A Geometric Approach to Comparing Treatments for Rapidly Fatal Diseases

Peter F. Thall,1,* Leiko H. Wooten,1 and Elizabeth J. Shpall2

1Department of Biostatistics and Applied Mathematics, The University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
2Department 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 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 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 TR and TD, early treatment success is the event S(t∗) = { TR < min(t∗, TD) }. This requires that the patient survive long enough to respond, and that response occurs by time t∗ from the start of therapy.

Although π = Pr(S(t∗)) depends on both TR and TD, comparing treatments in terms of π alone may ignore important information. For example, if treatments A and B have similar πA and πB, 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 πA and πB 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 TR, 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 π and μ, which is the conditional mean time to response, μ(Z) = E(TR | TR < TD, Z), where Z = (Z1, . . . , Zq) is a vector of patient prognostic covariates. We define μ(Z) conditionally because response and death without response are competing risks, that is, TR is observed only if TR < TD. Our method bases comparisons on these two parameters evaluated at a reference covariate vector, Z−, chosen by the physician. For convenience, we will denote π(Z−) by π and μ(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 π = Pr(TR < min(42, TD) | Z = Z−). Historical data (Shpall et al., 2002) give the means πA = 0.69 and μA = 30 days with a standard cord blood tx method. A new regimen is considered an improvement over the standard if E(π, μ) = (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 \( E(\mu) \) to remain the same but requires \( E(\pi) \) to increase from 0.69 to 0.90. This illustrates the motivation for our method, which compares treatments by considering \((\pi, \mu)\) together. Initially, several target \((\pi, \mu)\) pairs are elicited from the physician so that, for \( Z = Z^* \), all elicited pairs embody the same improvement compared to the historical mean, \((\pi_0, \mu_0)\). A curve is fit to the elicited pairs, and the curve is used to determine a two-dimensional region of \((\pi, \mu)\) 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 \((\pi, \mu)\) 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 \((T_2, T_2)\), 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 \( (T_2, T_2) \) 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 \( k = R, D, 1, 2 \), 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 \( S_k = 1 - F_k \). 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 \( S_{2|R} \). 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

\[
f_{D|R}(x \mid y) = F_1(y) f_{2|R}(x-y \mid y) I(x \geq y) + f_1(x) I(x < y), \quad x, y > 0.
\]

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

\[
f_D(x) = \int_{y=0}^{x} f_R(y) F_1(y) f_{2|R}(x-y \mid y) dy + F_1(x) f_1(x), \quad x > 0.
\]

Under this piecewise model, the hazard of death changes from \( h_1(x) = f_1(x)/F_1(x) \) for \( x < T_R \) to \( h_{2|R}(x-T_R \mid T_R) = f_{2|R}(x-T_R \mid T_R)/F_{2|R}(x-T_R \mid T_R) \) for \( x \geq T_R \).

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

\[
\pi = \Pr(T_R < t' \wedge T_D) = \int_{y=0}^{t'} f_R(y) F_1(y) dy, \tag{3}
\]

\[
\mu = E(T_R \mid T_R < T_1) = \frac{\int_{y=0}^{\infty} y f_R(y) F_1(y) dy}{\int_{y=0}^{\infty} f_R(y) F_1(y) dy}. \tag{4}
\]

The facts that \( \pi \) and \( \mu \) are determined by \((f_1, f_R)\) but do not involve \( f_{2|R}, \pi \) and \( \mu \) characterize different aspects of \((f_1, f_R)\), and the hazard of death decreases at \( T_R \), together motivate using \((\pi, \mu)\) to characterize early treatment efficacy. Moreover, the pair \((\pi, \mu)\) provides more information than \( \pi \) alone about \( T_1 \) and \( T_R \). While the ultimate therapeutic goal is to make \( \pi = 1 \) and \( T_1 \), \( T_R \) 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 \( T^* \) be the patient’s last follow-up time and denote the indicators \( Y_0 = I(T^* = T_R) \) that the time of death is observed and \( Y_R = I(T_R < T^*) \) that the time of response is observed. Each patient’s likelihood contribution may take one of four possible forms, depending on whether \((T_R, T_0), (T_R, T^*), (T_D, T_R), \) or \( T^* \) is observed. Accounting for these four possibilities, the likelihood may be written in the general form

\[
L = \{f_R(T_R) F_1(T_R)\}^{Y_R} \{f_R(T^*) f_1(T^*) \mid f(R(T^*)^{1-Y_0} F_1(T^*)^{1-Y_R})\}^{1-Y_R} \times \{f_{2|R}(T^* - T_R \mid T_R)^{Y_2} F_{2|R}(T^* - T_R \mid T_R)^{1-Y_2}\}^{Y_R}.
\]

For any given parametric model, we denote by \( \theta_R, \theta_1, \) and \( \theta_2 \) 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 \( L_{R,1}(\text{data} \mid \theta_R, \theta_1) \), and denote the third term involving the distribution of \( T_2 \) \( T_R \) by \( L_{R,2}(\text{data} \mid \theta_2) \). Since \( L = L_{R,1}(\text{data} \mid \theta_R, \theta_1) \times L_{R,2}(\text{data} \mid \theta_2) \), one may compute the maximum likelihood estimators (MLEs) and posterior of \((\theta_R, \theta_1)\) and \( \theta_2 \) separately, which reduces the dimensionality of the computational requirements for fitting data. Since \( \pi \) and \( \mu \) both are determined by \( f_R \) and \( f_1 \) and our clinical trial design is based on \((\pi, \mu)\), only the likelihood \( L_{R,1}(\text{data} \mid \theta_R, \theta_1) \) and the posterior of \((\theta_R, \theta_1)\) 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 non-monotone hazard function, and the second is the Weibull, which has a monotone hazard and includes the exponential as a special case. For real \( \eta \) and \( \sigma > 0 \), we denote the lognormal distribution with p.d.f.

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

by LN(\(\eta, \sigma^2\)). This has median \( e^\eta \), mean \( e^{\eta+\sigma^2/2} \), and variance \( e^{2\eta+\sigma^2}(e^\sigma - 1) \). Recall that \( Z^* \) is the physician’s reference
covariate vector. For each outcome \( k = 1, 2, \) or \( R, \) we account for covariate effects by
\[ \beta_j(Z - Z^*) = \beta_{k,1}(Z_1 - Z_1^*) + \cdots + \beta_{k,q}(Z_q - Z_q^*), \]\nwhere \( \beta_j = (\beta_{k,1}, \ldots, \beta_{k,q}). \) 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 \( \eta_{j,k} = \alpha_{j,k} + \beta_j(Z - Z^*) \) for \( k = 1, R, \) and
\[ \eta_{j,2} = \alpha_{j,2} + \beta_j(Z - Z^*) + \beta_{2,q+1} g(T_R), \] where \( g \) is a suitable transformation, as log or the identity. Under the lognormal model, for a patient with covariates \( Z \) randomized to treatment arm \( j, \) we assume that \( f_R, f_1, \) and \( f_{2,R} \) are given by
\[
\begin{align*}
T_R | j, Z &\sim \text{LN}(\eta_{j,R}, \sigma^2_{j,R}), \\
T_1 | j, Z &\sim \text{LN}(\eta_{j,1}, \sigma^2_{j,1}), \\
T_{2,j} | j, Z, T_R &\sim \text{LN}(\eta_{j,2}, \sigma^2_{j,2}).
\end{align*}
\]
For \( \lambda > 0 \) and \( \phi > 0, \) we denote the Weibull distribution with p.d.f.
\[
f(x | \lambda, \phi) = \phi \lambda x^{\phi-1} e^{-\phi x^\lambda}, \tag{10}\]
by \( \text{Weib}(\lambda, \phi). \) 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 \( \text{LN}(\eta_{j,k}, \sigma^2_{j,k}) \) replaced by a \( \text{Weib}(e^{\eta_{j,k}}, \phi_{j,k}). \)

Under both models, \( \beta_k \) 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 \( \Theta_{j,R} = (\alpha_{j,R}, \beta_{j,R}, \sigma^2_{j,R}), \) \( \Theta_{j,1} = (\alpha_{j,1}, \beta_{j,1}, \sigma^2_{j,1}), \) and \( \Theta_{j,2} = (\alpha_{j,2}, \beta_{j,2}, \beta_{2,q+1}, \sigma^2_{j,2}). \) For each \( j \) and \( k, \) since \( \eta_{j,k} = \Theta_j \{ \log(\tau_k) \} \) under the lognormal, and \( \eta_{j,k} = \Theta_j \{ \log(\tau_k) \} - \log(\Gamma(1 + \phi_{j,k})) \) 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 \( Z^*, \) 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, \( \pi_j = \pi(\Theta_{j,R}, \Theta_{j,1}, Z^*) \) and \( \mu_j = \mu(\Theta_{j,R}, \Theta_{j,1}, Z^*) \) are both highly nonlinear functions of the elements of \( \Theta_{j,R} \) and \( \Theta_{j,1}. \) Because they are functions of the same parameters, \( \pi_j \) and \( \mu_j \) are not independent. The mapping \( (\Theta_{j,R}, \Theta_{j,1}) \rightarrow (\pi_j, \mu_j) \) 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 Tsatis, 1987; Tang, Gnecco, and Geller, 1989), or specifies the alternative geometrically (William and Pater, 1985; Jennison and Turnbull, 1993; Bryant and Day, 1995; Conaway and Petroni, 1996; Thall and Cheng, 1999; Kosorok, Shu, 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 \((\pi, \mu)\) pairs from the physician, so that each elicited pair embodies the same desired improvement over \((\pi_0, \mu_0)\) for \( Z = Z^*. \) 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 \((\pi, \mu)\) 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 \( \mu \) as a function of \( \pi. \) The fitted function may be a line \( \hat{\mu}_a = a + b \pi, \) a quadratic \( \hat{\mu}_a = a + b \pi + c \pi^2, \) or some other simple function, where \( \hat{\mu}_a \) denotes the LS estimator of \( \mu \) at \( \pi. \) The particular function should provide a reasonably good fit to the elicited pairs, provided that \( d \hat{\mu}_a / d \pi \geq 0 \) for \( \pi \in [\pi, \bar{\pi}]. \) We define the target contour to be the fitted curve \( D = \{(\pi, \mu_0): \pi \leq \pi \leq \bar{\pi}\}. \) To facilitate the elicitation process, it is useful to plot the elicited pairs as they are specified, as well as \( D. \) The domain \([\pi, \bar{\pi}]\) should be the interval of \( \pi \) 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 \([\pi, \bar{\pi}]\), 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
\[
\Theta_D = \{(\pi, \mu): \pi \geq \pi' \text{ and } \mu \leq \mu' \text{ for some } (\pi', \mu') \in D \text{ and } \pi \leq \pi' \leq \bar{\pi}\}. \tag{11}\]

That is, \( \Theta_D \) is the set of all \((\pi, \mu)\) pairs at least as desirable as a pair on \( D. \) If \( 0 < \pi < \bar{\pi} \), then \( \Theta_D \) is bounded on the left by the vertical line segment from \((\pi, \mu_0)\) to \((\pi, 0)\), and if \( \bar{\pi} < 1 \) then \( \Theta_D \) is bounded above by the horizontal line segment from \((\bar{\pi}, \mu_0)\) to \((1, \mu_0). \) Figure 1 illustrates the historical mean, elicited pairs, fitted curve \( D, \) and set \( \Theta_D \) for the cord blood tx trial, for which \( \pi = 0.60 \) 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 \( \delta_{1,2} = (\delta_{1,2}^1, \delta_{1,2}^2) = (\pi_1 - \pi_2, \mu_1 - \mu_2) \) and...
The historical posterior mean and six elicited target values of \((\pi, \mu)\), and the fitted target curve \(D\) and set \(\Theta_D\) of desirable parameter pairs. All values correspond to a standard patient age of 38 years.

\[\delta_{2.1} = -\delta_{1.2}, \text{ and we compare experimental treatment } j \text{ to the historical treatment in terms of } \delta_{j,H} = (\delta_{j,H}^1, \delta_{j,H}^2) = (\pi_j - \pi_H, \mu_j - \mu_H), \text{ for } j = 1, 2.\]

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 \(\Theta_D\) has to \((\pi_0, \mu_0)\). Denote the historical data by \(\text{data}_H\), and the interim data obtained from the first \(n\) patients in the trial by \(\text{data}_n\), for \(1 \leq n \leq N\). Our design requires the following two decision rules:

1. Safety monitoring. Terminate treatment arm \(j = 1\) or 2 if
   \[\Pr\{\delta_{j,H} \in \Theta_D - (\pi_0, \mu_0) \mid \text{data}_n, \text{data}_H\} < p_L.\]  
   
   (12)

2. Treatment comparison. At the end of the trial, select treatment 1 if
   \[\Pr\{\delta_{1,2} \in \Theta_D - (\pi_0, \mu_0) \mid \text{data}_N\}\]
   \[> \Pr\{\delta_{2,1} \in \Theta_D - (\pi_0, \mu_0) \mid \text{data}_N\},\]
   and select treatment 2 if this inequality holds with \(\delta_{1,2}\) and \(\delta_{2,1}\) 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 \((\pi_1, \mu_1)\) and \((\pi_2, \mu_2)\) that represent a range of different clinical scenarios.

### 4.2 Establishing Priors and Design Parameters

We begin by assuming a noninformative prior on \(\theta_H = (\theta_{H,R}, \theta_{H,1}, \theta_{H,2})\), computing \(p(\theta_H \mid \text{data}_H)\), and using this posterior to establish priors on the two experimental treatments’ parameters, as follows. Since patients will be randomized, the priors on \((\theta_{H,R}, \theta_{1,1})\) and \((\theta_{H,2}, \theta_{2,1})\), 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 \(Z\) may affect \(R\) and \(T\), we use the posterior \(p(\beta_R, \beta_1 \mid \text{data}_H)\) as the prior on \((\beta_R, \beta_1)\) 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 \((\beta_R, \beta_1)\). Because \((\alpha_R, \alpha_1, \sigma_R, \sigma_1)\) characterize \((\pi, \mu)\) 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 \(\text{data}_H\), but inflate the posterior variances by setting \(\text{var}(\alpha_k) = c \text{ var}(\alpha_k \mid \text{data}_H)\) and \(\text{var}(\sigma_k) = c^{1/2} \text{ var}(\sigma_k \mid \text{data}_H)\) for \(k = 1, 2, R\), 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 \((\pi_2, \mu_2)\) 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, \(p(1 - p)/\text{var}_r(\pi) - 1\), obtained by equating the prior \(\text{var}_r(\pi)\) 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

Posterior distributions 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 \(L(\text{data} \mid \theta)\) prior\(\theta\) as a function of \(\theta\) 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 \(\mu\) and \(\pi\) 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 ≥500 cells per mm³, 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 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.

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 \(L_{Ri} = \prod_{j=1}^{30} f_{TR}(T_{Ri}, Z_i)f_{T_i}(T_{Ri}, Z_i)\) and \(L_i = \prod_{j=1}^{37} f_{TR}(T_{Ri}, Z_i)f_{T_i}(T_{Ri}, Z_i)\), respectively.

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, but if one arm is terminated early by (12) then all subsequent patients are treated on the remaining arm. Following the
general method described in Section 4.2, we initially examined the nine designs obtained for all pairs of \( \alpha = 10, 15, 20 \) and \( \beta > 0 \). These three values of \( \alpha \) 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 \( \pi \) 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 \((\pi_1, \mu_1)\) and \((\pi_2, \mu_2)\). These are given in Table 2, which summarizes the design’s OCs for the three scenarios with \( \alpha = 15 \). The numerical results for \( c = 10 \) and 20 are not tabulated to conserve space. Since \( E(T_1) = 99, 125, 164 \) 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 non-trivial probability we chose \( c = 15 \) to determine the prior variability.

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 \((\pi_2, \mu_2)\) at null or alternative values depending on the scenario. Within each arm, suppressing \( j \), we first fixed \( \beta \) and \( \text{var}(T_1 | Z = Z^* = e^{\alpha_k + \sigma_k^2/2}(e^{\alpha_k^2} - 1)) \) for \( k = 1, R \) at their historical posterior means and, given \((\pi, \mu)\), solved for \((\pi_1, \sigma_1, \sigma_R)\). 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 \((\pi, \mu)\) near the boundaries of \( \Theta_D \), with either \( \mu = 5 \) or \( \pi \geq 0.70 \) and \( \mu \geq 35 \), we incrementally reduced \( \text{var}(T_R | Z = Z^*) \) until the distributions of \((T_R, T_1)\) yielded the given \((\pi, \mu)\). 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 \( S_1 \) and \( S_6 \), which have \((\pi_2, \mu_2)\) values on the target curve \( D \), arm 2 is preferable to arm 1. In \( S_5 \), where \((\pi_2, \mu_2) = (0.90, 18)\) is in the interior of \( \Theta_D \), 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 \((\pi_2, \mu_2)\) are unsafe. The cutoffs \( p_L = 0.05 \) and 0.10 both give safe designs, but the selection probabilities for arm 2 under \( S_1 \) 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 1
Fit of the lognormal model to the historical data on 37 cord blood transplant patients

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Covariate</th>
<th>Parameter</th>
<th>Posterior values given data( \mu )</th>
<th>Mean (SD)</th>
<th>Pr(( \beta &gt; 0 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to engraftment ( T_R )</td>
<td>–</td>
<td>( \alpha_R )</td>
<td>–</td>
<td>-2.483 (0.069)</td>
<td>–</td>
</tr>
<tr>
<td>Time to death with engraftment ( T_1 )</td>
<td>–</td>
<td>( \alpha_1 )</td>
<td>–</td>
<td>-1.938 (0.169)</td>
<td>–</td>
</tr>
<tr>
<td>Time to death after engraftment ( T_2 )</td>
<td>age</td>
<td>( \beta_2 )</td>
<td>–</td>
<td>-0.067 (0.100)</td>
<td>0.25</td>
</tr>
<tr>
<td>( \text{Pr(engraftment in 42 days)} )</td>
<td>–</td>
<td>( \pi )</td>
<td>–</td>
<td>0.69 (0.07)</td>
<td>corr(( \pi, \mu )) = -0.39</td>
</tr>
<tr>
<td>( E(T_R</td>
<td>\text{engraft}) )</td>
<td>–</td>
<td>( \mu )</td>
<td>–</td>
<td>29.8 (2.24)</td>
</tr>
</tbody>
</table>

Note: Patient age is coded in each linear component as \( \text{age} - 38/10 \).

Figure 2. Estimated posterior median and 95 percent credible intervals for \( \pi \) and \( \mu \) as functions of patient age, based on the historical data.
with the opposite pattern, with high selection percentages from the target curve increases from 90%, when percentages (Figure 3a) increase sharply as $p = 0.05$ would stop the arm in the first two cases and continue in the third, all in agreement with the FDA rule. 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 $S_1$, then the criterion probability is 0.015. If all three engraft at day 40, three later than the maximum allowed 42 days in $S_1$, then the criterion probability in (12) 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 $(\pi_2, \mu_2)$ are varied over $\Theta_D$, with $(\pi_1, \mu_1) = (0.69, 30)$ in all cases. The early stopping percentages (Figure 3a) increase sharply as $(\pi_2, \mu_2)$ moves away from the target curve $D$. The selection percentages (Figure 3b) show the opposite pattern, with high selection percentages for desirable $(\pi_2, \mu_2)$ 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 $S_1$, 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 $(T_R, T_I)$. 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 $E(T_I)$ and $E(T_R)$ by $E(T_I) + \Delta_1$ and $E(T_R) - \Delta_1$ and $E(T_R) - \Delta_R$ with probability 1/2, and by $E(T_I) - \Delta_1$ and $E(T_R) + \Delta_R$ 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 $T_I$ and $T_R$ yielding the given $(\pi, \mu)$ in each arm, we performed a two-dimensional search in $(\Delta_1, \Delta_R)$. This yielded shift parameters in the ranges $4 \leq \Delta_1 \leq 7$ and $1 \leq \Delta_R \leq 3$ 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 $S_5$ but numerically identical under $S_2$, $S_4$, 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 $S_1$ and $S_6$. 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 $(\pi, \mu)$ pairs are inherently subjective. However, the conventional method of targeting a fixed improvement $\delta > 0$ from $\pi$ to $\pi + \delta$ in the one-dimensional case based on $\pi$ alone relies...
Table 3

Robustness studies

<table>
<thead>
<tr>
<th>Arm</th>
<th>True values of ((\pi_2, \mu_2))</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>0.69, 30</td>
<td>0.70, 40</td>
<td>0.55, 30</td>
<td>0.70, 18</td>
<td>0.90, 30</td>
<td>0.90, 18</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Stopped (%)</td>
<td>34</td>
<td>40</td>
<td>40</td>
<td>34</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Selected (%)</td>
<td>43</td>
<td>60</td>
<td>59</td>
<td>17</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>No. of patients</td>
<td>28</td>
<td>38</td>
<td>32</td>
<td>25</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>Stopped (%)</td>
<td>34</td>
<td>100</td>
<td>76</td>
<td>11</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Selected (%)</td>
<td>42</td>
<td>0</td>
<td>9</td>
<td>79</td>
<td>91</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>No. of patients</td>
<td>28</td>
<td>7</td>
<td>18</td>
<td>33</td>
<td>36</td>
<td>36</td>
</tr>
</tbody>
</table>

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 \((\pi_2, \mu_2)\), for \(T_1\) and \(T_2\) 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 \((\pi_2, \mu_2)\) when the fixed parameter values \((\pi_1, \mu_1)\) for arm 1 equal the historical posterior mean (0.69, 30).

on a similarly subjective value of \(\delta\). In any case, because the target \((\pi, \mu)\) 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


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