CLINICAL TRIALS ARTICLE Clinical Trials 2005; 2: 467–478

Monitoring event times in early phase clinical trials: some practical issues

Peter F Thall[a], Leiko H Wooten[a] and Nizar M Tannir[b]

Background In many early phase clinical trials it is scientifically inappropriate or logistically infeasible to characterize patient outcome as a binary variable. In such settings, it often is more natural to construct early stopping rules based on time-to-event variables. This type of design may involve a variety of complications, however.
Purpose The purpose of this paper is to illustrate by example how one may deal with various complications that may arise when monitoring time-to-event outcomes in an early phase clinical trial.
Methods We present a series of Bayesian designs for a phase II clinical trial in kidney cancer. Each design includes a procedure for monitoring the times to a severe adverse event, disease progression and death. The first design, which is the simplest, is based on the time to failure, defined as any of the three events, assuming exponentially distributed failure times with an inverse gamma prior on the mean. This design is compared by simulation to the CMAP design (Cheung and Thall, *Biometrics*, 2002; **58**: 89–97). The model and monitoring procedure are then extended successively to accommodate several common practical complications, and we also study the method's robustness.
Results Our simulations show that 1) one may apply the monitoring rule periodically, rather than continuously, without a substantive degradation of the design's reliability; 2) it is very important to account for interval censoring due to periodic evaluation of disease status; 3) it is important to account for the effect of disease progression on the subsequent death rate; 4) conducting a randomized trial presents little additional difficulty and provides unbiased comparisons; and 5) the exponential-inverse gamma model is surprisingly robust in most cases.
Limitations The methods discussed here do not account for patient heterogeneity. This is an important but complex issue that may be dealt with by extending the models and methods given here to accommodate patient covariates and treatment-covariate interaction.
Conclusions Bayesian procedures for monitoring time-to-event outcomes offer a practical way to conduct a variety of early phase trials. Considerable care must be given, however, to modeling the important aspects of the trial at hand, and to calibrating the prior and the design parameters to ensure that the design will have good operating characteristics. *Clinical Trials* 2005; **2**: 467–478. www.SCTjournal.com

Introduction

In many clinical trials the primary therapeutic outcomes are events that occur at random times. Examples of such events include a given amount of tumor shrinkage, disease progression, or regimen-related death in oncology; engraftment or graft-versus-host disease in bone marrow transplantation;

[a]Department of Biostatistics and Applied Mathematics, The University of Texas, M.D. Anderson Cancer Center, Houston, TX, USA, [b]Department of Genitourinary Medical Oncology, The University of Texas, M.D. Anderson Cancer Center, Houston, TX, USA
Author for correspondence: Peter F Thall, Department of Biostatistics and Applied Mathematics, The University of Texas, M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA. E-mail: rex@mdanderson.org

© Society for Clinical Trials 2005 10.1191/1740774505cn121oa

resolution of an infection with an antibiotic; and lowering systolic blood pressure by a specified amount when studying an anti-hypertension agent. Many statistical designs for early phase trials are based on the probability of a composite outcome, defined in terms of one or more event times, occurring within a specified time period from the start of treatment. For example, in oncology "response" may be defined as the event that the patient survives seven months without suffering disease progression. Denoting the times to progression and death by T_P and T_D, response is $R = \{7 < \min(T_P, T_D)\}$. However, if one wishes to make interim decisions based on the indicator X of R and the probability $\pi_R = \Pr(R) = \Pr(X = 1)$, then several logistical problems arise. The most severe problem created by this approach is that a failure (progression or death) may be observed at any time up to seven months, whereas a response can only be observed if the patient is followed for seven months to ensure that failure has not occurred. Consequently, for the first seven months of the trial only values of $X = 0$ may be observed, and this problem persists thereafter since the observed proportion of patients for whom $X = 0$ still over-represents the actual value of π_R. This renders any early stopping rule based on X very unreliable. While in theory this bias could be avoided by only scoring X for each patient at seven months, such an approach is very impractical. Another possible alternative might be to use a much shorter interval for defining response, say $R = \{1 < \min(T_P, T_D)\}$. This would have the undesirable effect of declaring patients for whom $1 < T < 7$ to be responders, contrary to the actual goal of the trial. In general, the use of an indicator such as X is only feasible if it can be observed very quickly and it provides a reasonable summary of patient outcome. Another possible way to overcome this problem might be to first apply the interim monitoring rule only after a sufficiently long time interval has elapsed so that a reasonable number of patients have been followed for seven months or longer. However, this defeats the purpose of safety monitoring since ignoring early failures that occur before first applying the stopping rule fails to protect against the case where the failure rate is unacceptably high.

The loss of information resulting from discretizing the time-to-event (TTE) variables T_P and T_D by using X to characterize the patient's outcome is illustrated by a patient who, at a given time during the trial when an interim decision must be made, is alive and has been progression-free for six months. Although X is not yet known for this patient, the observed event $\{6 < \min(T_P, T_D)\}$ clearly provides useful information about π_R. Similarly, if two patients both survive 12 months with $T_P = 6.5$ for one and $T_P = 7.5$ for the other, then $X = 0$ for the first patient while $X = 1$ for the other, despite the fact that their actual outcomes are very similar. The usual motivation for using discretized outcomes such as X to evaluate treatments in early phase trials is that it is impractical or undesirable to wait to observe the event times. The underlying scientific rationale, which usually is not stated explicitly, is that the early outcome is a reasonable surrogate for a TTE variable of primary interest [1, 2]. However, surrogacy typically is difficult to verify in practice [3–5].

Rather than basing clinical trial design on a binary outcome that is a surrogate for one or more TTE variables, a number of authors have advocated the more direct approach of formulating the underlying statistical model and interim decision rules in terms of the TTE variables themselves. Follman and Albert [6] propose monitoring the rate of an adverse event by using a Dirichlet process prior for the probabilities of the event on a large set of discretized event times. They compute an approximate posterior that is a mixture of Dirichlet processes by using a data augmentation algorithm. Rosner [7] takes a similar approach, but uses Gibbs sampling to generate posteriors. Cheung and Thall [8] propose a method, continuous monitoring based on an approximate posterior (CMAP), for constructing futility monitoring rules based on one or more event times used to define a composite event R in phase II. Thall *et al.* [9] use a hierarchical Bayesian model to account for multiple disease subtypes in a phase II trial based on event times. A general discussion of Bayesian methods in clinical trials is given by Spiegelhalter *et al.* [10].

In this paper, we present a series of Bayesian designs for a phase II trial based on three TTE variables. The designs are developed in the context of a trial to evaluate a new treatment for kidney cancer, and are based on the times to death, a severe adverse event (SAE) and disease progression. The first design is based on the time to failure, defined as any of these three events, assuming that failure time is exponentially distributed with mean following an inverse gamma prior. The performance of this design is evaluated via simulation and compared to the CMAP design. Once the initial TTE design based on the exponential-inverse gamma model is established, we successively refine the model and design to incorporate additional elements of complexity that may arise in practice. These refinements account, in turn, for the complications that disease progression is evaluated for each patient at a sequence of times rather than continuously, the hazard of death increases at the time of disease progression, and one may wish to randomize patients rather than conduct a single-arm trial. In each case we derive a general likelihood but, in order to focus on these particular

issues, we deal with the simple case where each event time is exponentially distributed with inverse gamma prior on the mean. Next, to account for the effect of progression on survival, we extend the survival time distribution to a piecewise exponential with hazard changing at progression. Since the assumption of exponentially distributed event times may be inadequate in many settings, we study the robustness of the initial design and of CMAP to departures from exponentiality, and we also include a more general version of the initial design based on the assumption that failure time follows a generalized gamma distribution. We close with a brief discussion of some additional issues, including patient heterogeneity and the use of multiple stopping rules.

Xeloda + Gemzar for kidney cancer

A single-arm phase II trial of the experimental (E) combination Xeloda + Gemzar (X + G) was conducted to obtain a preliminary evaluation of this combination for treatment of advanced kidney cancer. The trial was limited to patients who previously had received immunotherapy and either did not respond or achieved at least a partial disease remission but subsequently relapsed. For these patients, disease progression or death occur on average in about six to eight months with standard therapy (S), which consists of 5-fluorouracil + Gemzar (5-FU + G). While an improvement in progression-free survival (PFS) time was desired, another motivation for the trial was the fact that Xeloda is chemically similar to 5-FU but Xeloda is given orally, rather than intravenously as is required with 5-FU. Each patient's disease status was evaluated at the time of trial entry (baseline), and thereafter at eight week intervals until treatment failure up to 48 weeks, defined as progressive disease compared to the baseline evaluation, a regimen-related SAE at a level of severity precluding further treatment, or death. At this writing, the trial has accrued the maximum of 84 patients, and follow up currently is ongoing.

A simple design for the Xeloda + Gemzar trial

Our first design monitors the time to treatment failure, $T = \min\{T_P, T_D\}$, which we assume is exponentially distributed with mean μ following an inverse gamma (IG) prior. Formally, T has pdf $f(t|\mu) = \mu^{-1}e^{-t/\mu}$ and μ has prior $p(\mu|a, b) = e^{-b/\mu}\, b^a\, \mu^{-(a+1)}/\Gamma(a)$, where $\Gamma(\cdot)$ is the gamma function and $a, b > 0$ are fixed hyperparameters. We will denote this by $T|\mu \sim \text{Exp}(\mu)$ and $\mu|a, b \sim \text{IG}(a, b)$, and refer to this as the exponential-inverse gamma (E-IG) model. Since the IG(a, b) distribution has mean $b/(a-1)$ and variance $b^2/\{(a-1)^2(a-2)\}$, we require $a > 2$. Denoting the hazard by $\lambda = \mu^{-1}$, assuming that $\mu|a,b \sim \text{IG}(a, b)$ is equivalent to assuming that $\lambda|a, b \sim \text{Gam}(a, b)$, a gamma distribution with pdf $f(\lambda) = e^{-b\lambda}b^a\lambda^{a-1}/\Gamma(a)$, which has mean a/b and variance a/b^2. Since $\tilde{\mu} = \text{median}(T) = \log(2)\mu$, and in general if $X \sim \text{IG}(a, b)$ then $cX \sim \text{IG}(a, cb)$ for any $c > 0$, the above priors in terms of μ and λ are equivalent to specifying $\tilde{\mu} \sim \text{IG}(a, \log(2)\, b)$. It will be convenient to use these three equivalent forms of the distribution, and we will move freely between them. This model is especially tractable since the IG is a conjugate prior for the exponential. Given the above structure, we will index the historical standard and experimental treatments by $j = S, E$, so that μ_j, λ_j, $\tilde{\mu}_j$ are the parameters and a_j, b_j are the hyperparameters for treatment j. The time scale of all event times and their corresponding parameters will be in months. Whenever monitoring or observation intervals are given in terms of weeks this will be stated explicitly. We will provide details of prior elicitation later, in the context of a more complex model containing the models considered in this section and the next as special cases.

The historical standard treatment median failure time has prior $\tilde{\mu}_S \sim \text{IG}(53.477, 209.06)$, obtained from the elicited mean 4.0 for $\tilde{\mu}_S$ and 95% credible interval (CI) $\Pr(3.0 < \tilde{\mu}_S < 5.2) = 0.95$. Equivalently, since $\mu_S = \tilde{\mu}_S/\log(2)$, the historical mean time to failure has prior $\mu_S \sim IG(53.477, 301.61)$. On the event rate domain, λ_S has mean $a_S/b_S = 0.177$ and variance $a_S/b_S^2 = 0.00059$. We calibrated the prior of λ_E to have the same mean but $\text{var}(\lambda_E) = 10\text{var}(\lambda_S)$, to reflect the much greater prior uncertainty about E. Thus, $a_E/b_E = 0.177$ and $a_E/b_E^2 = 0.0059$, which implies that $\mu_E \sim \text{IG}(5.348, 30.161)$.

With the exception of the randomized trial to be discussed later, in each of the different cases that we will consider the distributions of any parameters corresponding to S do not change during the single-arm trial of E. Let n denote the number of patients who have been enrolled at the time of any given interim decision. For the ith patient, $i = 1, \ldots, n$, let T_i^o be the observed time of failure or administrative right censoring, and let $Y_i = I(T_i^o = T_i)$ indicate that T_i^o is a failure time. Denote the survivor function (sf) by $\mathcal{F}(t) = \Pr(T > t)$, which takes the form $\mathcal{F}(t) = e^{-t/\mu}$ under the Exp(μ) model. For $data_n = (T_1^o, Y_1, \ldots, T_n^o, Y_n)$, denoting the number of failures by $N_n = \Sigma_{i=1}^n Y_i$ and the total observation time by $T_n^+ = \Sigma_{i=1}^n T_i^o$, the likelihood is the well known expression

$$\begin{aligned}\mathcal{L}(data_n \,|\, \mu_E) &= \prod_{i=1}^{n} f_E(T_i^o|\mu_E)^{Y_i}\, \mathcal{F}_E(T_i^o|\mu_E)^{1-Y_i} \\ &= \mu_E^{-N_n} \exp(-T_n^+/\mu_E). \end{aligned} \quad (1)$$

Posterior computations are facilitated by conjugacy, since $\mu_E \sim \text{IG}(a_E, b_E)$ implies that $[\mu_E \mid N_n, T_n^+] \sim \text{IG}(a_E + N_n, b_E + T_n^+)$.

The design specified that a maximum of 84 patients were to be accrued, subject to the futility monitoring rule that the trial should be stopped early if, based on the current data,

$$Pr(\tilde{\mu}_s + 3 < \tilde{\mu}_E \mid data_n) < p_L. \quad (2)$$

Thus, the trial is stopped early if, given the current data, it is unlikely that the median failure time with E is at least a three month improvement over the historical median with S. This rule is similar to the futility stopping rule given by Thall and Simon [11], who deal with response probabilities $\pi_S = \text{Pr}_S(R)$ and $\pi_E = \text{Pr}_E(R)$ for binary outcomes, rather than median event times. Since $\tilde{\mu}_S$ and $\tilde{\mu}_E$ are independent parameters following IG distributions, (2) may be computed by numerical integration using a package such as S-PLUS, or software freely available at http://biostatistics.mdanderson.org.

In our simulations, we will use the superscript "true" to identify fixed parameter values that determine the probability distributions used to generate patient outcomes, to distinguish them from the random parameters in the Bayesian model. The cut-off p_L in (2) was calibrated to obtain probability of early termination (PET) 0.10 if the true median failure time with E is $\tilde{\mu}_E^{true} = 7$ months, the prior mean $E(\tilde{\mu}_S) = 4$ plus the desired 3 month improvement. This gave $p_L = 0.015$. To obtain the numerical value of p_L yielding a design with PET = 0.10 for a given $\tilde{\mu}_E^{true}$, we first evaluated the criterion probability $\text{Pr}(\tilde{\mu}_S + 3 < \tilde{\mu}_E)$ under the prior and used this as an upper limit $\bar{p}_L$ for p_L. Next, we evaluated the rule for $\tilde{\mu}_E^{true}$ using cut-off $\bar{p}_L/2$. If the resulting PET > 0.10 then $\bar{p}_L/4$ was used next; if PET < 0.10 then $3\bar{p}_L/4$ was used next. This method of bisection was iterated until PET = 0.10 was obtained within three decimal places of accuracy. Additionally, after the second bisection, the method was refined by linearly interpolating or extrapolating. For each cut-off studied, the trial was simulated 100 times, with this increased to 2000 to ensure the desired accuracy of the final value of p_L. In most cases, this required five to 10 iterations.

In the simulations, the rule (2) was applied continuously. Each time a new patient became available for enrollment, based on the most recent data at that time the posterior stopping criterion $\text{Pr}(\tilde{\mu}_S + 3 < \tilde{\mu}_E \mid data_n)$ was updated and applied. The maximum sample size of 84 was chosen, assuming an accrual rate of six patients per month, to ensure a maximum trial duration of about 14 months, to obtain desirable early stopping probabilities, and to obtain a reliable posterior for μ_E. For example, $N_{84} = 70$ failures and total observation time $T_n^+ = 706.3$ would give empirical mean failure time 706.3/70 = 10.10 months, which is what would be expected if $\tilde{\mu}_E^{true}$ = seven months, and these data would give a posterior 95% CI for μ_E of (8.12, 12.94). This design's operating characteristics (OCs) are summarized in the portion of Table 1 labeled "exponential-inverse gamma". For the simulation results summarized in Tables 1, 3, 5, 6 and 7, each case was simulated 2000 times, and the distributions of the number of patients and trial duration are summarized by their 25th, 50th and 75th percentiles.

As a basis for comparison, we simulated the trial using CMAP, which also constructs stopping rules using right-censored event times. In this case, CMAP is based on a probability of the form $\pi_E = \text{P}_E(t^* < T)$ for a fixed time t^*, and it relies on the decomposition $\text{P}_E\{A^o(t)\} = \text{P}_E\{A^o(t) | \text{t}^* < T\}\pi_E + \text{P}_E\{A^o(t) \mid t^* \geq T\}(1 - \pi_E)$, where $A^o(t)$ is the patient's observed data at time t in the trial. CMAP stops the trial if $\text{Pr}(\pi_S + \delta_\pi < \pi_E | data) < p_{L,\pi}$, with this rule applied continuously, using an approximate posterior for π_E obtained by treating $\text{P}_E\{A^o(t) \mid t^* < T\}$ and $\text{P}_E\{A^o(t) \mid t^* \geq T\}$ as nuisance parameters and estimating them empirically. Details are given in Cheung and Thall [8]. To make the two methods comparable, we constructed priors and the monitoring rule for CMAP as follows. We used $t^* = 7$ to define $\pi = \text{Pr}(7 < \text{T})$, and derived a beta($a_S$, b_S) prior on π_S by equating its mean $m_S = a_S/(a_S + b_S)$ and variance $m_S(1 - m_S)/(a_S + b_S + 1)$ to the mean 0.2934 and variance 0.0024 of $\pi_S = \exp(-7/\mu_S)$ under the IG(53.477, 301.61) prior

Table 1 Operating characteristics of the exponential-inverse gamma model-based design and CMAP for the Xeloda + Gemzar trial, based on exponentially distributed failure times

	Monitoring based on right-censored event times					
	Exponential-inverse gamma model			Approximate posteriors (CMAP)		
$\tilde{\mu}_E^{true}$	PET	No. pats.	Trial duration	PET	No. pats.	Trial duration
4	0.96	21 33 48	3.4 5.4 7.9	0.87	42 53 69	6.9 8.4 11.1
5	0.66	33 60 84	5.6 10.1 13.3	0.50	54 83 84	8.8 12.2 14.0
6	0.28	73 84 84	11.0 13.2 14.5	0.23	84 84 84	11.7 13.3 14.5
7	0.10	84 84 84	12.4 13.7 14.7	0.10	84 84 84	12.4 13.6 14.7

on μ_S. This yielded a beta(25.084, 60.406) prior for π_S. We used a beta(0.587, 1.413) prior for π_E, which has the same mean as π_S but effective sample size $a_E + b_E = 2$. Since $\tilde{\mu}_E^{true} = 7$ implies that $\pi_E^{true} = \exp[-7/\{7/\log(2)\}] = 0.50$, we used $\delta_\pi = 0.50 - 0.29 = 0.21$ in the CMAP stopping rule. As with (2), we calibrated the cut-off of the CMAP rule to obtain PET = 0.10 at $\tilde{\mu}_E^{true} = 7$, equivalently if $\pi_E^{true} = 0.50$, which yielded $p_{L,\pi} = 0.012$.

The simulation results are summarized in Table 1. Since both methods have PET = 0.10 at $\tilde{\mu}_E^{true}$ = seven months, differences can be seen as $\tilde{\mu}^{true}$ decreases from 7 to less desirable values, with the PET increasing much less rapidly for CMAP. In the undesirable case where $\tilde{\mu}_E^{true} = 4$, CMAP has PET = 0.87 and median sample size 53, compared to PET = 0.96 and median sample size 33 for the E-IG model-based method. Thus, in this case, CMAP has substantially less desirable OCs compared to the model-based method. This comparison is not entirely fair, however, since the E-IG model-based method is being evaluated assuming that the underlying model is correct. In the section Robustness, we will evaluate the robustness of these designs when the failure times are not exponentially distributed.

It is useful to consider how a comparable design based on the binary discretized outcome $X = I(7 < T)$, rather than T, would behave in this case. Suppose one assumes the same priors on π_E and π_S and applies the same stopping rule as used for the CMAP design, but now assuming the likelihood $\prod_i \pi_E^{X_i}(1 - \pi_E)^{1-X_i}$ and scoring the X_i's when they are first observed, so that $X_i = 1$ at seven months if $T_i \geq 7$ and $X_i = 0$ at T_i if $T_i < 7$. The posterior of π_E would be $\beta(0.587 + \Sigma_i X_i,\ 1.413 + \Sigma_i(1 - X_i))$, so computing the early stopping criterion probability $\Pr(\pi_S + \delta_\pi < \pi_E|\text{data})$ is straightforward. However, initially only values of $X_i = 0$ may be observed, so the data for estimating π_E are heavily biased. A consequence of this is that, if $\tilde{\mu}_E^{true} = 7$ so that $\pi_E^{true} = 0.50$, a desirable value, then the early stopping probability with this approach exceeds 0.99. That is, the rule is virtually certain to stop the trial in a desirable case where one would not want to stop.

The targeted three-month improvement of $\tilde{\mu}_E$ over $\tilde{\mu}_S$ in the stopping rule (2) is a subjective value determined by the clinician. In the illustrative trial, it is arguable that since Xeloda is given orally and has a lower risk of adverse effects, it may be reasonable to use the stopping criterion $\Pr(\tilde{\mu}_S < \tilde{\mu}_E \mid data_n)$ with no targeted improvement in median failure time. This may be considered a phase II equivalence trial, as defined by Thall and Sung for discrete outcomes ([12], Section III). In this case, the historical mean of 4.0 months for $\tilde{\mu}_S$ is considered a desirable value of $\tilde{\mu}_E^{true}$, and only very small values, such as $\tilde{\mu}_E^{true} = 1.0$ or 2.0 months, are undesirable. Calibrating this version of the rule to have PET = 0.10 for $\tilde{\mu}_E^{true} = 4.0$ yields a design with $p_L = 0.086$. For $\tilde{\mu}_E^{true} = 1$, 2 and 3, the respective early stopping probabilities are PET = 1.00, 1.00 and 0.64, with median sample sizes 13, 23 and 59.

Since it may not be feasible to monitor the data continuously in some trials, it is worthwhile to examine the behavior of the design if the early stopping rule is applied periodically. To do this, we repeated the simulations, but with (2) applied every k weeks, for k = 1, 2, 4, 6, 8, 12, 16, 20 or 24. The results, summarized in Table 2, show that there is a gradual decline in PET as the monitoring interval is increased, but even monitoring every eight weeks still maintains PET = 0.93 when $\tilde{\mu}_E^{true} = 4$. It thus appears that, if one accounts for the event times in this way, then applying the stopping rule periodically still provides a safe design while imposing less of a practical burden during trial conduct. Moreover, if periodic monitoring is planned initially then the value of p_L in the stopping rule (2) may be calibrated so that PET equals a given small value when $\tilde{\mu}_E^{true}$ equals a desirable target.

Accounting for interval censored progression times

Because each patient's disease status is evaluated at eight-week intervals, the actual time of any patients' disease progression is not available. Rather, it is only known whether progression occurred during each time interval between successive examinations. For example, if progression is first discovered at the week 24 examination, then it is only known that progression occurred between weeks 16 and 24. This sort of interval censoring of the time of transition between disease states is common in medical settings where the patient's disease status is evaluated periodically by tests such as magnetic resonance imaging or computed axial tomography scan. The previous probability model ignores this complication.

To account for interval censoring, first consider a single patient and temporarily suppress both the treatment and patient indices. Let T_P denote the time of disease progression, $T_{DA} = \min\{T_D, T_A\}$

Table 2 Effect of periodic rather than continuous monitoring. The column labeled "0"corresponds to continuous monitoring. Each entry is the probability of early termination

	Monitoring period, in weeks									
$\tilde{\mu}_E^{true}$	0	1	2	4	6	8	12	16	20	24
4	0.96	0.96	0.95	0.94	0.94	0.93	0.91	0.89	0.84	0.85
7	0.10	0.10	0.08	0.08	0.07	0.06	0.05	0.04	0.03	0.03

the time to death or a SAE, and identify the corresponding parameters and probability functions by the subscripts P and DA. Thus, we now account for two event times, rather than only one. For now, we will assume that T_P and T_{DA} are independent and exponentially distributed with parameters μ_P and μ_{DA}. Let $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots$ denote the successive times when the patient's disease status is evaluated, allowing the possibility that a patient's actual evaluation times may deviate from the scheduled times. We will assume that, like death, an SAE is a terminating event in that follow-up ends at the time of an SAE, but patients may be followed for some period of time after progression.

Each patient's likelihood contribution may take one of the following possible forms, which are a consequence of the fact that, while the actual value of T_{DA} is observed, only the interval during which T_P occurs can be known. Let T^o denote the time of the last observed event or follow up, with $Y_{DA} = I(T_{DA} = T^o)$ the indicator that the patient's time of death or an SAE is observed. Denote the last disease evaluation time by τ_{k^o}, let $Y_P = I(\tau_{k^o-1} < T_P \leq \tau_{k^o})$ indicate that progression is discovered at τ_{k^o}, and denote the probability that progression occurs between the $k-1$st and kth disease evaluation times by $\pi_k = \Pr(\tau_{k-1} < T_P \leq \tau_k) = \mathcal{F}_P(\tau_{k-1}) - \mathcal{F}_P(\tau_k)$. If the patient dies or has an SAE without previous observation of disease progression then, in addition to observing T_{DA} it is known that $\tau_{k^o} < T_P$, so in this case the likelihood contribution is $f_{DA}(T_{DA})\mathcal{F}_P(\tau_{k^o})$. If the evaluation at τ_{k-1} is negative, the patient survives to τ_k without an SAE, and the evaluation at τ_k shows progression, then $\tau_k = \tau_{k^o}$ and it is known that $\tau_{k^o-1} < T_P \leq \tau_{k^o} < T_{DA}$. Since in any case $\tau_{k^o} \leq T^o$, the likelihood contribution of this patient is either $\pi_{k^o}\, \mathcal{F}_{DA}(T^o)$ if T_{DA} is administratively censored at T^o, or $\pi_{k^o}\, f_{DA}(T_{DA})$ if the patient dies or has an SAE, i.e., if $T^o = T_{DA}$. If T_{DA} is administratively censored at T^o and progression was not observed at τ_{k^o}, then the likelihood contribution is $\mathcal{F}_{DA}(T^o)\mathcal{F}_P(\tau_{k^o})$. Since it is only known if a patient progresses during one of the intervals $(\tau_k - 1,\ \tau_k]$ and also survives to τ_k, it follows that $Y_P = 1$ if either the last follow time is $\tau_{k^o} = T^o < T_{DA}$, or if progression is discovered at τ_{k^o} and the patient later dies or has an SAE, in which case $\tau_{k^o} < T^o = T_{DA}$.

Accounting for all of this additional structure due to considering P and DA as separate events and accounting for the interval censoring of T_P, and now reintroducing the patient indices, the likelihood is given generally by

$$\mathcal{L}(data_n | f_{DA,}\, f_P) = \prod_{i=1}^{n} f_{DA}(T_i^o)^{Y_{i,DA}}\, \mathcal{F}_{DA}(T_i^o)^{1-Y_{i,DA}} \times \pi_{i,k^o}^{Y_{i,P}}\, \mathcal{F}_p(\tau_{i,k^o})^{1-Y_{i,P}}. \quad (3)$$

For the exponential case, denote the number of deaths or SAEs by $N_{n,DA} = \Sigma_{i=1}^{n} Y_{i,DA}$ and $T_{n,P}^{+} = \Sigma_{i=1}^{n} \tau_{i,k^o}(1 - Y_{i,P})$. The general likelihood (3) now takes the specific form

$$\mathcal{L}(data_n | \mu_{DA,}\, \mu_P) = \lambda_{DA}^{N_{n,DA}}\, e^{-T_{n,DA}^{+}\lambda_{DA} - T_{n,P}^{+}\lambda_P} \times \prod_{i=1}^{n} \left\{ e^{-\tau_{i,k^o-1}\lambda_P} - e^{-\tau_{i,k^o}\lambda_P} \right\}^{Y_{i,P}}. \quad (4)$$

Since $\mathcal{L}(data_n \,|\, \mu_{DA},\, \mu_P) = \mathcal{L}(data_n \,|\, \mu_{DA})\ \mathcal{L}(data_n \,|\, \mu_P)$, the posteriors of μ_{DA} and μ_P may be computed separately, and these two parameters also are independent *a posteriori*.

For this extended model, priors on the four parameters $\boldsymbol{\mu}_S = (\mu_{S,P},\ \mu_{S,DA})$ and $\boldsymbol{\mu}_E = (\mu_{E,P},\ \mu_{E,\ DA})$ are required. We assume independent priors with $\mu_{j,r} \sim \text{IG}(a_{j,r},\ b_{j,r})$ for each treatment $j = S,\ E$ and outcome $r = P,\ DA$. Some care must be taken when specifying the four hyperparameters $(a_{j,P},\ b_{j,P},\ a_{j,DA},\ b_{j,DA})$ for each j, since the fact that the hazards of independent exponentials are additive imposes some constraints. Specifically, $T_j = \min(T_{j,P},\ T_{j,DA})$ implies that $\lambda_j = \lambda_{j,P} + \lambda_{j,DA}$ for each j $= E,\ S$. This in turn implies that $\text{E}(\lambda_j) = \text{E}(\lambda_{j,P}) + \text{E}(\lambda_{j,DA})$ and, due to the independence of $\lambda_{j,P}$ and $\lambda_{j,DA}$, $\text{var}(\lambda_j) = \text{var}(\lambda_{j,P}) + \text{var}(\lambda_{j,D})$. Here, $(a_{S,P},\ b_{S,P},\ a_{S,DA},\ b_{S,DA}) = (20.391,\ 195.826,\ 95.401,\ 1303.670)$, which implies that $\text{E}(\tilde{\mu}_{S,P}) = 7$ and $\text{E}(\tilde{\mu}_{S,DA}) = 9.57$. As before, we assumed that the means of the hazards for E were the same as those seen historically but we inflated the variances by multiplying by 10, so that $\text{E}(\lambda_{E,r}) = \text{E}(\lambda_{S,r})$ and $\text{var}(\lambda_{E,r}) = 10\ \text{var}(\lambda_{S,r})$ for $r = P$ and DA. This yielded $(a_{E,P},\ b_{E,P},\ a_{E,DA},\ b_{E,DA}) = (2.039,\ 19.583,\ 9.540,\ 130.367)$.

To account for interval censoring of T_p, we write (2) in the form

$$Pr\left\{ \left(\tilde{\mu}_{S,DA}^{-1} + \tilde{\mu}_{S,P}^{-1}\right)^{-1} + 3 < \left(\tilde{\mu}_{E,DA}^{-1} + \tilde{\mu}_{E,P}^{-1}\right)^{-1} | data_n \right\} < p_L, \quad (5)$$

which accounts for all four elements of $(\boldsymbol{\mu}_S,\ \boldsymbol{\mu}_E)$ under the extended model. The posterior probability in (5) is based on the likelihood (4), with the distribution of $\boldsymbol{\mu}_S$ established as described above and fixed throughout the trial.

Posterior distributions under this model, and the models discussed below in the following two sections, were computed using iterative defensive importance sampling [13], which requires one to locate the mode of $\mathcal{L}(data \mid \boldsymbol{\theta})$ prior$(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ and compute the gradient at the mode at each iteration. We used the Nelder–Mead method [14] to find the mode. All programming was done in C++, which provides speed, reusability and flexibility. This language also was chosen to take advantage of the extensive in-house library of C++ computer programs in the M.D. Anderson Department of Biostatistics and Applied Mathematics, which are

Table 3 Operating characteristics of the Xeloda + Gemzar trial with T_P interval censored due to progression being evaluated at eight-week intervals for each patient

			Modelling interval censored T_p			Ignoring interval censoring		
$\tilde{\mu}_E^{true}$	$\tilde{\mu}_{E,P}^{true}$	$\tilde{\mu}_{E,DA}^{true}$	PET	No. pats.	Trial duration	PET	No. pats.	Trial duration
4	6.0	12.0	0.98	24 34 47	4.1 5.6 7.9	0.70	42 63 84	7.0 10.5 13.1
4	7.0	9.33	0.97	26 36 49	4.5 6.0 8.1	0.79	36 56 78	6.0 9.1 12.3
4	8.0	8.0	0.98	27 38 52	4.6 6.3 8.6	0.84	35 50 72	5.6 8.4 11.8
7	10.0	23.33	0.14	84 84 84	12.2 13.5 14.6	0.01	84 84 84	12.8 13.8 14.8
7	12.0	16.8	0.10	84 84 84	12.4 13.6 14.8	0.01	84 84 84	12.8 13.8 15.0
7	14.0	14.0	0.07	84 84 84	12.6 13.8 14.8	0.01	84 84 84	12.8 13.9 14.9

available from the second author on request. While many of the Bayesian computations described here could be carried out in WinBUGS, implementing the simulations described here using this approach would be highly complex.

To simulate the trial, we generated T_P and T_{DA} independently for several pairs of $(\tilde{\mu}_{E,P}^{true}, \tilde{\mu}_{E,DA}^{true})$ values such that $1/\tilde{\mu}_{E,P}^{true} + 1/\tilde{\mu}_{E,DA}^{true} = 1/\tilde{\mu}_E^{true}$, with either $\tilde{\mu}_E^{true} = 4$, the historical mean value, or $\tilde{\mu}_E^{true} = 7$, the desired target. For each patient, T_{DA} was observed continuously and T_P was observed at eight-week intervals up to a maximum of 48 weeks (six evaluations). The cut-off p_L was calibrated to obtain PET = 0.10 in the desirable case $(\tilde{\mu}_{E,P}^{true}, \tilde{\mu}_{E,DA}^{true}) = (12.0, 16.8)$, for which $\tilde{\mu}_E^{true} = 7$. This yielded $p_L = 0.019$. The stopping rule (5) was applied continuously. The simulations are summarized in the portion of Table 3 labeled "Modeling Interval Censored T_P".

To quantify what is gained by modeling T_P and T_{DA} as separate events and accounting for the fact that T_P is interval censored, we evaluated the OCs of the first design based on the simpler formulation with stopping rule (2) given previously, when in fact each patient's progression times are interval censored. That is, we simulated the observed event process for (T_P, T_{DA}) with T_P interval censored but used the rule (2) under the simple E-IG model for $T = \min(T_P, T_{DA})$. The results are summarized in the portion of Table 3 labeled "Ignoring Interval Censoring". These simulations show that ignoring the fact that T_P is interval censored greatly reduces the design's PET values. Thus, the OCs of a design that ignores interval censoring may be very misleading, and accounting for the fact that progression is only observed periodically provides a much safer design.

Accounting for the effect of disease progression on survival

A piecewise likelihood

Thus far, we have assumed that the three events occur independently. In kidney cancer and many other solid tumors, however, the hazard of death increases with disease progression. This additional complication, which is very important clinically and also may impact the way that a given monitoring rule behaves, may be modeled in a number of ways. Here, temporarily suppressing S and E for simplicity, we will use a piecewise distribution under which the pdf of T_D changes at T_P from $f_1(x)$ to $\mathcal{F}_1(T_P)f_2(x - T_P)$, where $f_1(x)$ and $f_2(x)$ are pdfs defined for $x > 0$. The joint distribution of T_D and T_P is given generally by

$$\begin{aligned} f_{D,P}(x,y) &= f_{D|P}(x\,|\,y) f_p(y) \\ &= \{f_1(x) I(x<y) + \mathcal{F}_1(y) f_2(x-y) I(x \geq y)\} \\ &\quad f_p(y),\ x, y > 0. \end{aligned} \tag{6}$$

Under this piecewise model, still accounting for the interval censoring of T_P, the likelihood takes one of four possible forms, summarized in Table 4. To express things more compactly, we combine the first two rows of Table 4, which correspond to the two cases where $Y_P = 1$. Under the piecewise model (6), the probability that progression is discovered at τ_{k^o} and is followed by either death or censoring at T^o is

$$\pi_{k^o,P}(T^o, Y_D) = \int_{\tau_{k^o-1}}^{\tau_k^o} f_P(y) \mathcal{F}_1(y) f_2(T^o - y)^{Y_D} \mathcal{F}_2(T^o - y)^{1-Y_D}\, dy. \tag{7}$$

Similarly, we combine the last two rows of Table 4, which correspond to the two cases where where $Y_P = 0$. The probability that the last disease evaluation at $\tau_k{}^o$ shows no progression and this is followed by death or censoring at T^o is

$$\begin{aligned} \pi_{k^o,\bar{P}}(T^o, Y_D) &= \int_{\tau_{k^o}}^{T_o} f_P(y) \mathcal{F}_1(y) f_2(T^o - y)^{Y_D} \mathcal{F}_2(T^o - y)^{1-Y_D} dy \\ &\quad + \mathcal{F}_P(T^o) f_1(T^o)^{Y_D} \mathcal{F}_1(T^o)^{1-Y_D}. \end{aligned} \tag{8}$$

The second summand in (8) is needed to include the event $T_P > T^o$, that the patient has not progressed by the last follow up time. Denote $T_A^o = \min(T_A, T^o)$ and $Y_A = I(T_A^o = T_A)$. Assuming that T_A and (T_P, T_D) are independent, the general

Table 4 Possible outcomes and likelihood contributions for the times to progression, T_P, and death, T_D, under the model with T_P interval censored and the hazard of death changing from from $f_1/\mathcal{F}_1$ before T_P to $f_2/\mathcal{F}_2$ after T_P

Y_P	Y_D	Outcome	Likelihood contribution
1	1	$\tau_{k^o-1} < T_P \le \tau_k^o < T^o = T_D$	$\int_{\tau_{k^o-1}}^{\tau_k^o} f_p(y)\mathcal{F}_1(y)f_2(T_D-y)dy$
1	0	$\tau_{k^o-1} < T_P \le \tau_k^o \le T^o < T_D$	$\int_{\tau_{k^o-1}}^{\tau_k^o} f_p(y)\mathcal{F}_1(y)\mathcal{F}_2(T_D-y)dy$
0	1	$\tau_{k^o} < T_P$ and $\tau_{k^o} < T^o = T_D$	$\int_{\tau_{k^o}}^{T_D} f_p(y)\mathcal{F}_1(y)f_2(T_D-y)dy + \mathcal{F}_p(T_D)f_1(T_D)$
0	0	$\tau_{k^o} < T_P$ and $\tau_{k^o} \le T^o < T_D$	$\int_{\tau_{k^o}}^{T^o} f_p(y)\mathcal{F}_1(y)\mathcal{F}_2(T^o-y)dy + \mathcal{F}_p(T^o)\mathcal{F}_1(T^o)$

likelihood for all of the possible observations of the three types of events now may be expressed as

$$\mathcal{L}\left(\tau_{k^o}, T^o, T_A^o, Y_P, Y_D, Y_A \mid f_D, f_A, f_P\right) =$$

$$\pi_{k^o,P}\left(T^o, Y_D\right)^{Y_P} \pi_{k^o,\bar{P}}\left(T^o, Y_D\right)^{1-Y_P} f_A\left(T_A^o\right)^{Y_A} \mathcal{F}_A\left(T_A^o\right)^{1-Y_A}. \quad (9)$$

Under the piecewise exponential model where the hazard of death changes from λ_1 to λ_2 at T_P, denoting $\gamma = \lambda_P + \lambda_1 - \lambda_2$, the general likelihood (6) takes the specific form

$$f_{D,P}(x,y) = \lambda_1\lambda_P e^{-x\lambda_1 - y\lambda_P} I(x < y) + \lambda_2\lambda_P e^{-x\lambda_2 - y\gamma} I(x \ge y), \quad (10)$$

the marginal pdf of T_D is

$$f_D\left(x \mid \lambda_1, \lambda_2, \lambda_P\right) = \lambda_1 e^{-(\lambda_1 - y\lambda_P)x} + \frac{\lambda_2\lambda_P\left\{e^{-\lambda_2 x} - e^{-(\lambda_1+\lambda_P)x}\right\}}{\lambda_P + \lambda_1 - \lambda_2}, \quad (11)$$

and the probabilities (7) and (8) take the forms

$$\pi_{k^o,P}\left(T^o, Y_D\right) = \gamma^{-1}\lambda_P\lambda_2^{Y_D} e^{-T^o\lambda_2}\left(e^{-\tau_{k^o-1}\gamma} - e^{-\tau_{k^o}\gamma}\right) \quad (12)$$

and

$$\pi_{k^o,\bar{P}}\left(T^o, Y_D\right) = \gamma^{-1}\lambda_P\lambda_2^{Y_D} e^{-T^o\lambda_2}\left(e^{-\tau_{k^o}\gamma} - e^{-T^o\gamma}\right) + \lambda_1^{Y_D} e^{-T^o(\lambda_P+\lambda_1)} \quad (13)$$

Combining these expressions with the fact that $f_A(T_A)^{Y_A}\mathcal{F}_A(T_A)^{1-Y_A} = \lambda_A^{Y_A} e^{-T_A^o\lambda_A}$ exponential case, the likelihood (9) may be written as

$$\mathcal{L}(data_n \mid \lambda_1, \lambda_2, \lambda_A, \lambda_P) = \lambda_A^{N_{n,A}} e^{-T_{n,A}^*\lambda_A} \prod_{i=1}^{n} \pi_{i,k,P}\left(T_i^o, Y_{i,D}\right)^{Y_{i,P}} \times \pi_{i,k,\bar{P}}\left(T_i^o, Y_{i,D}\right)^{1-Y_{i,P}}, \quad (14)$$

where $N_{n,A} = \Sigma_{i=1}^n Y_{i,A}$ and $T_{n,A} = \Sigma_{i=1}^n\{T_{i,A}Y_{i,A} + T_i^o(1 - Y_{i,A})\}$.

The hazard of death under the piecewise model is

$$h_D(x \mid \lambda_1, \lambda_2, \lambda_P) = \frac{(\lambda_1 - \lambda_2)(\lambda_1 + \lambda_P)e^{-(\lambda_1+\lambda_P)x} + \lambda_2\lambda_P e^{-\lambda_2 x}}{(\lambda_1 - \lambda_2)e^{-(\lambda_1+\lambda_P)x} + \lambda_P e^{-\lambda_2 x}}. \quad (15)$$

This expression reduces to λ under the simple exponential model where $\lambda_1 = \lambda_2 = \lambda$, and it converges to $\lambda_1 + \lambda_P$ as $\lambda_2 \to \infty$. The PFS time $\min\{T_P, T_D\} \sim \text{Exp}(\lambda_1^{-1} + \lambda_P^{-1})$, which is the same distribution as under the model where T_P and T_D are independent with $T_D \sim f_1$. Intuitively, this is the case because the hazard of death after progression has no effect on $\min\{T_P, T_D\}$. Consequently, the median PFS time equals $(\tilde{\mu}_1^{-1} + \tilde{\mu}_P^{-1})^{-1}$. Since T_A is independent of (T_P, T_D), it follows that the overall failure time $T \sim \text{Exp}(\lambda_1^{-1} + \lambda_P^{-1} + \lambda_A^{-1})$ so, in terms of the medians, the early stopping rule is

$$Pr\left\{\left(\mu_{S,1}^{-1} + \mu_{S,P}^{-1} + \mu_{S,A}^{-1}\right)^{-1} + 3 < \left(\tilde{\mu}_{E,1}^{-1} + \tilde{\mu}_{E,P}^{-1} + \tilde{\mu}_{E,A}^{-1}\right)^{-1} \mid data_n\right\} < p_L. \quad (16)$$

In particular, the post-progression death rate λ_2 plays no role in (16). However, if one wishes to monitor the death rate under the piecewise model, although median(T_D) cannot be computed in closed form, one may formulate an early stopping rule in terms of either the mean survival time, $E(T_D) = \{\lambda_1\lambda_2 + \lambda_P(\lambda_1 + \lambda_P + \lambda_2)\}/\{\lambda_2(\lambda_1 + \lambda_P)^2\}$, or the hazard function $h_D(x^*)$ evaluated at some fixed time $x = x^*$, since both quantities involve all three parameters $(\lambda_1, \lambda_2, \lambda_P)$ characterizing $f_{D,P}(x, y)$ and $f_D(x)$.

Establishing priors

In order to establish priors on $\boldsymbol{\lambda}_S = (\lambda_{S,1}, \lambda_{S,2}, \lambda_{S,A}, \lambda_{S,P})$ under the piecewise model (14), we proceeded in two stages. Recall that the priors on $\tilde{\mu}_{j,r}$, $\mu_{j,r}$ and $\lambda_{j,r}$ determine each other. We first established

priors on ($\lambda_{S,D}$, $\lambda_{S,A}$, $\lambda_{S,P}$) under the simpler model that assumes T_D, T_P and T_A are mutually independent, and we then extended this prior to account for the effect of progression on the hazard of death under the piecewise model. We proceeded in this way because, from a clinician's viewpoint, the piecewise model is rather complex. The simpler model thus serves as a conceptual bridge to check that the priors on $\lambda_{S,1}$, $\lambda_{S,2}$ and $\lambda_{S,P}$ yield a prior on $\lambda_{S,D}$ that makes sense.

For convenience, again temporarily suppress S and E. The possible outcomes described previously for P and DA now pertain to P and D. Allowing the possibility that a patient's follow-up may be continued beyond T_A, we define $T_A^o = \min\{T_A, T^o\}$ and the indicator $Y_A = I(T_A \leq T^o)$ that an SAE is observed. Accounting for three separate events and interval censoring of T_P, still assuming independence, the general likelihood is

$$\begin{aligned}\mathcal{L}(data_n \mid f_D, f_A, f_P) &= \prod_{i=1}^{n} f_D\left(T_i^o\right)^{Y_{i,D}} \mathcal{F}_D\left(T_i^o\right)^{1-Y_{i,D}} f_A\left(T_{i,A}^o\right)^{Y_{i,A}} \mathcal{F}_A\left(T_{i,A}^o\right)^{1-Y_{i,A}} \\ &\quad \times \pi_{i,k^o}^{\,Y_{i,P}}\, \mathcal{F}_P\left(\tau_{i,k^o}\right)^{1-Y_{i,P}}. \end{aligned} \tag{17}$$

Under the exponential model where each $T_r \mid \mu_r \sim \text{Exp}(\mu_r)$ with $\mu_r \sim \text{IG}(a_r, b_r)$, the above likelihood takes the specific form

$$\begin{aligned}&\mathcal{L}(data_n \mid \lambda_D, \lambda_A, \lambda_P) \\ &= \lambda_D^{N_{n,D}} \lambda_A^{N_{n,A}} e^{-T_{n,D}^{+}\lambda_D - T_{n,A}^{+}\lambda_A - T_{n,P}^{+}\lambda_P} \prod_{i=1}^{n} \left\{ e^{-\tau_{i,k^o-1}\lambda_P} - e^{-\tau_{i,k^o}\lambda_P} \right\}^{Y_{i,P}} \end{aligned} \tag{18}$$

Re-introducing S and E, the six hyperparameters ($a_{S,P}$, $b_{S,P}$, $a_{S,D}$, $b_{S,D}$, $a_{S,A}$, $b_{S,A}$) characterizing the three independent priors on ($\lambda_{S,D}$, $\lambda_{S,A}$, $\lambda_{S,P}$) may be elicited in many ways. See, for example, Chaloner *et al.* [15], or Kadane and Wolfson [16]. A straightforward approach is to elicit the mean and a 95% credible interval for each $\tilde{\mu}_{S,r}$, which together determine ($a_{S,r}$, $b_{S,r}$). However, one must proceed with caution when eliciting priors on $\tilde{\mu}_{S,P}$, $\tilde{\mu}_{S,D}$ and $\tilde{\mu}_{S,A}$ so that, when combined, they yield a reasonable prior on the overall failure time median, $\tilde{\mu}_S$. Since the hazards of independent exponentials are additive, $\lambda_j = \lambda_{j,P} + \lambda_{j,D} + \lambda_{j,A}$ for $j = E$, S, and taking means and variances gives the two equations

$$\frac{a_j}{b_j} = \frac{a_{j,P}}{b_{j,P}} + \frac{a_{j,D}}{b_{j,D}} + \frac{a_{j,A}}{b_{j,A}} \tag{19}$$

and

$$\frac{a_j}{b_j^2} = \frac{a_{j,P}}{b_{j,P}^2} + \frac{a_{j,D}}{b_{j,D}^2} + \frac{a_{j,A}}{b_{j,A}^2}. \tag{20}$$

Thus, given ($a_{S,P}$, $b_{S,P}$, $a_{S,D}$, $b_{S,D}$, $a_{S,A}$, $b_{S,A}$), it is important to check that the values (a_S, b_S) resulting from (19) and (20) give a prior on the overall failure time parameter that makes sense. Algebraically, one must determine eight hyperparameters subject to the two constraints (19) and (20), so there are really six pieces of information. In practice, instead of determining the six hyperparameters ($a_{S,P}$, $b_{S,P}$, $a_{S,D}$, $b_{S,D}$, $a_{S,A}$, $b_{S,A}$) on the right-hand sides of (19) and (20) and then hoping that the resulting (a_S, b_S) gives a reasonable prior on $\tilde{\mu}_S$, one may elicit the six pieces of information while taking advantage of the physician's familiarity with the overall failure rate. To do this, one may first elicit the mean and 95% CI of $\tilde{\mu}_S$ to determine (a_S, b_S), and then elicit four of the six event-specific hyperparameters. Substituting these values into (19) and (20), one may then check that the resulting prior of the remaining component event time is reasonable. This process may be iterated if needed to calibrate some of the hyperparameters. We took this approach, which produced prior means and 95% CI's $\text{E}(\tilde{\mu}_S) = 4$ (3, 5.2) for the overall failure time median, and $\text{E}(\tilde{\mu}_{S,P}) = 7$ (2, 10) and $\text{E}(\tilde{\mu}_{S,D}) = 12$ (9, 15) for the medians of T_P and T_D. The resulting hyperparameters were ($a_{S,P}$, $b_{S,P}$, $a_{S,D}$, $b_{S,D}$, $a_{S,A}$, $b_{S,A}$) = (20.391, 195.826, 61.827, 1053.050, 552.371, 38182.100) and (a_S, b_S) = (53.477, 301.61), as given earlier. This issue is still present when using a nonexponential event time distribution, such as a Weibull, lognormal or gamma, since in general the hazard of overall failure is determined by the hazards of the component events.

To extend this prior to accommodate the piecewise model, we next elicited priors on $\tilde{\mu}_{S,1}$ and $\tilde{\mu}_{S,2}$, subject to the constraint that the resulting prior on $\mu_{S,D}$ has the above mean and 95% CI. This required an iterative process of repeatedly specifying priors on $\tilde{\mu}_{S,1}$ and $\tilde{\mu}_{S,2}$ and evaluating the resulting prior of $\tilde{\mu}_{S,D}$ until this had mean 12 and 95% CI (9, 15). This gave prior mean and 95% CI of 14 (12, 16) for $\tilde{\mu}_{S,1}$ and 5.75 (2.75, 8.75) for $\tilde{\mu}_{S,2}$, which imply that ($a_{S,1}$, $b_{S,1}$, $a_{S,2}$, $b_{S,2}$) = (188.521, 3787.490, 14.939, 115.627). As before, we assumed that the means of the hazards for E were the same as for S, but we inflated the variances by multiplying by 10, so that $\text{E}(\lambda_{E,r}) = \text{E}(\lambda_{S,r})$ and $\text{var}(\lambda_{E,r}) = 10\ \text{var}(\lambda_{S,r})$ for $r = P$, 1, and A. For the post-progression death rate, we used the smaller multiplier $\text{var}(\lambda_{E,2}) = 2.5\ \text{var}(\lambda_{S,2})$ to stabilize the computations. This gives the $\text{var}(\mu_{E,2}) = 10.4$, which is very close to $\text{var}(\mu_{E,1}) = 12.8$. These values yielded the hyperparameters ($a_{E,P}$, $b_{E,P}$, $a_{E,1}$, $b_{E,1}$, $a_{E,2}$, $b_{E,2}$, $a_{E,A}$, $b_{E,A}$) = (2.039, 13.574, 18.852, 262.529, 5.974, 32.053, 55.237, 2646.580).

To simulate the design under the piecewise model, each scenario is determined by the four parameters $\boldsymbol{\lambda}_E^{true} = (\lambda_{E,1}, \lambda_{E,2}, \lambda_{E,P}, \lambda_{E,A})^{true}$. We chose

Table 5 Operating characteristics of the Xeloda + Gemzar trial based on a model accounting for the effect of progression on the hazard of death. The design is identical that summarized in Table 3, but here the hazard of death changes at T_P under the underlying piecewise hazard model. In the null case $(\tilde{\mu}_{E,1}, \tilde{\mu}_{E,2}, \tilde{\mu}_{E,P})^{true} = (14, 5.75, 7)$ with overall $\tilde{\mu}_E^{true} = 4.0$. In the alternative case $(\tilde{\mu}_{E,1}, \tilde{\mu}_{E,2}, \tilde{\mu}_{E,P})^{true} = (32, 12, 12.25)$ with overall $\tilde{\mu}_E^{true} = 7.0$. In both cases, $\tilde{\mu}_{E,A}^{true} = 48$

	Piecewise hazard model			Ignoring effect of P on λ_D		
$\tilde{\mu}_E^{true}$	PET	No. pats.	Trial dur.	PET	No. pats.	Trial dur.
4.0	0.96	31 42 56	5.2 7.0 9.3	0.99	25 34 46	4.2 5.7 7.6
7.0	0.10	84 84 84	12.4 13.6 14.7	0.15	84 84 84	12.3 13.6 14.6

numerical values of $\boldsymbol{\lambda}_E^{true}$ to correspond to the simpler cases studied previously, with overall median failure time $\tilde{\mu}_E^{true}$ either 4.0 or 7.0. As before, we calibrated the cut-off in the stopping rule to obtain PET = 0.10 in the desirable case, which gave $p_L = 0.006$. To assess the effect of accounting for the changing hazard of death, in each case we also simulated the trial using the rule (5) based on the previous model that assumes the hazard of death is not affected by progression. The simulations, summarized in Table 5, show that the design has very desirable properties, and that ignoring the fact that progression increases the subsequent hazard of death inflates the PET, with the PET increasing 50%, from 0.10 to 0.15, when $\tilde{\mu}_E^{true} = 7.0$.

A randomized phase II trial

Each of the early stopping rules (2), (5) and (16) is based on an *E*-versus-*S* comparison of event time parameters. An intrinsic problem with comparing data from a single-arm trial of *E* to an historical standard *S*, using either frequentist or Bayesian methods, is that any treatment effect is confounded by between-study effects [17]. This problem, which can be severe when trial effects are large relative to treatment effects, arises either when applying early stopping rules or when using the final data from the trial of *E* to estimate the *E*-versus-*S* treatment difference.

These concerns may motivate a randomized phase II trial of *E* versus *S*. The machinery used in each of the previous sections to conduct a single-arm trial of *E* may be used, with some simple modifications, to construct a randomized trial. To illustrate this in the general case considered in the previous section, we assume priors on $\boldsymbol{\mu}_S$ and $\boldsymbol{\mu}_E$ that are both identical to the noninformative prior on $\boldsymbol{\mu}_E$ specified in the *Establishing priors* section, randomize the 84 patients fairly between *E* and *S*, use the early stopping rule (16) for futility, and also use the additional rule that the trial will be stopped early with *E* declared promising if

$$Pr\left\{\left(\tilde{\mu}_{S,1}^{-1} + \tilde{\mu}_{S,P}^{-1} + \tilde{\mu}_{S,A}^{-1}\right)^{-1} < \left(\tilde{\mu}_{E,1}^{-1} + \tilde{\mu}_{E,P}^{-1} + \tilde{\mu}_{E,A}^{-1}\right)^{-1} \mid data_n\right\} > 0.99 \quad (21)$$

The three possible outcomes are that the trial is stopped early due to futility, the trial is stopped early with *E* declared promising, or the trial runs to completion without either decision. In the third case, the investigators may or may not decide to proceed with a phase III trial of *E* versus *S*. The simulation results are summarized in Table 6. As might be expected, since now a noninformative prior is assumed on $\boldsymbol{\mu}_S$ and there are on average 42 patients per arm, in the null case the PET for futility is smaller and the trial duration is longer than the comparable values for the single-arm trial in Table 5 assuming an informative prior on $\boldsymbol{\mu}_S$. However, a great advantage of randomizing in phase II is that the phase II data may be incorporated into subsequent phase III comparisons, provided that the patient entry criteria are the same [18, 19].

Table 6 Operating characteristics of a randomized trial of Xeloda + Gemzar (X + G) versus 5-FU + Gemzar (5-FU + G), accounting for interval censored progression time and the effect of progression on the hazard of death. The null parameter vector is $\tilde{\boldsymbol{\mu}}_0^{true} = (\tilde{\mu}_{E,1}, \tilde{\mu}_{E,2}, \tilde{\mu}_{E,P}, \tilde{\mu}_{E,A})^{true} = (14, 5.75, 7, 48)$. The alternative parameter vector is $\tilde{\boldsymbol{\mu}}_1^{true} = (32, 12, 12.25, 48)$. These give overall median failure times $\tilde{\mu}_0^{true} = 4.0$ and $\tilde{\mu}_1^{true} = 7.0$

		Early stopping probabilities		No. patients		
5-FU + G	X + G	Futility	Select X + G	5-FU + G	X + G	Trial duration
$\tilde{\boldsymbol{\mu}}_0^{true}$	$\tilde{\boldsymbol{\mu}}_0^{true}$	0.86	0.02	23 30 37	23 30 37	7.5 9.8 12.0
$\tilde{\boldsymbol{\mu}}_0^{true}$	$\tilde{\boldsymbol{\mu}}_1^{true}$	0.10	0.38	33 39 43	33 39 42	11.0 12.7 14.0

Robustness

Thus far, we have assumed exponential or piecewise exponential distributions in order to deal with the complications addressed in the previous three sections. If the event rates are not constant, however, a more complex distribution may be required. In this section, we examine the robustness of the E-IG model based method and CMAP, and also illustrate how a more complex event time model may be implemented. To do this, we first construct a new design for monitoring the overall failure rate, as before, but now assuming that T follows a generalized gamma (GG) distribution. Formally, we assume that T has pdf

$$f(t \mid \alpha, \phi, \beta) = \frac{\phi}{\alpha\Gamma(\beta)}\left(\frac{t}{\alpha}\right)^{\phi\beta-1} \exp\left\{-(t/\alpha)^{\phi}\right\} \qquad (22)$$

where α, ϕ and β are all positive-valued parameters following independent lognormal priors. Thus, six hyperparameters are required to determine the priors. Setting $\phi = 1$ yields a gamma distribution, and setting $\beta = 1$ yields a Weibull distribution. We shall refer to this as the generalized gamma-lognormal (GG-LN) model.

To establish lognormal priors under S, we used the same elicited mean 4 and 95% CI (3.0, 5.2) for $\tilde{\mu}_S$ as before, and also the elicited values 0.125 for the mean and 95% CI (0.05, 0.25) of $\mathcal{F}(12)$ and the elicited mean 0.30 and 95% CI (0.15, 0.60) for $\mathcal{F}(6)$. We solved for the six lognormal hyperparameters using the penalized least squares method of Thall and Cook [20]. This yielded $\log(\alpha_S)$ distributed normal with mean 1.340 and variance 0.091^2, denoted $\alpha_S \sim \text{LN}(1.340, 0.091^2)$, and $\phi_S \sim \text{LN}(-0.127, 0.189^2)$, $\beta_S \sim \text{LN}(0.292, 0.114^2)$. We assumed lognormal priors on α_E, ϕ_E and β_E having the same means but much larger variances. Specifically, we multiplied each prior variance under S by the smallest value so that the historical $\text{var}_S(T) = 2.1$ was inflated at least 10-fold, which yielded the multiplication factor 2.6 since the resulting $\text{var}_E(T) = 25.6$. Thus, we assumed $\alpha_E \sim \text{LN}(1.340, 2.6 \times 0.091^2) = \text{LN}(1.340, 0.146^2)$, and so on. We did not use an arbitrarily large multipier since this yields priors for which T is likely to take on unrealistically large values, in turn producing a design with poor properties. We used an early stopping rule of the same form as (2), with p_L calibrated to give PET = 0.10 when $\tilde{\mu}^{true} = 7.0$ for $T \sim$ GG with variance equal to that under the corresponding exponential distribution. This yielded $p_L = 0.03$. Posteriors were computed using the importance sampling method described earlier.

To study the robustness of the E-IG and GG-LN model based rules and CMAP, we simulated data from a Weibull distribution having $\tilde{\mu}^{true} = 4.0$ or 7.0 and shape parameter $\phi^{true} = 0.8$, 1.0, or 1.2, and also from a lognormal distribution having the given $\tilde{\mu}^{true}$ and variance equal to that of the corresponding exponential distribution. The results are summarized in Table 7. For Weibull data with shape parameter $\phi^{true} = 0.8$, all three methods have inflated PET values, in the range 0.23–0.26, when $\tilde{\mu}^{true} = 7.0$. This case is difficult because the hazard is initially high but monotone decreasing, so early in the trial a method must recognize that $\tilde{\mu}^{true} = 7.0$

Table 7 Robustness. Operating characteristics of the exponential-inverse gamma (E-IG) model-based design, the generalized gamma-lognormal (GG-LN) model-based design, and CMAP for the Xeloda + Gemzar trial, when failure times follow a Weibull distribution with shape parameter $0.80 \le \phi \le 1.2$, or a lognormal distribution

	$\tilde{\mu}^{true} = 4.0$			$\tilde{\mu}^{true} = 7.0$		
Design	PET	No. pats.	Trial duration	PET	No. pats.	Trial duration
			$T \sim$ Weibull, $\phi = 0.8$			
E-IG	0.94	17 25 41	2.7 4.3 7.0	0.25	84 84 84	10.7 13.2 14.6
GG-LN	0.94	12 24 45	2.1 4.0 7.6	0.26	55 84 84	9.2 13.1 14.5
CMAP	0.82	40 52 73	6.5 8.3 11.7	0.23	84 84 84	11.6 13.3 14.6
			$T \sim$ Weibull, $\phi = 1.0$			
E-IG	0.97	20 32 47	3.4 5.3 7.9	0.10	84 84 84	12.2 13.5 14.7
GG-LN	0.98	16 30 46	2.7 5.0 7.6	0.13	84 84 84	12.4 13.6 14.7
CMAP	0.87	42 53 70	6.9 8.5 11.2	0.10	84 84 84	12.4 13.6 14.7
			$T \sim$ Weibull, $\phi = 1.2$			
E-IG	0.99	26 37 51	4.3 6.3 8.5	0.04	84 84 84	12.6 13.7 14.8
GG-LN	0.99	21 33 46	3.6 5.6 7.7	0.08	84 84 84	12.4 13.6 14.8
CMAP	0.91	42 54 67	7.1 8.5 10.8	0.06	84 84 84	12.5 13.6 14.8
			$T \sim$ Lognormal			
E-IG	0.94	36 49 63	6.1 7.9 10.3	0.00	84 84 84	12.8 13.7 14.9
GG-LN	0.98	28 38 50	4.8 6.3 8.1	0.02	84 84 84	12.7 13.7 14.8
CMAP	0.79	50 62 80	7.9 10.1 12.3	0.03	84 84 84	12.7 13.7 14.8

despite the relatively large number of early failures that indicate an unacceptably high event rate. The case $\phi^{true} = 1.0$ is the exponential, studied in Table 1. Here the GG-LN model has a slightly inflated PET = 0.13 when $\tilde{\mu}^{true} = 7.0$. When $\phi^{true} = 1.2$, which produces more later events, all three methods have PET values smaller than the nominal 0.10 when $\tilde{\mu}^{true} = 7.0$. For lognormal data, this effect is more pronounced. In all cases, when $\tilde{\mu}^{true} = 4.0$ CMAP is less safe, with a substantially smaller PET, larger sample size and longer trial duration, compared to the other two methods. Despite the much greater flexibility of the GG-LN model, the original E-IG model based method performs very similarly and is remarkably robust in most of the cases studied.

Since the Xeloda trial data are available, it is of interest to assess the distribution of T. Of the 84 patients, as of 26 July 2005 there were 81 treatment failures (46 disease progressions and 35 SAEs), with sample median 15.7 weeks, virtually identical to the historical median with 5-FU + G. Goodness-of-fit analyses under each of the event time models discussed above showed that the lognormal gave an excellent fit. Denoting the Kaplan–Meier estimate by $\hat{S}_{KM}(t)$ and the normal pdf by Φ, under the lognormal $\Phi^{-1}\{1 - \hat{S}_{KM}(T_i^o)\}$ should be approximately linear in $\log(T_i^o)$. The plot showed good linearity, with $R^2 = 0.962$. For a Bayesian analysis of these data, assuming that $T \sim \text{LN}(\alpha, \sigma^2)$ with independent LN(0,10) priors for μ and $\log(\sigma^2)$, the posterior mean and 95% CI were 15.9 (13.2–19.3) weeks for the median $\tilde{\mu} = e^{\alpha}$ and 23.9 (19.3–30.9) weeks for the mean $\tilde{\mu} = e^{\alpha+\sigma^2/2}$.

Discussion

The examples discussed here were chosen to illustrate how one may deal with particular complications that commonly occur when monitoring event times in clinical trials. There are a several important issues that we have not addressed. A very important problem is patient heterogeneity. While in principle this may be dealt with by including prognostic covariates in the model and monitoring procedure, it raises the practical issues of dealing with possible treatment-covariate interactions [21] and specifying priors. This may be difficult for covariate parameters, and raises the additional issue of using empirical versus elicited priors. Finally, it may be desirable to use multiple stopping rules, e.g., by specifying a separate rule for the SAE rate.

Acknowledgements

Peter Thall's research was partially supported by NCI grant RO1 CA 83932.

References

1. **Prentice R.** Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 1989; **8**: 431–40.
2. **Fleming TR, Prentice RL, Pepe MS and Glidden D.** Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* 1994; **13**: 955–68.
3. **Buyse M and Molenberghs G.** Criteria for validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**: 1014–29.
4. **Begg CB and Leung DHY.** On the use of surrogate end points in randomized trials. *J Royal Statistical Society, Ser A* 2000; **163**: 15–24.
5. **Cowles MK.** Bayesian estimation of the proportion of treatment effect captured by a surrogate marker. *Statistics in Medicine* 2002; **21**: 811–34.
6. **Follman DA and Albert PS.** Bayesian monitoring of event rates with censored data. *Biometrics* 1999; **55**: 603–607.
7. **Rosner GL.** Bayesian monitoring of clinical trials with failure-time endpoints. *Biometrics* 2005; **61**: 239–45.
8. **Cheung YK and Thall PF.** Monitoring the rates of composite events with censored data in phase II clinical trials. *Biometrics* 2002; **58**: 89–97.
9. **Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LO and Benjamin RS.** Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* 2003; **22**: 763–80.
10. **Spiegelhalter DJ, Abrams KR, and Myles JP.** *Bayesian approaches to clinical trials and health care evaluation*. New York: Wiley, 2004.
11. **Thall PF and Simon R.** Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50**: 337–49.
12. **Thall PF and Sung H-G.** Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine* 1998; **17**: 1564–80.
13. **Owen A and Zhou Y.** Safe and effective importance sampling. *Journal of the American Statistical Association* 1999; **95**: 135–40.
14. **Nelder JA and Mead R.** A simplex method for function minimization. *Computer Journal* 1965; **7**: 308.
15. **Chaloner KM, Church T, Louis TA and Matts JP.** Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* 1993; **42**: 341–53.
16. **Kadane JB and Wolfson LJ.** Priors for the design and analysis of clinical trials. In Berry D and Stangl D eds. *Bayesian biostatistics*, New York: Dekker, 1996: 157–84.
17. **Estey EH and Thall PF.** New designs for phase II clinical trials. *Blood* 2003; **102**: 442–48.
18. **Inoue LYT, Thall PF and Berry DA.** Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* 2002, **58**: 823–31.
19. **Liu Q and Pledger G.** Phase 2 and 3 combination designs to accelerate drug development. *Journal of the American Statistical Association* 2005; **100**: 493–502.
20. **Thall PF and Cook JD.** Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 2004; **60**: 684–93.
21. **Thall PF, Sung H-G and Estey EH.** Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *Journal of the American Statistical Association* 2002; **97**: 29–39.