STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:763–780 (DOI: 10.1002/sim.1399)

Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes

Peter F. Thall[1,*,†], J. Kyle Wathen[1], B. Nebiyou Bekele[1], Richard E. Champlin[2], Laurence H. Baker[3] and Robert S. Benjamin[4]

[1]*Department of Biostatistics, Box 447, University of Texas, M.D. Anderson Cancer Center, Houston, Texas, U.S.A.*
[2]*Department of Blood and Marrow Transplantation, Box 423, University of Texas, M.D. Anderson Cancer Center, Houston, Texas, U.S.A.*
[3]*University of Michigan Comprehensive Cancer Center, 1500 East Medical Drive, 7216 CCGC, Ann Arbor, Michigan 48109-0948, U.S.A.*
[4]*Department of Sarcoma Medical Oncology, Box 450, University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.*

SUMMARY

We propose a methodology for conducting phase II clinical trials in settings where the disease is categorized into multiple subtypes. A hierarchical Bayesian model is assumed for treatment effects within the subtypes. The hierarchical model, which is tailored to each particular application, allows treatment effects to differ across subtypes while assuming *a priori* that the effects are exchangeable and correlated. Two applications are described. The first is a trial of imatinib for sarcoma in which treatment activity is characterized by a binary indicator of tumour response. The second is a phase II trial of a new preparative regimen for allogeneic bone marrow transplantation in patients with haematologic malignancies, with treatment effect characterized by the mean time from transplant to disease progression or death. The applications illustrate how the hierarchical Bayesian model borrows strength across subtypes.

KEY WORDS: activity trial; bone marrow transplantation; heterogeneity; phase II clinical trial; sarcoma

1. INTRODUCTION

A common problem in oncology and other medical settings is to determine whether a new treatment is sufficiently promising to warrant further study in a large scale randomized trial. This typically is addressed by conducting a single-arm, 'phase II' study with patient outcome

* Correspondence to: Peter F. Thall, Department of Biostatistics, Box 447, University of Texas, M.D. Anderson Cancer Center, Houston, TX 77030-4095, U.S.A.
† E-mail: rex@manderson.org

Contract/grant sponsor: NIH; contract/grant number: CA83932.

Received April 2002
Accepted August 2002

characterized by a binary 'response/no response' variable observed relatively soon after the start of therapy. For example, response may be defined as 50 per cent or greater shrinkage of a solid tumour, resolution of infection, or reduction of post-operative pain below a given threshold. The use of such early events to evaluate treatment activity often is motivated by the belief that it is impractical to wait to evaluate each patient's survival time, and it relies on the implicit assumption that response may be predictive of improved survival. In trials of rapidly fatal diseases, however, it may be practical to use survival time or disease-free survival (DFS) time as the outcome, rather than a binary indicator of response.

In patient-disease subgroups where there is no treatment that provides any substantive beneficial anti-disease effect, it usually is appropriate to conduct a phase IIA, or 'activity' trial of a new agent. Statistical designs for phase IIA trials based on binary response are straightforward, usually consisting of a method for determining maximum sample size and outcome-adaptive rules for stopping the trial early if the observed interim response rate is not promising. Such rules generally are based on comparison of the response probability, π, to a fixed target response probability, π^*, with values in the range $\pi^* = 0.10$ to 0.30 used most commonly. The first phase IIA design was proposed by Gehan [1], who specified a '0 out of n' rule to control the false negative rate in a first stage with n patients and, if the trial is not stopped, criteria for specifying a second stage sample size to obtain a confidence interval for π having given reliability. Thall and Sung [2] proposed a Bayesian phase IIA design that begins by assuming that π follows an uninformative beta prior and stops the trial if the posterior probability that the response rate is at least π^* falls below a fixed lower probability cut-off, p_L. Formally, the Bayesian stopping rule in terms of the random π and fixed π^* is

$$\Pr(\pi > \pi^* \mid \text{data}) < p_L \tag{1}$$

In settings where one or more 'standard' treatment regimens having substantive anti-disease effect are available, it is appropriate to conduct a phase IIB trial, where the question is not whether the experimental treatment is active but whether it provides an improvement over the standard treatments. For example, if a typical standard treatment provides on average a 20 per cent response probability, then a new treatment may be considered promising if there is a non-negligible probability that it will increase this to at least 35 per cent. When survival time or DFS time is the outcome, the phase IIB goal is to determine whether there is any promise of increasing the mean or median event time by a specified value, usually in the range 25 per cent to 75 per cent. Numerous phase IIB designs have been proposed, including frequentist group-sequential methods for binary outcomes [3–5], Bayesian methods for binary outcomes [6–8], and designs accommodating multivariate outcomes using frequentist methods [9–11] and Bayesian methods [2, 12, 13].

This paper is motivated by the problem of designing phase II trials in settings where the disease has multiple subtypes, $S_1, \ldots, S_k$. We will consider the two common cases where patient outcome is either a binary response or a time-to-event variable. In such settings, one possible approach is simply to assume that the parameters, $\theta_1, \ldots, \theta_k$, characterizing the response probabilities or the event rates in the k subtypes, are mutually independent, and conduct the trial using a separate design within each subtype. Because this approach does not share data between subtypes, it fails to accommodate the possibility that, because the subtypes belong to the same disease, knowledge about the agent's effect in one subtype may provide information about its effects in the other subtypes. The opposite approach would be to ignore the subtypes entirely by assuming a common θ. This does not allow for the possibility

that the agent may achieve the desired activity level in some subtypes but not in others. We will take an approach that may be considered a compromise between these two extremes. Specifically, we assume that $\theta_1, \ldots, \theta_k$ follow a Bayesian hierarchical model [14–16] while applying a separate early stopping rule for each subtype. Because the θ_j's are correlated under the hierarchical model, the data from each subtype provide information about all of the θ_j's. For example, a response or a longer survival time for a patient in a given subtype on average increases the posteriors of all the θ_j's, while a subtype with no observed responses or having shorter survival times decreases all of the posteriors.

The remainder of the paper is organized as follows. In Section 2 we briefly describe two motivating examples, one a trial based on a binary response outcome and the other based on DFS time. A general formulation of the hierarchical Bayesian model is given in Section 3. We describe application of the method to a trial of imatinib for sarcoma in Section 4, and to a trial of a new preparative regimen for allogeneic bone marrow transplantation in Section 5. Section 6 describes an extension of the method, in the discrete outcome case, to accommodate both response and toxicity. We close with a discussion on Section 7.

2. MOTIVATING APPLICATIONS

2.1. Imatinib for Sarcoma

Our first example is a trial of imatinib in sarcoma. Although soft tissue and bone sarcomas represent less than 1 per cent of all malignancies, the morbidity is great in that the peak incidence of many sarcomas is seen in children and young adults [17]. Sarcomas are a highly heterogeneous group of tumours that are often classified according to the normal adult tissue that they resemble [18]. For example, fibromas and fibrosarcomas resemble fibrous tissue while haemangioma and angiosarcomas resemble vascular tissues. Thus, there are many subtypes of both soft tissue and bone sarcomas. While the primary treatment of these tumours has improved, with limb-sparing surgery and radiation therapy resulting in improved functional ability, the treatment of metastatic disease is still unsatisfactory and systemic chemotherapy is of limited value [19].

Recently, the success of imatinib, also known as Gleevec or STI-571, in treating gastrointestinal stromal sarcoma has provided proof of principle that patients with a solid tumour may benefit from treatment targeting a specific small molecule [20]. Many sarcoma subtypes overexpress one or more of the tyrosine kinase activated oncogenes affected by imatinib. Since sarcomas are so uncommon and accrual to individual subtypes of sarcomas is often difficult at any single institution, a multi-centre trial was designed to determine the efficacy of daily imatinib for patients with locally incurable or metastatic sarcomas that have failed one or more prior treatment regimens. The sarcoma subtypes in the trial and their anticipated accrual rates are summarized in Table I.

For the purpose of evaluating treatment activity in the course of trial conduct, patient outcome is defined as follows. All evaluations of the extent of the patient's disease are carried out by computerized tomography (CT) or magnetic resonance imaging (MRI). Each patient's disease is evaluated at baseline, at two months after the start of therapy, and possibly at four months, according to the following scheme. At the two- and four-month evaluations, compared to baseline, a complete response (CR) is defined as the complete absence of detectable disease;

Table I. Sarcoma subtypes and anticipated accrual rates.

Sarcoma subtype	Monthly accrual rate
Synovial sarcoma	3.0–6.0
Leiomyosarcoma	3.0–6.0
Malignant fibrous histiocytoma	3.0–6.0
Fibrosarcoma	3.0–6.0
Liposarcoma	3.0–6.0
Ewing's sarcoma	0.5–2.0
Osteosarcoma	0.5–2.0
Rhabdomyosarcoma	0.5–2.0
Peripheral nerve sheath sarcoma	0.5–2.0
Angiosarcoma	0.5–2.0

a partial response (PR) is a reduction in tumour volume between 50 per cent and 100 per cent; stable disease (SD) is the condition that the extent of disease has not changed substantively; and progressive disease (PD) occurs if the extent of disease has increased. Thus, the four outcomes $\{\mathrm{CR}, \mathrm{PR}, \mathrm{SD}, \mathrm{PD}\}$ comprise an ordinal scale. At the two-month evaluation, a CR or PR is scored as a response and PD or death is scored as a failure. Patients with SD at month two are re-evaluated at month four, and at this evaluation SD, PR or CR is scored as a response and PD or death is a failure. While this definition is to some extent arbitrary, it satisfies the requirement that however 'response' is defined, it should characterize anti-disease activity. This is the case here since without treatment a patient is virtually certain to have PD by month four. Given this definition of response, within each sarcoma subtype the goal is to detect a response probability of 0.30 or larger.

2.2. A new preparative regimen for allogeneic BMT

Allogeneic bone marrow transplantation is an effective treatment for haematologic malignancies [21]. Patients initially receive a 'preparative regimen' of high dose chemotherapy with two objectives: (i) to produce immunosuppression in order to prevent rejection of the transplant; (ii) to eradicate the malignancy. The normal bone marrow is also ablated by the preparative regimen, and the bone marrow transplant is given to restore blood cell production and immunity. The preparative regimen also causes toxicity in other tissues. Consequently, there is considerable interest in identifying less toxic, more effective preparative regimens to improve disease-free survival time after bone marrow transplantation.

Our second application is a phase II trial of fludarabine + busulfan as a preparative regimen in allogeneic (donor derived) bone marrow transplantation, hereafter 'allotx'. The eligible patients may be from any of three general disease subgroups: acute myelogenous leukaemia (AML) in remission, AML in relapse, or myelodysplastic syndromes (MDS). The priors for the historical mean DFS times summarized in Table II show that these subgroups have rather different DFS times with standard preparative regimens, such as cyclophosphamide + busulfan. The goal of the trial is to determine whether there is a non-negligible probability of improving the mean DFS by 50 per cent over what is achieved with standard preparative regimens. Thus, the goals in the three subgroups are to detect increases in the mean DFS time from 18 to 27 months for patients with AML in remission, from 5 to 7.5 months for patients with AML in

Table II. Prior mean and 95 per cent credibility interval (CI) of the historical mean disease-free survival times, and monthly accrual rates, in each patient-disease subgroup of the allogeneic bone marrow transplantation trial.

Disease subgroup	Priors		IG parameters		Patients/month
	$E(\lambda_j)$	95% CI	α_j	β_j	
AML in remission	18.1	12.1–27.0	24.0	416.3	1
AML in relapse	5.1	4.0–6.5	66.0	331.0	2
MDS	6.0	4.4–8.2	40.0	234.0	1

relapse, and from 6 to 9 months for patients with MDS. Patients will be accrued and treated for 24 months, with an additional 12 months follow up thereafter.

3. PROBABILITY MODEL

A general formulation of the hierarchical model that includes both the binary and time-to-event outcome cases is as follows. For the k disease subtypes, let $\mathbf{Y}=(\mathbf{Y}_1,\ldots,\mathbf{Y}_k)$ denote the vector of observed data values and $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_k)$ the corresponding parameters of their distributions. Let $\boldsymbol{\phi}$ denote the vector of hyperparameters. We assume that the k observed random variables are conditionally independent given $\boldsymbol{\theta}$

$$\mathbf{Y}\,|\,\boldsymbol{\theta}\sim f(\mathbf{Y}\,|\,\boldsymbol{\theta})=\prod_{j=1}^{k} f(\mathbf{Y}_j\,|\,\theta_j) \tag{2}$$

and that the first level prior parameters are conditionally independent given $\boldsymbol{\phi}$

$$\boldsymbol{\theta}\,|\,\boldsymbol{\phi}\sim f(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})=\prod_{j=1}^{k} f(\theta_j\,|\,\boldsymbol{\phi}) \tag{3}$$

For simplicity, we abuse notation by using $\mathbf{Y}_j$, θ_j and $\boldsymbol{\phi}$ to denote both random quantities and the arguments of their PDFs. The hyperprior $f(\boldsymbol{\phi})$ induces association among the θ_j's since, unconditionally, the prior of $\boldsymbol{\theta}$ is

$$f(\boldsymbol{\theta})=\int \prod_{j=1}^{k} f(\theta_j\,|\,\boldsymbol{\phi})\, f(\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi} \tag{4}$$

This association carries through to the posterior $f(\boldsymbol{\theta}\,|\,\mathbf{Y})$, which will be the basis for decision-making during the trial. The posterior of $\boldsymbol{\theta}\,|\,\mathbf{Y}$ is given generally by

$$f(\boldsymbol{\theta}\,|\,\mathbf{Y})=\frac{\int f(\mathbf{Y},\boldsymbol{\theta},\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\phi}}{\int f(\mathbf{Y},\boldsymbol{\theta}',\boldsymbol{\phi})\,\mathrm{d}\boldsymbol{\theta}'\,\mathrm{d}\boldsymbol{\phi}} \tag{5}$$

where the joint distribution of the data and all the parameters is

$$f(\mathbf{Y},\boldsymbol{\theta},\boldsymbol{\phi})=f(\mathbf{Y}\,|\,\boldsymbol{\theta})\, f(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})\, f(\boldsymbol{\phi})=\left\{\prod_{j=1}^{k} f(\mathbf{Y}_j\,|\,\theta_j)\, f(\theta_j\,|\,\boldsymbol{\phi})\right\} f(\boldsymbol{\phi}) \tag{6}$$

The model given by (2)–(6) has been termed a conditionally independent hierarchical model by Kass and Steffey [15].

4. THE SARCOMA TRIAL

Let π_j be the jth response probability and assume the logistic model $\theta_j = \log\{\pi_j/(1-\pi_j)\}$, for $j = 1, \ldots, k$. We will assume, *a priori*, that θ_j's are i.i.d. normal with mean μ and precision (inverse variance) τ, denoted

$$\theta_1, \ldots, \theta_k \mid \mu, \tau \sim \text{i.i.d. } \mathrm{N}(\mu, \tau^{-1}) \tag{7}$$

so that $\boldsymbol{\phi} = (\mu, \tau)$. Denote the gamma distribution with mean α/β and variance α/β^2 by $\mathrm{G}(\alpha, \beta)$. For the hyperpriors, we assume that

$$\mu \sim \mathrm{N}(-1.386, 10) \quad \text{and} \quad \tau \sim \mathrm{G}(2, 20) \tag{8}$$

Thus, τ has prior mean 0.10 and variance 0.005. The mean of the hyperprior of μ is set equal to the logit of 0.20 to represent the prior belief that the average response rate is between the targeted 0.30 and the uninteresting value 0.10. The numerical hyperprior parameters were elicited as follows. Let X_j denote the number of responders and m_j the number of patients evaluated in S_j at any point in the trial, so that $\mathbf{Y}_j = (X_j, m_j)$. The elicited prior probabilities were $\Pr(\pi_1 > 0.30) = 0.45$, $\Pr(\pi_1 > 0.30 \mid X_1/m_1 = 2/6) = 0.525$, and $\Pr(\pi_1 > 0.30 \mid X_2/m_2 = 2/6) = 0.47$. Thus, *a priori*, while observing 2/6 responses in subtype S_1 would increase $\Pr(\pi_1 > 0.30)$ from 0.45 to 0.525, observing this in another subtype, S_2, would raise this probability from 0.45 to 0.47, that is, about 27 per cent as much.

To compute the posterior (5), we begin with the joint distribution of the data and parameters,

$$f(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \left\{ \prod_{j=1}^{k} \binom{m_j}{X_j} \frac{\mathrm{e}^{\theta_j X_j}}{(1+\mathrm{e}^{\theta_j})^{m_j}} \frac{\mathrm{e}^{-(\theta_j-\mu)^2\tau/2}\, \tau^{1/2}}{(2\pi)^{1/2}} \right\} \frac{\mathrm{e}^{-(\mu+1.386)^2/20}}{(20\pi)^{1/2}}\, 400\, \tau\, \mathrm{e}^{-20\tau} \tag{9}$$

To see how the hyperprior induces correlation among the θ_j's, first denote the k-vector of 1's by $\mathbf{1}_k$ and the $k \times k$ identity matrix by $\mathbf{I}_k$, the $k \times k$ matrix with all entries 1 by $\mathbf{J}_k$, and for convenience denote the mean and variance of the hyperprior of μ by $\tilde{\mu}$ and $\tilde{\sigma}^2$. Since $\mu \sim \mathrm{N}(\tilde{\mu}, \tilde{\sigma}^2)$ and $\boldsymbol{\theta} \mid \mu, \sigma^2$ is k-variate normal with mean vector $\mu \mathbf{1}_k$ and variance-covariance matrix $\sigma^2 \mathbf{I}_k$, which we write $\boldsymbol{\theta} \mid \mu, \sigma^2 \sim \mathrm{MN}(\mu \mathbf{1}_k, \sigma^2 \mathbf{I}_k)$, it follows that $\boldsymbol{\theta} \mid \sigma^2 \sim \mathrm{MN}(\tilde{\mu}\mathbf{1}, \sigma^2 \mathbf{I}_k + \tilde{\sigma}^2 \mathbf{J}_k)$. In particular, this implies that

$$\theta_1 \mid \theta_2,\ \sigma^2 \sim \mathrm{N}\left(\frac{\sigma^2\tilde{\mu} + \tilde{\sigma}^2\theta_2}{\sigma^2 + \tilde{\sigma}^2}, \frac{\sigma^4 + 2\sigma^2\tilde{\sigma}^2}{\sigma^2 + \tilde{\sigma}^2} \right) \tag{10}$$

Thus, *a priori*, the effect of the association between θ_1 and θ_2 on the early stopping criterion in S_1 is quantified by

$$\Pr(\theta_1 > \theta^* \mid \theta_2) = \int_0^\infty \left\{ 1 - \Phi\left(\frac{\theta^* - (\sigma^2\tilde{\mu} + \tilde{\sigma}^2\theta_2)/(\sigma^2 + \tilde{\sigma}^2)}{\{(\sigma^4 + 2\sigma^2\tilde{\sigma}^2)/(\sigma^2 + \tilde{\sigma}^2)\}^{1/2}} \right) \right\} \frac{\mathrm{e}^{-20/\sigma^2}\, 400}{\sigma^6}\, \mathrm{d}\sigma^2 \tag{11}$$

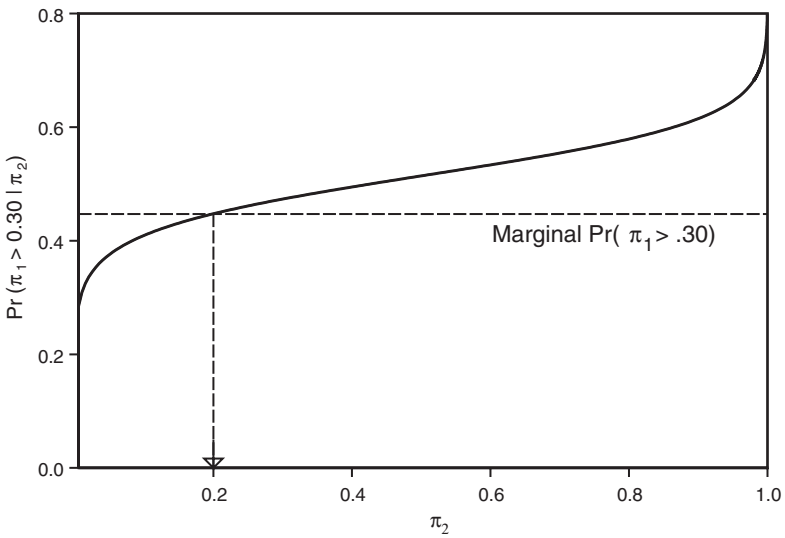

Figure 1. Prior early stopping probability for π_1 as a function of π_2 under the hierarchical model for the sarcoma trial.

Figure 1 illustrates how $\Pr(\pi_1 > 0.30 \mid \pi_2)$ varies as a function of π_2 *a priori*, as an illustration of the correlation among the π_j's under the hierarchical model. The values in the figure may be compared to the marginal probability $\Pr(\pi_1 > 0.30) = \Pr(\theta_1 > \log(0.30/0.70)) = 0.47$. As a basis for comparison, we will also consider the more conventional Bayesian model that assumes the π_j's are independent with identical $\text{beta}(0.20, 0.80)$ priors. Under this model, $\Pr(\pi_1 > 0.30) = 0.26$. The posteriors are independent conjugate betas

$$\pi_j \mid X_j \sim \text{beta}(0.20 + X_j, 0.80 + m_j - X_j), \quad j = 1, \ldots, k \tag{12}$$

Within subtype S_j, a maximum of 30 patients are treated with accrual terminated if

$$\Pr(\pi_j > 0.30 \mid \text{data}) < 0.005 \tag{13}$$

This rule is first applied after a minimum of eight patients in S_j have been evaluated. An essential point is that the 'data' to the right of the conditioning bar in (13) include the data from all ten subtypes, since all of this information affects the posterior of each π_j under the hierarchical model. To avoid treating an unacceptably large number of patients in a subtype having a relatively high accrual rate but a poor response rate, the following rule is applied: during the first three months of the trial, if S_j has accrued 15 patients but not all 15 have been evaluated, then accrual will be suspended temporarily in S_j unless (13) would not be met even if all patients not yet evaluated were to fail. For example, suppose nine subtypes have $0/8$ responses, while the tenth has $1/8$ responses with an additional seven patients treated but not yet evaluated. In this case, accrual would be suspended temporarily in the tenth subtype since $\Pr(\pi_j > 0.30 \mid \text{data}) < 0.005$, where here 'data' is the future outcome that nine subtypes have $0/8$ responses and the tenth has $1/15$. In contrast, if the tenth subtype has $3/8$ responses with seven patients treated but not yet evaluated, then accrual would not be suspended in

Table III. Comparison of some early termination decisions under the conventional and hierarchical Bayesian models in the sarcoma trial.

Case	Outcomes	Decision	
		Conventional model	Hierarchical model
1	5 subtypes with 0/8	Stop	Stop
	5 subtypes with 1/8	Continue	Continue
2	3 subtypes with 0/8	Stop	Continue
	2 subtypes with 1/8	Continue	Continue
	5 subtypes with 2/8	Continue	Continue
3	2 subtypes with 1/17	Stop	Continue
	3 subtypes with 5/17	Continue	Continue
	5 subtypes with 7/23	Continue	Continue
4	3 subtypes with 0/8	Stop	Stop
	2 subtypes with 1/8	Continue	Stop
	5 subtypes with 2/23	Stop	Stop
5	3 subtypes with 1/8	Continue	Continue
	2 subtypes with 2/22	Continue	Stop
	5 subtypes with 3/30	Stop	Stop

the tenth subtype. This is because $\Pr(\pi_j > 0.30 \mid \text{data}) = 0.09 > 0.005$, where now 'data' is the future outcome that nine subtypes have $0/8$ responses and the tenth has $3/15$.

Under this hierarchical model, the joint posterior distribution is not readily available in closed form and, due to the high dimension of the posterior, numerical integration is not practical. Thus, we used Markov chain Monte Carlo (MCMC) to compute the posteriors [22]. Straightforward but tedious computations show that the full conditional distributions of some model parameters can be obtained in closed form, and sampling from these is straightforward in the MCMC. For the conditionals not available in closed form, since these full conditionals are log-concave we used adaptive rejection sampling (ARS) [23]. To evaluate the algorithm's performance we generated many data sets, ran parallel chains for each with different starting points, and evaluated the convergence of the chains using the software CODA [24]. It was determined that a short burn-in and a chain of length 1000 met reasonable convergence criterion. However, since the stopping probability criterion in (13) must be compared to the small cut-off value 0.005, a burn-in of 1000 with a chain of length 5000 was used when simulating the trial. In addition, if the probability in (13) was determined to be less than 0.20 with a chain of length 5000, then an additional 5000 samples were drawn to obtain a more precise estimate of the criterion probability.

The manner in which the data in the different subtypes affect the early stopping decisions through the hyperprior is illustrated by the examples in Table III, which also includes the corresponding decisions based on the conventional independent beta priors. Comparison of case 2 to case 1 shows that an apparently small improvement in five of the ten subtypes leads to the decision to continue the subtypes having $0/8$ responses, rather than terminating them. Case 3 illustrates a similar effect later in the trial, where the conventional model would

Table IV. Operating characteristics of the sarcoma trial design. Subtypes 1–5 accrue patients quickly (3 to 6 per month), subtypes 6–10 accrue patients slowly (0.5 to 2 per month). Tabled values are the mean per subtype. Standard deviations are given in parentheses.

Scenario 1	Sarcoma subtype			
	1–5		6–10	
True Pr(response)	0.10		0.10	
Number of patients	23.2 (5.7)		12.6 (2.5)	
Number of months	5.4 (1.5)		10.5 (1.9)	
Per cent stopped early	77		47	
Scenario 2	1	2–5	6–10	
True Pr(response)	0.30	0.10	0.10	
Number of patients	29.2 (2.9)	23.2 (5.6)	12.8 (2.3)	
Number of months	7.1 (1.0)	5.4 (1.5)	10.7 (1.7)	
Per cent stopped early	9.5	76	45	
Scenario 3	1–2	3–5	6–10	
True Pr(response)	0.30	0.10	0.10	
Number of patients	29.3 (2.7)	23.9 (5.5)	13.1 (2.1)	
Number of months	7.1 (0.9)	5.6 (1.5)	10.9 (1.5)	
Per cent stopped early	8.0	73	42	
Scenario 4	1–5	6	7–10	
True Pr(response)	0.10	0.30	0.10	
Number of patients	23.4 (5.6)	14.2 (1.8)	12.9 (2.3)	
Number of months	5.5 (1.4)	11.8 (0.9)	10.7 (1.7)	
Per cent stopped early	75	7	45	
Scenario 5	1–5	6–7	8–10	
True Pr(response)	0.10	0.30	0.10	
Number of patients	23.2 (5.6)	14.3 (1.8)	13.0 (2.2)	
Number of months	5.5 (1.4)	11.8 (0.8)	10.8 (1.6)	
Per cent stopped early	75	6.0	44	
Scenario 6	1	2–5	6	7–10
True Pr(response)	0.30	0.10	0.30	0.10
Number of patients	29.2 (2.8)	23.5 (5.6)	14.3 (1.7)	13.1 (2.1)
Number of months	7.1 (1.0)	5.5 (1.5)	11.9 (0.7)	10.9 (1.5)
Per cent stopped early	8.5	73	5.0	43

terminate the subtypes with 1/17 responses but the hierarchical model continues them based on the positive results in the other subtypes. Cases 4 and 5 show that this effect also works in the opposite direction, with subtypes that would have been continued under the conventional model terminated under the hierarchical model.

Table IV summarizes a simulation study of the design. Under each scenario in the study, the responses were simulated with probability fixed at either 0.10 or 0.30 within each sarcoma subtype. A distinction is made between the quickly accruing (QA) subtypes, numbered 1–5,

Table V. Effect of the number of quickly accruing subtypes with response probabilities 0.30 versus 0.10 on the early stopping rates in the sarcoma trial. All five slowly accruing subtypes have response probability 0.10. Each entry is the mean percentage of early stops per subtype.

	Number of subtypes with ($p = 0.30$, $p = 0.10$)					
	(0, 5)	(1, 4)	(2, 3)	(3, 2)	(4, 1)	(5, 0)
$p = 0.30$	—	9.6	6.6	6.4	5.7	4.3
$p = 0.10$	78	75	72	70	67	—

and the slowly accruing (SA) subtypes, numbered 6–10. In general, the SA subtypes have lower early stopping probabilities because there is less information per unit time compared to the QA subtypes. Under scenario 1, all ten subtypes have the unpromising response rate of 0.10. As a basis for comparison, if the hierarchical model were not assumed, and the θ_j's were independent but with the same prior means and variances, then the early stopping probability would be 0.67 for each of subtypes 1–5 and 0.42 for each of subtypes 6–10. The higher values 0.77 and 0.47 under the hierarchical model are an advantage of borrowing strength. Scenarios 2 and 3 show that, while one QA subtype with desirable response probability 0.30 has a 9.5 per cent false negative rate, if two have response probability 0.30 then each has false negative rate 8.0 per cent due to the fact that they borrow strength. Scenarios 4 and 5 show the analogous effect in the SA subtypes, and scenario 6 shows this effect when one QA and one SA subtype have response probability 0.30. Table V gives a more complete picture of how the hierarchical model causes the false negative rate of each subtype to drop as the number of QA subtypes with true success probability 0.30 increases from 1 to 5.

5. THE ALLOGENEIC TRANSPLANTATION TRIAL

Let $X_{j,i}$ denote the event time of the ith of m_j patients in S_j. The time-to-event case is not strictly analogous to the binary outcome case in that the null, historical mean event time in a patient group cannot equal 0, but rather it must be positive real-valued. We denote the historical mean event times in the k subgroups by $\lambda_1, \ldots, \lambda_k$ and assume that $E(X_{j,i} \mid \boldsymbol{\theta}_j) = \lambda_j \, e^{\rho_j}$ for real-valued treatment effects $\rho_1, \ldots, \rho_k$, denoting $\boldsymbol{\theta}_j = (\lambda_j, \rho_j)$. We assume that the event times are exponential, formally

$$X_{j,1}, \ldots, X_{j,m_j} \mid \boldsymbol{\theta}_j \sim \text{i.i.d. } \mathrm{G}(\lambda_j \, e^{\rho_j}, 1), \quad j = 1, \ldots, k \tag{14}$$

To accommodate the different historical event rates of the patient-disease subgroups, we assume that $\lambda_1, \ldots, \lambda_k$ are independent but not identically distributed, while the treatment effects are i.i.d. given the hyperparameters. Formally, we assume that

$$f(\boldsymbol{\lambda}, \boldsymbol{\rho} \mid \boldsymbol{\phi}) = f(\boldsymbol{\lambda}) f(\boldsymbol{\rho} \mid \boldsymbol{\phi}) = \prod_{j=1}^{k} f_j(\lambda_j) f(\rho_j \mid \boldsymbol{\phi}) \tag{15}$$

Denote the inverse gamma distribution by $\mathrm{IG}(\alpha, \beta)$, where we define $W \sim \mathrm{IG}(\alpha, \beta)$ if and only if $1/W \sim \mathrm{G}(\alpha, \beta)$. The priors are given by

$$\lambda_j \sim \mathrm{IG}(\alpha_j, \beta_j), \quad j = 1, \ldots, k \quad \text{and} \quad \rho_1, \ldots, \rho_k \mid \mu, \sigma^2 \sim \text{i.i.d. } \mathrm{N}(\mu, \sigma^2) \tag{16}$$

where now $\boldsymbol{\phi}=(\mu,\sigma^2)$. For the hyperpriors, we assume that $\mu\sim \mathrm{N}(\tilde{\mu},\tilde{\sigma}^2)$ and $\sigma^2\sim \mathrm{IG}(\tilde{\alpha},\tilde{\beta})$. Denote the time to the event or censoring for patient i in S_j by $X^0_{j,i}$, let $\delta_{j,i}$ be the indicator that $X^0_{j,i}=X_{j,i}$, so that $\mathbf{Y}_j=\{(X^0_{j,i},\ \delta_{j,i}),\ i=1,\ldots,m_j\}$. The total time to failure or censoring is $X^0_{j,+}=X^0_{j,1}+\cdots+X^0_{j,m_j}$, the number of events out of the m_j patients is $\delta_{j,+}=\delta_{j,1}+\cdots+\delta_{j,m_j}$, and the likelihood of $\mathbf{Y}_j\mid\boldsymbol{\theta}_j$ is

$$f(X_{j,1},\delta_{j,1},\ldots,X_{j,m_j},\delta_{j,m_j}\mid\lambda_j,\rho_j)=(\lambda_j^{-1}\,\mathrm{e}^{-\rho_j})^{\delta_{j,+}}\,\mathrm{e}^{-\lambda_j^{-1}\mathrm{e}^{-\rho_j}X^0_{j,+}} \tag{17}$$

Thus, the joint distribution of all data and parameters takes the form

$$f(\mathbf{Y},\boldsymbol{\delta},\boldsymbol{\theta},\boldsymbol{\phi})=\left\{\prod_{j=1}^{k}(\lambda_j^{-1}\,\mathrm{e}^{-\rho_j})^{\delta_{j,+}}\mathrm{e}^{-\lambda_j^{-1}\mathrm{e}^{-\rho_j}X^0_{j,+}}\frac{\lambda_j^{-\alpha-1}\beta^{\alpha}\mathrm{e}^{-\beta/\lambda_j}}{\Gamma(\alpha)}\,\frac{\mathrm{e}^{-(\rho_j-\mu)^2/2\sigma^2}}{(2\pi)^{1/2}\sigma}\right\}$$
$$\times\frac{\mathrm{e}^{-(\mu-\tilde{\mu})^2/2\tilde{\sigma}^2}}{(2\pi)^{1/2}(\tilde{\sigma}^2)^{1/2}}\,\frac{\tilde{\beta}^{\tilde{\alpha}}\,\mathrm{e}^{-\tilde{\beta}/\sigma^2}}{(\sigma^2)^{\tilde{\alpha}+1}\Gamma(\tilde{\alpha})} \tag{18}$$

Table II summarizes the marginal prior distributions of the λ_j's in terms of their prior means and 95 per cent credibility intervals, along with the corresponding inverse gamma parameters, for each disease subgroup. The hyperprior parameters reflect the prior belief that the new preparative regimen should provide a 25 per cent increase in the mean DFS in each subgroup, formally $E(\mathrm{e}^{\rho_j})=1.25$, with 95 per cent credibility interval 0.65 to 2.20 for e^{ρ_j}. To determine the prior association among the ρ_j's, the hyperprior was calibrated so that $E(\mathrm{e}^{\rho_2}\mid\mathrm{e}^{\rho_1}=1.50)=1.275$. That is, given that there is a 50 per cent improvement in subgroup 1, the expected improvement in subgroup 2 is increased by $1/10$ of the difference between the prior mean 1.25 and 1.50. This yielded the parameters of the hyperpriors $f(\mu)$ and $f(\sigma^2)$.

The early stopping rule in patient-disease subgroup S_j is

$$\Pr(\mathrm{e}^{\rho_j}>1.50\mid\mathrm{data})<0.075 \tag{19}$$

The improvements $\{\mathrm{e}^{\rho_j},\ j=1,2,3\}$ in average DFS time over the historical means in the patient-disease subgroups in the allotx trial thus play a role that is analogous to that of the ten response probabilities in the sarcoma trial. Conduct of the allotx trial differs from that of the sarcoma trial in that the allotx patients are accrued, treated and their outcomes are monitored continuously without any reference to specific time intervals or surrogate events. The event time data, including the times of relapse, death or current follow-up (censoring), are monitored and the decision criteria (19) are updated continuously over the 24 month accrual period of the trial. Computation of the posterior early stopping criteria (19) under the hierarchical time-to-event model was done similarly to that under the model for binary outcomes in the sarcoma trial.

Figure 2 illustrates how $\Pr(\mathrm{e}^{\rho_1}>1.50\mid\rho_2)$ varies with ρ_2 *a priori* under the hierarchical model. The values in the figure may be compared to the marginal prior probability $\Pr(\mathrm{e}^{\rho_1}>1.50)=0.447$. The effect of the data from patient-disease subgroups other than S_1 on the decision probability $\Pr\{\mathrm{e}^{\rho_1}>1.50\mid\mathrm{data}\}$ is illustrated by the hypothetical examples in Table VI, which is analogous to Table III. The 'conventional' method in Table VI was obtained by applying the stopping rules (19) independently in the three disease subtypes based on three independent priors. For comparability, these priors were calibrated to have the same

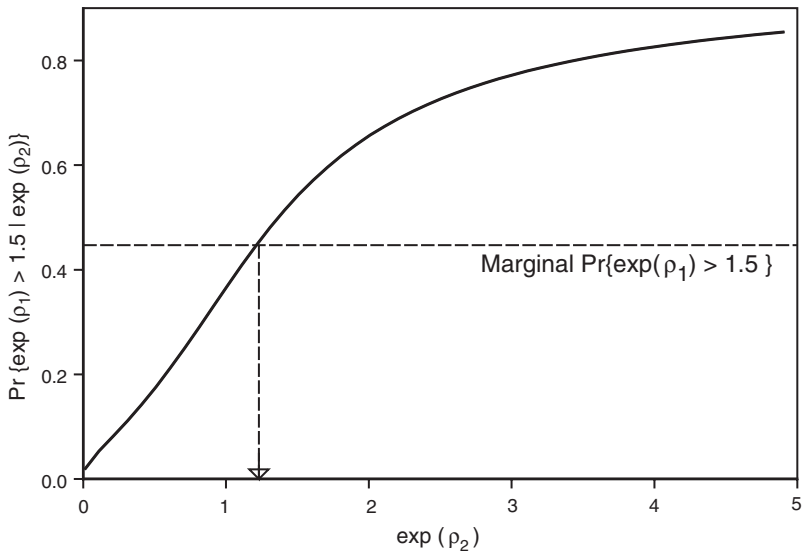

Figure 2. Prior early stopping probability for $\exp(\rho_1)$ as a function of $\exp(\rho_2)$ under the hierarchical model for the bone marrow transplantation trial.

Table VI. Comparison of some early termination decisions under the conventional and hierarchical Bayesian models in the allogeneic transplantation trial. Within each subgroup, $N =$ number of patients, $T_+ =$ total time on test, in months, and $F =$ number of failures.

Case		AML in remission	AML in relapse	Myelodysplastic syndromes
1	$(N,\ T_+,\ F)$	(10, 54, 2)	(20, 88, 12)	(10, 54, 6)
	Estimated mean DFS	27	7.3	9.0
	Conventional	Continue	Continue	Continue
	Hierarchical	Continue	Continue	Continue
2	$(N,\ T_+,\ F)$	(10, 54, 2)	(20, 70, 14)	(10, 54, 6)
	Estimated mean DFS	27	5.0	9.0
	Conventional	Continue	Stop	Continue
	Hierarchical	Continue	Continue	Continue
3	$(N,\ T_+,\ F)$	(10, 48, 5)	(20, 66, 16)	(10, 40, 8)
	Estimated mean DFS	9.6	4.2	5.0
	Conventional	Stop	Stop	Continue
	Hierarchical	Stop	Stop	Stop

means and variances as the corresponding marginals of $\lambda_j e_j^{\rho},\ j = 1, 2, 3$ under the hierarchical model. In Case 1, the empirical DFS in each group is consistent with a 50 per cent increase over the historical rate in each subgroup, and the two methods agree to continue in all three subgroups. Case 2 differs from case 1 in that the empirical DFS time in the second subgroup is much smaller. Here, the conventional method would stop accrual in the second subgroup

Table VII. Operating characteristics of the allogeneic transplantation trial. For each scenario, the fixed mean DFS in each patient-disease subgroup is $\exp(\rho^{F})$ times the historical subgroup mean.

Scenario		Patient–disease subgroup		
		AML in remission	AML in relapse	Myelodysplastic syndromes
1	$\exp(\rho^{F})$	1.0	1.0	1.0
	Number of patients	19.9(6.2)	24.4(11.8)	16.2(7.1)
	Per cent stopped early	58(1.6)	89(1.0)	71(1.4)
2	$\exp(\rho^{F})$	1.5	1.0	1.0
	Number of patients	23.1(6.1)	25.2(11.4)	17.0(7.2)
	Per cent stopped early	18(1.2)	83(1.2)	62(1.5)
3	$\exp(\rho^{F})$	1.0	1.5	1.0
	Number of patients	21.0(5.9)	43.4(12.3)	18.1(6.8)
	Per cent stopped early	40(1.5)	19(1.2)	42(1.6)
4	$\exp(\rho^{F})$	1.0	1.0	1.5
	Number of patients	21.2(5.9)	26.1(11.9)	22.4(6.6)
	Per cent stopped early	42(1.6)	84(1.2)	18(1.2)
5	$\exp(\rho^{F})$	1.5	1.5	1.0
	Number of patients	23.6(5.2)	44.0(11.4)	18.5(6.7)
	Per cent stopped early	7(0.8)	14(1.1)	51(1.6)
6	$\exp(\rho^{F})$	1.5	1.0	1.5
	Number of patients	23.3(5.3)	26.9(12.6)	22.5(6.2)
	Per cent stopped early	8(0.9)	80(1.3)	13(1.1)
7	$\exp(\rho^{F})$	1.0	1.5	1.5
	Number of patients	21.9(5.8)	43.8(11.5)	23.6(5.7)
	Per cent stopped early	27(1.4)	14(1.1)	7(0.8)
8	$\exp(\rho^{F})$	1.5	1.5	1.5
	Number of patients	23.9(5.1)	45.0(10.9)	23.3(5.5)
	Per cent stopped early	4(0.6)	11(1.0)	6(0.8)

whereas the hierarchical model-based method would continue accrual in that subgroup, essentially because it takes the higher observed rates in the other subgroups into account. Case 3 illustrates the opposite effect, in that the poor results in all three subgroups combine to cause accrual to be terminated in subgroup 3 rather than continued.

To simulate the allotx trial, we assumed that the mean DFS time within the jth subgroup was the fixed value $\lambda_j^{F}\exp(\rho_j^{F})$ where $\lambda_1^{F}=18.1$, $\lambda_2^{F}=5.1$, and $\lambda_3^{F}=6$, the historical means, and each of the effects $\{\exp(\rho_j^{F}),\ j=1,2,3\}$ was set equal to either 1 or 1.50, depending on the particular scenario. Patients within each subgroup were simulated to arrive according to a Poisson process having rate equal to the anticipated accrual rates given in Table II, and their times to failure were simulated as i.i.d. exponential random variables with the above fixed means. The simulation results are summarized in Table VII. When interpreting these results, it is important to bear in mind that the fixed mean DFS times in the three subgroups are

18.1, 5.1 and 6.0 months, and that the corresponding monthly accrual rates are 1, 2 and 1 (Table II). Scenarios 2, 3 and 4 show that if the 50 per cent improvement is the case in only one of the three subgroups then there is about an 18 per cent chance of wrongly terminating that subgroup. Scenarios 5, 6 and 7 show that if the improvement is obtained in two of the three subgroups then they borrow strength to greatly reduce the false negative rates. This reduction is even greater if all three subgroups have 50 per cent improvement (scenario 8).

6. ACCOUNTING FOR TOXICITY AND RESPONSE

The model and method used for the imatinib–sarcoma trial may be extended to accommodate toxicity as well as response in settings where both of these events are important. To show how this may be done, we first describe a phase IIA trial of intra-prostatic PS-341 for prostate cancer patients who have relapsed after external beam radiation therapy. Response was defined as a >50 per cent reduction in PSA or a >25 per cent reduction in tumour mass as measured by trans-rectal ultrasound, or both. Toxicity was defined as any grade 3 or 4 non-haematologic local or systemic toxicity. Denoting the probabilities of the four possible outcomes (Response, Toxicity) = (Yes, Yes), (Yes, No), (No, Yes) and (No, No) by $\theta_1, \theta_2, \theta_3$, and $\theta_4 = 1-\theta_1-\theta_2-\theta_3$, it was assumed that $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ followed a Dirichlet prior with parameters $\mathbf{a} = (0.08, 0.72, 0.32, 2.88)$, denoted $\boldsymbol{\theta} \sim \text{Dir}(\mathbf{a})$. This reflected the prior belief that the probabilities of response, $\theta_R = \theta_1 + \theta_2$, and toxicity, $\theta_T = \theta_1 + \theta_3$, would on average equal 0.20 and 0.10, but with a high degree of uncertainty since $\sum_{j=1}^{4} a_j = 4$. Assuming that, given $\boldsymbol{\theta}$, the four-category outcome count vector $\mathbf{X} = (X_1, X_2, X_3, X_4)$ is multinomial in $\boldsymbol{\theta}$ and $X_+ = \sum_{i=1}^{4} X_i$, the design specified that the trial be stopped early if

$$\Pr(\theta_R > 0.20 \mid \text{data}) < 0.025 \quad \text{or} \quad \Pr(\theta_T > 0.10 \mid \text{data}) > 0.925 \tag{20}$$

This model and composite early stopping rule are of the family proposed by Thall *et al.* [12], but using the fixed limits 0.20 and 0.10 rather than random parameters.

We now extend this design to accommodate k patient-disease subtypes, as in the previous examples. Let $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3, \phi_4)$ be independent hyperparameters corresponding to the four response–toxicity outcomes, with $\phi_i \sim \text{G}(\alpha_i, \beta)$ for $i = 1, 2, 3, 4$, and denote $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ be the k four-outcome probability vectors corresponding to the patient–disease subtypes. For the first-level prior, assume that $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k \mid \boldsymbol{\phi} \sim$ i.i.d. $\text{Dir}(\boldsymbol{\phi})$ and that, given $\boldsymbol{\theta}_j$, the four-outcome count vector $\mathbf{X}_j = (X_{j,1}, X_{j,2}, X_{j,3}, X_{j,4})$ observed in subtype S_j is multinomial with parameters $\boldsymbol{\theta}_j$ and $X_{j,+} = \sum_{i=1}^{4} X_{j,i}$. Since the four independent gamma hyperpriors have the same second parameter, β, it follows that $\boldsymbol{\phi}/\phi_+ \sim \text{Dir}(\boldsymbol{\alpha})$, and consequently $E(\theta_j) = E\{E(\theta_j \mid \boldsymbol{\phi})\} = E\{\boldsymbol{\phi}/\phi_+\} = \boldsymbol{\alpha}/\alpha_+$. Thus, to apply this model one must elicit the gamma hyperpriors of the α_i's and β so that $\boldsymbol{\alpha}/\alpha_+$ equals the physician's prior mean of $\boldsymbol{\theta}$, such as $\mathbf{a} = (0.08, 0.72, 0.32, 2.88)$, above, and so that the induced correlation among the θ_j's appropriately reflects the physician's prior belief regarding association among the subtypes. The two early stopping rules (20) would then be applied for each pair $(\theta_{R,j}, \theta_{T,j})$ within its patient-disease subgroup S_j, for $j = 1, \ldots, k$, with the important provision that the posterior of each $(\theta_{R,j}, \theta_{T,j})$ be computed based on the data from all k subtypes, so that the k pairs of rules may borrow strength from each other via the hierarchical model. This multinomial-Dirichlet-gamma hierarchical model may be applied in similar settings involving an arbitrary number

of elementary events, although from our experience the great majority of applications would have outcomes with at most five categories.

7. DISCUSSION

We have described a general method for constructing early stopping rules in a trial of a new agent or treatment combination that accounts for multiple disease subtypes. Our two applications, for trials based on a binary response and on survival time, illustrate both the method's generality and its range of practical application.

Conducting a multi-centre trial that requires repeated application of safety monitoring rules in multiple disease subtypes is not logistically straightforward. We thus developed a web-based interface for conducting the sarcoma trial. The interface provides the participating institutions with a facility for enrolling patients and viewing current data, and it also is the basis for efficient safety monitoring. At this writing, the first two monthly evaluations in the sarcoma trial have been carried out with no interruption of patient accrual.

A major issue in designing each trial was the amount of time required to conduct the computer simulations. For each simulated trial under each scenario, the appropriate subgroup stopping rule must be checked before each patient is enrolled. Evaluation of the stopping rules is expensive in terms of computing time, since it involves MCMC. For example, based on initial runs, simulating the transplantation trial under each of the eight scenarios that we considered would have required well over six weeks of continuous processing on one dual Pentium (r) 41.8GHz computer. However, by making a few small modifications to the original simulation program we were able to use a distributed processing system that utilizes all of the idle computers in our environment, including 25 PCs of varying speeds. For each of the two applications, this allowed us to complete the simulations in less than two days. In general, we have found that simulation studies that would be impractically long on a single computer can be completed in our environment with distributed processing in at most a few days. The total run time of a given simulation study carried out using distributed processing varies with the statistical model and method, the number and nature of the scenarios, and the type and availability of the PCs.

The main logistical problem in conducting the imatinib–sarcoma trial is that it involves ten institutions. In the initial discussions of how this trial would be conducted, it was proposed that each institution would send the data to the second author of this paper. To facilitate this process, a website was constructed for data collection and real time evaluation of the stopping rules. While the latter currently is being done manually due to the simplicity of the data structure, we are developing a version of the program that will do this automatically. Implementating the transplantation trial design requires a user interface, however, because time-to-event variables must be monitored. These interfaces look like patient logs that communicate with the user, usually a research nurse or physician, by eliciting data and showing which subgroups have been terminated or continued. The main difficulty in writing such interfaces, which we have used at M.D. Anderson in recent years for various adaptive clinical trial designs, is linking the program that does the statistical computing to the program that interacts with the user. Appendices A and B contain WinBUGS programs for computing the stopping probability criteria used to conduct each of the two types of trials. Computer programs for simulating each type of trial are available from the second author on request.

APPENDIX A: WinBUGS CODE FOR TRIAL CONDUCT WITH BINARY OUTCOMES

```
model
{
   for (i in 1:numGroups)
   {

# numGroups is k, the number of different probabilities
      x[i] ~ dbin(p[i],n[i]);
# In each group, x is the number of responses and n is the number of patients
      logit(p[i]) <- rho[i];
      rho[i] ~ dnorm(mu,tau)
      pg[i] <- step(p[i] - targetResp)

# Probability that the response rate for each group is > than targetResp,
which is fixed response probability π*
      }

#Priors
      mu ~ dnorm(mean.Mu, perc.Mu)
      tau ~ dgamma(tau.alpha, tau.beta)
}

# Example data
list(x = c(0,0,1,3,5,0,1,2,0,0), n = c(0,2,1,7,5,0,2,3,1,0), numGroups = 10,
targetResp = 0.30,
mean.Mu = -1.3863, perc.Mu = .10, tau.alpha = 2, tau.beta = 20)

# Example of initial values
list(mu = 1, tau = .10)
```

APPENDIX B: WinBUGS CODE FOR TRIAL CONDUCT WITH TIME-TO-EVENT OUTCOMES

```
model
{
   #Loop over groups
   for (j in 1:numGroups)
   {
      lamdainv[j] ~ dgamma(alpha[j],beta[j])
      lamda[j] <- 1/lamdainv[j]
      rho[j] ~ dnorm(mu,tau)
      erho[j] <- exp(rho[j])
```



```
    invtheta[j] <- 1/theta[j]
    theta[j] <- exp(rho[j] / lamdainv[j])
    erhoGr[j]<-step(erho[j]- target)
    # mean(erhoGr[i]) contains Pr(e^rho > target | data)
    }
  # Loop for data
  for(i in 1:numPats)
  {
    x[i] ~ dexp(invtheta[group[i]])I(censvec[i],)
  }
  mu ~ dnorm(mu.mean, mu.pre)
  tau ~ dgamma(tau.alpha, tau.beta)
}

# Example data

list(numPats = 30, numGroups = 3, target = 1.5
x = c(NA, 15.1, 16.2, NA, 18, 12, 18.8, 18, 5.8, 7.8, 7, 9, 3, 10, NA, NA,
15, 8.6, 3.2, NA, NA, 8.8, 9.1, 5, 3, 4.6, 9, 4, 3, NA),
censvec = c(14.2, 0, 0, 13.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, NA, 0, 0, 0,
11.1, 13.2, 0, 0, 0, 0, 0, 0, 0, 0, 12),
group = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3),
alpha = c(24, 66, 40), beta = c(416.3, 331, 234), mu = 1, tau = .1,
mu.mean = .1856, mu.pre = .5, tau.alpha = 3, tau.beta = 3)

# Example initial values

list(lamdainv = c(18, 6, 5), rho = c(1, 1, 1)))

# Parameter Explanation
```

In this WinBUGS code, x is the array of observed event times for all patients. The length of this array equals the total number of patients (numPats) who have currently been observed. If an event has not been observed for the jth patient (that is, that patient is censored) then enter an NA in the jth position of x. The array of censored times is denoted by censvec. If the jth patient is censored then place the censoring time in the jth position of censvec, otherwise place a 0 in that position. The group array indicates the group to which a given patient belongs. To illustrate this data structure, we consider the first two patients in our example data. The first patient was censored at time 14.2 and belongs to group 1, thus the first position in the x array is NA, the first position in the censvec array is 14.2 and the first position in the group array is 1. The second patient had an event at time 15.1 and also belongs to group 1, thus the x array has a 15.1 in the second position, the censvec has 0 in the second position and the group array has a 1 in the second position. The connection between arguments in the above code and the model parameters in Section 5 is as follows: (mu, tau, mu.mean, mu.pre, alpha, beta, alpha[j], beta[j]) correspond to $(\mu, 1/\sigma^2, \tilde{\mu}, 1/\tilde{\sigma}^2, \tilde{\alpha}, \tilde{\beta}, \alpha_j, \beta_j)$, respectively.

ACKNOWLEDGEMENTS

The first author's work was partially supported by NIH grant CA 83932. We thank two referees for their thoughtful and constructive comments.

REFERENCES

1. Gehan EA. The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* 1961; **13**:346–353.
2. Thall PF, Sung H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine* 1998; **17**:1563–1580.
3. Fleming TR. One sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**:143–151.
4. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**:1–10.
5. Chang M, Therneau T, Wieand HS. Designs for group sequential phase II clinical trials. *Biometrics* 1987; **43**:865–874.
6. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50**:337–349.
7. Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine* 1995; **16**:1791–1802.
8. Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* 1998; **54**:279–294.
9. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 1995; **51**:1372–1383.
10. Conaway MR, Petroni GR. Bivariate sequential designs for phase II trials. *Biometrics* 1995; **51**:656–664.
11. Zee B, Melnychuk D, Dancey J, Eisenhauer E. Multinomial phase II cancer trials incorporating response and early progression. *Journal of Biopharmaceutical Statistics* 1999; **9**:351–363.
12. Thall PF, Simon R, Estey E. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 1995; **14**:357–379.
13. Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* 1999; **55**:971–977.
14. Lindley DV, Smith AFM. Bayesian estimates for the linear model (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:1–41.
15. Kass RE, Steffey D. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* 1989; **84**:717–726.
16. Skene AM, Wakefield JC. Hierarchical models for multicentre binary response studies. *Statistics in Medicine* 1990; **9**:919–929.
17. Wingo PA, Tong T, Bolden S. Cancer statistics, 1995. *CA J. Clin.* 1995; **45**:8–30.
18. Enzinger FM, Weiss SW. *General Considerations*, *Soft Tissue Tumors*. Mosby: St. Louis, 1995; 1–16.
19. Patel SR, Benjamin RS. Sarcomas of soft tissue and bone. In *Harrison's Principles of Internal Medicine*, 14th edn, Fauci AS, Braunwald E, Isselbacher KJ *et al.* (eds). McGraw Hill: New York, 1997; 611–614.
20. Blanke CD, von Mehren M, Joensuu H *et al.* Evaluation of the safety and efficacy of an oral molecularly targeted therapy, STI571, in patients with unresectable or metastatic gastrointestinal stromal tumours expressing C-KIT. *Proceedings of the American Society of Clinical Oncology* 2001; **20**:1a.
21. Thomas ED. Marrow transplantation for malignant diseases. *Journal of Clinical Oncology* 1983; **1**:517–531.
22. Gilks W, Richardson S, Spiegelhalter D. *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London, 1996.
23. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 1992; **41**:337–348.
24. Best N, Cowles MK, Vines K. CODA—Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output. Medical Research Council, Biostatistics Unit, Cambridge, 1995.

