

**Covariate-adjusted adaptive randomization  
in a multi-stage sarcoma trial**

Peter F. Thall and J. Kyle Wathen

*Department of Biostatistics, Box 447  
The University of Texas, M.D. Anderson Cancer Center  
1515 Holcombe Blvd., Houston, Texas 77030, USA  
E-mail: [rex@mdanderson.org](mailto:rex@mdanderson.org)*

June 3, 2003

## SUMMARY

Medical therapy often requires multiple stages. In such settings, physicians routinely make successive treatment decisions adaptively. In each stage, the decision of whether to continue the patient's therapy, and if so the choice of treatment, may be based on the patient's previous treatments and outcomes, baseline prognostic covariates, as well as data from other patients. Most statistical designs used for clinical trials in such settings waste or distort information by ignoring the multi-stage structure. We present a statistical framework that accounts for patient heterogeneity while making decisions adaptively in clinical trials involving multi-stage therapies. This is illustrated by an ongoing multi-center trial to compare gemcitabine plus docetaxel to gemcitabine alone for unresectable soft tissue sarcoma. The design uses Bayesian covariate-adjusted adaptive randomization based on a categorical variable that characterizes the patient's disease severity after each stage of therapy. A simulation study of the design in the context of the sarcoma trial is presented.

**KEY WORDS:** adaptive design; adaptive randomization; Bayesian design; clinical trial; covariate adjustment; sarcoma

## 1. INTRODUCTION

This paper is motivated by a multi-center, randomized trial of gemcitabine + docetaxel ( $G + D$ ) versus gemcitabine alone ( $G$ ) for treating patients with unresectable soft tissue sarcoma [1]. Each patient receives up to four stages of chemotherapy, each stage lasting six weeks. In multi-stage treatment of solid tumors, the patient's outcome at the end of each stage is scored on an ordinal scale of disease severity. A categorization commonly used for sarcomas is {complete remission ( $CR$ ), partial remission ( $PR$ ), stable disease ( $SD$ ), progressive disease ( $PD$ ), death}. Observation of  $CR$  or  $PR$  may motivate the physician to either repeat the same treatment in the next stage ("consolidation") or terminate therapy,  $SD$  often motivates repeating the treatment, and  $PD$  usually causes the physician either to switch to a different treatment or terminate anti-disease therapy and switch to palliative care. In the sarcoma trial,  $CR$  is defined as disappearance of all lesions,  $PR$  as a decrease in tumor mass between 30% and 99% compared to baseline,  $PD$  as a 20% or greater increase in tumor mass compared to baseline, and  $SD$  as the patient being alive without  $CR$ ,  $PR$ , or  $PD$ . For the purpose of adaptive decision-making, these events are combined to obtain the three-category variable  $Y = R$  (response) if  $CR$  or  $PR$  is observed,  $Y = S$  if  $SD$  is observed, and  $Y = F$  (failure) if  $PD$  is observed or the patient dies. Thus, in each stage,  $Y$  records whether the patient's disease status has improved, stayed the same, or gotten worse, compared to baseline.

Many medical settings involve multiple stages of therapy in which the patient is treated and evaluated repeatedly. After the first stage, it is common practice for the physician to choose the patient's subsequent treatments adaptively, selecting the treatment in each stage only after taking into account the treatments and observed outcomes in previous stages. The set of possible outcomes in each stage and the algorithm for choosing treatments depend on the particular disease and clinical goals. In addition to the methods based on  $CR$ ,  $PR$ ,  $SD$ ,  $PD$  and death in solid tumors, there are many other cases. Physicians often treat a life-threatening infection by consecutively trying different antibiotics until either one of them resolves the infection or the patient dies. The same sort of strategy may be employed with the use of chemical agents for treating clinical depression. In leukemia chemotherapy,

*CR* is defined very differently, usually in terms of the counts of neutrophils, blastic cells, and platelets in the patient's blood or bone marrow. Typically, both *CR* and toxicity are recorded after each course of therapy, with *CR* a binary variable and toxicity graded on an ordinal scale. If, for example, *CR* has not been achieved and toxicity is at most low grade after the first course of a particular combination chemotherapy, the physician may either give another course of the same combination or declare the patient's disease to be "resistant" and switch to a different combination (Thall, Sung and Estey [2,3]). In allogeneic stem cell transplantation, it is standard practice to record whether the transplanted cells have engrafted, and whether the patient has had a disease recurrence, experienced graft-versus-host disease, or died within the first 30 days post transplant and, if the patient survives, record these outcomes again at 100 and 180 days. Different combinations of these events in each stage may motivate a wide variety of therapeutic decisions. Examples of multi-stage adaptive clinical trial designs are given by Thall, Millikan and Sung [4] and Lavori and Dawson [5].

This paper is motivated by the desire to make decisions adaptively during the sarcoma trial while accounting for multi-stage nature of each patient's therapy. The design problem is further complicated by the fact that the observed outcome after each stage is trinary rather than binary, as well as patient heterogeneity. A point routinely ignored in the design and conduct of clinical trials involving multi-stage therapy is that the probability of a given event within a given stage is not the same as the probability of that event over all stages of therapy. For example, consider a very simple version of the sarcoma trial where the respective per-stage probabilities of response, failure and stable disease are 0.20, 0.40 and 0.40, these probabilities do not change over the stages, and treatment is terminated if either a response or a failure is observed in up to four stages of therapy. A trivial probability calculation shows that the overall four-stage outcome probabilities are 0.325 for response and 0.65 for failure. Unfortunately, statisticians often ignore this simple point, e.g. that  $0.325 \neq 0.20$  and  $0.65 \neq 0.40$ . This leads to confusion among physicians, who typically are not trained in probability theory, with a consequent distortion of therapeutic goals. Thus, a major goal of this paper is to address the fundamental issue that a reasonable probability model, accounting for the

multiple treatment stages, is required in the design, conduct and analysis of clinical trials involving multi-stage therapies. Beyond this, in the context of the sarcoma trial, we will present an outcome-adaptive, covariate-adjusted randomization method. To accommodate the fact that patient outcome in each course is trinary, the method incorporates a parameter, elicited from the physicians conducting the trial, that quantifies the relative importance of decreasing  $\Pr(Y = F)$  and increasing  $\Pr(Y = R)$ .

There are numerous statistical methods for making decisions adaptively in clinical trials. These methods use the data from patients previously treated in the trial to make decisions for current and future patients. The decisions may include individual treatment assignments, as is done by dose-finding algorithms or outcome-adaptive randomization (AR), dropping a treatment arm, or terminating the trial due to either futility or superiority of a treatment, as is done by group-sequential procedures [6]. Between-patient and within-patient adaptive decision-making may be combined in the multi-stage setting by using data from other patients, as well as the patient's own data, as a basis for choosing a patient's treatments after the first stage of therapy. There is a substantial literature on outcome-adaptive treatment assignment [7-11]. Most of it is theoretical, however, and until recently there have been few applications. Surveys are given by Rosenberger and Lachin [12] and Rosenberger [13].

Our proposed method uses randomization probabilities biased in favor of the treatment having superior interim results. From the viewpoint of bandit strategies [9,10,14], such a method cannot be optimal with regard to any loss function. Our use of AR, rather than an optimal, non-random allocation strategy, is motivated by several considerations. The first is the desire to reduce bias in the treatment comparison, especially in the form of selection bias on the part of the physicians. While 50:50 (balanced) randomization eliminates bias in the estimation of treatment effect, many physicians are hesitant or unwilling to enter their patients into randomized trials [15]. The clinical trial described here is one of numerous adaptively randomized trials that we have initiated in recent years, including single-institution trials at M.D. Anderson Cancer Center as well as multi-institution trials. We have found that physicians who are hesitant to enter patients into a trial with balanced randomization are

very often much more favorably inclined to participate in a trial with AR. A major practical issue is that, even in the simplest settings, the computational requirements to design and implement an optimal allocation procedure are substantial, and in many cases such methods are impossible to realize. The main difficulty is that optimal decision-theory-based methods require backward induction, or a computationally intensive approximation to optimality [16,17]. Thus, AR provides a practical compromise between optimal allocation and balanced randomization. In Section 6, we will discuss some of literature on covariate-adjusted AR.

## 2. PROBABILITY MODELS

### 2.1 Models for multi-stage therapy

We first provide a general probability framework for trials involving multi-stage therapy. At any given point in the trial, for the  $i^{\text{th}}$  patient, let  $K_i$  denote the number of stages for which data are available, for  $i = 1, \dots, n$ . Let  $\mathbf{Z}_i = (Z_{i,1}, \dots, Z_{i,p})$  be the baseline covariate vector, and denote by  $\mathbf{T}_{i,k} = (T_{i,1}, \dots, T_{i,k})$  the treatments assigned and  $\mathbf{Y}_{i,k} = (Y_{i,1}, \dots, Y_{i,k})$  the outcomes observed in the first  $k$  stages of therapy, for  $k = 1, \dots, K_i$ . Thus, if  $K$  is the maximum number of stages, then each  $K_i \leq K$ . In practice, most multi-stage settings may be accommodated quite well by assuming that each  $Y_{i,k}$  is a categorical variable, with possible values indexed by  $j = 1, \dots, J$ . A patient's therapy may be terminated early due to either an adaptive decision made by the physician or a criterion specified in the trial protocol, or because the patient has died or dropped out of the trial. In general,  $K_i$  may depend on the patient's own data as well data from other patients, depending on the particular within-patient and between-patient decision rules being used.

Denote the model parameter vector by  $\boldsymbol{\theta}$ . To reflect the sequentially adaptive nature of the within-patient medical decision-making, we first define the conditional outcome probabilities

$$\pi_{k,y_k}(\mathbf{Y}_{i,k-1}, \mathbf{T}_{i,k}, \mathbf{Z}_i, \boldsymbol{\theta}) = \Pr(Y_{i,k} = y_k \mid \mathbf{Y}_{i,k-1}, \mathbf{T}_{i,k}, \mathbf{Z}_i, \boldsymbol{\theta}), \quad (1)$$

for  $k \geq 2$ , with  $\pi_{1,y_1}(T_{i,1}, \mathbf{Z}_i, \boldsymbol{\theta}) = \Pr(Y_{i,1} = y_1 \mid T_{i,1}, \mathbf{Z}_i, \boldsymbol{\theta})$ , where each  $y_k \in \{1, \dots, J\}$ . Thus,  $[\mathbf{Y}_{i,k} \mid \mathbf{Y}_{i,k-1}, \mathbf{T}_{i,k}, \mathbf{Z}_i, \boldsymbol{\theta}]$  is multinomial with probabilities  $(\pi_{k,1}, \dots, \pi_{k,J})$  given by (1). The dependence of  $\pi_{k,j}$  on  $\mathbf{Y}_{i,k-1}$ ,  $\mathbf{T}_{i,k}$  and  $\mathbf{Z}_i$  may be accommodated by incorporating

these variables as predictors in a parametric regression model. Denote the unconditional probabilities corresponding to (1) by

$$\xi_{k,y_k}(\mathbf{T}_{i,k}, \mathbf{Z}_i, \boldsymbol{\theta}) = \Pr(Y_{i,k} = y_k \mid \mathbf{T}_{i,k}, \mathbf{Z}_i, \boldsymbol{\theta}). \quad (2)$$

In the first stage,  $\pi_{1,y_1}(T_{i,1}, \mathbf{Z}_i, \boldsymbol{\theta}) = \xi_{1,y_1}(T_{i,1}, \mathbf{Z}_i, \boldsymbol{\theta})$ , while for  $k \geq 2$  the probabilities (1) and (2) are related by the equations

$$\xi_{k,y_k}(\mathbf{T}_{i,k}, \mathbf{Z}_i, \boldsymbol{\theta}) = \sum_{y_1=1}^J \cdots \sum_{y_{k-1}=1}^J \pi_{1,y_1}(T_{i,1}, \mathbf{Z}_i, \boldsymbol{\theta}) \prod_{r=2}^k \pi_{r,y_r}(\mathbf{y}_{r-1}, \mathbf{T}_{i,r}, \mathbf{Z}_i, \boldsymbol{\theta}), \quad (3)$$

denoting  $\mathbf{y}_r = (y_1, \dots, y_r)$ . The likelihood of the data  $\mathbf{X}_n = (\mathbf{Y}_{1,K_1}, \mathbf{Z}_1, \mathbf{T}_1, \dots, \mathbf{Y}_{n,K_n}, \mathbf{Z}_n, \mathbf{T}_n)$  from  $n$  patients is the product multinomial

$$\mathcal{L}(\mathbf{X}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^n \prod_{j=1}^J \left\{ \pi_{1,j}(T_{i,1}, \mathbf{Z}_i, \boldsymbol{\theta}) \right\}^{\delta_{i,1,j}} \prod_{k=2}^{K_i} \left\{ \pi_{k,j}(\mathbf{Y}_{i,k-1}, \mathbf{T}_{i,k}, \mathbf{Z}_i, \boldsymbol{\theta}) \right\}^{\delta_{i,k,j}}, \quad (4)$$

where  $\delta_{i,k,j} = 1$  if  $Y_{i,k} = j$  and 0 otherwise. In general, each patient's sequence of treatments is random and, depending on the particular adaptive treatment assignment algorithm being used in the trial, for  $k \geq 2$ ,  $T_{i,k}$  may depend on  $\mathbf{Y}_{i,k-1}$ ,  $\mathbf{T}_{i,k-1}$  and  $\mathbf{Z}_i$ , as well as the outcomes and covariates of other patients previously treated in the trial.

## 2.2 Model for the sarcoma trial

The three possible outcomes at each stage of therapy in the sarcoma trial are indexed by  $j = R, F, S$ , as defined above in Section 1. Index treatment by  $T = 1$  for gemcitabine + docetaxel and  $T = -1$  for gemcitabine alone. Let  $\mathbf{Z} = (Z_1, Z_2)$  be the vector of binary covariates with  $Z_1 = 1$  if the patient received prior pelvic radiation (PPR),  $-1$  if not, and  $Z_2 = 1$  if the patient's disease is leiomyosarcoma (LMS),  $-1$  if not. The first covariate is important because a patient who has received PPR is given a reduced dose of  $G$ , which in turn motivates inclusion of  $T \times Z_1$  interaction terms in the model. The indicator of LMS also is included because the physicians anticipate that the treatment effects on this particular type of sarcoma may differ from the effects on the other types. We define  $T, Z_1$  and  $Z_2$  symmetrically, with values  $\pm 1$  rather than as a 0/1 variables, to ensure that corresponding probabilities in the eight treatment-covariate subgroups have the same variance under the

Bayesian model, because variability plays a key role in the AR. Since therapy is continued to stage  $k$  only if all elements of  $\mathbf{Y}_{k-1}$  equal  $S$ , and moreover each patient's treatment is chosen only once, at the start of therapy, we simplify notation by writing  $\pi_{k,j}(T_i, \mathbf{Z}_i, \boldsymbol{\theta})$ .

We will assume the following generalized logistic model, using  $j = S$  as the baseline outcome. Temporarily suppress the patient index  $i$ . For stages  $k = 1, 2, 3, 4$  and outcomes  $j = F, R$ , we define the linear terms

$$\eta_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \mu_j + \alpha_j T + \gamma_{k,j} + \sum_{r=1}^2 (\beta_{j,r} + \tau_{j,r} T) Z_r \quad (5)$$

with  $\gamma_{j,1} = 0$  and  $\eta_{k,S} = 0$ . Writing  $\boldsymbol{\gamma}_j = (\gamma_{j,2}, \gamma_{3,j}, \gamma_{4,j})$ ,  $\boldsymbol{\beta}_j = (\beta_{j,1}, \beta_{j,2})$  and  $\boldsymbol{\tau}_j = (\tau_{j,1}, \tau_{j,2})$ , the 18-dimensional parameter vector is  $\boldsymbol{\theta} = (\mu_F, \boldsymbol{\gamma}_F, \alpha_F, \boldsymbol{\beta}_F, \boldsymbol{\tau}_F, \mu_R, \boldsymbol{\gamma}_R, \alpha_R, \boldsymbol{\beta}_R, \boldsymbol{\tau}_R)$ . The conditional outcome probabilities given generally by (1) now take the specific form

$$\pi_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \frac{\exp\{\eta_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta})\}}{1 + \exp\{\eta_{k,F}(T, \mathbf{Z}, \boldsymbol{\theta})\} + \exp\{\eta_{k,R}(T, \mathbf{Z}, \boldsymbol{\theta})\}} \quad (6)$$

for  $j = F, R$ , with  $\pi_{k,S}(T, \mathbf{Z}, \boldsymbol{\theta}) = [1 + \exp\{\eta_{k,F}(T, \mathbf{Z}, \boldsymbol{\theta})\} + \exp\{\eta_{k,R}(T, \mathbf{Z}, \boldsymbol{\theta})\}]^{-1}$ . Thus, for  $j = F$  or  $R$ ,  $\gamma_{k,j}$  is the effect of stage  $k$  versus stage 1,  $\alpha_j$  is the  $G + D$  versus  $G$  (docetaxel, “treatment”) effect,  $\boldsymbol{\beta}_j$  are the covariate main effects, and  $\boldsymbol{\tau}_j$  are the treatment-covariate interactions. Since  $\eta_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \log\{\pi_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta})/\pi_{k,S}(T, \mathbf{Z}, \boldsymbol{\theta})\}$  is the log odds of outcome  $j$  relative to  $S$  for a patient with covariates  $\mathbf{Z}$  treated with  $T$ , it follows that  $\alpha_j + \boldsymbol{\tau}_j \mathbf{Z} = \frac{1}{2}\{\eta_{k,j}(1, \mathbf{Z}, \boldsymbol{\theta}) - \eta_{k,j}(-1, \mathbf{Z}, \boldsymbol{\theta})\}$  is the log odds ratio of outcome  $j$  for  $G+D$  compared to  $G$  alone for a patient with covariates  $\mathbf{Z}$ . As before,  $\xi_{1,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \pi_{1,j}(T, \mathbf{Z}, \boldsymbol{\theta})$ , but now

$$\xi_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \pi_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) \prod_{r=1}^{k-1} \pi_{r,S}(T, \mathbf{Z}, \boldsymbol{\theta}) \quad (7)$$

for  $(k, j) = (2, R), (2, F), (3, R), (3, F), (4, R), (4, F)$  or  $(4, S)$ .

Re-introducing the patient index  $i$ , since  $\delta_{i,k-1,S} \neq 1$  implies that  $\delta_{i,k,j} = 0$  for any  $k > 1$ , the likelihood (4) may now be written in the form

$$\mathcal{L}(\mathbf{X}_n | \boldsymbol{\theta}) = \prod_{i=1}^n \left[ \prod_{k=1}^3 \prod_{j=F,R} \left\{ \xi_{k,j}(T_i, \mathbf{Z}_i, \boldsymbol{\theta}) \right\}^{\delta_{i,k,j}} \right] \left[ \prod_{j=F,S,R} \left\{ \xi_{4,j}(T_i, \mathbf{Z}_i, \boldsymbol{\theta}) \right\}^{\delta_{i,4,j}} \right]. \quad (8)$$

The possible likelihood contributions are summarized in Table I.

### 2.3 Calibrating Priors

To complete the Bayesian model formulation, we assumed *a priori* that all parameters are Gaussian. The real-valued parameter means were calibrated to correspond to mean response and failure probabilities, within each patient subgroup, elicited from the physicians. The variances of the real-valued parameters were determined, within each subgroup, by first eliciting probabilities of the form  $\Pr(\pi_R > \pi_R^*)$  for two or three fixed values of  $\pi_R^*$ , and similarly for  $\Pr(\pi_F > \pi_F^*)$ . A range of values of parameter variances were then considered, calibrated to be larger than the values that agreed with these prior probabilities on the  $\pi_R$  and  $\pi_F$  domains. As formalized by Thall, Sung and Estey [2,3], in practical Bayesian clinical trial design, historical data or elicited information may be used to construct informative priors on non-treatment parameters, such as the covariate and stage effects in the present setting. However, to ensure that the trial design is both ethical and widely acceptable, we feel quite strongly that no undue information on treatment effects should be introduced into the prior before the trial is begun. In the sarcoma trial, this means that  $E(\alpha_R) = E(\alpha_F) = 0$ , and moreover that the variances must be calibrated to ensure that neither very large nor very small values of the treatment effects are highly likely *a priori*. In our experience, physicians' prior opinions regarding treatment effects may be quite strong, and they often provide a poor reflection of any actual historical data that may exist. Consequently, priors carefully and honestly elicited from physicians are not unlikely to lead to a design with such poor properties that it simply should not be run. For the sarcoma trial, after preliminary sensitivity analyses, the prior variances that we used were  $\sigma_\mu^2 = \sigma_\alpha^2 = \sigma_\tau^2 = 1.0$  and  $\sigma_\beta^2 = \sigma_\gamma^2 = 0.50$ . Once the trial is completed, a range of priors may be used for analysis of the final data, and these priors may be very different from that used to conduct the trial.

## 3. ADAPTIVE RANDOMIZATION

We will use an AR procedure that generalizes a method originally proposed by Thompson [18]. Thompson addressed the problem in the context of two independent binomial samples from a randomized trial comparing treatments  $A$  and  $B$ . Assuming that the success probabilities

$\theta_A$  and  $\theta_B$  follow independent beta priors, Thompson proposed that, given the current data, the next patient should be randomized to treatment  $A$  with probability  $\rho_A(data) = \Pr(\theta_B < \theta_A \mid data)$  and to  $B$  with probability  $\rho_B(data) = 1 - \rho_A(data)$ . A strong practical motivation for our using a generalization of this approach is given by Berry and Eick [19], who provide numerical comparisons of four treatment assignment methods in the special case of binary outcomes. They study Thompson’s method, a procedure proposed by Bather [20], a play-the-winner strategy proposed by Robins [21] and later recommended by Zelen [22], and a robust Bayes method that maximizes the expected number of successes in the trial and over a horizon of future patients, with the optimization implemented via backward induction. Berry and Eick’s numerical results show that, for a trial of 100 patients with a horizon of either 100 or 1000 future patients, the Thompson method is nearly identical to the robust Bayes procedure with respect to total number of successes. The sarcoma trial has 120 patients, and the horizon of future patients who likely would be treated with the better treatment from this trial is very likely to be within this range. Moreover, backward induction is extremely computationally intensive and non-trivial to implement in most settings with any degree of complexity, including the sarcoma trial. Thus, we chose to generalize Thompson’s method to accommodate the present setting.

Compared to the two-sample binomial case, the sarcoma trial has the three complications that patients are heterogeneous, each patient’s overall outcome is a vector of random length between one and four, and the outcome in each stage is trinary rather than binary. Both patient heterogeneity and the multi-stage structure are accounted for explicitly in the generalized logistic model by the term  $\gamma_{j,k} + (\beta_{j,1} + \tau_{j,1}T)Z_1 + (\beta_{j,2} + \tau_{j,2}T)Z_2$  in the linear component  $\eta_{j,k}(T, \mathbf{Z}, \boldsymbol{\theta})$ . For a patient with covariates  $\mathbf{Z}$ , we will use as AR criteria the probabilities  $\xi_{4,R}(T, \mathbf{Z}, \boldsymbol{\theta})$  and  $\xi_{4,F}(T, \mathbf{Z}, \boldsymbol{\theta})$  of each outcome over all four stages of therapy. In the sarcoma trial, for  $j = R$  or  $F$  these take the form

$$\xi_{4,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \pi_{1,j}(T, \mathbf{Z}, \boldsymbol{\theta}) + \sum_{k=2}^4 \pi_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) \prod_{r=1}^{k-1} \pi_{r,S}(T, \mathbf{Z}, \boldsymbol{\theta}). \quad (9)$$

For each  $j$ , the probabilities  $\xi_{4,j}(1, \mathbf{Z}, \boldsymbol{\theta})$  and  $\xi_{4,j}(-1, \mathbf{Z}, \boldsymbol{\theta})$  may be regarded as the covariate-

adjusted, multi-course generalizations of the simple binomial probabilities  $\theta_A$  and  $\theta_B$  in Thompson's regime. As noted earlier, in typical practice, the distinction between the conditional probabilities  $\pi_{1,j}$ ,  $\pi_{2,j}$ ,  $\pi_{3,j}$ ,  $\pi_{4,j}$  and  $\xi_{4,j}$ , and the elementary probability calculus given by (9), often are ignored.

Given this structure, a dimension reduction problem still remains, since the two probabilities  $\{\xi_{4,R}(T, \mathbf{Z}, \boldsymbol{\theta}), \xi_{4,F}(T, \mathbf{Z}, \boldsymbol{\theta})\}$  must be expressed as a single criterion for each  $T$  in order to compare treatments. To address this problem, we set  $\lambda_S = 0$  and  $\lambda_R = 1.0$  and elicited the value  $\lambda_F$ , from a group of three physicians who were the opinion leaders in planning the trial, that quantifies the relative importance of decreasing  $\xi_{4,F}$  compared to increasing  $\xi_{4,R}$ . The value  $\lambda_F = 1.3$  was elicited, and there was a strong consensus among the physicians. In turn,  $\lambda_F$  determines the weights  $\omega_F = 1.3/(1.0+1.3) = 0.565$  and  $\omega_R = 1 - \omega_F = 1.0/(1.0+1.3) = 0.435$ . The pair  $(\omega_R, \omega_F)$  may be used to obtain a weighted linear combination of  $\xi_{4,R}(T, \mathbf{Z}, \boldsymbol{\theta})$  and  $1 - \xi_{4,F}(T, \mathbf{Z}, \boldsymbol{\theta})$ , which may be used in turn to compute the posterior probability that is the basis for the AR. Specifically, we define

$$\zeta(T, \mathbf{Z}, \boldsymbol{\theta}) = \omega_R \xi_{4,R}(T, \mathbf{Z}, \boldsymbol{\theta}) + \omega_F \{1 - \xi_{4,F}(T, \mathbf{Z}, \boldsymbol{\theta})\}, \quad (10)$$

and we use  $\zeta(1, \mathbf{Z}, \boldsymbol{\theta})$  and  $\zeta(-1, \mathbf{Z}, \boldsymbol{\theta})$  as the AR criteria, as follows. Given data  $\mathbf{X}_n$ , a patient with covariates  $\mathbf{Z}$  may be randomized to  $G + D$  with probability

$$\nu(\mathbf{Z}, \mathbf{X}_n) = Pr\{\zeta(1, \mathbf{Z}, \boldsymbol{\theta}) > \zeta(-1, \mathbf{Z}, \boldsymbol{\theta}) \mid \mathbf{X}_n\} \quad (11)$$

and to  $G$  with probability  $1 - \nu(\mathbf{Z}, \mathbf{X}_n)$ . This AR criterion is formally equivalent to

$$Pr\{\xi_{4,R}(1, \mathbf{Z}, \boldsymbol{\theta}) - \lambda_F \xi_{4,F}(1, \mathbf{Z}, \boldsymbol{\theta}) > \xi_{4,R}(-1, \mathbf{Z}, \boldsymbol{\theta}) - \lambda_F \xi_{4,F}(-1, \mathbf{Z}, \boldsymbol{\theta}) \mid \mathbf{X}_n\}. \quad (12)$$

Since the posterior probability (11), (12) may be somewhat variable early in the trial, a simple modification that improves the stability and the overall behavior of the AR procedure is to replace it with the corresponding stabilized version

$$\nu^*(\mathbf{Z}, \mathbf{X}_n) = \frac{\{\nu(\mathbf{Z}, \mathbf{X}_n)\}^{1/2}}{\{1 - \nu(\mathbf{Z}, \mathbf{X}_n)\}^{1/2} + \{\nu(\mathbf{Z}, \mathbf{X}_n)\}^{1/2}} \quad (13)$$

as the AR probability for  $G + D$ , with  $1 - \nu^*(\mathbf{Z}, \mathbf{X}_n)$  the probability for  $G$ . The criterion (11) is similar to that used by Thall, Inoue and Martin [23] to adaptively randomize patients with advanced hematologic malignancies among five different donor lymphocyte infusion times. Denoting the probabilities of treatment success with the five infusion times by  $\xi_1, \dots, \xi_5$ , the AR probabilities are  $\nu_j(data) = \Pr(\xi_j = \max\{\xi_1, \dots, \xi_5\} \mid data)$  for  $j = 1, \dots, 5$ .

## 4. SIMULATION STUDY

### 4.1 Trial Design

Based on an historical accrual rate of 15 patients per month, the sarcoma trial design specified a maximum sample size of 120 patients, with 7 additional months of follow up, for a maximum trial duration of roughly 15 months. To ensure that the AR criterion was reasonably informative, the trial was begun by first randomizing 30 patients fairly to the two treatment arms, restricted so that there were 15 patients per arm. Thereafter, the AR criterion (13) was used. Recall that each patient is evaluated in up to four 6-week intervals, with the patient scored as  $R$  or  $F$  if this event is observed at any evaluation. If at any time during the trial either  $\nu(\mathbf{Z}, \mathbf{X}_n) < 0.01$  or  $> 0.99$ , then the trial is terminated within the subgroup of patients having covariate vector  $\mathbf{Z}$  and the superior treatment is selected for that subgroup.

### 4.2 Simulation Design

The simulation scenarios in Table II were formulated in terms of fixed stage 1 outcome probabilities  $(p_R, p_F)$ . The value of the AR criterion  $\zeta = 0.435 \xi_{4,R} + 0.565 (1 - \xi_{4,F})$  corresponding to  $(p_R, p_F)$  for each combination of prognostic subgroup and treatment group under each scenario also is tabled. In the null case (Scenario 1), where  $(p_R, p_F)$  are identical for  $G$  and  $G + D$  within each prognostic group,  $(p_R, p_F) = (.27, .23)$  in the subgroup of (No PPR, LMS) patients, for which  $\mathbf{Z} = (-1, 1)$ . In each scenario, to account for effects of the two covariates and the successive stages of therapy in the simulated trials, we determined fixed values of the covariate and stage effect parameters by asking the physicians to specify changes in  $(p_R, p_F)$  compared to the stage 1 values in the baseline prognostic subgroup. As shown in Table II,

under Scenario 1,  $(p_R, p_F)$  changed the null vector to  $(.22, .28)$  for  $\mathbf{Z} = (1, 1)$ , to  $(.17, .32)$  for  $\mathbf{Z} = (1, -1)$ , and to  $(.21, .27)$  for  $\mathbf{Z} = (-1, -1)$ . This information provided the fixed values  $\beta_F$  and  $\beta_R$  in each scenario. For the stage effects, the physicians hypothesized no stage 2 effect, but they changed the null vector for (No PPR, LMS) patients to  $(p_R, p_F) = (.05, .23)$  for stage 3 and to  $(p_R, p_F) = (.05, .50)$  for stage 4. This yielded the fixed stage effect parameters  $\gamma_R$  and  $\gamma_F$ . These fixed parameter values were used throughout to quantify the stage effects in the simulated patients, with patients generated from the four prognostic subgroups according to the historical proportions noted above. As shown by the values of  $\zeta$  in Table II,  $G + D$  is better than  $G$  under Scenario 2 and much better than  $G$  under Scenario 3. Scenario 4 includes treatment-covariate interactions, with  $G + D$  superior to  $G$  if the patient has LMS ( $Z_2 = 1$ ), but inferior to  $G$  if the patient has a type of sarcoma other than LMS ( $Z_2 = -1$ ). Because the first three prognostic subgroups of patients who have either PPR or LMS on average comprise only 31% of the sample, and moreover the entire of sample of 120 is rather limited for a four-subgroup trial, there is relatively little information to allow any method to detect treatment-covariate interactions in these subgroups. Thus, to examine the AR method's behavior when treatment-covariate interaction is present, we simulated Scenario 4 three different ways: using the actual subgroup proportions with maximum trial sample size 120 (4a), using equal proportions of 25% for each subgroup and maximum trial sample size 120 (4b), and using equal subgroup proportions with maximum trial sample size 240 (4c). The trial was simulated 1000 times under each case.

### 4.3 Simulation Results

To interpret the simulation results in Tables II - IV, it is helpful to consider the overall treatment effect within each treatment-subgroup combination in terms of the weighted average  $\zeta$ . The simulations summarized in Table II show that the AR method behaves as expected under the null case (Scenario 1), with the sample sizes divided equally between  $G$  and  $G + D$ . Under Scenario 2, about 14 (12%) more patients are randomized to the superior  $G + D$  arm, and the correct selection probabilities in the subgroups are all 60% to 81%. This is a desirable

result in the first three subgroups, each of which has a rather small overall sample size, and it reflects the fact that the regression model borrows strength across the patient subgroups. With the comparatively greater superiority of  $G + D$  over  $G$  under Scenario 3, as quantified by  $\zeta$ , the behavior of the AR method becomes substantially more favorable. The imbalance in treatment assignment increases to 28 (14%) more patients given the superior treatment, and the method selects the better treatment in each subgroup 75% to 96% of the time. Under Scenario 4, the treatment-covariate interactions make  $G$  superior for  $\mathbf{Z} = (1, -1)$  or  $(-1, -1)$ , while  $G + D$  is superior for the other two subgroups,  $\mathbf{Z} = (1, 1)$  or  $(-1, 1)$ . Under 4a, which has the actual sample size of 120 and the historical subgroup proportions, owing to the small sample sizes in each of the first three subgroups, each has a moderate difference in sample size and selection probability. For the subgroup  $\mathbf{Z} = (-1, -1)$ , which has on average 69% of the 120 patients, the AR procedure behaves quite well, with better than 2 to 1 odds over the course of the trial of randomizing a patient to the superior arm, and a 100% correct selection probability. If the sample of 120 is balanced among the four subgroups with on average 30 patients in each (4b), then the method has correct selection percentages within all subgroups of 93% to 96%, and the sample size imbalance in favor of the superior arm is 3 to 2 in the  $\mathbf{Z} = (1, 1)$  subgroup and just under 2 to 1 in each of the other subgroups. Finally, if the sample size is doubled to 240 patients and the four subgroups are of equal size then, in each subgroup, the correct selection percentage is at least 97% and the sample size is unbalanced over 2 to 1 in favor of the superior treatment.

Recall that the design includes rules for stopping the trial early and selecting the superior arm within each subgroup if the interim data show that one arm is highly likely to have a higher four-course success probability. In the  $\mathbf{Z} = (-1, -1)$  subgroup, the probability of stopping early and selecting  $G + D$  was 3% under Scenario 1, 8% under Scenario 2, and 30% under Scenario 3. In each of the other three subgroups, the early selection percentages were all 1% to 11% under each of these three scenarios. Under Scenario 4, where  $G$  is the superior arm for the  $\mathbf{Z} = (-1, -1)$  subgroup, the  $G$  arm was selected early 50% of the time under 4a, 25% of the time under 4b, and 40% of the time under 4c. These early selection percentages

reflect the average sample sizes of 83, 30 and 60 for this subgroup in these three cases. For the other three subgroups, the early selection percentages were trivially small under 4a and all in the range 20% to 25% under 4b and 40% to 45% under 4c.

#### 4.4 Sensitivity analyses

A useful sensitivity analysis is to assess the effect on the AR procedure of borrowing strength through the regression model. To do this, we repeated the simulations under each of two extreme approaches, suggested to us by Don Berry in a personal communication. One extreme is to simply ignore covariates, so that  $\mathbf{Z} = \mathbf{0}$ , the parameters  $(\boldsymbol{\beta}_F, \boldsymbol{\beta}_R, \boldsymbol{\tau}_R, \boldsymbol{\tau}_R)$  are dropped from the model, and  $\eta_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \mu_j + \alpha_j + \gamma_{k,j}$  for all  $\mathbf{Z}$ . If the outcome probabilities do in fact vary with  $\mathbf{Z}$ , then this model may be regarded as borrowing too much strength across prognostic subgroups. At the other extreme, one may assume a fully interactive model under which  $\eta_{k,j}$  accounts for stage and treatment effects but is specific within each prognostic subgroup, that is,  $\eta_{k,j}(T, \mathbf{Z}, \boldsymbol{\theta}) = \mu_j(\mathbf{Z}) + \gamma_{k,j}(\mathbf{Z}) + \alpha_j(\mathbf{Z})T$  for each  $\mathbf{Z}$ . This model does not borrow strength across prognostic subgroups at all, and consequently the AR procedure is implemented separately within each subgroup. Effectively, this results in four simultaneous but separate trials being conducted. The model (5) and (6) and the AR method actually used may be regarded as lying between these two extremes.

Table III summarizes simulation results of the AR method ignoring covariates. With this approach, under each scenario the selection probabilities are identical for the four subgroups. The method is balanced under the null scenario, as expected. Under Scenarios 2 and 3, while the sample size imbalances are somewhat smaller but nearly the same as those obtained under the full regression model (Table II), due to the enforced homogeneity of the model without covariates the correct selection percentages in the first three subgroups are 84% under Scenario 2 and 98% under Scenario 3. These are all slightly higher than the corresponding values for the first three subgroups in Table II. While this might seem to indicate that it is actually slightly advantageous to ignore covariates when implementing the AR procedure, the results under Scenario 4 show that things may go very wrong if this is done. Under all three cases

of Scenario 4, the selection percentage of 71% to 72% for  $G + D$  is of course very undesirable in the two subgroups with  $Z_2 = -1$ , where  $G + D$  is the inferior treatment arm. Moreover, the desirable imbalance of the AR procedure in terms of the numbers of patients randomized to the superior treatment arm in each subgroup, which is the motivation for using the AR method in the first place, is lost. Taken together, these simulation results indicate that it is very important to account for covariates, and moreover to accommodate the possibility of treatment-covariate interactions when implementing an AR procedure.

Table IV summarizes simulation results of the AR method conducted separately within the four subgroups. Comparing these results to those in Table II, the properties are very similar under the null case, Scenario 1. When there are treatment differences (Scenarios 2 and 3), the sample size imbalances are roughly the same, while the correct selection percentages drop by 0% to 5%. When there is treatment-covariate interaction (Scenario 4), the fully interactive model has smaller sample size imbalances in favor of the superior arm within each subgroup, and the loss in numbers of patients randomized to the superior treatment arm increases going from cases 4a to 4b to 4c. The fully interactive model also has smaller correct selection percentages, with the drop relative to the values in Table II largest (11% to 17%) for the two subgroups with  $Z_2 = -1$ . Taken together, these results suggest that a great deal may be lost if one fails to use a regression model to borrow strength across prognostic subgroups when implementing an AR procedure.

A natural question is how sensitive the AR method is to the parameter  $\lambda_F$ , elicited from the physicians, that quantifies the relative importance of decreasing  $\xi_{4,F}$  compared to increasing  $\xi_{4,R}$ . To explore sensitivity of the method to  $\lambda_F$ , we considered two additional scenarios. The first is obtained from Scenario 1 by keeping all parameters the same for  $G$  while doubling the value of  $p_R$  for the  $G + D$  arm with all of the increase in  $p_R$  obtained by decreasing  $p_S$  while leaving  $p_F$  unchanged. Thus, for example, in the  $\mathbf{Z} = (1,1)$  subgroup,  $(p_R, p_F) = (.22, .28)$  for the  $G$  arm and  $(.44, .28)$  for the  $G + D$  arm. The second scenario is obtained by halving  $p_F$  for the  $G + D$  arm with all of the decrease in  $p_F$  obtained by increasing  $p_S$ . For example, in the  $\mathbf{Z} = (1,1)$  subgroup,  $(p_R, p_F) = (.22, .14)$  for the  $G + D$

arm. We simulated the trial for values of  $\lambda_F$  ranging from 0.10 to 10.0, under each of these two scenarios with equal subgroup sample sizes and a maximum sample size of 120. These values of  $\lambda_F$  correspond to  $\omega_R$  ranging from 0.10 to 0.91, which covers any reasonable range that would require accounting for a trinary rather than a binary outcome. The AR method proved to be remarkably insensitive to  $\lambda_F$  over this range, with all subgroup-treatment sample sizes differing by at most two patients and all subgroup-treatment selection percentages differing by at most 3% over the range of  $\lambda_F$  in each scenario.

#### *4.5 Bias and Mean Squared Error*

Table V gives the bias and mean squared error (MSE) of the posterior mean of each parameter, with the average value over the 1000 simulations tabulated. When considered relative to the absolute magnitude of the true parameter value, the bias seems to be negligible in most cases. Two exceptions are the posterior mean of  $\beta_{R,1}$ , which has negative biases of 32%, 29%, 22% and 13% under Scenarios 3, 4a, 4b and 4c; and the posterior mean of  $\tau_{R,2}$ , which has positive biases of 22%, 20% and 6% under Scenarios 4a, 4b and 4c. Given the large number of parameters and cases considered, however, these selected extreme results do not seem to indicate any systematic pattern of substantive bias due to the adaptive procedure under the Bayesian model. The issue of bias introduced into treatment effect estimates by outcome-adaptive procedures is very important, however. If such Bayesian methods are to be widely accepted, the frequentist properties of Bayesian estimators based on data from outcome-adaptive designs is an area for future research

## **5. USER INTERFACE**

An essential requirement for conducting a multi-center clinical trial that uses an outcome-adaptive design is a high quality, real-time user interface. A user interface is a computer program that provides an environment for enrolling patients, entering patient data at each participating institution, as well as interfacing with the computer program that performs the statistical computations underlying the adaptive decision rules. These activities must be performed in real time throughout the trial so that the most current data are available to

perform the statistical computations each time a new patient is enrolled and there is no delay in enrollment.

The interface that is being used to conduct the sarcoma trial has several desirable properties. Physicians and research nurses at the 11 clinical centers participating in the trial are each given a user name and a password to access a web site that has modules for both training and trial conduct. Each module essentially looks like a computer-based patient log that tells the user what actions to take, including what treatment to give each new patient, when it is necessary to update patient outcome data, and when a treatment arm has been closed within a patient subgroup. The training module is provided to allow each user to become familiarized with the system before actually beginning to enroll patients.

The computer program that performs the statistical computations for the adaptive decisions during the trial was written in C++, and the user interface was written in ASP. Both programs reside on a secure server located in the Biostatistics Department at M.D. Anderson Cancer Center.

## 6. DISCUSSION

We have proposed and studied by computer simulation an outcome-adaptive, covariate-adjusted method for randomizing patients between treatments in a multi-stage clinical trial. The simulation results show that the method does a good job of unbalancing the randomization in favor of a superior treatment arm, and that it does this reliably within subgroups when there is treatment-covariate interaction. The method also has good selection probabilities. Moreover, the method is consistent in that, within each subgroup, both the degree of imbalance in favor of a better treatment and the correct selection probability increase as the sample size increases. Sensitivity analyses indicated that there is a substantial advantage from borrowing strength across subgroups through a regression model that includes parameters for treatment-covariate interactions, and that this is preferable to either ignoring covariates or conducting separate trials within subgroups.

There is a statistical literature on various theoretical aspects of covariate-adjusted adap-

tive treatment assignment methods. Woodroffe [24] derives approximately optimal Bayesian treatment assignment rules in the case of a univariate outcome with one covariate. Similarly, Clayton [25] considers the case of binary outcomes with one covariate under a bandit model, where the goal is to maximize the number of successes in the trial. Both of these papers deal with so called “bandit strategies,” which are essentially allocation schemes based on an expected gain function computed using backward induction, rather than randomization schemes. In contrast, Rosenberger, Vidyashankar and Agarwal [26] address the same problem as Clayton, but propose using adaptive randomization probabilities based on parameters estimated using maximum likelihood. They model the probability of success as the logit of a linear component including treatment and a vector of covariates, similarly to our model (5) in the generalized logistic case. Yang and Zhu [27] consider a multi-armed bandit framework with covariates, using a non-parametric estimator of the regression function, but they introduce randomization “... to appropriately balance the tendency to use the currently most promising arm with the desire to try other arms.” The key point in this regard is that, if one wishes to maximize the number of successes over a future horizon of patients, then the value in future information must be considered along with what is optimal for the next patient [9]. In theory, an appropriate loss function in a decision-theoretic framework that accounts for an appropriate future horizon should do this. However, given the computational limitations of carrying out a full backward induction [16,17] and the subjectivity of any loss function, introducing randomness into the treatment assignment algorithm is a very reliable way to ensure that information on the comparative treatment effects is obtained. The papers noted above are devoted to theoretical issues, however, with little attention to practical issues such as interaction with physicians, multivariate or multi-stage outcomes, logistics of trial conduct, or user interfaces.

Our design does not include re-randomization after the first stage of therapy. This certainly can be incorporated into the design, provided that the physicians wish this to be done. An important, closely related issue with both scientific and therapeutic implications is that patients whose therapy fails but who are still alive are often treated with salvage therapies.

Such a treatment is generally chosen by the physician based on clinical experience. Because a patient whose initial treatment has failed typically has a worse prognosis than (s)he had at the start of therapy, salvage treatment is more likely to be a more novel agent or combination than was given initially. Thus, patients whose initial treatment has failed are often eligible for a phase II trial. This motivates conducting several single-arm phase II trials together as part of the original trial. This may be done by re-randomizing patients whose initial therapy has failed among a set of salvage therapies. While this may appear to be straightforward in principle, the technical details required to do this are far from straightforward. Design and implementation of such a trial would require one to account for possible synergistic or antagonistic effects of different treatments given consecutively, e.g. as done by Thall, Millikan and Sung [4], as well as the issues of feasible model parameterization, sample size, early stopping for inactive salvage treatments, and calibrating multiple adaptive rules for randomization, re-randomization, selection of treatment strategies, and early stopping due to futility or unacceptable adverse event rates. Additionally, such a re-randomization may be done in several ways, depending on the particular clinical requirements and scientific goals. This is an important area for future research.

#### ACKNOWLEDGEMENTS

We thank Donald Berry for his helpful comments on an earlier draft of this paper. This research was partially supported by NIH grant R01-CA-83932.

#### REFERENCES

1. Patel, S. (2002) Phase III randomized trial of 120 minutes infusion of GEMZAR versus 90 minutes infusion of GEMZAR plus docetaxel in unresectable soft tissue sarcoma. M.D. Anderson protocol ID 02-633.
2. Thall, P.F., Sung, H-G., and Estey, E.H. (2002) Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *Journal of the American Statistical Association* **97**, 29-39.

3. Thall, P.F., Sung, H-G., and Estey, E.H. (2002) Multi-course treatment strategies for clinical trials of rapidly fatal diseases (with discussion). Invited paper, *Case Studies in Bayesian Statistics, VI*, Lecture Notes in Statistics 167, Springer, New York, pp. 33-89.
4. Thall, P.F., Millikan, R.E. and Sung, H.G. (2000) Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*, **19**, 1011-1028.
5. Lavori, P.W. and Dawson, R. (2000). A design for testing clinical strategies: Biased adaptive within-subject randomization. *J. Royal Statistical Society, A* **163**, 29-38.
6. Jennison, C and Turnbull, B.W. *Group Sequential Methods With Application to Clinical Trials*, Chapman and Hall, New York, 2000.
7. Louis, T. A. (1977). Sequential allocation in clinical trials comparing two exponential survival curves. *Biometrika* **33**, 627-634.
8. Wei, L.J. and Durham, S. (1978) The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association* **73**, 840-843.
9. Gittins, J.C. (1979) Bandit processes and dynamic allocation indices. (with discussion) *Journal of the Royal Statistical Society, Series B* **48**, 148-177.
10. Eick, G. E. (1988). The two-armed bandit with delayed responses. *The Annals of Statistics* **16**, 254-264.
11. Ware, J.H. (1989) Investigating therapies of potentially great benefit: ECMO (with discussion) *Statistical Science* **4**, 298-340.
12. Rosenberger, W. F. and Lachin, J. M. (1993).The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials* **14**, 471-484.
13. Rosenberger, W. F. (1996). New directions in adaptive designs. *Statistical Science* **11**, 137-149.

14. Berry, D.A. and Fristedt, B. *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, New York, 1985.
15. Emanuel, E.J. and Patterson, W.B. (1998) Ethics of randomized clinical trials. *Journal of Clinical Oncology* **48**, 6-29.
16. Carlin, B.P Kadane, J.B. and Gelfand, A.E (1998) Approaches for optimal sequential decision analysis in clinical trials *Biometrics* **54**, 964-975.
17. Mueller, P. and Parmigiani, G. (2000) Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association* **99**, 1329-1330.
18. Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285–294.
19. Berry, D. A. and Eick, G. E. (1995). Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in Medicine* **14**, 231–246.
20. Bather, J.A. (1981) Randomized allocation of treatments in sequential medical trials (with discussion) *Journal of the Royal Statistical Society, Series B* **43**, 265-292.
21. Robins, H. (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58**, 527-536.
22. Zelen, M. (1969) Play the winner and the controlled clinical trial. *Journal of the American Statistical Association* **64**, 131-146.
23. Thall, P.F., Inoue, L.Y.T. and Martin, T. (2002) Adaptive decision making in a lymphocyte infusion trial. *Biometrics* **58**, 560–568.
24. Woodroffe, M. (1979) A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association* **74**, 799-806.
25. Clayton, M.K. (1989) Covariate models for Bernoulli bandits. *Sequential Analysis* **8**, 405-426.

26. Rosenberger, W. F., Vidyashankar, A.N. and Agarwal, D.K. (2001) Covariate-adjusted response-adaptive designs for binary response. *Biopharmaceutical Statistics*, **11**, 227-236.
27. Yang, Y. and Zhu, D. (2002) Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals of Statistics* **30**, 100-121.

**Table I.** Possible outcomes and likelihood contributions. Each  $\pi_{k,j}$  has argument  $(T, \mathbf{Z}, \boldsymbol{\theta})$ .

Case	Stage of Therapy				Overall Outcome	Contribution $\xi_{k,j}$ to the Likelihood
	1	2	3	4		
1	R				R	$\pi_{1,R}$
2	F				F	$\pi_{1,F}$
3	S	R			R	$\pi_{1,S} \pi_{2,R}$
4	S	F			F	$\pi_{1,S} \pi_{2,F}$
5	S	S	R		R	$\pi_{1,S} \pi_{2,S} \pi_{3,R}$
6	S	S	F		F	$\pi_{1,S} \pi_{2,S} \pi_{3,F}$
7	S	S	S	R	R	$\pi_{1,S} \pi_{2,S} \pi_{3,S} \pi_{4,R}$
8	S	S	S	F	F	$\pi_{1,S} \pi_{2,S} \pi_{3,S} \pi_{4,F}$
9	S	S	S	S	S	$\pi_{1,S} \pi_{2,S} \pi_{3,S} \pi_{4,S}$

**Table II.** Operating characteristics of the adaptive randomization procedure

		G			G+D			
<i>Scenario 1</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	4.8	49	(.22, .28)	0.41	4.8	51
(1, -1)	(.17, .32)	0.33	3.6	49	(.17, .32)	0.33	3.5	51
(-1, 1)	(.27, .23)	0.50	10	50	(.27, .23)	0.50	10	50
(-1, -1)	(.21, .27)	0.41	41	50	(.21, .27)	0.41	41	50
Totals			59 ( 50 % )				60 ( 50 % )	
<i>Scenario 2</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	4.5	36	(.29, .25)	0.49	4.9	64
(1, -1)	(.17, .32)	0.33	3.4	40	(.23, .29)	0.41	3.8	60
(-1, 1)	(.27, .23)	0.50	9.5	30	(.35, .20)	0.59	11	70
(-1, -1)	(.21, .27)	0.41	36	19	(.28, .24)	0.50	47	81
Totals			53 ( 44 % )				67 ( 56 % )	
<i>Scenario 3</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	4.4	23 ( 0 )	(.34, .19)	0.59	5.3	77
(1, -1)	(.17, .32)	0.33	3.2	25 ( 0.1 )	(.27, .23)	0.51	4	75
(-1, 1)	(.27, .23)	0.50	8.7	19 ( 0.1 )	(.40, .15)	0.67	12	81
(-1, -1)	(.21, .27)	0.41	30	4 ( 0 )	(.33, .18)	0.59	53	96
Totals			46 ( 38 % )				74 ( 62 % )	

**Table II. (continued)**

		G				G+D			
<i>Scenario 4 a : <math>N_{max} = 120</math>, True Subgroup Percentages</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	8	(.22, .28)	0.41	4.1	21	(.37, .18)	0.62	5.3	79
(1, -1)	6	(.17, .32)	0.33	4.2	85	(.07, .59)	0.12	3	15
(-1, 1)	17	(.27, .23)	0.50	8.6	15	(.42, .13)	0.71	12	85
(-1, -1)	69	(.21, .27)	0.41	58	100	(.10, .50)	0.17	25	0
<i>Scenario 4 b: <math>N_{max} = 120</math>, 25% in Each Subgroup</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	25	(.22, .28)	0.41	12	7	(.37, .18)	0.62	18	93
(1, -1)	25	(.17, .32)	0.33	19	96	(.07, .59)	0.12	11	4
(-1, 1)	25	(.27, .23)	0.50	11	7	(.42, .13)	0.71	19	93
(-1, -1)	25	(.21, .27)	0.41	19	96	(.10, .50)	0.17	11	4
<i>Scenario 4 c: <math>N_{max} = 240</math>, 25% in Each Subgroup</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	25	(.22, .28)	0.41	18	2	(.37, .18)	0.62	42	98
(1, -1)	25	(.17, .32)	0.33	44	99	(.07, .59)	0.12	16	1
(-1, 1)	25	(.27, .23)	0.50	18	3	(.42, .13)	0.71	42	97
(-1, -1)	25	(.21, .27)	0.41	44	99	(.10, .50)	0.17	16	1

**Table III.** Operating characteristics of the adaptive randomization procedure, ignoring prognostic covariates.

	G				G+D			
<i>Scenario 1</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	4.7	51	(.22, .28)	0.41	4.7	49
(1, -1)	(.17, .32)	0.33	3.4	51	(.17, .32)	0.33	3.4	49
(-1, 1)	(.27, .23)	0.50	9.6	51	(.27, .23)	0.50	9.8	49
(-1, -1)	(.21, .27)	0.41	40	51	(.21, .27)	0.41	40	49
Totals			57 (50%)				58 (50%)	
<i>Scenario 2</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	3.9	16	(.29, .25)	0.49	5.1	84
(1, -1)	(.17, .32)	0.33	2.9	16	(.23, .29)	0.41	3.7	84
(-1, 1)	(.27, .23)	0.50	8.3	16	(.35, .20)	0.59	11	84
(-1, -1)	(.21, .27)	0.41	34	16	(.28, .24)	0.50	43	84
Totals			49 (44%)				63 (56%)	
<i>Scenario 3</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	3.1	2	(.34, .19)	0.59	4.8	98
(1, -1)	(.17, .32)	0.33	2.5	2	(.27, .23)	0.51	3.7	98
(-1, 1)	(.27, .23)	0.50	6.7	2	(.40, .15)	0.67	10	98
(-1, -1)	(.21, .27)	0.41	27	2	(.33, .18)	0.59	42	98
Totals			40 (39%)				61 (61%)	

**Table III. (continued)**

		G				G+D			
<i>Scenario 4 a : <math>N_{max} = 120</math>, True Subgroup Percentages</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	8	(.22, .28)	0.41	4.2	29	(.37, .18)	0.62	4.2	71
(1, -1)	6	(.17, .32)	0.33	3.2	29	(.07, .59)	0.12	3.0	71
(-1, 1)	17	(.27, .23)	0.50	9.7	29	(.42, .13)	0.71	9.7	71
(-1, -1)	69	(.21, .27)	0.41	40.7	29	(.10, .50)	0.17	42.1	71
<i>Scenario 4 b: <math>N_{max} = 120</math>, 25% in Each Subgroup</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	25	(.22, .28)	0.41	14	29	(.37, .18)	0.62	15	71
(1, -1)	25	(.17, .32)	0.33	13	29	(.07, .59)	0.12	15	71
(-1, 1)	25	(.27, .23)	0.50	13	29	(.42, .13)	0.71	15	71
(-1, -1)	25	(.21, .27)	0.41	14	29	(.10, .50)	0.17	15	71
<i>Scenario 4 c: <math>N_{max} = 240</math>, 25% in Each Subgroup</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	25	(.22, .28)	0.41	25	28	(.37, .18)	0.62	30	72
(1, -1)	25	(.17, .32)	0.33	25	28	(.07, .59)	0.12	30	72
(-1, 1)	25	(.27, .23)	0.50	25	28	(.42, .13)	0.71	30	72
(-1, -1)	25	(.21, .27)	0.41	25	28	(.10, .50)	0.17	30	72

**Table IV.** Operating characteristics of the adaptive randomization procedure, conducted separately within prognostic subgroups.

	G				G+D			
<i>Scenario 1</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	5	51	(.22, .28)	0.41	5	49
(1, -1)	(.17, .32)	0.33	3.9	47	(.17, .32)	0.33	4.1	53
(-1, 1)	(.27, .23)	0.50	10	52	(.27, .23)	0.50	11	48
(-1, -1)	(.21, .27)	0.41	41	51	(.21, .27)	0.41	40	49
Totals			60 (50%)				60 (50%)	
<i>Scenario 2</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	4.8	38	(.29, .25)	0.49	5.2	62
(1, -1)	(.17, .32)	0.33	3.8	40	(.23, .29)	0.41	4.2	60
(-1, 1)	(.27, .23)	0.50	9.8	34	(.35, .20)	0.59	11	66
(-1, -1)	(.21, .27)	0.41	36	21	(.28, .24)	0.50	44	79
Totals			54 (46%)				64 (54%)	
<i>Scenario 3</i>								
<b>Z</b>	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	(.22, .28)	0.41	4.8	28	(.34, .19)	0.59	5.2	72
(1, -1)	(.17, .32)	0.33	3.8	30	(.27, .23)	0.51	4.2	70
(-1, 1)	(.27, .23)	0.50	9.2	18	(.40, .15)	0.67	11	82
(-1, -1)	(.21, .27)	0.41	30	4.1	(.33, .18)	0.59	44	96
Totals			48 (42%)				65 (58%)	

**Table IV. (continued)**

		G				G+D			
<i>Scenario 4 a : <math>N_{max} = 120</math>, True Subgroup Percentages</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	8	(.22, .28)	0.41	4.6	24	(.37, .18)	0.62	5.4	76
(1, -1)	6	(.17, .32)	0.33	5.2	68	(.07, .59)	0.12	4.8	32
(-1, 1)	17	(.27, .23)	0.50	9.1	15	(.42, .13)	0.71	11.0	85
(-1, -1)	69	(.21, .27)	0.41	45.0	90	(.10, .50)	0.17	33.0	10
<i>Scenario 4 b: <math>N_{max} = 120</math>, 25% in Each Subgroup</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	25	(.22, .28)	0.41	12	11	(.37, .18)	0.62	17	89
(1, -1)	25	(.17, .32)	0.33	16	79	(.07, .59)	0.12	13	21
(-1, 1)	25	(.27, .23)	0.50	12	12	(.42, .13)	0.71	16	88
(-1, -1)	25	(.21, .27)	0.41	16	78	(.10, .50)	0.17	13	22
<i>Scenario 4 c: <math>N_{max} = 240</math>, 25% in Each Subgroup</i>									
<b>Z</b>	%	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.	$(p_R, p_F)$	$\zeta$	# Pats.	% Sel.
(1, 1)	25	(.22, .28)	0.41	19	4	(.37, .18)	0.62	31	96
(1, -1)	25	(.17, .32)	0.33	33	88	(.07, .59)	0.12	23	12
(-1, 1)	25	(.27, .23)	0.50	19	4	(.42, .13)	0.71	32	96
(-1, -1)	25	(.21, .27)	0.41	33	88	(.10, .50)	0.17	23	12

**Table V.** Bias and mean squared error (MSE) of each posterior mean parameter value.

	<i>Scenario 3</i>			<i>Scenario 4</i>						
	True Value	Bias	MSE	True Value	a		b		c	
					Bias	MSE	Bias	MSE	Bias	MSE
$\alpha_R$	0.252	0.033	0.058	0.036	-0.002	0.1	0.022	0.056	0.0	0.04
$\beta_{R,1}$	-0.103	-0.033	0.048	-0.083	-0.024	0.05	-0.017	0.033	-0.011	0.02
$\beta_{R,2}$	0.142	0.035	0.008	0.402	-0.008	0.05	-0.003	0.036	0.01	0.029
$\tau_{R,1}$	0.00	-0.001	0.001	0.02	0.013	0.107	0.0	0.042	-0.003	0.021
$\tau_{R,2}$	0.00	0.000	0.001	0.257	0.056	0.077	0.051	0.054	0.068	0.045
$\alpha_F$	-0.168	-0.015	0.044	0.136	-0.001	0.075	0.012	0.034	-0.002	0.023
$\beta_{F,1}$	0.097	0.004	0.037	0.131	0.004	0.043	0.004	0.027	-0.003	0.014
$\beta_{F,2}$	-0.055	-0.009	0.031	-0.396	0.018	0.04	-0.011	0.028	0.001	0.017
$\tau_{F,1}$	0.00	0.002	0.001	0.032	-0.001	0.072	0.0	0.029	-0.006	0.017
$\tau_{F,2}$	0.00	0.000	0.002	-0.336	-0.02	0.055	-0.028	0.031	-0.034	0.02
$\gamma_{R,2}$	0.000	0.010	0.059	0.0	0.02	0.061	0.016	0.06	0.017	0.047
$\gamma_{R,3}$	-2.051	-0.016	0.04	-2.051	-0.022	0.035	-0.015	0.041	-0.029	0.056
$\gamma_{R,4}$	-1.581	-0.031	0.035	-1.581	-0.031	0.03	-0.018	0.036	-0.045	0.048
$\gamma_{F,2}$	0.000	-0.02	0.055	0.0	0.005	0.054	-0.013	0.061	-0.009	0.043
$\gamma_{F,3}$	-0.365	-0.016	0.061	-0.365	-0.006	0.062	0.006	0.069	-0.023	0.059
$\gamma_{F,4}$	0.882	-0.145	0.036	0.882	-0.145	0.036	-0.147	0.037	-0.119	0.03
$\mu_R$	-0.612	-0.084	0.064	-0.827	-0.156	0.095	-0.116	0.07	-0.114	0.053
$\mu_F$	-0.786	0.020	0.050	-0.482	0.053	0.059	0.046	0.036	0.046	0.025