Biometrics 62, 193–201
March 2006
DOI: 10.1111/j.1541-0420.2005.00434.x

A Geometric Approach to Comparing Treatments for Rapidly Fatal Diseases

Peter F. Thall,[1,*] Leiko H. Wooten,[1] and Elizabeth J. Shpall[2]

[1]Department of Biostatistics and Applied Mathematics, The University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
[2]Department of Blood and Marrow Transplantation, The University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
[*]*email:* rex@mdanderson.org

Summary. In therapy of rapidly fatal diseases, early treatment efficacy often is characterized by an event, "response," which is observed relatively quickly. Since the risk of death decreases at the time of response, it is desirable not only to achieve a response, but to do so as rapidly as possible. We propose a Bayesian method for comparing treatments in this setting based on a competing risks model for response and death without response. Treatment effect is characterized by a two-dimensional parameter consisting of the probability of response within a specified time and the mean time to response. Several target parameter pairs are elicited from the physician so that, for a reference covariate vector, all elicited pairs embody the same improvement in treatment efficacy compared to a fixed standard. A curve is fit to the elicited pairs and used to determine a two-dimensional parameter set in which a new treatment is considered superior to the standard. Posterior probabilities of this set are used to construct rules for the treatment comparison and safety monitoring. The method is illustrated by a randomized trial comparing two cord blood transplantation methods.

Key words: Adaptive design; Bayesian design; Clinical trials; Competing risks; Cord blood transplantation; Simulation.

1. Introduction

When evaluating treatments for rapidly fatal diseases, early therapeutic efficacy often is characterized in terms of an event, "response," which occurs continuously in time. Some examples are engraftment in a blood or marrow cell transplantation (tx) trial, resolution of a life-threatening infection by antibiotics, and complete remission of acute leukemia by chemotherapy. In each of these examples, achieving response is the first therapeutic goal because the hazard of death is initially high but decreases substantially once response is achieved. In many settings, it also is the case that a response achieved more quickly is associated with a lower subsequent hazard of death. For example, Estey, Shen, and Thall (2000) show that, in chemotherapy of acute leukemia, patients who achieve a complete remission more quickly are more likely to have a longer subsequent survival. Let t^* denote a fixed time limit for achieving response, after which the initial therapy may be replaced by a salvage regimen. Denoting the times to response and death by T_R and T_D, early treatment success is the event $S_{t^*} = \{T_R < \min(t^*, T_D)\}$. This requires that the patient survive long enough to respond, and that response occurs by time t^* from the start of therapy.

Although $\pi = \Pr(S_{t^*})$ depends on both T_R and T_D, comparing treatments in terms of π alone may ignore important information. For example, if treatments A and B have similar π_A and π_B, but on average treatment A achieves a response more quickly than B, then A will have a lower overall death rate. This is because, with either treatment, the risk of death decreases once a response is achieved. A statistical comparison based on π_A and π_B would be likely to conclude that the two treatments have similar efficacy, despite the fact that, on average, A has longer overall survival time. If the death rate after response increases with T_R, then the superiority of A over B is even greater.

We propose a Bayesian approach to treatment comparison in this setting based on a two-dimensional parameter consisting of $\pi(\mathbf{Z}) = \Pr(S_{t^*} \mid \mathbf{Z})$ and the conditional mean time to response, $\mu(\mathbf{Z}) = E(T_R \mid T_R < T_D, \mathbf{Z})$, where $\mathbf{Z} = (Z_1, \ldots, Z_q)$ is a vector of patient prognostic covariates. We define $\mu(\mathbf{Z})$ conditionally because response and death without response are competing risks, that is, T_R is observed only if $T_R < T_D$. Our method bases comparisons on these two parameters evaluated at a reference covariate vector, $\mathbf{Z}^*$, chosen by the physician. For convenience, we will denote $\pi(\mathbf{Z}^*)$ by π and $\mu(\mathbf{Z}^*)$ by μ. Evidently, larger π or smaller μ is more desirable. If π and μ are considered together, however, treatment comparison is problematic. When one treatment has both larger π and larger μ than the other, it is not obvious which treatment is preferable. For example, in our application, S_{t^*} is the event that a cord blood tx patient achieves engraftment within 42 days, so $\pi = \Pr\{T_R < \min(42,\ T_D) \mid \mathbf{Z} = \mathbf{Z}^*\}$. Historical data (Shpall et al., 2002) give the means $\pi_0 = 0.69$ and $\mu_0 = 30$ days with a standard cord blood tx method. A new regimen is considered an improvement over the standard if $\mathrm{E}(\pi,\ \mu) = (0.70, 18)$,

which requires engraftment with about the same probability as the standard but on average 12 days sooner. A different improvement is achieved by $(0.90, 30)$, which allows $\mathrm{E}(\mu)$ to remain the same but requires $\mathrm{E}(\pi)$ to increase from 0.69 to 0.90. This illustrates the motivation for our method, which compares treatments by considering (π, μ) together. Initially, several target (π, μ) pairs are elicited from the physician so that, for $\mathbf{Z} = \mathbf{Z}^*$, all elicited pairs embody the same improvement compared to the historical mean, (π_0, μ_0). A curve is fit to the elicited pairs, and the curve is used to determine a two-dimensional region of (π, μ) pairs for which a new treatment is considered superior to the standard. This region is used to compute posterior probabilities that form the basis for treatment comparison and safety monitoring.

We establish a probability model in Section 2. The method for constructing a two-dimensional parameter set for comparing treatments in terms of (π, μ) is described in Section 3. In Section 4, we present decision rules based on this construction and a design for trial conduct. The method is illustrated in Section 5 by application to a randomized trial to compare two cord blood tx strategies. We close with a discussion in Section 6.

2. Probability Models

2.1 *A Competing Risks Model*

Since response and death without response cannot both occur in the same patient, these events are competing risks. To account for this while also allowing the hazard of death to change when response occurs, we assume the following piecewise model. Let T_1 denote the time to death without response and, among the patients who respond, let T_2 be the time from response to subsequent death. Aside from administrative censoring, for each patient either T_1 or the pair (T_R, T_2), but not both, may be observed, and the patient's survival time is either $T_D = T_1$ if $T_1 < T_R$, or $T_D = T_R + T_2$ if $T_1 \geq T_R$. We assume that T_1 and (T_R, T_2) are independent. To accommodate the common case where T_2 is stochastically decreasing in T_R, we model the conditional distribution of T_2 given T_R. For $k = R, D$, 1, or 2, denote the probability density function (p.d.f.), cumulative distribution function (c.d.f.), and survivor function (s.f.) of T_k by f_k, F_k, and $\mathcal{F}_k = 1 - F_k$. Denote the conditional p.d.f., c.d.f., and s.f. of T_2 given T_R by $f_{2|R}$, $F_{2|R}$, and $\mathcal{F}_{2|R}$. Let $I(A)$ be the indicator of the event A. We assume that the conditional p.d.f. of T_D given T_R is the piecewise distribution

$$f_{D|R}(x \,|\, y) = \mathcal{F}_1(y) f_{2|R}(x - y \,|\, y) I(x \geq y) + f_1(x) I(x < y), \quad x, y > 0. \tag{1}$$

The joint distribution $f_{D,R}(x, y) = f_{D|R}(x \,|\, y) f_R(y)$ thus is determined by $f_{2|R}$, f_1, and f_R, and averaging over f_R yields the marginal survival time distribution,

$$f_D(x) = \int_{y=0}^{x} f_R(y) \mathcal{F}_1(y) f_{2\,|\,R}(x - y \,|\, y)\, dy + \mathcal{F}_R(x) f_1(x), \quad x > 0. \tag{2}$$

Under this piecewise model, the hazard of death changes from $h_1(x) = f_1(x)/\mathcal{F}_1(x)$ for $x < T_R$ to $h_{2|R}(x - T_R \,|\, T_R) = f_{2|R}(x - T_R \,|\, T_R)/\mathcal{F}_{2|R}(x - T_R \,|\, T_R)$ for $x \geq T_R$.

The two key parameters that will form the basis for our design strategy are

$$\pi = \Pr(T_R < t^* \wedge T_D) = \int_{y=0}^{t^*} f_R(y) \mathcal{F}_1(y)\, dy, \tag{3}$$

and

$$\mu = E(T_R \,|\, T_R < T_1) = \frac{\displaystyle\int_0^\infty y f_R(y) \mathcal{F}_1(y)\, dy}{\displaystyle\int_0^\infty f_R(y) \mathcal{F}_1(y)\, dy}. \tag{4}$$

The facts that π and μ are determined by (f_1, f_R) but do not involve $f_{2|R}$, π and μ characterize different aspects of (f_1, f_R), and the hazard of death decreases at T_R, together motivate using (π, μ) to characterize early treatment efficacy. Moreover, the pair (π, μ) provides more information than π alone about T_1 and T_R. While the ultimate therapeutic goal is to make $\pi = 1$ and T_2 very large with high probability, that is, to cure the patient, this goes far beyond the goals of the sort of trials that we have in mind.

Let T^o be the patient's last follow-up time and denote the indicators $Y_D = I(T^o = T_D)$ that the time of death is observed and $Y_R = I(T_R < T^o)$ that the time of response is observed. Each patient's likelihood contribution may take one of four possible forms, depending on whether (T_R, T_D), (T_R, T^o), T_D, or T^o is observed. Accounting for these four possibilities, the likelihood may be written in the general form

$$\mathcal{L} = \{f_R(T_R) \mathcal{F}_1(T_R)\}^{Y_R} \left\{\mathcal{F}_R(T^o) f_1(T^o)^{Y_D} \mathcal{F}_1(T^o)^{1-Y_D}\right\}^{1-Y_R} \times \left\{f_{2|R}\big(T^o - T_R \,|\, T_R\big)^{Y_D} \mathcal{F}_{2|R}\big(T^o - T_R \,|\, T_R\big)^{1-Y_D}\right\}^{Y_R}. \tag{5}$$

For any given parametric model, we denote by $\boldsymbol{\theta}_R, \boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ the parameter vectors characterizing f_R, f_1, and $f_{2\,|\,R}$. Denote the first two terms in (5) involving the distributions of T_R and T_1 by $\mathcal{L}_{R,1}(\mathrm{data} \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_R)$, and denote the third term involving the distribution of $T_2 \,|\, T_R$ by $\mathcal{L}_{2|R}(\mathrm{data} \,|\, \boldsymbol{\theta}_2)$. Since $\mathcal{L} = \mathcal{L}_{R,1}(\mathrm{data} \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_R) \times \mathcal{L}_{2|R}(\mathrm{data} \,|\, \boldsymbol{\theta}_2)$, one may compute the maximum likelihood estimators (MLEs) and posteriors of $(\boldsymbol{\theta}_R, \boldsymbol{\theta}_1)$ and $\boldsymbol{\theta}_2$ separately, which reduces the dimensionality of the computational requirements for fitting data. Since π and μ both are determined by f_R and f_1 and our clinical trial design is based on (π, μ), only the likelihood $\mathcal{L}_{R,1}(\mathrm{data} \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_R)$ and the posterior of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_R)$ given the interim data are required for computing adaptive decision rules during the trial.

2.2 *Parametric Models*

We will consider two families of parametric distributions for f_R, f_1, and $f_{2|R}$. The first is the lognormal, which has a nonmonotone hazard function, and the second is the Weibull, which has a monotone hazard and includes the exponential as a special case. For real η and $\sigma > 0$, we denote the lognormal distribution with p.d.f.

$$f(x \,|\, \eta, \sigma) = \frac{1}{x\, \sigma (2\pi)^{1/2}} \exp[-\{\log(x) - \eta\}^2 / 2\sigma^2], \tag{6}$$

by $\mathrm{LN}(\eta, \sigma^2)$. This has median e^η, mean $e^{\eta + \sigma^2/2}$, and variance $e^{2\eta + \sigma^2}(e^{\sigma^2} - 1)$. Recall that $\mathbf{Z}^*$ is the physician's reference

covariate vector. For each outcome $k = 1$, 2, or R, we account for covariate effects by $\boldsymbol{\beta}_k(\mathbf{Z} - \mathbf{Z}^*) = \beta_{k,1}(Z_1 - Z_1^*) + \cdots + \beta_{k,q}(Z_q - Z_q^*)$, where $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,q})$. We will index the two experimental treatments in the trial by $j = 1$, 2 and the historical treatment by $j = H$. For each j, we denote the linear terms $\eta_{j,k} = \alpha_{j,k} + \boldsymbol{\beta}_k(\mathbf{Z} - \mathbf{Z}^*)$ for $k = 1$, R, and $\eta_{j,2} = \alpha_{j,2} + \boldsymbol{\beta}_2(\mathbf{Z} - \mathbf{Z}^*) + \beta_{2,q+1}\, g(T_R)$, where g is a suitable transformation, such as log or the identity. Under the lognormal model, for a patient with covariates $\mathbf{Z}$ randomized to treatment arm j, we assume that f_R, f_1, and $f_{2|R}$ are given by

$$T_R \,|\, j, \mathbf{Z} \sim \mathrm{LN}\big(\eta_{j,R},\ \sigma^2_{j,R}\big), \tag{7}$$

$$T_1 \,|\, j, \mathbf{Z} \sim \mathrm{LN}\big(\eta_{j,1},\ \sigma^2_{j,1}\big), \tag{8}$$

$$T_2 \,|\, j, \mathbf{Z}, T_R \ \sim \mathrm{LN}\big(\eta_{j,2},\ \sigma^2_{j,2}\big). \tag{9}$$

For $\lambda > 0$ and $\phi > 0$, we denote the Weibull distribution with p.d.f.

$$f(x \,|\, \lambda, \phi) \ = \ \phi\lambda^{-\phi} t^{\phi-1} e^{-(t/\lambda)^\phi}, \tag{10}$$

by Weib(λ, ϕ). This distribution has median $\lambda\{\log(2)\}^{1/\phi}$, mean $\lambda\,\Gamma(1 + \phi^{-1})$, and variance $\lambda^2\{\Gamma(1 + 2\phi^{-1}) - \Gamma^2(1 + \phi^{-1})\}$. Under the Weibull, we will assume the same regression structures as in (7)–(9), but with each $\mathrm{LN}(\eta_{j,k}, \sigma^2_{j,k})$ replaced by a Weib($e^{\eta_{j,k}}$, $\phi_{j,k}$).

Under both models, $\boldsymbol{\beta}_k$ accounts for the covariate effects on T_k for $k = 1$, 2, and R, and $\beta_{2,q+1}$ accounts for the effect of the time to achieve a response on the patient's subsequent survival time. The parameters characterizing f_R, f_1, and $f_{2|R}$ under treatment j are $\boldsymbol{\theta}_{j,R} = (\alpha_{j,R}, \boldsymbol{\beta}_R, \sigma_{j,R})$, $\boldsymbol{\theta}_{j,1} = (\alpha_{j,1}, \boldsymbol{\beta}_1, \sigma_{j,1})$, and $\boldsymbol{\theta}_{j,2} = (\alpha_{j,2}, \boldsymbol{\beta}_2, \beta_{2,q+1}, \sigma_{j,2})$. For each j and k, since $\eta_{j,k} = \mathrm{E}_j\{\log(T_k)\}$ under the lognormal, and $\eta_{j,k} = \mathrm{E}_j\{\log(T_k)\} - \log\{\Gamma(1 + \phi_{j,k}^{-1})\}$ under the Weibull, with both models a larger value of T_k is associated with larger $\eta_{j,k}$. When evaluated at the reference covariate vector $\mathbf{Z}^*$, the $\eta_{j,k}$'s in (7)–(9) reduce to $\alpha_{j,R}$, $\alpha_{j,1}$, and $\alpha_{j,2} + \beta_{2,q+1} T_R$. If $\beta_{2,q+1} < 0$ then T_2 is stochastically decreasing in T_R. If $\beta_{2,q+1} = 0$, then T_2 and T_R are independent. If $\beta_{2,q+1} > 0$, then T_2 is stochastically increasing in T_R, although this case does not correspond to any of the medical settings that we have in mind, since longer T_R typically shortens T_2 in treatment of rapidly fatal diseases. While this is the case if $\Pr(\beta_{2,q+1} < 0)$ is reasonably large, even if $\beta_{2,q+1} \equiv 0$ the hazard of death should still drop at T_R due to an improvement in the patient's medical status. In the three examples given in Section 1, this would be recovery of white cell count to a minimally functional level, absence of infection, and absence of leukemia, respectively.

Under either parametric model, $\pi_j = \pi(\boldsymbol{\theta}_{j,R}, \boldsymbol{\theta}_{j,1}, \mathbf{Z}^*)$ and $\mu_j = \mu(\boldsymbol{\theta}_{j,R}, \boldsymbol{\theta}_{j,1}, \mathbf{Z}^*)$ are both highly nonlinear functions of the elements of $\boldsymbol{\theta}_{j,R}$ and $\boldsymbol{\theta}_{j,1}$. Because they are functions of the same parameters, π_j and μ_j are not independent. The mapping $(\boldsymbol{\theta}_{j,R}, \boldsymbol{\theta}_{j,1}) \longrightarrow (\pi_j, \mu_j)$ reduces the $(4 + 2q)$-dimensional parameter vector under treatment j to a two-dimensional parameter whose elements have a natural clinical interpretation. This will provide a basis for our use of $(\pi_1,\ \mu_1)$ and $(\pi_2,\ \mu_2)$ for treatment comparison.

3. Characterizing Treatment Differences

3.1 *Methods Based on Multivariate Outcomes*

There are many approaches to the problem of comparing treatments based on multidimensional outcomes. Most methods are based on hypothesis tests, and involve a test statistic that differentially weights the outcomes (O'Brien, 1984; Pocock, Geller, and Tsiatis, 1987; Tang, Gnecco, and Geller, 1989), or specifies the alternative geometrically (Willan and Pater, 1985; Jennison and Turnbull, 1993; Bryant and Day, 1995; Conaway and Petroni, 1996; Thall and Cheng, 1999; Kosorok, Shi, and DeMets, 2004). Thall, Simon, and Shen (2000) compared treatments by partitioning the parameter space into four sets characterizing the desirability of the two treatments and computing the posterior probabilities of these sets.

3.2 *Constructing a Desirable Parameter Set*

Because we characterize the efficacy of treatment j by the two-dimensional parameter $(\pi_j,\ \mu_j)$, treatment comparison requires a further dimension reduction to obtain a one-dimensional decision criterion. We begin by eliciting a set of fixed target (π, μ) pairs from the physician, so that each elicited pair embodies the same desired improvement over $(\pi_0,\ \mu_0)$ for $\mathbf{Z} = \mathbf{Z}^*$. The elicited values, $(\pi_1^*,\ \mu_1^*), \ldots, (\pi_M^*,\ \mu_M^*)$, are used to obtain a smooth curve, D, which is the boundary of a set of desirable (π, μ) pairs in the two-dimensional parameter space. We construct D by treating the elicited values like data and using conventional least squares (LS) to fit μ as a function of π. The fitted function may be a line $\hat{\mu}_\pi = \mathrm{a} + \mathrm{b}\pi$, a quadratic $\hat{\mu}_\pi = \mathrm{a} + \mathrm{b}\pi + \mathrm{c}\pi^2$, or some other simple function, where $\hat{\mu}_\pi$ denotes the LS estimator of μ at π. The particular function should provide a reasonably good fit to the elicited pairs, provided that $d\hat{\mu}_\pi/d\pi \geq 0$ for $\pi \in [\underline{\pi},\ \bar{\pi}]$. We define the *target contour* to be the fitted curve $D = \{(\pi,\ \hat{\mu}_\pi) :\ \underline{\pi} \leq \pi \leq \bar{\pi}\}$. To facilitate the elicitation process, it is useful to plot the elicited pairs as they are specified, as well as D. The domain $[\underline{\pi},\ \bar{\pi}]$ should be the interval of π values where the physician considers D to be a valid representation of equivalent targeted improvements over $(\pi_0,\ \mu_0)$, and this may simply be the range of $\{\pi_1^*, \ldots, \pi_M^*\}$. If desired, rather than using a smooth curve, D could be a continuous, nondecreasing piecewise linear function on $[\underline{\pi},\ \bar{\pi}]$, similar to the polygonal boundaries of the alternatives used in the two-dimensional hypothesis testing settings discussed by Conaway and Petroni (1996) in the one sample case, and by Thall and Cheng (1999) in the context of randomized trials.

DEFINITION 1: Given target contour D, the *set of desirable parameter pairs* is

$$\Theta_D = \{(\pi,\ \mu) :\ \pi \geq \pi' \text{ and } \mu \leq \mu' \text{ for some } (\pi',\ \mu') \in D \text{ and } \underline{\pi} \leq \pi \leq \bar{\pi}\}. \tag{11}$$

That is, Θ_D is the set of all (π, μ) pairs at least as desirable as a pair on D. If $0 < \underline{\pi}$, then Θ_D is bounded on the left by the vertical line segment from $(\underline{\pi}, \hat{\mu}_{\underline{\pi}})$ to $(\underline{\pi}, 0)$, and if $\bar{\pi} < 1$ then Θ_D is bounded above by the horizontal line segment from $(\bar{\pi}, \hat{\mu}_{\bar{\pi}})$ to $(1, \hat{\mu}_{\bar{\pi}})$. Figure 1 illustrates the historical mean, elicited pairs, fitted curve D, and set $\boldsymbol{\Theta}_D$ for the cord blood tx trial, for which $\underline{\pi} = 0.60$ and $\bar{\pi} = 1$.

4. Trial Design and Conduct

4.1 *Decision Rules*

Our rules for comparing treatments 1 and 2 will be based on the differences $\boldsymbol{\delta}_{1,2} = (\delta^\pi_{1,2},\ \delta^\mu_{1,2}) = (\pi_1 - \pi_2,\ \mu_1 - \mu_2)$ and

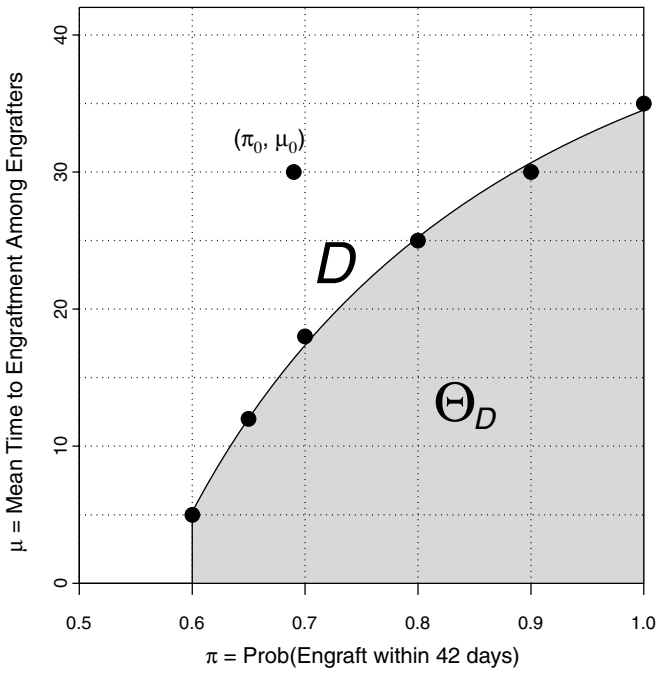

Figure 1. The historical posterior mean and six elicited target values of (π, μ), and the fitted target curve D and set Θ_D of desirable parameter pairs. All values correspond to a standard patient age of 38 years.

$\boldsymbol{\delta}_{2,1} = -\boldsymbol{\delta}_{1,2}$, and we compare experimental treatment j to the historical treatment in terms of $\boldsymbol{\delta}_{j,H} = (\delta^{\pi}_{j,H}, \delta^{\mu}_{j,H}) = (\pi_j - \pi_H, \mu_j - \mu_H)$, for $j = 1, 2$. We do this by evaluating the posterior probabilities that these differences are in the shifted set $\Theta_D - (\pi_0, \mu_0)$, which has the same relationship to $(0, 0)$ as Θ_D has to (π_0, μ_0). Denote the historical data by data_H, and the interim data obtained from the first n patients in the trial by data_n, for $1 \le n \le N$. Our design requires the following two decision rules:

1. Safety monitoring. Terminate treatment arm $j = 1$ or 2 if
$$\Pr\{\boldsymbol{\delta}_{j,H} \in \Theta_D - (\pi_0, \mu_0) \,|\, \text{data}_n, \text{data}_H\} < p_L. \quad (12)$$
2. Treatment comparison. At the end of the trial, select treatment 1 if
$$\Pr\{\boldsymbol{\delta}_{1,2} \in \Theta_D - (\pi_0, \mu_0) \,|\, \text{data}_N\} > \Pr\{\boldsymbol{\delta}_{2,1} \in \Theta_D - (\pi_0, \mu_0) \,|\, \text{data}_N\}, \quad (13)$$
and select treatment 2 if this inequality holds with $\boldsymbol{\delta}_{1,2}$ and $\boldsymbol{\delta}_{2,1}$ reversed.

The stopping rule (12) may be applied continuously, each time a new patient is accrued, or group-sequentially at a prespecified sequence of interim times or sample sizes. In some trials, if one arm is terminated by (12) then the investigator may wish to treat all remaining patients, up to N, on the remaining arm. In other trials, if the criterion (12) is met by either arm then it may be appropriate to stop the entire trial. The trial's operating characteristics (OCs) may be evaluated by simulation, with the values of p_L or N modified on that basis to obtain a design with desirable properties. The OCs consist of the selection probabilities, early stopping probabilities, and sample sizes of the two treatment arms under a set of fixed values of (π_1, μ_1) and (π_2, μ_2) that represent a range of different clinical scenarios.

4.2 *Establishing Priors and Design Parameters*

We begin by assuming a noninformative prior on $\boldsymbol{\theta}_H = (\boldsymbol{\theta}_{H,R}, \boldsymbol{\theta}_{H,1}, \boldsymbol{\theta}_{H,2})$, computing $p(\boldsymbol{\theta}_H \,|\, \text{data}_H)$, and using this posterior to establish priors on the two experimental treatments' parameters, as follows. Since patients will be randomized, the priors on $(\boldsymbol{\theta}_{1,R}, \boldsymbol{\theta}_{1,1})$ and $(\boldsymbol{\theta}_{2,R}, \boldsymbol{\theta}_{2,1})$, should be identical, so to simplify notation we temporarily drop the first subscript $j = 1, 2$. For convenience, we consider the lognormal model, since the regime for establishing a prior under the Weibull is similar. To utilize historical information on how $\mathbf{Z}$ may affect T_R and T_1, we use the posterior $\text{p}(\boldsymbol{\beta}_R, \boldsymbol{\beta}_1 \,|\, \text{data}_H)$ as the prior on $(\boldsymbol{\beta}_R, \boldsymbol{\beta}_1)$ for the trial. This relies on the assumption that the covariate effects in the trial will be the same as seen historically, which makes sense if the patient populations are reasonably similar. Otherwise, it may be more realistic to assume a noninformative prior on $(\boldsymbol{\beta}_R, \boldsymbol{\beta}_1)$. Because $(\alpha_R, \alpha_1, \sigma_R, \sigma_1)$ characterize (π, μ) and we wish the data to dominate all inferences and interim decisions, we require the multivariate normal prior on $\{\alpha_R, \alpha_1, \log(\sigma_R), \log(\sigma_1)\}$ to be uninformative. We thus set the means and correlations of this prior equal to their posterior values given data_H, but inflate the posterior variances by setting $\text{var}(\alpha_k) = c\ \text{var}(\alpha_k \,|\, \text{data}_H)$ and $\text{var}(\sigma_k) = c^{1/2}\ \text{var}(\sigma_k \,|\, \text{data}_H)$ for $k = 1, R$, where c is a suitably large positive number.

The decision rules, cutoff p_L, and prior together determine the design's OCs. One may calibrate the prior and p_L together by considering an array of suitable (c, p_L) pairs, for reasonably large values of the variance multiplier c. This may be done by simulating the trial for each (c, p_L) pair with $(\pi_1, \mu_1) = (\pi_0, \mu_0)$, the historical posterior means, and (π_2, μ_2) varying over several fixed alternatives. As an additional check, since a beta(a, b) distribution has mean $p = a/(a + b)$, variance $p(1 - p)/(a + b + 1)$, and effective sample size $a + b$, it is useful to compute the approximate effective prior sample size, $p(1 - p)/\text{var}_c(\pi) - 1$, obtained by equating the prior $\text{var}_c(\pi)$ to this beta variance for each value of c considered. This provides a basis for choosing c and p_L so that, together, they yield a design with desirable properties while also ensuring that the prior is reasonably noninformative.

4.3 *Numerical Methods*

Posteriors were computed using the iterative defensive importance sampling method of Owen and Zhou (1999). At each iteration, this method requires choosing the posterior mode of $\mathcal{L}(\text{data} \,|\, \boldsymbol{\theta})\ \text{prior}(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ and then computing the gradient at the mode. We used the method of Nelder and Mead (1965) to determine the mode at each iteration. The numerical integrations required to compute μ and π as defined by (3) and (4) were done using the method of Takahasi and Mori (1974). All programming was done in `C++`. Computer programs for implementing the method are available from the second author on request.

5. Application

5.1 *A Cord Blood Transplantation Trial*

In blood or bone marrow cell tx for treatment of leukemia and other cancers, high-dose chemotherapy is given to kill the tumor cells, but it also destroys all of the patient's normal bone marrow cells. In conventional allogeneic tx, bone marrow or peripheral blood progenitor cells (PBPCs) from a human leukocyte antigen (HLA)-matched donor are given after the chemotherapy to restore the patient's marrow with healthy cells. These include neutrophils (white blood cells) to fight infection, red blood cells, and platelets. Response in tx therapy occurs when the absolute neutrophil count (ANC) in the patient's peripheral blood is $\geq$500 cells per mm^3, an event called engraftment, which indicates that the transplanted cells have begun functioning normally in the bone marrow. Because tx carries a substantial risk of fatal infection due to low ANC, and engraftment is essential to give the patient a chance of long-term survival, it is critical to achieve engraftment as soon as possible. When an HLA-matched donor cannot be found, an alternative is to use umbilical cord blood, which has been used since 1988 to treat patients worldwide for a variety of diseases (Kurtzberg et al., 1996). The major disadvantage of cord blood tx is its low cell dose, which results in a longer time to engraftment. Whereas bone marrow and PBPC recipients engraft in approximately 14–17 days, cord blood recipients often take 25–30 days to engraft and have higher rates of engraftment failure and death due to complications, such as infections and bleeding.

We illustrate our method with a small-scale clinical trial to compare two experimental strategies for increasing the cell dose in cord blood tx. Both strategies involve cord blood expansion, in which the cord blood cells are manipulated ex vivo (outside the patient's body) to grow more cells prior to tx. A different approach is double cord blood tx, in which unexpanded cord blood grafts obtained from two different placentae are infused simultaneously (Rubinstein et al., 1998). Patients are randomized to receive either two unexpanded ($j = 1$), or one unexpanded plus a second, expanded cord blood graft ($j = 2$). There is one predictive covariate, the patient's age, and $Z^* = 38$ years is the reference value used for eliciting the target pairs and computing the decision criteria. The key parameters are thus the probability of engraftment within 42 days, $\pi_j = \Pr_j(T_R < 42 \wedge T_1 \,|\, Z = 38)$, and the conditional mean time to engraftment, $\mu_j = \mathrm{E}_j(T_R \,|\, T_R < T_1, Z = 38)$ for $j = 1, 2$.

5.2 *Analysis of the Historical Data*

To establish a standard for comparison in the trial being planned, we first analyze a data set (Shpall et al., 2002) including the times to engraftment and death in 37 patients who received an allogeneic cord blood tx using an ex vivo expansion method similar to those in the trial being planned. Of the 37 patients, 30 (81.1%) engrafted ($Y_R = 1$), with 28 of 37 (75.7%) engrafting within 42 days, and seven died without engraftment ($Y_R = 0$ and $Y_D = 1$). Among the 30 patients who engrafted, the mean time to engraftment was 28.1 days with variance 75.8 and range 15 to 49. For the seven patients who died without engraftment, the mean survival time was 35.7 days with variance 451.6 and range 14 to 74. Indexing the patients who engrafted by $i = 1, \ldots, 30$, the early outcome likelihoods in these two groups are $\mathcal{L}_R = \prod_{i=1}^{30} f_R(T_{i,R} \,|\, Z_i)\mathcal{F}_1(T_{i,R} \,|\, Z_i)$ and $\mathcal{L}_1 = \prod_{i=31}^{37} \mathcal{F}_R(T_{i,1} \,|\, Z_i) f_1(T_{i,1} \,|\, Z_i)$. After engraftment, 18 of 30 (60%) died ($Y_R = 1$ and $Y_D = 1$), and the values of T_2 for the remaining 12 patients were censored ($Y_R = 1$ and $Y_D = 0$). Indexing the 18 patients who died after engraftment by $i = 1, \ldots, 18$, the respective likelihood contributions of these 30 subsequent post-transplant observations are $\mathcal{L}_{R,2} = \prod_{i=1}^{18} f_{2|R}(T_{i,2} - T_{i,R} \,|\, T_{i,R},\, Z_i)$ and $\mathcal{L}_{R,c} = \prod_{i=19}^{30} \mathcal{F}_{2|R}(T_i^o - T_{i,R} \,|\, T_{i,R}, Z_i)$. No patient had both T_R and T_1 censored, so the case $Y_R = 0$ and $Y_D = 0$ did not occur. Thus, the likelihood for the historical data is $\mathcal{L} = \mathcal{L}_R \times \mathcal{L}_1 \times \mathcal{L}_{R,2} \times \mathcal{L}_{R,c}$. For these data, only $Z =$ patient age is predictive of outcome, $Z^* = 38$ years was chosen as the reference age, and there was only one treatment. Under either parametric model, there are three linear terms, $\eta_1(Z) = \alpha_1 + \beta_1(Z - 38)/10$, $\eta_R(Z) = \alpha_R + \beta_R(Z - 38)/10$, and $\eta_2(Z, T_R) = \alpha_2 + \beta_{2,1}(Z - 38)/10 + \beta_{2,2} \log(T_R)$. Under the lognormal model, $\boldsymbol{\theta}_R = (\alpha_R, \beta_R, \sigma_R)$, $\boldsymbol{\theta}_1 = (\alpha_1, \beta_1, \sigma_1)$, and $\boldsymbol{\theta}_2 = (\alpha_2, \beta_{2,1}, \beta_{2,2}, \sigma_2)$, and under the Weibull each σ_k is replaced by ϕ_k. Since $\mathcal{L}_R \times \mathcal{L}_1$ is parameterized by $(\boldsymbol{\theta}_R, \boldsymbol{\theta}_1)$ and $\mathcal{L}_{R,2} \times \mathcal{L}_{R,c}$ by $\boldsymbol{\theta}_2$, these two subvectors may be treated separately when computing MLEs and posteriors.

For the historical data, the maximized log likelihood is –10.02 under the lognormal and –15.01 under the Weibull. Since both models have 10 parameters, we chose the lognormal for our analysis of data$_H$ and as a basis for the trial design. For the prior used to analyze the historical data, we assumed that each α_k and β_k was normal with mean 0 and variance 100, and that each $\sigma_k \sim \mathrm{LN}(0, 100)$, with all parameters independent. The posterior is summarized in Table 1. In these 37 patients, older age was predictive of a longer time to engraftment, with $\Pr(\beta_R > 0 \,|\, \mathrm{data}_H) = 0.86$, and moderately predictive of a quicker time to death without engraftment, with $\Pr(\beta_1 < 0 \,|\, \mathrm{data}_H) = 0.75$. Among patients who engrafted, older age was strongly predictive of a quicker time to death after engraftment, with $\Pr(\beta_{2,1} < 0 \,|\, \mathrm{data}_H) = 0.96$, and a longer time to achieve engraftment was moderately predictive of a shorter subsequent survival time, with $\Pr(\beta_{2,2} < 0 \,|\, \mathrm{data}_H) = 0.80$. The posterior means of the two design parameters for a 38-year-old patient were $\mathrm{E}(\pi \,|\, \mathrm{data}_H) = 0.69$ and $\mathrm{E}(\mu \,|\, \mathrm{data}_H) = 30$ days and, as might be expected from their definitions, π and μ were negatively correlated. Figure 2 gives the posterior medians and 95% credible intervals of π and μ as functions of age. Although there was substantial variability due to the small sample size, the plots show that older patients were less likely to engraft and, given that they did engraft, on average they required a longer time to achieve engraftment. We used $\log(T_R)$ in $\eta_2(Z, T_R)$ since it was moderately predictive of T_2, whereas the untransformed T_R was not predictive, with $\Pr(\beta_{2,2} < 0 \,|\, \mathrm{data}_H) = 0.52$.

5.3 *Prior, Trial Design, and Simulations*

The trial has a maximum of $N = 60$ patients. The scientific goal is to select the better of the two treatments, provided that it does not compare unfavorably with the historical treatment. Patients are randomized between the two treatments arms, but if one arm is terminated early by (12) then all subsequent patients are treated on the remaining arm. Following the

Table 1
Fit of the lognormal model to the historical data on 37 *cord blood transplant patients*

Outcome	Covariate	Parameter	Posterior values given data_H	
			Mean (SD)	$\Pr(\beta > 0)$
Time to	–	α_R	−2.483 (0.069)	–
engraftment (T_R)	age	β_R	0.041 (0.038)	0.86
	–	σ_R	0.375 (0.045)	–
Time to death	–	α_1	−1.938 (0.169)	–
without engraftment (T_1)	age	β_1	−0.067 (0.100)	0.25
	–	σ_1	0.556 (0.112)	–
Time to death	–	α_2	−1.530 (1.213)	–
after engraftment (T_2)	age	$\beta_{2,1}$	−0.487 (0.276)	0.04
	$\log(T_R)$	$\beta_{2,2}$	−0.409 (0.479)	0.20
	–	σ_2	2.304 (0.423)	–
Pr(engraftment in 42 days)	–	π	0.69 (0.07)	$\text{corr}(\pi, \mu) = -0.39$
$\text{E}(T_R \mid \text{engraft})$	–	μ	29.8 (2.24)	–

Note: Patient age is coded in each linear component as (age − 38)/10.

general method described in Section 4.2, we initially examined the nine designs obtained for all pairs of $c = 10$, 15, 20 and $p_L = 0.01$, 0.05, 0.10. These three values of c give $\text{var}(\mu) = 60$, 100, 140 and $\text{var}(\pi) = 0.039$, 0.051, 0.060, so the approximate effective sample sizes based on a beta distribution having the same mean and variance as π are 4.5, 3.2, and 2.6 patients. For each (c, p_L) pair, we simulated the trial under each of six scenarios defined in terms of fixed values of (π_1, μ_1) and (π_2, μ_2). These are given in Table 2, which summarizes the design's OCs for the three scenarios with $c = 15$. The numerical results for $c = 10$ and 20 are not tabled to conserve space. Since $\text{E}(T_1) = 99$, 125, 164 days for these three values of c, and 125 was considered a very large value for T_1, to avoid unrealistically large values of T_1 occurring with nontrivial probability we chose $c = 15$ to determine the prior variability.

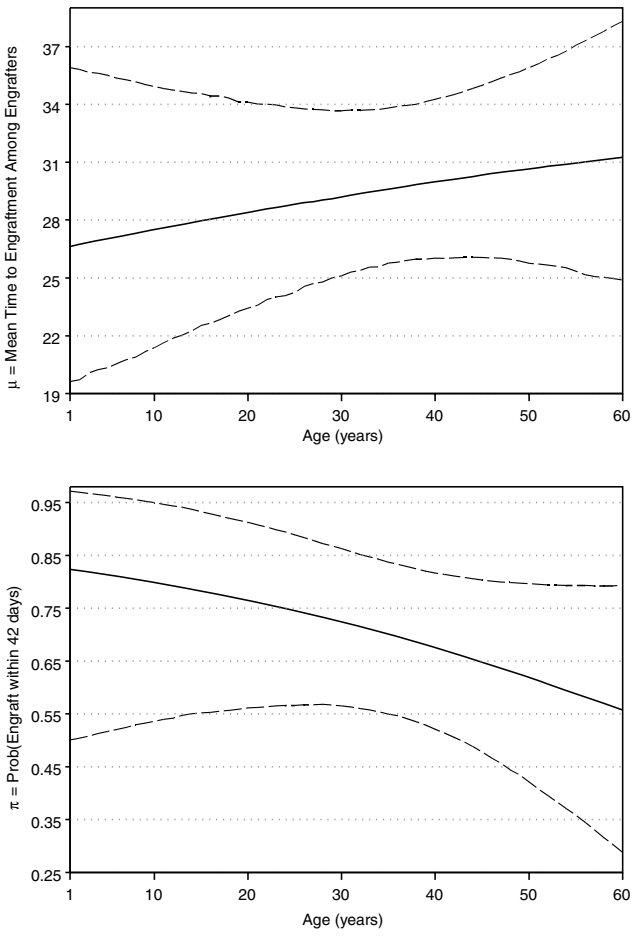

Figure 2. Estimated posterior median and 95 percent credible intervals for π and μ as functions of patient age, based on the historical data.

For the simulations, we generated the data as follows. In all scenarios studied, we fixed $(\pi_1, \mu_1) = (0.69, 30)$, the null values, and fixed (π_2, μ_2) at null or alternative values depending on the scenario. Within each arm, suppressing j, we first fixed $\boldsymbol{\beta}$ and $\text{var}(T_k \mid Z = Z^*) = e^{\alpha_k + \sigma_k^2/2}(e^{\sigma_k^2} - 1)$ for $k = 1$, R at their historical posterior means and, given (π, μ), solved for $(\alpha_1, \alpha_R, \sigma_1, \sigma_R)$. We simulated each patient's outcomes (T_1, T_R) from the lognormal distributions determined by these fixed parameter values, with age sampled from a smoothed density based on the historical data. For (π, μ) near the boundaries of Θ_D, with either $\mu = 5$ or $\pi \geq 0.70$ and $\mu \geq 35$, we incrementally reduced $\text{var}(T_R \mid Z = Z^*)$ until the distributions of (T_R, T_1) yielded the given (π, μ). Each case was simulated 1000 times.

Denote the scenarios in Table 2 by $S_1, \ldots, S_6$. The null case, S_1, is given by $(\pi_2, \mu_2) = (\pi_1, \mu_1) = (0.69, 30)$, the historical posterior mean. Arm 2 is unsafe in S_2, where $\mu_2 = 40$ is unacceptably large, and in S_3, where $\pi_2 = 0.55$ is unacceptably small. In S_4 and S_5, which have (π_2, μ_2) values on the target curve D, arm 2 is preferable to arm 1. In S_6, where $(\pi_2, \mu_2) = (0.90, 18)$ is in the interior of Θ_D, arm 2 is highly preferable to arm 1. Table 2 shows that $p_L = 0.01$ gives a design with early stopping probabilities for arm 2 that are far too small in S_2 and S_3, where (π_2, μ_2) are unsafe. The cutoffs $p_L = 0.05$ and 0.10 both give safe designs, but the selection probabilities for arm 2 under S_4 and S_5, where arm 2 is more desirable, were higher for $p_L = 0.05$, so this was chosen as the cutoff. These results illustrate the general fact that smaller p_L gives higher selection probabilities and smaller stopping probabilities. Thus, p_L must be chosen to balance safety with the ability to select desirable treatments.

Table 2
Stopping and selection percentages and sample sizes for true $(\pi_1, \mu_1) = (0.69, 30)$ *and six different true values of* (π_2, μ_2)

			True values of (π_2, μ_2)					
	Arm		1 0.69, 30	2 0.70, 40	3 0.55, 30	4 0.70, 18	5 0.90, 30	6 0.90, 18
$p_L = 0.01$	1	Stopped (%)	8	9	7	8	9	7
		Selected(%)	51	90	79	5	7	0
		No. of patients	30	35	34	29	28	29
	2	Stopped (%)	8	[46]	[34]	2	0	0
		Selected (%)	48	6	17	[95]	[93]	[100]
		No. of patients	30	24	25	31	32	31
$p_L = 0.05$	1	Stopped (%)	49	54	55	46	46	46
		Selected (%)	38	46	44	6	7	0
		No. of patients	26	34	30	23	22	22
	2	Stopped (%)	49	[100]	[90]	11	3	0
		Selected (%)	36	0	5	[87]	[91]	[99]
		No. of patients	26	7	17	35	37	38
$p_L = 0.10$	1	Stopped (%)	79	83	81	71	71	70
		Selected (%)	18	17	19	8	6	0
		No. of patients	20	23	21	18	17	16
	2	Stopped (%)	78	[100]	[98]	25	12	1
		Selected (%)	18	0	0	[73]	[85]	[99]
		No. of patients	19	5	10	35	40	44

Note: Correct decision percentage is enclosed in box.

An additional safety rule imposed by FDA reviewers terminates an arm if none of the first three patients in that arm engraft. We applied this rule to each arm 42 days after the third patient was accrued, with no patients enrolled in the interim. Thereafter, the trial may continue with (12) applied at 30-day intervals until the last patient has been accrued. However, it appears that (12) with $p_L = 0.05$ subsumes the FDA rule. If the first three patients die at day 74 without engrafting, then the criterion probability in (12) is 0.0003. If all engraft at day 49, the historical maximum and 7 days later than the maximum allowed 42 days in S_t^*, then the criterion probability is 0.015. If all three engraft at day 40, three late successes, then the criterion probability is 0.075. Thus, $p_L = 0.05$ would stop the arm in the first two cases and continue in the third, all in agreement with the FDA rule.

Figure 3 gives plots of the early stopping and selection percentages of arm 2 as fixed values of (π_2, μ_2) are varied over Θ_D, with $(\pi_1, \mu_1) = (0.69, 30)$ in all cases. The early stopping percentages (Figure 3a) increase sharply as (π_2, μ_2) moves away from the target curve D. The selection percentages (Figure 3b) show the opposite pattern, with high selection percentages for desirable (π_2, μ_2) pairs. While early stopping and selection are disjoint events for each treatment arm, they are not complementary since it may be the case that an arm is neither stopped early nor selected. Thus, the two plots are not redundant.

To examine the design's sensitivity to N, we repeated the simulations for $N = 90$, 120, and 150. These showed that the correct decision probabilities increase with N. For example, under S_3 the correct early stopping percentage for arm 2 increases from 90%, when $N = 60$, to 99% when $N = 150$. Under S_4, S_5, and S_6, when $N = 150$ the correct selection percentages for arm 2 increase to 88%, 93%, and 100%, respectively.

5.4 *Robustness*

To examine the method's sensitivity to deviations from the assumed underlying model, we repeated the simulations under the six scenarios in Table 2 with $p_L = 0.05$, but generating the data from each of two qualitatively different distributions for (T_R, T_1). The first is a Weibull as described in Section 2. The second is a mixture of two lognormals, which might result from the effects of an unobserved binary covariate. To construct the mixtures, we replaced $\mathrm{E}(T_1)$ and $\mathrm{E}(T_R)$ by $\mathrm{E}(T_1) + \Delta_1$ and $\mathrm{E}(T_R) - \Delta_R$ with probability 1/2, and by $\mathrm{E}(T_1) - \Delta_1$ and $\mathrm{E}(T_R) + \Delta_R$ with probability 1/2. Thus, the latent variable was either beneficial or harmful with probability 1/2 each. For each scenario, to obtain mixture distributions of T_1 and T_R yielding the given (π, μ) in each arm, we performed a two-dimensional search in (Δ_1, Δ_R). This yielded shift parameters in the ranges $4 \le \Delta_1 \le 7$ and $1 \le \Delta_R \le 3$ for the six scenarios. The additional simulations are summarized in Table 3. Comparing these results to those in Table 2 with $p_L = 0.05$, under the Weibull the method is less likely to terminate an arm early in most scenarios, and the decisions for arm 2 are less reliable under S_3 and S_4 but numerically identical under S_2, S_5, and S_6. Under the mixture distributions, the method is, generally, more likely to terminate an arm early, but has higher correct selection probabilities for arm 2 under S_4 and S_5. Thus, the method is sensitive to deviations from the assumed model, but still maintains good OCs in all of the cases studied.

6. Discussion

A potential criticism of our method is that the elicited target (π, μ) pairs are inherently subjective. However, the conventional method of targeting a fixed improvement $\delta > 0$ from π to $\pi + \delta$ in the one-dimensional case based on π alone relies

Table 3
Robustness studies

			True values of (π_2, μ_2)					
			1	2	3	4	5	6
	Arm		0.69, 30	0.70, 40	0.55, 30	0.70, 18	0.90, 30	0.90, 18
Weibull	1	Stopped (%)	34	40	40	34	34	34
		Selected (%)	43	60	59	17	8	0
		No. of patients	28	38	32	25	24	23
	2	Stopped (%)	34	[100]	[76]	11	3	0
		Selected (%)	42	0	9	[79]	[91]	[100]
		No. of patients	28	7	18	33	36	36
Mixture model	1	Stopped (%)	84	87	87	78	75	78
		Selected (%)	14	13	13	2	5	0
		No. of patients	19	22	19	16	17	15
	2	Stopped (%)	83	[100]	[98]	11	7	1
		Selected (%)	14	0	1	[89]	[89]	[99]
		No. of patients	18	5	11	40	43	45

Note: Stopping and selection percentages and sample sizes for true $(\pi_1, \mu_1) = (0.69, 30)$ and six different true values of (π_2, μ_2), for T_1 and T_R following either a Weibull distribution or a mixture of lognormal distributions. Correct decision percentage is enclosed in box.

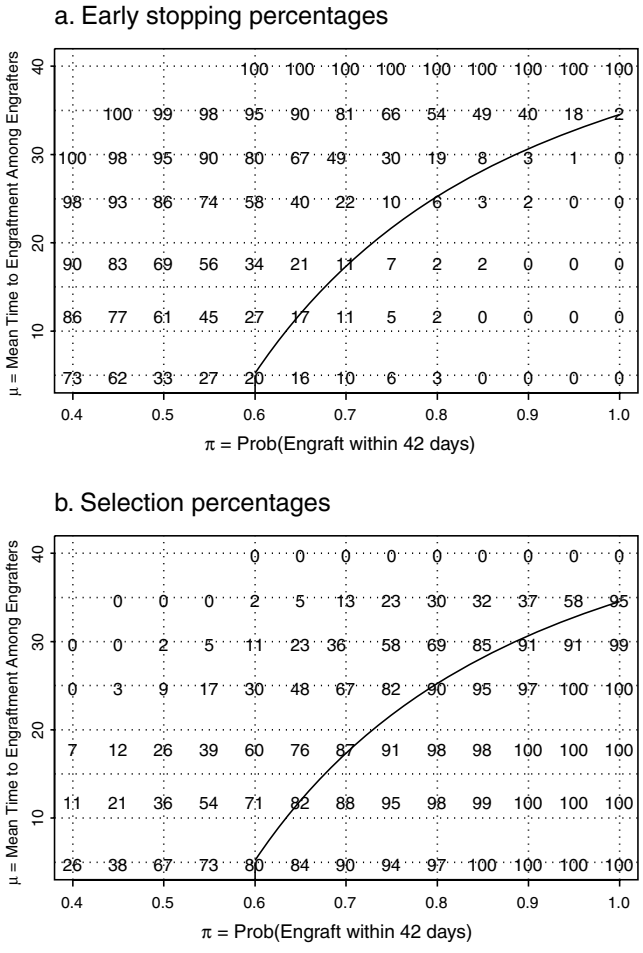

Figure 3. Early stopping percentages (a) and selection percentages (b) for treatment arm 2 under six different fixed values of (π_2, μ_2) when the fixed parameter values (π_1, μ_1) for arm 1 equal the historical posterior mean $(0.69, 30)$.

on a similarly subjective value of δ. In any case, because the target (π, μ) values play a critical role in our method, these should be elicited carefully. Our use of a contour constructed from elicited values as a basis for decision making is similar to the method of Thall, Sung, and Estey (TSE; 2002), who also begin by determining a target contour in a two-dimensional parameter space. A fundamental difference between the two approaches is that TSE use the target contour to generate a family of contours, and they base treatment comparison on the posterior expected value of a utility function that varies numerically over the contours. In contrast, we fix a single target contour, use it to define a region of desirable parameter pairs, and rely on the posterior probabilities of this region, given in (12) and (13), as decision criteria.

Acknowledgements

The authors are grateful to a referee and an associate editor, whose comments led to a substantial improvement in the manuscript. We also thank Peter Mueller for helpful discussions and advice. Dr Thall's work was partially supported by NCI grant RO1-CA-83932. Dr Shpall's work was partially supported by NCI grant RO1-CA-75163.

References

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* **51,** 1372–1383.

Conaway, M. R. and Petroni, G. R. (1996). Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics* **52,** 1375–1386.

Estey, E. H., Shen, Y., and Thall, P. F. (2000). Effect of time to complete remission on subsequent survival and disease-free survival time in AML, RAEB-t and RAEB. *Blood* **95,** 72–77.

Jennison, C. and Turnbull, B. W. (1993). Group sequential tests for bivariate response: Interim analyses of clinical

trials with both efficacy and safety endpoints. *Biometrics* **49,** 741–752.

Kosorok, M. R., Shi, Y., and DeMets, D. L. (2004). Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics* **60,** 134–145.

Kurtzberg, J., Laughlin, M., Graham, M. L., Smith, C., Olson, J. F., Halperin, E. C., Ciocci, G., Carrier, C., Stevens, C. E., and Rubinstein, P. (1996). Placental blood as a source of hematopoietic stem cells for transplantation into unrelated recipients. *New England Journal of Medicine* **335,** 157–166.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7,** 308.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40,** 1079–1087.

Owen, A. and Zhou, Y. (1999). Safe and effective importance sampling. *Journal of the American Statistical Association* **95,** 135–140.

Pocock, S. J., Geller, N. J., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43,** 487–498.

Rubinstein, P., Carrier, C., Scaradavou, A., et al. (1998). Outcomes among 562 recipients of placental-blood transplants from unrelated donors. *New England Journal of Medicine* **339,** 1565–1577.

Shpall, E. J., Quinones, R., Giller, R., et al. (2002). Transplantation of ex vivo expanded cord blood. *Biology of Blood and Marrow Transplantation* **8,** 368–376.

Takahasi, H. and Mori, M. (1974). Double exponential formulas for numerical integration. *Publication RIMS Kyoto University* **9,** 721–741.

Tang, D. I., Gnecco, C., and Geller, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association* **84,** 776–779.

Thall, P. F. and Cheng, S. C. (1999). Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics* **55,** 746–753.

Thall, P. F., Simon, R. M., and Shen, Y. (2000). Approximate Bayesian evaluation of multiple treatment effects. *Biometrics* **56,** 213–219.

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.

Willan, A. R. and Pater, J. L. (1985). Hypothesis testing and sample size for bivariate binomial response in the comparison of two groups. *Journal of Chronic Diseases* **38,** 603–608.

Received February 2005. *Revised May* 2005.
Accepted June 2005.