BIOMETRICS 57, 1089–1095
December 2001

Decision-Theoretic Designs for Pre-Phase II Screening Trials in Oncology

Nigel Stallard

Medical and Pharmaceutical Statistics Research Unit, The University of Reading,
P.O. Box 240, Earley Gate, Reading, RG6 6FN, U.K.
email: n.stallard@reading.ac.uk.

and

Peter F. Thall

Department of Biostatistics, M. D. Anderson Cancer Center, University of Texas,
Box 447, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.

SUMMARY. A Bayesian decision-theoretic method is proposed for conducting small, randomized pre-phase II selection trials. The aim is to improve on the design of Thall and Estey (1993, *Statistics in Medicine* **12,** 1197–1211). Designs are derived that optimize a gain function accounting for current and future patient gains, per-patient cost, and future treatment development cost. To reduce the computational burden associated with backward induction, myopic versions of the design that consider only one, two, or three future decisions at a time are also considered. The designs are compared in the context of a screening trial in acute myelogenous leukemia.

KEY WORDS: Bayesian design; Decision theory; Phase II clinical trial; Selection.

1. Introduction

Phase II cancer clinical trials typically are small to moderate sized exploratory studies to assess whether a new therapy is sufficiently promising to warrant evaluation in a large-scale randomized comparative phase III trial. In many areas of oncology, it is common for several new treatments to be available for phase II evaluation at the same time. Because resources are limited, it is thus necessary to select among potential therapies for phase II evaluation. This is usually done informally based on data from preclinical studies and phase I trials. A Bayesian design for trials aimed at selecting treatments for phase II was proposed by Thall and Estey (1993; hereafter TE). They suggested a small, randomized trial of several candidate treatments in patients with poorer prognosis than those to be treated in phase II, as is ethically appropriate for newer treatments. The design drops inferior treatments early, with the best remaining treatment selected for phase II if it shows a likely improvement over a fixed standard.

Sequential designs are important in early phase trials due to the ethical need to stop a trial if an experimental therapy appears likely to be either unsafe or ineffective. This is very important in oncology, where substantive treatment advances are rare and adverse side effects may be quite severe. Many authors have proposed sequential designs for phase II cancer trials, including Fleming (1982), Simon (1989), Thall and Simon (1994), Bryant and Day (1995), Thall, Simon, and Estey (1995), Stallard (1998), and Stallard, Thall, and Whitehead (1999). In this article, we propose decision-theoretic sequential designs for randomized pre-phase II screening trials. Our aim is to improve on the non-decision-theoretic TE design. We develop decision-theoretic designs in Section 3 and, in Section 4, apply them to a trial in acute myelogenous leukemia, comparing them with the TE design. We close with a discussion in Section 5.

2. Preliminaries and Notation

Suppose that m experimental therapies, T_t, $t = 1, \ldots, m$, are to be comparatively evaluated in terms of a success/failure patient outcome observed soon enough after the start of the patient's treatment to enable sequential monitoring. Denote by θ_t the treatment success probability for patients receiving treatment T_t. We assume that θ_t follows a beta prior with parameters a_t and b_t, denoted $\theta_t \sim \text{beta}(a_t, b_t)$, with $\theta_1, \ldots, \theta_m$ independent. In practice, we will usually take all $(a_t, b_t) = (a, b)$, with $a + b \leq 2$ to reflect the relative lack of knowledge about $T_1, \ldots, T_m$.

We consider designs in which patients are randomized among the treatments in blocks, with one patient per treatment in each block. After observing the outcomes for each block, a decision is made whether to drop any or all of the treatments from the study. Dropped treatments may not be reinstated. The trial continues until either no treatments remain or the total number of patients in the trial reaches a prescribed maximum, N. At the end of the study, a decision is made about which of the remaining treatments is best and whether this treatment should be evaluated in a phase II trial. In practice, N is determined based on patient accrual rate, feasible trial duration, monetary costs, and drug availability in addition to statistical properties of the design and reliability of parameter estimates.

Let p_0 be a fixed, minimum clinically acceptable success rate specified by the physician(s) conducting the trial. The TE approach bases decisions on the posterior probabilities $\Pr(\theta_t < p_0 \mid \text{data})$ for $t = 1, \ldots, m$. If this probability exceeds π_1, say, for $\pi_1 = 0.90$ or 0.85, then T_t is dropped. At the end of the trial, the treatment T_{t^*} having largest posterior mean among those remaining is selected, provided that $\Pr(\theta_{t^*} > p_0 \mid \text{data}) > \pi_2$, say, for $\pi_2 = 0.90$ or 0.95.

Similar to TE, our proposed decision-theoretic designs are based on gain functions that depend on the values of $\theta_t - p_0$ for $t = 1, \ldots, m$. Criticisms of decision-theoretic methods are that the gain function underlying the decisions is subjective and that the computational task of constructing optimal designs is substantial. To address these issues, we illustrate how gain function parameters may be chosen to obtain a design with desirable frequentist properties and we consider myopic strategies that reduce computational requirements.

3. Decision-Theoretic Designs

3.1 *Computing the Expected Gain*

At any interim point during the trial, possible actions are to continue with some subset of the set of treatments remaining or terminate the trial. If all N patients are treated, a final decision must be made about which treatment, if any, should be developed further in a subsequent phase II trial.

Let $A \subseteq \{1, \ldots, m\}$ denote the action of continuing with treatments $\{T_t,\ t \in A\}$. We will abuse usual set notation by writing $A = t$ for the action of selecting treatment T_t for future evaluation after the trial and $A = 0$ if no treatment is selected. Let $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im})$ denote the data observed through the ith stage of the trial, where $\mathbf{x}_{it} = (x_{it1}, x_{it2})'$ with x_{it1} and x_{it2} the numbers of successes and failures, respectively, among n_{it} patients receiving T_t. Let $D(\mathbf{X}_i)$ denote the set of possible actions following observation of $\mathbf{X}_i$. If the number of patients treated, $n_i = \Sigma_{t=1}^{m} n_{it}$, equals N, then $D(\mathbf{X}_i)$ contains zero and all t such that $n_{it} = \max_{t'=1,\ldots,m} n_{it'}$ since the latter are the treatments that have not been dropped. If $n_i < N$, then $D(\mathbf{X}_i)$ contains all subsets of the set $\{t \in \{1, \ldots, m\} : n_{it} = \max_{t'=1,\ldots,m} n_{it'}\}$. Let $g(A, \boldsymbol{\theta}, \mathbf{X}_i)$ denote the gain from taking action A after observing $\mathbf{X}_i$ when the true state of nature is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$. The posterior expectation of g is $G(A, \mathbf{X}_i) = \mathrm{E}(g(A, \boldsymbol{\theta}, \mathbf{X}_i) \mid \mathbf{X}_i)$. Specific forms for g will be discussed in Section 3.3.

At each point in the trial, the action in $D(\mathbf{X}_i)$ maximizing the expected gain $G(A, \mathbf{X}_i)$ will be chosen. Values of $G(A, \mathbf{X}_i)$ corresponding to termination of the trial, i.e., for $A \in \{0, \ldots, m\}$ or ϕ, can be evaluated directly. Values corresponding to continuation of the trial can be obtained by backward induction (cf., Berger, 1985) as follows. Given observed data $\mathbf{X}_i$, the overall expected gain, $G(A, \mathbf{X}_i)$, from taking action $A \subseteq \{1, \ldots, m\}, A \neq \phi$, is the posterior mean with respect to $\boldsymbol{\theta}$ of $\Sigma \Pr(\mathbf{X}_{i+1} \mid \boldsymbol{\theta}) \max_{A' \in D(\mathbf{X}_{i+1})} \{G(A', \mathbf{X}_{i+1})\}$, where the sum is taken over all data sets $\mathbf{X}_{i+1}$ possible at the $i+1$th stage following action A given $\mathbf{X}_i$ and each $\Pr(\mathbf{X}_{i+1} \mid \boldsymbol{\theta})$ is a product of relevant θ_{tj} and $(1 - \theta_{tj})$ values giving the probability of $\mathbf{X}_{i+1}$ given $\mathbf{X}_i$. Since this gives $G(A, \mathbf{X}_i)$ in terms of $G(A, \mathbf{X}_{i+1})$, expected gains from each action can be obtained. Consider first the end of the trial, when there are N patients. The gains from all possible actions can be calculated directly. The optimal final action is that with the largest expected gain. Consider next data sets $\mathbf{X}_i$ with $n_i = N - 1$. The set $D(\mathbf{X}_i)$ contains ϕ and single element sets corresponding to treatments with $n_{it} = \max_{t'=1,\ldots,m} n_{it'}$. The expected gain from ϕ can be calculated directly. That from each other action can be obtained by averaging expected gains for data sets $\mathbf{X}_i + 1$ with $n_{i+1} = N$, which have already been obtained. The optimal action can therefore be selected. Working backward through the entire trial in this way yields the expected gain for the optimal action in $D(\mathbf{X})$ for all possible data sets $\mathbf{X}$.

3.2 *Myopic Designs*

We derived the optimal designs described below by complete enumeration of all possible data sets, as required in the backward induction algorithm. This requires many expected gains to be calculated and stored. Computational requirements are a major limitation for trials much larger than that considered below or for trials involving more than three treatments. Currently, the problem remains hardly feasible.

An approximation to determining the optimal decision by backward induction is the use of a so-called myopic strategy in which, rather than considering possible outcomes at all subsequent stages in the trial, decisions are based on examination of at most, say, r future stages.

Taking $r = 1$ leads to a decision at each stage in the trial assuming that the next stage will be the last. This means that optimal actions depend only on the data observed so far and expected gains need not be stored. Suppose that data $\mathbf{X}_i$ have been observed and at most $N - n_i$ patients remain to be treated. The myopic strategy with $r = 1$ first assumes that, for some $s \geq 0$, the best s of the remaining treatments will be selected to continue in the study, with the remaining $N - n_i$ patients divided equally among them. For simplicity, denote the remaining treatments by $T_1, \ldots, T_s$. From the myopic viewpoint with $r = 1$, after one more stage, the trial will end and one of the actions $A' \in \{0, \ldots, s\}$ will be taken. Denoting the data at the end of the trial by $\mathbf{X}_{i+1}$, the expected gain from the optimal action at the end of the trial would be the posterior mean with respect to $\mathbf{X}_i$ of $\Sigma_{\mathbf{X}_{i+1}} \Pr(\mathbf{X}_{i+1} \mid \boldsymbol{\theta}) \max_{A' \in D(\mathbf{X}_{i+1})} \{G(A', \mathbf{X}_{i+1})\}$. The optimal value for s and action A having s elements are chosen to maximize this gain. If $A \neq \phi$, then one patient is allocated to each of these treatments, the data are observed, and the process repeated, again assuming that the next stage of the trial will be the last. Taking $r = 2$ or 3 corresponds to the expected gain for each action being calculated assuming that there will be one or two further opportunities to drop treatments, with remaining patients assumed to be equally divided between these decision points.

3.3 *Possible Gain Functions*

The proposed method requires a gain function, $g(A, \boldsymbol{\theta}, \mathbf{X}_i)$, for each $A \in \{0, \ldots, m\}$. This should be obtained through discussion with clinical collaborators.

Forms for gain functions in similar settings with a single experimental treatment were proposed by Stallard (1998) and Stallard et al. (1999). Here we proceed in a fashion similar to the latter paper to give possible forms for the gain functions, which are used in the comparison of the different approaches in Section 4. While the gain functions proposed may be suitable for many trials, other qualitative forms will sometimes be more appropriate. It is important that careful consideration be given to the choice of the gain functions used.

Following Stallard et al. (1999), we consider the gains to patients who will receive the experimental treatments either in this trial or in the future. In terms of the probability of success, the gain to a patient receiving treatment T_t rather than a standard treatment with success rate p_0 is $(\theta_t - p_0)$. If T_t is selected at the end of the study, then this also is the gain to a future patient treated with T_t. If all treatments are dropped from the study, future gains will equal zero. We suppose that there is a per patient cost, c, during the study and a total cost K for future investigation of a recommended treatment, with these costs also expressed on the scale of probability of success. After observing data $\mathbf{X}_i$, the gain from taking action 0 if the true state of nature is $\boldsymbol{\theta}$ is $g(0, \boldsymbol{\theta}, \mathbf{X}_i) = \Sigma_{t=1}^{m} \{n_{it}(\theta_t - p_0)\} - n_i c$. The gain from selecting T_{t^*}, which can only occur if $n_i = N$, is

$$g(t^*, \boldsymbol{\theta}, \mathbf{X}_i) = \sum_{t=1}^{m} \{n_{it}(\theta_t - p_0)\} - Nc - K + \Pi(\theta_{t^*} - p_0), \tag{1}$$

where Π represents the patient horizon, i.e., the number of potential future patients. The value of K may be restricted to be at most Π since the cost of future investigations can be compared with the expected gain in the number of patient successes from Π future patients, which cannot exceed Π. If $K > \Pi$, the future gain from further development cannot be positive, so no phase II trial will ever be conducted and the pre-phase II screen is futile. Similarly, we may restrict c to be at most one.

Aside from computational issues, a major criticism of the decision-theoretic approach is the subjective nature of the gain function. While the above form of g is qualitatively reasonable for many trials, in practice, it is very difficult to determine numerical values for the parameters c, K, and Π. Subjectivity is also present in other methods, e.g., in the choice of π_1 and π_2 in the TE approach. TE address this problem by considering several (π_1, π_2) pairs and choosing values that provide a design with desirable frequentist properties. A similar approach was taken by Thall and Simon (1994) in designing a single-treatment phase II trial. They considered scenarios where the experimental treatment had either a high or a low success probability. A design was sought that would lead to further development with high probability in the former case and low probability in the latter while having a small expected sample size in the latter case, where early stopping would be particularly desirable for ethical reasons. We extend this approach to the multiple treatment case, with consideration of the correct selection probability.

4. Comparison of the Procedures

In this section, we evaluate and compare the design maximizing the gain function (1) described in Section 3.3 (OPT); the myopic designs with $r = 1$ (MYO1), $r = 2$ (MYO2), and $r = 3$ (MYO3); and the TE design. We base this on an example, given by TE, of a trial with at most $N = 30$ patients to compare three treatments as salvage therapy for poor-prognosis patients with acute myelogenous leukemia who either were resistant to initial therapy or have relapsed after achieving complete remission (CR). The primary patient outcome is CR, with θ_t the probability of CR for patients receiving treatment T_t for $t = 1, 2, 3$. Since $p_0 = 0.20$, we took $\{\theta_1, \theta_2, \theta_3\}$ to be independent with beta(0.4, 1.6) prior, which has mean 0.20.

For the OPT and MYO designs, we considered a wide range of gain function parameterizations, with $0 \leq c \leq 1$, $100 \leq \Pi \leq 100{,}000$, and $0 \leq K \leq \Pi$. For each combination of c, K, and Π, the OPT design was obtained. For each of these designs and the TE design, we computed the probability that no treatment was selected, PNS, and the expected sample size, EN0, under the null, $(p_1, p_2, p_3) = (0.2, 0.2, 0.2)$, and, for $(p_1, p_2, p_3) = (0.4, 0.2, 0.2)$, the probability that treatment T_1 was correctly selected, PCS.

Figure 1 shows how PNS, PCS, and EN0 vary with (Π, c), $(\Pi, K/\Pi)$, and $(c, K/\Pi)$ for the OPT design. The contour lines for PNS and PCS are close together and are nearly parallel to the axis in the plots varying with $(\Pi, K/\Pi)$ and $(c, K/\Pi)$. In contrast, the contour lines for PNS and PCS varying with (Π, c) are far apart (note that, in these plots, contours are spaced at intervals of 0.1 rather than 0.2 as used elsewhere). It is clear that K has a much larger effect on PNS and PCS than either c or Π, although both PNS and PCS are reduced by taking Π to be small. Both Π and K/Π have substantial effects on EN0, with a much smaller effect due to c. Not surprisingly, smaller values of EN0 are generally associated with smaller values of PNS and PCS.

Figure 2 shows how PNS, PCS, and EN0 vary with K/Π for fixed c and Π ($c = 0.4$ and $\Pi = 1000$). This figure illustrates the problem that the choice of K, Π, and c to obtain a design with desirable overall properties inevitably must involve a compromise between the desire to make both PNS and PCS as large as possible. This is similar to the choice of π_1 and π_2 in the TE design. To facilitate the comparisons, we obtained an OPT design similar to the TE design with $\pi_1 = \pi_2 = 0.9$. The TE design has PNS and PCS of 0.821 and 0.504, respectively, and so seems a reasonable choice. Figure 2 suggests that the OPT design with $K/\Pi = 0.15$ has both PNS and PCS slightly higher than these values and so might be comparable with the TE design. The results reported for the OPT design given below are for the design with $c = 0.4$, $\Pi = 1000$, and $K/\Pi = 0.15$.

Table 1 summarizes frequentist properties of the OPT, MYO1, MYO2, and MYO3 designs under this gain function parameterization and of the TE design with $\pi_1 = \pi_2 = 0.9$ for $p_2 = p_3 = 0.2$ and $0.1 \leq p_1 \leq 0.5$ (the rows labeled TE* are explained below). All the designs have the desirable properties that the probability of selecting no treatments decreases and the probability of selecting T_1 increases with p_1 and the prob-

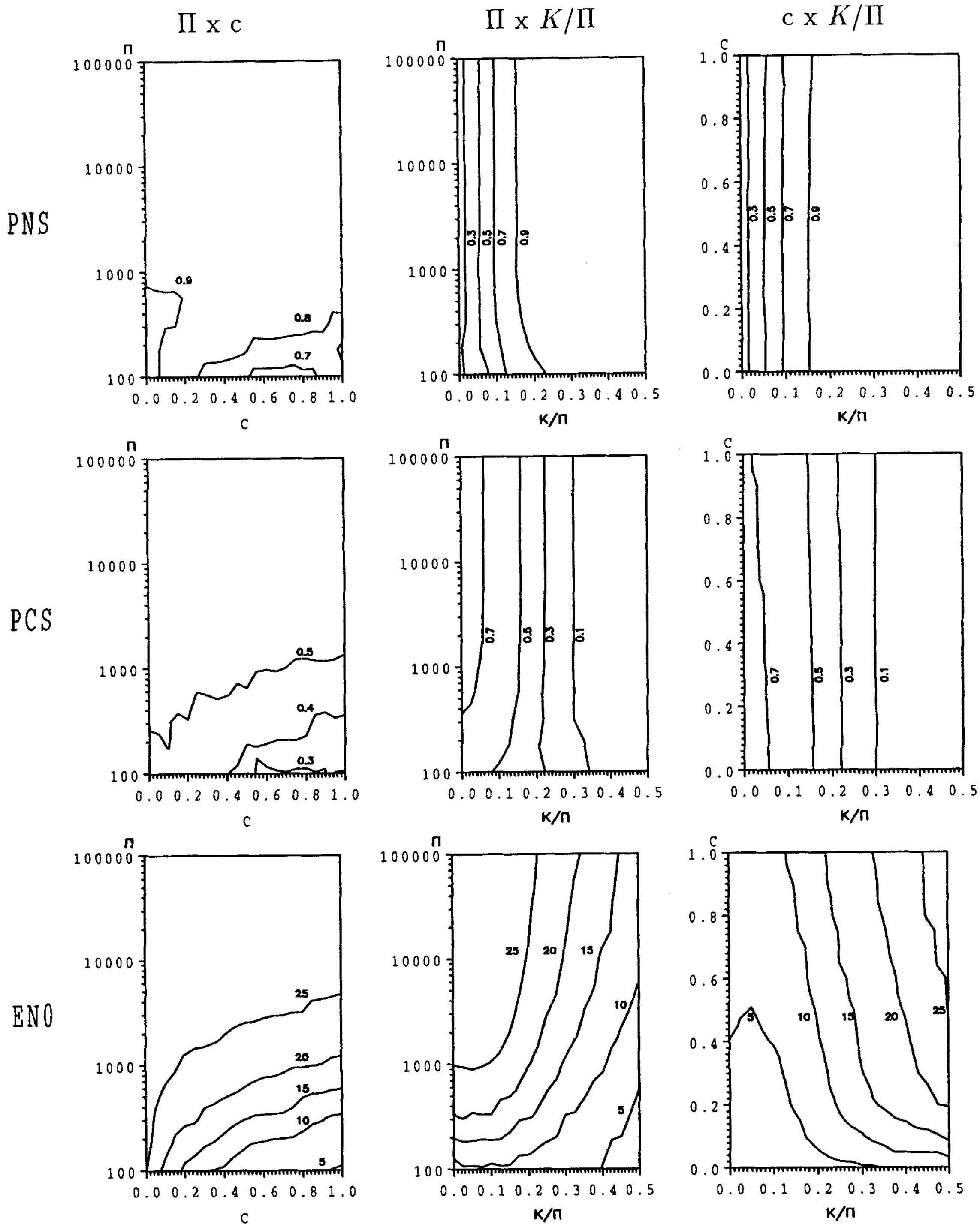

Figure 1. Contour plots showing how PNS, PCS, and EN0 vary with c, Π, and K for the OPT design.

ability of incorrectly selecting T_1 when $p_1 \leq 0.2$ is small. The OPT design is slightly more likely than the TE design both to select no treatments for $0.1 \leq p_1 \leq 0.4$ and to select T_1 for $p_1 \geq 0.4$. In all cases, the number of patients treated is smaller for the OPT design than for the TE design. In the MYO1 design, the gain function from continuing is clearly underestimated. This leads to a design that stops with a sample size much smaller than that for OPT and has a larger probability of selecting no treatment for all values of p_1 considered. The MYO3 design provides a much better approximation to the OPT design, though it still has a slightly larger probability of selecting no treatment for all values of p_1. The MYO3 design

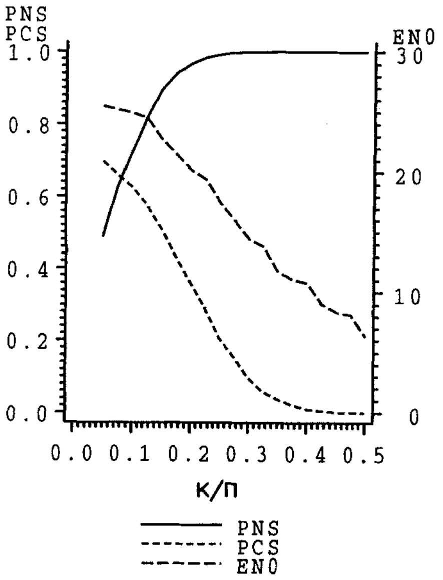

Figure 2. Properties of the OPT design for a range of values of K with c and Π fixed.

Table 1

Frequentist properties for OPT and MYO designs with $u = 1$, $c = 0.4$, $\Pi = 1000$, $K = 150$, and the TE and TE designs with $\pi_1 = \pi_1{}^* = \pi_2 = 0.90$ for $p_2 = p_3 = p_0 = 0.2$ and p_1 ranging from 0.1 to 0.5*

		Probability of selecting		Expected number of patients	
p_1	Design	None	T_1	Receiving T_1	Total
0.1	OPT	0.930	0.001	5.24	21.24
	MYO1	0.976	0.000	3.02	14.02
	MYO2	0.960	0.000	3.68	16.95
	MYO3	0.942	0.001	4.62	20.40
	TE	0.867	0.003	7.06	27.53
	TE*	0.742	0.011	8.66	29.47
0.2	OPT	0.894	0.036	7.84	22.76
	MYO1	0.965	0.011	5.07	15.23
	MYO2	0.940	0.020	6.20	18.62
	MYO3	0.915	0.028	7.33	22.04
	TE	0.821	0.057	9.49	28.48
	TE*	0.651	0.110	9.80	29.40
0.3	OPT	0.727	0.208	10.44	24.03
	MYO1	0.865	0.113	8.72	17.96
	MYO2	0.806	0.154	10.05	21.46
	MYO3	0.759	0.186	11.09	24.46
	TE	0.654	0.238	11.16	29.14
	TE*	0.464	0.330	11.13	29.41
0.4	OPT	0.439	0.510	11.48	23.66
	MYO1	0.626	0.354	13.17	21.44
	MYO2	0.538	0.428	13.21	24.51
	MYO3	0.475	0.480	14.87	26.80
	TE	0.412	0.504	12.23	29.54
	TE*	0.257	0.594	12.86	29.50
0.5	OPT	0.194	0.772	10.73	21.56
	MYO1	0.386	0.599	17.12	24.41
	MYO2	0.292	0.684	17.65	26.82
	MYO3	0.227	0.742	17.88	28.41
	TE	0.204	0.740	12.87	29.78
	TE*	0.107	0.807	15.04	29.64

also has a smaller expected sample size than the TE design for all values of p_1. The probability of selecting T_1 when $p_1 \geq 0.4$ is similar for the OPT and MYO3 designs.

Compared with the TE design, a strength of the decision-theoretic approach lies in the fact that decisions of whether or not to drop any treatment are made in the light of data available from the other treatments. This advantage is illustrated by Table 2, which gives properties of the designs under the scenario $(p_1, p_2, p_3) = (0.4, 0.3, 0.3)$, where all three treatments provide an improvement over p_0 and T_1 is best but by the small increment of 0.1. The OPT design has a slightly higher probability of selecting T_1 and a slightly smaller sample size compared with both the MYO3 and TE designs. It can be seen that, with the OPT design, when T_1 is superior to T_2 and T_3, the latter will be dropped early in the study even if they represent an improvement relative to p_0 (the unequal expected numbers on T_2 and T_3 arise because of inequality in the breaking of ties in the design evaluated). In the TE design, by contrast, almost as many patients are recruited to T_2 and T_3 as to T_1. In particular, fewer patients receive inferior treatments under the OPT design.

The TE approach may be modified to allow direct comparisons of the different arms by dropping treatment T_t if either $\Pr(\theta_t < p_0 \mid \text{data}) > \pi_1$ or $\Pr(\theta_t < max_{t'=1,\ldots,m}\{\theta_{t'}\}) \mid \text{data}) > \pi_1{}^*$ for some $\pi_1{}^*$. Results for such a procedure with $\pi_1 = \pi_2 = \pi_1{}^* = 0.9$ are given in the rows of Tables 1 and 2 labeled TE*. It can be seen that the effect of the modification is to increase both the number of patients receiving the best treatment and the probability that this treatment is selected. Indeed, the latter is larger than that for the OPT design. This is achieved, however, at the cost of a large increase in the probability of erroneously making a selection when $p_1 = p_2 = p_3 = p_0$.

In practice, patients are often treated in cohorts rather than one at a time, with decisions made only after the results from each cohort have been observed. When there is an appreciable delay between treatment and the observation of a response, this is a practical necessity in order to avoid suspending accrual while waiting to observe the outcomes of patients who have been treated. This logistical problem has been addressed in the phase I dose-finding setting by Thall et al. (1999) and Cheung and Chappell (2000) and in the single-arm phase II

Table 2

Frequentist properties of designs as in Table 1 when $(p_0, p_1, p_2, p_3) = (0.2, 0.4, 0.3, 0.3)$

	Probability of selecting		Expected number of patients receiving		
Design	None	T_1	T_1	T_2	T_3
OPT	0.318	0.416	9.83	7.60	7.43
MYO3	0.356	0.387	12.21	8.06	8.08
TE	0.319	0.372	10.78	9.54	9.54
TE*	0.140	0.449	11.32	9.04	9.04

Table 3
Frequentist properties of designs as in Table 1 with patients treated in cohorts

Design	Cohort size	PNS	PCS	EN0
OPT	1	0.894	0.510	22.76
	2	0.867	0.530	23.75
	3	0.863	0.542	24.15
MYO1	1	0.965	0.354	15.23
	2	0.933	0.431	18.93
	3	0.914	0.454	19.63
MYO2	1	0.940	0.428	18.62
	2	0.926	0.454	20.94
	3	0.889	0.491	28.56
MYO3	1	0.915	0.480	22.04
	2	0.895	0.497	24.02
	3	0.884	0.507	25.39
TE	1	0.821	0.504	28.48
	2	0.850	0.474	27.70
	3	0.786	0.495	28.55

setting by Follman and Albert (1999). Suppose that patients are treated in cohorts of size k, so that, if at some point in the trial there are m' remaining treatments, a total of km' patients are randomized, with k receiving each remaining treatment. Table 3 gives PNS, PCS, and EN0 for the OPT, MYO1, MYO2, MYO3, and TE designs for cohorts of size one, two, or three.

In general, treating patients in cohorts reduces the opportunity to stop the trial early. Because in the designs considered early stopping necessarily leads to no treatment being selected, both the expected sample size and the probability of a treatment being selected are increased as k increases for all designs. This can be seen in Table 3; as k increases, PNS decreases and PCS and EN0 increase for both decision-theoretic designs. If there are m' remaining treatments, the study is terminated as soon as more than $30 - km'$ patients have been treated. This means that, in some cases, the sample size is reduced by increasing the cohort size k. For the TE design, in which the expected sample size is close to 30, this results in PNS, PCS, and EN0 being nonmonotone in k.

5. Discussion

Although there is a considerable literature on selection and screening designs (cf., Bechhofer, Santner, and Goldsman, 1995), this mostly focuses on preserving error rates and requires large sample sizes. There has been relatively little work of relevance for phase II or pre-phase II screening trials. In this article, we have used Bayesian decision theory to improve the designs proposed by Thall and Estey (1993). Our approach optimizes a decision process that allows inferior treatments to be dropped early, with the possible final selection of a single treatment for phase II testing. A strength of the method is that decisions about each treatment are made in light of the data available from the other treatment arms rather than comparing each treatment only with some critical threshold. Thus, if one or two treatments are considerably superior to the others, the latter will be dropped early in the study even if they represent an improvement relative to a target success rate.

Wang and Leung (1998) propose Bayesian decision-theoretic designs for pre-phase II screening studies in oncology in which a sequence of single-arm trials are conducted. Their approach is similar to that developed here, although they assume that an infinite number of potential treatments are available for testing. This means that, as in the TE design, decisions about when to stop testing a treatment T_t or to start a phase II trial are based on the data on T_t alone.

The major disadvantage of the decision-theoretic approach developed in this article is the computational requirement for the full backward induction approach, which is considerable. Calculations for the examples given in Section 4 were conducted using a Sun Ultra 5 workstation with 384 MB of memory. Although the calculation of the optimal design took only about 2 minutes, a very large amount of memory is required to store the expected gains for use later in the algorithm. For $n = 35$, the time was increased to 20 minutes because of the need to write large arrays to disk, and for $n = 40$, insufficient memory space was available to complete the calculations. Reducing the storage space needed by calculating the expected gains each time they are required increases the time required very dramatically, from a few minutes to several days. Even if a very large amount of memory was available, the large number of outcomes considered in the backward induction algorithm means that the calculations would rapidly become computationally infeasible as n was increased further.

The MYO3 design requires much less memory space and takes about 1 minute, so this approach remains feasible for larger studies. The fact that optimal actions are not found for all possible data sets in the algorithm does mean, however, that the time taken to calculate frequentist properties for the MYO3 approach exceeds considerably that needed for the OPT design.

While the gain functions proposed might be used in many studies, there are some circumstances in which alternative forms may be more appropriate. A simple modification to those proposed is to include specific consideration of the uncertainty regarding the outcome of further development in phase II and phase III trials. The gain from potential future patients could then be replaced by $\xi(\theta_{t^*})\Pi(\theta_{t^*} - p_0)$, where $\xi(\theta_{t^*})$ is an increasing function representing the probability, if the true success rate is θ_{t^*}, of successful further development. An alternative is to consider the financial gains and losses associated with further testing and successful development. Such an approach was used by Stallard (1998).

Inevitably, the construction of a gain function requires comparison of the cost of treatment with the gain to patients from success. Although we have chosen to attempt to present the costs c and K on the same scale as the probability of success, there is really no simple solution to this problem. A major theme of the approach suggested in this article is the choice of parameters in whatever gain functions are used in light of the frequentist properties of the design obtained. Sensitivity of the design to the gain function parameters used is illustrated by the contour plots in Figure 1. The values of Π and c do not have a very large effect on the properties of the design obtained, so the choice of K/Π is the most important decision. In particular, e.g., the design with $\Pi = 1000$ and $K = 150$ has properties almost identical to that with $\Pi = 100{,}000$ and $K = 15{,}000$.

The designs we have considered do not stop early if one treatment appears, based on interim data, to be superior to all the others. This is because the trial is very small and even the data from the completed trial can at best only suggest that a particular treatment may provide a real therapeutic advance. An important purpose of such small-scale early-phase trials is to estimate parameters about which, *a priori*, little is known. In the example considered in Section 4, CR is a necessary but not sufficient condition for long-term survival in acute leukemia. While it is reasonable to drop arms with low CR rates, it is desirable to continue recruitment to superior arms in order to gain as much information as possible from the trial prior to commitment to phase II testing.

We have focused on trials aimed at screening treatments for later study in a single-arm phase II trial. An alternative is to also conduct a randomized selection trial in phase II, possibly also including a control treatment. There are two important differences between such a trial and the pre-phase II screening trial considered here. First, the phase II trial might be much larger than the 30-patient trial considered in Section 4; second, in phase II, safety is often as important as efficacy, and both must be considered. A larger sample size means that, ethically, it is important that an arm be dropped not only for lack of efficacy but also if it shows a high rate of treatment-related adverse events.

In principle, extension of the method described here to a larger (phase II) trial should be straightforward. In practice, however, the backward induction required to construct the optimal design becomes computationally infeasible as the maximum sample size approaches even that of a moderately sized phase II study. In such circumstances, the myopic approach considered above would seem particularly suitable.

Phase II screening designs that consider both efficacy and safety endpoints have been considered by Thall and Sung (1998), using an approach that extends the TE method to incorporate multinomial endpoints and randomization. Stallard et al. (1999) obtained decision-theoretic designs for single-treatment phase II studies incorporating both safety and efficacy data. Their gain function could be used in the approach described here to give designs for phase II selection trials although, again, the increased complexity would increase the computational burden.

Acknowledgements

The authors are grateful to two referees and an associate editor for their helpful comments.

Résumé

Dans le contexte de petits essais randomisés pré-phase II, conçus pour établir une première sélection de traitements, et dans le but d'améliorer le plan expérimental de Thall et Estey (1993, *Statistics in Medicine* **12,** 1197–1211), nous proposons une approche bayésienne, issue de la théorie de la décision, qui permet d'optimiser une fonction de gain prenant simultanément en compte les bénéfices immédiats et futurs attendus pour les patients, les coûts unitaires par patient et le montant total du développement du traitement considéré. Afin de réduire la charge de calcul entraînée par ce processus inductif, nous présentons également des versions "myopes" de cette méthode, "myopes" en ce qu'elles ne considèrent à la fois qu'une, deux ou trois des décisions futures. Plusieurs de ces plans expérimentaux sont comparés entre eux, dans le contexte d'un essai de screening chez des patients atteints de leucémie myélogène aiguë.

References

Bechhofer, R. E., Santner, T. J., and Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* New York: Wiley.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

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

Cheung, Y. K. and Chappell, R. (2000). Sequential designs for phase 1 clinical trials with late-onset toxicities. *Biometrics* **56,** 1177–1182.

Fleming, T. R. (1982). One sample multiple testing procedure for phase II clinical trials. *Biometrics* **38,** 143–151.

Follman, D. A. and Albert, P. S. (1999). Bayesian monitoring of event rates with censored data. *Biometrics* **53,** 603–607.

Simon, R. (1989). Optimal 2-stage designs for phase-II clinical trials. *Controlled Clinical Trials* **10,** 1–10.

Stallard, N. (1998). Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* **54,** 279–294.

Stallard, N., Thall, P. F., and Whitehead, J. (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* **55,** 971–977.

Thall, P. F. and Estey, E. H. (1993). A Bayesian strategy for screening cancer treatments prior to phase II clinical evaluation. *Statistics in Medicine* **12,** 1197–1211.

Thall, P. F. and Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* **50,** 337–349.

Thall, P. F. and Sung, H.-G. (1998). Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine* **17,** 1563–1580.

Thall, P. F., Simon, R., and Estey, E. H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* **14,** 357–379.

Thall, P. F., Lee, J. J., Tseng, C.-H., and Estey, E. H. (1999). Accrual strategies for phase I trials with delayed patient outcome. *Statistics in Medicine* **18,** 1155–1169.

Wang, Y.-G. and Leung, D. H.-Y. (1998). An optimal design for screening trials. *Biometrics* **54,** 243–250.

Received August 2000. *Revised July* 2001.
Accepted July 2001.