Contemporary Clinical Trials Communications 36 (2023) 101225

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Research paper

Sample size determination and evaluation for two-stage adaptive designs of single arm clinical trials based on median event time test

Yeonhee Park [*], Yi Chen

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, United States of America

ARTICLE INFO

Keywords:
Median
Phase II
Single-arm
Survival

ABSTRACT

Clinical trials play a critical role in drug development which involves a series of phases and requires a significant amount of time and effort. Efficient clinical trial designs are necessary to investigate a new drug. Investigators strongly desire to use the time-to-event endpoint as the primary endpoint for Phase II studies, which evaluates the therapeutic efficacy of the new drug, with the hypothesis that the new drug improves the median survival time. The one-sample log-rank test has been used for single-arm Phase II trials, but it generally requires more samples. Recently, the median event time test was proposed to provide a simple, straightforward decision rule, which compares the observed median survival time for the new drug with the threshold, which is determined through the numerical search. We improve the computation of the method for the two-stage design of single-arm clinical trials based on the median event time test. By utilizing the large sample theory of order statistics, we provide the explicit formulas to calculate the sample size for the first and second stages and propose the testing procedure. The performance of the proposed method is evaluated through simulations and a trial example.

1. Introduction

Drug development is a complex process that involves a series of phases, requiring a significant time and effort. Clinical trials play a critical role investigating new drugs. The therapeutic efficacy of the new drug is assessed in Phase II and then further confirmed with usually hundreds or thousands of patients in Phase III. This confirmation is necessary to establish whether the experimental treatment provides a longer overall survival compared to a placebo or standard control. Phase II clinical trial designs have traditionally relied heavily on a binary response endpoint. For example, in cancer clinical trials, a binary indicator of having a complete response or a partial response to therapy is used. According to Response evaluation criteria in solid tumors (RECIST) forms, a complete response is defined as the removal of all target lesions, and a partial response is defined as a reduction of at least 30% in the total number of target lesions [1]. However, it may not be desirable to use the response rate as a surrogate for the desired survival time, which is a gold standard endpoint for Phase III studies. For example, the tumor shrinkage may not be measurable in rapidly lethal cancers or may not be available in uveal cancer [2]. Moreover, it is critical to avoid unexpected failures in Phase III studies after successful Phase II results. To address these issues, it is recommended to evaluate the new drugs based on the time-to-event endpoint in Phase II studies.

The one-sample log-rank test has been commonly used to compare the survival distribution of the experimental treatment with a historical control, without making parametric survival assumptions. It was first introduced by Breslow [3]. Tu and Gross [4] derives a Bartlett-type correction for the two-sided one-sample log-rank test, and Sun et al. [5] derives the correction to the one-sided one-sample log-rank test based on Edgeworth expansion. Kerschke et al. [6] improves the one-sample log-rank test based on the transformation of the underlying counting process martingale to correct the conservativeness. Several studies, including Sun et al. [5],Finkelstein et al. [7],Wu [8], and Schmidt et al. [9], provide sample size formulas to achieve the desired power for the one-sample log-rank test. The one-stage procedure and two-stage designs based on the one-sample log-rank test are proposed by Sun et al. [5],Kwak and Jung [10], and Belin et al. [11]. They compare the survival curves or hazard rates to evaluate the promising nature of the new drug. Instead of comparing entire survival curves, another approach to assess the efficacy of drugs is to compute the survival probability at a clinically meaningful time point, e.g., one-year survival probability [2,12–14].

Our previous work uses the parametric survival distributions to develop the median event time test for single-arm Phase II trials whose primary endpoint is time-to-event [15]. It extends the idea of Simon's

[*] Corresponding author.
E-mail address: ypark56@wisc.edu (Y. Park).

https://doi.org/10.1016/j.conctc.2023.101225
Received 26 April 2023; Received in revised form 4 September 2023; Accepted 23 October 2023
Available online 29 October 2023

two-stage design [16] for a time-to-event endpoint based on the median event time test. In clinical practice, clinical investigators and practitioners report the median survivals as a measure of the time-to-event endpoints. The median information is easily accessible from historical trials, and it provides a straightforward understanding of the therapeutic efficacy of the drugs based on median survival. For clinical purposes, the median event time test is useful as it allows for a simple and straightforward interpretation. It tests the improvement of median survival and can be easily implemented by comparing the observed median survival time for the new drug with a prespecified threshold. Park [15] determines sample size and threshold of clinical trial design based on median event time test through the numerical search for the given target type I and II error rates.

This paper aims to develop a new method for the median event time test, which provides explicit formulas for determining the sample size in two-stage designs. Compared with the existing literature, we make three major contributions. First, we derive explicit sample size formulas for the two-stage design in order to attain the desired power of the one-sided median event time test with a significance level. The explicit formulas eliminate the computational burden associated with our earlier work. Second, we determine the decision rule for the median event time test with the optimal choice of design parameters, which minimizes the expected total sample size under the null hypothesis. Third, we provide a user-friendly shiny application which is freely available and easy to use.

This article is organized as follows. Section 2 describes the two-stage design of single-arm trials whose primary endpoint is time-to-event and provides explicit formulas determining the sample size based on the median event time test. We also provide the practical considerations and software which is freely available to use the proposed method for the trial implementation. The performance of the proposed method is investigated through simulations in Section 3. We illustrate the application of the method with a real trial example in Section 4 and provide the concluding remarks in Section 5.

2. Sample size determination

Consider a two-stage design of single-arm Phase II clinical trials whose clinical endpoint is time-to-event. For the first stage, n_1 patients are accrued, and an interim analysis is performed to determine the go/no-go of the trial when the total accrual reached n_1 patients. If the trial continues, we determine the required sample size, say n_2, for the second stage. In the second stage, the n_2 patients are enrolled while the first stage patients are being followed for the efficacy evaluations at the final analysis. The final analysis will be performed after the follow-up of all n patients, where $n = n_1 + n_2$.

Let $Y_1, \ldots, Y_n$ be random variables from an exponential distribution with mean μ. Then, the median ϕ of the exponential distribution is $\mu \log 2$. We want to test the null hypothesis $H_0 : \phi = \phi_0$ versus the alternative hypothesis $H_a : \phi = \phi_1$ based on the one-sided median event time test (METT) at a significance level of α to attain the power $\geq 1 - \beta$. The null and alternative values are specified by clinicians. For example, in intervention clinical trials, the null value of ϕ_0 is the median time of the standard drug and the alternative value of ϕ_1 is the expected median time from the intervention (for improvement). Let Z_1 and Z_2 be test statistics for a two-stage design at the interim and final analyses, respectively. The test statistic at the interim, denoted by Z_1, is calculated by the observed median survival time of the first stage, and the test statistic at the final analysis, denoted by Z_2, is calculated by the observed median survival time of all enrolled patients after the follow-up. Let $\hat{S}_n(t)$ be the estimate of the survival function based on the sample of size n. The observed median survival time based on the sample of size n, denoted by M_n, is the time at which 50% of individuals are expected to have survived, i.e., $M_n = \min\{t \in T : \hat{S}_n(t) \leq 0.5\}$. Then, we have $Z_1 = M_{n_1}$ and $Z_2 = M_{n_2}$ based on the estimate of the survival function at interim and final analyses, respectively. The nonparametric estimator with the Kaplan–Meier estimator can be used to calculate the test statistics Z_1 and Z_2. We assume that a censoring rate is less than 50% at the interim so that we obtain reasonably precise estimates of the median survival time Z_1. The assumption can be relaxed to handle the case when we do not observe Z_1 (See practical considerations below). Let t_1 and t_2 be the threshold to be compared with the observed median survival at the interim and final analysis, respectively. Then, METT compares the observed median survivals Z_1 and Z_2 with the corresponding thresholds t_1 and t_2 at the interim and final analysis, respectively. We stop the trial for futility at the interim analysis if $Z_1 \leq t_1$. Otherwise (i.e., $Z_1 > t_1$), the trial continues to enroll the second-stage patients. At the final analysis, we compare Z_2 with t_2 to argue that the experimental treatment shows sufficient improvement in efficacy. When $Z_2 > t_2$, the experimental treatment is considered promising. This framework follows Simon's two-stage design but extends Simon's two-stage design to a time-to-event endpoint based on the median event time test.

Park [15] determines the decision rule of METT by using the empirical search, which requires a lot of computation time. In this work, to save the computational cost, we use the true survival distribution to determine the decision rule for the two-stage design and show that the proposed decision rule is not sensitive to censoring information, provided the median survival time is observed at interim. Specifically, we use the theoretical results in Mosteller [17] for the observed survival time. The median of $Y_1, \ldots, Y_n$ follows asymptotical normal distribution with mean ϕ and variance $0.25/[n\{f(\phi)\}^2]$, where $f(y) = (\log 2/\phi)\exp(-y \log 2/\phi)$ denotes the density function of the exponential distribution with median ϕ.

Let α_1 and β_1 be given such that $1 - \alpha_1$ denotes the probability of correct decision at the interim analysis under the null hypothesis, i.e., $\Pr(Z_1 \leq t_1 | H_0) = 1 - \alpha_1$, and $1 - \beta_1$ denotes the probability of correct decision at the interim analysis under the alternative hypothesis, i.e., $\Pr(Z_1 > t_1 | H_a) = 1 - \beta_1$. Since Z_1 asymptotically follows normal distribution with mean ϕ_0 and variance $0.25/[n_1\{f(\phi_0)\}^2]$ under the null hypothesis, we obtain

$$t_1 = \frac{0.5 z_{\alpha_1}}{\sqrt{n_1} f(\phi_0)} + \phi_0, \tag{1}$$

where z_{α_1} denotes the critical value of the standard normal distribution at α_1, i.e., $\Pr(Z \geq z_{\alpha_1}) = \alpha_1$, where Z is a standard normal variable. Therefore, by the normality of Z_1 under H_a, we have

$$n_1 = \left(\frac{f(\phi_1)}{f(\phi_0)} z_{\alpha_1} + z_{\beta_1}\right)^2 \left(\frac{0.5}{f(\phi_1)(\phi_1 - \phi_0)}\right)^2, \tag{2}$$

where z_{β_1} denotes the critical value of the standard normal distribution at β_1.

Since the type I error rate is at most α and the expected power is at least $1 - \beta$, we want to have $\Pr(Z_2 > t_2, Z_1 > t_1 | H_0) \leq \alpha$ and $\Pr(Z_2 > t_2, Z_1 > t_1 | H_a) \geq 1 - \beta$. We notice that $\Pr(Z_2 > t_2, Z_1 > t_1 | H_0) \geq \Pr(Z_2 > t_2 | H_0) - \Pr(Z_1 \leq t_1 | H_0)$ and $\Pr(Z_2 > t_2, Z_1 > t_1 | H_a) \geq \Pr(Z_2 > t_2 | H_a) - \Pr(Z_1 \leq t_1 | H_a)$, because

$$\Pr(Z_2 > t_2 \text{ or } Z_1 > t_1) = \Pr(Z_2 > t_2) + \Pr(Z_1 > t_1) - \Pr(Z_2 > t_2, Z_1 > t_1) \leq 1.$$

Then, it suffices to have $\Pr(Z_2 > t_2 | H_0) = \alpha$ and $\Pr(Z_2 > t_2 | H_a) - \Pr(Z_1 \leq t_1 | H_a) = 1 - \beta$ in order to attain $\Pr(Z_2 > t_2, Z_1 > t_1 | H_0) \leq \alpha$ and $\Pr(Z_2 > t_2, Z_1 > t_1 | H_a) \geq 1 - \beta$. Setting the quantity of $\Pr(Z_2 > t_2 | H_0)$ to be α, we obtain

$$t_2 = \frac{0.5 z_{\alpha}}{\sqrt{n} f(\phi_0)} + \phi_0. \tag{3}$$

Setting the quantity of $\Pr(Z_2 > t_2 | H_a) - \Pr(Z_1 \leq t_1 | H_a)$ to be $1 - \beta$, we obtain

$$n_2 = \left(\frac{f(\phi_1)}{f(\phi_0)} z_{\alpha} + z_{\beta_2}\right)^2 \left(\frac{0.5}{f(\phi_1)(\phi_1 - \phi_0)}\right)^2 - n_1, \tag{4}$$

where $\beta_2 = \beta - \beta_1$. The details of the derivation of (1)–(4) are provided in Appendix A.

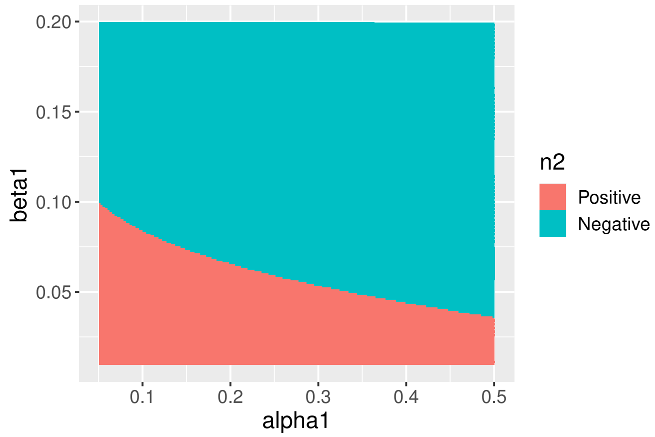


Fig. 1. Sign of the sample size for the second stage when $\alpha \le \alpha_1 \le 0.5$ and $0 < \beta_1 < \beta$. The sample size n_2 is calculated from (4) when $\phi_0 = 3, \phi_1 = 6$, $\alpha = 0.05$, and $\beta = 0.2$.

Optimal choice of α_1 **and** β_1 To calculate the sample size using (1)–(4), α_1 and β_1 should be specified. We choose optimal values of α_1 and β_1 in that they minimize the expected total sample size under the null hypothesis. In the proposed two-stage design, the sample size is random. Specifically, the sample size for the first stage is fixed as n_1, but the sample size for the second stage, denoted by $\tilde{n}_2$, is random. If we observe at interim $Z_1 \le t_1$, the trial is terminated due to the futility and $\tilde{n}_2 = 0$. If we observe at interim $Z_1 > t_1$, then the trial continues to enroll n_2 patients for the second stage and $\tilde{n}_2 = n_2$. Then, the expected total sample size under the null hypothesis is $\mathrm{E}(n|H_0) = n_1 + \mathrm{E}(\tilde{n}_2|H_0) = n_1 + n_2\alpha_1$, where n_1 and n_2 are functions of α_1 and β_1. We notice that the threshold t_1 for the first stage lies between ϕ_0 and ϕ_1. It implies that z_{α_1} is nonnegative. Therefore, the optimal values of α_1 and β_1 are identified by minimizing the expected total sample size $\mathrm{E}(n|H_0)$ over $\alpha \le \alpha_1 \le 0.5$ and $0 < \beta_1 < \beta$. Since n_2 is nonnegative, the values of α_1 and β_1 which lead to $n_2 < 0$, i.e., $\left\{(\alpha_1, \beta_1) : (z_\alpha - z_{\alpha_1}) f(\phi_1)/f(\phi_0) < z_{\beta_1} - z_{\beta - \beta_1}\right\}$, are excluded to consider possible expected total sample size. For example, when the median progression-free survival (PFS) of the standard drug is 3 months and the investigational drug is expected to improve PFS to 6 months, we have $\phi_0 = 3$ and $\phi_1 = 6$. Using the target error rates $\alpha = 0.05$ and $\beta = 0.2$, we observe the sign of n_2, which is displayed in Fig. 1. The green colored region of (α_1, β_1) which yields the negative value of n_2 is not considered to obtain the optimal values of α_1 and β_1.

Practical considerations and preservation of type I error rate The null and alternative median values could be specified by adopting clinical background and knowledge (e.g., pilot results or historical trials). For the given hypothetical median values, the proposed method determines the sample size and thresholds assuming that the test statistics Z_1 and Z_2 exist, i.e., the median survivals at interim and final analysis are observed. However, in practice, depending on the effect size and accrual rate, the median survival would not be observed because of a few events observed. In addition, some superior drugs cure patients and it could be infeasible to obtain a median survival even after sufficient follow-up. In these cases, the smallest fitted survival of the Kaplan–Meier curve is larger than 50%. Therefore, when Z_1 does not exist, we do not want to stop the trial for futility based on n_1 patients but rather continue the trial to enroll the patients for the second stage. This allows us to have more evidence to evaluate the therapeutic efficacy at the end of the trial.

When the median Z_1 based on the first n_1 patients is not observed at the interim, we do not make any decisions for early stopping, which implies that we do not spend any errors at the interim. Thus, we suggest replacing t_2 with the threshold denoted by t^*. In other words, when Z_1 does not exist, we continue to enroll additional n_2 patients. At the final analysis, we compare the observed median Z_2 based on all $n = n_1 + n_2$ enrolled patients to the threshold t^* defined as

$$t^* = \frac{0.5 z_\alpha}{\sqrt{n^*} f(\phi_0)} + \phi_0, \tag{5}$$

where

$$n^* = \left(\frac{f(\phi_1)}{f(\phi_0)} z_\alpha + z_\beta\right)^2 \left(\frac{0.5}{f(\phi_1)(\phi_1 - \phi_0)}\right)^2. \tag{6}$$

The quantities t^* and n^* in (5) and (6) are obtained from a single-stage design based on METT yielding the power $1 - \beta$ at the significance level of α.

We evaluate the performance through simulations and observe that the approach using t^* works well to control error rates for faster accrual rates. In practice, we recommend running preliminary simulations to report the operating characteristics of the designs using threshold t_2 or t^* for the suggested accrual rate of the trial. The preliminary studies will be helpful to choose the decision rule which preserves the overall type I error rate of the trial design and avoid unnecessarily losing power from using a relatively stringent cutoff.

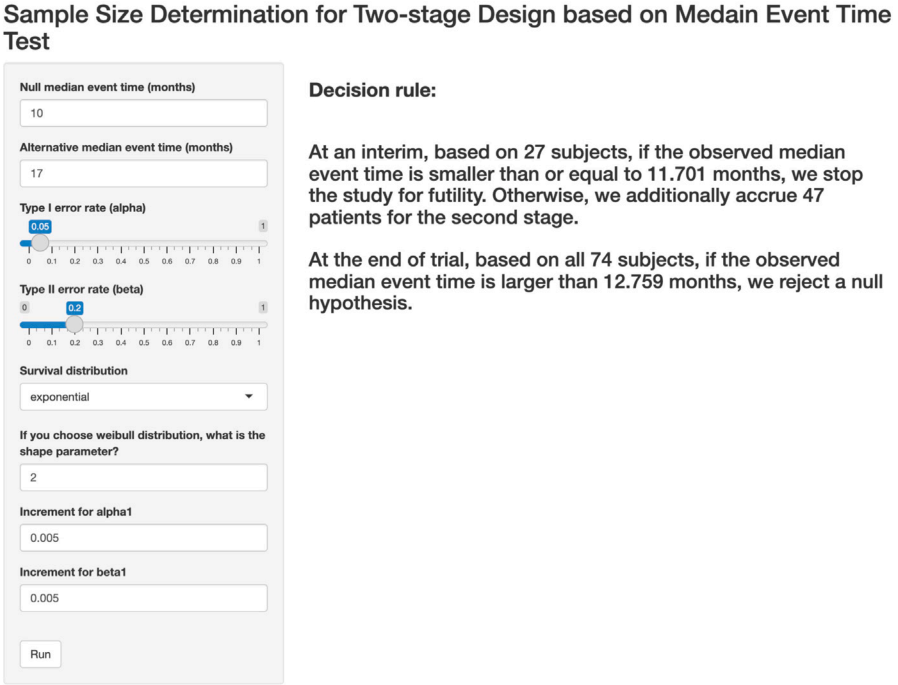
Software and trial implementation To facilitate the use of the two-stage design based on METT, we develop a shiny application, which is freely available at https://yeonhee.shinyapps.io/METTSS/. It allows users to obtain a decision rule for the trial. Fig. 2 shows the screenshot of the application website. As seen in Fig. 2, users need to specify the following parameters

- *Null median event time* denoted by ϕ_0 in months
- *Alternative median event time* denoted by ϕ_1 in months
- *Type I error rate* denoted by α
- *Type II error rate* denoted by β
- *Survival distribution*. Users choose one of the parametric survival distributions among exponential, uniform, and Weibull distributions based on knowledge or background of the study.
- *If you choose Weibull distribution, what is the shape parameter?* Users specify a positive value for the shape parameter of the Weibull distribution. The shape parameter is known as the Weibull slope. For a shape parameter being larger than 1, the survival time follows the Weibull distribution with an increasing hazard. For a shape parameter smaller than 1, it follows the Weibull distribution with a decreasing hazard. For a shape parameter of 1, it follows the Weibull distribution with a constant hazard, which is equivalent to an exponential distribution. The value of the shape parameter can be specified by historical trial data or literature information.
- *Increment for alpha1*. It provides a sequence from α to 0.5 by the specified increment for searching the optimal value of α_1. The default is 0.005.
- *Increment for beta1*. It provides a sequence from 0 to β by the specified increment for searching the optimal value of β_1. The default is 0.005.

After hitting the button "Run" with the specifications, the application returns the decision rule in the right-side panel.

3. Evaluation of the two-stage design

The proposed method was conducted for the hypothesis testing $H_0 : \phi = \phi_0$ versus $H_a : \phi = \phi_1$, where ϕ_0 denotes the null median survival time and ϕ_1 denotes the alternative median survival time. We set the type I error rate $\alpha = 0.05$ and the type II error rate $\beta = 0.2$ for the hypothesis testing. Design parameters such as α_1 and β_1 were numerically searched with the increment 0.005 over the α_1 between 0.05 and 0.5 and β_1 between 0.001 and 0.199. Using formulas (1)–(4), the required sample sizes n_1 and n_2 for stages 1 and 2, respectively, and



Sample Size Determination for Two-stage Design based on Medain Event Time Test

Null median event time (months)
10

Alternative median event time (months)
17

Type I error rate (alpha)
0.05

Type II error rate (beta)
0.2

Survival distribution
exponential

If you choose weibull distribution, what is the shape parameter?
2

Increment for alpha1
0.005

Increment for beta1
0.005

Decision rule:

At an interim, based on 27 subjects, if the observed median event time is smaller than or equal to 11.701 months, we stop the study for futility. Otherwise, we additionally accrue 47 patients for the second stage.

At the end of trial, based on all 74 subjects, if the observed median event time is larger than 12.759 months, we reject a null hypothesis.

Fig. 2. METTSS shiny application to determine sample size and threshold for the two-stage design based on METT.

Table 1
Decision rules for two-stage design considering hypothesis testing of median survivals ϕ_0 months versus ϕ_1 months based on target error rates of $\alpha = 0.05$ and $\beta = 0.2$.

ϕ_0	ϕ_1	Decision rule				
		n_1	t_1	n_2	t_2	t^*
3	5	28	3.501	54	3.786	4.073
3	6	17	3.692	29	4.050	4.453
3	7	13	4.219	26	4.140	4.780
8	14	25	9.557	44	10.285	11.164
8	17	15	10.728	33	10.740	12.245
10	17	27	11.701	47	12.759	13.706

Table 2
Computing time to determine decision rules for two-stage design considering hypothesis testing of median survivals ϕ_0 months versus ϕ_1 months based on target error rates of $\alpha = 0.05$ and $\beta = 0.2$.

ϕ_0	ϕ_1	Numeric search [15]	Asymptotic theory (Proposed method)
3	5	13.50 h	0.053 s
3	6	4.17 h	0.053 s
3	7	69.91 min	0.053 s
8	14	8.39 h	0.051 s
8	17	2.82 h	0.123 s
10	17	11.92 h	0.046 s

the thresholds t_1 and t_2 at the interim and final analysis, respectively, are provided in Table 1. For example, when the null median ϕ_0 is 3 months and the alternative median ϕ_1 is 6 months, the proposed method yields $n_1 = 17$, $t_1 = 3.692$, $n_2 = 29$, and $t_2 = 4.050$. It implies that the two-stage design accrues $n_1 = 17$ patients for the first stage. At the interim based on data of the first 17 patients, we stop the study for futility if the observed median Z_1 is less than or equal to the first threshold $t_1 = 3.692$ months. Otherwise, i.e., if $Z_1 > t_1$, we additionally accrue $n_2 = 29$ patients for the second stage, which results in a total sample size of $n = n_1 + n_2 = 46$ patients. At the end of the trial, based on all $n = 46$ patients data, if the observed median Z_2 is larger than $t_2 = 4.050$ months, the null hypothesis is rejected and we claim that the experimental treatment is sufficiently promising. When Z_1 is not observed, we continue to the second stage enrolling $n_2 = 29$ patients and we replace the threshold $t_2 = 4.050$ with $t^* = 4.453$ determined by the formulas (5)–(6) for final analysis.

We measured the computing time to determine the decision rules for the two-stage design and provided it in Table 2. The computing time was measured with Apple M1 Ultra and 128 GB memory. Using the proposed method, the decision rules in Table 1 were obtained very quickly with less than 0.2 s. However, using the numeric search proposed in Park [15], it took several hours to obtain the decision rules in Table 2 of Park [15], which was computationally expensive. Table 2 shows that the proposed method improves the computation to determine the decision rules for the two-stage design.

We investigated the operating characteristics of the proposed design through simulations. We considered both null and alternative scenarios for each hypothesis testing described in Table 1. Null scenario indicates the case where there is no improvement from the experimental treatment, i.e., the true median for the experimental treatment is the same as the null hypothesized value ϕ_0, and the alternative scenario indicates that experimental treatment improves the median survival time and the true median is the same as the alternative hypothesized value ϕ_1. True survivals were generated from the exponential distribution whose median is the null or alternative hypothesized value described in Table 1 for the null or alternative scenario, respectively. We assumed that patients arrived according to a Poisson process with the accrual rate of 1.04 patients per month, and we continued follow-up for 24 months after the last patient was enrolled. We replicated 10000 times to obtain simulation results such as the rejection probability (denoted by $\hat{\alpha}$ or $1-\hat{\beta}$ under null or alternative scenarios, respectively), the expected sample size (denoted by EN_0 or EN_a under null or alternative scenarios, respectively), and the probability of early termination (denoted by PET_0 or PET_a under null or alternative scenarios, respectively).

Table 3 shows that both type I and II error rates were preserved at the target rates of $\alpha = 0.05$ and $\beta = 0.2$, respectively. Since the decision rules were derived by the asymptotic method, the overall error rates in Table 1 were not necessary to be the same as the exact test's results. For example, $\hat{\alpha} = 0.039$ or $\hat{\beta} = 1 - 0.869 = 0.131$ was acceptable because they were smaller or equal to the target error rates (or even target error

Table 3
The operating characteristics of the two-stage design considering hypothesis testing of median survivals ϕ_0 months versus ϕ_1 months based on target error rates of $\alpha = 0.05$ and $\beta = 0.2$. The results of $\hat{\alpha}$ and $1 - \hat{\beta}$ have two columns: the results without parenthesis are obtained by using t_2 at the final analysis while the results in the parenthesis are obtained by using the threshold t^* at the final analysis when Z_1 does not exist at the interim.

ϕ_0	ϕ_1	$\hat{\alpha}$	$1 - \hat{\beta}$	EN$_0$	EN$_a$	PET$_0$	PET$_a$
3	5	0.043(0.043)	0.864(0.864)	43.6	75.9	0.711	0.112
3	6	0.051(0.051)	0.869(0.868)	25.5	42.8	0.708	0.110
3	7	0.041(0.039)	0.849(0.844)	18.6	35.3	0.786	0.144
8	14	0.046(0.045)	0.841(0.836)	38.0	63.0	0.705	0.137
8	17	0.039(0.019)	0.805(0.786)	25.4	41.8	0.684	0.187
10	17	0.044(0.042)	0.835(0.827)	41.7	67.4	0.688	0.140

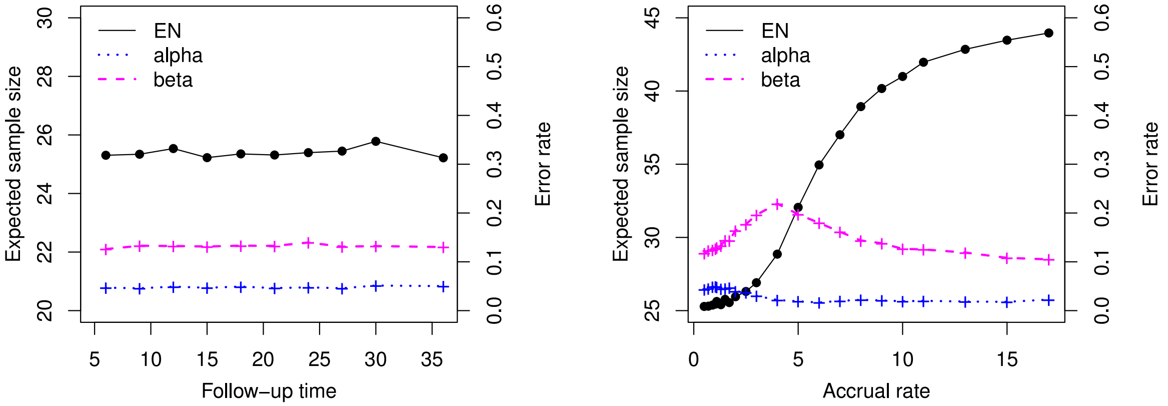


Fig. 3. The plot of operating characteristics of the proposed design using the threshold t^* at the final analysis when the median Z_1 does not exist at the interim for the testing $\phi_0 = 3$ versus $\phi_1 = 6$ months. The expected sample size EN is obtained under the null scenario (i.e., EN$_0$). The left panel shows the result when the follow-up time is varied, and the right panel shows the results when the accrual rate is varied.

rates plus some reasonable margin bound). We also observed that EN$_0$ is much smaller than the total planned number of patients, i.e., $n = n_1 + n_2$, and EN$_a$ is close to n. Specifically, the two-stage design used on average 53.5% of the total sample size when the drug is not good enough, which saves a lot of patients treated with subtherapeutic drugs and could avoid experiencing unnecessary adverse events. This resulted from the effective futility monitoring to stop the trial earlier when the drug is not good enough. As seen in Table 3, PET$_0$ is actually relatively higher to stop the trial earlier. On average, the probability of early stopping when the drug is not good enough is 0.714. It implies that our futility monitoring works well across all scenarios.

As sensitivity analyses, we evaluated the sensitivity of the proposed two-stage design to follow-up time and accrual rate. Back to the setting of the second scenario in Table 1, i.e., $H_0 : \phi_0 = 3$ and $H_a : \phi_1 = 6$, we varied the follow-up time $6, 9, 12, 15, 18, 21, \mathbf{24}, 27, 30, 36$ months and accrual rate 0.5, $0.7, 0.9, \mathbf{1.04}$, $1.1, 1.3, 1.5, 1.7$, $2, 2.5, 3, 4$, $5, 6, 7, 8, 9$, $10, 11$, $13, 15, 17$ patients per month to reflect the different clinic situations. We replicated 10,000 times and reported the operating characteristics, such as EN$_0$, $\hat{\alpha}$, and $\hat{\beta}$, of the two-stage design using the identified decision rule with $n_1 = 17$, $t_1 = 3.692$, $n_2 = 29$, $t_2 = 4.050$, and $t^* = 4.453$. The results are summarized in Fig. 3. It shows that the proposed two-stage design is robust to the follow-up time but impacted by the accrual rate. When the accrual rate is high (i.e., patients are accrued quickly), it is less likely to observe the median at interim because of few events, and we do not perform an early stopping test but continue the trial to the second stage. Specifically, in the figure, we observed that the black solid line indicating the expected sample size under the null hypothesis suddenly increases when the accrual rate is higher than 2 patients per month and then reaches the close value to the total sample size of 46. As described in Fig. 3, even though the proposed design was impacted by the accrual rate, error rates are generally controlled in most cases. Specifically, when we varied the accrual rate, the overall type I error rate ranges from 0.0166 to 0.0496, and the overall type II error rate ranges from 0.105 to 0.219.

Further investigations for accrual distributions and true survival distributions

As a further investigation of the proposed method, we considered three different accrual distributions from Table 3: (1) uniform accrual which ranges from 0 to 3 (2) nonhomogeneous Poisson process with slow accrual rate (i.e., 0.5 patients per month) at the first stage and then fast accrual rate (i.e., 1.5 patients per month) at the second stage (3) nonhomogeneous Poisson process with fast accrual rate (i.e., 3 patients per month) at the first stage and slow accrual rate (i.e., 0.5 patients per month) at the second stage. Using the decision rules in Tables 1 and 4 shows the simulation results with the different accrual distributions. We included the results of Table 3 assuming the homogeneous accrual rate following the Poisson process for the comparison of the performance. We observed that the proposed method works well to preserve the error rates at the target rates regardless of the accrual distributions.

We learned several lessons from the simulation studies. As seen in Table 4, the proposed design using the decision rule (n_1, t_1, n_2, t_2) works well once the median survival exists at the interim. However, when the median survival does not exist at the interim, the proposed design using the practical consideration, i.e., (n_1, t_1, n_2, t_2, t^*), works well. For example, suppose that we consider the hypothesis test of the null median of 8 months versus the alternative median of 17 months, and the accrual rate is fast in the first stage and then low in the second stage. We could have trouble observing the median survival at interim because of the fast accrual rate, and we observed the overall type I error rate of 0.057 when (n_1, t_1, n_2, t_2) was used. However, following the practical suggestion using the cutoff t^* when the median was not observed at interim, both type I and II error rates are controlled (i.e., $\hat{\alpha} = 0.012$ and $\hat{\beta} = 0.135$). This illustrates how to use the proposed method for the practice. Using the cutoff t^* is helpful but it might be stringent to apply to all cases. As seen in Table 3, the rejection probability decreases by replacing t_2 with t^*. Therefore, we strongly

Table 4
Simulation results considering different accrual distributions when the hypothesis testing of median survivals ϕ_0 months versus ϕ_1 months is conducted based on target error rates of $\alpha = 0.05$ and $\beta = 0.2$.

ϕ_0	ϕ_1	$\hat{\alpha}$	$1-\hat{\beta}$	EN_0	EN_a	PET_0	PET_a
Accrual rate follows from homogeneous Poisson process							
3	5	0.043	0.864	43.6	75.9	0.711	0.112
3	6	0.051	0.869	25.5	42.8	0.708	0.110
3	7	0.041	0.849	18.6	35.3	0.786	0.144
8	14	0.046	0.841	38.0	63.0	0.705	0.137
8	17	0.039	0.805	25.4	41.8	0.684	0.187
10	17	0.044	0.835	41.7	67.4	0.688	0.140
Accrual rate follows from uniform distribution							
3	5	0.046	0.855	42.7	75.6	0.728	0.118
3	6	0.048	0.879	25.4	43.1	0.709	0.099
3	7	0.039	0.866	18.1	35.7	0.802	0.126
8	14	0.047	0.857	37.4	63.8	0.719	0.118
8	17	0.035	0.828	23.0	42.5	0.757	0.166
10	17	0.043	0.851	40.8	68.2	0.707	0.124
Accrual rate is slow in the first stage and then fast in the second stage							
3	5	0.043	0.863	43.5	75.9	0.712	0.113
3	6	0.044	0.878	25.4	43.0	0.709	0.103
3	7	0.040	0.866	18.3	35.7	0.794	0.126
8	14	0.045	0.855	37.4	63.7	0.718	0.121
8	17	0.039	0.830	22.9	42.6	0.759	0.163
10	17	0.049	0.854	41.1	68.3	0.699	0.121
Accrual rate is fast in the first stage and then slow in the second stage							
3	5	0.048	0.849	43.7	75.3	0.710	0.124
3	6	0.049	0.854	25.9	42.3	0.693	0.127
3	7	0.039	0.844	20.4	35.1	0.714	0.149
8	14	0.049	0.822	45.1	62.5	0.544	0.149
8	17	0.057	0.909	37.0	45.5	0.334	0.076
8	17[a]	0.012	0.865	37.5	45.4	0.318	0.078
10	17	0.049	0.838	52.9	68.0	0.450	0.129

[a] Obtained from using practical consideration with t^* instead of t_2 when the median Z_1 does not exist at the interim.

suggest biostatisticians run simulations with the determined decision rules, (n_1, t_1, n_2, t_2) and $(n_1, t_1, n_2, t_2, t^*)$, to ensure that which decision rule is appropriate for the given accrual rate, hypothetical values, and target errors.

The proposed method is applicable to any parametric survival distribution with the density function $f(y)$ in order to determine the decision rules using (1)–(4). We considered non-exponential survival assumptions such as uniform survivals or Weibull survivals with an increasing hazard function. Using the simulation setting used for Table 3, the simulation results are provided in Table 5. The proposed method works well to control error rates regardless of the parametric distribution of the survivals.

Comparison with existing methods

Lastly, we compared the proposed two-stage design (called METTSS) with existing two-stage designs such as (1) a restricted KJ design, r-KJ, which tests for survival curves based on one-sample log-rank test [11]; (2) a two-stage design minimizing expected sample size called OES [14]; (3) a two-stage design minimizing the expected total study length called OETSL [14]. The r-KJ uses one-sample log-rank test statistics proposed by Kwak and Jung [10], and both OES and OETSL use the normalized Z-statistic to test and determine decision rules at each stage. The methods of r-KJ, OES, and OETSL require specifying the clinically meaningful time point. In our simulations, we used 6 months. The null and alternative values of the 6 months survival probability were determined by the survival distribution with hypotheses ϕ_0 and ϕ_1, respectively. For the proposed two-stage design METTSS, we considered three designs assuming exponential, uniform, and Weibull survivals. The decision rules for each parametric survival assumption are present in Tables 1 and 5, and Appendix (See Table 7). We assumed the accrual rate is 1.04 patients per month and the follow-up time is 24 months.

Table 6 provides the comparison results in terms of EN_0 and PET_0 for hypothesis tests. Note that r-KJ does not restrict to certain survival distributions but specifies the hazard ratio and 6-month survival probability from the exponential distribution. Both OES and OETSL assume the Weibull survivals. In most cases, METTSSW used a smaller expected sample size and stopped the trial early for futility when therapeutic intervention is not effective. We observed that METTSSW, OES, and OETSL yielded very similar values of EN_0, but METTSSW led to smaller PET_0 than OETSL. Specifically, when $\phi_0 = 3$, OETSL stopped the trial early with an average of 74.7% while METTSSW led to smaller PET_0 of 55% on average. This is because METTSS designs passed the interim monitoring when the median did not exist, which could lead to smaller PET_0 when the true median is large enough and the sample size is small. So, the quantitative difference in PET_0 is not much meaningful to compare. We also found that the benefit of using METTSS becomes greatly increasing as the null median is large. Under METTSS assuming exponential survivals, on average 29.2, 35.2, 73.7 patients were expected to be tested for $\phi_0 = 3, 6, 10$, respectively; under METTSS assuming uniform survivals, on average 14.3, 18.8, 38.1 patients were expected to be tested for $\phi_0 = 3, 6, 10$, respectively; under METTSSW, on average 8.8, 11.7, 22.7 patients were expected to be tested for $\phi_0 = 3, 6, 10$, respectively; under r-KJ, on average 20.0, 39.4, 122.5 patients were expected to be tested for $\phi_0 = 3, 6, 10$, respectively; under OES, on average 8.4, 19.2, 51.7 patients were expected to be tested for $\phi_0 = 3, 6, 10$, respectively; and under OETSL, on average 8.7, 19.5, 51.8 patients were expected to be tested for $\phi_0 = 3, 6, 10$, respectively.

4. An example trial

To illustrate the application of the sample size determination for a two-stage design, we consider a study with ClinicalTrials.gov Identifier NCT00871923. The goal of the study was to determine the effect of combining whole brain radiation therapy with Tarceva (Erlotinib hydrochloride) in patients with brain metastases from Non-Small Cell Lung Cancer. The historical controls are patients who have not met criteria (1) Karnofsky Performance Status Scale (KPS) < 70 with an average survival of 2.3 months or (2) KPS > 70 and age < 65 years to live approximately 7.1 months. They have an average survival of 4.2 months, which indicates a median survival of 2.9 months. Investigators assumed that survival times are exponentially distributed and the therapeutic approach will improve the mean survival by 43% to 6 months, which indicates a median survival of 4.2 months. The study started on March 26, 2009, and was completed to collect final data for the primary outcome measure on December 4, 2019. The study results in ClinicalTrials.gov show a median survival of 11.8 months with a 95% confidence interval (7.4, 19.1) based on 40 patients treated with Tarceva after radiation therapy completed.

We redesigned the trial with a two-stage design using METT. Using our shiny application with $\phi_0 = 2.9$ and $\phi_1 = 11.8$, we obtained $n_1 = 6$, $t_1 = 5.556$, $n_2 = 19$, and $t_2 = 4.276$ based on type I error rate of 0.05 and type II error rate of 0.2. At the interim, based on 6 subjects, if the observed median event time is smaller than or equal to 5.556 months, we stop the study for futility. Otherwise, we additionally accrue 19 patients for the second stage. At the end of the trial, based on all 25 subjects, if the observed median event time is larger than 4.276 months, we reject a null hypothesis.

The study NCT00871923 was designed with a uniform accrual rate of 1.7 patients per month and a 9-month follow-up. Using the information on the accrual rate and follow-up, we conducted simulations to evaluate the two-stage design. It led to the overall type I error rate of 0.027 and power of 82.4% to detect the improvement of median survivals from the intervention.

Table 5
Decision rules and operating characteristics of two-stage design when survivals are assumed to follow non-exponential distributions. The hypothesis testing of median survivals ϕ_0 months versus ϕ_1 months is conducted based on target error rates of $\alpha = 0.05$ and $\beta = 0.2$.

	ϕ_0	ϕ_1	Decision rule				
			n_1	t_1	n_2	t_2	t^*
Uniform	3	5	14	3.432	22	3.822	4.077
	3	6	8	3.666	14	4.052	4.424
	3	7	6	4.169	13	4.132	4.745
	8	14	12	9.558	21	10.291	11.102
	8	17	7	10.491	15	10.805	12.161
	10	17	13	11.914	25	12.668	13.678
Weibull	3	5	7	3.465	13	3.796	4.073
	3	6	4	3.854	10	3.951	4.453
	3	7	3	3.986	6	4.187	4.780
	8	14	6	9.552	12	10.237	11.164
	8	17	4	10.641	7	10.862	12.245
	10	17	7	11.671	11	12.797	13.577

	ϕ_0	ϕ_1	Operating characteristics					
			$\hat{\alpha}$	$1 - \hat{\beta}$	EN_0	EN_a	PET_0	PET_a
Uniform	3	5	0.033	0.821	20.7	32.9	0.693	0.139
	3	6	0.029	0.803	12.4	19.6	0.688	0.171
	3	7	0.024	0.787	9.8	16.4	0.71	0.202
	8	14	0.031	0.786	20.5	29.2	0.595	0.180
	8	17	0.032	0.851	16.8	20.1	0.350	0.124
	10	17	0.030	0.788	24.5	33.5	0.541	0.180
Weibull	3	5	0.036	0.804	11.6	17.9	0.644	0.165
	3	6	0.034	0.810	8.4	12.2	0.559	0.176
	3	7	0.045	0.863	6.3	8.3	0.446	0.117
	8	14	0.043	0.866	14.6	16.8	0.284	0.102
	8	17	0.051	0.932	9.8	10.7	0.175	0.040
	10	17	0.047	0.872	15.3	17.1	0.242	0.082

Table 6
Comparison of the expected sample size and probability of early termination under the null hypothesis with existing designs. The hypothesis testing of median survivals ϕ_0 months versus ϕ_1 months is conducted based on target error rates of $\alpha = 0.05$ and $\beta = 0.2$. "Exp" indicates an exponential distribution.

ϕ_0	ϕ_1	EN_0					
		METTSS			r-KJ	OES	OETSL
		Exp	Uniform	Weibull			
3	5	43.6	20.7	11.6	29.7	9.8	10.6
3	6	25.5	12.4	8.4	17.6	6.9	8.4
3	7	18.6	9.8	6.3	12.8	–*	7.0
6	9	66.5	32.6	18.1	74.5	27.6	27.8
6	11	32.6	16.7	11.0	37.5	18.5	18.8
6	13	22.5	14.5	9.8	25.5	15.8	16.2
6	15	19.1	11.4	7.9	19.9	14.9	15.3
10	13	153.8	74.6	40.5	250.0	–**	–**
10	15	67.3	33.9	21.2	114.4	64.7	64.8
10	17	41.9	24.7	15.4	72.4	48.7	48.8
10	19	31.8	19.3	13.5	53.2	41.6	41.7
ϕ_0	ϕ_1	PET_0					
		METTSS			r-KJ	OES	OETSL
		Exp	Uniform	Weibull			
3	5	0.71	0.69	0.64	0.56	0.40	0.72
3	6	0.71	0.69	0.56	0.54	0.18	0.76
3	7	0.79	0.71	0.45	0.54	–[a]	0.76
6	9	0.66	0.69	0.55	0.57	0.63	0.70
6	11	0.71	0.58	0.34	0.57	0.58	0.70
6	13	0.70	0.50	0.28	0.56	0.58	0.71
6	15	0.68	0.36	0.19	0.55	0.56	0.71
10	13	0.65	0.66	0.66	0.57	–[b]	–[b]
10	15	0.65	0.65	0.38	0.57	0.66	0.68
10	17	0.68	0.53	0.24	0.57	0.65	0.68
10	19	0.70	0.36	0.15	0.59	0.65	0.69

[a] Not applicable because all patients can be accrued before the event time.

[b] Not applicable because sample size exceeds specified accrual rates/time.

5. Discussion

The median event time test proposed by Park [15] provides a valuable tool for study developers to determine the required sample size in order to achieve sufficient power (at least $1 - \beta$) at a significance level of α for hypothesized median survivals. This approach is appealing due to its simplicity in decision-making and the ease of understanding the drug's efficacy based on the median event time. However, the method determines the sample size and threshold through the numerical search, which would require much time to search over the fine grid and be a computational burden. In this article, we addressed this challenge and propose explicit formulas for calculating the sample size and determining the threshold for the decision rules. We evaluated the performance of the proposed method in various clinical scenarios, considering the presence of censoring events to reflect the characteristics of the time-to-event endpoint. Simulation studies demonstrated that the proposed method effectively maintains the overall type I and II error rates and performs favorably compared to existing methods.

The design parameters α_1 and β_1 are searched over the grids to find the optimal pair minimizing the expected total sample size under the null hypothesis. The searching algorithm can be modified by applying the error spending function such as the O'Brien–Fleming spending function [18] to specify β_1. It means that the expected total sample size becomes a function of α_1 for the fixed β_1. The idea of using the error spending function allows us to extend the two-stage design to multi-stage clinical trials by splitting the type II error rate β into $\beta_k, k = 1, \ldots, K$, where K indicates the number of analyses and $1 - \beta_k$ indicates the probability of correct decision at the kth analysis under the alternative hypothesis, i.e., $\sum_{k=1}^{K} \beta_k = \beta$. In addition, the idea can also be used to extend the two-stage design used for Phase II to group sequential design for Phase III randomized controlled trials.

Since the proposed method uses the large sample theory of order statistics, it is likely to result in a large maximum sample size. To obtain a smaller maximum sample size, the minimax approach can be

Table 7
Decision rules for two-stage design and operating characteristics for hypothesis testing of median survivals ϕ_0 months versus ϕ_1 months at target error rates of $\alpha = 0.05$ and $\beta = 0.2$.

	ϕ_0	ϕ_1	Decision rule				Operating characteristics					
			n_1	t_1	n_2	t_2	$\hat{\alpha}$	$1-\hat{\beta}$	EN_0	EN_a	PET_0	PET_a
Exp	6	9	43	6.581	69	7.345	0.043	0.864	66.5	105.8	0.659	0.090
	6	11	21	7.274	40	7.823	0.042	0.843	32.6	55.6	0.710	0.136
	6	13	14	7.826	28	8.197	0.049	0.827	22.5	37.5	0.695	0.16
	6	15	11	8.650	25	8.373	0.042	0.817	19.1	31.5	0.677	0.179
	10	13	94	10.63	173	11.45	0.047	0.846	153.8	249.1	0.654	0.104
	10	15	43	10.97	69	12.24	0.047	0.849	67.3	105.0	0.648	0.102
	10	17	27	11.70	47	12.76	0.044	0.836	41.9	67.6	0.684	0.137
	10	19	20	12.66	39	13.09	0.043	0.814	31.8	52.2	0.696	0.174
Uniform	6	9	20	6.742	40	7.274	0.034	0.806	32.6	53.6	0.685	0.159
	6	11	10	6.914	16	7.935	0.030	0.789	16.7	23.2	0.582	0.175
	6	13	7	8.119	15	8.104	0.026	0.815	14.5	19.5	0.497	0.166
	6	15	5	8.072	10	8.548	0.031	0.854	11.4	13.8	0.361	0.117
	10	13	44	10.71	91	11.42	0.040	0.813	74.6	122.2	0.664	0.141
	10	15	20	11.24	40	12.12	0.036	0.780	33.9	52.6	0.652	0.185
	10	17	13	11.91	25	12.67	0.034	0.793	24.7	33.6	0.531	0.177
	10	19	9	11.89	16	13.29	0.035	0.825	19.3	22.9	0.356	0.131
Weibull	6	9	10	6.491	18	7.345	0.039	0.783	18.1	24.9	0.552	0.170
	6	11	5	6.932	9	7.903	0.042	0.85	11.0	13.0	0.337	0.112
	6	13	4	8.438	8	8.055	0.043	0.916	9.8	11.4	0.281	0.071
	6	15	3	8.936	6	8.373	0.050	0.943	7.9	8.7	0.191	0.042
	10	13	24	10.79	48	11.40	0.036	0.790	40.5	63.8	0.656	0.171
	10	15	10	10.82	18	12.24	0.042	0.798	21.2	25.4	0.377	0.147
	10	17	7	11.67	11	12.80	0.047	0.868	15.4	17.0	0.237	0.088
	10	19	5	12.66	10	13.06	0.047	0.925	13.5	14.5	0.152	0.046

considered, but it would require a constraint because the maximum sample size depending on β_2 is attained when β_2 goes to β (i.e., β_1 goes to 0).

As a future study, it is of interest to investigate the decision rule of the METT describing in terms of the number of events rather than the number of patients. This would address the practical issue of not being able to observe the median at the interim. Recently, Kundu et al. [19] provides the variance of the sample median based on a certain number of events for Weibull and exponential survivals. The idea of Kundu et al. [19] can be applied to improve the proposed monitoring rule. It is also interesting to extend the median event time test in the Bayesian framework and apply it to the two-stage adaptive clinical trials.

Supplementary materials

R codes are provided at https://github.com/funnypyh/METTSS and a shiny application is available at https://yeonhee.shinyapps.io/METTSS/.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

Dr. Park is partially supported by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education, United States.

Appendix A. Derivation of sample size determination

The details of the derivation of (1)–(4) are given as follows. By the definition of α_1 and β_1, we have $\Pr(Z_1 \leq t_1|H_0) = 1 - \alpha_1$ and $\Pr(Z_1 > t_1|H_a) = 1 - \beta_1$. Using the asymptotic normality of Z_1 under the null hypothesis, we have

$$\frac{t_1 - \phi_0}{\sqrt{\frac{0.25}{n_1\{f(\phi_0)\}^2}}} = z_{\alpha_1},$$

which implies Eq. (1). Similarly, using the asymptotic normality of Z_1 under the alternative hypothesis, we have

$$\frac{t_1 - \phi_1}{\sqrt{\frac{0.25}{n_1\{f(\phi_1)\}^2}}} = -z_{\beta_1},$$

which implies that

$$t_1 = \frac{-0.5z_{\beta_1}}{\sqrt{n_1}f(\phi_1)} + \phi_1. \tag{7}$$

From Eqs. (1) and (7), we have

$$\frac{0.5z_{\alpha_1}}{\sqrt{n_1}f(\phi_0)} + \phi_0 = \frac{-0.5z_{\beta_1}}{\sqrt{n_1}f(\phi_1)} + \phi_1.$$

We solve the last equation for the sample size n_1. Thus, we obtain the formula in (2). The derivation of (3) and (4) is the same as the derivation of (1) and (2). It uses the fact that $\Pr(Z_2 > t_2|H_0) = \alpha$ and $\Pr(Z_2 > t_2|H_a) - \Pr(Z_1 \leq t_1|H_a) = 1 - \beta$.

Appendix B. Decision rules of METT for Table 6

We provided in Table 7 the decision rule of the proposed two-stage design for the hypothesis tests we considered in the comparison. We also reported the operating characteristics for the different parametric assumptions of the survivals such as exponential, uniform, and Weibull distribution. Note that the decision rule and operating characteristics of the two-stage designs for the first three hypothesis testings in Table 6 are described in Tables 1 and 5.

References

[1] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, et al., New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1), Eur. J. Cancer 45 (2) (2009) 228–247.

[2] J. Whitehead, One-stage and two-stage designs for phase II clinical trials with survival endpoints, Stat. Med. 33 (22) (2014) 3830–3843.

[3] N.E. Breslow, Analysis of survival data under the proportional hazards model, Int. Stat. Rev. (Revue Int. Stat.) (1975) 45–57.

[4] D. Tu, A.J. Gross, A bartlett-type correction for the subject-years method in comparing survival data to a standard population, Statist. Probab. Lett. 29 (2) (1996) 149–157.

[5] X. Sun, P. Peng, D. Tu, Phase II cancer clinical trials with a one-sample log-rank test and its corrections based on the edgeworth expansion, Contemp. Clin. Trials 32 (1) (2011) 108–113.

[6] L. Kerschke, A. Faldum, R. Schmidt, An improved one-sample log-rank test, Stat. Methods Med. Res. 29 (10) (2020) 2814–2829.

[7] D.M. Finkelstein, A. Muzikansky, D.A. Schoenfeld, Comparing survival of a sample to that of a standard population, J. Natl. Cancer Inst. 95 (19) (2003) 1434–1439.

[8] J. Wu, Sample size calculation for the one-sample log-rank test, Pharm. Statist. 14 (1) (2015) 26–33.

[9] R. Schmidt, R. Kwiecien, A. Faldum, F. Berthold, B. Hero, S. Ligges, Sample size calculation for the one-sample log-rank test, Stat. Med. 34 (6) (2015) 1031–1040.

[10] M. Kwak, S.-H. Jung, Phase II clinical trials with time-to-event endpoints: optimal two-stage designs with one-sample log-rank test, Stat. Med. 33 (12) (2014) 2004–2016.

[11] L. Belin, Y. De Rycke, P. Broët, A two-stage design for phase II trials with time-to-event endpoint using restricted follow-up, Contemp. Clin. Trials Commun. 8 (2017) 127–134.

[12] L.D. Case, T.M. Morgan, Design of phase II cancer trials evaluating survival probabilities, BMC Med. Res. Methodol. 3 (1) (2003) 6.

[13] K. Owzar, S.-H. Jung, Designing phase II studies in cancer with time-to-event endpoints, Clin. Trials 5 (3) (2008) 209–221.

[14] B. Huang, E. Talukder, N. Thomas, Optimal two-stage phase II designs with long-term endpoints, Stat. Biopharm. Res. 2 (1) (2010) 51–61.

[15] Y. Park, Optimal two-stage design of single arm Phase II clinical trials based on median event time test, PLoS One 16 (2) (2021) e0246448.

[16] R. Simon, Optimal two-stage designs for phase II clinical trials, Control. Clin. Trials 10 (1) (1989) 1–10.

[17] F. Mosteller, On some useful “inefficient” statistics, Ann. Math. Stat. 17 (4) (1946) 377–408.

[18] P.C. O'Brien, T.R. Fleming, A multiple testing procedure for clinical trials, Biometrics (1979) 549–556.

[19] M.G. Kundu, S. Samanta, S. Mondal, Review of calculation of conditional power, predictive power and probability of success in clinical trials with continuous, binary and time-to-event endpoints, Health Serv. Outcomes Res. Methodol. (2023) 1–32.