard Ratio for Survival Outcomes

For censored exponential survival data, we evaluated equation (1) for K = 2, 3, and 4, median survival = 0.75 for all K groups, and a uniform [0.5, 1.5]

**Table 1** Tail Probability of the Observed Largest Hazard Ratio (OLHR) When There Are No Survival Differences

K	N per group	Probability of OLHR		
		> 1.3	> 1.5	> 1.7
2	40	0.37	0.16	0.07
3	58	0.52	0.21	0.07
4	70	0.63	0.25	0.08

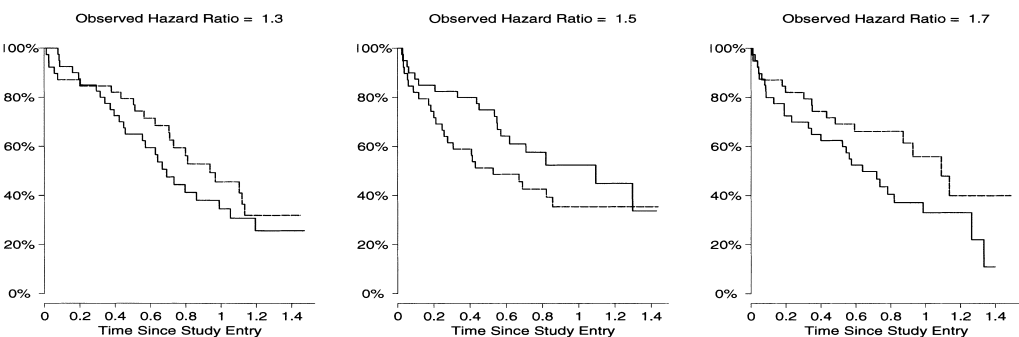
K = number of groups.

censoring distribution, which corresponds to a censoring rate of 41%. The sample sizes are such that the correct selection probability would be 0.90 if there were a best treatment with respect to which the death hazard ratios were 1.5 when other groups were compared to it [2]. Although we evaluated only one survival and censoring combination for each K, the results can be generalized. This is so because the 90% correct selection probability for the alternative case dictates a fixed number of observed deaths for each K, which in turn determines the distribution of the estimated hazard ratios. In addition, we also performed simulations (3000 replications), the results were generally within 1% of those from equation (1).

Table 1 presents the probabilities of observing a hazard ratio greater than 1.3, 1.5, and 1.7 when the true hazard ratio is 1 between all treatments. Hazard ratios of 1.3 or greater often represent significant findings in adequately powered Phase III trials. For example, in a study of 546 stage III ovarian cancer patients, Alberts et al. concluded that, compared to intravenous (IV) cisplatin, the same agent delivered through the intraperitoneal (IP) route significantly improved patient survival [8]. The IV over IP death hazard ratio was 1.32 with a *p*-value of 0.02 in Cox model regression analysis. Investigators found similar survival hazard ratios for tamoxifen and CMF adjuvant therapy in breast cancer patients [9]. Additionally, in the pivotal study leading to the wide adoption of high-dose interferon as the standard treatment for high-risk melanoma, Kirkwood et al. reported observation-versus-interferon hazard ratios of 1.49 (*p* = .01) and 1.64 (*p* = 0.001) for overall survival and relapse-free survival respectively [10]. As Table 1 indicates, in the absence of treatment differences, the chances of observing a hazard ratio greater than 1.3 are 0.37, 0.52, and 0.63 for 2, 3, and 4 samples in a selection design. The chances of seeing a hazard ratio greater than 1.5 are 0.16, 0.21, and 0.25, respectively. There is a 0.07 to 0.08 chance of observing hazard ratios as high as 1.7. Thus, the false-positive rates are very high if one treats randomized Phase II trial results as conclusive. Using the Liu et al. selection design [2], we illustrate in Figure 2 survival curves with observed hazard ratios 1.3, 1.5, and 1.7 generated from the same exponential survival and uniform censoring distributions as used in Table 1, *K* = 2, and *n* = 40 per group.

EXAMPLES

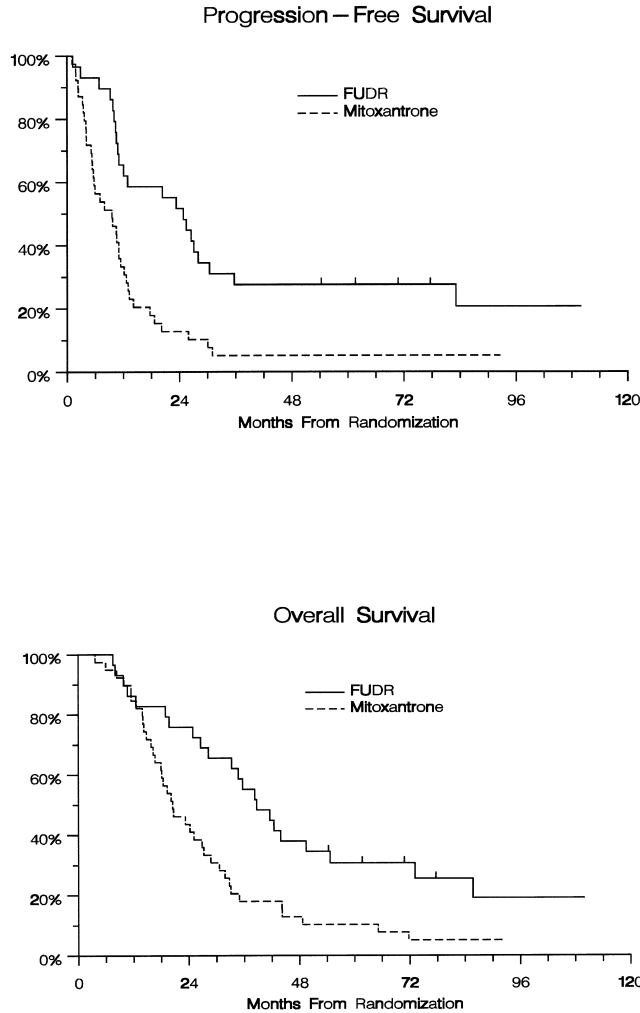
In this section, we give two actual examples of randomized Phase II selection studies. The Cancer and Leukemia Group B conducted a study in adults with



**Figure 2** Survival curves from the same exponential distribution with  $K = 2$ ,  $n = 40$  per group, and observed hazard ratios 1.3, 1.5, and 1.7.

relapsed or refractory acute myeloid leukemia (AML) [11]. The trial randomized patients to the three possible two-drug combinations of diaziquone, etoposide, and mitoxantrone in order to select the combination with the highest complete response (CR) rate for further testing. The sample size target of 52 patients per arm was chosen on the basis of the Simon design and a baseline CR rate of 30%. The design has a 90% correct selection probability if the underlying CR rates are 45% for best treatment and 30% for the other two treatments. The trial enrolled 166 evaluable patients between 1987 and 1990. The observed CR rates were 30% (17/57) for the diaziquone/mitoxantrone combination and 23% (12/53 and 13/56) for the other two. As a result, the Group further investigated the diaziquone and mitoxantrone combination for newly diagnosed AML patients in a single-arm pilot study [12], before launching a Phase III comparison between two post-remission treatment strategies: sequential cycles of high-dose Ara-C versus high-dose Ara-C followed by high-dose cyclophosphamide/etoposide followed by diaziquone/mitoxantrone/G-CSF. The Phase III study is ongoing at the time of this writing.

Between 1989 and 1994, the Southwest Oncology Group performed a selection study (S8835) of intraperitoneal mitoxantrone or intraperitoneal floxuridine (FUDR) for the treatment of minimal residual epithelial ovarian cancer found at second-look laparotomy after initial platinum-based chemotherapy [13]. The trial would select for further testing the agent with the higher 1-year progression-free survival (PFS) rate. The investigators used the Simon design with a sample size target of 37 patients per arm. The correct selection probability is at least 90% should there be an absolute 15% difference in the underlying 1-year PFS rates. In terms of the hazard ratios projected for the disease setting, a 50% versus 65% 1-year PFS rate would correspond to an exponential hazard ratio of 1.6. Thirty-nine and 28 evaluable patients were enrolled to mitoxantrone and FUDR, respectively. The discrepancy in sample size resulted from a higher ineligibility rate on FUDR that was due to violations of prestudy requirements on the second-look surgery. The progression-free survival and overall survival curves are presented in Figure 3. The 1-year progression-free survival rates were 38% for patients on mitoxantrone and 69% for those on FUDR. The hazard ratios comparing mitoxantrone to FUDR were 2.5 and 1.75, respectively, for



**Figure 3** Kaplan–Meier curves for Southwest Oncology Group Study S8835.

progression-free survival and overall survival. The results were virtually identical when all 83 registered patients (including the 16 ineligible patients) were analyzed. Therefore, FUDR was to receive further investigation. When these results were being prepared for publication, however, efficacy data of paclitaxel and intraperitoneal cisplatin therapy for treating ovarian cancer became available [8, 14]. Since then, Phase III evaluations for ovarian cancer treatments have focused on the various schedule, dose, and route of administration combinations of paclitaxel and platinum compounds. The role of FUDR in this disease is still unknown.

**DISCUSSION**

The promise and premise of the selection design is a high chance of correct selection when and only when there is a superior treatment. Its principal



advantage of small sample sizes stems from the hypothesized distance between a potential superior treatment and the others, and its qualitative decision rule—that is, selection of the treatment with the best observed outcome for further testing regardless of how small the difference is over the other treatments. Yet because of this advantage in sample size, results of the selection study itself give little clue as to whether it has successfully identified a superior treatment.

There would be no need for Phase III comparisons if such answers were within the prowess of a Phase II design. Using the properties of normal order statistics and a simple integration formula, we have shown that impressive differences can easily arise from the null case when there are no efficacy differences between the arms tested. The “pick-the-winner” nickname for this design is a misnomer because any also-ran can appear to be a leader in a pack without winners. The truth about “winning” is not unveiled until the selected treatment is evaluated in a Phase III study. Unfortunately, even with the best intentions the Phase III comparison may be indefinitely delayed by unforeseen circumstances, as in the case of Southwest Oncology Group Study S8835. If not followed by Phase III evaluations, the stand-alone randomized Phase II study wastes resources when all treatments are experimental, while it can be extremely harmful if a control arm is included. In the absence of a true treatment difference, all treatments have equal chance of being the “best looking.” The chance that an experimental treatment will appear better than the control is  $(K - 1)/K$ , or 0.50 when  $K = 2$ , 0.67 when  $K = 3$ , and 0.75 when  $K = 4$ . The chance of seeing an impressive difference between an experimental treatment and the control is at least  $2/K$  of the probabilities presented. In this case the control arm’s looking impressively better can be just as detrimental. A new treatment with similar efficacy as the control but less severe side effects can be dismissed as ineffective once and for all. In no circumstances should a control arm be included in a selection design, because the temptation to draw error-prone inferences from it is too great.

As for the practice of performing hypothesis tests and presenting  $p$ -values when they are less than 0.05, the investigators forget the purpose of selection and confuse the issue at hand. If the purpose is selection, then  $p$ -values are irrelevant because the best-appearing treatment will be further tested regardless of the magnitude of its observed advantage over the other treatments. If the purpose is to reach relative efficacy conclusions between the treatments then a Phase III study should be designed with appropriate statistical error rates. Presenting  $p$ -values to lend validity to Phase II comparisons forfeits the nature of selection and all Phase III requirements consequently apply. For instance, the selection sample sizes are typically equivalent to Phase III sample sizes at the earliest interim analysis stage, when  $p$ -values much more stringent than 0.05 are required to “stop the study.” In addition, if more than two treatments are tested, type I error rates need further adjustment to account for the multiple comparisons [3]. Although we have focused our attention to selection designs, the same principles apply to all small randomized trials.

Finally, our results have implications for comparisons between nonrandomized, separate Phase II studies. Given the difference in patient populations and clinical practice, such comparisons are subject to even more variations than randomized Phase II comparisons [15]. Ironically, we believe nonrandomized comparisons to be less dangerous because their limitations have long been



recognized and they do not supplant Phase III comparisons. In conclusion, there are no substitutes for large Phase III trials with properly controlled error rates. It is not our purpose to discuss strategies on how best to gather evidence justifying Phase III trials; suffice to say this is a complex issue with no easy answers. Still, examining carefully the function and limitations of different approaches before choosing one, then staying the course at every step provides the best chance for reducing costly misinformation and lengthy detours.

This research was supported by U.S. National Cancer Institute Grant CA53996.

## REFERENCES

1. Simon R, Wittes RE, Ellenberg SS. Randomized Phase II clinical trials. *Cancer Treat Rep* 1985;69:1375–1381.
2. Liu PY, Dahlberg S, Crowley J. Selection designs for pilot studies based on survival. *Biometrics* 1993;49:391–398.
3. Liu PY, Dahlberg S. Design and analysis of multiarm clinical trials with survival endpoints. *Control Clin Trials* 1995;16:119–130.
4. Green S, Benedetti J, Crowley J. *Clinical Trials in Oncology*. London: Chapman Hall; 1997.
5. Gibbons JD, Okin I, Sobel M. *Selecting and Ordering Populations: A New Statistical Methodology*. New York: Wiley; 1977.
6. Kendall MG, Stuart A. *The Advanced Theory of Statistics Vol 1*. 3rd ed. New York: Hafner; 1969.
7. Cox DR. Regression models and life tables. *J Roy Stat Soc* 1972;34:187–202.
8. Alberts DS, Liu PY, Hannigan EV, et al. Intraperitoneal cisplatin plus intravenous cyclophosphamide versus intravenous cisplatin plus intravenous cyclophosphamide for stage III ovarian cancer. *N Engl J Med* 1996;335:1950–1955.
9. Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy, 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet* 1992;339:1–15, 71–85.
10. Kirkwood JM, Strawderman MH, Ernstoff MS, et al. Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: The Eastern Cooperative Oncology Group Trial EST 1684. *J Clin Oncol* 1996;14:7–17.
11. Lee EJ, George SL, Amrein PC, et al. An evaluation of combinations of diaziquone, etoposide and mitoxantrone in the treatment of adults with relapsed or refractory acute myeloid leukemia: Results of 8722, a randomized phase II study conducted by Cancer and Leukemia Group B. *Leukemia* 1998;12:139–143.
12. Moore JO, Dodge RK, Amrein PC, et al. Granulocyte colony stimulating factor (Filgrastim) accelerates granulocyte and platelet recovery following intensive post-remission chemotherapy for acute myeloid leukemia with aziridinybenzoquinone (AZQ) and mitoxantrone: Cancer and Leukemia Group B study 9022. *Blood* 1997;89:780–788.
13. Miggia FM, Liu PY, Alberts DS, et al. Intraperitoneal mitoxantrone or floxuridine: Effects on time-to-failure and survival in patients with minimal residual ovarian cancer after second-look laparotomy—A randomized phase II study by the Southwest Oncology Group. *Gynecol Oncol* 1996;61:395–402.

14. McGuire WP, Hoskins WJ, Brady MF, et al. Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer. *N Engl J Med* 1996;334:1–6.
15. Flaherty LE, Liu PY, Unger J, Sondak VK. Comparison of patient characteristics and outcome between a single-institution phase II trial and a cooperative-group phase II trial with identical eligibility in metastatic melanoma. *Am J Clin Oncol* 1997;20:600–604.