STATISTICS IN MEDICINE, VOL. 14, 357–379 (1995)

BAYESIAN SEQUENTIAL MONITORING DESIGNS FOR SINGLE-ARM CLINICAL TRIALS WITH MULTIPLE OUTCOMES

PETER F. THALL

Department of Biomathematics, Box 237, M.D. Anderson Cancer Center, University of Texas, 1515 Holocombe Boulevard, Houston, TX 77030, U.S.A.

RICHARD M. SIMON

Biometric Research Branch, Division of Cancer Treatment, National Cancer Institute, 6130 Executive Boulevard, Rockville, MD 20892, U.S.A.

AND

ELIHU H. ESTEY

Department of Hematology, Box 061, M.D. Anderson Cancer Center, University of Texas, 1515 Holcombe Boulevard, Houston, TX 77030, U.S.A.

SUMMARY

We present a Bayesian approach for monitoring multiple outcomes in single-arm clinical trials. Each patient's response may include both adverse events and efficacy outcomes, possibly occurring at different study times. We use a Dirichlet-multinomial model to accommodate general discrete multivariate responses. We present Bayesian decision criteria and monitoring boundaries for early termination of studies with unacceptably high rates of adverse outcomes or with low rates of desirable outcomes. Each stopping rule is constructed either to maintain equivalence or to achieve a specified level of improvement of a particular event rate for the experimental treatment, compared with that of standard therapy. We avoid explicit specification of costs and a loss function. We evaluate the joint behaviour of the multiple decision rules using frequentist criteria. One chooses a design by considering several parameterizations under relevant fixed values of the multiple outcome probability vector. Applications include trials where response is the cross-product of multiple simultaneous binary outcomes, and hierarchical structures that reflect successive stages of treatment response, disease progression and survival. We illustrate the approach with a variety of single-arm cancer trials, including bio-chemotherapy acute leukaemia trials, bone marrow transplantation trials, and an anti-infection trial. The number of elementary patient outcomes in each of these trials varies from three to seven, with as many as four monitoring boundaries running simultaneously. We provide general guidelines for eliciting and parameterizing Dirichlet priors and for specifying design parameters.

1. INTRODUCTION

Patient response in clinical trials is an inherently multidimensional phenomenon, with the possibility of both adverse and desirable events. In this paper we present a Bayesian approach to the conduct of single-arm trials of experimental treatments in which patient response is multinomial. Single-arm trials range from conventional phase II evaluations of a new drug to studies of

CCC 0277–6715/95/040357–23

Received October 1993
Revised March 1994

complex multi-stage treatment regimens. Such trials frequently are used to determine whether an experimental treatment is sufficiently safe and efficacious to warrant evaluation in a large randomized trial. Our proposed strategy provides a practical framework for monitoring multiple outcomes continuously, based on multiple simultaneous stopping rules that protect future patients against treatments with unacceptably high rates of adverse events or low rates of desirable treatment responses. We incorporate historical data or clinical experience with 'standard' treatment into a multivariate prior distribution on the patient outcome probabilities, and we evaluate the joint operating characteristics of the stopping rules using frequentist criteria. The underlying model and monitoring strategy account for inherent interdependencies among the various outcomes. This extends the designs of Thall and Simon[1,2] from single to multiple outcomes, and thereby accommodates a broad variety of clinical trials. We are motivated by experiences with trials of highly innovative and aggressive treatments for rapidly fatal diseases, such as acute leukaemia, where the primary clinical concern is the trade-off between the chance of improved efficacy and increased risk of adverse treatment effects, such as acute toxocity or death.

Although patient response in any medical setting is multivariate, the basis for statistical methods for the design, monitoring and analysis of clinical trials has generally consisted of a single endpoint. With this approach, typically one relegates all other patient responses to the status of 'secondary' endpoints. As such one may observe them and analyse them informally, for example, to test hypotheses nominally of incidental or secondary importance, but one ignores them in the formal statistical design and associated size, power and sample size computations. Defects of this approach are that it does not provide specific guidelines for safety monitoring, it does not account for the interrelationships among endpoints, it does not account for the effects of monitoring adverse events on inference for a primary efficacy endpoint, and it is not realistic concerning the broad use of multiple endpoints in reporting the results of clinical trials. A review of 67 published clinical trials (Smith *et al.*[3]), which found a mean of 21·7 different endpoints analysed per trial, illustrates the seriousness of this issue. Furthermore, the trade-off between toxicity and improved efficacy is a major issue in the evaluation of most new chemotherapies, and safety concerns are rarely of secondary importance. Standard designs based on a single binary or time-to-event endpoint essentially ignore this fact.

Many clinical settings involve multiple outcomes. A simple example is a cancer chemotherapy trial of an experimental treatment in which the major outcomes are disease remission and acute toxicity, where it is essential to terminate the trial if the observed toxicity rate is too high or the remission rate is too low. We accommodate such settings by providing stopping rules for adverse events to protect future patients, and stopping rules for efficacy events to reduce the probability of continuing a trial of a new treatment unlikely to provide an improvement over standard therapy. These rules help clear the way for testing other, potentially more effective new treatments. Finally, we provide a rule for determining whether an experimental treatment is sufficiently efficacious to warrant termination of a phase II trial and commencement of a large-scale phase III trial. Our approach also accommodates situations where observation of certain endpoints depends conditionally on the occurrence of earlier events. The structure is thus quite general, and it accommodates rather complicated clinical settings where, to our knowledge, no other effective monitoring strategy exists.

Several authors have recently addressed the problem of formulating and testing hypotheses based on multiple endpoints in clinical trials. For settings in which each element of a multivariate response vector is a measure of treatment efficacy, O'Brien[4] examined existing methods and proposed a global test directed at alternative hypotheses that have treatment effects in the same direction, essentially to conserve power. Pocock *et al.*,[5] and Tang *et al.*[6,7] provided extensions, the

latter two papers dealing with group sequential tests. Lehmacher *et al.*[8] extended O'Brien's approach to accommodate a sequence of hypotheses in a closed multiple test procedure. Gelber *et al.*[9] proposed a method for combining toxicity and survival outcomes into a single endpoint.

A limitation of these procedures in settings where both efficacy and adverse events must be monitored is that they combine all outcomes into a single test statistic. One notable exception is the group-sequential testing procedure of Jennison and Turnbull,[10] who propose use of a bivariate test statistic for trials with two outcome variables which characterize different aspects of treatment response. This includes the important case of an efficacy and an adverse outcome. Our monitoring strategy is motivated by similar considerations, with the essential differences that we consider only single-arm trials, an arbitrary number of outcomes may be monitored, the data are monitored continuously, and our framework for constructing stopping rules is Bayesian.

In general, we characterize each patient's outcome as one of K possible elementary events. We use a Dirichlet-multinomial model for the event probabilities and corresponding counts. Continuous variables are discretized. We base stopping boundaries on posterior probabilities of the incidences of adverse and favourable events with the experimental regimen, compared to prior experience with standard therapy. We do not use loss functions or decision theory. Rather, we evaluate the behaviour of the monitoring bounds under fixed values of the multiple outcome probability vector.

Several considerations motivate our use of Bayesian criteria to construct decision rules combined with frequentist evaluation of their operating characteristics under fixed values of the event probabilities. The first is that in general one interprets the results of early clinical trials of a new regimen subjectively based on informal comparison to prior experience with other, standard therapies. The monitoring strategy described in this paper provides a formal basis for this process. Whereas the use of external data or prior opinion is problematic in major randomized trials, it is inherent in the interpretation of early developmental studies.

The second motivation for our approach is that many clinicians involved in the development of improved therapies find themselves comfortable with Bayesian concepts. Clinicians asked to provide a single value of a parameter required to implement a frequentist design often respond by giving a range of values, describing the parameter's distribution along that range and citing data from previous trials. Moreover, we have received extremely positive responses from clinicians at M.D. Anderson Cancer Center to whom we have provided Bayesian designs based upon this approach.

Decision-theoretic methods have seen little practical application in clinical trials, due to the difficulty in quantifying loss functions and the often elaborate mathematical framework. Moreover, the nature of decision-making at the end of a trial is generally difficult to quantify.[11] The use of frequentist criteria to evaluate a Bayesian monitoring design is a scientifically sound and extremely practical alternative to the use of formal decision theory in conjunction with Bayesian probability criteria for monitoring clinical trials. Ho[12] used frequentist criteria to evaluate a group sequential Bayesian rule for comparing two Gaussian samples. Recently, Etzioni and Pepe[13] proposed a Bayesian model for jointly monitoring two adverse outcomes in a clinical trial, combined with the use of frequentist inferences at the end of the trial. Other Bayesian approaches to multiple testing and estimation problems are described by Dixon and Duncan,[14] Louis[15] and Berry.[16,17]

Section 2 presents the general monitoring approach for single-arm trials with multiple discrete outcomes, including descriptions of the Dirichlet-multinomial model, stopping criteria, and guidelines for constructing monitoring boundaries. Section 3 describes five applications that illustrate the general approach. We discuss general issues and extensions in Section 4.

2. THE GENERAL APPROACH

2.1. The Dirichlet–multinomial model

Let $A_1, \ldots, A_K$ denote all possible combinations of patient response, with corresponding category probabilities $\theta = (\theta_1, \ldots, \theta_{K-1})$, and $\theta_K = 1 - \theta_1 - \ldots - \theta_{K-1}$. For example, if one monitors both complete remission (CR) and acute toxicity (TOX) in a cancer chemotherapy trial then, denoting the complement of CR by $\overline{\text{CR}}$, the four elementary response categories are $A_1 = [\text{CR and TOX}]$, $A_2 = [\text{CR and } \overline{\text{TOX}}]$, $A_3 = [\overline{\text{CR}} \text{ and TOX}]$ and $A_4 = [\overline{\text{CR}} \text{ and } \overline{\text{TOX}}]$. We consider only trials in which it is reasonable to treat patient response as discrete. In particular, we accommodate continuous variables by discretizing them, for example, replacing the time T of disease progression by the indicator of the event $[T \geqslant s]$ for a particular fixed s, or more generally by $[s_1 \leqslant T < s_2]$ and $[T \geqslant s_2]$ if $s_1 < s_2$ are clinically important times.

Let $i = 1, 2, \ldots$ index patients, $j = 1, \ldots, K$ index the categories of response, and $t = \text{E}, \text{S}$ index treatment, where E denotes the experimental and S the standard treatment. Our first model assumption is that (1) conditional on θ_{E} the observed patient responses are independent with Pr [patient i has outcome A_j when treated with E] $= \theta_{\text{E},j}$, for all i and j. This implies in particular that the response rates, while random, do not change in some systematic manner during the course of the trial, which might occur due to a change in some aspect of treatment or supportive care. We denote by $X_{n,j}$ the number of patients out of the first n scored who experience elementary outcome A_j. Conditional on θ_{E}, the vector $\mathbf{X}_n = (X_{n,1}, \ldots, X_{n,K})$ follows a multinomial distribution in n and θ_{E}. Our second model assumption is (2) *a priori*, θ_{E} and θ_{S} follow independent Dirichlet distributions $\text{Dir}(\mathbf{a}_t) = \text{Dir}(a_{t,1}, \ldots, a_{t,K})$, $t = \text{E}, \text{S}$, written $\theta_t \sim \text{Dir}(\mathbf{a}_t)$ for brevity. Denoting the probability vector $\mathbf{p} = (p_1, \ldots, p_{K-1})$, with $p_K = 1 - p_1 - \ldots - p_{K-1}$, the Dir(**a**) PDF is

$$f(\mathbf{p}; \mathbf{a}) = \frac{\Gamma(a_1 + \ldots + a_K)}{\Gamma(a_1) \ldots \Gamma(a_K)} p_1^{a_1 - 1} \ldots p_K^{a_K - 1},$$

where each $p_j \geqslant 0$, $p_1 + \ldots + p_{K-1} \leqslant 1$ and $\Gamma(\cdot)$ is the gamma function. The important special case where $K = 2$ is that when $\mathbf{X}_n = (X_{n,1}, X_{n,2})$ is binomial and $\theta = \theta_1$ is beta, and we write $\text{Dir}(a, b) = \text{Beta}(a, b)$. Denoting $a_\cdot = a_1 + \ldots + a_K$, if $\theta \sim \text{Dir}(\mathbf{a})$ then $\text{E}(\theta_j) = a_j / a_\cdot = \mu_j$, $\text{var}(\theta_j) = \mu_j(1 - \mu_j)/(a_\cdot + 1)$ and $\text{cov}(\theta_j, \theta_i) = -\mu_j \mu_i / (a_\cdot + 1)$. We also require the following additional properties of the Dirichlet family.

Theorem 1: Under assumptions (1) and (2) above,

$$\theta_{\text{E}} | \mathbf{X}_n \sim \text{Dir}(a_{\text{E},1} + X_{n,1}, \ldots, a_{\text{E},K} + X_{n,K}). \tag{1}$$

Theorem 2: If $(\theta_1, \ldots, \theta_{K-1}) \sim \text{Dir}(a_1, \ldots, a_K)$, then for any $r = 1, \ldots, K - 1$,

$$(\theta_1, \ldots, \theta_r) \sim \text{Dir}\left(a_1, \ldots, a_r, \sum_{j=r+1}^{K} a_j\right), \tag{2}$$

$$\left(\sum_{j=1}^{r} \theta_j, \theta_{r+1}, \ldots, \theta_{K-1}\right) \sim \text{Dir}\left(\sum_{j=1}^{r} a_j, a_{r+1}, \ldots, a_K\right), \tag{3}$$

and

$$\left(\sum_{j=1}^{r+1} \theta_j\right)^{-1} (\theta_1, \ldots, \theta_r) \sim \text{Dir}(a_1, \ldots, a_{r+1}). \tag{4}$$

Theorem 1 says that the Dirichlet is a conjugate prior for the multinomial. Statements (2) and (3) of theorem 2 say that the Dirichlet family is closed under collapsing of categories, while (4) says that it is closed under conditioning. For example, if $K = 4$ and we combine categories 1 and 2, then the corresponding distribution of $(\theta_1 + \theta_2, \theta_3)$ is $\mathrm{Dir}(a_1 + a_2, a_3, a_4)$, while $(\theta_1 + \theta_2)/(\theta_1 + \theta_2 + \theta_3) \sim \mathrm{Beta}(a_1 + a_2, a_3)$ and $\theta_1/(\theta_1 + \theta_2 + \theta_3) \sim \mathrm{Beta}(a_1, a_2 + a_3)$.

For a phase II trial of E, we require an informative prior on $\boldsymbol{\theta}_S$, based on some combination of historical data and clinical experience (see Freedman and Spiegelhalter[18]). In practice, it is often appropriate simply to take $a_{S,j}$ as the number of responses in the jth category from historical data on S. We also require that the prior of $\boldsymbol{\theta}_E$ be at most slightly informative, to reflect properly the fact that we usually know little about E at the outset of a phase II trial. Since we can regard $a_{E.} = a_{E,1} + \ldots + a_{E,K}$ as a dispersion parameter, with larger values corresponding to smaller variances of the θ_j's, we set $a_{E.} = K$ so that the prior amount of information on $\boldsymbol{\theta}_E$ corresponds to that of the uniform distribution $\mathrm{Dir}(1, \ldots, 1)$ on the $(K-1)$-dimensional square.

Sometimes it is useful to reparameterize a $\mathrm{Dir}(\mathbf{a})$ distribution as follows. For a given simple or compound event, say A_1, marginally θ_1 is distributed $\mathrm{Beta}(a_1, a_2 + \ldots + a_K)$. Generalizing Thall and Simon,[1] first note that $(a_1, a_2 + \ldots + a_K)$ corresponds on a one-to-one basis to $(a_1, a_.)$, which in turn corresponds to $\mu_1 = a_1/a_.$ and $W_{1,90}$ = the width of the 90 per cent probability interval of the $\mathrm{Beta}(a_1, a_2 + \ldots + a_K)$ distribution, running from the 5th to the 95th percentiles. Given μ_1 and $W_{1,90}$, there are $K - 2$ parameters that remain among $a_2, \ldots, a_K$ from the original $\mathrm{Dir}(\mathbf{a})$, and we may specify these either as a_j's or as $K - 2$ means from among $\mu_2, \ldots, \mu_K$. This reparameterization is useful if the clinician wishes to describe the prior in terms of the response category means. When W corresponds to a compound event C, one implements this approach simply by referring to the marginal distribution of θ_C to compute W_C.

2.2. Stopping criteria

In general, our objective is to monitor all clinically important events. We thus consider trials where the clinical focus is two or more simple or compound events obtained from the elementary outcomes $A_1, \ldots, A_K$. We define the monitoring criteria for each event marginally, that is, in terms of that event alone, both for simplicity and because clinicians think in terms of these event rates. Since the monitoring rules operate simultaneously, however, it is essential to evaluate their joint behaviour, based on a consistent probability model for $\boldsymbol{\theta}_E$ and $\mathbf{X}_n$.

Let $(j_1, \ldots, j_r)$ be r distinct indices from $(1, \ldots, K)$ such that $C = A_{j_1} \cup \ldots \cup A_{j_r}$ is a given outcome of interest, and denote $\eta = \theta_{j_1} + \ldots + \theta_{j_r} = \Pr(C)$. In a single-arm trial of E the prior of $\boldsymbol{\theta}_S$ is unchanged, whereas we update the prior on $\boldsymbol{\theta}_E$ repeatedly as we observe patient responses. The decision rule for monitoring the incidence of C is one of three types, each of a general form based on comparison of the posterior distribution of η_E given $\mathbf{X}_n$ to the distribution of η_S. We monitor the data for each n, beginning at a minimum sample size m and continuing until we either make a decision or we reach a predetermined maximum sample size M. The monitoring criterion is the posterior probability

$$\Pr[\eta_S + \delta < \eta_E \mid \mathbf{X}_n] = \lambda(X_{n,j_1} + \ldots + X_{n,j_r}, n; \mathbf{a}_S, \mathbf{a}_E, \delta),$$

where $\delta \geqslant 0$ is a design parameter that quantifies the desired increase (for efficacy outcomes) or largest allowable increase (for adverse events) in the probability of C. Denoting the beta density and CDF by b and B, respectively, the above probability equals

$$\int_0^{1-\delta} \{1 - B_{\eta E}(p + \delta)\} b_{\eta S}(p)\, \mathrm{d}p. \tag{5}$$

Denote $\alpha = a_{j_1} + \ldots + a_{j_r}$, $\beta = \{a_\cdot - (a_{j_1} + \ldots + a_{j_r})\}$ and $X_n(C) = X_{n,j_1} + \ldots + X_{n,j_r}$ for brevity. We can easily evaluate expression (5) by numerical integration using the facts that $\eta_S \sim \text{Beta}(\alpha_S, \beta_S)$ and, by (1), that $\eta_E | \mathbf{X}_n \sim \text{Beta}(\alpha_E + X_n(C), \beta_E + n - X_n(C))$.

The first two types of monitoring boundaries correspond to an *efficacy* event C, which we define as any outcome for which a higher probability is clinically desirable. An efficacy event in a phase II trial typically characterizes short – or intermediate – term treatment success, and increasing its likelihood is usually the primary clinical goal of the trial. If a targeted improvement of $\delta(C)$ in the mean of η_S is the efficacy goal, then we terminate the trial and declare E 'not promising' compared to S if

$$\Pr[\eta_S + \delta(C) < \eta_E | \mathbf{X}_n] \leqslant p_L(C) \tag{6}$$

for a given small value of the lower criterion probability $p_L(C)$. Essentially, (6) ensures early termination of the trial if E is unlikely to provide the desired $\delta(C)$ improvement. We obtain the corresponding upper boundary from the criterion that the trial be terminated and E declared 'promising' compared to S if

$$\Pr[\eta_S < \eta_E | \mathbf{X}_n] \geqslant p_U(C) \tag{7}$$

for a given large value of $p_U(C)$. This rule simply says that we should declare E efficacious, in terms of the event C, if *a posteriori* it becomes likely that the probability of achieving the clinical outcome C when we treat patients with E exceeds the corresponding probability associated with S. Thall and Simon[1] proposed the criteria (6) and (7) to monitor phase II trials with a single binary outcome. Freedman and Spiegelhalter[19] suggested a similar approach, that does not use decision theory, for randomized trials with one outcome under a Gaussian model.

We use the third type of stopping boundary to maintain approximate equivalence in the rate of a given *adverse* event, which we define as any outcome for which a lower probability is clinically desirable, equivalently as the complement of an efficacy event. For an adverse event T with probability $\eta(T)$, the rule is to terminate the trial if

$$\Pr[\eta_S(T) + \delta(T) < \eta_E(T) | \mathbf{X}_n] \geqslant p_U(T) \tag{8}$$

for large upper criterion probability $p_U(T)$. We have found this rule highly desirable when used together with the efficacy rule (6) in trials where the efficacy event C and the adverse event T have non-empty intersection. This situation corresponds to that described in Section 2.1 where the innovative aspect of E is likely to increase the rates of both C and T, and one regards an increase of $\delta(T)$ in the rate of the adverse event as the largest clinically acceptable price that one can pay for a $\delta(C)$ increase in the rate of the efficacy event.

For example, in the (CR, TOX) example noted earlier, $C = \text{CR} = A_1 \cup A_2$ denotes complete remission and $T = \text{TOX} = A_1 \cup A_3$ denotes acute toxicity, hence $A_1 = [\text{CR and TOX}]$ is both desirable and undesirable. In particular, the probability of A_1 is likely to be increased by a more aggressive therapy, that is, in many trials of new combination bio-chemotherapies we may anticipate that the rates of both CR and TOX will increase. If, for example, we target an improvement of $\delta(\text{CR}) = 0{\cdot}15$ in CR rate and we consider an increase of $\delta(\text{TOX}) = 0{\cdot}10$ in the TOX rate an acceptable tradeoff for the desired CR rate improvement, then we would use the rules (6) and (8) together to monitor both CR and TOX, with the possibility of using the 'upper' efficacy criterion (7) as well. The use of stopping rules for one or more adverse events in such circumstances helps to reduce the probability of outcomes in which the best one can say is 'The treatment was a success but the patient died'.

2.3. Constructing the Stopping Boundaries

Computation of a stopping boundary that corresponds to an event C relies on the facts that the posterior of θ_E given $\mathbf{X}_n$ is $\text{Dir}(\mathbf{a}_E + \mathbf{X}_n)$, that λ depends on $\mathbf{X}_n$ only through $x = X_{n,j_1} + \ldots + X_{n,j_r}$, and that $\lambda(x, n; \mathbf{a}_S, \mathbf{a}_E, \delta)$ is an increasing function of x. Thall and Simon[1,2] discuss this in detail in the context of single-arm trials with one binary efficacy outcome. Given the criterion probabilities $p_L(C)$ and $p_U(C)$ for the stopping criteria (6) and (7), where $p_L(C)$ is a small value such as 0·01–0·20 and $p_U(C)$ is a large value such as 0·80–0.99, we define the lower and upper decision cutoffs, respectively, for monitoring the efficacy endpoint C as

$$L_n(C) = \text{the largest integer } x \text{ such that } \lambda(x, n; \pi_S, \pi_E, \delta(C)) \leqslant p_L(C),$$

$$U_n(C) = \text{the smallest integer } x \text{ such that } \lambda(x, n; \pi_S, \pi_E, 0) \geqslant p_U(C).$$

The corresponding decision rules for C at stage n, each applied under the condition that we have not hit a stopping boundary prior to stage n, are as follows:

$$\text{If } X_{n,j_1} + \ldots + X_{n,j_r} \leqslant L_n(C), \text{ then stop the trial and declare E not promising.} \quad (9)$$

$$\text{If } X_{n,j_1} + \ldots + X_{n,j_r} \geqslant U_n(C), \text{ then stop the trial and declare E promising.} \quad (10)$$

As discussed in Thall and Simon,[2] in some trials one may consider it desirable to use the lower efficacy boundary, since it is clinically more protective, but not the upper bound. The point here is that one may use either of the two rules (9) or (10) without the other. For monitoring an adverse event $T = A_{k,1} \cup \ldots \cup A_{k,l}$ based on (8), the upper decision cutoff is

$$U_n(T) = \text{the smallest integer } x \text{ such that } \lambda(x, n; \pi_S, \pi_E, \delta(T)) \geqslant p_U(T)$$

and the corresponding decision rule in terms of the data is if

$$X_{n,k_1} + \ldots + X_{n,k_l} \geqslant U_n(T) \quad (11)$$

then stop the trial. As we observe each patient response, the multinomial vector $\mathbf{X}_n$ is updated to $\mathbf{X}_{n+1}$ and thus we update the counts $X_{n,j_1} + \ldots + X_{n,j_r}$ of C and $X_{n,k_1} + \ldots + X_{n,k_l}$ of T and compare them to their stopping bounds, with the obvious elaboration if we monitor more than two events.

We choose design parameters to obtain monitoring boundaries which have desirable properties when used jointly. To do this, we first evaluate the marginal operating characteristics of the design that corresponds to each single outcome of interest while ignoring the others. We then use these results to construct several joint design parameterizations for monitoring all of the events together and evaluating each design under relevant fixed values of the multiple outcome probability vector. We repeat these steps in collaboration with the clinician until we obtain a design which is ethically, medically and statistically desirable.

3. APPLICATIONS

The operating characteristics for each design considered here are based on 10,000 simulated trials. We performed all computations in C on a Solbourne 5/600 computer, using the Bays-Durham shuffling algorithm (see Press *et al.*,[20] Chapter 7.1) to generate random numbers for the simulations. Each run of 10,000 took about 30 to 120 seconds, depending upon machine load, so that we could evaluate even the most complicated designs very quickly under multiple

parameterizations. A menu driven computer program which carries out the necessary computations is available from the first author on request.

3.1. The HLA non-identical donor BMT trial

In patients diagnosed with the haematologic malignancies leukaemia, lymphoma or myelodysplastic syndrome, BMT using marrow cells from a human leukocyte antigen (HLA) identical sibling offers a potentially curative treatment. Unfortunately, only about one-third of such patients have HLA-identical siblings. An alternative for the other two-thirds is to transplant marrow from donors whose cells match the patient's at several of the HLA loci. Graft-versus-host disease (GVHD) and transplant rejection (TR) are major complications associated with this approach. The following design has been used at M.D. Anderson Cancer Center for a phase II trial of XomaZyme-CD5 +, cyclosporine and methylprednisone given as a post-transplant prophylaxis for GVHD in patients receiving partially T-cell depleted marrow from an HLA-matched unrelated or one-antigen-mismatched related donor.

Both GVHD and TR were monitored for 100 days post transplant, producing the 2×2 structure given in Table I, which also appears in Thall and Simon.[26] We obtained the standard therapy Dirichlet prior parameters $\mathbf{a}_S = (a_{S,1}, a_{S,2}, a_{S,3}, a_{S,4})$ by first eliciting the elementary outcome means and the dispersion parameter $W_{S,90} = 0{\cdot}20$ for $\Pr[\overline{\text{GVHD}}] = \theta_{S,1} + \theta_{S,2}$ from the clinician, then converting $(\mu_{S,1}, \mu_{S,2}, \mu_{S,3}, W_{S,90})$ to $\mathbf{a}_S$ as described in Section 2.1. The efficacy event is $A_1 \cup A_2 = [\overline{\text{GVHD}}]$, and $A_2 \cup A_4 = \text{TR}$ is the adverse event. The study objectives were to obtain an improvement of 0·20 in $\Pr[\overline{\text{GVHD}}] = \theta_{E,1} + \theta_{E,2}$ while maintaining with high posterior probability a TR rate no more than 0·05 above that of standard therapy. We chose a maximum sample size of 75 to ensure that if the trial ran to completion the posterior of $\theta_{E,1} + \theta_{E,2}$ will have 92·5 per cent probability interval of width 0·20. The formal decision rules are to stop the trial if

$$\Pr[\theta_{S,1} + \theta_{S,2} + 0{\cdot}20 < \theta_{E,1} + \theta_{E,2} \mid \mathbf{X}_n] \leqslant 0{\cdot}02, \tag{12}$$

or

$$\Pr[\theta_{S,2} + \theta_{S,4} + 0{\cdot}05 < \theta_{E,2} + \theta_{E,4} \mid \mathbf{X}_n] \geqslant 0{\cdot}80. \tag{13}$$

Table II gives this design's operating characteristics. We obtained the design parameters by first evaluating the design which monitors only GVHD for various numerical values of $p_L(\overline{\text{GVHD}})$ and $\delta(\overline{\text{GVHD}})$, and we likewise evaluated the design that monitors only TR for several values of $p_U(\text{TR})$ and $\delta(\text{TR})$. We then chose the criterion probabilities $p_L = 0{\cdot}02$ and $p_U = 0{\cdot}80$ to obtain desirable operating characteristics when the two rules are used jointly. We began this process with $\delta(\text{TR}) = 0{\cdot}10$, that is, to allow 0·10-equivalence in the TR rate. The clinician's reaction to the fact that the two monitoring boundaries jointly produced a stopping probability of 0·37 for fixed values $p(\overline{\text{GVHD}}) = 0{\cdot}40$ and $p(\text{TR}) = 0{\cdot}30$, however, was that 0·37 was too low, that is, the design was not sufficiently protective if the TR rate increased from 0·20 to 0·30, even with the desired improvement in GVHD rate. Decreasing $\delta(\text{TR})$ to 0·05 produced the desired operating characteristics, with a termination probability of 0·68 and a median of 33 patients in the case noted. This is the sort of approach which we recommend in general, since one may obtain the numerical properties of several design parameterizations very quickly via simulation. In the best case given in Table II, namely with the desired 0·20-improvement in GVHD rate and a 0·10 drop in rejection rate, that is, $p(\overline{\text{GVHD}}) = 0{\cdot}40$ and $p(\text{TR}) = 0{\cdot}10$, the design has a 91 per cent chance of continuing to conclusion with 75 patients.

Table I. Outcomes and standard prior for HLA non-identical donor BMT trial

Patient response	Probability	Mean	$a_{S,I}$
A_1 = [No GVHD and No TR]	θ_1	0·05	2·037
A_2 = [No GVHD and TR]	θ_2	0·15	6·111
A_3 = [GVHD and No TR]	θ_3	0·75	30·555
A_4 = [GVHD and TR]	θ_4	0·05	2·037

Table II. HLA non-identical donor BMT trial operating characteristics

True probabilities		Stopping probabilities							Achieved sample size		
No GVHD	TR	Due to: GVHD	+	TR	−	Both	=		25th	50th percentiles	75th
0·20	0·10	0·94	+	0·00	−	0·00	=	0·94	11	18	33
0·20	0·20	0·87	+	0·09	−	0·01	=	0·95	11	14	30
0·20	0·30	0·67	+	0·36	−	0·05	=	0·98	11	14	22
*0·20	0·40	0·48	+	0·66	−	0·14	=	1·00	11	12	18
†0·40	0·10	0·08	+	0·01	−	0·00	=	0·09	75	75	75
0·40	0·20	0·08	+	0·12	−	0·00	=	0·20	75	75	75
0·40	0·30	0·08	+	0·60	−	0·00	=	0·68	14	33	75
0·40	0·40	0·05	+	0·93	−	0·01	=	0·97	11	13	23

* Worst outcome: No improvement in GVHD rate and mean rejection rate increases from 20 per cent to 40 per cent
† Best outcome: Mean GVHD-free rate increases from 20 per cent to 40 percent and rejection rate drops to 10 per cent

Graphical representations of the design's operating characteristics are given by contour plots of the probability of early termination (Figure 1) and of sample size (Figure 2), which show how these design properties vary with fixed values of $p(\overline{\text{GVHD}})$ and $p(\text{TR})$. In these plots the most and least desirable pairs of these probabilities are in the lower right and upper left portion of the graph, respectively. The design is highly likely to terminate early with a relatively small number of patients when it is desirable to do so, and it is likely to accrue the maximum 75 patients when the true rates of GVHD and TR are more desirable.

An important provision is that one must score each patient's outcomes at day 100. If one scores GVHD or TR at the calendar times of their occurrence, then a bias will result because, by definition, these events occur sooner than the 'success' events of lasting the 100 days without GVHD or TR. In general, to avoid such bias one should score the binary indicator of $[T \geqslant t_0]$ for any waiting-time variable T for each patient at t_0 after the patient's entry date, not at the calendar time of occurrence when $T < t_0$. To see the potential problem, consider a trial of T = time to relapse or death in which the true probability of $[T \geqslant$ one year$]$ is 0·50, this is considered an acceptable rate, and 40 patients are entered simultaneously. By month nine of the trial, about 15 events should have occurred, but no patients can yet be scored as reaching the success goal of one year. If one scores events at their calendar times, then at month nine the summary statistic is 15 events out of 15 scored and the trial surely will terminate, even though the true rate of $[T \geqslant$ one year$]$ is an acceptable 0·50. In practice, this provision presents minimal difficulty, since one scores patient outcomes in exactly the same time sequence as the patients enter the trial, shifted t_0 into the future.

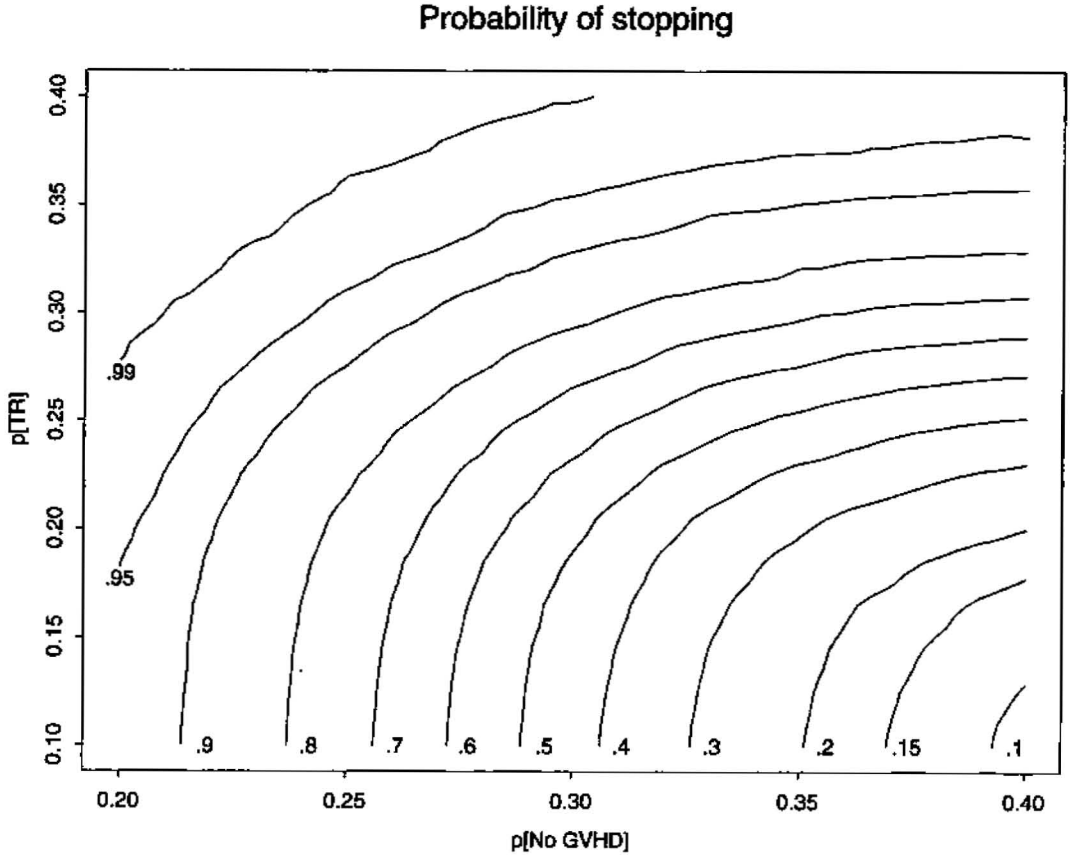


Figure 1. Contour plot of Pr[Early Stopping] as a function of fixed values of Pr[No GVHD] and Pr[Transplant Rejection] for HLA non-identical donor BMT trial

An additional rule to use in conjunction with the above for monitoring each event is the following: if, at the calendar time t^* of any individual patient 'failure', even under the assumption that all patients accrued but not yet evaluated will have successes, the trial will meet a future stopping criterion for this failure event, then one should terminate the trial at t^*. This simply applies a well-known advantage of sequential monitoring, and in our setting it protects those patients whom we would have accrued and treated with E after calendar time t^*.

3.2. The IAG trial

Patients with newly diagnosed acute myelogenous leukaemia (AML) are heterogenous with respect to prognosis, depending primarily upon cytogenetic abnormalities, presence or absence of an antecedent acute haematologic disorder, and patient age. A phase II trial of idarubicin (I) + ara-C (A) + granulocyte colony-stimulating factor (G-CSF) for both remission induction and remission maintainance was carried out at M.D. Anderson Cancer Center in 'intermediate' prognosis AML patients. The rationale for this combination was the success of I + A (IA) in an earlier trial, and that both *in vitro* and clinical evidence suggested that the growth factor G-CSF would increase the sensitivity of AML blast cells to chemotherapy.

Traditionally, the binary variable that indicates whether a patient has achieved complete remission (CR) by one month has been used to define patient response in phase II bio-chemotherapy trials in acute leukaemia. One problem with this approach is that it scores

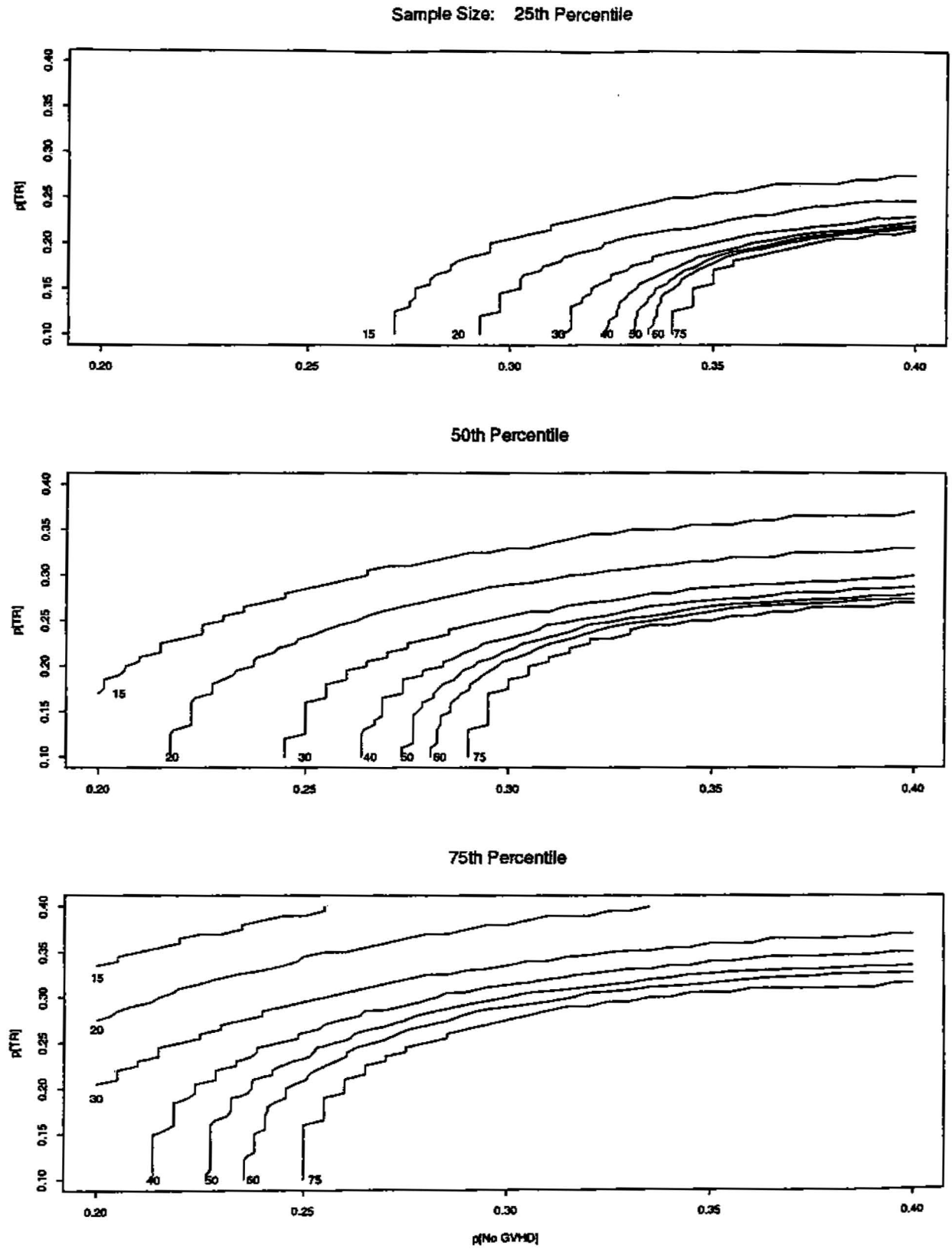


Figure 2. Contour plots of 25th, 50th and 75th percentiles of sample size as functions of fixed values of Pr[No GVHD] and Pr[Transplant Rejection] for HLA non-identical donor BMT trial

Table III. Outcomes and standard prior for IAG trial

Patient response	Probability	Mean	$a_{IA,j}$
A_1 = [CR and RD $\geq$ 6 months]	θ_1	0·5849	31
A_2 = [CR and RD < 6 months]	θ_2	0·2642	14
A_3 = [No CR]	θ_3	0·1509	8

a patient who achieves CR by day 30 post induction but relapses or dies shortly thereafter as a treatment success. Moreover, patients who achieve CR but subsequently relapse have a much lower overall survival rate, largely due to the reduced probability of achieving a second remission after relapse. The goal of the I + A + G − CSF (IAG) trial was to achieve more durable remissions compared to IA, hence the usual one-month timeframe for defining patient response was extended to seven months. Denoting RD = first remission duration, the specific goals were to increase $\Pr[\text{RD} \geq 6 \text{ months} \mid \text{CR}]$ by $\delta(\text{RD}) = 0{\cdot}15$ while maintaining the CR rate within $\delta(\text{CR}) = 0{\cdot}10$. We thus defined the three response categories given in Table III, with standard therapy defined as IA and the Dirichlet prior on θ_{IA} determined by the response category counts from the earlier IA trial. Tables I and III together illustrate the flexibility of the general approach, since we determine the standard treatment prior parameters by the event probability mean vector μ_S and a dispersion parameter W_S in the former, and in terms of the Dirichlet parameters $\mathbf{a}_S = \mathbf{a}_{IA}$ in the latter.

All patients in the trial were evaluated for CR, the equivalence outcome, at 30 days post initiation of therapy, and each patient in the subgroup who achieved CR was subsequently evaluated for the binary event [RD $\geq$ 6 months] at 1 + 6 = 7 months after initiation of therapy. Regarding $\overline{\text{CR}}$ as an adverse outcome and applying (3) and (4), the probabilities that serve as the basis for the stopping rules are thus $\theta_3 = \Pr[\overline{\text{CR}}]$, which is $\sim \text{Beta}(a_3, a_1 + a_2)$, and $\tau = \theta_1/(\theta_1 + \theta_2) = \Pr[\text{RD} \geq 6 \text{ months} \mid \text{CR}]$, which is $\sim \text{Beta}(a_1, a_2)$. Denote $X_{n,3}$ = the number of patients out of the first n evaluated who fail to achieve CR by day 30, etc. For the subgroup of patients entering CR, the effective sample size for the number who achieve a six-month remission duration is $X_{n,1} + X_{n,2} = X_{n,\text{CR}}$, rather than n. That is, for given $\theta = (p_1, p_2)$ and n, we first observe $X_{n,\text{CR}}$ which is $\sim$ binomial in $(n, p_1 + p_2)$, and subsequently observe $X_{n,1}$, which, given $X_{n,\text{CR}}$ is $\sim$ binomial in $(X_{n,\text{CR}}, p_1/(p_1 + p_2))$. As each patient's study time reaches day 30, we update $X_{n,3}$ and $X_{n+1,3} = X_{n,3} + 1$ or $X_{n,3}$ depending upon whether the patient did or did not enter CR. For patients entering CR, if $X_{n,\text{CR}} = k$ at the time a patient reaches the seven-month endpoint, then $X_{k+1,1} = X_{k,1}$ or $X_{k,1} + 1$ depending upon whether the patient did or did not relapse prior to six months after achieving CR.

The early termination criteria are to stop the trial if

$$\Pr[\theta_{S,3} + 0{\cdot}10 < \theta_{E,3} \mid X_{n,3} \text{ out of } n] \geq 0{\cdot}90, \tag{14}$$

or

$$\Pr[\tau_S + 0{\cdot}15 < \tau_E \mid X_{n,1} \text{ out of } X_{n,\text{CR}}] \leq 0{\cdot}10. \tag{15}$$

The termination rule that corresponds to (14) is of the form $X_{n,3} \geq U_n(\overline{\text{CR}})$. To monitor six-month remission duration among the subgroup of patients achieving CR, the inequality in (9) takes the form $X_{n,1} \leq L_{X_{n,\text{CR}}}(\text{RD})$.

The rationale for using equivalence $\delta(\mathrm{CR}) = 0{\cdot}10$ and efficacy $\delta(\mathrm{RD}) = 0{\cdot}15$ is as follows. Since the mean CR rate with IA was 0·85 while the mean six-month RD rate among those achieving CR was 0·69, a drop of 0.10 in the CR rate and increase of 0·15 in conditional [RD $\geqslant$ 6 months] rate would produce a mean Pr[CR and RD $\geqslant$ 6 months] = $0{\cdot}75 \times 0{\cdot}84 = 0{\cdot}63$, which is a modest improvement over the mean of 0·59 for the rate of this most desirable outcome obtained with IA. If we can maintain the mean CR rate at 0·85 with IAG, however, then the mean Pr[CR and RD $\geqslant$ 6 months] = 0·71, a substantial improvement over IA. Again, we chose the criterion probabilities $p_U(\overline{\mathrm{CR}}) = 0{\cdot}90$ and $p_L(\mathrm{RD}) = 0{\cdot}10$ to obtain a design with desirable operating characteristics.

We chose the maximum sample size to ensure that, if the trial runs to completion, a posterior 95 per cent probability interval for τ_{IAG} will have width 0·20, which requires 50 patients to enter CR for evaluation of six-month remission duration. If the observed CR rate is 0·85 or 0·75, then we will accrue 50/0·85 = 59 or 50/0·75 = 67 patients, respectively. For true CR rates much lower, the trial is likely to terminate early. Table IV gives operating characteristics of the IAG trial design. Since [RD $\geqslant$ 6 months] and $\overline{\mathrm{CR}}$ cannot both occur in the same patient, each patient's outcomes can contribute to at most one of the stopping events. For true CR rate $\leqslant$ 0·65 there is at least a 77 per cent chance the trial will terminate early, and the probability of early termination is much higher if remission duration does not improve. In the best case, where true CR rate is maintained at 0·85 and we achieve the targeted improvement of 0·15 in the conditional probability of six-month remission duration, there is a probability of 0·85 that the trial will not stop early. In this case the median sample size is 58 patients with the expectation that 41/58 = 71·4 per cent of these will achieve the most desirable outcome, compared to 31/53 = 58·5 per cent with IA.

3.3. A double-intensification BMT trial

In treatment of non-Hodgkin's lymphoma by BMT, prior to transplant the patient first undergoes conventional-dose chemotherapy to reduce the number of cancer cells, then receives *intensification* with high-dose chemotherapy, followed by transplant and a post-transplant regimen to reduce the rates of GVHD and infection. An innovation in this process is to repeat the intensification stage, a more aggressive approach which may have a higher risk of early death but also an increased chance of long-term survival in those who do not die early. In a phase II double-intensification BMT trial in patients with malignant lymphoma, conventional-dose chemotherapy was followed by intensification with cyclophosphamide (CYC) + etoposide + cisplatin, followed by G-CSF to accelerate recovery of white blood cell and platelet counts. Patients next received a second intensification with thiotepa + busulfan + CYC, and then underwent transplantation and standard post-transplant therapy. High risk, typically chemotherapy refractory patients having HLA-compatible donors, were given allogeneic BMT (from a donor's bone marrow), with all others in the trial receiving autologous cells (from the patient's own marrow). Patients who received autologous transplant were divided into three risk groups, defined by the pathologic characteristics (grade) of their lymphoma.

With this approach, early success was defined as 75-day survival and late success as one-year disease-free survival. The design thus must accommodate monitoring of 75-day survival in the combined subgroups, with long-term relapse and survival monitored separately in each subgroup, as illustrated by Figure 3. Denote X = time from the initiation of treatment to death and R = time to relapse. Since relapse can occur only in patients who survive the initial 75 day double-intensification regimen, the three elementary outcomes are $A_1 = [X < 75 \text{ days}]$, $A_2 = [75 \text{ days} \leqslant \min(X, R) < \text{one year}]$, and $A_3 = [\min(X, R) \geqslant \text{one year}]$. This patient group is homogeneous with respect to short-term survival, specifically $\theta_1 = \Pr[A_1]$ is the same for all

Table IV. IAG trial operating characteristics

Assumed true probabilities Pr[CR]	Pr[RD ⩾ 6 | CR]	Pr[stop] CR + RD = TOTAL	Achieved sample size N_{25}	N_{50}	N_{75}
0·85	0·69	0·01 + 0·85 = 0·86	14	19	43
0·85	0·84	0·01 + 0·14 = 0·15	55	58	60
0·75	0·69	0·12 + 0·75 = 0·87	15	20	40
0·75	0·84	0·16 + 0·13 = 0·29	22	62	67
0·65	0·69	0·48 + 0·48 = 0·96	14	18	26
0·65	0·84	0·67 + 0·09 = 0·76	15	33	69
0·55	0·69	0·80 + 0·19 = 0·99	12	14	18
0·55	0·84	0·96 + 0·03 = 0·99	12	14	23

Standard (*IA*) mean Pr[CR] = 0·85
Standard mean Pr[RD ⩾ 6 | CR] = 0·69 (target = 0·84)

patient subgroups, but patient heterogeneity is a factor in long-term survival or relapse. Index the four patient subgroups by $j = 1$ for allogeneic transplant and $j = 2$, 3 and 4, respectively, for autologous transplant with high, intermediate and low grade lymphoma, so that *a priori* long-term prognosis improves as j increases. By theorem 2, $(\theta_1, \theta_{j,2}) \sim \text{Dir}(a_1, a_{j,2}, a_{j,3})$ in subgroup j, with the long-term survival probability $\theta_{j,3} = 1 - \theta_1 - \theta_{j,2}$ also stratum-specific. Note that $a_{j,2} + a_{j,3} = a_{2,3}$ does not vary with j; otherwise, the distribution of θ_1 would not be homogeneous across patient groups. The measure of treatment efficacy was one-year disease-free survival, A_3, monitored in subgroup j in terms of $\tau_j = \text{Pr}_j[\min(T, R) \geqslant \text{one year} \mid T \geqslant 75 \text{ days}] = \text{Pr}_j[A_3 \mid A_2 \cup A_3] = \theta_{j,3}/(\theta_{j,2} + \theta_{j,3}) \sim \text{Beta}(a_{j,3}, a_{j,2})$, $1 \leqslant j \leqslant 4$. The adverse outcome $A_1 = [T \leqslant 75 \text{ days}]$, that is, early death, has probability $\theta_1 \sim \text{Beta}(a_1, a_{2,3})$, and this was monitored in the combined subgroups.

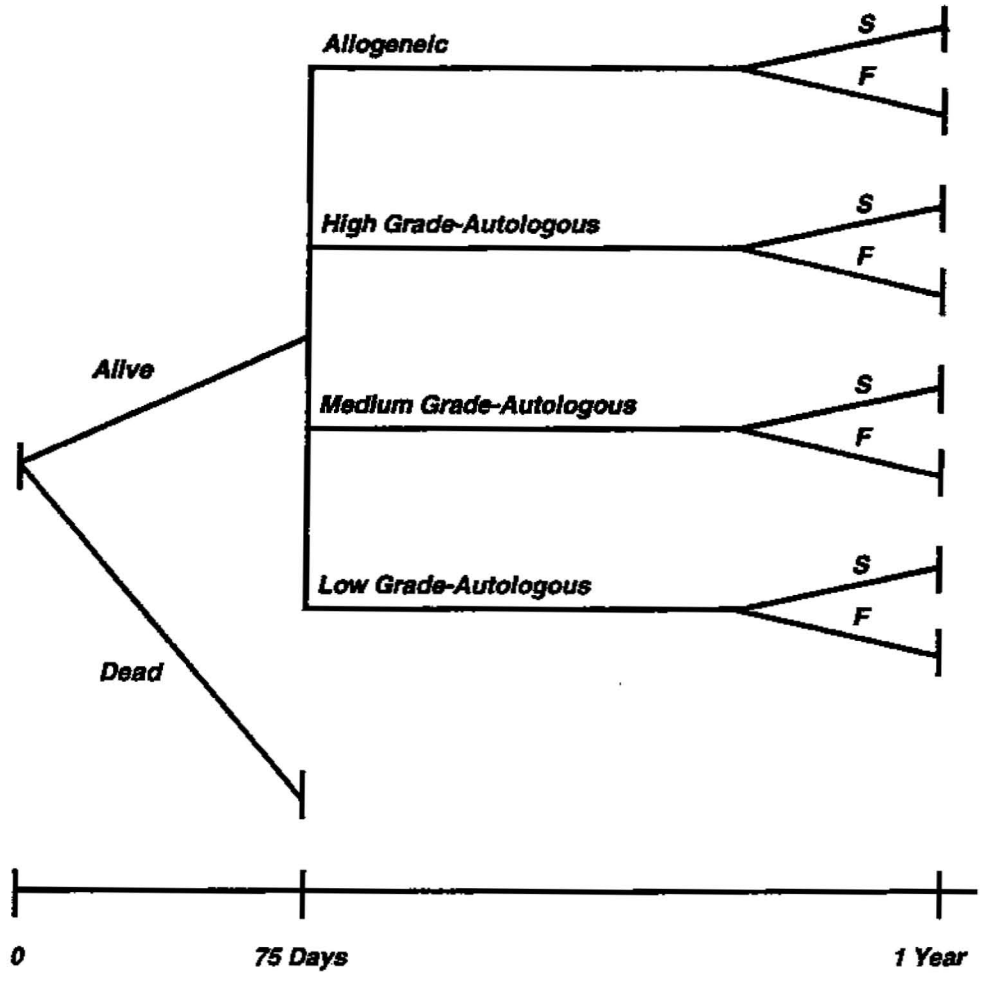
The clinician specified the priors for standard (single-intensification) therapy in terms of $\mu_1 = 0{\cdot}85$ and $W_{1,90} = 0{\cdot}20$ for the distribution of $1 - \theta_1$, which determines a_1 and $a_{2,3}$, and then the means $E(\tau_j) = a_{j,3}/(a_{2,3})$ of the conditional one-year survival probabilities, which were $E(\tau_1) = E(\tau_2) = 0{\cdot}20$, $E(\tau_3) = 0{\cdot}30$, and $E(\tau_4) = 0{\cdot}40$. Although $\tau_{S,1}$ and $\tau_{S,2}$ have identical priors, the first two subgroups were monitored separately to allow for the possibility of different response rates. As before, we used a flat prior for θ_E with means equated to those of θ_S in each subgroup. The criterion to terminate the entire trial was

$$\text{Pr}[\theta_{S,1} + 0{\cdot}05 < \theta_{E,1} \mid \mathbf{X}_n] \geqslant 0{\cdot}85, \tag{16}$$

and the criterion to terminate subgroup j *per se* was

$$\text{Pr}[\tau_{S,j} + 0{\cdot}20 < \tau_{E,j} \mid \mathbf{X}_{n,j}] \leqslant p_{L,j}(A_3), \tag{17}$$

where $p_{L,j}(A_3) = 0{\cdot}05$ for subgroups 1, 2 and 3, and 0·075 for subgroup 4. As in all of our applications, we examined a range of values for p_U and $p_{L,j}$ to obtain a design with good operating characteristics. Termination of the trial due to (16) corresponds to failure of the double-intensification regimen due to an unacceptably high early death rate, compared to single-intensification. The clinician considered an increase of $\delta(A_1) = 0{\cdot}05$ in the probability of death during the first 75 days an acceptable trade-off for a 0·20-improvement in the conditional probability of one-year survival, since with single-intensification on average the latter is only 0·40 even in the most



S = Alive and Not Relapsed at 1 Year

F = $\bar{S}$ = Dead or Relapsed Prior to 1 Year

Figure 3. Schematic of double-intensification BMT trial design

favourable subgroup. We chose the maximum sample size in each subgroup subject to practical limitations in accrual rates, with each M_j chosen to obtain a posterior 90 per cent probability interval for τ_j having width 0·90, so that $M_1 = M_2 = M_4 = 39$ and $M_3 = 40$. Total maximum sample size thus is $157/0{\cdot}85 = 185$ or $157/0{\cdot}75 = 210$ for true 75-day survival probability $p_{75} = 0{\cdot}85$ or $0{\cdot}75$, if the trial runs to completion in all subgroups, with a high likelihood of early termination if p_{75} is much below 0·75. The maximum trial duration is nearly five years. The trial thus has two stages, with stage 2 (one-year disease-free survival) monitoring beginning for each patient at day 75, provided (s)he has survived that long. Again, to avoid bias one scores 75-day survival at day 75 post initiation of treatment, not at the time of death, for patients dying during the initial period. Likewise, for patients surviving the initial 75 days, one scores subsequent long-term disease-free survival at one-year. Table V summarizes the operating characteristics of this design. We used a maximum sample size of $157/0{\cdot}65 = 242$ in the stage 1 computations. All

Table V. Double-intensification BMT trial operating characteristics

Stage 1

True Pr[$T^* \geq 75$ days]	Pr[stop]	N_{25}	N_{50}	N_{75}
0·65	0·93	11	15	29
0·75	0·49	18	242	242
0·85	0·06	242	242	242

Stage 2

Patient subgroup	Prior mean	P_L	Maximum sample size	True conditional Pr[$T^* \geq 1$ year]	Pr[stop]	N_{25}	N_{50}	N_{75}
Allogeneic or high grade lymphoma autologous	0·20	0·05	39	0·20	0·80	10	12	23
				0·40	0·14	39	39	39
Intermediate grade lymphoma autologous	0·30	0·05	40	0·30	0·80	10	16	33
				0·50	0·11	40	40	40
Low grade lymphoma autologous	0·40	0·075	39	0·40	0·82	10	14	29
				0·60	0·14	39	39	39

T^* = time to relapse or death

stage 2 probabilities in Table V are conditional on $A_2 \cup A_3$, reflecting the way BMT specialists view each stage of patient response in this clinical setting. For example, to obtain an overall stopping probability in patient subgroup 4 when the true $p_{75} = 0{\cdot}85$ and the conditional probability of one-year survival is 0·40, denoting S_j = [stop at stage j], one simply computes $\text{pr}[S_1] + \text{Pr}[S_2 \mid \bar{S}_1] \times \text{Pr}[\bar{S}_1] = 0{\cdot}06 + 0{\cdot}13(1-0{\cdot}06) = 0{\cdot}18$.

Each of the final two examples has a more complex structure than those considered thus far, and they illustrate how we can implement the general approach when there are more possible patient outcomes and three or four monitoring bounds.

3.4. A two-agent anti-infection trial

Serious infections are a major complication of cancer chemotherapy, and trials of new antibiotics are constantly ongoing. No one antibiotic kills all infection-causing micro-organisms, and administration of one antibiotic to kill a pathogen may predispose the patient to infection by another, a phenomenon known as 'supra-infection'. Generally, the effects of an antibiotic can be evaluated within three days. These include cure of the infection, supra-infection or persistence of the original infection, or death.

The following design is based on the common idea in medicine that if a treatment proves to be effective in a patient then it should be continued, but if it is ineffective then a different treatment should be tried. Denote two antibiotics by D_1 and D_2. At each of two consecutive three-day periods, the patient has one of the three possible outcomes $I_- =$ [Alive and No Infection],

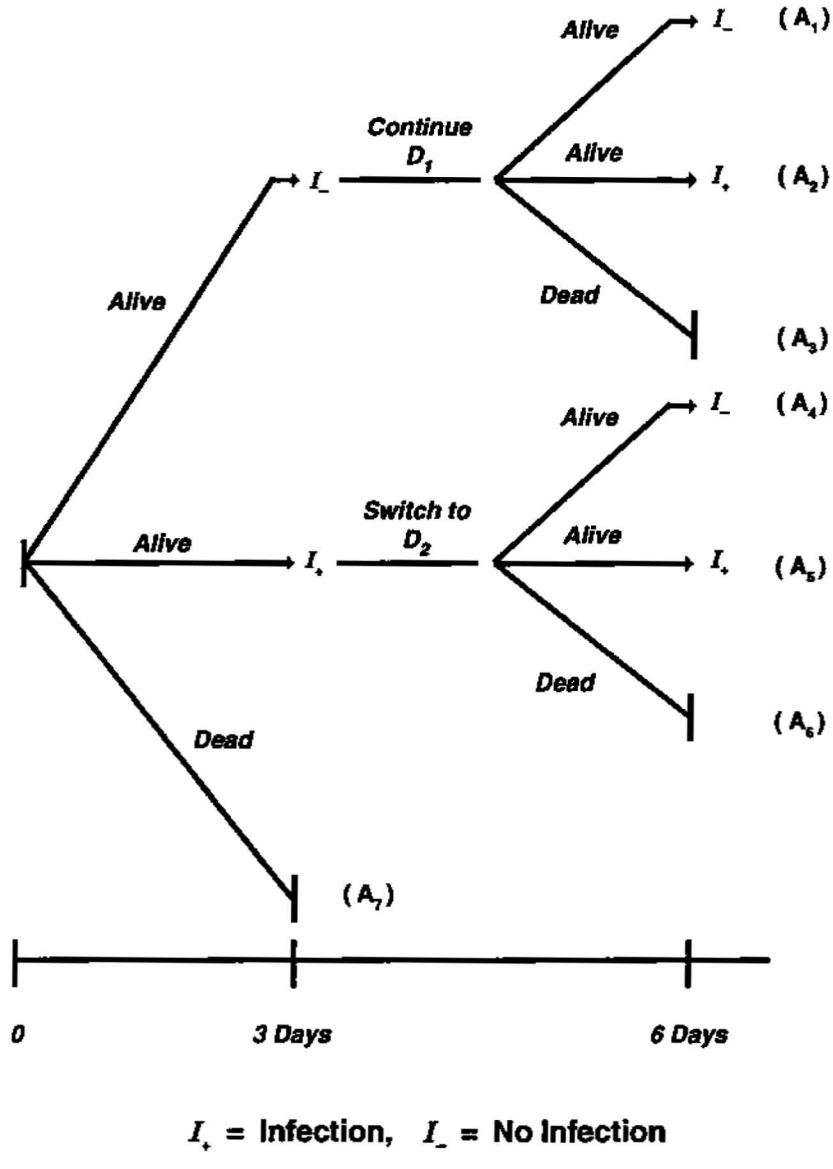


Figure 4. Schematic of anti-infection trial design

I_+ = [Alive and Infection] or [Dead]. All patients receive the anti-infection agent D_1 at the start of the trial, then are evaluated at day 3. If a patient is still alive and infected after three days (I_+), then D_1 is discontinued and D_2 is employed; otherwise, D_1 is continued. Patients with early success I_- receive D_1 again during the second stage. Each patient is re-evaluated at day 6. Figure 4 illustrates this scheme. Note that this design uses D_1 as a first choice with D_2 as a substitute if D_1 fails. If the clinician desires a symmetric comparison between D_1 and D_2, one could randomize patients to two arms, apply the above strategy in the first arm and reverse the roles of D_1 and D_2 in the second.

Here we use three monitoring criteria: (i) to improve the rate $\theta_1 = \Pr[A_1]$ of complete success with D_1; (ii) to improve the conditional rate $\tau_4 = \theta_4/(\theta_4 + \theta_5 + \theta_6)$ of stage 2 success among those with infections at stage 1, that is, of switching to D_2 if D_1 is not successful; and (iii) to

Table VI. Anti-infection trial operating characteristics

Case	True probabilities			Overall early stopping probability	Achieved sample size		
	p_1	c_4	p_{DEATH}		N_{25}	N_{50}	N_{75}
(1)*	0·48	0·43	0·17	0·82	15	31	61
(2)	0·455	0·375	0·27	0·94	11	21	38
(3)	0·48	0·58	0·14	0·71	17	37	82
(4)	0·63	0·43	0·15	0·42	43	82	82
(5)	0·63	0·58	0·09	0·12	82	82	82

$p_1 = \Pr[I_-$ at day 3 and at day 6], $c_4 = \Pr[I_-$ at day $6 \mid I_+$ at day 3]
* Null case

control the overall death rate $\theta_3 + \theta_6 + \theta_7$. The three stopping criteria are thus

$$\Pr[\theta_{S,1} + \delta(A_1) < \theta_{E,1} \mid \mathbf{X}_n] \leqslant p_L(A_1), \tag{18}$$

$$\Pr[\tau_{S,4} + \delta(A_4) < \tau_{E,4} \mid \mathbf{X}_n] \leqslant p_L(A_4), \tag{19}$$

and

$$\Pr[\theta_{S,3} + \theta_{S,6} + \theta_{S,7} + \delta(\text{Death}) < \theta_{E,3} + \theta_{E,6} + \theta_{E,7} \mid \mathbf{X}_n] \geqslant p_U(\text{Death}). \tag{20}$$

Table VI gives operating characteristics for this design with $(p_L(A_1), p_L(A_4), p_U(\text{Death})) = (0{\cdot}025, 0{\cdot}05, 0{\cdot}80)$ and $(\delta(A_1), \delta(A_4), \delta(\text{Death})) = (0{\cdot}15, 0{\cdot}15, 0{\cdot}05)$. The prior for the standard treatment probability vector had mean vector (0·48, 0·10, 0·02, 0·15, 0·10, 0·10, 0·05) with $W_{1,90} = 0{\cdot}20$. We determined the criterion probabilities following the general approach described in Section 2.3. In the null case 1, we set the true probabilities of complete success (p_1), stage 2 success among those with infections at stage 1 (c_4), and death (p_{DEATH}) equal to the corresponding means of θ_S; here the design has an 82 per cent chance of stopping early with a median of 31 patients. In case 2, where we increased p_{DEATH} by 0·10 over the standard mean of 0·17, the early termination probability is 0·94 with a median sample size of 21, so the design is highly protective against an increase in overall death rate. In case 3, we increased c_4 by the targeted $\delta(A_4) = 0{\cdot}15$ but left p_1 at its null value. In case 4, we increased p_1 by the targeted $\delta(A_1) = 0{\cdot}15$ but left c_4 at its null value. We regard each of cases 3 and 4 as a partial treatment improvement. In case 5, the best state of nature considered, we increased both p_1 and c_4 by their targeted values, and the trial has an 88 per cent chance of running to completion with 82 patients.

3.5. A general leukaemia chemotherapy trial

The following structure accommodates many bio-chemotherapy trials in acute leukaemia. It consists of two stages, each lasting three months, with CR and survival monitored during stage 1, and relapse and survival monitored during stage 2. Figure 5 gives the general structure and elementary events. In particular, we consider $\text{CR}_3 = [\text{Alive and in CR at 3 Months}] = A_3 \cup A_4 \cup A_5 \cup A_6$ early treatment success. An important point here is that a patient who enters CR but relapses prior to three months is in $A_2 = [\text{No CR, Alive}]$, hence is a treatment failure. Under the usual way of scoring in terms of early CR, one would count such an outcome as a success. We monitor patients in CR_3 for an additional three months, and partition CR_3 into one of the four elementary outcomes $A_3 = \text{CR}_3 \cap [\text{Relapsed, Dead}]$, $A_4 = \text{CR}_3 \cap [\text{No Relapse,}$

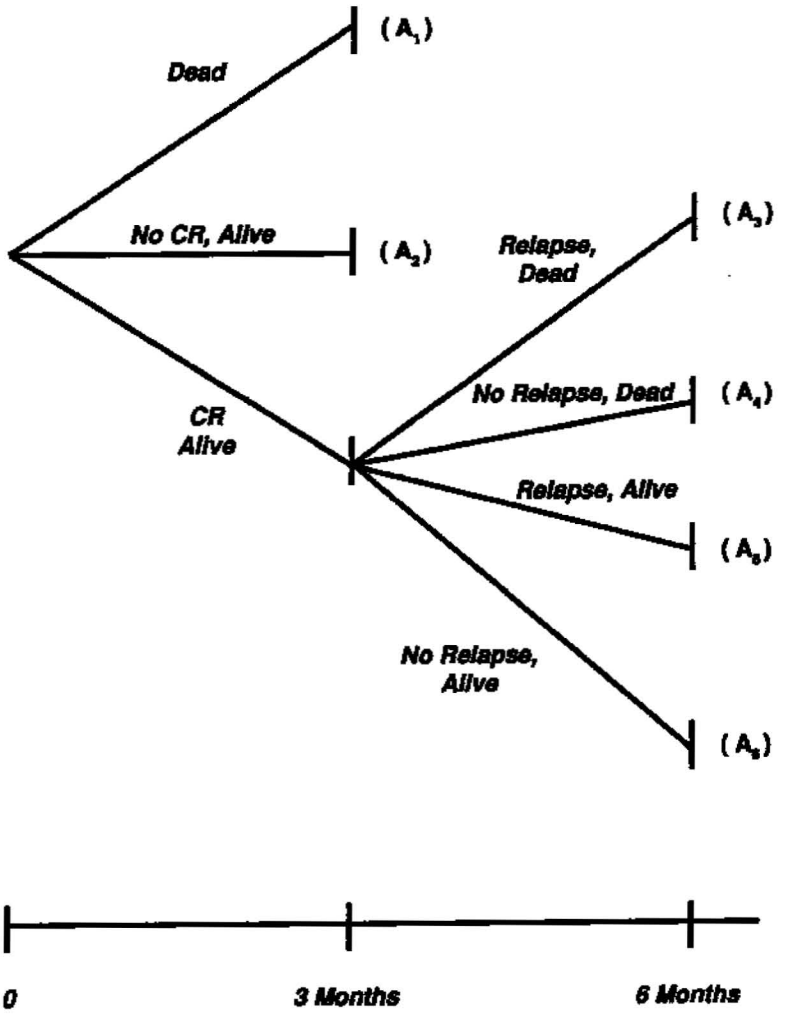


Figure 5. Schematic of general leukaemia bio-chemotherapy trial design

Dead], $A_5 = \mathrm{CR}_3 \cap$ [Relapse, Alive], $A_6 = \mathrm{CR}_3 \cap$ [No Relapse, Alive]. As before, for a patient who dies during the first three months, we score A_1 at month three and not at time of death. Likewise, we score stage 2 relapses and deaths at month 6, and not at the times of their occurrences. For example, we categorize a patient in CR_3 who subsequently relapses and then dies prior to month 6 as A_3 at month 6, with $X_{n+1,3} = X_{n,3} + 1$ at that time. This structure generalizes that used in the IAG trial, aside from the stage 1 timeframe, and we can similarly collapse or modify it to accommodate a particular clinical situation.

Suppose that the goals of the trial are to increase the conditional probability of six-month success $\Pr[A_6 \mid \mathrm{CR}_3] = \tau_6 = \theta_6/(\theta_3 + \theta_4 + \theta_5 + \theta_6)$ among those achieving CR, while controlling the early rates θ_1 and θ_2 of death and resistance and also the conditional six-month death rate $\tau_{3,4} = (\theta_3 + \theta_4)/(\theta_3 + \theta_4 + \theta_5 + \theta_6)$. The four stopping criteria are thus

$$\Pr[\theta_{S,1} + \delta(A_1) < \theta_{E,1} \mid \mathbf{X}_n] \geqslant p_U(A_1), \tag{21}$$

$$\Pr[\theta_{S,2} + \delta(A_2) < \theta_{E,2} \mid \mathbf{X}_n] \geqslant p_U(A_2), \tag{22}$$

Table VII. General leukaemia bio-chemotherapy trial operating characteristics

Case	True probabilities				Overall early stopping probability	Achieved sample size		
	p_1	p_2	$c_{3,4}$	c_6		N_{25}	N_{50}	N_{75}
(1)*	0·317	0·142	0·062	0·753	0·94	19	28	64
(2)	0·317	0·142	0·062	0·903	0·16	159	170	179
(3)	0·317	0·142	0·112	0·728	0·978	18	25	47
(4)	0·417	0·142	0·062	0·712	0·996	18	24	36
(5)	0·317	0·242	0·062	0·712	0·998	17	23	34
(6)	0·417	0·242	0·062	0·646	1·000	15	23	30

p_1 = Pr[Dead at month 3], p_2 = Pr[Alive, No CR at month 3],
$c_{3,4}$ = Pr[Dead at month 6 | CR_3], c_6 = Pr[Alive, in CR at Month 6 | CR_3]
* Null case

$$\Pr[\tau_{S,3,4} + \delta(A_3 \cup A_4) < \theta_{E,3,4} \mid \mathbf{X}_n] \geqslant p_U(A_3 \cup A_4) \tag{23}$$

and

$$\Pr[\tau_{S,6} + \delta(A_6) < \tau_{E,6} \mid \mathbf{X}_n] \leqslant p_L(A_6), \tag{24}$$

with monitoring carried out by comparing each of $X_{n,1}, X_{n,2}$ and $X_{n,3} + X_{n,4} \mid (n - X_{n,1}, - X_{n,2})$ to appropriate upper (equivalence) bounds, and comparing $X_{n,6} \mid (n - X_{n,1} - X_{n,2})$ to an apropriate lower (efficacy) bound, each at the time of update. If the clinician prefers to think of the stage 2 outcomes unconditionally, we can formulate (23) and (24) in terms of the corresponding unconditional probabilities $\theta_3 + \theta_4$ and θ_6.

To illustrate this general four-boundary monitoring design, we use historical data from 120 patients treated with idarubicin + ara-C (standard therapy) at M.D. Anderson Cancer Center during 1992 and 1993 to obtain the standard prior parameters $a_S = (38, 17, 2, 2, 12, 49)$. Thus the standard mean rates of early (three month) death and resistance are, respectively, 31·7 per cent and 14·2 per cent, while the conditional mean rates of stage 2 death and success, among those alive and in CR at three months, are, respectively, 6·2 per cent and 75·4 per cent. The trial was designed to achieve a $\delta(A_6) = 0{\cdot}15$ improvement in τ_6 to 90·4 per cent while maintaining 0·05-equivalence in the each of θ_1, θ_2 and in the late death rate $\tau_{3,4}$. We specified a maximum stage 1 sample size of 94 patients alive and in CR at month 3, that is, in CR_3, to ensure that a 90 per cent posterior probability interval for the conditional stage 2 success probability τ_6 would have width 0·10. As before, we determined the probability criteria $p_U(A_1) = p_U(A_2) = p_U(A_3 \cup A_4) = 0{\cdot}90$ and $p_L(A_6) = 0{\cdot}05$ following the general approach described in Section 2.3.

Table VII gives this design's operating characteristics for various fixed values of the probabilities of the events monitored. These are p_1 = Pr[Dead at Month 3], p_2 = Pr[Alive But Not in CR (Resistant) at Month 3], $c_{3,4}$ = Pr[Dead at Month 6 | CR_3], and c_6 = Pr[Alive and in CR at Month 6 | CR_3]. Case 1 is the null case, as in evaluation of the anti-infection trial design, and here the trial is highly likely to terminate early, although the sample size distribution is somewhat skewed to the right. Case 2 represents treatment success, with the conditional stage 2 success rate c_6 increased by the targeted 0·15, and in this case the trial runs to completion with probability 0·84. In case 3 the late death rate $c_{3,4}$ increases by 0.05, in cases 4 and 5 the early death and resistance rates each increase by 0·10 while the other remains at its null rate, and in case 6 both p_1 and p_2 increase by 0·10. Cases 3–6 represent different ways in which the rates of death or

resistance for E are higher than those of S, and in all of these cases the trial almost certainly terminates with a relatively small number of patients.

4. DISCUSSION

Phase II clinical trials are medical studies that assist in the determination of what treatments to study in large-scale randomized comparative (phase III) trials, and in the design of phase III trials. Phase II trials often utilize short-term endpoints, such as tumour shrinkage, in contrast to phase III trials where survival or disease progression are usually the primary measures of treatment effectiveness. Moreover, most statistical methodologies for phase II trials assume that there is a single endpoint of interest (see Gehan,[21] Fleming,[22] Sylvester,[23] Simon,[24] Thall and Simon[1,2,25,26]). The determination of whether a new regimen is sufficiently promising for phase III study is usually a complex, multi-faceted process, however, and the actual conduct of many phase II trials is more complicated than a design based on a single binary efficacy outcome variable may indicate. There are often several intermediate measures of treatment efficacy, as well as important adverse outcomes such as toxicity.

Although toxicity is generally dealt with informally in the design and monitoring of clinical trials, it is often a key issue in the decision of whether to terminate a trial early or to continue development of a new treatment. Such early stopping rules sometimes are mentioned in trial protocols, but typically they are ignored in the computation of the design's operating characteristics and in planning the sample size. Whereas there is generally an urgent need to terminate a trial if the observed toxicity rate is unacceptably high or if the efficacy event rate is too low, a larger sample size is often desirable when such problems do not occur. We have adopted a design philosophy that takes estimation of the primary efficacy endpoint probability distribution as the main objective of the trial, with early termination should the rate of any adverse event prove to be unacceptably high. In this context adverse events include both toxicity and failure to achieve an efficacy outcome. Our approach accounts for the distinction between adverse and desirable outcomes, and has the simultaneous goals of controlling the rate of the former while improving the rate of the latter. Moreover, by accounting for multiple outcomes , it provides a framework for monitoring both early and late patient responses, so that the sequential, interactive nature of treatment and response may be accommodated. Use of this methodology *per se* in co-operative studies involving many hospitals would be problematic, however, in that continuous monitoring would likely be difficult, hence a group sequential version might be more appropriate in such circumstances. For trials involving rapidly fatal diseases, however, one would then lose much of the protective aspect of the method.

Our proposed monitoring strategies use Bayesian criteria to construct early stopping rules, but we perform frequentist evaluation of their operating characteristics under fixed values of the event probabilities. This is Bayesian inference, because it is based on the information in the posterior. We do not use a decision-theoretic framework, however. We make a distinction between the probability distributions on θ_S and θ_E, which reflect the investigators' prior experience and possibly historical data, and an assumed state of nature expressed as a fixed value of θ_E. Given the decision boundaries, frequentist evaluation of the design under fixed parameter values is objective, and moreover it is easily communicated to both statisticians and physicians. Evaluation of the operating characteristics under an array of possible design parameterizations is a simple, practical means of obtaining a design which appeals to the clinician, reflects actual clinical practice, and has good statistical properties. Naturally, a Bayesian is free to evaluate the final data in any manner desired, based on the posterior distribution of θ, or based on a set of posteriors

corresponding to other priors of interest. A frequentist may form confidence intervals or test hypotheses conditional on the monitoring process and trial outcome.

Several important issues still remain. These include analysis of the method's sensitivity to the Dirichlet prior and possible extension to a more complex model for categorical outcomes, generalization of the model to accommodate continuous responses without discretizing them, and incorporation of individual patient prognostic variables. We chose the Dirichlet-multinomial model because it quantifies prior information and accumulating data in a simple and reasonable manner, and it is highly tractable. In our experience applying the method, we have found the categorical structure to be quite adaptable to a broad variety of clinical settings, the discetization of continuous variables notwithstanding. We are not aware of any other method for dealing effectively with multiple outcomes at the level of complexity illustrated by our applications. Moreover, we regard single-outcome phase II designs as the standard of statistical practice upon which we wish to improve. Consequently, we believe that our proposed method provides a substantial improvement over existing methods currently employed in the design and conduct of single-arm trials.

Still, the simplicity and tractability of our approach must be weighed against the advantages of models that have more parameters or that accommodate time-to-event variables directly. One limitation of our model is that θ_S and θ_E may not be independent, and an extension that accounts for their joint distribution would be more appropriate in such settings. Another problem is that, in specifying a prior for θ_S through elicitation of the dispersion parameter W_S, different reference events will likely lead to different priors. A multivariate normal prior on the logits of the entries of (θ_S, θ_E) is one reasonable way to deal with these problems, although the associated numerical computations would be considerably more complex.

The extension to continuous-time models is straightforward in the univariate case. Thall and Simon[26] describe the use of a gamma-exponential model for monitoring a single time-to-event outcome. Extension of our approach to multiple outcomes is also straightforward in some cases. For example, the application in Section 3.1 could be modelled with a bivariate log-normal distribution for the times to transplant rejection and GVHD. Situations involving competing risks or outcomes that depend on the occurrence of previous events are more difficult to model in a general continuous-time framework, however. These problems are obviated by discretizing continuous variables and categorizing outcomes exhaustively.

Another limitation of our method, as with most clinical trial designs, is that it does not account for individual patient covariates. Between-patient variability typically is quite large in clinical trials, and it may mask treatment effects. We are currently investigating an extension which incorporates patient covariate data while providing a more refined parameterization.

ACKNOWLEDGEMENTS

The authors thank Derek Jacoby for computer programming. We also thank Dennis Dixon, Joan Staniswalis, two referees and an editor for numerous constructive comments and suggestions.

REFERENCES

1. Thall, P. F. and Simon, R. 'Practical Bayesian guidelines for phase IIB clinical trials', *Biometrics*, **50**, 337–349 (1994).
2. Thall, P. F. and Simon, R. 'A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials', *Controlled Clinical Trials*, (1994). (In press)
3. Smith, D. G., Clemens, J., Crede, W., Harvey, M., and Gracely, H. J 'The impact of multiple comparisons in randoized clinical trials', *The American Journal of Medicine*, **83**, 545–550 (1987).

4. O'Brien, P. C. Procedures for comparing samples with multiple endpoints', *Biometrics*, **40**, 1079–1087 (1984).
5. Pocock, S. J., Geller, N. L. and Tsiatis, A. A. 'The analysis of multiple endpoints in clinical trials', *Biometrics*, **43**, 487–498 (1987).
6. Tang, D.-I., Gnecco, C. and Geller, N.L. 'Design of group sequential clinical trials with multiple endpoints', *Journal of the American Statistical Association*, **84**, 577–583 (1989).
7. Tang, D.-I., Geller, N. L. and Pocock, S. J. 'On the design and analysis of clinical trials with multiple endpoints', *Biometrics*, **49**, 23–30 (1993).
8. Lehmacher, W., Wassmer, G. and Reitmeir, P. 'Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate', *Biometrics*, **47**, 511–521 (1991).
9. Gelber, R. D., Gelman, R. S. and Goldhirsch, A. 'A quality-of-life-oriented endpoint for comparing therapies', *Biometrics*, **45**, 781–795 (1989).
10. Jennison, C. and Turnbull, B. W. 'Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints', *Biometrics*, **49**, 741–752.
11. Spiegelhalter, D. J. and Freedman, L. S. 'Bayesian approaches to clinical trials (with discussion)', *in* Bernardo, J. M. DeGroot, M. H., Lindley, D. V. and Smith A. F. M. (eds), *Bayesian Statistics 3*, Clarendon Press, Oxford, 1988, pp. 453–477.
12. Ho, C. H. 'Some frequentist properties of a Bayesian method in clinical trials', *Biometrical Journal*, **33**, 735–740 (1991).
13. Etzioni, R. and Pepe, M. S. 'Monitoring of a pilot toxicity study with two adverse outcomes', *Statistics in Medicine* (In press). 1994.
14. Dixon, D. O. and Duncan, D. B. 'Minimum Bayes risk t-intervals for multiple comparisons', *Journal of the American Statistical Association*, **68**, 117–130 (1975).
15. Louis, T. A. 'Estimating a population of parameter values using Bayes and empirical Bayes methods', *Journal of the American Statistical Association*, **79**, 393–398 (1984).
16. Berry, D. A. 'Interim analyses in clinical trials: Classical vs. Bayesian approahces', *Statistics in Medicine*, **4**, 521–526 (1985).
17. Berry, D. A. 'Multiple comparisons, multiple tests and data dredging: A Bayesian perspective', *in* Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith A. F. M. (eds), *Bayesian Statistics 3*, Clarendon Press, Oxford, 1988, pp. 79–94.
18. Freedman, L. S. and Spiegelhalter, D. J. 'The assessment of subjective opinion and its use in relation to stopping rules for clinical trials', *The Statistician*, **32**, 153–160 (1983).
19. Freedman, L. S. and Spiegelhalter, D. J. 'Comparison of Bayesian with group sequential methods for monitoring clinical trials', *Controlled Clinical Trials*, **10**, 357–367 (1989).
20. Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. *Numerical Recipes in C*, Cambridge University Press, New York, 1988.
21. Gehan, E. A. 'The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent', *Journal of Chronic Diseases*, **13**, 346–353 (1961).
22. Fleming, T. R. 'One sample multiple testing procedure for phase II clinical trials', *Biometrics*, **38**, 143–151 (1982).
23. Sylvester, R. J. 'A Bayesian approach to the design of phase II clinical trials', *Biometircs*, **44**, 823–836 (1989).
24. Simon, R. 'Optimal two-stage designs for phase II clinical trials', *Controlled Clinical Trials*, **10**, 1–10 (198).
25. Thall, P. F. and Simon, R. 'Incorporating historical control data in planning phase II clinical trials', *Statistics in Medicine*, **9**, 215–228 (1990).
26. Thall, P. F. and Simon, R. 'Bayesian design and monitoring of phase II clinical trials', *Proceedings of the XVIth International Biometric Conference*, Hamilton, New Zealand, 1992, pp. 205–220.