BIOMETRICS 55, 971–977
September 1999

Decision Theoretic Designs for Phase II Clinical Trials with Multiple Outcomes

Nigel Stallard,[1,*] Peter F. Thall,[2] and John Whitehead[1]

[1]Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading RG6 6FN, U.K.

[2]Department of Biostatistics, M. D. Anderson Cancer Center, University of Texas, Box 237, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.

*email: n.stallard@reading.ac.uk

SUMMARY. In many phase II clinical trials, it is essential to assess both efficacy and safety. Although several phase II designs that accommodate multiple outcomes have been proposed recently, none are derived using decision theory. This paper describes a Bayesian decision theoretic strategy for constructing phase II designs based on both efficacy and adverse events. The gain function includes utilities assigned to patient outcomes, a reward for declaring the new treatment promising, and costs associated with the conduct of the phase II trial and future phase III testing. A method for eliciting gain function parameters from medical collaborators and for evaluating the design's frequentist operating characteristics is described. The strategy is illustrated by application to a clinical trial of peripheral blood stem cell transplantation for multiple myeloma.

KEY WORDS: Backward induction; Clinical trial design; Optimal stopping; Safety and efficacy monitoring; Sequential procedure.

1. Introduction

Phase II clinical trials are usually small studies conducted to evaluate the antidisease effect of an experimental therapy and to obtain safety information. If the new therapy is determined to be both efficacious and safe compared to existing standard treatments, then it may be studied further in a randomized phase III trial.

Ethical concerns that a trial must be stopped early if the experimental treatment appears to be unsafe or ineffective have led to the development of sequential designs for phase II trials. Most of the development of phase II designs has been in the area of oncology, where the severity of the disease and possible side effects make early stopping particularly desirable. Designs have been proposed by numerous authors, including Fleming (1982), Simon (1989), Bellisant, Benichou, and Chastang (1990), Ensign et al. (1994), Thall and Simon (1994a,b), Heitjan (1997), and Stallard (1998). All of these designs are based on a single binary indicator of treatment efficacy with safety considerations ignored. More recently, designs that monitor safety explicitly, with rules to stop the trial if the treatment is either inefficacious or too toxic, have been proposed by Bryant and Day (1995), Conaway and Petroni (1995, 1996), Thall, Simon, and Estey (1995, 1996), and Thall and Sung (1998).

In this paper, we propose a method for constructing sequential designs based on both efficacy and safety using Bayesian decision theory. This approach, in which the design optimizes some gain function, has been described in general settings by Raiffa and Schlaifer (1961), DeGroot (1970), and Berger (1985), among others, and more specifically in the setting of a phase II trial by Berry and Stangl (1996). The decision theoretic approach is briefly described at the beginning of Section 2. A "pure Bayesian" could use the approach described to obtain designs optimal for their choice of a prior and a gain function. In many cases, however, it is difficult to elicit both the prior information and the gain function from the clinical physicians who are planning the trial. As an alternative, we propose that a particular form be chosen for the gain function and that prior information, trial goals, and certain gain function parameters be elicited from the clinician, with other design parameters chosen in view of the frequentist properties of the design. As the application in Section 3 will illustrate, the ability to select numerical values for the gain function parameters in this way provides a method by which the clinician may obtain a design that will be acceptable in practice. Thall and Simon (1994a,b), Rosner and Berry (1995), Thall et al. (1995, 1996), and Thall and Sung (1998) have also constructed stopping rules from a Bayesian viewpoint and have evaluated their frequentist properties. Fisher (1996) has referred to this as a *stylized Bayesian* approach.

2. A Bayesian Decision Theoretic Approach

Although most phase II studies in oncology require that all patients receive the experimental treatment, E, the trial is inherently comparative because decisions involving E must be made relative to some standard, S. We consider the patient outcome to be characterized by a multinomial random vari-

able with k possible values comprising a partition of all relevant combinations of efficacy and adverse events. For $t = E, S$ and $i = 1, \ldots, k$, we denote the probability of outcome i for patients receiving treatment t by θ_{ti} so that $\theta_{t1} + \cdots + \theta_{tk} = 1$. Following the Bayesian approach, the probability vectors $\boldsymbol{\theta}_E = (\theta_{E1}, \ldots, \theta_{Ek})'$ and $\boldsymbol{\theta}_S = (\theta_{S1}, \ldots, \theta_{Sk})'$ corresponding to E and S are treated as random. Prior distributions for $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_S$ will be assumed to take the Dirichlet form. The priors reflect expert opinion concerning standard and experimental therapies. Typically, the prior for $\boldsymbol{\theta}_S$ reflects historical data or clinical experience of S, whereas the prior of $\boldsymbol{\theta}_E$ reflects little or no clinical experience. In a single-arm trial, no patients receive S, so the prior for $\boldsymbol{\theta}_S$ is never updated.

At any stage during the trial, three possible actions are envisaged:

Action P: Stop the study and declare E promising.
Action N: Stop the study and declare E not promising.
Action C: Continue the study.

A maximum sample size, M, is chosen so that after observation of M patients, only actions P and N are available. We also limit the availability of action P to avoid the possibility of declaring E promising on the basis of little or no data from the trial. Two cases will be considered. In the first, action P may be taken only after the maximum of M patients have been observed. The second case generalizes the first by allowing P after at least M_1 patients for some $M_1 \leq M$. Taking $M_1 = 0$ allows early stopping as a result of superiority of E over S at any stage in the trial. This is equivalent to using the upper boundary in the Thall et al. (1995) design. In practice, a larger value of M_1 seems to be more appropriate to avoid declaring E promising on the basis of data from very few patients. Reasonable values of M_1 might be similar to the stage 1 sample size of a two-stage design, such as that of Simon (1989). In practice, both M_1 and M will be specified during consultation with the clinician. Action N will be allowed at any stage during the trial, so no minimum sample size needs be specified in advance.

To decide between P, N, and C, these possible actions must be assigned gains indicative of their desirability. The gains for the actions P and N depend on $\boldsymbol{\theta}_E$, $\boldsymbol{\theta}_S$, and n and will be denoted by $G_P(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S, n)$ and $G_N(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S, n)$, respectively. The expected gain from continuing beyond the nth patient to the $(n+1)$st depends on the action that would be taken after observing the response from that patient and can be obtained by backward induction (DeGroot, 1970).

Because it may be difficult for a clinician to specify G_P and G_N, we now suggest a form for the gain functions that summarizes the concerns of those with an interest in the trial. Suppose that some function $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ gives the gain for a patient being treated with E under the pair $(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$, where $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S) = 0$ if E and S are equally desirable treatments. We will call this the *patient gain*. A possible form for g is given. The total gain to the n patients in a trial is $ng(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$.

Everything else being equal, it seems sensible to stop the study earlier rather than later. This reflects a cost per patient for conducting the trial, besides the possible loss arising from negative values of $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$. Although it is difficult to quantify this cost, which could be either financial or a consequence of a delay in development of E or of alternative therapies, on the same scale as the patient gain, it cannot be ignored. Denoting the cost per patient by c, we define the *trial gain* to be $ng(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S) - nc$. The declaration of E as promising or not does not affect the trial gain, which depends on the conduct of the trial itself.

If E is declared promising, then there will be a cost, denoted by K, associated with further development and testing of E in phase III. There will also be a benefit to future patients depending on the true value of $\boldsymbol{\theta}_E$, which we take to be $\Pi\, g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ for some $\Pi > 0$. Note that if $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ is negative, then $\Pi\, g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ is the loss arising from erroneously declaring E promising. Because this gain is on the same scale as the patient gain, Π may be interpreted as the number of future patients to benefit from treatment with E, the so-called *patient horizon*. However, the values of Π might also reflect the gains to clinicians or pharmaceutical companies and thus need not be equated to a number of potential patients. The gain functions for the actions P and N are thus given by

$$G_P(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S, n) = ng(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S) - nc + \Pi g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S) - K, \quad (1)$$

$$G_N(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S, n) = ng(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S) - nc. \quad (2)$$

To specify a form for the patient gain function $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$, we assign utilities $u_1, \ldots, u_k$ to the k possible outcomes. As the expected utility under E would be $\mathbf{u}'\boldsymbol{\theta}_E = u_1\theta_{E1} + \cdots + u_k\theta_{Ek}$, the patient gain could thus be defined as

$$g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S) = \mathbf{u}'(\boldsymbol{\theta}_E - \boldsymbol{\theta}_S). \quad (3)$$

In practice, $u_1, \ldots, u_k$ may be elicited from the clinician. Because it is generally straightforward to identify the best and worst outcomes and because the problem is invariant to scaling and shifting the vector $\mathbf{u}$, for convenience, we assign utilities $+1$ and -1 to the best and worst outcomes, respectively, with the others taking on values in the interval $[-1, +1]$. The form of $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ given by (3) provides an explicit trade-off between desirable and undesirable outcomes, where an increase in the probability of the latter is tolerated if accompanied by an increase in that of the former, the relative size of which is determined by the choice of $\mathbf{u}$.

The patient gain given by (3) is equal to zero when $\boldsymbol{\theta}_E = \boldsymbol{\theta}_S$, indicating that E and S are equally attractive. Many clinicians, however, are accustomed to thinking that E should be considered promising compared with S only if $\boldsymbol{\theta}_E = \boldsymbol{\theta}_S + \boldsymbol{\delta}$, for some targeted improvement, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)'$. An alternative definition of $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ accommodating this is

$$g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S) = \mathbf{u}'(\boldsymbol{\theta}_E - (\boldsymbol{\theta}_S + \boldsymbol{\delta})). \quad (4)$$

Substituting this form for $g(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ into (1) and (2) gives

$$G_P(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S, n) = n\{\mathbf{u}'(\boldsymbol{\theta}_E - \boldsymbol{\theta}_S) - \mathbf{u}'\boldsymbol{\delta} - c\} + \Pi\mathbf{u}'(\boldsymbol{\theta}_E - \boldsymbol{\theta}_S) - \Pi\mathbf{u}'\boldsymbol{\delta} - K, \quad (5)$$

$$G_N(\boldsymbol{\theta}_E, \boldsymbol{\theta}_S, n) = n\{\mathbf{u}'(\boldsymbol{\theta}_E - \boldsymbol{\theta}_S) - \mathbf{u}'\boldsymbol{\delta} - c\}. \quad (6)$$

Because backward induction requires comparing $E(G_P)$ with $E(G_N)$, including $\boldsymbol{\delta}$ in (4) is equivalent to adding $\Pi\mathbf{u}'\boldsymbol{\delta}$ to the cost K. That is, the requirement that E must show some improvement over S contributes to the future development cost of E, and formulation (4) provides no generalization over (3). As clinicians may find it easier to express a required improvement directly rather than in terms of cost, we will use forms (5) and (6) with $K = 0$.

As an alternative to direct specification of the parameters in the gain functions (5) and (6), we propose a stylized Bayesian approach in which $\mathbf{u}$ is elicited from the clinician along with the priors, M and M_1, whereas $\boldsymbol{\delta}$, c, and Π are chosen by consideration of the frequentist properties of designs obtained. This approach is illustrated by the example given in Section 3.

3. Application

Although the approach described previously is applicable for any k, two cases are of particular interest. The first case has $k = 3$ with the possible outcomes *toxicity*, *efficacy without toxicity*, and *neither efficacy nor toxicity*. The second case has binary responses for both efficacy and toxicity so that $k = 4$. Although these two cases are not comprehensive, they encompass a very large proportion of phase II trials. The case $k = 4$ also reduces to the case $k = 3$ in settings where the occurrence of the adverse event renders efficacy either irrelevant or impossible. We will illustrate the method described in the case $k = 3$ with reference to a trial in peripheral blood stem cell transplantation in multiple myeloma patients.

The trial discussed in this section was originally designed for use at the M. D. Anderson Cancer Center using the method of Thall et al. (1995), although for administrative reasons, the trial was never actually conducted. The decision theoretic designs described next are based on the outcomes, priors, maximum sample sizes, and trial goals originally specified by the clinician.

The phase II trial was designed to assess autologous CD 34-selected peripheral blood stem cell transplantation, E, in patients with poor prognosis multiple myeloma. The three clinically relevant outcomes, scored at 4 weeks post-transplant, were complete remission (CR), transplant-related mortality (TRM), or neither of these (NCT). Historical data were available for 43 patients given the standard, S, of conventional bone marrow transplantation. Of these, 33 had outcome NCT, 2 achieved CR, and 8 experienced TRM. For $t = E, S$, we denote the components of the probability vector $\boldsymbol{\theta}_t$ by $(\theta_{t,\mathrm{NCT}}, \theta_{t,\mathrm{CR}}, \theta_{t,\mathrm{TRM}})$. A Dirichlet prior with parameters 33, 2, and 8 was thus assumed for $\boldsymbol{\theta}_S$. A Dirichlet prior with parameters 2.302, 0.140, and 0.558 was assumed for $\boldsymbol{\theta}_E$ because this prior has the same mean as that for $\boldsymbol{\theta}_S$ but a weight equivalent to only three observations, reflecting the relative lack of prior knowledge of E. The clinician indicated that E would be considered promising compared with S if an improvement of 0.15 in θ_{CR} could be achieved, from a mean of 0.047 with S, but that no increase in θ_{TRM} would be acceptable. A design was obtained using the method of Thall et al. (1995), based on the stopping criteria $Pr(\theta_{S,\mathrm{CR}} + .15 < \theta_{E,\mathrm{CR}} \mid \mathrm{data}) < .025$ and $Pr(\theta_{S,\mathrm{TRM}} < \theta_{E,\mathrm{TRM}} \mid \mathrm{data}) > .975$, with a maximum sample size of $M = 40$.

The decision theoretic designs obtained next will be evaluated by consideration of their frequentist properties. These will be computed under five scenarios, each characterized by fixed values of the vectors $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_S$ and representative of a clinical circumstance of interest to the clinician. Abusing notation, for the remainder of this section $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_S$ will be used to denote these fixed values, which should not be confused with the random parameters considered under the Bayesian model during the conduct of the trial. The scenarios are given in Table 1, where in each case, $\boldsymbol{\theta}_S$ is equal to $(0.767, 0.047, 0.186)'$, the mean value of the prior distribution described earlier. In scenario 1, $\boldsymbol{\theta}_E$ equals $\boldsymbol{\theta}_S$. In scenarios 2, 3, and 4, $\boldsymbol{\theta}_E$ achieves the required increase of 0.15 in θ_{CR}, with the gain coming equally from θ_{NCT} and θ_{TRM} in scenario 2, solely from θ_{NCT} in scenario 3, and solely from θ_{TRM} in scenario 4. In scenario 5 there is no improvement in θ_{CR}, but θ_{TRM} rises by 0.10.

Table 1
Scenarios considered in the frequentist evaluation of designs in Section 4

Scenario	Description	$\theta_{E,\mathrm{NCT}}$	$\theta_{E,\mathrm{CR}}$	$\theta_{E,\mathrm{TRM}}$
1	$E(\boldsymbol{\theta}_S)$	0.767	0.047	0.186
2	$\theta_{\mathrm{CR}} \uparrow .15$, $\theta_{\mathrm{TRM}} \downarrow .075$	0.692	0.197	0.111
3	$\theta_{\mathrm{CR}} \uparrow .15$, $\theta_{\mathrm{NCT}} \downarrow .15$	0.617	0.197	0.186
4	$\theta_{\mathrm{CR}} \uparrow .15$, $\theta_{\mathrm{TRM}} \downarrow .15$	0.767	0.197	0.036
5	$\theta_{\mathrm{TRM}} \uparrow .10$	0.667	0.047	0.286

Given a stopping rule, such as that obtained using the backward induction described, in which the number of patients is limited, it is possible to conduct an exhaustive enumeration of all possible outcomes from the patients in the study, together with the calculation of the probability of each outcome for fixed values of $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_S$. This provides a complete calculation of the discrete probability distribution of the vector of responses, which in turn allows the derivation of the frequentist operating characteristics of the design. These exact calculations serve the purpose of simulations performed by other authors. The frequentist properties, consisting of $pr(P)$ and $E(n)$ under the five scenarios, are computed in this way for each design obtained. To develop a decision theoretic design for this trial, we first note that any acceptable design must have a large probability of declaring E promising, $pr(P)$, under scenarios 2, 3, and 4 and a small $pr(P)$ under scenarios 1 and 5; furthermore ethical considerations dictate that expected sample size, $E(n)$, must be small under scenario 1 and very small under scenario 5. To achieve this using the decision theoretic structure, we first specify $\mathbf{u}$ and choose $\boldsymbol{\delta}$ so as to make $pr(P \mid \text{scenario 1})$ small and $pr(P \mid \text{scenario 2})$ large for a wide range of Π and c values. We then choose Π and c to minimize $E(n \mid \text{scenario 1})$ subject to the constraint that $pr(P \mid \text{scenario 2})$ must be bounded below by some large probability. The values of u_{NCT}, the lower bound on $pr(P \mid \text{scenario 2})$, and possibly $\boldsymbol{\delta}$ are then calibrated to obtain an acceptable design. The following account is intended to serve as a guide for others using this methodology.

As CR and TRM are the best and worst outcomes, we assign them utilities $+1$ and -1, respectively. Because specifying u_{NCT} may not be straightforward, the meaning of a particular numerical value might, in practice, be assessed by evaluating the operating characteristics of the resulting design and calibrating u_{NCT} on that basis. We initially set $u_{\mathrm{NCT}} = 0$ and studied the design using three different values for the required improvement, $\boldsymbol{\delta}$, specifically, $(0, 0, 0)'$, $(-0.075, 0.15, -0.075)'$, denoted by $\boldsymbol{\Delta}$ and equal to the difference between $\boldsymbol{\theta}_E$ and $\boldsymbol{\theta}_S$ under scenario 2, and $\boldsymbol{\Delta}/2$. For each $\boldsymbol{\delta}$, we obtained designs over a range of c and Π, using the priors given before with $M = 40$ and action P allowed only after observation of exactly

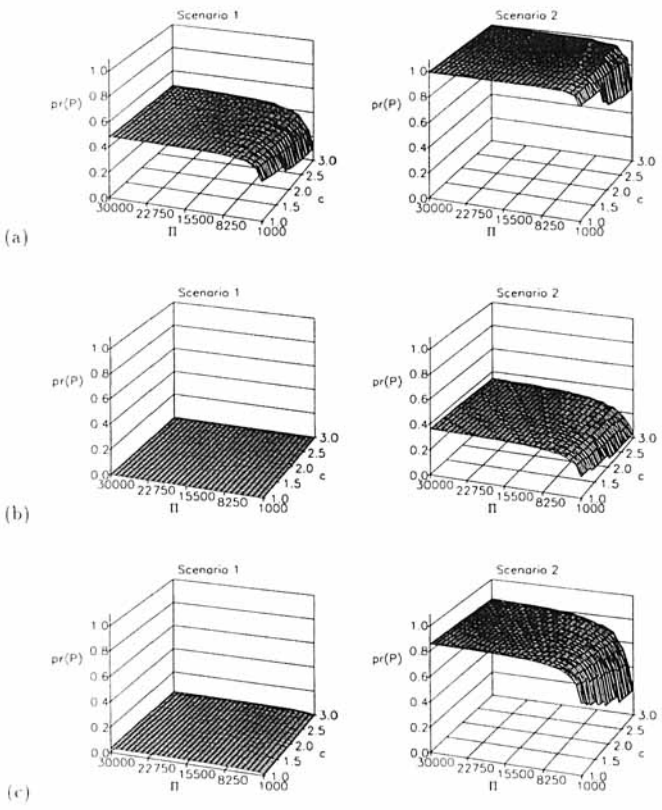

Figure 1. $pr(P)$ under scenarios 1 and 2 for a range of values of Π and c for (a) $\boldsymbol{\delta} = \mathbf{0}$, (b) $\boldsymbol{\delta} = \boldsymbol{\Delta}$, and (c) $\boldsymbol{\delta} = \boldsymbol{\Delta}/2$.

40 patients ($M_1 = M$). Figure 1 shows $pr(P)$, under scenarios 1 and 2 over the (c, Π) study domain for the three choices of $\boldsymbol{\delta}$. Setting $\boldsymbol{\delta} = \mathbf{0}$ leads to a relatively high $pr(P \mid \text{scenario } 1)$, whereas setting $\boldsymbol{\delta} = \boldsymbol{\Delta}$ leads to a relatively low $pr(P \mid \text{scenario } 2)$. Thus, $\boldsymbol{\Delta}/2$ appears to be the most appropriate choice for $\boldsymbol{\delta}$. Values of $E(n)$ under scenarios 1 and 2 over the same (c, Π) domain with $\boldsymbol{\delta} = \boldsymbol{\Delta}/2$ are given in Figure 2.

Larger values of Π increase the likelihood that E will be declared promising, as the reward for doing so is greater. However, values of $E(n)$ produced are also larger. Increasing c has a lesser effect; because observations are more expensive, it leads to a smaller trial and hence to a slightly smaller $pr(P)$. Taken together, these results suggest that c and Π may be chosen to achieve a compromise between a large $pr(P)$ under scenario 2 and a small $E(n)$ under scenario 1. A simple search can be used to find, e.g., the (c, Π) pair that minimizes $E(n \mid \text{scenario } 1)$ subject to $pr(P \mid \text{scenario } 2)$ being bounded below by 0.85. This design is achieved by $(c, \Pi) = (1.15,\ 20{,}000)$. The properties of the design under the five scenarios are given in the columns labeled "Decision Theoretic, $M_1 = 40$" in Table 2, with the corresponding properties for the Thall et al. (1995) design, described earlier, given in the columns labeled "Thall et al. (1995)" for comparison. It can be seen that the decision theoretic approach yields a design with $pr(P)$ being smaller under scenario 1 and larger under scenario 2 than that of the Thall et al. (1995) design. However, in this case, the cost is an increased expected sample size under scenario 1. The decision theoretic design also has $pr(P)$ being much larger under scenario 4 and smaller under

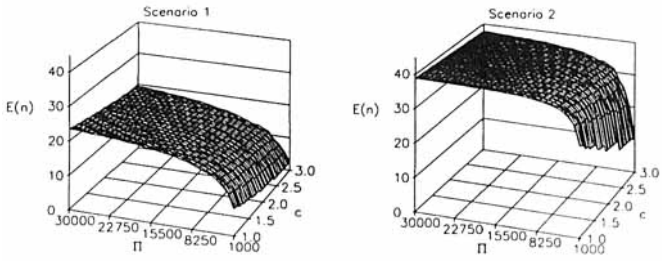

Figure 2. Expected sample size under scenario 1 for a range of values of Π and c for $\boldsymbol{\delta} = \boldsymbol{\Delta}/2$.

scenario 3. As discussed in more detail later, this follows from comparison of the scenarios when considering the vector of utilities, $\mathbf{u}$.

In designing the trial using the method of Thall et al. (1995), the clinician initially planned a maximum sample size of 20, but consideration of the operating characteristics of the design led to the use of $M = 40$. With this in mind, we next obtain a design as that shown above but with $M_1 = 20$, i.e., a design in which the treatment can be declared promising, provided that at least 20 patients have been included in the trial. For this extended class of designs, the effect of c is more marked because the decision to stop early no longer amounts to a decision that E is not promising. Frequentist properties are given in the last two columns of Table 2, in this case for the design with $c = 1.5$ and $\Pi = 22{,}000$. Comparing the properties of this design with those of the design with $M_1 = M = 40$ shows an unsurprising drop in the expected sample size, particularly under scenarios 2, 3, and 4, together with slightly lower values of $pr(P)$ under the "good" scenarios and slightly higher values under the "bad" ones. This may be interpreted as a penalty for the smaller sample sizes.

The probabilities in Table 2 show that, although the $pr(P)$ for the Thall et al. (1995) design is similar for scenarios 2, 3, and 4, these probabilities differ markedly for the decision theoretic designs. A contour plot of $pr(P)$ for values of $\theta_{E,\text{TRM}}$ and $\theta_{E,\text{CR}}$, with $\boldsymbol{\theta}_S$ fixed at $(0.767, 0.047, 0.186)'$ for the design with $M_1 = 20$, is given in Figure 3 and illustrates this difference. The probability contours in Figure 3 follow the direction of contours of the patient utility when $u_{\text{NCT}} = 0$. The decision theoretic designs thus discriminate to the greatest extent between scenarios farthest apart in terms of their patient gain. The direction of the contours of the patient utility is determined by the value chosen for u_{NCT}; decreasing

Table 2
Properties of the designs obtained in Section 3 $u_{\text{NCT}} = 0$

			Decision theoretic			
	Thall et al. (1995)		$M_1 = 40$		$M_1 = 20$	
Scenario	$pr(P)$	$E(n)$	$pr(P)$	$E(n)$	$pr(P)$	$E(n)$
1	0.051	15.6	0.035	21.7	0.039	21.1
2	0.810	35.3	0.850	38.5	0.850	28.6
3	0.799	35.0	0.571	35.5	0.574	30.6
4	0.811	35.3	0.988	39.8	0.987	23.4
5	0.046	15.2	0.003	15.9	0.003	15.5

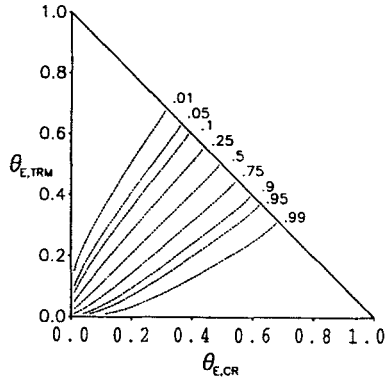

Figure 3. Contours of $pr(P)$ for the design in Table 2, with $M_1 = 20$.

u_{NCT} from 0 increases the slope of the contour lines. This acts to make scenarios 2, 3, and 4 more similar and scenarios 1 and 2 more different, with this difference greatest when $u_{\mathrm{NCT}} = -1/3$. Columns 2 and 3 of Table 3 give properties for the design with $u_{\mathrm{NCT}} = -1/3$, $\boldsymbol{\delta} = \boldsymbol{\Delta}/2$, and $M_1 = 20$, with c and Π chosen so as to minimize $E(n \mid \text{scenario 1})$ subject to $pr(P \mid \text{scenario 2}) \geq 0.85$. The values of c and Π for this design are 2.6 and 26,000, respectively. The effect of changing u_{NCT} is as anticipated; namely, it increases the difference between $pr(P)$ under scenarios 1 and 2 and decreases the differences between scenarios 2, 3, and 4. The choice of $\mathbf{u}$ is thus important and is related to the selection of the scenarios at which the design is to be evaluated, thereby illustrating that the goal of a trial cannot be considered separately from the desirability of possible outcomes. A discussion of this point with clinical collaborators could be instructive.

Depending on the views of the clinician, a more or less exhaustive search for appropriate $\boldsymbol{\delta}$ could be performed in practice. Figure 2 shows that, when $pr(P \mid \text{scenario 1})$ is small, it is relatively unaffected by the choice of c and Π. Thus, $\boldsymbol{\delta}$ is important in determining this probability. Table 2 shows that $pr(P \mid \text{scenario 1})$ is smaller for the decision theoretic designs than for the Thall et al. (1995) design. By a careful selection of $\boldsymbol{\delta}$, under both scenarios 1 and 2 $pr(P)$ can be increased slightly so that $pr(P \mid \text{scenario 1})$ remains small, but $pr(P \mid \text{scenario 2})$ attains larger values for smaller expected sample sizes than in the designs given in Table 2. The fourth and fifth columns of Table 3 give such a design, with $\boldsymbol{\delta} = 0.38\boldsymbol{\Delta}$, $u_{\mathrm{NCT}} = -1/3$, $c = 1.7$, and $\Pi = 4000$. It can be seen that this design has properties under both scenarios 1 and 2 that are preferable to those of the Thall et al. (1995) design.

Table 3
Properties of the designs obtained in Section 3 with $u_{\mathrm{NCT}} = -1/3$ *and* $M_1 = 20$

	$\boldsymbol{\delta} = \boldsymbol{\Delta}/2$		$\boldsymbol{\delta} = 0.38\boldsymbol{\Delta}$	
Scenario	$pr(P)$	$E(n)$	$pr(P)$	$E(n)$
1	0.021	19.9	0.051	14.0
2	0.852	26.6	0.812	20.5
3	0.695	28.8	0.695	21.4
4	0.951	23.1	0.896	19.7
5	0.004	16.4	0.012	11.6

Application of the method to the $k = 4$ case is conceptually straightforward. In this case, however, the fact that both u_2 and u_3 must be specified and that the effect of choices of u_2 and u_3 is harder to visualize in the four-outcome case leads to additional practical difficulties in eliciting parameters and calibrating the resulting designs. These difficulties would naturally increase for $k > 4$ with the specification of patients' utilities for so many outcomes. A computer program written in C to calculate designs in the cases $k = 3$ and $k = 4$ is available via anonymous ftp from odin.mdacc.tmc.edu as decBayes97.tar.gz in the subdirectory/pub/source.

4. Discussion

Although the decision theoretic approach has been available for many years, there have been relatively few attempts to apply it to clinical trials. The reluctance to use decision theory may be due to the difficulty of specifying a suitable gain function (Efron, 1986). Indeed, Spiegelhalter and Freedman (1988) specifically rejected such an approach in the Bayesian analysis of phase III trials for this reason. Staquet and Sylvester (1977), Sylvester (1988) (see also Hilden, 1990), and Brunier and Whitehead (1994) have used the Bayesian decision theory to design phase II studies with a fixed sample size. Berry and Ho (1988), Berry, Wolff, and Sack (1994), Cressie and Biele (1994), Lewis and Berry (1994), and Stallard (1998) have applied decision theory to construct optimal sequential designs for a variety of clinical trials. However, their designs consider only univariate outcomes. We have proposed a general form for the gain function that explicitly quantifies a trade-off between efficacy and toxicity for the patients in the trial and that includes the costs and benefits associated with conducting the phase II trial itself and with continuing the development process in the future. Aside from the decision theoretic structure, the substantive practical advances over the approach of Thall et al. (1995) are that the assignment of utilities to patient outcomes and the calibration of design parameters together may produce a more attractive design.

Because the actual values used for the gain function parameters determine the design obtained, they must be chosen with care. The values used ideally reflect the collective views of those involved in the trial. Because a single design must be obtained, and hence a single gain function used, it is envisaged that the gain be specified by the clinician overseeing the study with the interests of the patients in mind. A pure Bayesian approach, in which the clinician specifies prior distributions and the gain function, could be based on the gain functions described. We propose, however, a *stylized Bayesian* method in which the form of the gain function is chosen in advance and some parameters are specified by the clinician, whereas the others are chosen in light of the frequentist properties of the designs obtained.

The choice of the extent to which a pure Bayesian or a stylized Bayesian approach is adopted must depend on the clinicians' confidence in their ability to choose attractive designs on the basis of either gain functions or frequentist properties evaluated under a small number of scenarios. Consideration of the latter alone might suggest that one should obtain a design that minimizes the expected sample size under a "null"

scenario similar to scenario 1 in the application given earlier, subject to constraints on $pr(P)$ under scenarios 1 and 2. The sequential probability ratio test (SPRT) of Wald (1947) has such properties, but it does not arise from the methodology described here, thereby showing that the designs obtained here are not optimal in this sense. The fact that the SPRT is not generally used in phase II studies shows that frequentist properties evaluated under only two scenarios are not sufficient to judge a design.

Although suitable for any k, the approach described here is likely to be most commonly used with either $k = 3$ or $k = 4$. The latter case has been discussed extensively by Bryant and Day (1995), Conaway and Petroni (1995, 1996), and Thall et al. (1996). In contrast to the approach adopted here, however, all these authors focus on the marginal probabilities of toxicity and efficacy, thereby effectively reducing the dimensionality of the problem. This maneuver can be misleading, as given marginals may correspond to very different joint probabilities and hence very different clinical scenarios. These differences are reflected by the values of the gain function we have proposed.

An obvious extension could be to the design of randomized phase II trials in which the approach described here could be used with the prior for $\boldsymbol{\theta}_S$ that is updated after each observation in a way similar to that for $\boldsymbol{\theta}_E$.

Acknowledgements

This work was conducted while Professor Thall was visiting the University of Reading and was funded by an EPSRC fellowship. The work of the MPS Research Unit is sponsored by Amgen.

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

Résumé

Dans beaucoup d'essais thérapeutiques de phase II, il est essentiel d'évaluer à la fois l'efficacité et la sécurité. Bien que plusieurs schémas de phase II assumant des résultats multiples aient été récemment proposés, aucun n'a été obtenu à partir de la théorie de la décision. Ce papier décrit une approche théorique de décision bayésienne pour construire des schémas de phase II basés à la fois sur l'efficacité et les réactions indésirables. La fonction de gain inclut les utilités associées au résultat du patient, une prime pour l'intérêt du nouveau traitement, et les coûts associés à l'essai de phase II et à la future étude de phase III. Une méthode pour obtenir les paramètres de la fonction de gain des collaborateurs médicaux, et évaluer les caractéristiques opératoires du schéma fréquentiel est décrite. La stratégie est illustrée par une application à un essai clinique de transplantation de cellules souches sanguines dans le myélome multiple.

References

Bellisant, E., Benichou, J., and Chastang, C. (1990). Application of the triangular test to phase II cancer trials. *Statistics in Medicine* **9,** 907–917.

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

Berry, D. A. and Ho, C.-H. (1988). One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics* **44,** 219–227.

Berry, D. A. and Stangl, D. K. (1996). Bayesian methods in health-related research. In *Bayesian Biostatistics*, D. A. Berry and D. K. Stangl (eds), 3–66. New York: Dekker.

Berry, D. A., Wolff, M. C., and Sack, D. (1994). Decision making during a phase III randomized controlled trial. *Controlled Clinical Trials* **15,** 360–378.

Brunier, H. C. and Whitehead, J. (1994). Sample sizes for phase II clinical trials derived from Bayesian decision theory. *Statistics in Medicine* **13,** 2493–2502.

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

Conaway, M. R. and Petroni, G. R. (1995). Bivariate sequential designs for phase II trials. *Biometrics* **51,** 656–664.

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

Cressie, N. and Biele, J. (1994). A sample-size-optimal Bayesian decision procedure for sequential pharmaceutical trials. *Biometrics* **50,** 700–711.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

Efron, B. (1986). Why isn't everyone a Bayesian? *American Statistician* **40,** 1–5.

Ensign, L. G., Gehan, E. A., Kamen, D. S., and Thall, P. F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine* **13,** 1727–1736.

Fisher, L. D. (1996). Comments on Bayesian and frequentist analysis and interpretation of clinical trials. *Controlled Clinical Trials* **17,** 423–434.

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

Heitjan, D. F. (1997). Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine* **16,** 1791–1802.

Hilden, J. (1990). Corrected loss calculation for phase II trials. Reader reaction. *Biometrics* **46,** 535–538.

Lewis, R. J. and Berry, D. A. (1994). Group sequential clinical trials: A classical evaluation of Bayesian decision-theoretic designs. *Journal of the American Statistical Association* **89,** 1528–1534.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory.* Cambridge, Massachusetts: MIT Press.

Rosner, G. L. and Berry, D. A. (1995). A Bayesian group sequential design for a multiple-arm randomized clinical trial. *Statistics in Medicine* **14,** 381–394.

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

Spiegelhalter, D. J. and Freedman, L. S. (1988). Bayesian approaches to clinical trials. In *Bayesian Statistics 3*, J. M. Bernado, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (eds), 453–477. Oxford: Oxford University Press.

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

Staquet, M. J. and Sylvester, R. J. (1977). A decision theory approach to phase II clinical trials. *Biomedicine* **26,** 262–266.

Sylvester, R. J. (1988). A Bayesian approach to the design of phase II clinical trials. *Biometrics* **44,** 823–836.

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

Thall, P. F. and Simon, R. (1994b). A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Controlled Clinical Trials* **15,** 463–481.

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., Simon, R., and Estey, E. H. (1996). New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *Journal of Clinical Oncology* **14,** 296–303.

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.

Wald, A. (1947). *Sequential Analysis.* New York: John Wiley.

Received December 1997. *Revised August* 1998.
Accepted August 1998.