Received: 14 October 2021 | Revised: 2 May 2022 | Accepted: 30 May 2022

DOI: 10.1002/pst.2256

WILEY

MAIN PAPER

A Bayesian piecewise exponential phase II design for monitoring a time-to-event endpoint

Yun Qing[1,2] | **Peter F. Thall**[1] | **Ying Yuan**[1]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

[2]Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Correspondence
Ying Yuan, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
Email: yyuan@mdanderson.org

Funding information
NIH/NCI, Grant/Award Numbers: P50CA127001, P50CA221707, R01 CA261978

Abstract

A robust Bayesian design is presented for a single-arm phase II trial with an early stopping rule to monitor a time to event endpoint. The assumed model is a piecewise exponential distribution with non-informative gamma priors on the hazard parameters in subintervals of a fixed follow up interval. As an additional comparator, we also define and evaluate a version of the design based on an assumed Weibull distribution. Except for the assumed models, the piecewise exponential and Weibull model based designs are identical to an established design that assumes an exponential event time distribution with an inverse gamma prior on the mean event time. The three designs are compared by simulation under several log-logistic and Weibull distributions having different shape parameters, and for different monitoring schedules. The simulations show that, compared to the exponential inverse gamma model based design, the piecewise exponential design has substantially better performance, with much higher probabilities of correctly stopping the trial early, and shorter and less variable trial duration, when the assumed median event time is unacceptably low. Compared to the Weibull model based design, the piecewise exponential design does a much better job of maintaining small incorrect stopping probabilities in cases where the true median survival time is desirably large.

KEYWORDS

Bayesian adaptive design, futility monitoring, go/no-go decision, interim analysis

1 | INTRODUCTION

Most phase II clinical trial designs are based on a binary or ordinal response outcome that is assumed to be observed relatively quickly after the start of treatment. For settings where no effective therapy is available, Gehan[1] proposed a design for single-arm phase IIA cancer trials that aim to detect any anti-disease effect, in terms of response probability, with an experimental treatment, *E*. Trials where a standard therapy *S* exists, but it is desired to obtain an improvement over *S* with *E*, are called "phase IIB." While most phase II trials in oncology are single-arm, randomized phase II trials have been proposed. See Simon, Wittes and Ellenberg,[2] Thall and Sung,[3] or Rubinstein et al.[4] Fleming[5] proposed a group sequential test based single-arm phase IIB design with rules to stop early for futility or efficacy. Simon proposed test-based two-stage phase IIB designs,[6] which have been used widely in practice. Many other phase IIB designs have been proposed, including a two-stage design monitoring both efficacy and toxicity by Bryant and Day,[7] and three-stage

designs for binary response outcomes by Ensign et al.[8] and Chen et al.[9] Zhou et al.[10] proposed the Bayesian optimal phase II (BOP2) design that can handle various categorical endpoints under a uniformed framework.

An important practical issue is that, in many settings, a primary clinical outcome cannot be characterized adequately by a binary or other discrete variable defined so that it can be observed quickly enough to apply a design's outcome-adaptive monitoring rules. Rather, the time, Z, to a particular treatment failure event may be the most important outcome. Examples of failure events in oncology include severe toxicity, disease progression, or death. Other examples are rejection of a transplanted organ in a patient treated with an immunosuppressive agent, or infection for a patient who has received a prophylactic antibiotic. In each case, a smaller observed time to the failure event is less desirable. In such settings, a common practice is to dichotomize Z by defining response as the event $[Z > z^*]$ for a short follow up time z^* so that a monitoring rule based on the probability of this event can be applied. This practice, while convenient, discards important information. As an illustration, suppose that Z = progression-free survival (PFS) time, Y is defined as the indicator that $Z > 28$ days, and a conventional phase II design is constructed based on the probability $\Pr(Y = 1)$. This leads to the problem that, for example, $Y = 1$ if $Z = 29$ days while $Y = 0$ if $Z = 27$ days, despite the fact that these two PFS times differ by a trivial amount. Moreover, in most clinical settings, a much longer follow up time than 28 days is needed to evaluate a PFS time distribution. While one might define $Y = 1$ (response) if $Z > 12$ months, this approach leads to the problem that Y cannot be scored as a "response" until the patient has been followed for 12 months without observing disease progression or death. This renders any outcome-adaptive rule based on Y of no practical use, since most or all patients in a trial are likely to have been accrued before such a stopping rule can be applied. Moreover, two patients who have been followed for one and 10 months, respectively, without experiencing the event both are considered inevaluable, despite the fact that the second patient is much more likely to be a responder.

Much more informative data can be obtained by monitoring Z for each patient over a prespecified follow up period, $[0, t^*]$, for a reasonably large value of t^*. Denoting the time of administrative right censoring by U, $Z^o = \min\{Z, U\}$, and $\delta = I(Z < U)$, a patient's observed outcome data at any follow up time consist of the pair (Z^o, δ), which is the same information used to construct a Kaplan–Meier (KM)[11] estimator. Several single-arm phase II designs for an experimental treatment have been proposed to monitor time-to-event outcomes subject to administrative right-censoring based on (Z^o, δ). Thall, Wooten, and Tannir[12] proposed a design based on a Bayesian exponential-inverse gamma model with an early stopping rule defined by a posterior probability comparing the medians of Z with E and with a historical standard treatment, S. We will refer to this as the E-IG design. Huang et al.[13] proposed an optimal two-stage phase II design based on the Nelson-Aalen estimator[14] of the event-free rate at each time point. Kwak and Jung[15] proposed a two-stage phase II design based on a one-sample version of the log-rank test.[16] Zhou et al.[17] extended the BOP2 design to handle the time-to-event endpoint based on a Bayesian exponential-inverse gamma model.

The E-IG design assumes that Z follows an exponential distribution with mean μ_E, and that μ_E follows a non-informative inverse gamma (IG) prior. We denote this by $Z \mid \mu_E \sim \text{Exp}(\mu_E)$ and $\mu_E \sim \text{IG}(a_E, b_E)$. For this prior, mean $(\mu_E) = b_E/(a_E - 1)$ and $\text{var}(\mu_E) = b_E^2/\left((a_E - 1)^2(a_E - 2)\right)$ with the requirement that $a_E > 2$. Denoting mean failure time with S by μ_S, an informative prior $\mu_S \sim IG(a_S, b_S)$ is assumed. In practice, the two priors are specified to have the same mean, $b_E/(a_E - 1) = b_S/(a_S - 1)$, and thus they differ in that the IG(a_E,b_E) prior is non-informative, to reflect little or no knowledge about $[Z \mid E]$, while the IG(a_S,b_S) prior is informative, to reflect experience from treating patients with S. Many physicians like to think in terms of median rather than mean failure time. Under the exponential distribution, the median of Z is $\widetilde{\mu}_E = \ln(2)\mu_E$ with prior $\widetilde{\mu}_E \sim \text{IG}(a_E, \ln(2)b_E)$. The E-IG design thus may be described equivalently in terms of the median, and its early stopping rule takes the general form

$$\Pr(\widetilde{\mu}_S + \delta < \widetilde{\mu}_E \mid \mathcal{D}_n) < p_L, \tag{1}$$

where $\delta \geq 0$ is a targeted improvement, and $\mathcal{D}_n = \left\{\left(Z_i^o, \delta_i\right), i = 1, \cdots, n\right\}$ denotes the observed outcome data from the first n patients in the trial. The decision cut-off p_L is a fixed parameter that is calibrated to obtain a design with good operating characteristics (OCs), which are computed by simulation. A well-calibrated design should have a small probability of early termination, PET, such as.10, for a desirably large assumed true fixed median $\widetilde{\mu}_E^{\text{true}}$, such as the targeted value $E(\widetilde{\mu}_S) + \delta$, and a large PET for an undesirably small $\widetilde{\mu}_E^{\text{true}}$, such as $E(\widetilde{\mu}_S)$. The OCs include the PET as well as the trial's sample size distribution and trial duration distribution for each value of $\widetilde{\mu}_E^{\text{true}}$ studied.

A limitation of the E-IG design is that, in many settings, assuming an exponential distribution may be an over-simplification. To illustrate this problem, we consider data from a retrospective study of transplant-eligible patients with transformed indolent B-cell lymphoma (Tr-iNHL) in Australia and the United States.[18] For patients who had no

autologous stem cell transplantation in first complete remission (CR1) ($n = 98$), the distribution of Z = PFS time was estimated by the KM method. Plugging in the median obtained from the KM estimate, we estimated the survival distribution of Z assuming that it was exponential. We also estimated the survival distribution assuming that it was Weibull, with this model fit to the study data using the R function *survreg* from the survival package.[19] Figure 1 shows that the estimates of the PFS time distribution obtained from the estimated Weibull distribution are closer to the KM curve than those obtained assuming an exponential distribution. These results suggest that the exponential assumption is not valid for this dataset.

To address this problem in the context of a phase II trial, in this paper we propose an extension of the E-IG design that assumes a more robust model. Our extension, which we call the PE-G design, replaces the Bayesian E-IG model with a Bayesian piecewise exponential (PE) distribution for Z with gamma priors on the hazard parameters in the subintervals of the PE model. As an additional comparator, we also define and evaluate a version of the design based on an assumed underlying Weibull distribution, with truncated normal priors on its shape and scale parameters. We call this the W-TN design.

Aside from the assumed models, all other elements of the PE-G and W-TN designs are the same as those of the E-IG design. All three designs use (1) as an early stopping criterion, with the fixed cutoff p_L calibrated for each design so that, when event times are simulated from a log-logistic distribution with desirably large true median event time $\widetilde{\mu}$, which is the value 6 months in our simulations, the probability of stopping early incorrectly is.10. The Bayesian PE-G model provides much greater flexibility and robustness for a large set of possible true event time distributions, including the Weibull, log-logistic, and gamma, as well as for time-to-event data that do not follow any simple parametric model, such as multi-modal data. The PE model has been applied widely. See, for example, Aslanidou et al.[20]; Aitkin et al.[21]; Breslow[22]; and Ibrahim et al.[23]

The remainder of the paper is organized as follows. In Section 2, we describe the general Bayesian phase II design paradigm, which includes the E-IG design. In Section 3, details of the Bayesian PE-G and W-TN models are provided. In Section 4, we summarize simulations to study and compare the OCs of the PE-G, W-TN, and E-IG designs, and also perform sensitivity analyses including the ones to cohort size. Section 5 gives an illustrative example of a single-arm phase II trial designed and conducted using the PE-G design, followed by general guidelines for using the Shiny application to implement a design. We conclude with a discussion in Section 6.

2 | BAYESIAN PHASE II TRIAL DESIGN PARADIGM

A well-designed single-arm phase IIB trial of an experimental treatment, E, should include an interim futility monitoring rule to stop the trial early if the treatment is found to be ineffective compared to standard therapy based on the observed data. A phase IIB design also may include a superiority rule to stop early if E is seen to be greatly superior to S. Because phase IIB superiority rules seldom are used in practice, however, we will not consider them here. To construct a Bayesian single-arm phase IIB design, an early stopping rule for futility must be calibrated to have a low

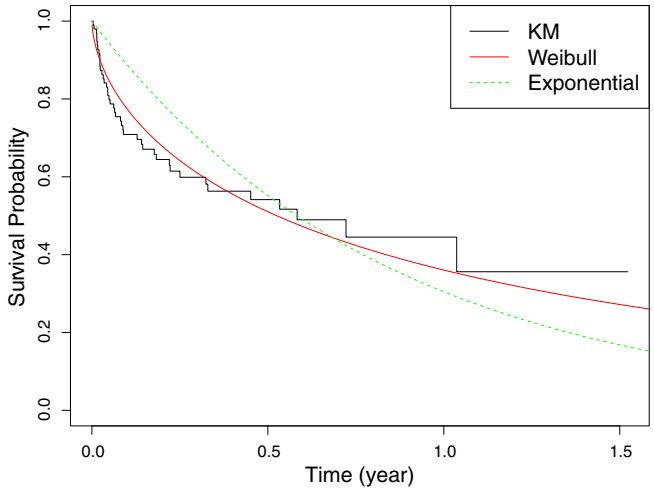

FIGURE 1 Estimation of the survival distribution by the Kaplan–Meier method, or assuming either a Weibull or an exponential distribution

probability of early termination (PET) if E is sufficiently effective in terms of an assumed fixed true parameter value, θ_2^{true}, compared to a smaller value θ_1^{true} which typically is an estimate of the mean of θ obtained from historical data on S. Given a fixed targeted improvement $\delta > 0$, one may define $\theta_2^{\text{true}} = \theta_1^{\text{true}} + \delta$. In general, we will denote by PET(θ^{true}) the probability of early termination by a design if the assumed true parameter value is θ^{true}. For example, consider a phase IIB trial with a binary response outcome, where θ_E and θ_S denote the respective response probabilities with treatments E and S. If a point estimate is $\widehat{\theta}_S = 0.30$ and it is considered desirable to achieve an improvement of $\delta = 0.20$, then the desired fixed target response probability is $\theta_2^{\text{true}} = \widehat{\theta}_S + \delta = 0.50$. The design parameters include p_L in a rule of the form (1), but with the random probabilities θ_E and θ_S in place of $\widetilde{\mu}_E$ and $\widetilde{\mu}_S$. Additional design parameters are the monitoring schedule, which may be specified in terms of a time interval such as 1 month or a cohort size such as 10 patients, and maximum sample size, N. These design parameters should be calibrated to obtain a small PET(θ_2^{true}), such as.10, for $\theta_2^{\text{true}} = \theta_1^{\text{true}} + \delta$, the targeted value. It also is desirable to have large PET(θ_1^{true}) for θ_1^{true} an undesirably low value. Given p_L, this may be obtained by choosing sufficiently large N. If, instead, N is fixed, then p_L may be calibrated by preliminary simulations to obtain a specified small PET($\theta_1^{\text{true}} + \delta$) value.

Under a Bayesian model, prior distributions are assumed for θ_E and θ_S. The prior $p(\theta_S)$ is informative, computed from either historical data or elicited values, whereas the prior $p(\theta_E)$ is non-informative to reflect little or no knowledge about E. Because a single-arm trial treats all patients with E, $p(\theta_S)$ does not change, but the posterior $p(\theta_E \mid \mathcal{D}_n)$ based on data $\mathcal{D}_n$ becomes increasingly more informative as n increases. The early stopping rule (1) may be applied after successive cohorts of a given size, or periodically at scheduled times during the trial. Under this Bayesian design paradigm, there are no hypotheses and no tests. Rather, there is a stopping rule with a monitoring schedule. To validate the design's behavior, values of PET(θ_E^{true}), the achieved sample size distribution, and trial duration distribution are computed via simulation for two or more assumed values of θ_E^{true} and fixed values of the other parameters of the assumed underlying distribution.

This Bayesian phase II design paradigm first was established by Thall and Simon[24] for phase IIB trials with binary outcomes. Extensions have been given, for example, by Thall, Simon, and Estey[25] for multiple discrete outcomes using a multinomial-Dirichlet model, by Jiang et al.[26] to incorporate sample-size dependent decision cut-offs, and by Zhou et al.[10] to accommodate complex multiple outcomes and maximize power by evaluating a set of posterior probabilities computed under a multinomial-Dirichlet model. The PE-G, W-TN, and E-IG designs all follow the same general paradigm given above based on medians, with the only difference being the assumed underling model.

A phase IIB design also may be constructed by considering θ_E to be a fixed unknown constant, casting the early stopping rule in the context of frequentist group sequential hypothesis testing with null hypothesis $H_0 : \theta_E = \theta_1^{\text{true}}$, one-sided alternative $H_1 : \theta_E > \theta_1^{\text{true}}$, and computing the test's power at a value $\theta_2^{\text{true}} > \theta_1^{\text{true}}$. The trial is stopped early if H_0 is accepted interimly. As noted above, this hypothesis test based approach has been taken by Fleming[5] with group sequential tests, by Simon[6] with two-stage tests, and many others. Using this frequentist formulation, setting $\theta_1^{\text{true}} = 0.3$ and $\theta_2^{\text{true}} = 0.5$, in the above example one would define $1 - \text{PET}(0.30)$ to be the test's type I error probability and $1 - \text{PET}(0.50)$ to be the test's power, both computed by also including a final test after the maximum of N patients have been treated and evaluated if the trial is not stopped early.

3 | BAYESIAN PIECEWISE EXPONENTIAL MODEL

Let $n < N$ denote an interim sample size where the monitoring rule (1) will be applied, with corresponding dataset $\mathcal{D}_n$. Among the first n patients, denote the maximum of the observed times to failure or censoring by $T_n = \max\left\{Z_i^o : i = 1, \cdots, n\right\}$, and denote the number of observed failure times by $m_n = \sum_{i=1}^{n} \delta_i$. To specify a PE model, for fixed integer J, we partition the interval $[0, T_n]$ into J disjoint sub-intervals, $[0, d_1), [d_1, d_2), ..., [d_{J-1}, d_J = T_n]$ by using the $1/J, 2/J, \cdots, (J-1)/J$ quantiles of the set of m_n observed failure times. Thus, while J is fixed, T_n and the sub-intervals change stochastically as the interim sample size n increases and the dataset $\mathcal{D}_n$ expands. Let t_j denote the $(100 \times j/J)$th percentile of the m_n observed event times in $\mathcal{D}_n$, and let $t_j' > t_j$ denote the smallest observed event time following t_j. The jth interval cut-off is defined as the statistic $d_j = \left(\frac{1}{2}\right)\left(t_j + t_j'\right)$, where $j = 1, ..., J-1$. The PE model assumes a constant hazard, λ_j, on the jth interval, and that $\lambda_1, ..., \lambda_J$ are independent a priori. We denote $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J)$. To complete the PE model's hazard specification on $[0, \infty)$, since there are no event times observed beyond T_n based on the data from n patients, one may assume any fixed value $\lambda_{J+1} > 0$ for the event rate on (T_n, ∞) since this will not affect the posterior computations for $(\lambda_1, ..., \lambda_J)$.

Under the Bayesian PE-G model, we assume that each λ_j follows a non-informative gamma prior with hyperparameters α_j and β_j, denoted by $\lambda_j \sim \text{Gamma}\left(\alpha_j, \beta_j\right)$. Due to the relatively small sample size of most phase IIB studies,

we recommend $J = 3$ to 5 for applications. When no event is observed interimly, it is possible that either a small number of patients are enrolled, or that the true event rate with E is very low. In such cases, we will continue the trial due to a lack of information or potential effectiveness of E.

Let $\delta_{i,j} = 1$ if the ith patient experiences the event in the jth interval $\left[d_{j-1}, d_j\right)$, and $\delta_{i,j} = 0$ otherwise. Let $e_{i,j}$ denote the observed event time in $\left[d_{j-1}, d_j\right)$, formally

$$e_{i,j} = \begin{cases} d_j - d_{j-1} & \text{if } Z_i^o \geq d_j \\ Z_i^o - d_{j-1} & \text{if } Z_i^o \in \left[d_{j-1}, d_j\right) \\ 0 & \text{otherwise.} \end{cases}$$

Denoting the data vector for event times in the jth interval by $\mathcal{D}_{n,j} = \left(e_{1,j}, \delta_{1,j}, ..., e_{n,j}, \delta_{n,j}\right)$, the likelihood for all of the data $\mathcal{D}_J = \{\mathcal{D}_{n,1}, ..., \mathcal{D}_{n,J}\}$ is

$$L(\mathcal{D}_J|\boldsymbol{\lambda}) = \prod_{j=1}^{J} L\left(\mathcal{D}_{n,j}|\lambda_j\right) = \prod_{j=1}^{J} \prod_{i=1}^{n} \lambda_j^{\delta_{i,j}} \mathrm{e}^{-\lambda_j e_{i,j}}.$$

Due to conjugacy of the exponential distribution and gamma prior, the posterior distribution of the failure rate on the jth sub-interval is

$$\lambda_j \mid \mathcal{D}_{n,j} \sim \text{Gamma}\left(\alpha_j + \Delta_{n,j}, \beta_j + M_{n,j}\right),$$

where $\Delta_{n,j} = \sum_{i=1}^{n} \delta_{i,j}$ is the total number of events and $M_{n,j} = \sum_{i=1}^{n} e_{i,j}$ is the total time to an event or last follow up in the jth time interval. The posterior of the median $\widetilde{\mu}$ is calculated as follows. The distribution of the hazard $\lambda(u)$ is derived from posterior sampling of $\boldsymbol{\lambda}$. Since the survivor function is $\Pr(T > t) = S(t) = \exp\left\{-\int_0^t \lambda(u)\mathrm{d}u\right\}$, we solve for $\widetilde{\mu} = S^{-1}\left(\frac{1}{2}\right)$ numerically.

On the jth sub-interval, we assume that $\lambda_j \sim \text{Gamma}\left(\frac{\widehat{\lambda}_j}{c}, \frac{1}{c}\right)$ a priori, so the hyperparameter $\widehat{\lambda}_j$ is the prior mean, the hyperparameter c quantifies dispersion, and $c\widehat{\lambda}_j$ is the variance. We recommend using a large value, such as $c = 100$, to ensure that the prior is non-informative. We determine $\widehat{\lambda}_1, \cdots, \widehat{\lambda}_J$ as follows. First, we approximate the PE distribution by a Weibull or log-logistic distribution. For the remainder in this section, we illustrate the design using a Weibull distribution with survival function $S(t) = \exp\left(-(t/\beta_W)^{\alpha_W}\right)$, hazard function $\lambda(t) = \alpha_W \beta_W^{-\alpha_W} t^{\alpha_W - 1}$, and median $\beta_W\{\ln(2)\}^{1/\alpha_W}$. The design with the log-logistic distribution can be shown similarly. For two distinct fixed time points, t_1 and t_2, we elicit survival probabilities $\widehat{S}(t_1)$ and $\widehat{S}(t_2)$ from the physician(s) planning the trial, and solve the two resulting equations for the hyperparameter estimates $\widehat{\alpha}_W$ and $\widehat{\beta}_W$. Approximating the PE with the resulting Weibull $\left(\widehat{\alpha}_W, \widehat{\beta}_W\right)$ distribution, we solve for the prior means $\widehat{\lambda}_1, \cdots, \widehat{\lambda}_J$ of $\lambda_1, \cdots, \lambda_J$ using the sub-interval average

$$\widehat{\lambda}_j = \frac{1}{(d_j - d_{j-1})} \int_{d_j-1}^{d_j} \widehat{\lambda}(t)dt = \frac{\widehat{d_j^{\alpha_W}} - \widehat{d_{j-1}^{\alpha_W}}}{\widehat{\beta}_W^{\widehat{\alpha}_W}\left(d_j - d_{j-1}\right)},$$

for each j. Because $\widehat{\lambda}(t) \to \infty$ as $t \to 0$ when the shape parameter is less than 1, $\widehat{\lambda}_1$ is defined as the median of $\widehat{\lambda}(t)$ for the first interval $j = 1$.

4 | SIMULATION STUDY

This section describes a computer simulation study to compare the OCs of the PE-G, W-TN, and E-IG designs. Our simulation study design follows the design structured similarly as Thall et al.,[12] which was based on a trial of

Xeloda + Gemzar for patients with advanced kidney cancer who previously received immunotherapy and either did not respond or relapsed. We assume throughout that the maximum sample size is 104, and that accrual follows a Poisson process with rate two patients per month. Using a cohort size of $c = 26$, three interim analyses applying the early stopping rule are performed, when 1/4 ($n = 26$), 2/4 ($n = 52$), and 3/4 ($n = 78$) patients are enrolled. The event times are generated from a log-logistic distribution, with pdf given by $f(t|\alpha_{LL},\beta_{LL}) = (\beta_{LL}/\alpha_{LL})(t/\alpha_{LL})^{\beta_{LL}-1}\left(1+(t/\alpha_{LL})^{\beta_{LL}}\right)^{-2}$, where α_{LL} and β_{LL} are the scale and shape parameters, respectively. Each simulation scenario is characterized by assumed fixed values of the shape parameter β_{LL}^{true} and median survival time, $\widetilde{\mu}_E^{\text{true}}$, given in Table 1. The value $\widetilde{\mu}_E^{\text{true}} = 6$ months is considered promising, while $\widetilde{\mu}_E^{\text{true}} = 3$ months is considered unacceptably low. We calibrated the probability cutoff p_L of each design so that PET(6) = 0.10 under the assumed log-logistic distribution with shape parameter $\beta_{LL}^{\text{true}} = 0.8$. These assumed true values, used for simulating data, should not be confused with the random parameters in the Bayesian E-IG and PE-G models.

For the E-IG design, we assumed a vague inverse-gamma prior $\widetilde{\mu}_E \sim IG(4.442, 16.326)$, obtained by assuming that median survival time has prior mode 3 months and the prior predictive probability $\Pr(T > 6 \text{ months } |\alpha_E,\beta_E) = 0.365$. For the PE-G design, the priors of $(\lambda_1,...,\lambda_J)$ were determined using the method with log-logistic distribution described in Section 3, calibrated for comparability with the E-IG design so that the expected median survival time = 3 months and the prior predictive probability $\Pr(T > 6 \text{ months } | \boldsymbol{\lambda}) = 0.365$. Similarly, for the Weibull priors, α_W was assumed to follow a truncated normal distribution on $(0,+\infty)$, calibrated to have mean 0.54 and standard deviation 10, and β_W was assumed to follow a truncated normal distribution on $(0,+\infty)$, calibrated to have mean 5.91 and standard deviation 10. MCMC methods were applied to compute posteriors under each model. Each combination of design and scenario was simulated 1000 times.

To avoid confusion, we note that three different roles are played by event time distributions in our simulations. The first is the assumed underlying distribution used by the stopping rule, which here are exponential, piecewise exponential, or Weibull. The second is the distribution used to simulate the data used to calibrate the cutoff p_L of each rule, and we used the log-logistic throughout the simulations reported in Tables 1 and 2 to do this. The third is the assumed true distribution, with given fixed median either 3 or 6, used to simulate the data in each scenario, and for this we used either a log-logistic (Table 1) or a Weibull (Table 2).

Table 1 summarizes the simulation results for data generated from a log-logistic with shape $\beta_{LL}^{\text{true}} = 0.8$. Thus, in the desirable scenario where $\widetilde{\mu}_E^{\text{true}} = 6$, all three methods have the same PET = 0.10, as designed. In most cases, for $\widetilde{\mu}_E^{\text{true}} = 3$, the PE-G design outperforms the E-IG design, with substantially larger PET values, smaller average sample size (number of patients, "No. Pts."), and shorter average trial duration ("Trial Duration"). While the PE-G design has a much

TABLE 1 Operating characteristics of designs with assumed exponential-inverse gamma (E-IG), piecewise exponential-gamma (PE-G), or Weibull-truncated normal (W-TN) model, based on data generated from a log-logistic distribution with median either 3 or 6 and shape parameter between 0.5 and 1.2. Each design was calibrated to have lower cutoff p_L giving PET = 0.10 under a log-logistic distribution with median $\widetilde{\mu}^{\text{true}} = 6$ and shape parameter $\beta_{LL}^{\text{true}} = 0.8$

Scenario		PET			No. Pt. (SD)			Trial duration (SD)		
β_{LL}^{true}	$\widetilde{\mu}_E^{\text{true}}$	E-IG	PE-G	W-TN	E-IG	PE-G	W-TN	E-IG	PE-G	W-TN
0.8	3	0.62	0.82	0.84	58.7 (36.8)	54.1 (29.1)	51.7 (28.6)	31.5 (21.8)	28.1 (16.8)	26.8 (16.6)
	6	0.10	0.10	0.10	96.3 (23.3)	97.3 (20.7)	97.1 (21.2)	53.4 (14.6)	53.9 (13.2)	53.8 (13.5)
0.5	3	0.46	0.60	0.73	68.8 (38.6)	68.2 (32.8)	57.6 (32.8)	37.6 (23.1)	36.4 (19.7)	30.3 (19.2)
	6	0.17	0.13	0.20	90.9 (29.1)	95.2 (23.3)	90.2 (28.3)	50.4 (17.9)	52.7 (14.6)	49.7 (17.2)
0.7	3	0.54	0.75	0.79	63.1 (38)	58.6 (31.1)	55 (30.7)	34.2 (22.7)	30.8 (18.3)	28.7 (17.8)
	6	0.12	0.10	0.11	95 (24.8)	97 (21.2)	96.4 (22.2)	52.7 (15.4)	53.7 (13.4)	53.3 (13.9)
1.0	3	0.76	0.93	0.93	50.4 (32.5)	45.6 (23.8)	44.3 (23.9)	26.6 (19.1)	23.3 (13.4)	22.6 (13.5)
	6	0.07	0.09	0.09	98.5 (19.9)	98.3 (19.3)	98 (20)	54.7 (12.7)	54.4 (12.5)	54.3 (12.8)
1.1	3	0.83	0.96	0.96	45.6 (29.3)	41.8 (20.7)	40.5 (21)	23.8 (17.1)	21.2 (11.4)	20.5 (11.7)
	6	0.05	0.07	0.08	100 (16.9)	99 (18.1)	98.7 (18.8)	55.5 (11.1)	54.9 (11.9)	54.8 (12.2)
1.2	3	0.90	0.99	0.98	41.8 (25.5)	39.3 (17.9)	37.6 (18.7)	21.5 (14.7)	19.8 (9.7)	19 (10.4)
	6	0.05	0.07	0.08	100.3 (16.3)	99.1 (17.9)	98.5 (19.4)	55.7 (10.9)	54.9 (11.7)	54.7 (12.5)

TABLE 2 Operating characteristics of designs with assumed exponential-inverse gamma (E-IG), piecewise exponential-gamma (PE-G), or Weibull truncated normal (W-TN) model, based on data generated from a Weibull distribution with median 3 or 6 and shape parameter between 0.4 and 1.3. Each design was calibrated to have lower cutoff p_L giving PET = 0.10 under a log-logistic distribution with median $\widetilde{\mu}^{\text{true}} = 6$ and shape parameter $\beta_{LL}^{\text{true}} = 0.8$

Scenario		PET			No. Pt. (SD)			Trial duration (SD)		
α_W^{true}	$\widetilde{\mu}_E^{\text{true}}$	E-IG	PE-G	W-TN	E-IG	PE-G	W-TN	E-IG	PE-G	W-TN
0.4	3	0.53	0.72	0.85	64 (38.3)	61.8 (31.5)	50.4 (28.5)	34.8 (22.9)	32.5 (18.5)	25.9 (16.1)
	6	0.17	0.13	0.26	90.8 (29.2)	95.3 (23.1)	86.9 (30.3)	50.3 (17.9)	52.7 (14.4)	47.8 (18.2)
0.6	3	0.81	0.95	0.98	47.1 (30.4)	45.6 (22.6)	41.9 (19.7)	24.6 (17.8)	23.2 (12.6)	21.1 (10.6)
	6	0.12	0.12	0.17	94.9 (24.8)	96.3 (21.6)	94.4 (23.4)	52.6 (15.4)	53.3 (13.7)	52.1 (14.6)
0.8	3	1.00	1.00	1.00	33.6 (14)	36.2 (14.7)	33.9 (12.9)	16.8 (7.6)	18.2 (7.8)	17 (7)
	6	0.10	0.12	0.15	96.9 (22)	97 (20.6)	95.3 (22.3)	53.8 (13.8)	53.7 (13.1)	52.6 (14)
1.0	3	1.00	1.00	1.00	30.4 (10)	31.5 (11.2)	29.8 (9.2)	15.2 (5.6)	15.7 (6.2)	14.9 (5.2)
	6	0.11	0.12	0.17	97.3 (20.1)	97 (20.3)	94.6 (22.7)	53.8 (12.7)	53.6 (12.9)	52.1 (14.1)
1.1	3	1.00	1.00	1.00	28.9 (8.2)	30.1 (9.5)	28.2 (7.2)	14.5 (4.7)	15.1 (5.4)	14.2 (4.3)
	6	0.11	0.10	0.14	97.9 (18.8)	98.2 (18.7)	95.6 (22.2)	54.1 (12.1)	54.3 (12.1)	52.8 (13.9)
1.3	3	1.00	1.00	1.00	27.6 (6.2)	27.9 (6.8)	26.8 (4.6)	13.8 (3.8)	14 (4.1)	13.5 (3.3)
	6	0.20	0.09	0.16	95.8 (18.4)	98.8 (17.6)	95.1 (22.4)	52.6 (12.1)	54.7 (11.7)	52.5 (14.1)

larger PET than the E-IG design, differences in mean sample size and trial duration are small. The standard deviations of the sample size and trial duration are smaller for the PE-G design. Compared to the W-TN design, the PE-G design generally had similar or smaller PET values for both $\widetilde{\mu}_E^{\text{true}} = 6$ and 3. The most prominent exception is the case where $\beta_{LL}^{\text{true}} = 0.5$, with the W-TN model giving PET(3) = 0.73 and PET(6) = 0.20, compared to PET(3) = 0.60 and PET (6) = 0.13 for the PE-G. Thus, the W-TN model gives a larger PET for either $\widetilde{\mu}_E^{\text{true}}$ value, so calibrating it for good performance across all scenarios does not appear to be possible.

Table 2 shows the OCs of the three designs under several Weibull distributions. The PET(6) values of the PE-G design are reasonably close to 0.10 in all scenarios as α_W^{true} is varied, but the E-IG and W-TN designs have much larger PET(6) values in many cases. For $\widetilde{\mu}_E^{\text{true}} = 3$, compared to the E-IG, the PE-G design again has much larger PET values for shape parameter values $\alpha_W = 0.4$ and 0.6, and both methods have PET ≥0.99 for larger values of α_W. The W-TN design generally has both larger PET(6) and larger PET(3) values than the PE-G design, so it appears that calibrating the W-TN design using data generated from a log-logistic gives a design with undesirable properties. The achieved sample sizes and trial durations reflect the PET values in each case.

We evaluated the sensitivity of the proposed PE-G design to (i) the cohort size (or equivalently the number of interim looks), and (ii) the number of partition intervals for the piecewise exponential model. Tables 3 and 4 provide the results, showing that the PE-G design is generally robust to these two factors.

5 | ILLUSTRATION

In this section, we illustrate how a trial may be designed and conducted using either the PE-G or E-IG design. Suppose that the PE-G design includes up to $R - 1$ interim applications of the rule (1) when the proportions $1/R$, $2/R$, ..., $(R - 1)/R$ of the planned maximum sample size of N patients have been enrolled. To calibrate p_L, in the illustration of the PE-G method, we use the estimated Weibull parameters $\widehat{\alpha}_W$ and $\widehat{\beta}_W$, for each candidate p_L value, simulate the trial assuming that the Z_i's follow a Weibull $\left(\widehat{\alpha}_W, \widehat{\beta}_W\right)$ distribution, compute PET($\widetilde{\mu}^{\text{true}}$), and iterate this until a value of p_L is obtained that gives the pre-specified desired small value of PET($\widetilde{\mu}^{\text{true}}$) = 0.10, where $\widetilde{\mu}^{\text{true}} = 6.5$ is the desirably large value in this application. The value of p_L for the E-IG method was calibrated similarly, to give PET($\widetilde{\mu}^{\text{true}}$) = PET (6.5) = 0.10, simulating the data from a Weibull.

Suppose from historical data on standard treatment, S, the mean of the median survival time is 2.5 months and the PFS rate at 6.5 months is 32.7%. Assuming that PFS time follows a Weibull distribution, we solved for the shape and

TABLE 3 Operating characteristics of the designs with exponential-inverse gamma (E-IG) or piecewise exponential-gamma (PE-G) models, when patients are monitored in cohorts of size 2 or 4, under log-logistic distributions with median 3 or 6 and shape parameter $\beta_{LL} = 0.8$

Scenario	PET		No. Pts.		SD (No. Pts.)		Trial duration		SD (Trial duration)	
Cohorts of size 2										
$\widetilde{\mu}_E^{\text{true}}$	E-IG	PE-G	E-IG	PE-G	E-IG	PE-G	E-IG	PE-G	E-IG	PE-G
3	0.54	0.69	57.6	59.0	44.12	38.19	31.3	31.2	25.43	21.75
6	0.10	0.10	94.7	94.8	27.86	27.88	52.6	52.6	16.45	16.51
Scenario	**PET**		**No. Pts.**		**SD (No. Pts.)**		**Trial duration**		**SD (Trial duration)**	
Cohorts of size 4										
$\widetilde{\mu}_E^{\text{true}}$	E-IG	PE-G	E-IG	PE-G	E-IG	PE-G	E-IG	PE-G	E-IG	PE-G
3	0.54	0.74	58.1	55.9	43.63	36.79	31.6	29.4	25.17	20.83
6	0.10	0.10	94.9	94.9	27.40	27.47	52.7	52.7	16.31	16.32

TABLE 4 Operating characteristics of PE-G design, when number of subinterval partition (J) is 4 or 5, based on data generated from a log-logistic distribution with median either 3 or 6 and shape parameter between 0.5 and 1.2. The design was calibrated to have lower cutoff p_L giving PET = 0.10 under a log-logistic distribution with median $\widetilde{\mu}_E^{\text{true}} = 6$ and shape parameter $\beta_{LL}^{\text{true}} = 0.8$

Scenario		PET		No. Pt. (SD)		Trial duration (SD)	
β_{LL}^{true}	$\widetilde{\mu}_E^{\text{true}}$	$J = 4$	$J = 5$	$J = 4$	$J = 5$	$J = 4$	$J = 5$
0.8	3	0.83	0.84	55.2 (28.3)	54.7 (28.1)	28.6 (16.3)	28.3 (16.2)
	6	0.10	0.10	97.7 (19.8)	98 (19.1)	54.1 (12.8)	54.2 (12.4)
0.5	3	0.55	0.56	72.1 (32.4)	72.2 (32.3)	38.6 (19.5)	38.6 (19.4)
	6	0.11	0.11	96.8 (21.4)	97.1 (20.7)	53.6 (13.6)	53.8 (13.2)
0.7	3	0.74	0.76	60.5 (30.7)	59.9 (30.2)	31.7 (18)	31.3 (17.6)
	6	0.09	0.10	98.2 (19.4)	97.8 (19.4)	54.4 (12.6)	54.2 (12.6)
1.0	3	0.94	0.95	45.9 (23.2)	45.7 (22.7)	23.3 (12.8)	23.1 (12.5)
	6	0.10	0.11	98.1 (19.2)	97.5 (19.8)	54.3 (12.4)	53.9 (12.9)
1.1	3	0.98	0.97	41.9 (20.2)	41.9 (20.1)	21.1 (11)	21.1 (10.8)
	6	0.09	0.11	98.3 (19)	97.3 (20.1)	54.4 (12.3)	53.8 (12.9)
1.2	3	0.99	0.99	39.2 (17.8)	39.3 (18.3)	19.7 (9.6)	19.7 (9.9)
	6	0.09	0.12	98.5 (18.5)	97.1 (20.5)	54.5 (12.1)	53.6 (13.1)

scale parameters 0.50 and 5.2025. A maximum of 78 patients will be enrolled with an assumed accrual rate of 3 patients per month, and the last patient will be followed for an additional 6 months. We conduct interim analyses for cohorts of $c = 26$, when 1/3 ($n = 26$) and 2/3 ($n = 52$) of the patients have been enrolled. The prior of the PE-G model is based on the historical data using the Weibull distribution, as described in Section 3. The prior of the median PFS for E-IG is assumed to be IG(3,5), which has mean 2.5 and variance 6.25. We calibrated p_L to obtain PET = 0.10 at $\widetilde{\mu}_E^{\text{true}} = 6.5$ for both the PE-G and E-IG design. Given these calibrated p_L values, we calculated the PET($\widetilde{\mu}_E^{\text{true}}$) = PET(2.5) = 0.85 for the PE-G design and 0.72 for E-IG design. Thus, in this case, the PE-G design is much more desirable than the E-IG design in terms of PET for an undesirably small $\widetilde{\mu}_E^{\text{true}}$.

To mimic a real-world scenario for trial conduct, we simulated a dataset with underlying median time-to-disease progression or death following a Weibull distribution with shape parameter 0.5 and scale parameter 5.2025, which corresponds to a median PFS of 2.5 months. Based on the interim data from 52 patients, the KM estimate of the median PFS is 2.41. The posterior mean of the median PFS under the E-IG model is 4.66. For the PE-G model, the estimated median PFS is 2.41, derived from the posterior mean of the empirical hazards in the sub-intervals. Figure 2 shows that

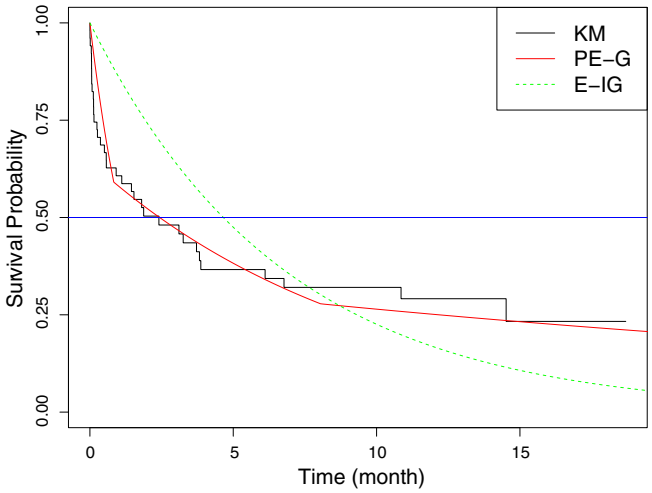

FIGURE 2 Estimation of the progression-free survival distribution by the Kaplan–Meier method, or assuming either a piecewise exponential or exponential distribution

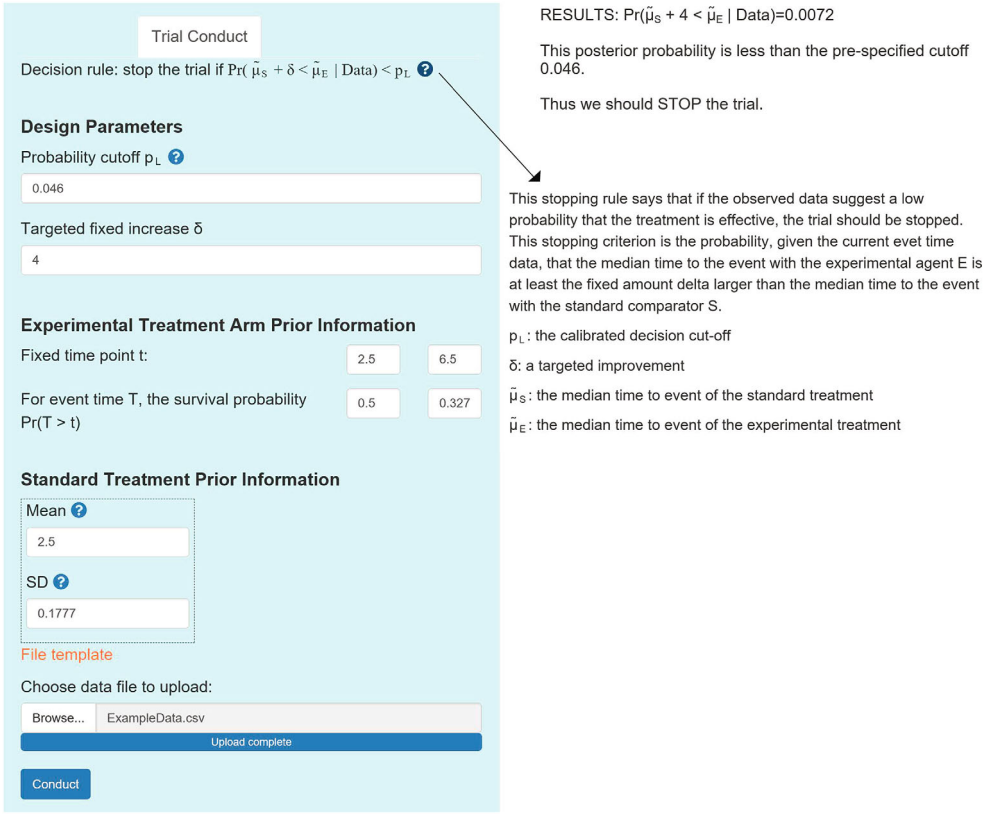

FIGURE 3 User interface for the Shiny application of PE-G method

the PFS probability estimates under the PE-G model are much closer to the KM estimates than the estimates under the E-IG model. The early stopping thresholds p_L are 0.046 for the PE-G design and 0.0015 for the E-IG design. The posterior probability $\Pr(\tilde{\mu}_S + \delta < \tilde{\mu}_E \mid \mathcal{D}_n)$ is 0.0072 for the PE-G design and 0.0218 for the E-IG design. Therefore, the trial would be stopped early at $n = 52$ by the PE-G design, but not by the E-IG design. Since the treatment is ineffective if the underlying median PFS of $\tilde{\mu}_E^{\text{true}} = 2.5$, stopping the trial early is the correct decision in this case. This example illustrates that the PE-G and E-IG designs may behave very differently, with different median estimates and different go/no-go decisions.

We are developing a Shiny app to facilitate the application of the proposed design. Figure 3 shows the user interface of the software under development, which will be deployed at www.trialdesign.org.

6 | DISCUSSION

We have proposed a Bayesian phase II design for a single-arm trial to monitor one event time. The design is motivated by the desire to obtain a more robust version of the design proposed by Thall, Wooten, and Tannir,[12] which is based on an exponential-inverse gamma model. By replacing this model with a piecewise exponential with gamma priors on the hazard parameters in the model's subintervals, a design with much more desirable OCs is obtained. We also defined and studied a third design, based on an assumed Weibull distribution with parameters following truncated normal priors. Simulations show that, compared to the E-IG, the PE-G based design has much larger PET values in many cases where the true median event time is small and it is desirable to stop the trial for futility. Additionally, the new design obtains smaller sample sizes and shorter, less variable trial durations when the assumed median event time is unacceptably low. Compared to the Weibull based design, the PE-G design does a better job of controlling incorrect stopping probabilities to be small when the true median event time is desirably large. Computer software, including a user interface, will be provided for implementing the new design.

While the proposed PE-G design is useful for the simple setting that it addresses, there are several limitations. It accommodates one time-to-event outcome, so trials with multiple event times or some combination of discrete and continuous outcomes would require a more complex design. Our results suggest that assuming a piecewise exponential likelihood for each event time in such a setting is likely to produce a robust design. A final point is that we have assumed patients are homogeneous. Constructing a generalization that accommodates patient heterogeneity with subgroup-specific stopping rules is an area for future research.

ACKNOWLEDGMENTS

Peter Thall's research was supported by NIH/NCI grant R01 CA261978. Ying Yuan's research was supported by NIH/NCI grant P50CA221707 and P50CA127001.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Peter F. Thall https://orcid.org/0000-0002-7293-529X
Ying Yuan https://orcid.org/0000-0003-3163-480X

REFERENCES

1. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chronic Dis*. 1961;13(4):346-353.
2. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treat Rep*. 1985;69(12):1375-1381.
3. Thall PF, Sung HG. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat Med*. 1998;17(14):1563-1580.
4. Rubinstein L, Crowley J, Ivy P, Leblanc M, Sargent D. Randomized phase II designs. *Clin Cancer Res*. 2009;15(6):1883-1890.
5. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*. 1982;38(1):143-151.
6. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10(1):1-10.
7. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*. 1995;51(4):1372-1383.
8. Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. *Stat Med*. 1994;13(17):1727-1736.
9. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Stat Med*. 1997;16(23):2701-2711.
10. Zhou H, Lee JJ, Yuan Y. BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Stat Med*. 2017;36(21):3302-3314.
11. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
12. Thall PF, Wooten LH, Tannir NM. Monitoring event times in early phase clinical trials: some practical issues. *Clin Trials*. 2005;2(6):467-478.
13. Huang B, Talukder E, Thomas N. Optimal two-stage phase II designs with long-term endpoints. *Stat Biopharm Res*. 2010;2(1):51-61.
14. Nelson W. Hazard plotting for incomplete failure data. *J Qual Technol*. 1969;1(1):27-52.

15. Kwak M, Jung SH. Phase II clinical trials with time-to-event endpoints: optimal two-stage designs with one-sample log-rank test. *Stat Med*. 2014;33(12):2004-2016.
16. Finkelstein DM, Muzikansky A, Schoenfeld DA. Comparing survival of a sample to that of a standard population. *J Natl Cancer Inst*. 2003;95(19):1434-1439.
17. Zhou H, Chen C, Sun L, Yuan Y. Bayesian optimal phase II clinical trial design with time-to-event endpoint. *Pharm Stat*. 2017;36(19):776-786.
18. Chin CK, Lim KJ, Lewis K, et al. Autologous stem cell transplantation for untreated transformed indolent B-cell lymphoma in first remission: an international, multi-centre propensity-score-matched study. *Br J Haematol*. 2020;191(5):806-815.
19. Therneau T. A Package for Survival Analysis in R. *R package version 3.2–13*; 2021. https://CRAN.R-project.org/package=survival
20. Breslow N. Covariance analysis of censored survival data. *Biometrics*. 1974;30(1):89-99.
21. Aitkin M, Laird N, Francis B. A reanalysis of the Stanford heart transplant data. *J Am Stat Assoc*. 1983;78(382):264-274.
22. Aslanidou H, Dey DK, Sinha D. Bayesian analysis of multivariate survival data using Monte Carlo methods. *Can J Stat*. 1998;26(1):33-48.
23. Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. Springer; 2001.
24. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*. 1994;50(2):337-349.
25. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med*. 1995;14(4):357-379.
26. Jiang L, Yan F, Thall PF, Huang X. Comparing Bayesian early stopping boundaries for phase II clinical trials. *Pharm Stat*. 2020;19(6):928-939.

How to cite this article: Qing Y, Thall PF, Yuan Y. A Bayesian piecewise exponential phase II design for monitoring a time-to-event endpoint. *Pharmaceutical Statistics*. 2022;1-11. doi:10.1002/pst.2256