STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2802–2815
Published online 19 October 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3109

Accounting for patient heterogeneity in phase II clinical trials

J. Kyle Wathen[1,*,†], Peter F. Thall[1], John D. Cook[1] and Elihu H. Estey[2]

[1]*Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, Box 447, 1515 Holcombe Boulevard, Houston, TX 77030, U.S.A.*
[2]*Department of Leukemia, University of Texas, M.D. Anderson Cancer Center, Box 428, 1515 Holcombe Boulevard, Houston, TX 77030, U.S.A.*

SUMMARY

Phase II clinical trials typically are single-arm studies conducted to decide whether an experimental treatment is sufficiently promising, relative to standard treatment, to warrant further investigation. Many methods exist for conducting phase II trials under the assumption that patients are homogeneous. In the presence of patient heterogeneity, however, these designs are likely to draw incorrect conclusions. We propose a class of model-based Bayesian designs for single-arm phase II trials with a binary or time-to-event outcome and two or more prognostic subgroups. The designs' early stopping rules are subgroup specific and allow the possibility of terminating some subgroups while continuing others, thus providing superior results when compared with designs that ignore treatment–subgroup interactions. Because our formulation requires informative priors on standard treatment parameters and subgroup main effects, and non-informative priors on experimental treatment parameters and treatment–subgroup interactions, we provide an algorithm for computing prior hyperparameter values. A simulation study is presented and the method is illustrated by a chemotherapy trial in acute leukemia. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: adaptive design; Bayesian design; futility rule; phase II clinical trial; simulation

1. INTRODUCTION

Phase II clinical trials are usually single-arm studies conducted to decide whether an experimental treatment, E, is sufficiently promising relative to a standard treatment, S, to be studied in a

*Correspondence to: J. Kyle Wathen, Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, Box 447, 1515 Holcombe Boulevard, Houston, TX 77030, U.S.A.
†E-mail: jkwathen@mdanderson.org

Contract/grant sponsor: NIH grant; contract/grant number: CA-83932

Received 27 June 2007
Accepted 17 September 2007

large-scale randomized trial. Interim monitoring rules are usually employed to stop the trial early for futility, and to avoid giving an unacceptable number of patients an inferior treatment. Taylor *et al*. [1] provide a simulation-based rationale for conducting such single-arm trials, rather than randomizing, when there is substantial information on S. Numerous designs have been proposed for phase II trials [2–7]. Several methods that use regression models to account for covariates in particular phase II settings have been proposed [8–10]. When the response probabilities in prognostic subgroups are exchangeable one can use a hierarchical Bayesian model [11]. In a setting where the assumptions required for using a hierarchical model are not appropriate, however, a general method for phase II trials that accounts for patient heterogeneity between prognostic subgroups is not currently available.

To see why such a method is needed, consider a trial including good (G) and poor (P) prognosis patients where the historical response probabilities with S (or their means under a Bayesian formulation) are $\pi_{S,G}=0.45$ and $\pi_{S,P}=0.25$. Suppose that an improvement of 0.15 within each subgroup is desired, that is, response probabilities $\pi_{E,G}\geqslant 0.60$ and $\pi_{E,P}\geqslant 0.40$ with E are targeted. Such a trial often is designed, based on either frequentist or Bayesian criteria, by conducting simultaneous but separate trials within the subgroups, using two separate sets of rules that stop the trial and declare E not promising within the G subgroup if $\pi_{E,G}\geqslant 0.60$ is unlikely, and within the P subgroup if $\pi_{E,P}\geqslant 0.40$ is unlikely. This approach is inherently flawed since, if an agent that does not provide an improvement in one subgroup is unlikely to provide an improvement in the other subgroup, then the design cannot exploit this fact since the data from the two subgroups are not combined in any way. A common alternative approach that does use the combined data is simply to ignore patient prognosis, assume average historical and targeted probabilities, say $(0.45+0.25)/2=0.35$ and $(0.60+0.40)/2=0.50$, and conduct a conventional single-arm trial. As an alternative, London and Chang [12] propose a stratified phase II design that adjusts for patient heterogeneity. However, these approaches are both flawed since, in the presence of treatment–subgroup interactions, they lead to designs with very large false-positive error and false-negative error (FNR) rates within each subgroup.

In this article we present a Bayesian methodology for the design and conduct of single-arm phase II clinical trials with binary or time-to-event (TTE) outcomes that accounts for heterogeneity between patient prognostic subgroups when treatment–subgroup interactions are present. Our formulation generalizes the approaches of Thall and Simon [5] and Thall *et al*. [13]. In each case, we assume a regression model with linear term including treatment–subgroup interactions. Owing to the importance of specifying priors that accurately reflect knowledge about the standard treatment and patient prognosis, while also providing a design with good properties, we provide an algorithm for computing numerical values of the prior hyperparameters. We propose monitoring rules, which may be applied continuously, periodically or after successive cohorts of patients, that allow accrual to be terminated in some subgroups while continuing in the others. We employ computer simulation to calibrate design parameters in order to obtain a design having good frequentist operating characteristics. Our aim is to provide a practical methodology with a wide range of potential applications and provide free software for implementing the method.

In Section 2 we present the regression model underlying the method and provide an algorithm for prior specification. Section 3 defines the decision criteria. Section 4 presents the results of a simulation study comparing our method to three alternative methods. Section 5 describes an application of our method to a clinical trial in relapsed acute myeloid leukemia (AML) and we conclude with a discussion in Section 6.

DOI: 10.1002/sim

2. PROBABILITY MODEL

2.1. A parametric likelihood

To account for patient heterogeneity in a parsimonious but reliable way, we assume K prognostic subgroups, identified by the categorical covariate $Z \in \{0, 1, \ldots, K-1\}$. Denote the subgroup membership indicators $I(Z=k)$, for $k=0, 1, \ldots, K-1$. For $t=S$ or E, we assume linear components of the form

$$\eta_{t,Z}(\boldsymbol{\theta}) = \xi + \sum_{k=0}^{K-1} \{\beta_k + \tau_k I(t=E)\} I(Z=k) \tag{1}$$

denoting $\beta_0 = 0$ for convenience. The kth element of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{K-1})$ is the historical effect of subgroup k compared with the baseline subgroup 0, and the kth element of $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{K-1})$ is the *E-versus-S* treatment effect within subgroup k. The model parameter vector $\boldsymbol{\theta} = (\xi, \boldsymbol{\beta}, \boldsymbol{\tau})$ is $2K$-dimensional. In terms of the parameters, the purpose of the trial is to learn about $\boldsymbol{\tau}$. We allow the treatment effects to differ between subgroups in order to provide a basis for a design that allows the trial to reach different conclusions within different subgroups. As we will show in Section 5, assuming the simpler model with a single *E-versus-S* effect, $\tau = \tau_0 = \tau_1 = \cdots = \tau_{K-1}$, may lead to a design with very poor properties when treatment–subgroup interactions are present. The linear terms characterizing treatment and prognostic subgroup effects are summarized as follows:

Subgroup	$\eta_{S,Z}(\boldsymbol{\theta})$	$\eta_{E,Z}(\boldsymbol{\theta})$
0	ξ	$\xi+\tau_0$
1	$\xi+\beta_1$	$\xi+\beta_1+\tau_1$
$\vdots$	$\vdots$	$\vdots$
$K-1$	$\xi+\beta_{K-1}$	$\xi+\beta_{K-1}+\tau_{K-1}$

The parametrization used here is critical to how well the method performs. It was chosen because it allows borrowing strength across subgroups and reflects the fact that much more is known *a priori* about the standard treatment. Specifically, by including the parameter ξ in the linear term for all subgroups and the same β_j for each subgroup j, the model borrows information in a desirable way. One consequence of the parametrization is that the prior variance is smallest for the 'baseline' subgroup (subgroup 0, treated with S). It ensures that, within each subgroup j, the linear term $\eta_{S,j}(\boldsymbol{\theta})$ of the treatment group S will have a smaller variance than $\eta_{E,j}(\boldsymbol{\theta})$ for the corresponding experimental subgroup. This reflects the knowledge available, since much more is known about S. If an alternative parametrization were used where the prior for $\eta_{t,Z}$ is normal with mean $\mu_{t,Z}$ and variance $\sigma^2_{t,Z}$, then there would be no borrowing of information between the subgroups and one would effectively be conducting separate trials for each subgroup.

This formulation accommodates trials with either binary or TTE outcomes. For the binary case, denoting the treatment response indicator by Y, we assume the response probabilities $\pi_{t,Z}(\boldsymbol{\theta}) = P(Y=1|t, Z, \boldsymbol{\theta}) = \text{logit}^{-1}\{\eta_{t,Z}(\boldsymbol{\theta})\}$ are defined in terms of $\eta_{t,Z}(\boldsymbol{\theta})$. Denoting the data on the first n patients in the trial by $\mathscr{D}_n = \{(Y_1, Z_1), \ldots, (Y_n, Z_n)\}$, the likelihood is the product

$$\mathscr{L}(\mathscr{D}_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \{\pi_{E,Z_i}(\boldsymbol{\theta})\}^{Y_i} \{1-\pi_{E,Z_i}(\boldsymbol{\theta})\}^{1-Y_i} \tag{2}$$

DOI: 10.1002/sim

For TTE outcomes where Y may be either the time to treatment failure or the time to clinical response, we denote $Y^0 =$ time to the event or right censoring, $\varepsilon = I(Y = Y^0)$, with f_t the probability density function (pdf) and $\mathscr{F}_t$ the survivor function for treatment $t = S$ or E. The likelihood of data $\mathscr{D}_n = \{(Y_1^0, \varepsilon_1, Z_1), \ldots, (Y_n^0, \varepsilon_n, Z_n)\}$ takes the form

$$\mathscr{L}(\mathscr{D}_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{E,Z_i}(Y_i^0|\boldsymbol{\theta})^{\varepsilon_i} \mathscr{F}_{E,Z_i}(Y_i^0|\boldsymbol{\theta})^{1-\varepsilon_i} \tag{3}$$

For many applications with TTE outcomes, it is reasonable to assume exponentially distributed event times with means $E(Y|t, Z, \boldsymbol{\theta}) = \mathrm{e}^{\eta_{t,Z}(\boldsymbol{\theta})}$, so that the pdf and survival functions are $f_{t,Z}(y|\boldsymbol{\theta}) = \mathrm{e}^{-\eta_{t,Z}(\boldsymbol{\theta})} \exp\{-y\,\mathrm{e}^{-\eta_{t,Z}(\boldsymbol{\theta})}\}$ and $\mathscr{F}_{t,Z}(y|\boldsymbol{\theta}) = \exp\{-y\,\mathrm{e}^{-\eta_{t,Z}(\boldsymbol{\theta})}\}$. To facilitate exposition, we will present and illustrate the method in the binary outcome case. Modifications needed to deal with TTE outcomes will be described in Section 6.

2.2. *Establishing priors*

To reflect prior knowledge about S and the effects of patient prognosis, we assume an informative prior on $(\xi, \boldsymbol{\beta})$. In contrast, since little is known about the effects of E at the start of the trial, to ensure that the trial design is ethical and its results persuasive, no undue information should be introduced into the prior for E. We therefore assume a vague prior on $\boldsymbol{\tau}$. This generalizes the approach of Thall and Simon [5], who dealt with the homogeneous case with two response probabilities, π_E and π_S, assuming beta priors. In the present regime, for tractability we assume normal priors on all parameters: $\xi \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2), \beta_k \sim \mathrm{N}(\mu_{\beta_k}, \sigma_{\beta_k}^2)$ for $k = 1, 2, \ldots, K-1$, and $\tau_j \sim \mathrm{N}(\mu_{\tau_j}, \sigma_{\tau_j}^2)$, $j = 0, 1, \ldots, K-1$, where σ_ξ^2, $\sigma_{\beta_1}^2, \ldots, \sigma_{\beta_{K-1}}^2$ must be 'small' and $\sigma_{\tau_0}^2, \ldots, \sigma_{\tau_{K-1}}^2$ must be 'large' in some suitable sense. Denote the hyperparameter mean vector by $\boldsymbol{\psi}_\mu = (\mu_\xi, \mu_{\beta_1}, \ldots, \mu_{\beta_{K-1}}, \mu_{\tau_0}, \ldots, \mu_{\tau_{K-1}})$, the variance vector by $\boldsymbol{\psi}_{\sigma^2} = (\sigma_\xi^2, \sigma_{\beta_1}^2, \ldots, \sigma_{\beta_{K-1}}^2, \sigma_{\tau_0}^2, \ldots, \sigma_{\tau_{K-1}}^2)$, and the vector of all $4K$ hyperparameters by $\boldsymbol{\psi} = (\boldsymbol{\psi}_\mu, \boldsymbol{\psi}_{\sigma^2})$. When applying Bayesian methods to conduct small- to moderate-sized clinical trials with outcome-adaptive decision rules, the numerical values of the hyperparameters are very important because they may substantively affect the decisions. Since the prior on $(\xi, \boldsymbol{\beta}|\boldsymbol{\psi})$ is informative but the prior on $(\boldsymbol{\tau}|\boldsymbol{\psi})$ is non-informative, we provide an algorithm for specifying numerical values of $\boldsymbol{\psi}$ that reflect these general requirements.

In some Bayesian models, there is a direct link between the prior hyperparameter values and a putative number of patients on whom the prior is based, i.e. its effective sample size (ESS). For our model, there is no clear link between $\boldsymbol{\psi}$ and such a prior ESS. Since the ESS of the standard Beta(a, b) distribution is $a+b$, we will exploit this to determine $\boldsymbol{\psi}$ by matching prior moments of the $\pi_{t,j}(\boldsymbol{\theta})$'s to those of specified Betas. To do this, we first impose the K constraints $E\{\pi_{E,j}(\boldsymbol{\theta})\} = E\{\pi_{S,j}(\boldsymbol{\theta})\}$ for $j = 0, \ldots, K-1$, so that the prior expected response rate of E equals that of S within each prognostic subgroup. These constraints are only valid for superiority trials and may be too optimistic for non-inferiority trials. As an alternative, one could impose the K constraints $E\{\pi_{E,j}(\boldsymbol{\theta})\} = \hat{\pi}_{E,j} = E\{\pi_{S,j}(\boldsymbol{\theta})\} - \varepsilon_j$, for example, one may use $\varepsilon_j = 0.05$ to 0.10, for $j = 0, \ldots, K-1$, so that the prior expected response rate of E is less than that of S within each prognostic subgroup. For each t and j, let $f_{t,j}(p)$, $0 \leqslant p \leqslant 1$, denote the prior density on $\pi_{t,j}(\boldsymbol{\theta})$ induced by the prior on $\boldsymbol{\theta}$. Given the above constraints on the mean probabilities, we determine $\boldsymbol{\psi}$ by matching the first two moments of each $f_{t,j}$ to those of a corresponding beta distribution, Beta$(a_{t,j}, b_{t,j})$, with pdf denoted by $f_{t,j}^*$, having mean $a_{t,j}/(a_{t,j}+b_{t,j})$ and ESS $a_{t,j}+b_{t,j}$. We

DOI: 10.1002/sim

assume that estimates of the historical mean response rate and the number of historical patients in each subgroup are available for S, denoted by $\hat{\pi}_{S,j}$ and $N_{S,j}$, respectively, for $j=0,1,\ldots,K-1$. That is, we assume that these values may be computed from informative priors on the $\pi_{S,j}(\boldsymbol{\theta})$'s. Let $N_{E,j}$ denote the desired prior ESS for the distribution of $\pi_{E,j}(\boldsymbol{\theta})$. Given these values, we will use the following algorithm to determine $\boldsymbol{\psi}$.

Algorithm for determining prior hyperparameters

Step 1: For each $j=0,\ldots,K-1$, set $a_{S,j}=N_{S,j}\hat{\pi}_{S,j}$ and $b_{S,j}=N_{S,j}(1-\hat{\pi}_{S,j})$.

Step 2: Find (μ_ξ,σ^2_ξ) to minimize $\int_0^1|f_{S,0}(p)-f^*_{S,0}(p)|\,\mathrm{d}p$, subject to (s.t.) the constraint $E\{\pi_{S,0}(\boldsymbol{\theta})\}=\hat{\pi}_{S,0}$.

Step 3: Given (μ_ξ,σ^2_ξ) from Step 2, for each $j=1,\ldots,K-1$ find $(\mu_{\beta_j},\sigma^2_{\beta_j})$ to minimize $\int_0^1|f_{S,j}(p)-f^*_{S,j}(p)|\,\mathrm{d}p$, s.t. the constraint $E\{\pi_{S,j}(\boldsymbol{\theta})\}=\hat{\pi}_{S,j}$.

Step 4: For each $j=0,\ldots,K-1$, set $a_{E,j}=N_{E,j}\hat{\pi}_{S,j}$ and $b_{E,j}=N_{E,j}(1-\hat{\pi}_{S,j})$.

Step 5: Given $(\mu_\xi,\mu_{\beta_1},\ldots,\mu_{\beta_{K-1}},\sigma^2_\xi,\sigma^2_{\beta_1},\ldots,\sigma^2_{\beta_{K-1}})$ from Steps 2 and 3, for each $j=0,\ldots,K-1$ find $(\mu_{\tau_j},\sigma^2_{\tau_j})$ to minimize $\int_0^1|f_{E,j}(p)-f^*_{E,j}(p)|\,\mathrm{d}p$, s.t. the constraint $E\{\pi_{E,j}(\boldsymbol{\theta})\}=\hat{\pi}_{S,j}$.

The algorithm exploits the facts that, given Step 1, $f_{S,0}$ depends on $\boldsymbol{\psi}$ only through (μ_ξ,σ^2_ξ); given Steps 1 and 2, for each $j=1,\ldots,K-1$, $f_{S,j}$ depends on $\boldsymbol{\psi}$ only through $(\mu_{\beta_j},\sigma^2_{\beta_j})$; and given Steps 1–4, for each $j=0,\ldots,K-1$, $f_{E,j}$ depends on $\boldsymbol{\psi}$ only through $(\mu_{\tau_j},\sigma^2_{\tau_j})$. The minimizations in Steps 2, 3, and 5 are carried out using the simplex algorithm of Nelder and Mead [14]. To reflect prior uncertainty about E in Step 5, reasonable values for $N_{E,j}$ are $\frac{1}{2}$, 1 or 2, which yield suitably large values for each $\sigma^2_{\tau_j}$. Although the algorithm equates the prior mean response rates for S and E within subgroups, alternatively one may assume either more skeptical or more optimistic prior means for the $\pi_{E,j}(\boldsymbol{\theta})$'s. For example, if the more optimistic targets $\hat{\pi}_{S,j}+\varepsilon_j$ with all $\varepsilon_j>0$ are desired, e.g. $\varepsilon_j=\delta_j$ or $\delta_j/2$, then these values would be used as the prior means $E\{\pi_{E,j}(\boldsymbol{\theta})\}$ and in place of $\hat{\pi}_{S,j}$ in Steps 4 and 5 of the algorithm for determining $\boldsymbol{\psi}$.

2.3. Computing posteriors

Let Σ denote the $K\times K$ diagonal matrix with diagonal elements equal to $\boldsymbol{\psi}_{\sigma^2}$. Combining the likelihood and priors from the previous sections, the posterior of $\boldsymbol{\theta}$ after observing $\mathscr{D}_n$ is

$$
\begin{aligned}
f(\boldsymbol{\theta}|\mathscr{D}_n) &\propto \mathscr{L}(\mathscr{D}_n|\boldsymbol{\theta})\,f(\boldsymbol{\theta})\\
&= \prod_{i=1}^{n}\{\pi_{E,Z_i}(\boldsymbol{\theta})\}^{Y_i}\{1-\pi_{E,Z_i}(\boldsymbol{\theta})\}^{1-Y_i}\times\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\psi}_\mu)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\psi}_\mu)\right\}\\
&= \prod_{i=1}^{n}\frac{\exp[Y_i\{\xi+(\boldsymbol{\beta}^{\mathrm{T}}+\boldsymbol{\tau}^{\mathrm{T}})\mathbf{X}(Z_i)\}]}{1+\exp\{\xi+(\boldsymbol{\beta}^{\mathrm{T}}+\boldsymbol{\tau}^{\mathrm{T}})\mathbf{X}(Z_i)\}}\times\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\psi}_\mu)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\psi}_\mu)\right\}
\end{aligned}
$$

where $\mathbf{X}(K)=I(Z=K)$ and we denote $\mathbf{X}(Z)=(X_0(Z),\ldots,X_{K-1}(Z))$. Since $f(\boldsymbol{\theta}\,|\,\mathscr{D}_n)$ is not conjugate to the prior distribution, we use Monte Carlo Markov chain (MCMC) methods [15] to compute posterior probabilities.

DOI: 10.1002/sim

3. DECISION CRITERIA

Typically, E must provide evidence of a response rate greater than that of S to be considered for a randomized phase III study. Formally, in subgroup j, E must provide a specified improvement δ_j over S. We thus apply the subgroup-specific early stopping rules

$$\lambda(\mathscr{D}_n, \boldsymbol{\theta}, j, \delta_j) = \Pr\{\pi_{E,j}(\boldsymbol{\theta}) > \pi_{S,j}(\boldsymbol{\theta}) + \delta_j \mid \mathscr{D}_n\} < p_j \quad \text{for } j = 0, 1, \ldots, K-1 \tag{4}$$

where $\mathscr{D}_n$ denotes the currently available data, and each decision cut-off p_j is a fixed lower probability cut-off, usually in the range of 0.01–0.10. In practice, each p_j may be calibrated to obtain a pre-specified FNR, defined as the probability of stopping the trial early for futility in subgroup j when the fixed 'true' value of $\pi_{E,j}$, denoted by $\pi_{E,j}^{\text{true}}$, equals the fixed target $E\{\pi_{S,j}(\boldsymbol{\theta})\} + \delta_j$. The FNR corresponds to the type II error probability of a frequentist hypothesis-test-based design. These rules may be applied continuously, periodically, or after successive cohorts of patients. An important property of our proposed method is that, because we allow treatment–subgroup interactions, accrual may be stopped within one subgroup but continue in another. Using these rules with posterior quantities computed under regression model (1), the method borrows strength between subgroups. Our simulations will demonstrate the advantages of this approach over simpler methods that ignore treatment–subgroup interactions, or that do not assume a regression model, or that ignore patient subgroups entirely.

4. SIMULATION STUDIES

Although there are many phase II clinical trial designs, in order to fairly evaluate our method while focusing on the advantages of accounting for between-subgroup heterogeneity, we will compare it with three simpler versions of our design. All four methods are based on Bayesian models and use early stopping rules constructed to be as similar as possible to ensure comparability. Since our proposed method accounts for subgroups and, moreover, treatment–subgroup interactions, we denote it by S-TI for brevity. The first simplified method, S-NTI, assumes that there are no treatment–subgroup interactions, with the underlying regression model simplified by assuming $\tau_0 = \cdots = \tau_{K-1}$, so that $\eta_{t,Z}(\boldsymbol{\theta}) = \xi + \sum_{k=1}^{K-1} \beta_j X_k(Z) + \tau_0 I(t = E)$. The second simplification, SEP, is to simultaneously conduct K separate trials, one within each subgroup, without using a regression model to borrow strength. The third simplification, NS, is to assume that there are no subgroup effects and conduct one trial. Since SEP does not borrow strength between subgroups and conducts separate trials within subgroups, we assume independent beta-binomial models within subgroups, with $\pi_{t,j}(\boldsymbol{\theta}) \sim \text{Beta}(a_{t,j}, b_{t,j})$ for each $j = 0, 1, \ldots, K-1$ and $t = S, E$. For NS, we also assume beta-binomial models, but now there are only two Beta distributions, $\pi_E \sim \text{Beta}(a_E, b_E)$ and $\pi_S \sim \text{Beta}(a_S, b_S)$, as in Thall and Simon [5]. To evaluate the designs over a realistic range of cases, we present the results of two simulation studies, one with $K = 2$ subgroups and another with $K = 4$. While each design's decision boundaries are computed based on a Bayesian model that regards the parameters as random, we evaluated each design's behavior over a range of fixed values of $\pi_{E,0}^{\text{true}}, \ldots, \pi_{E,K-1}^{\text{true}}$. Since patient outcomes are rarely observed immediately, we assumed that there is one month from the start of treatment to the evaluation of response, with patients arriving at a rate of 30 per year according to a Poisson process, distributed equally among the subgroups. To ensure comparability, all four methods were implemented using a total maximum sample size of

DOI: 10.1002/sim

100 and cohorts of 10 patients, with the data currently available when patients 11, 21, etc. arrived used to compute the stopping criteria.

To compute $\lambda(\mathscr{D}_n, \boldsymbol{\theta}, j, \delta_j)$, denoted for brevity by $\hat{\lambda}$, in the S-TI and S-NTI designs, we sampled the posterior and calculated $\pi_{E,j}(\boldsymbol{\theta})$ and $\pi_{S,j}(\boldsymbol{\theta})$ for each sampled $\boldsymbol{\theta}$. For the MCMC sampling procedure, we implemented four parallel chains, with an initial burn-in of 5000 samples followed by an additional 25 000 samples that were retained for computing $\hat{\lambda}$ for each chain. To ensure convergence, we monitored the potential scale reduction [15], and if this value was greater than 1.15 we obtained an additional 25 000 samples for each chain. Since the sampling procedure can be very time consuming and $\hat{\lambda}$ must be computed repeatedly during each simulated trial, we stored the $\hat{\lambda}$ values using a method similar to Wathen and Christen [16]. Stored $\hat{\lambda}$ values corresponding to previously simulated data that matched the current data were retrieved to increase the speed of the simulations. In particular, using stored $\hat{\lambda}$ values greatly increases the speed of the process of calibrating p_j in (3) to obtain a pre-specified FNR. Since $\hat{\lambda}$ must be computed and stored for many $\mathscr{D}_n$, once the values $\hat{\lambda}$ are stored for an initial p_j it becomes likely that much of the time-consuming MCMC sampling is avoided on all subsequent runs when calibrating p_j, thus substantially increasing the speed of the simulations.

4.1. Two prognostic subgroups

In the first simulation study we assumed $K=2$ subgroups, with $j=0$ for Poor (P) and $j=1$ for Good (G) prognosis. Since SEP conducts two separate trials in this case, for this design a maximum of 50 patients were enrolled in each trial. We used the same targeted improvement for both subgroups, $\delta_0=\delta_1=0.15$. All designs were calibrated to have nominal FNR$=0.10$, depending on the design's assumptions. For the S-TI and S-NTI designs, this meant that when $\pi_{E,j}^{\text{true}}=E\{\pi_{S,j}(\boldsymbol{\theta})\}+\delta_j$ the trial should declare E inferior (reject E) at most 10 per cent of the time within subgroup j, regardless of the true response rate in the other subgroup. For the SEP and NS designs, each of the K designs using SEP and the single design using NS had FNR$=$ 0.10. The priors for all four methods were calibrated to be equivalent to having observed 100 historical patients on S ($N_{S,0}=N_{S,1}=100$) and one patient on E ($N_{E,0}=N_{E,1}=1$). For the S-TI, S-NTI, and SEP designs the prior means were $E\{\pi_{E,0}(\boldsymbol{\theta})\}=E\{\pi_{S,0}(\boldsymbol{\theta})\}=0.25$ and $E\{\pi_{E,1}(\boldsymbol{\theta})\}=E\{\pi_{S,1}(\boldsymbol{\theta})\}=0.45$. Since NS ignores subgroups the single prior mean was assumed to be 0.35, the average of the two subgroups, and thus for the NS method *a priori* $\pi_S(\boldsymbol{\theta})\sim\text{Beta}(35, 65)$ and $\pi_E(\boldsymbol{\theta})\sim\text{Beta}(0.35, 0.65)$, with the subscript j omitted because NS ignores subgroups. For SEP, we assumed $\pi_{S,0}(\boldsymbol{\theta})\sim\text{Beta}(25, 75)$, $\pi_{E,0}(\boldsymbol{\theta})\sim\text{Beta}(0.25, 0.75)$, $\pi_{S,1}(\boldsymbol{\theta})\sim\text{Beta}(45, 55)$ and $\pi_{E,1}(\boldsymbol{\theta})\sim$ Beta$(0.45, 0.55)$. Using each of the four methods, we simulated the trial 5000 times under each of the four scenarios. The results are summarized in Table I.

In the 'treatment–subgroup interaction' Scenario 1, E achieves the targeted improvement in the G subgroup but not in the P subgroup. Thus, given desired FNR$=0.10$, a design should reject E at most 10 per cent of the time in the G group but terminate accrual with reasonably high probability in the P subgroup. Both S-TI and SEP do this, although SEP has a smaller stopping probability for the P subgroup (0.65 *versus* 0.73 for S-TI) due to the fact that SEP does not borrow strength between subgroups. In sharp contrast, although S-NTI and NS both are constructed to control FNP$=0.10$, due to the facts that S-NTI ignores treatment–subgroup interactions and NS ignores subgroups entirely, both of these designs have greatly inflated FNP values in the G subgroup, 0.47 with S-NTI and 0.42 with NS, as well as much smaller stopping probabilities

DOI: 10.1002/sim

Table I. Simulation results for a trial with two prognostic subgroups, poor (P) and good (G), $N_{\max}=100$ patients with monitoring in cohorts of size 10.

			Pr(Reject E) (Mean no. of patients)			
Scenario	Subgroup	$\pi^{\text{true}}_{E,Z}$	S-TI	S-NTI	NS	SEP
Scenarios with treatment–subgroup interactions						
1	G	0.60	0.10 (64)	0.47 (38)	0.42 (38)	0.10 (47)
	P	0.25	0.73 (32)	0.50 (38)	0.42 (38)	0.65 (33)
2	G	0.45	0.72 (33)	0.53 (35)	0.41 (38)	0.65 (34)
	P	0.40	0.10 (64)	0.51 (39)	0.41 (38)	0.10 (47)
Scenarios with no treatment–subgroup interaction						
3	G	0.60	0.10 (50)	0.10 (47)	0.10 (47)	0.10 (47)
	P	0.40	0.10 (50)	0.10 (47)	0.10 (47)	0.10 (47)
4	G	0.45	0.76 (38)	0.93 (21)	0.86 (24)	0.65 (34)
	P	0.25	0.78 (37)	0.93 (23)	0.86 (24)	0.65 (33)

In all scenarios, $E\{\pi_{S,P}(\boldsymbol{\theta})\}=0.25$, $E\{\pi_{S,G}(\boldsymbol{\theta})\}=0.45$, $\delta_P=\delta_G=0.15$ and the design is calibrated to have an FNR $=0.10$. Values of $\pi^{\text{true}}_{E,Z}$ equal to the targeted value in subgroup Z are enclosed in boxes.

in the P subgroup, 0.50 with S-NTI and 0.42 with NS. The clear messages here are that either ignoring patient heterogeneity or accounting for patient subgroups but not accounting for possible treatment–subgroup interactions greatly increases the risks of missing an actual treatment advance in a subgroup and of wrongly concluding that a treatment advance exists in a subgroup where it does not. The same messages are given by Scenario 2, where E achieves the targeted improvement in the P subgroup but not in the G subgroup. Another important advantage of the S-TI design is that, because it reliably controls both the FNRs and false-positive rates within subgroups, in the presence of treatment–subgroup interaction S-TI allocates a much larger number of patients to the subgroup where there is an actual treatment advance. For example, in Scenario 1 on average 64 patients are treated in the G subgroup by S-TI, compared with 38 with either S-NTI or NS and 47 with SEP. Thus, S-TI not only maintains the nominal FNR within subgroups but also uses patient resources much more effectively.

In the 'no treatment–subgroup interaction' Scenario 3, E achieves the targeted improvement in both the G and P subgroups and all four methods achieve the nominal 0.10 FNR. In the worst-case Scenario 4, where there is no improvement over S in either subgroup, all four methods have high probabilities of stopping the trial early, with much higher values 0.93 for S-NTI and 0.86 for NS, compared with 0.76–0.78 for S-TI and 0.65 for SEP. Given the very poor performance of S-NTI and NS in the presence of treatment–subgroup interactions, however, the larger stopping probabilities in Scenario 4 hardly make these methods desirable in situations where a treatment–subgroup interaction is likely. The smaller stopping probability with S-TI compared with the simpler S-NTI and NS methods in Scenario 4 is the price that is paid for its greatly superior performance when treatment–subgroup interactions are present.

To compare the four designs over a wider range of $\pi^{\text{true}}_{E,0}$ and $\pi^{\text{true}}_{E,1}$ values, for the null response rates 0.25 in P and 0.45 in G, we simulated cases with $\pi^{\text{true}}_{E,1}=0.60$, the targeted improvement for

DOI: 10.1002/sim

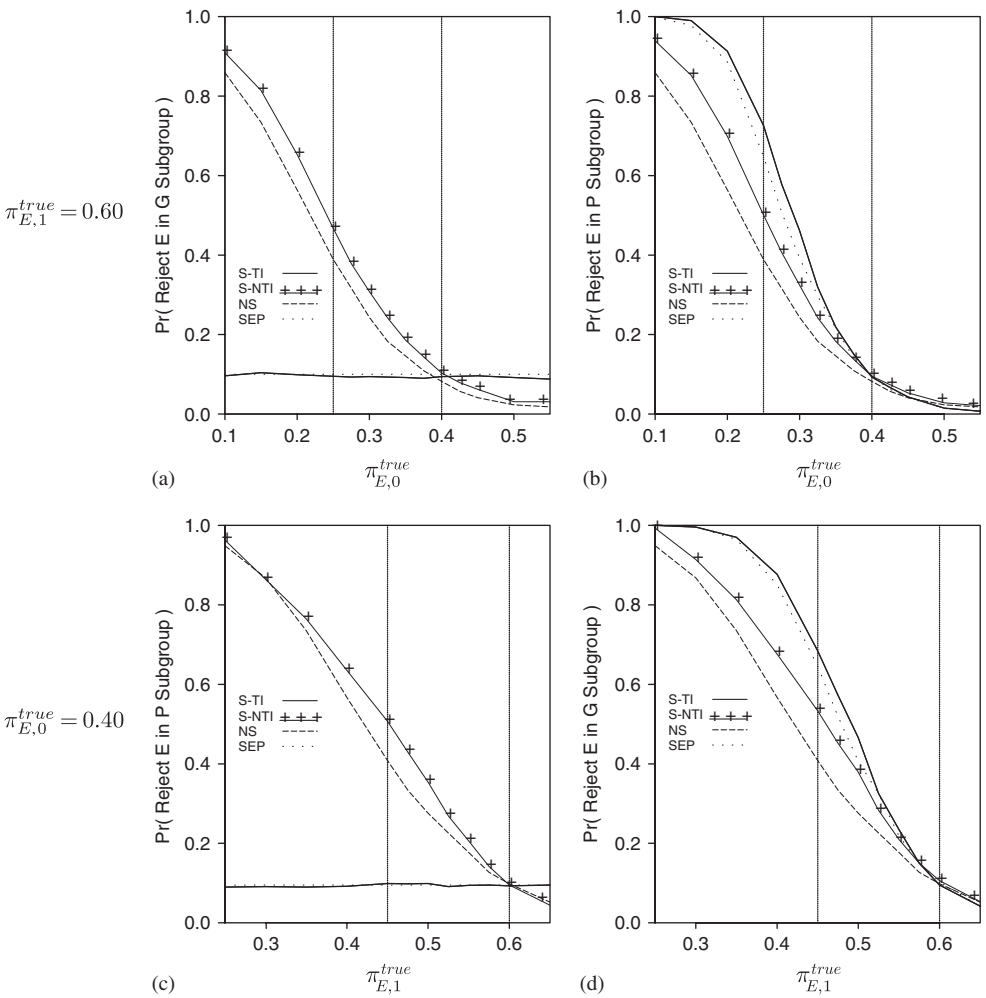

Figure 1. Simulation results when E obtains the targeted improvement in the G subgroup (top row) and P subgroup (bottom row). Vertical lines represent $E\{\pi_{S,0}(\boldsymbol{\theta})\}$ and $E\{\pi_{S,0}(\boldsymbol{\theta})\}+\delta_0$ in the top row and $E\{\pi_{S,1}(\boldsymbol{\theta})\}$ and $E\{\pi_{S,1}(\boldsymbol{\theta})\}+\delta_1$ in the bottom row.

the G group, but varying $\pi_{E,0}^{\text{true}}$ over the range 0.10–0.60. Similarly, we fixed $\pi_{E,0}^{\text{true}}=0.40$ and varied $\pi_{E,1}^{\text{true}}$ over the range 0.20–0.70. The within-subgroup probability of rejecting E, as a function of $\pi_{E,0}^{\text{true}}$, is shown for the G subgroup in Figure 1(a) and for the P subgroup in Figure 1(b). These figures may be interpreted as varying the amount of treatment–subgroup interaction, with the 'no interaction' case corresponding to the values where $\pi_{E,0}^{\text{true}}=0.40$. Figure 1(a) illustrates how poorly both S-NTI and NS perform, since nominally all curves should equal 0.10 for all values of $\pi_{E,0}^{\text{true}}$. In Figure 1(b), larger (smaller) values of Pr(reject E in P subgroup) are more desirable for values of $\pi_{E,0}^{\text{true}}<0.40$ ($\pi_{E,0}^{\text{true}}>0.40$), with all four methods calibrated to stop with probability 0.10 when $\pi_{E,0}^{\text{true}}=0.40$. This figure shows that S-TI and SEP are much more likely to correctly stop trial in

DOI: 10.1002/sim

the P subgroup for small values of $\pi_{E,0}^{\text{true}}$ compared with S-NTI or NS, and that S-TI has the best overall performance. Figures 1(c) and (d) display similar curves, but with the roles of the two subgroups reversed.

4.2. Four prognostic subgroups

For a simulation study with $K=4$ subgroups, we assumed a maximum sample size of 200 patients so that the per-subgroup sample sizes would be comparable to those in the simulations for

Table II. Simulation results for a trial with four prognostic subgroups, $N_{\max}=200$ patients with monitoring in cohorts of size 10.

			Pr(Reject E) (Mean no. of patients)			
Scenario	Subgroup	$\pi_{E,Z}^{\text{true}}$	S-TI	S-NTI	NS	SEP
Scenarios with treatment–subgroup interactions						
1	0	0.25	0.75 (35)	0.74 (30)	0.64 (30)	0.65 (33)
	1	0.45	0.67 (36)	0.67 (31)	0.64 (30)	0.65 (34)
	2	0.70	0.10 (64)	0.64 (34)	0.64 (30)	0.10 (47)
	3	0.75	0.10 (64)	0.70 (30)	0.64 (30)	0.10 (47)
2	0	0.40	0.10 (65)	0.67 (35)	0.64 (30)	0.10 (47)
	1	0.45	0.69 (35)	0.67 (31)	0.64 (30)	0.65 (34)
	2	0.55	0.73 (34)	0.65 (32)	0.64 (30)	0.73 (32)
	3	0.75	0.10 (65)	0.66 (31)	0.64 (30)	0.10 (47)
3	0	0.25	0.74 (39)	0.94 (21)	0.90 (20)	0.65 (33)
	1	0.45	0.73 (38)	0.92 (22)	0.90 (20)	0.65 (34)
	2	0.55	0.74 (37)	0.91 (22)	0.90 (20)	0.73 (32)
	3	0.75	0.10 (83)	0.92 (22)	0.90 (20)	0.10 (47)
Scenarios with no treatment–subgroup interactions						
4	0	0.40	0.10 (50)	0.10 (47)	0.10 (46)	0.10 (47)
	1	0.60	0.10 (50)	0.10 (47)	0.10 (46)	0.10 (47)
	2	0.70	0.10 (50)	0.10 (47)	0.10 (46)	0.10 (47)
	3	0.75	0.10 (50)	0.10 (47)	0.10 (46)	0.10 (47)
5	0	0.25	0.83 (42)	0.99 (16)	0.99 (14)	0.65 (33)
	1	0.45	0.76 (46)	0.99 (16)	0.99 (14)	0.65 (34)
	2	0.55	0.81 (41)	0.99 (16)	0.99 (14)	0.73 (32)
	3	0.60	0.82 (41)	0.99 (16)	0.99 (14)	0.75 (31)

In all scenarios, the respective null values are 0.25, 0.35, 0.55, 0.60 in the four subgroups, all $\delta_j = 0.15$ and the design is calibrated to have an FNR $=0.10$. Values of $\pi_{E,j}^{\text{true}}$ equal to the targeted value in subgroup j are enclosed in boxes.

DOI: 10.1002/sim

$K=2$. As before, since SEP conducts four separate trials, a maximum of 50 patients per trial were enrolled. We assumed the same targeted improvement for all subgroups, $\delta_0=\delta_1=\delta_2=\delta_3=0.15$, and all designs were calibrated to have FNR$=0.10$, as in the study with $K=2$, and the same accrual rate of 30 patients per year was used. For all the four methods, the priors were calibrated to have ESS of 100 for S ($N_{S,0}=N_{S,1}=N_{S,2}=N_{S,3}=100$) and 1 for E ($N_{E,0}=N_{E,1}=N_{E,2}=N_{E,3}=1$). Subgroup-specific null values $\hat{\pi}_{S,0}=0.25$, $\hat{\pi}_{S,1}=0.45$, $\hat{\pi}_{S,2}=0.55$, and $\hat{\pi}_{S,3}=0.60$ were assumed. Since NS ignores subgroups its prior mean response rates were assumed to be 0.46, the average over the four subgroups, and thus $\pi_S(\boldsymbol{\theta})\sim$Beta$(46,54)$ and $\pi_E(\boldsymbol{\theta})\sim$Beta$(0.46,0.54)$. For SEP, we assumed $\pi_{S,0}(\boldsymbol{\theta})\sim$Beta$(25,75)$, $\pi_{E,0}(\boldsymbol{\theta})\sim$Beta$(0.25,0.75)$, $\pi_{S,1}(\boldsymbol{\theta})\sim$Beta$(45,55)$, $\pi_{E,1}(\boldsymbol{\theta})\sim$Beta$(0.45,0.55)$, $\pi_{S,2}(\boldsymbol{\theta})\sim$Beta$(55,45)$, $\pi_{E,2}(\boldsymbol{\theta})\sim$Beta$(0.55,0.45)$, $\pi_{S,3}(\boldsymbol{\theta})\sim$Beta$(60,40)$ and $\pi_{E,3}(\boldsymbol{\theta})\sim$Beta$(0.60,0.40)$. The results of this simulation study are given in Table II.

In Scenario 1, E achieves the targeted improvement in subgroups 2 and 3 but not in 0 or 1. Thus, it is desirable to have small rejection rates for E in subgroups 2 and 3 but large rates in 0 and 1. While both S-TI and SEP maintain the FNR$=0.10$ within subgroups 2 and 3, S-NTI incorrectly rejects E 64 and 70 per cent of the time for groups 2 and 3, respectively, and NS incorrectly rejects E 64 per cent of the time for both subgroups. It thus appears that the probability of S-NTI and NS incorrectly stopping the trial and rejecting E in subgroups where E actually provides an advance over S increases substantially as the number of subgroups increases. In subgroup 0, S-TI has the largest chance of correctly rejecting E, 75 per cent compared with 74, 64 and 65 per cent for S-NTI, NS and SEP, respectively. In subgroup 1, all methods have approximately the same chance of rejecting E. In this case, S-TI reliably detects that E has not met the targeted improvement in subgroups 0 and 1 and terminates patient accrual to these groups, consequently allowing more patients in subgroups 2 and 3 to be treated with an effective treatment. Thus, the substantive conclusions are the same as those for the case $K=2$, but the advantages of S-TI over the other methods are greater for $K=4$. In Scenario 2, E achieves the targeted improvement in subgroups 0 and 3 but not in 1 or 2. The rest of the results are substantively similar to those of Scenario 2.

In Scenario 3, E only achieves the targeted improvement in subgroup 3. While both S-TI and SEP maintain FNR$=0.10$ in subgroup 3, S-NTI and NS incorrectly reject E in this subgroup 92 and 90 per cent of the time, respectively. Since the stopping probabilities in subgroups 0, 1 and 2 with S-TI uniformly dominate those of SEP, this scenario provides an extremely strong motivation for accounting for subgroups and treatment–subgroup interaction.

In Scenario 4, E achieves the targeted improvement in all subgroups and thus, by design, all four methods have a 0.10 FNR. In Scenario 5, E fails to achieve the targeted improvement in all subgroups and thus it is desirable for a design to reject E in all subgroups with high probability. Because S-NTI and NS do not include a treatment–subgroup interaction, both procedures combine the information for the subgroups and thus, in this case, are nearly certain to stop early and reject E. However, S-TI has subgroup-specific stopping probabilities ranging from 0.76 to 0.83 in this case.

5. AN APPLICATION TO A TRIAL IN RELAPSED AML

Patients with AML in relapse after an initial complete remission (CR) have second CR rates ranging from 0 to 60 per cent when given 'salvage' therapy. The principal predictor of second CR is first CR duration (CRD1). Patients with CRD1 <1 year have second CR rates of only approximately 10 per cent when given standard chemotherapy (containing cytosine arabinoside,

DOI: 10.1002/sim

'ara-C'), whereas patients with CRD1 $\geqslant$1 year have second CR rates of about 45 per cent when treated similarly. The most effective therapy for relapsed AML appears to be an allogeneic stem cell transplant from an HLA-matched donor (allotx). Since allotx is not feasible for many patients, there is a need to develop new chemotherapy regimens for patients with AML in first relapse. Gemtuzumab ozogamycin (GO) is a combination of an antibody against CD33, an antigen found on the surface of AML cells, and a toxin. The antibody is hypothesized to direct the toxin to the AML cells rather than their normal counterparts. Although approved by the FDA for treatment of AML in first relapse, GO is only slightly more effective than ara-C in this circumstance. However, decitabine, an agent that is believed to allow re-expression of genes that are silenced by AML cells, has been found to increase sensitivity to GO in AML model systems. This finding provides the basis for the current trial, which administers decitabine for 5 days, with GO given on day 5.

We obtained historical data on 171 AML patients in first relapse enrolled between 1994 and 2003 at M.D. Anderson Cancer Center. The empirical overall probability of response to historical treatment was 0.21. However, upon classifying patients by CRD1 as either poor prognosis (P), if CRD1 <52 weeks, or good prognosis (G), if CRD1 $\geqslant$52 weeks, the 118 (69 per cent) patients in the P subgroups had a response rate of 0.11 and 53 (31 per cent) patients in the G subgroup had a response rate of 0.43.

A typical phase II study would employ a Simon's 2-stage (S2S) design [4] and would assume patient homogeneity. However, based on the heterogeneity seen in the historical data we applied our proposed design, using *P-versus-G* as a binary prognostic variable. The goal of the study is to determine if GO can increase the response rate by 0.15 in the G subgroup, from 0.43 to 0.58 and 0.20 in the P subgroup, from 0.11 to 0.31. Using a null response probability equal to the overall

Table III. Simulation results for a trial with two prognostic subgroups, P (first CR duration $\leqslant$52 weeks) and G (first CR duration $>$52 weeks), for 2-stage ($N_{\max}=40$) and multi-stage ($N_{\max}=80$).

			2-Stage $N_{\max}$ = 40 Pr(Reject GO)		$N_{\max}=80$ Pr(Reject GO)
Scenario	Subgroup	$\pi_{E,Z}^{\text{true}}$	S-TI	S2S	S-TI
Scenarios with treatment–subgroup interactions					
1	G	0.58	0.10 (21)	0.75 (10)	0.10 (37)
	P	0.11	0.90 (19)	0.75 (25)	0.96 (25)
2	G	0.43	0.50 (13)	0.26 (11)	0.71 (29)
	P	0.31	0.10 (27)	0.26 (30)	0.10 (39)
Scenarios with no treatment–subgroup interaction					
3	G	0.58	0.10 (13)	0.11 (12)	0.10 (39)
	P	0.31	0.10 (27)	0.11 (30)	0.10 (37)
4	G	0.43	0.50 (20)	0.92 (9)	0.71 (29)
	P	0.11	0.90 (19)	0.92 (22)	0.96 (25)

In all scenarios, $E\{\pi_{S,P}(\boldsymbol{\theta})\}=0.11$, $E\{\pi_{S,G}(\boldsymbol{\theta})\}=0.43, \delta_P=0.2$, $\delta_G=0.15$ and the design is calibrated to have a false-negative rate$=0.10$. Values of $\pi_{E,Z}^{\text{true}}$ equal to the targeted value in subgroup Z are enclosed in boxes. Numbers in parenthesis are the average number of patients enrolled.

DOI: 10.1002/sim

historical rate of 0.21 (ignoring CRD1), both FNR and false-positive rate of 0.10, and a targeted improvement $(0.31\times0.15)+(0.69\times0.20)=0.1845$ to ensure comparability with the S-TI design, an S2S design would enroll 21 patients in the first stage and a maximum of 43 patients. At the end of the first stage, if five or more of the first 22 patients respond the trial would continue to the second stage and would reject (accept) GO at the end of the trial if 12 or fewer (13 or more) patients responded. Again to ensure comparability, we constrained the S-TI design to look at the data at 22 patients and at the end of the trial. Simulation results are given in Table III. In the cases where the targeted improvement is reached only in one subgroup but not the other (Scenarios 1 and 2) S2S is very likely to reject GO and thus miss a substantive treatment advance. However, if GO does not reach the targeted improvement in the G subgroup then S-TI is less likely to reject GO than S2S. Since neither 2-stage design has desirable properties in this trial we constructed an S-TI design with a maximum of 80 patients and monitoring done once each month, controlling all FNRs to be 0.10. Simulation results are given in Table III. This design has superior properties compared with either of the 2-stage designs.

6. DISCUSSION

We have proposed a flexible Bayesian design for monitoring single-arm phase II clinical trials that accounts for prognostic subgroups and treatment–subgroup interactions, with the decision of whether to stop patient accrual made separately within each subgroup. Our simulations show that S-TI maintains the FNR, and in the presence of treatment–subgroup interactions has much more desirable properties than the alternative designs considered. However, in studies where a subgroup–treatment interaction is very unlikely the simpler alternative methods S-NTI or SEP presented here may be sufficient, and more desirable because they are less complex.

Our simulation studies assumed an equal number of patients in both subgroups and this may not be the case for all trials. Thus, we investigated cases with a subgroup imbalance. For example, simulating the four scenarios in Table I with 70 per cent of the patients in the G subgroup and 30 per cent in the P subgroup, both S-TI and NS maintained the nominal FNR. In the treatment–subgroup interaction Scenarios 1 and 2 the results were similar to those in Table I, but in general the stopping probabilities were lower(higher) for the $P(S)$ subgroup, due to smaller(larger) sample size. In Scenario 1, where the targeted improvement was reached only in the G subgroup, S-NTI and NS incorrectly stopped the trial in this subgroup 17–20 per cent of the time. In Scenario 2, where the targeted improvement was reached only in the P group, S-NTI and NS incorrectly stopped the trial in this subgroup nearly 80 per cent of the time. Thus, the pathological behaviors of S-NTI and NS become more likely if the subgroup showing an improvement comprises a smaller proportion of the sample.

In our development we focused on binary data. However, using the likelihood from equation (3) similar procedures can be used in the context of TTE data under an exponential model, the historical data for the TTE case may be summarized by the number of events observed, $N_{S,j}$, the total time on test, Y^{+}, and the historical mean TTE, denoted by $\hat{\mu}_{S,j}$ for each subgroup $j=0,\ldots,K-1$. To accommodate TTE outcomes, the following changes are required in Section 2.3 and the algorithm for calibrating the prior: use an Inverse Gamma(a,b) with mean $b/(a-1)$ for $a>1$ instead of the Beta(a,b); in step 1 set $a_{S,j}=N_{S,j}$ and $b_{S,j}=N_{S,j}*\hat{\mu}_{S,j}$; in steps 2 and 3 replace the constraint $E\{\pi_{S,j}(\boldsymbol{\theta})\}=\hat{\pi}_{S,j}$ by $E\{e^{\eta_{S,Z}\boldsymbol{\theta})}\}=\hat{\mu}_{S,j}$; in step 4 set $a_{E,j}=N_{E,j}$ and $b_{E,j}=N_{E,j}*\hat{\mu}_{S,j}$; in step 5

DOI: 10.1002/sim

replace the constraint $E\{\pi_{E,j}(\boldsymbol{\theta})\}=\hat{\pi}_{S,j}$ by $E\{e^{\eta_{E,Z}(\boldsymbol{\theta})}\}=\hat{\mu}_{S,j}$. In addition, the decision criterion in equation (4) would be $\Pr\{e^{\eta_{E,Z}(\boldsymbol{\theta})}>e^{\eta_{S,Z}(\boldsymbol{\theta})}+\delta_j|\mathscr{D}_n\}$.

ACKNOWLEDGEMENTS

We thank the reviewers for their constructive comments that helped to improve this manuscript.

REFERENCES

1. Taylor JM, Braun TM, Li Z. Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm phase ii design. *Clinical Trials* 2006; **3**:335–348.
2. Gehan EA. The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* 1961; **13**:346–353.
3. Fleming TR. One sample multiple testing procedure for phase ii clinical trials. *Biometrics* 1982; **38**:143–151.
4. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**:1–10.
5. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50**:337–349.
6. Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine* 1995; **16**:1791–1802.
7. Chen TT. Optimal three-stage designs for phase ii cancer clinical trials. *Statistics in Medicine* 1998; **16**:2701–2711.
8. Ibrahim J, Ryan L, Chen MH. Use of historical controls to adjust for covariates in trend tests for binary data. *Journal of the American Statistical Association* 1998; **93**:1282–1293.
9. Thall PF, Sung HG, Estey EH. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *Journal of American Statistical Association* 2002; **97**:29–39.
10. Thall PF, Wooten LH, Shpall EJ. A geometric approach to comparing treatments for rapidly fatal diseases. *Biometrics* 2006; **62**:193–201.
11. Thall PF, Wathen JK, Bekele NB, Champlin RE, Baker LE, Benjamin LH. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* 2003; **22**:763–780.
12. London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine* 1995; **24**:2597–2611.
13. Thall PF, Wooten LE, Tannier T. Monitoring event timer in early phase clinical trials: some practical issues. *Clinical Trials* 2005; **2**:467–478.
14. Nelder JA, Mead R. A simplex method for function minimization. *The Computer Journal* 1965; **7**:308–313.
15. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC: London, Boca Raton, FL, 1996.
16. Wathen JK, Christen JA. Implementation of backward induction for sequentially adaptive clinical trials. *Journal of Computational and Graphical Statistics* 2006; **15**:398–413.

DOI: 10.1002/sim