STATISTICS IN MEDICINE
Statist. Med. **17**, 1563–1580 (1998)

SOME EXTENSIONS AND APPLICATIONS OF A BAYESIAN STRATEGY FOR MONITORING MULTIPLE OUTCOMES IN CLINICAL TRIALS

PETER F. THALL* AND HSI-GUANG SUNG

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

SUMMARY

We present some practical extensions and applications of a strategy proposed by Thall, Simon and Estey for designing and monitoring single-arm clinical trials with multiple outcomes. We show by application how the strategy may be applied to construct designs for phase IIA activity trials and phase II equivalence trials. We also show how it may be extended to incorporate the use of mixture priors in settings where a Dirichlet distribution does not adequately quantify prior experience, randomized phase II selection trials involving two or more experimental treatments, and trials with group-sequential monitoring for applications involving multiple institutions. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

This paper presents some practical extensions and applications of a strategy proposed by Thall, Simon and Estey,[1] hereafter TSE, for designing and monitoring single-arm clinical trials with multiple outcomes. The strategy generalizes the method of Thall and Simon[2,3] for phase II trials with one binary outcome. Rather than proposing a particular design, TSE presented a general strategy for constructing designs. The strategy allows one to tailor the design of a given trial to accommodate its specific patient outcome structure, scientific goals and safety requirements. TSE used Bayesian criteria to determine sample size and generate early stopping boundaries, and evaluated the design's frequentist operating characteristics via simulation. Owing to its practicality and generality, and the availability of a menu-driven computer program for implementation, this strategy has been used to design numerous clinical trials conducted at M.D. Anderson Cancer Center and other medical institutions.

This broad usage has generated a great deal of feedback. In communicating with statisticians and physicians using the TSE method, we found that some important applications of the strategy are not apparent from the five illustrations provided by TSE. These applications include phase IIA 'activity' trials of potential new anti-cancer agents and phase II 'equivalence' trials. Our experience also has shown the need for certain practical extensions, including the use of mixture

* Correspondence to: Peter Thall, Department of Biomathematics, Box 237, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, U.S.A. E-mail: rex@odin.mdacc.tmc.edu

CCC 0277–6715/98/141563–18$17.50

Received February 1997
Revised September 1997

priors in settings where a Dirichlet distribution may not adequately quantify prior experience, randomized phase II selection trials involving two or more experimental treatments and trials with discontinuous monitoring. We illustrate how the strategy may be applied or extended to provide designs in each of these settings. The orientation throughout is toward practical application. Our aim is to enable practitioners to use the strategy in a wider variety of clinical scenarios than those described in TSE.

While the focus of this paper is statistical methodology, out of necessity we also address computational issues. First, several clinical trial scenarios that have arisen in various applications are not readily accommodated by the original computer program that we provided. In addition, Lazaridis and Gonin[4] pointed out some inaccuracies in the numerical methods used in the program. To address these problems, we have written a new menu-driven computer program which includes all of the previous program's capabilities, accommodates the extensions described in this paper, and uses more reliable numerical integration and simulation routines. All of the computations described in the sequel were carried out on a Sun SPARC Station 20 or DEC AlphaServer 2100 using this new program. Each simulated trial scenario reported here is based on 10,000 replications. The program is freely available via anonymous ftp from odin.mdacc.tmc.edu as 'multcomp97.tar.gz' in the subdirectory/pub/source.

The remainder of the paper is organized as follows. We first describe the TSE strategy, including its underlying philosophy and guidelines for practical application. We then illustrate by example application of the strategy to phase II equivalence trials and phase IIA trials, the use of discrete mixture priors, trials with discontinuous monitoring, and randomized phase II trials. We close with a brief discussion.

2. REVIEW OF THE BASIC STRATEGY

We first review the TSE strategy in the context of a continuously monitored single-arm trial of an experimental treatment evaluated relative to a standard treatment. Subsequently, we build on this basic structure as necessary.

2.1. Motivation

Our experience arises primarily in oncology, where adverse treatment effects such as severe toxicity or regimen-related death occur routinely and the case of multiple patient outcomes is quite common. Consequently, the strategy is oriented toward safety monitoring and accommodates multiple events. The general goal is to provide physicians who conduct a phase II trial of an experimental treatment with an ethical and scientifically reasonable basis for deciding whether to stop the trial early.

For the case of a single binary efficacy outcome, the phase IIA design of Gehan[5] and the group sequential phase II designs of Schultz *et al.*,[6] Fleming,[7] Chang *et al.*,[8] Therneau *et al.*[9] and Simon[10] each includes interim stopping rules, while Thall and Simon[2,3] provide designs with continuous monitoring. Monitoring efficacy alone may be inadequate, however. A common phase II scenario is one where a physician has been provided with a statistical design based on a binary efficacy outcome, but must rely on intuition and clinical experience to decide whether to stop a trial early if an adverse event rate seems too high. Simple probability computations often show that early stopping rules actually used by physicians in such circumstances may have very undesirable properties. Moreover, the nominal operating characteristics of any design based on

evaluation of response but ignoring likely adverse events are a fiction. Regardless of what may appear in a written protocol, physicians do not use designs that do not reflect clinical reality, especially with regard to safety monitoring. Oncologists will certainly terminate a trial if they believe that the observed rate of an adverse event is unacceptably high, or that the response rate is too low, and they will do so whether or not the statistical design provides rules for such decisions.

There is an extensive literature on group sequential methods for monitoring randomized trials.[11–14] A discussion of the distinction between formal statistical rules and practical aspects of data monitoring is given by Friedman *et al.*,[11] chapter 12. In recent years there has been a substantial increase in papers proposing practical Bayesian methods.[15–20] There are few practical methods available that provide physicians with a means to quantify the clinical experience and standards underlying early stopping decisions based on multiple outcomes, however. In practice, clinicians who conduct trials often must rely on informal judgement alone for safety monitoring. The available methods that accommodate multiple outcomes in phase II appear to be the general Bayesian strategy of TSE[1,21] and the hypothesis test-based designs for trials with a bivariate binary (efficacy, toxicity) outcome proposed by Bryant and Day[22] and Conaway and Petroni.[23,24] Designs for randomized trials with (efficacy, toxicity) outcome have been proposed by Jennison and Turnbull[25] and Cook and Farewell.[26]

2.2. Model

Early stopping decisions are inherently comparative. We formalize this by first specifying a standard treatment S against which the experimental treatment E will be evaluated, even though there is no S arm in the trial. In practice, S may refer to one treatment or a composite of several. The trial of E is terminated if, compared to prior clinical experience with S, the observed rate of one or more adverse outcomes is unacceptably high, the observed rate of an efficacy outcome is unacceptably low, or possibly if the efficacy rate is so high that it is desirable to report the results immediately and organize a phase III trial. In making these comparisons, an important scientific requirement is that the variability or uncertainty about the patient outcome probabilities $\boldsymbol{\theta}_S$ under S be quantified honestly. Thus, a prior for $\boldsymbol{\theta}_S$ is required. A maximum of M patients are treated if the trial is not stopped early, with M chosen to obtain posterior probability estimates that have a given level of reliability.

The following model was chosen by TSE to accommodate a broad array of clinical scenarios, hence it is rather simple. Many elaborations are possible, and we discuss some possibilities in Sections 5 and 8. The design process begins by working with the physician to specify the partition $\{A_1, \ldots, A_K\}$ of all possible elementary patient outcomes. This must be done in such a way that each event to be monitored is the union of one or more of the A_j's. For a patient treated with $t =$ E or S, the probability vector corresponding to the elementary events is $\boldsymbol{\theta}_t = (\theta_{t,1}, \ldots, \theta_{t,K-1})$, with $\theta_{t,K} = 1 - \theta_{t,1} \cdots - \theta_{t,K-1}$. The probability $\eta_t(C)$ that a patient treated with t experiences a given compound event C is the sum over the subvector of $\boldsymbol{\theta}_t$ corresponding to the elementary events comprising C. Each patient's outcome is characterized by a vector $\mathbf{Y} = (Y_1, \ldots, Y_K)$ with a single indicator $Y_j = 1$ if the elementary outcome A_j occurred and the remaining $K - 1$ entries 0. The sum over the first n patients is $\mathbf{Y}_1 + \cdots + \mathbf{Y}_n = \mathbf{X}_n = (X_{n,1}, \ldots, X_{n,K})$.

The model assumptions are (i) $\mathbf{X}_n \mid \boldsymbol{\theta}_E$ is multinomially distributed with parameters n and $\boldsymbol{\theta}_E$ and (ii) *a priori* $\boldsymbol{\theta}_t$ follows a Dirichlet distribution with parameters $\mathbf{a}_t = (a_{t,1}, \ldots, a_{t,K})$, denoted $\boldsymbol{\theta}_t \sim \mathrm{Dir}(\mathbf{a}_t)$, for $t =$ E, S. The probability density function of the $\mathrm{Dir}(a_1, \ldots, a_K)$ distribution for

the probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ of a partition of K events is given by

$$f_{\boldsymbol{\theta}}(p_1, \ldots, p_k \,|\, a_1, \ldots, a_K) = \Gamma(a_0) \prod_{j=1}^{K} \frac{p_j^{a_j - 1}}{\Gamma(a_j)}$$

where we denote $a_0 = \sum_{i=1}^{K} a_j$, $\Gamma(\cdot)$ is the gamma function, $p_1 + \cdots + p_K = 1$, $0 \leqslant p_j \leqslant 1$ for all j and each parameter $a_j > 0$. The mean vector $\boldsymbol{\mu} = E(\boldsymbol{\theta}) = (a_1, \ldots, a_K)/a_0$ and $\mathrm{var}(\theta_j) = \mu_j(1 - \mu_j)/(1 + a_0)$. Due to the constraint that the arguments sum to 1 this is a $(K - 1)$-variate distribution, and when $K = 2$ this is the well-known univariate beta distribution with parameters a_1 and a_2. In particular, each compound event probability $\eta_t(C)$ follows a beta distribution, each marginal count $X_n(C)$ is conditionally binomial in n and $\eta_{\mathrm{E}}(C)$, and *a posteriori* $\boldsymbol{\theta}_{\mathrm{E}} | \mathbf{X}_n \sim \mathrm{Dir}(\mathbf{a}_{\mathrm{E}} + \mathbf{X}_n)$.

The $\mathrm{Dir}(\mathbf{a}_{\mathrm{S}})$ prior should reflect clinical experience with S whereas the $\mathrm{Dir}(\mathbf{a}_{\mathrm{E}})$ prior should reflect the fact that E is a new treatment. Thus, an informative prior on $\boldsymbol{\theta}_{\mathrm{S}}$ is used. When historical data consisting of the K elementary outcome counts are available they may be used as the Dirichlet parameters $\mathbf{a}_{\mathrm{S}}$. Otherwise, one must work harder to elicit the $\mathrm{Dir}(\mathbf{a}_{\mathrm{S}})$ prior. A non-informative or weakly informative prior on $\boldsymbol{\theta}_{\mathrm{E}}$ is used, typically characterized by $a_{\mathrm{E},1} + \cdots + a_{\mathrm{E},K} = K$ and $\boldsymbol{\mu}_{\mathrm{E}} = \boldsymbol{\mu}_{\mathrm{S}}$, although $\boldsymbol{\mu}_{\mathrm{E}}$ may be adjusted to reflect either an optimistic or pessimistic prior.

The posterior probability $\lambda(C, \delta) = \Pr[\eta_{\mathrm{S}}(C) + \delta < \eta_{\mathrm{E}}(C) | \mathbf{X}_n]$ is used as an early stopping criterion for each compound event C to be monitored, where δ is a fixed constant specified by the clinician. Although $\lambda(C, \delta)$ depends on $\mathbf{X}_n$ and the priors, for brevity we suppress these additional arguments. For an adverse event T the trial stops if $\lambda(T, \delta_T) > p_{\mathrm{U}}(T)$, where $p_{\mathrm{U}}(T)$ is a fixed upper probability cut-off and δ_T is a small non-negative *slippage*, typically in the range $0 \leqslant \delta_T \leqslant 0{\cdot}10$. The slippage quantifies the maximum amount of increase in $\eta_{\mathrm{S}}(T)$ that one will tolerate. For an efficacy outcome R the trial stops if $\lambda(R, \delta_R) < p_{\mathrm{L}}(R)$, corresponding to an unacceptably low rate of R, where δ_R is a targeted efficacy improvement, typically $0{\cdot}15 \leqslant \delta_R \leqslant 0{\cdot}25$. If it is also desired to stop the trial early if E is promising compared to S with regard to $\eta(R)$, then one uses the additional criterion $\lambda(R, \delta_R) > p_{\mathrm{U}}(R)$. If a goal of the trial is to decrease the probability $\eta(T)$ of an adverse event T by a target $\delta_T > 0$, then one may simply use the efficacy criterion $\lambda(\bar{T}, \delta_{\bar{T}}) < p_{\mathrm{L}}(\bar{T})$, defined in terms of the complement $\bar{T}$ of T. This is identical to $\lambda(T, -\delta_{\bar{T}}) > 1 - p_{\mathrm{L}}(T)$. In oncology, it is often the case that E is likely to increase the probabilities of both an adverse outcome T and an efficacy outcome R. An approach that we find useful in such settings[21] is to ask the physician to specify δ_T as the maximum allowed increase in $\eta(T)$ which is an *acceptable trade-off* for a desired increase δ_R in $\eta(R)$. Given minimum sample size m, the probability criterion for each C generates a stopping boundary in terms of $\{X_n(C), n = m, \ldots, M\}$. We use the multiple stopping bounds together, and E is considered promising compared to S in terms of the defined safety and efficacy outcomes if the trial does not terminate early. To obtain confirmatory comparative evaluation of treatment effects, however, comparison to historical controls is not a substitute for a randomized trial.

2.3. Evaluation of operating characteristics

First, several vectors $\mathbf{p} = (p_1, \ldots, p_K)$ of fixed probabilities, each characterizing a clinical scenario of interest, are specified. The trial is simulated and its operating characteristics (OCs) are evaluated under each scenario. The OCs consist of early stopping probabilities, possibly broken down by reason for stopping, and achieved sample size distribution, which may be summarized

conveniently in terms of selected percentiles. Thus, we are concerned with the likelihood that the trial stops early and the distribution of the number of patients treated, rather than the conventional frequentist type I and type II error probabilities typically associated with a test of hypothesis.

If some aspect of either the OCs or the boundaries is undesirable to the clinician, then one modifies the design parameters accordingly and repeats the simulations. This usually involves adjusting the probability cut-offs p_U and p_L, although evaluation of OCs may provide insights that motivate changes in other design parameters. This is iterated until the clinician is happy with all aspects of the design. In our experience, the design process may move from an individual physician to a group at the section or department level, where more extensive experience may motivate changes in fundamental model components such as the outcome set, standard prior or trial goals. Naturally, the design and its operating characteristics depend on the two priors, and a sensitivity analysis may be carried out if desired.

Specification of each fixed vector $\mathbf{p}$ may not be entirely straightforward. For example, in the bivariate (response, toxicity) outcome case where $\mathbf{p} = (p_1, p_2, p_3, p_4)$ and the marginal probabilities are $p_1 + p_2 = \Pr[\text{response}]$ and $p_2 + p_4 = \Pr[\text{toxicity}]$, there are infinitely many $\mathbf{p}$ for each pair $(p_1 + p_2, p_2 + p_4)$. For example, the scenario in which Pr[response] increases by 0·15 and Pr[toxicity] increases by 0·05, compared to their standard mean values obtained from $\mathbf{p} = \boldsymbol{\mu}_S$, we require a third parameter to specify $\mathbf{p}$ fully. This general problem was pointed out by Lazaridis and Gonin[4] in studying the bone marrow transplantation (BMT) application described by TSE. They argued that, when evaluating OCs in the 2×2 setting, one must account for the sensitivity to the third parameter, which may be characterized as the joint probability p_2 of both response and toxicity, as a conditional probability, or as an odds ratio. In more complex settings, we suggest working with the physician to characterize each scenario by starting with $\boldsymbol{\mu}_S$ and moving probability mass from one elementary outcome to another in a way that makes sense clinically.

Although the frequentist properties of our designs with which we are primarily concerned consist of the probability that the trial terminates early and the sample size distribution, the designs also have a more common frequentist interpretation. One may consider the parameters as being the fixed values $\boldsymbol{\mu}_E$ and $\boldsymbol{\mu}_S$, the null hypothesis to be $\boldsymbol{\mu}_E = \boldsymbol{\mu}_S$, and the alternative to be some value of $\boldsymbol{\mu}_E$ which is associated with the general conclusion that E is promising compared to S. If one then considers early termination of the trial as acceptance of the null and continuation to the maximum M as acceptance of the alternative, then under continuous monitoring what we call the early stopping probability π is a type II error under the alternative and $1 - \pi$ is a type I error under the null. To complete this frequentist interpretation under group-sequential monitoring, which we discuss in Section 7, an additional final decision would be required at the completion of a trial that does not terminate early. This could be done by simply recording whether or not the probability criteria for early stopping are met at the end of the trial when computing the overall stopping probability of the trial under the given null and alternative values of $\boldsymbol{\mu}_E$. Alternatively, one could require that a more specific criterion such as $\Pr[\eta_S(C) < \eta_E(C) | \text{final data}] > p_U$ be satisfied for primary efficacy outcome C and some large cut-off probability p_U to declare E promising, or to 'reject the null' under the frequentist interpretation. Thus, our Bayesian criteria could be used to develop purely frequentist designs. In this regard, however, we feel that such a hypothesis testing framework may be somewhat misleading in the context of a phase II trial. In general, the best that can be accomplished in a single arm trial of an experimental treatment is that the patients in the trial are protected from an unsafe or inefficacious treatment by formal

stopping rules, and, if the trial does not terminate early, reasonably reliable estimates of the various outcome probabilities can be obtained to determine if E is promising.

3. PHASE II EQUIVALENCE TRIALS

To avoid ambiguity, we first make the following distinction between phase III, or *confirmatory* equivalence, and phase II equivalence. Let η denote the probability of a single binary efficacy outcome. For given slippage $\delta \geqslant 0$, typically in the range $0{\cdot}05 \leqslant \delta \leqslant 0{\cdot}10$, we say that *E is δ-equivalent to S* in the confirmatory sense if *a posteriori* $\lambda(-\delta) = \Pr[\eta_S - \delta < \eta_E \,|\, \text{data}]$ is large and moreover this probability is based on data from a randomized trial of E versus S. This is a Bayesian analogue of the common frequentist approach to establishing equivalence in which, given a fixed standard μ_0, a randomized trial is conducted to test the null hypothesis $\mu_E \leqslant \mu_0 - \delta$ versus the alternative $\mu_E > \mu_0 - \delta$, with the size computed at $\mu_E = \mu_0 - \delta$ and the power usually computed at $\mu_E = \mu_0$. We conduct a *phase II equivalence trial* in such a way that, to continue the trial to the $n + 1$st patient, we require only that $\lambda_n(-\delta) = \Pr[\eta_S - \delta < \eta_E \,|\, \mathbf{X}_n] \geqslant p_L$ for small p_L. The practical point is simply that the phase II trial terminates early if the posterior probability of a slippage no larger than δ is small. Otherwise, the completed phase II trial provides evidence that there is some hope of subsequently establishing confirmatory equivalence, or even superiority of E over S, via a large randomized trial.

The following application illustrates settings where E embodies a qualitative innovation over S, hence it is appropriate to require only phase II equivalence rather than improvement over S to continue the trial. Currently available treatments for patients with metastatic breast cancer are unlikely to provide a cure. Although autologous BMT provides a complete remission rate over 50 per cent, median remission duration is only about one year. Success in treating other diseases by transplanting autologous peripheral blood progenitor cells (PBPCs) rather than bone marrow cells motivated a trial of this therapeutic modality for metastatic breast cancer patients. The experimental treatment began with a myeloablative regimen consisting of doxorubicin, paclitaxel and cyclophosphamide plus cytokine support followed by PBPC infusion. Patient outcomes were scored over the first four months post-transplant. As illustrated in Figure 1, the relevant compound events were death ($D = A_5$) and, among patients who survived four months, complete remission ($CR = A_2 \cup A_4$) and severe, grade $\geqslant 3$ toxicity ($TOX = A_3 \cup A_4$). The outcome space was constructed in this way because it was desired to score CR and TOX only among patients who survived, rather than using the combined adverse outcome D$\cup$TOX as is often done in other trials. The clinician specified prior means (0·34, 0·55, 0·02, 0·03, 0·06) based on experience with approximately 300 patients, hence a Dir(102, 165, 6, 9, 18) prior was used for $\boldsymbol{\theta}_S$. The standard mean probabilities of the three outcomes to be monitored were thus $\mu_S(CR) = 0{\cdot}58$, $\mu_S(TOX) = 0{\cdot}05$ and $\mu_S(D) = 0{\cdot}06$. The trial goals were to maintain equivalent rates of each of these events, with long-term goal to estimate disease-free survival if the trial did not terminate early. Thus, the early stopping criteria were $\lambda(CR, 0) < p_L(CR)$, $\lambda(TOX, 0) > p_U(TOX)$ and $\lambda(D, 0) > p_U(D)$. The values $p_U(TOX) = 0{\cdot}99$, $p_U(D) = 0{\cdot}98$ and $p_L(CR) = 0{\cdot}06$ were used to obtain desirable OCs. A maximum sample size of $M = 54$ was specified to ensure a 95 per cent posterior probability interval for $\eta_E(CR)$ of width $\leqslant 0{\cdot}25$. Specifically, if 31/54 (57·4 per cent) CRs are observed, then $\Pr[\xi_{0{\cdot}025} < \eta_E(CR) < \xi_{0{\cdot}975} \,|\, X_{54}(CR) = 31] = 0{\cdot}95$ with the percentiles satisfying $\xi_{0{\cdot}975} - \xi_{0{\cdot}025} = 0{\cdot}697 - 0{\cdot}448 = 0{\cdot}249$. A minimum sample size $m = 6$ was used, although the stopping bound $X_6(TOX) \geqslant 3$ required that $X_n(TOX) \geqslant 3$ also be applied for $n = 3$, 4 and 5, with a similar runback for D.

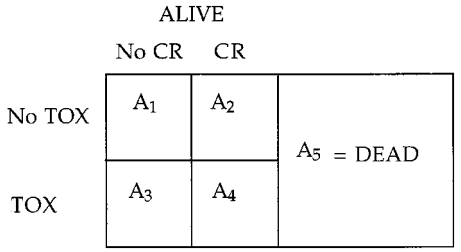

Figure 1. Patient outcomes for the metastatic breast PBSC transplantation trial

Table I. Metastatic breast cancer PBSC transplantation equivalence trial operating characteristics

Clinical scenario						Sample size percentiles				
p_1	p_2	p_3	p_4	p_5	π	N_{10}	N_{25}	N_{50}	N_{75}	N_{90}
		Null case								
0·34	0·55	0·02	0·03	0·06	0·20	17	54	54	54	54
		$p_{\text{DEATH}} \uparrow 0{\cdot}15$								
0·265	0·475	0·02	0·03	0·21	0·92	6	10	18	31	49
0·29	0·50	0	0	0·21	0·91	6	10	18	31	50
		$p_{\text{TOX}} \uparrow 0{\cdot}15$								
0·265	0·475	0·095	0·105	0·06	0·89	6	12	21	37	54
0·265	0·475	0·17	0·03	0·06	0·91	6	11	19	32	51
		$p_{\text{CR}} \downarrow 0{\cdot}15$								
0·49	0·40	0·02	0·03	0·06	0·81	7	11	21	41	54
0·46	0·43	0·05	0	0·06	0·81	7	11	21	43	54

Table I summarizes the OCs. The notation '$p_{\text{DEATH}} \uparrow 0{\cdot}15$' means that the fixed probability vector $(p_1, \ldots, p_5)$ characterizing the clinical scenario is one in which $p_{\text{DEATH}} = p_5 = \mu_S + 0{\cdot}15$, that is, the probability of death is 0·15 larger than the mean probability of death under S. Similarly, '$p_{\text{CR}} \downarrow 0{\cdot}15$' means that $p_{\text{CR}} = p_2 + p_4$ is 0·15 smaller than $\mu_2 + \mu_4$, that is, the CR probability under the scenario is 0·15 smaller than the mean probability under S. The notation '$N_{10}, N_{25}, N_{50}, N_{75}, N_{90}$' refers to the specified 10th to 90th percentiles of the achieved sample size distribution. In a phase II trial where the goals include an improvement in an efficacy outcome or a drop in an adverse outcome it is desirable to have a high early stopping probability π in the null case where $\mathbf{p} = \boldsymbol{\mu}_S$. In contrast, for a phase II equivalence trial it is desired to have a low π under the null case and high π if any adverse (efficacy) event rate is too high (low), in terms of the fixed vector $\mathbf{p}$. Although the stopping probability $\pi = 0{\cdot}20$ when $\mathbf{p} = \boldsymbol{\mu}_S$ may seem high, the price of a smaller null π is a drop in one or more of the π values under one of the other scenarios. Because $\mu_{S,1} + \mu_{S,2} = 0{\cdot}89$, alternative scenarios obtain mainly by moving probability from

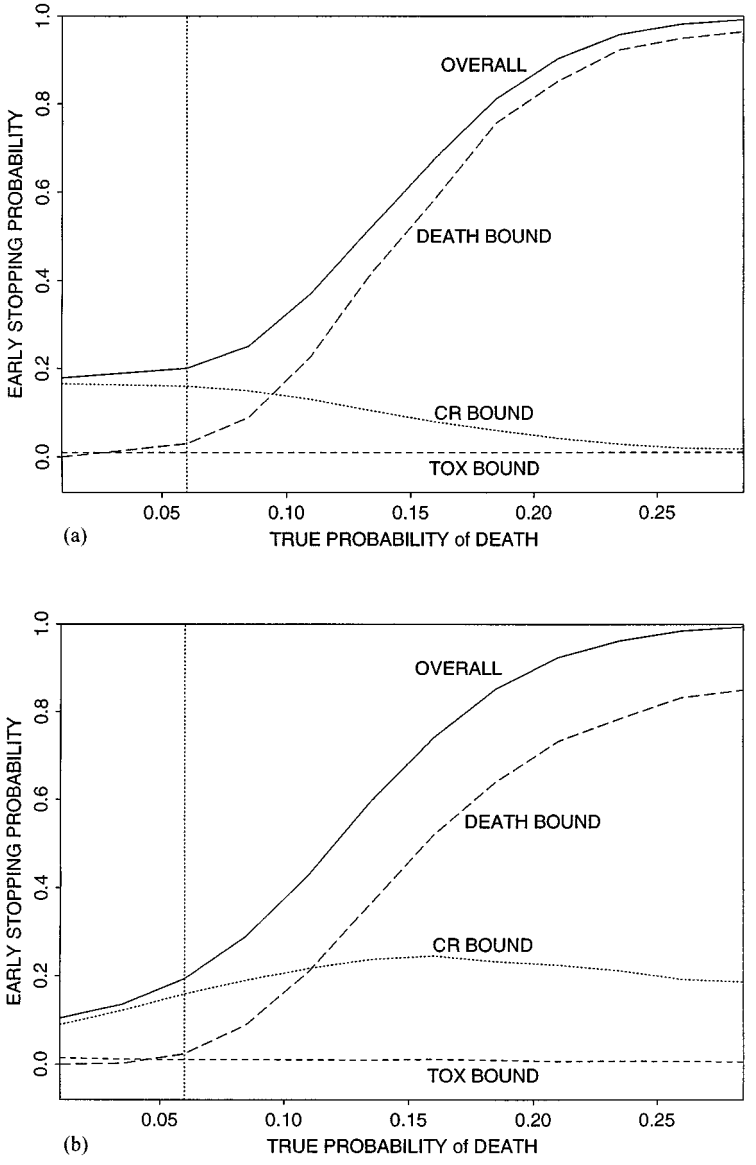

Figure 2. Early stopping probabilities by reason for stopping in the metastatic breast PBSC tansplantation trial, with $p(CR)$, $p(TOX)$ and their odds ratio fixed at their null values under μ_S in (a) and $p(TOX)$ and $p(CR|\overline{TOX})/p(\overline{CR}|\overline{TOX})$ fixed at their null values in (b)

$A_1 \cup A_2$ to the three other elementary events. We include two possibilities for each undesirable scenario in Table I to illustrate how this may be done, although for given p_{DEATH}, p_{TOX} and p_{CR} the OCs were insensitive to variations in **p**.

Figure 2(*a*) illustrates the manner in which π varies with the true probability of death, overall and by reason for stopping. The standard mean value $\mu_S(\mathrm{DEATH}) = 0{\cdot}06$ is shown by the vertical

Table II. Some Bayesian phase IIA designs which stop early if *a posteriori* $\Pr[p_0 \leqslant \theta_E | \text{data}] < p_L$ for fixed target p_0

p_0	p_L	Stopping bounds*						Prob[stop early]		
		0	1	2	3	4	5	$M = 20$	$M = 30$	$M = 40$
0·15	0·005	19	36	–	–	–	–	0·046	0·046	0·055
	0·010	15	32	–	–	–	–	0·087	0·087	0·103
	0·020	12	28	39	–	–	–	0·142	0·163	0·171
	0·040	9	23	34	–	–	–	0·234	0·270	0·290
0·20	0·005	15	28	37	–	–	–	0·037	0·044	0·051
	0·010	13	24	33	–	–	–	0·055	0·069	0·077
	0·020	10	21	29	37	–	–	0·107	0·142	0·164
	0·040	7	18	26	33	–	–	0·237	0·252	0·272
0·25	0·005	13	22	29	36	–	–	0·026	0·038	0·042
	0·010	11	19	26	33	39	–	0·060	0·069	0·072
	0·020	9	17	24	30	36	–	0·096	0·111	0·128
	0·040	7	14	21	26	32	37	0·186	0·209	0·239

* $p_0 = 0{\cdot}20$, $p_L = 0{\cdot}010 \Rightarrow$ stop if [# responses]/n patients $\leqslant 0/13, 1/24$ etc. up to n = maximum sample size M

dotted line. We computed each probability vector by fixing $p(\text{CR}) = p_2 + p_4 = 0{\cdot}58$, $p(\text{TOX}) = p_3 + p_4 = 0{\cdot}05$, and the odds ratio $p(\text{CR})p(\overline{\text{TOX}})/p(\overline{\text{CR}})p(\text{TOX}) = 51/55$, their null values, and varying $p(\text{DEATH}) = p_5$ over the domain 0·01 to 0·285. Most of the change in **p** occurs due to p_1 decreasing as p_5 increases. Alternatively, we may allow $p(\text{CR})$ to decrease as $p(\text{DEATH})$ increases by fixing $p(\text{TOX})$ and $p(\text{CR} \mid \overline{\text{TOX}})/p(\overline{\text{CR}} \mid \overline{\text{TOX}})$ as their null values. This produces Figure 2(*b*). The three stopping probabilities may sum to a value larger than the overall stopping probability because more than one bound may be hit simultaneously in a given trial. This sort of empirical analysis suggests deeper issues pertaining to the manner in which the multiple bounds interact, although we do not pursue them further here.

4. PHASE II ACTIVITY TRIALS

Once one has established the maximum tolerated dose of a new agent in a phase I trial, the next step is to determine whether it has any anti-disease effects in a phase IIA, or 'activity' trial. The goal is to decide whether the response probability is at least a given level p_0, usually in the range $0{\cdot}15 \leqslant p_0 \leqslant 0{\cdot}25$. Gehan[5] proposed the first phase IIA design, consisting of two stages. At stage 1, one tests $p < p_0$ versus $p \geqslant p_0$ to achieve a given power at p_0. If $p < p_0$ is accepted then the trial terminates, otherwise additional patients are treated in a second stage, with the stage 2 sample size determined to estimate p with a given reliability.

We can achieve these goals by a simple adaptation of the Bayesian method for a single binary outcome. Owing to its flexibility, this yields a broad and intuitively appealing set of designs. Since the effective standard response rate is 0 or possibly a small value in the range 0·05–0·10, to obtain a phase IIA design we simply replace $\eta_S + \delta$ in the definition of λ with the fixed value p_0, set $m = 1$, specify a reasonable value of M, and stop the trial early if $\Pr[p_0 < \eta_E | X_n] < p_L$. Table II provides designs for an array of parameterizations corresponding to some likely phase IIA settings. For example, the parameters $p_0 = 0{\cdot}20$, $M = 40$ and $p_L = 0{\cdot}01$ give a design with early

stopping probability $\pi = 0{\cdot}077$ if the true response rate is 0·20. Equivalently, 0·077 is the false negative probability. This design stops early if there are at most 0 responses in the first 13 patients, 1 in the first 24 or 2 in the first 33. An interesting aspect of constructing stopping rules in this way is that the number of stages is a consequence of the numerical values of p_0, M and p_L, rather than an additional design parameter specified separately as in a typical group-sequential design.

An important ethical consideration in designing any phase II trial is the question of whether E is appropriate for the particular patient group. Phase IIA trials typically evaluate new agents that may not even have any anti-disease activity. Thus, it is only appropriate to conduct the first trial of a new agent in a patient group where either there is no effective treatment or the best available treatment has a very low response rate, typically near 0·05. This is the primary reason why phase IIA trials are conducted most often in salvage patients with very poor prognosis. It is unethical to test a new agent, which may be inactive, in a patient group where there is an established treatment with a high response rate. For example, if $\mu_S = 0{\cdot}60$ in a group of untreated patients, then it is unethical to conduct a phase IIA trial of a completely new agent in that patient group.

5. MIXTURE PRIORS

In a trial to evaluate the efficacy of a new vaccine in late stage melanoma, patient outcome was characterized by a single binary indicator of response. The prior for the response probability θ_S with interferon, the standard treatment, reflected experience with several hundred patients in numerous trials. The clinician specified a mean of $\mu_S = 0{\cdot}15$, but also said that observed rates varied from 0 to 50 per cent. A beta prior with parameters beta(0·15 N, 0·85 N) for large historical sample size N is not consistent with this upper limit, however, since it does not spread enough probability mass on the upper limit of the domain from 0·40 to 0·50. Thus, a prior for θ_S with mean 0·15 but with a heavier tail was needed.

We obtain a simple extension that accommodates this sort of setting as follows. In general, one may specify m component priors $\mathrm{Dir}(\mathbf{a}_{S,1}), \ldots, \mathrm{Dir}(\mathbf{a}_{S,m})$ and weights $w_1, \ldots, w_m$ such that the discrete mixture which is $\mathrm{Dir}(\mathbf{a}_{S,j})$ with probability w_j has the desired properties. If historical data on S are available, one may simple let $\mathbf{a}_{S,j}$ be the event counts from the jth historical trial, with the w_j's reflecting their relative sample sizes. Otherwise, one may mimic this structure by specifying m Dirichlet components such that $w_1\boldsymbol{\mu}_{S,1} + \cdots + w_m\boldsymbol{\mu}_{S,m}$ equals a specified overall mean vector and the dispersion parameters reflect the amount of clinical experience. An ordinary Dirichlet prior is used for $\boldsymbol{\theta}_E$. Since no additional data on S are observed the prior and posterior of $\boldsymbol{\theta}_S$ are the same, hence $\lambda(C, \delta) = \sum_1^m w_j\lambda_{S,j}(C, \delta)$. Extending the model in this way thus provides a much broader family of priors while not introducing any new computational problems.

In the above application, we used five beta(a, b) component priors with means 0·05, 0·15, 0·25, 0·35 and 0·45, each with dispersion $a + b = 100$, hence $(a_{S,j}, b_{S,j}) = 100(\mu_{S,j}, 1 - \mu_{S,j})$. The mixture weight $\mathbf{w} = (0{\cdot}6, 0{\cdot}1, 0{\cdot}1, 0{\cdot}1, 0{\cdot}1)$ yield the desired overall mean of $\sum_1^5 w_j\mu_{S,j} = 0{\cdot}15$. Figure 3 shows a plot of this prior. Although this particular formulation certainly is not unique, it corresponds more closely to the stated clinical experience than does a simple beta prior since it yields $\Pr[\theta_S \geqslant 0{\cdot}50] = 0{\cdot}016$ rather than 0.

The trial goals were to: (i) stop and declare the vaccine not promising compared to interferon if an improvement of 0·30 over θ_S was unlikely; (ii) stop and declare the vaccine promising if an improvement in θ_S was likely, and otherwise treat $M = 30$ patients. Thus, both upper and

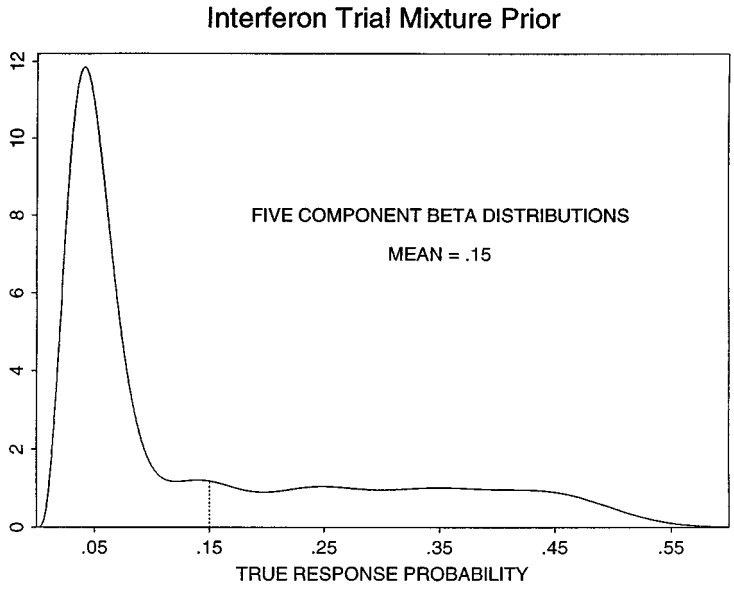

Figure 3. Five-component beta mixture prior for the vaccine trial

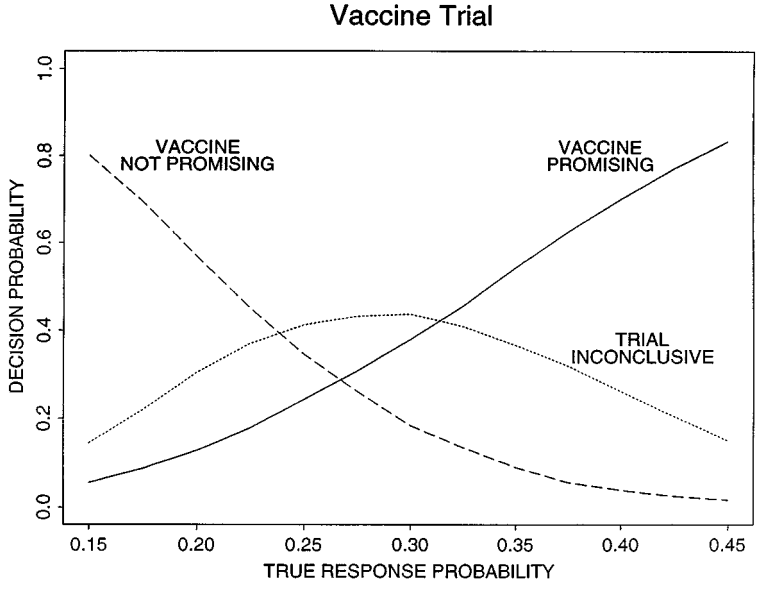

Figure 4. Decision probabilities for the vaccine trial

lower early stopping bounds were desired. The formal stopping criteria were $\lambda(\text{RES}, 0{\cdot}30) < 0{\cdot}02$ and $\lambda(\text{RES}, 0) > 0{\cdot}92$. These criteria are an application of Thall and Simon.[2,3] Figure 4 summarizes the design's OCs, with the probabilities that the vaccine is declared not promising, declared promising or that the trial is inconclusive represented, respectively, by dashed, solid and dotted

lines. The goals of this design are rather optimistic in that $\delta = 0{\cdot}30$ is quite large. A large sample size would be required to obtain comparable OCs for the more typical values $\delta = 0{\cdot}15$ or $0{\cdot}20$.

6. RANDOMIZED PHASE II SELECTION TRIALS

In many settings it is desired to carry out phase II evaluation of two or more experimental treatments simultaneously, with each compared to a common fixed or random standard. There is an extensive literature on the general problem of ranking and selection.[27,28] The goal here is not to obtain confirmatory comparative results, however, as done for example in the two-stage selection and testing designs of Thall *et al.*[29,30] or Schaid *et al.*[31] Rather, the aim is to select one or more of the experimental treatments for subsequent evaluation, as in Simon *et al.*[32] or Thall and Estey.[33] We show by example how to adapt the TSE strategy to achieve this type of goal. Our approach is simply to randomize patients among the experimental treatments while applying the same early stopping criteria to each arm. Depending upon one's goals, one may use any appropriate criterion at the end to select among the treatments not terminated early.

If an acute myelogenous leukaemia (AML) patient either fails to achieve CR with initial chemotherapy (is 'resistant') or has achieved CR but has relapsed in less than a year, subsequently it becomes more difficult to achieve a CR. These are referred to as *salvage* patients, since the next round of treatment is an attempt to save patients who have failed initial therapy. In an attempt to improve on the CR rate of 11 per cent achieved in this patient group with cytosine arabinoside (ara-C), it was decided to test three new chemotherapy combinations. These were topotecan (topo) + ara-C, topo + etoposide (VP-16) given before the topo, and VP-16 given after the topo. A randomized phase II trial was conducted, with each arm compared to the historical experience with S = ara-C. Figure 5 illustrates the outcomes monitored, each at two months after initiation of treatment. CR was scored only among patients alive at two months and a distinction was made between death with and without TOX. A Dir(25, 3, 35, 6, 2, 10) standard prior was used, based on event counts from historical data on 81 AML salvage patients treated with ara-C at M.D. Anderson. For each of the three experimental treatment probability vectors, we used a Dirichlet prior with mean vector $\boldsymbol{\mu}_S$ and $a_1 + \cdots + a_6 = 6$. The goals were the same in each experimental arm, namely to improve the CR rate while controlling the death and toxicity rates. The specific early stopping criteria used in each arm were $\lambda(\text{CR}, 0{\cdot}20) < 0{\cdot}005$, $\lambda(\text{TOX}, 0{\cdot}05) > 0{\cdot}98$ and $\lambda(\text{D}, 0) > 0{\cdot}95$. Thus, a slippage of $0{\cdot}05$ in $\eta_S(\text{TOX}) = \theta_{S,3} + \theta_{S,4}$ was a trade-off for a $0{\cdot}20$ increase in $\eta_S(\text{CR}) = \theta_{S,2} + \theta_{S,4}$. A maximum of 120 patients were randomized among the treatment arms. The selection criterion was simply to choose the treatment, among those not terminated early, having the highest posterior mean $\eta(\text{CR})$.

Tables III and IV summarize the operating characteristics of this design. Since the three experimental arms have identical designs, their usual OCs, consisting of within-arm stopping probabilities and sample sizes, also are identical. We summarize these values based on $M = 40$ in Table III. Table III shows that this design has very desirable within-arm OCs. We used the null odds ratio OR = $1{\cdot}43$ between p_{TOX} and p_{CR} to determine $\mathbf{p}$ in scenario 2 and varied the OR from $1{\cdot}1$ to 20 in scenarios 3 and 4, although the OCs were insensitive to the OR over this range.

The selection probabilities are given in Table IV. We chose the three scenarios in Table IV as extensions of three scenarios often evaluated in the case of a single parameter.[17–19] Consider the simpler setting with one binary outcome where the goal is to determine if one of K experimental treatments $E_1, \ldots, E_K$ provides at least a δ improvement over a standard null value p_0. Assuming that no experimental success probability p_j is in the interval $(p_0, p_0 + \delta)$ and that at least one

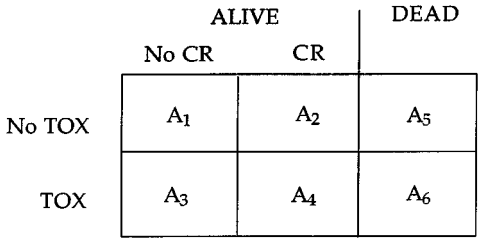

Figure 5. Patient outcomes for the randomized topotecan trial

Table III. Within-arm operating characteristics of randomized phase II topotecan trial

Clinical scenario*	π	Sample size percentiles				
		N_{10}	N_{25}	N_{50}	N_{75}	N_{90}
(1) Null case: $\mathbf{p} = \boldsymbol{\mu}_S$	0·85	10	10	15	32	40
(2a) $p_{DEATH} \uparrow 0{\cdot}10$, $p_{CR} = \mu_{S,CR}$	0·89	10	10	15	27	40
(2b) $p_{DEATH} \uparrow 0{\cdot}10$, $p_{CR} \uparrow 0{\cdot}20$	0·45	11	18	40	40	40
(3) $p_{TOX} \uparrow 0{\cdot}15$	0·87–0·88	10	10	15	27	40
(4a) $p_{CR} \uparrow 0{\cdot}20$, $p_{TOX} \uparrow 0{\cdot}05$, $p_D = \mu_{S,D}$	0·12	27–30	40	40	40	40
(4b) $p_{CR} \uparrow 0{\cdot}20$, $p_{TOX} \uparrow 0{\cdot}05$, $p_D \downarrow 0{\cdot}05$	0·08	40	40	40	40	40

* Odds ratio of p_{CR} and p_{TOX} varied from 1·1 to 20 in cases 2–4

Table IV. Selection probabilities for randomized phase II topotecan trial

Scenario	Selection probabilities			
	E_1	E_2	E_3	None
Randomize all patients among arms not terminated				
(1) $\mathbf{p}_{E_1} = \mathbf{p}_{E_2} = \mathbf{p}_{E_3} = \boldsymbol{\mu}_S$	0·05	0·05	0·05	0·85
(2) $\mathbf{p}_{E_1} = \mathbf{p}_{E_2} = \boldsymbol{\mu}_S$, $\mathbf{p}_{E_3} = \mathbf{p}_{4a}$*	0·01	0·01	0·87	0·11
(3) $\mathbf{p}_{E_1} = \boldsymbol{\mu}_S$, $\mathbf{p}_{E_2} = \mathbf{p}_{E_3} = \mathbf{p}_{4a}$	0·00	0·49	0·49	0·02
Treat at most $M = 40$ patients in each arm				
(1) $\mathbf{p}_{E_1} = \mathbf{p}_{E_2} = \mathbf{p}_{E_3} = \boldsymbol{\mu}_S$	0·13	0·13	0·13	0·61
(2) $\mathbf{p}_{E_1} = \mathbf{p}_{E_2} = \boldsymbol{\mu}_S$, $\mathbf{p}_{E_3} = \mathbf{p}_{4a}$	0·02	0·02	0·88	0·08
(3) $\mathbf{p}_{E_1} = \boldsymbol{\mu}_S$, $\mathbf{p}_{E_2} = \mathbf{p}_{E_3} = \mathbf{p}_{4a}$	0·00	0·49	0·49	0·02

* $p_{CR} \uparrow 0{\cdot}20$, $p_{TOX} \uparrow 0{\cdot}05$ and $p_D = \mu_{S,D}$

$p_j \geqslant p_0 + \delta$, then the vector of experimental success probabilities $\mathbf{p} = (p_1, \ldots, p_K)$ that minimizes the probability of selecting an E_j having $p_j \geqslant p_0 + \delta$ is the *least favourable configuration* (LFC), characterized by $p_1 = \cdots = p_{K-1} = p_0$ and $p_K = p_0 + \delta$. The three cases in Table IV generalize the three analogous one-parameter cases: (1) $p_1 = p_2 = p_3$; (2) the LFC where $p_1 = p_2 = p_0$ and $p_3 = p_0 + \delta$; and (3) the lucky case where $p_1 = p_0$ and $p_2 = p_3 = p_0 + \delta$, by replacing the respective one-dimensional values p_0 and $p_0 + \delta$ with the vectors $\boldsymbol{\mu}_S$ and $\mathbf{p}_{4a}$. Although this leads

to deeper issues regarding definition or derivation of a LFC in the multidimensional setting, we do not pursue this further here.

If a treatment arm terminates after $n < M$ patients, one may either randomize the remaining $M - n$ patients among the other arms or not do so and thus have a smaller overall sample size. The former approach is similar to what is done in adaptive or multi-arm bandit designs. For each of the three scenarios described above, we computed the selection probabilities under both approaches. The results comprise the upper and lower portions of Table IV. It may seem that the second approach is more desirable since there is a savings in sample size. However, under the first scenario this approach has the undesirable effect of greatly increasing the probability of incorrectly selecting a treatment which is on average equivalent to S, while under the second scenario it increases the probability of correctly selecting E_3 slightly, from 0·87 to 0·88.

In addition to accommodating multiple outcomes, this approach to phase II selection extends the two-stage selection designs cited above by replacing one interim test with continuous monitoring. Thus, we may terminate an arm early if it is not promising, in terms of either safety or efficacy, at any point in the trial. We can easily modify or extend the design in various ways. Two important extensions are (1) inclusion of the possibility of selecting more than one treatment, as in Schaid *et al.*,[19] and (2) incorporation of a standard arm[17] rather than relying only on an informative prior on $\boldsymbol{\theta}_S$. This latter extension produces a rather different type of trial, since one would require the priors on $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_{E_1}, \ldots, \boldsymbol{\theta}_{E_K}$ to be much more similar to each other for the randomization to be ethical. Additionally, the randomization would allow confirmatory evaluation of $\eta_{E_j}(\mathrm{CR}) - \eta_S(\mathrm{CR})$ for those E_j not terminated, given a sufficiently large sample size M.

7. DISCONTINUOUS MONITORING

Because TSE present the method in the context of continuous monitoring, it seems impractical for use in multi-centre trials. A simple modification that may accommodate such settings is to update the posterior and apply the decision criteria only at the interim times when the data from successive patient cohorts of a given size c become available. Applying this approach with, say, $c = 6$ thus would require monitoring after data became available from the 6th, 12th, 18th patient etc. This less intensive requirement might be more feasible with multiple institutions involved.

Thall and Simon[3] examined the effects of discontinuous monitoring in the univariate binary outcome case. To determine how this may work when it is necessary to monitor multiple outcomes, we examine the effects of discontinuous monitoring for the PBPC transplantation trial discussed in Section 3. Table V presents the OCs for this design with cohort sizes varying from 1 to 18. Naturally, the smaller value of π in the case $\mathbf{p} = \boldsymbol{\mu}_S$ obtained with $c > 1$ is desirable. For the other three cases where a large value of π is desirable the decline in π seems acceptable up to $c = 9$, but the design with $c = 18$ clearly is unsafe. We can deal with this by simply adjusting the probability cut-offs. If we use the values $p_L(\mathrm{CR}) = 0{\cdot}15$, $p_U(\mathrm{TOX}) = 0{\cdot}97$ and $p_U(\mathrm{D}) = 0{\cdot}95$ with $c = 18$, then the respective early stopping probabilities under the four scenarios in Table V become 0·19, 0·85, 0·81 and 0·78, which seems like a reasonable design. The decision rules are to stop the trial if there are $\leqslant 7/18$ or $\leqslant 17/36$ CRs, $\geqslant 4/18$ or $\geqslant 6/36$ toxicities, or $\geqslant 4/18$ or $\geqslant 6/36$ deaths. Note that, for example, we should terminate the trial if we observe 4 deaths at any point up to the 18th patient. Failure to take advantage of this simple aspect of the sequential design would result in patients being treated with a regimen that is certain to be declared unsafe. This design is similar to a conventional group-sequential trial derived using frequentist hypothesis testing criteria in the univariate binary case,[7–9] with the additional feature that now we monitor

Table V. PBPC trial operating characteristics with varying cohort size

Clinical scenario					c	π	Sample size percentiles				
p_1	p_2	p_3	p_4	p_5			N_{10}	N_{25}	N_{50}	N_{75}	N_{90}
Null case											
0·34	0·55	0·02	0·03	0·06	1	0·20	17	54	54	54	54
					3	0·17	21	54	54	54	54
					6	0·11	42	54	54	54	54
					9	0·12	36	54	54	54	54
					18	0·06	54	54	54	54	54
$p_{\text{DEATH}} \uparrow 0{\cdot}15$											
0·265	0·475	0·02	0·03	0·21	1	0·92	6	10	18	31	49
					3	0·91	6	12	21	36	51
					6	0·86	6	12	24	42	51
					9	0·82	9	18	27	45	54
					18	0·70	18	18	36	54	54
$p_{\text{TOX}} \uparrow 0{\cdot}15$											
0·265	0·475	0·095	0·105	0·06	1	0·89	6	12	21	37	54
					3	0·88	6	12	24	39	54
					6	0·84	6	12	30	42	54
					9	0·77	9	18	27	45	54
					18	0·63	18	18	36	54	54
$p_{\text{CR}} \downarrow 0{\cdot}15$											
0·49	0·40	0·02	0·03	0·06	1	0·82	7	11	21	41	54
					3	0·80	9	15	24	45	54
					6	0·71	12	18	30	54	54
					9	0·71	9	18	27	54	54
					18	0·55	18	18	36	54	54

all of three of the outcomes CR, toxicity and death. Since our focus is early stopping we declare E promising if the trial does not terminate early. Thus, if the trial does not terminate by the 36th patient then we treat the last 18 patients simply to obtain reasonably reliable posterior probability estimates. This type of design could be made more similar to conventional group-sequential trials by simply adding an upper probability criterion $\lambda(\text{CR}) > p_{\text{U}}(\text{CR})$ in order to declare E promising once the data from all 54 patients are available.

8. DISCUSSION

We have presented some applications and extensions of the TSE monitoring strategy that we feel practitioners will find useful. We chose these five cases based on feedback from others and our own experiences applying the method. One objective has been to provide designs that we can derive using the strategy but that are not entirely apparent from TSE. In addition, we have provided extensions accommodating mixture priors, randomized trials and monitoring by cohorts of size greater than one.

To apply the TSE strategy, the physician must be closely involved in specifying the patient outcomes to monitor, the standard treatment prior and the goals of the trial. We have found that the Bayesian model provides physicians with a rational framework for what they must do in any case. While designing trials in this way entails considerably more work than is required for most conventional designs, the response by clinicians has been extremely favourable. Our experience has shown that, when physicians collaborate in the process of constructing the design, they are much more likely to adhere to it in the conduct of the trial. From both a scientific and a clinical viewpoint, the use of a practical design that realistically reflects the medical phenomenon and that actually will be followed is highly preferable to an unrealistic design that will be violated in practice.

We have used historical counts for the elementary events from a trial or trials of S as the parameters of the Dirichlet prior on $\boldsymbol{\theta}_S$. Alternatively, in settings where the clinician specifies the prior mean probabilities $\boldsymbol{\mu}_S$ and the number N of historical patients upon which this mean is based we have used $N\boldsymbol{\mu}_S$ as the Dirichlet parameters. In either case, the sum of the parameters could be reduced if it is felt that, aside from the fact that E and S are different treatments, the trial of E will be conducted differently from the manner in which the trials that produced the prior on $\boldsymbol{\theta}_S$ were conducted. This really addresses the fact that, unavoidably, there is always the possibility of a trial effect, and this effect may differ between the historical trial or trials and the planned trial of E. This issue is addressed formally in an empirical Bayes setting by Thall and Simon.[34] In the present context, the essential practical issues are whether the prior on $\boldsymbol{\theta}_S$ is an honest representation of knowledge about the standard, and how downweighting the counts of this prior will affect the operating characteristics of the design.

Despite its flexibility and generality, the method still has some practical limitations. The first is that it does not accommodate dose changing during the course of the trial. Although the conventional clinical trial model is that an appropriate dose is first determined in a phase I trial, a typical phase II trial often involves one or more dose modifications. A hybrid phase I/II design allowing interim dose changes while also monitoring efficacy and toxicity has been proposed by Thall and Russell[35] for a particular application, but a general method as yet does not exist. Another important issue is accounting for patient prognostic factors. This motivates the use of regression models, which adds another level of complexity. If one is willing to extend or replace the Dirichlet-multinomial model, then one can use joint priors that account for dependency between $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_E$ or that allow one to quantify $\text{var}\{\eta(\text{CR})\}$ and $\text{var}\{\eta(\text{TOX})\}$ separately, both of which are desirable properties not enjoyed under the Dirichlet formulation. We have not examined the sensitivity of the method to the priors, and this sort of conventional Bayesian robustness analysis might provide additional insights. This is related to an approach recently proposed by Heitjan[36] for the univariate binary case in which one declares E promising if the posterior convinces someone with a sceptical prior that E is superior to S, and not promising if the posterior convinces one with an optimistic prior that E is inferior to S. Extension of this type of criterion to the multivariate case might prove quite useful. Finally, it would be more desirable from a philosophical viewpoint to construct a fully Bayesian version based on decision theory, and ideally this may lead to designs with better practical features.

REFERENCES

1. Thall, P. F., Simon, R. and Estey, E. H. 'Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes', *Statistics in Medicine*, **14**, 357–379 (1995).
2. Thall, P. F. and Simon, R. 'Practical Bayesian guidelines for phase IIB clinical trials', *Biometrics*, **50**, 337–349 (1994).

3. Thal, P. F. and Simon, R. 'A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials', *Controlled Clinical Trials*, **15**, 463–481 (1994).
4. Lazaridis, E. and Gonin, R. 'Continuously monitored stopping boundary methodologies: The issues of sensitivity, association and trial suspension', *Statistics in Medicine*, **16**, 1925–1941 (1997).
5. 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).
6. Schultz, J. R., Nichol, F. R., Elfring, G. L. and Weed, S. L. 'Multiple stage procedures for drug screening', *Biometrics*, **29**, 293–300 (1973).
7. Fleming, T. R. 'One sample multiple testing procedure for phase II clinical trials', *Biometrics*, **38**, 143–151 (1982).
8. Chang, M., Therneau, T. and Wieand, H. S. 'Designs for group sequential phase II clinical trials', *Biometrics*, **43**, 865–874 (1987).
9. Therneau, T., Wieand, H. S. and Chang, M. 'Optimal designs for a grouped sequential binomial test', *Biometrics*, **46**, 771–781 (1990).
10. Simon, R. 'Optimal two-stage designs for phase II clinical trials', *Controlled Clinical Trials*, **10**, 1–10 (1989).
11. Friedman, L. M., Furberg, C. D. and DeMets, D. L. *Fundamentals of Clinical Trials*, John Wright, PSG Ltd., Boston, 1981.
12. Workshop on Early Stopping Rules in Cancer Clinical Trials, Robinson College, Cambridge, U.K., 13–15 April, 1993 (R. L. Souhami and J. Whitehead, guest editors), *Statistics in Medicine*, **13**, No. 13/14 (1994).
13. DeMets, D. L. and Lan, K. K. G. 'Interim analysis: The alpha spending approach', *Statistics in Medicine*, **13**, 1341–1356 (1994).
14. Jennison, C. and Turnbull, B. W. 'Interim analyses: The repeated confidence interval approach, (with discusion)', *Journal of the Royal Statistical Society*, *Series B*, **51**, 305–361 (1989).
15. Freedman, L. S. and Spiegelhalter, D. J. 'The assessment of subjective opinion and its use in relation to stopping rules for clinical trials', *Statistician*, **32**, 153–160 (1983).
16. 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).
17. Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. 'Bayesian approaches to randomized trials, (with discussion)', *Journal of the Royal Statistical Society*, *Series A*, **157**, 357–416 (1994).
18. Berry, D. A. 'A case for Bayesianism in clinical trials (with discussion)', *Statistics in Medicine*, **12**, 1377–1404 (1993).
19. Berry, D. A. 'Decision analysis and Bayesian methods in clinical trials', *in* Thall, P. F. (ed.), *Recent Advances in Clinical Trial Design and Analysis*, Kluwer, Boston, 1995, pp. 125–154.
20. Rosner, G. L. and Berry, D. A. 'A Bayesian group sequential design for a multiple arm randomized clinical trial', *Statistics in Medicine*, **14**, 381–394 (1995).
21. Thall, P. F., Simon, R. and Estey, E. H. 'New statistical strategy for monitoring safety and efficacy in single-arm clinical trials', *Journal of Clinical Oncology*, **14**, 296–303 (1995).
22. Bryant, J. and Day, R. 'Incorporating toxicity considerations into the design of two-stage phase II clinical trials', *Biometrics*, **51**, 1372–1383 (1995).
23. Conaway, M. R. and Petroni, G. R. 'Bivariate sequential designs for phase II clinical trials', *Biometrics*, **51**, 656–664 (1995).
24. Conaway, M. R. and Petroni, G. R. 'Designs for phase II trials allowing for a trade-off between response and toxicity', *Biometrics*, **52**, 1375–1386 (1996).
25. 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 (1993).
26. Cook, R. J. and Farewell, V. T. 'Guidelines for monitoring efficacy and toxicity response in clinical trials', *Biometrics*, **50**, 1146–1152 (1994).
27. Gibbons, J. D., Olkin, I. and Sobel, M. *Selecting and Ordering Populations*: *A New Statistical Methodology*, Wiley, New York, 1977.
28. Bechhofer, R. E., Santner, T. J. and Goldsman, D. *Design and Analysis of Experiments for Statistical Selection*, *Screening and Multiple Comparisons*, Wiley, New York, 1995.
29. Thall, P. F., Simon, R. and Ellenberg, S. S. 'Two stage selection and testing designs for comparative clinical trials', *Biometrika*, **75**, 303–310 (1988).

30. Thall, P. F., Simon, R. and Ellenberg, S. S. 'A two-stage design for choosing among several experimental treatments and a control in clinical trials', *Biometrics*, **45**, 537–547 (1989).
31. Schaid, D. J., Wieand, H. S. Therneau, T. M. 'Optimal two-stage screening designs for survival comparisons', *Biometrika*, **77**, 507–513 (1990).
32. Simon, R., Wittes, R. E. and Ellenberg, S. S. 'Randomized phase II clinical trials', *Cancer Treatment Reports*, **69**, 1375–1381 (1985).
33. Thall, P. F. and Estey, E. H. 'A Bayesian strategy for screening cancer treatments prior to phase II clinical evaluation', *Statistics in Medicine*, **12**, 1197–1211 (1993).
34. Thall, P. F. and Simon, R. 'Incorporating historical data in planning Phase II clinical trials', *Statistics in Medicine*, **9**, 215–228 (1990).
35. Thall, P. F. and Russell, K. E. 'A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials', *Biometrics* (1998) to appear.
36. Heitjan, D. F. 'Bayesian interim analysis of phase II cancer clinical trials', *Statistics in Medicine*, **16**, 1791–1802 (1997).

