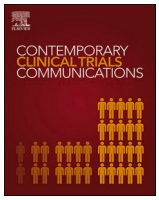
Contemporary Clinical Trials Communications 40 (2024) 101315

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

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



Group sequential design for time-to-event outcome with non-proportional hazards using the concept of relative time utilizing two different Weibull distributions

Milind A. Phadnis *, Nadeesha Thewarapperuma, Matthew S. Mayo

Department of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS, USA

ARTICLE INFO

Keywords:
Efficacy
Error spending
Futility
Non-proportional hazards
Sample size
Weibull

ABSTRACT

A group sequential design allows investigators to sequentially monitor efficacy and safety as part of interim testing in phase III trials. Literature is well developed in the case of continuous and binary outcomes, however, in case of trials with a time-to-event outcome, popular methods of sample size calculation often assume proportional hazards. In situations where the proportional hazards assumption is inappropriate as indicated by historical data, these popular methods are very restrictive. In this paper, a novel simulation-based group sequential design is proposed for a two-arm randomized phase III clinical trial with a survival endpoint for the non-proportional hazards scenario. By assuming that the survival times for each treatment arm follow two different Weibull distributions, the proposed method utilizes the concept of Relative Time to calculate the efficacy and safety boundaries at selected interim testing points. The test statistic used to generate these boundaries is asymptotically normal, allowing p-value calculation at each boundary. Many design features specific to time-to-event data can be incorporated with ease. Additionally, the proposed method allows the flexibility of having the accelerated failure time model and the proportional hazards model as constrained special cases. Real life applications are discussed demonstrating the practicality of the proposed method.

1. Introduction

A group sequential design (GSD) aims to incorporate interim testing at prespecified time points called 'looks' to collect early evidence for efficacy and/or safety and is often conducted as a phase III randomized clinical trial (RCT). That is, at each interim look, the decision to stop or continue the RCT is taken based on whether a test statistic (or p-value) exceeds or does not exceed a well-defined boundary value. Such GSDs enjoy a rich history starting with the works of Wald [26] and Armitage [1] with a vast literature available on the subject in books by Whitehead [29], Jennison and Turnbull [6], Proschan et al. [20], Dmitrienko et al. [2], Wassner and Brannath [28]. Likewise, many articles provide an excellent overview (see Whitehead [30], Todd [25], Mazumdar and Bang [11]) and stress the importance of considering ethical, financial and administrative requirements in designing such GSDs (see Enas et al. [4], Jennison and Turnbull [5], Ellenberg et al. [3]). Historically, GSDs are well developed in the case of continuous and binary endpoints and are available in popular statistics software using the Repeated Significance Testing (RST) approach. This method incorporates a rich family of designs proposed by Pocock [19], O'Brien and Fleming [13], Wang and Tsiatis [27] while also allowing flexible data monitoring strategies using the error spending method of Lan and DeMets [10]. In the case of a time-to-event outcome, however, the available literature only discusses scenarios where the survival times in the two treatment arms of an RCT are exponentially distributed or when the proportional hazards (PH) assumption is satisfied. For example, Jennison and Turnbull [6] have discussed examples of using the log-rank and stratified log-rank tests, and separately, using the PH assumption. The same assumptions are made by popular statistics software such as PASS [14] in using the method proposed by Reboussin et al. [21] – Fortran 77 program using the framework of Lan and DeMets [10], and Kim and DeMets [8] – to implement GSDs by deploying the weighted and unweighted variations of the log-rank test. More recently, Wu and Xiong [32] have proposed a GSD using a Weibull model that satisfies the PH property but provides better results than the usual log-rank test at very early look points where the available data could be small. Likewise, Jiang et al. [7] has proposed a simulation-based SAS macro for a GSD using the exponential and Weibull distributions. While such methods do incorporate many design

* Corresponding author. 3901 Rainbow Boulevard, Kansas City, KS, 66160, USA.
E-mail address: mphadnis@kumc.edu (M.A. Phadnis).

https://doi.org/10.1016/j.conctc.2024.101315
Received 11 November 2023; Received in revised form 22 May 2024; Accepted 25 May 2024
Available online 27 May 2024

features specific to time-to-event outcomes such as loss to follow-up, limited accrual and follow-up times, myriad accrual patterns, equal/-unequal allocation to groups, equally or unequally spaced looks, adjustments for non-compliance, and the fact that a patient surviving till the end of study contributes to the test statistic computed at each of the interim looks, they are based on the restrictive PH assumption or on the assumption of exponentially distributed survival times.

In scenarios where results from previously conducted historical studies or earlier phase II trials are used to guide the design of a current phase III trial, the choice of PH assumption may not be appropriate. For example, a previously conducted moderately sized phase II trial may have indicated that a new investigative treatment outperformed a standard control by improving the median survival time by 50 % (say, 18 months vs 12 months) but that the assumption of proportional hazards was not appropriate. In this situation, it would not be correct to design a phase III trial with the PH assumption or by assuming that survival times in both treatment arms follow the exponential distribution (constant hazards in each arm leading to a constant hazard ratio). The situation would further get compromised in the case of a GSD where the decision for early evidence of efficacy or futility is based on the construction of decision boundaries which themselves are calculated utilizing a test statistic based on the PH assumption. Secondly, researchers working in specific disease areas may find it more comfortable to define an effect size using the paradigm of 'improvement in longevity' instead of a 'reduction in hazard'. That is, an effect size defined as – 'improvement in 25th, 50th (median) and 75th percentile of survival time' – may be more informative for patients while consenting to take part in a RCT compared to, say, a 25 % or 50 % reduction in hazard. Here, it is important to note that only in the case of exponentially distributed survival times, a halving of hazard automatically implies doubling of longevity whereas for other survival distributions explicit calculations need to be done to relate the two effect size definitions. Thirdly, a GSD incorporates multiple looks and hence it is important to correctly represent the number of events occurring at each interim look to best encapsulate the underlying biological phenomenon of a disease more accurately.

To counter the limitations mentioned above, Phadnis and Mayo [16] developed a parametric GSD using a generalized gamma (GG) distribution after extending the work on two-arm fixed RCT design of Phadnis et al. [18]. Their recommended method based on the proportional time (PT) assumption provides flexibility of modeling various hazard shapes in the treatment arms and does not require the PH assumption. The PT assumption implies that for all quantiles of survival time, the life course of a disease (or event of interest) in one group is accelerated (or decelerated) by a constant factor compared to another group and is therefore essentially an Accelerated Failure Time (AFT) model. While it offers the practical benefit of an easily interpretable *treatment effect* in terms of a *percent improvement in longevity*, it also has two notable limitations. First, although the point estimates of the GG shape parameters to be used as input values in the current study GSD can be obtained from previous studies, their accuracy cannot be ascertained in all situations (the most general case of the GG) if the previous studies were small or moderately sized. This may lead to a situation where the decision boundaries are sensitive to the choice of the GG shape parameters. Second, when the early interim looks for the phase II trial involve relatively small samples sizes (say in the 30–50 range) the recommended method may sometimes run into convergence problems. Authors like Klein and Moeschberger [9] have noted that the three-parameter GG distribution is often used to choose a simpler two-parameter distribution (special case of the GG family such as Weibull, lognormal, and gamma) while modeling time-to-event data except when dealing with large sample sizes thereby restricting the method of Phadnis and Mayo [16] to only large size phase III RCTs. Third, it is possible that in a real-life biomedical application neither the PH nor the PT assumption is appropriate and to the best of our knowledge there is no method available in the literature for this general scenario. In our paper, we aim to fill in this gap in the literature.

Two motivating examples discussed in Section 2 elucidate the need for developing a GSD where neither the PH nor the PT assumption is appropriate. Section 3 discusses the proposed GSD by extending the recent work of Phadnis and Mayo [17] using the concept of Relative Time (RT). The derivations by Phadnis and Mayo [17] are discussed briefly in the online Appendix A with Section 3 detailing the main GSD method by means of a combination of analytical formulas and simulations. Results are presented in Section 4 for various scenarios of the input parameters of the GSD. The discussion in the final Section 5 provides practical insights for the proposed method and deliberates about its advantages and limitations.

2. Motivating examples

We discuss two motivating examples representing two different scenarios pertaining to the construction of a GSD using our proposed method. Many variations of these two scenarios are possible and some of them are discussed in the Results section showing how the sample size changes depending on varying user inputs.

First, we consider the example where researchers intend to construct a two-arm phase III GSD for treating patients afflicted with chemotherapy refractory advanced metastatic biliary cholangiocarcinoma – a rare but a very aggressive neoplasm. Such patients have metastatic disease and undergo an initial treatment followed by a second-line of treatment. Trialists are interested in comparing the performance of a new experimental (E) second-line treatment to a standard control (C) second-line treatment using progression-free survival (PFS) as the time-to-event endpoint of interest. In a previous phase II study, the PFS for the C arm has been reported using a Kaplan Meier (KM) curve with a median PFS of 4 months and an interquartile range (IQR) of 2–7 months. In the current phase III trial under consideration, researchers hypothesize that the E arm will show an improvement in median PFS compared to the C arm, but that this improvement measured as a metric of *longevity* will be gradual. That is, the nature of the disease is such that the improvement for 10th percentile (denoted p_1 in later sections) of PFS will be by a factor of 1.5 and the improvement for 90th percentile (denoted p_2 in later sections) of PFS will be by a factor of 2. Thus, the effect of treatment improves with the passage of time with improvement in median PFS being the target of the research. Both accrual and follow-up time is taken as 12 months (leading to a total study time of 24 months) with a type I error of 5 % for a one-sided test (acceptable as an option for rare cancers as discussed by Renfro et al. [22], and by Sposto and Stram [24]) and the target power is 80 %. This example represents a frequently occurring real-life scenario in cancer trials where researchers anticipate long-term survivors to get the maximum benefit from a new treatment but expect only a small realistic improvement for short-term survivors (see Fig. 1a). Therefore, the GSD sample size calculations should be done keeping in mind that the hypothesized *treatment effect* changes over time and cannot be expressed through a single constant number such as a simple ratio of mean or median survival time. Additionally, the researchers are not comfortable with the proportional hazards assumption because published results from some previous observational studies related to the current disease area suggested that this assumption was not valid. Thus, the *effect size* cannot be defined by a single number such as a constant hazard ratio (HR). Further, due to the high cost associated with the treatment regimens for this rare form of cancer, researchers would like to conduct a GSD with interim testing for evidence of early efficacy or futility at the equally spaced intervals of 8 months – two interim analyses at 8 and 16 months, followed by the final analysis at 24 months.

The second example represented by Fig. 1b pertains to a real-life scenario with surgery as an experimental treatment whose performance is to be compared to a non-surgical standard-of-care control. Here, the trialists hypothesize that patients who receive surgery will experience a substantial benefit very soon after surgery compared to those who do not receive surgery, but that this benefit measured in terms of improvement in *longevity* will gradually wane with the passage of

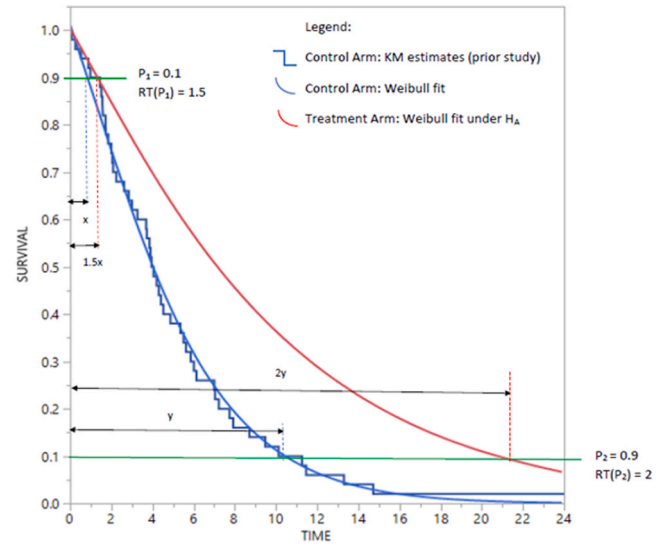


Fig. 1a. Scenario #1 with effect size defined as RT(0.1) = 1.5 and RT(0.9) = 2. Reprinted from *Sample size calculation for two-arm trials with time-to-event endpoint for non-proportional hazards using the concept of Relative Time when inference is built on comparing Weibull distributions*, by M.A. Phadnis and M.S. Mayo, Biometrical Journal 63 (2021), Pg. 1409. Copyright [2021] by John Wiley & Sons, Inc. Reprinted with permission.

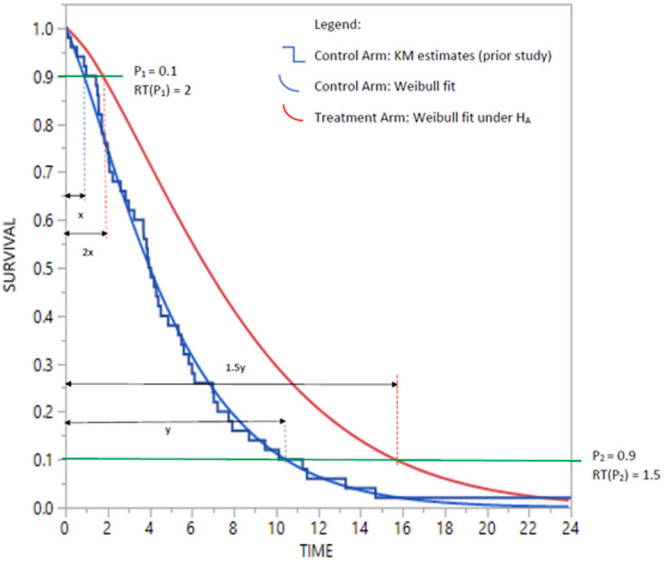


Fig. 1b. Scenario #2 with effect size defined as RT(0.1) = 2 and RT(0.9) = 1.5. Reprinted from *Sample size calculation for two-arm trials with time-to-event endpoint for non-proportional hazards using the concept of Relative Time when inference is built on comparing Weibull distributions*, by M.A. Phadnis and M.S. Mayo, Biometrical Journal 63 (2021), Pg. 1409. Copyright [2021] by John Wiley & Sons, Inc. Reprinted with permission.

time. Thus, the improvement for the 10th percentile (denoted p_1 in later sections) of Overall Survival (OS) will be by a factor of 2 but as the effect of surgery diminishes over time, the improvement for the 90th percentile (denoted p_2 in later sections) of OS will be by a factor of 1.5. Analogous to the first example, researchers would like to conduct the sample size calculations for a GSD with three look points (8, 16, and 24 months) with the main research question targeted at improvement in median OS time.

In both the examples described above, researchers would also like to conduct a sensitivity analysis by varying some of the input design parameters and assess how the sample size calculations change. That is, if the calculations suggest a very large sample size which is not realistically feasible, they would like to consider alternate design inputs such as longer values of accrual and/or follow-up time. They would also like to assess design efficiency by comparing the expected sample size obtained from a GSD to that of a regular two-arm fixed sample design. In such examples with non-constant effect size definitions, Phadnis and Mayo [17] have discussed the methodology for a *fixed* two-arm design and in the remainder of this paper we show how their method can be extended to conduct a GSD for the non-PH and non-PT scenarios. The reader is recommended to read this paper before proceeding to the next section and get acquainted with key notations used throughout the text. Appendix A discusses these notations in brief.

3. Methods

To implement a GSD in the case of non-PH and non-PT situations described in the examples discussed above, we propose a method using the concept of Relative Time (see online Appendix A before proceeding further). This method combines the analytical results obtained by Phadnis and Mayo [17] using an asymptotically normally distributed test statistic with a proposed simulation-based approach to conduct the sample size calculations. Phadnis and Mayo [17] have demonstrated maintenance of type I error, low average relative bias, and adequacy of power for a variety of scenarios and hence their method can be readily adopted to construct a GSD.

3.1. Calculation of sample size

Interim testing in a GSD may involve - testing for efficacy only, testing both efficacy and futility at all looks, testing for efficacy at all looks but testing for futility after skipping a few looks, binding or non-binding futility rules, and equally spaced or unequally spaced look times. The myriad variations of these input combinations can be incorporated along with various aspects of time-to-event data and are summarized using a stepwise algorithm given in online Appendix B. SAS software [23] is used to evaluate the design characteristics using 10 000 simulations by adopting the GSD algorithm discussed in Reboussin et al. [21] but with extra adjustments for parameter estimation using PROC LIFEREG (parametric regression procedure for time-to-event data in SAS software) with efficacy and futility boundaries determined by using the normally distributed test statistic (see online Appendix A) and corresponding p-values.

A detailed ten-step procedure detailing the algorithms using a simulation-based procedure are explained in online Appendix B. In section 1 of this appendix, the main GSD algorithm incorporating both efficacy and futility testing (with or without skipping, binding or non-binding) with corresponding boundaries, is explained. In section 2 of this appendix, a much simpler GSD algorithm with interim testing for efficacy-only is outlined. Additionally, sample size calculations can also be conducted for a two-sided hypothesis with minor adjustments to the algorithms detailed in online Appendix B section 1 and 2 (by having two boundaries for an 'efficacy-only GSD'). Analogously, simultaneous efficacy/futility testing with a two-sided H_A would warrant two separate sets of two boundaries. For all scenarios mentioned above, efficacy and/or futility boundaries can be constructed on the z-scale or equivalently on the p-value scale.

3.2. Calculation of stop probabilities and expected number of events

The calculation of stop probabilities (probability of ending a trial at a given look) under H_0 and H_A can be conducted using the simulation-based approach proposed by us (see online Appendix B section 1 and 2). Consider a GSD with a one-sided hypothesis incorporating both efficacy and futility boundaries. Here the stopping probability under H_0 is the summation of the stopping probability for efficacy under H_0 and

stopping probability for futility under H_0. At look j, this can be calculated as the proportion of samples under the null hypothesis that are above Q_{0j} or are below Q_{1j}. This in turn facilitates the calculation of the cumulative stopping probability under H_0. Likewise, the stopping probability under H_A is the summation of the stopping probability for efficacy under H_A and stopping probability for futility under H_A. At look j this can be calculated as the proportion of samples under the alternate hypothesis that are above Q_{0j} or below Q_{1j}. This in turn facilitates the calculation of the cumulative stopping probability under H_A. At look m (the final look), both cumulative stopping probabilities should be equal to 1. However, due to the random nature of the simulations and the discrete nature of the sample size n, a tolerance of 0.001 is permitted in this calculation by us.

In case of an efficacy-only GSD with a one-sided hypothesis, the stop probability under H_0 at each look j is, by definition, equal to the amount of alpha spent α_j. The cumulative stop probability under H_0 is thus equal to α. Likewise, the stop probability under H_A at each look j is the proportion of samples under H_A that exceed Q_{0j}. This in turn facilitates calculation of the cumulative stop probability under the alternate hypothesis H_A. For the last look m, this should be equal to $1-\beta$, but due to n being a whole number, the power will slightly exceed $1-\beta$.

The calculation of the expected number of events under H_0 and H_A is as follows. Let n_{c,j,H_0} and n_{e,j,H_0} be the simulated number of events at look j under H_0 in the control arm and experimental treatment arm respectively. Let P_{j,H_0} be the stopping probability under H_0 at look j. Then the expected number of events in the control arm and experimental arm under H_0 is calculated as:

$$E(n_{H_0}) = \sum_{j=1}^{m-1} \left\{ \left(\frac{n_{c,j,H_0} + n_{e,j,H_0}}{2} \right) P_{j,H_0} \right\} + \left(\frac{n_{c,j,H_0} + n_{e,j,H_0}}{2} \right) \left(1 - \sum_{j=1}^{m-1} P_{j,H_0} \right) \tag{1}$$

Let n_{c,j,H_A} and n_{e,j,H_A} be the simulated number of events at look j under H_A in the control arm and experimental treatment arm respectively. Let P_{j,H_A} be the stopping probability under H_A at look j. Then the expected number of events in the control arm under H_A is calculated as:

$$E(n_{c,H_A}) = \sum_{j=1}^{m-1} \left(n_{c,j,H_A} . P_{j,H_A} \right) + n_{c,j,H_A} \left(1 - \sum_{j=1}^{m-1} P_{j,H_A} \right) \tag{2}$$

Note that $n_{c,j,H_A} = n_{c,j,H_0}$ because under H_A, the effect size definition of $RT(p_{mid}) > 1$ affects only the number of events in the experimental treatment arm.

The expected number of events in the experimental treatment arm under H_A is calculated as:

$$E(n_{e,H_A}) = \sum_{j=1}^{m-1} \left(n_{e,j,H_A} . P_{j,H_A} \right) + n_{e,j,H_A} \left(1 - \sum_{j=1}^{m-1} P_{j,H_A} \right) \tag{3}$$

Other important quantities of interest to the trialists such as the cumulative subject time under H_0 and H_A can also be computed using the simulated datasets.

4. Results

4.1. Model validation

Before proceeding to discuss the two examples mentioned in Section 2, we provide validation for a GSD deploying our method in the special case of exponentially distributed times as well as proportional hazards ($\beta_0 = \beta_1 = 1$) by comparing the obtained results to that of standard sample size software. We have chosen the PASS 20 statistical software [14] for this comparison as it is one of the specialized commercial software built for sample size and power calculations and provides an equivalent simulation-based approach for exponentially distributed data. The design features used for the purpose of validation are as given below:

- *Number of simulations* $B = 10\,000$
- *Type I error* $\alpha = 0.025$ *(one-sided test),*
- *Power* $1 - \omega = 0.80$
- *Mean survival time in Control arm* $= 1$ *year*
- *Effect size* $RT(p_1) = RT(p_2) = constant = 1.75$ *(Hence choice of* p_1, p_2 *does not matter*).
- *Control arm shape* $= \beta_0 =$ *Treatment arm shape* $= \beta_1 = 1$ *(exponential distribution)*
- *Allocation ratio* $r = 1$
- *Proportion loss to follow-up* $= 0$
- *Accrual time* $a = 1$ *(year),*
- *Accrual pattern = Uniform,*
- *Total time* $t = 4$ *(years)*
- *Number of looks* $m = 4$ *(equally spaced at 1, 2, 3, and 4 years)*
- *Number of skips for futility (binding)* $= 2$
- *Alpha spending function = Hwang-Shih-DeCani (with* $\rho_0 = 1$ *i.e. Pocock type)*
- *Beta spending function = Hwang-Shih-DeCani (with* $\rho_1 = 1$ *i.e. Pocock type)*

Table 1A displays the output obtained using our proposed method and Table 1B displays the output using PASS 20 (note that the displayed output may change slightly owing to the choice of random seed used for data simulation). For example, Table 1B (PASS output) shows the total sample size as $n_c = 64, n_e = 65$ (see values mentioned below the tables). In reality with 10 000 simulations using different random seeds, we saw sample sizes ranging from $n_c = 63, n_e = 63$ to $n_c = 66, n_e = 66$ and we selected $n_c = 64, n_e = 65$ as a middle option. For the GSD using our proposed method, we get $n_c = 64, n_e = 64$ (see right-hand bottom corner of Table 1A). Comparing the results of Table 1A to Table 1B, we find that most of the column entries such as the amount of alpha and beta spent at each look, expected number of events under the null and alternate hypotheses, stop probability under the null and alternate hypotheses, Z test statistic (and p-values) defining the efficacy and futility boundaries, and cumulative observation time under the null and alternate hypotheses match quite well with each other. It should be noted that although both methods have efficacy and futility boundaries defined using an asymptotically normal distributed test statistic, the underlying methods are different. While PASS uses a logrank test in calculating the Z test statistic, our proposed method uses a Z test statistic based on the concept of Relative Time RT(p). Thus, the Z test statistic from the proposed method is associated with a specific combination of $\widehat{RT}(p)$ and $SE\{\widehat{RT}(p)\}$ and therefore provides a practically meaningful interpretation of efficacy and futility boundaries as a measure of improvement in longevity for treatment arm relative to the control arm. On the other hand, the Z test statistic from the logrank test yields efficacy and futility boundary values for reduction in hazard (and consequently improvement in survival) for treatment arm relative to control arm but does not provide a direct interpretation of its magnitude. We also tried many different design scenarios (such as those discussed in Sections in 4.2 and 4.3) and in all scenarios we found that the results from the two methods are consistent for exponentially distributed survival times, that is for the case where the proportional hazards assumption automatically holds.

4.2. Clinical trial with cholangiocarcinoma with progression-free survival (PFS) as endpoint

For the cholangiocarcinoma example of Section 2, we consider designing a GSD with the HSD error spending using $\rho_0 = -4$ for both type I and type II errors to approximate the O'Brien-Fleming approach. The Weibull shape parameter for the control arm β_0 is varied from 0.25, 0.5, 0.75, 1.0, 1.25 to 1.50 in Table 2A through 2F with all other design features remaining the same. As an example, the design features corresponding to Table 2A are:

Table 1A
GSD validation output, exponential distn, (proposed method), RT(0.1) = RT(0.9) = 1.75, HSD(OBF) spending, 10 000 simulations, 2 futility skips).

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	1	23.51	23.51	0.00875	0.00875	2.259	0.0119	0.0088	0.0088	23.530
2	2	49.09	49.09	0.00681	0.01556	2.236	0.0127	0.0003	0.0091	49.070
3	3	58.51	58.51	0.00531	0.02087	2.238	0.0126	0.9777	0.9868	58.458
4	4	61.96	61.96	0.00413	0.02500	2.089	0.0183	0.0131	0.9999	61.886

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	1	23.56	15.22	0.0000	0.0000	–	–	0.2808	0.2808	26.690
2	2	49.13	36.45	0.0000	0.0000	–	–	0.3534	0.6342	63.855
3	3	58.54	48.41	0.1646	0.1646	1.934	0.0265	0.2858	0.9200	84.852
4	4	62.02	55.25	0.0326	0.1972	2.089	0.0183	0.0798	0.9998	96.695

$E(n_{H_0}) = 58.25$ $E(n_{c,H_A}) = 45.61$ $E(n_{e,H_A}) = 35.40$ Sample Size: $n_c = n_e = 64$

Table 1B
GSD output for exponential distribution using PASS 2020, HSD (OBF) spending, Mean in control = 1 year, 10 000 simulations, 2 futility skips.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	1	23.59	23.96	0.00875	0.00875	2.371	0.0089	0.0087	0.0087	23.54
2	2	49.14	49.91	0.00681	0.01556	2.330	0.0099	0.0068	0.0155	49.09
3	3	58.53	59.44	0.00531	0.02087	2.259	0.0120	0.9703	0.9858	58.48
4	4	62.00	62.97	0.00413	0.02500	2.119	0.0170	0.0051	0.9909	61.92

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	1	23.59	15.50	0.0000	0.0000	–	–	0.2566	0.2566	27.10
2	2	49.14	37.02	0.0000	0.0000	–	–	0.3708	0.6274	64.81
3	3	58.53	49.20	0.1525	0.1525	1.901	0.0287	0.2872	0.9146	86.12
4	4	62.00	56.09	0.0302	0.1827	2.119	0.0170	0.0854	1.0000	98.14

$E(n_{H_0}) = 58.64$ $E(n_{c,H_A}) = 46.38$ $E(n_{e,H_A}) = 36.63$ Sample Size: $n_c = 64$; $n_e = 65$

- *Number of simulations B = 10 000*
- *Type I error $\alpha = 0.025$ (one-sided test)*
- *Power $1 - \omega = 0.80$*
- *Median survival time in Control arm = 4 (months)*
- *Quantiles at which the effect size is defined: $p_1 = 0.10$, $p_2 = 0.90$.*
- *Effect size at p_1 defined as: $RT(p_1) = 1.52$.*
- *Effect size at p_2 defined as: $RT(p_2) = 1.98$.*
- *Control arm shape $= \beta_0 = 0.25$.*
- *Allocation ratio r = 1*
- *Proportion loss to follow-up = 0.20*
- *Accrual time a = 12 (months), Accrual pattern = Uniform*
- *Total time t = 24 (months)*
- *Number of looks m = 3 (equally spaced at 8, 16, and* 24 months*)*
- *Number of skips for futility (binding) = 0*
- *Alpha spending function = Hwang-Shih-DeCani (with $\rho_0 = -4$ i.e. OBF type)*
- *Beta spending function = Hwang-Shih-DeCani (with $\rho_1 = -4$ i.e. OBF type)*

Note that 'RT(p_1) = 1.52 and RT(p_2) = 1.98' have been adjusted from 'RT(p_1) = 1.5 and RT(p_2) = 2' following a discussion in Phadnis and Mayo [17].

Table 2A
GSD – Weibull shape $\beta = 0.25$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	412.18	412.18	0.00130	0.00130	3.041	0.0012	0.3896	0.3896	1811.58
2	16	731.93	731.93	0.00494	0.00625	2.465	0.0069	0.4621	0.8517	5840.63
3	24	796.21	796.21	0.01875	0.02500	1.944	0.0259	0.1482	0.9999	9295.89

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	412.28	375.42	0.01009	0.01009	−0.276	0.6088	0.1604	0.1604	1981.79
2	16	732.19	672.24	0.03828	0.04837	1.001	0.1585	0.4740	0.6344	6536.76
3	24	796.54	735.66	0.14523	0.19360	1.944	0.0260	0.3652	0.9996	10596.85

$E(n_{H_0}) = 616.91$ $E(n_{c,H_A}) = 703.83$ $E(n_{e,H_A}) = 647.52$ Sample Size: $n_c = n_e = 1392$

Table 2B
GSD – Weibull shape $\beta = 0.50$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	90.05	90.05	0.00130	0.00130	3.004	0.0013	0.3050	0.3050	450.17
2	16	183.36	183.36	0.00494	0.00625	2.497	0.0063	0.5213	0.8263	1298.28
3	24	209.62	209.62	0.01875	0.02500	1.962	0.0249	0.1736	0.9999	1813.40

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	89.92	74.85	0.01038	0.01038	−0.497	0.6903	0.1217	0.1217	516.81
2	16	183.25	157.15	0.03937	0.04975	0.921	0.1784	0.4671	0.5888	1583.80
3	24	209.55	184.60	0.14935	0.19910	1.962	0.0249	0.4110	0.9998	2362.87

$E(n_{H_0}) = 159.39$ $E(n_{c,H_A}) = 182.77$ $E(n_{e,H_A}) = 158.38$ Sample Size: $n_c = n_e = 313$

Table 2C
GSD – Weibull shape $\beta = 0.75$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	36.73	36.73	0.00130	0.00130	2.906	0.0018	0.2558	0.2558	200.133
2	16	83.05	83.05	0.00494	0.00625	2.530	0.0057	0.5583	0.8141	523.673
3	24	96.26	96.26	0.01875	0.02500	1.969	0.0245	0.1858	0.9999	654.905

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	36.65	28.15	0.01036	0.01036	−0.653	0.7431	0.1053	0.1053	236.318
2	16	82.93	67.48	0.03931	0.04967	0.885	0.1881	0.4545	0.5598	691.786
3	24	96.20	83.62	0.14913	0.19880	1.969	0.0245	0.4401	0.9999	965.703

$E(n_{H_0}) = 73.60$ $E(n_{c,H_A}) = 83.97$ $E(n_{e,H_A}) = 70.43$ Sample Size: $n_c = n_e = 130$

Table 2D
GSD – Weibull shape $\beta = 1.00$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	19.93	19.93	0.00130	0.00130	3.019	0.0013	0.2264	0.2264	114.626
2	16	48.30	48.30	0.00494	0.00625	2.611	0.0045	0.5918	0.8182	278.808
3	24	55.31	55.31	0.01875	0.02500	1.994	0.0231	0.1816	0.9998	319.278

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	19.82	14.10	0.01034	0.01034	−0.752	0.7740	0.0694	0.0694	137.899
2	16	48.25	38.00	0.03923	0.04957	0.906	0.1824	0.4551	0.5245	391.368
3	24	55.28	48.78	0.14883	0.19840	1.994	0.0231	0.4754	0.9999	517.358

$E(n_{H_0}) = 43.11$ $E(n_{c,H_A}) = 49.66$ $E(n_{e,H_A}) = 41.47$ Sample Size: $n_c = n_e = 71$

Table 2E
GSD – Weibull shape $\beta = 1.25$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	12.56	12.56	0.00130	0.00130	2.784	0.0027	0.2272	0.2272	75.076
2	16	31.96	31.96	0.00494	0.00625	2.500	0.0062	0.5818	0.8090	172.488
3	24	35.80	35.80	0.01875	0.02500	1.945	0.0259	0.1909	0.9999	186.879
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	12.55	8.29	0.01011	0.01011	−0.775	0.7809	0.0748	0.0748	91.243
2	16	31.94	24.73	0.03836	0.04847	0.839	0.2008	0.4816	0.5564	253.616
3	24	35.75	32.49	0.14553	0.19400	1.945	0.0259	0.4436	1.0000	320.519
$E(n_{H_0}) = 28.27$			$E(n_{c,H_A}) = 32.21$			$E(n_{e,H_A}) = 26.94$		Sample Size: $n_c = n_e = 45$		

Table 2F
GSD – Weibull shape $\beta = 1.50$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	8.70	8.70	0.00130	0.00130	2.735	0.0031	0.1927	0.1927	52.919
2	16	22.63	22.63	0.00494	0.00625	2.607	0.0046	0.6206	0.8133	116.976
3	24	24.78	24.78	0.01875	0.02500	1.965	0.0247	0.1866	0.9999	122.904
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	8.68	5.29	0.01010	0.01010	−0.891	0.8134	0.0591	0.0591	65.079
2	16	22.64	17.41	0.03830	0.04840	0.871	0.1919	0.4627	0.5218	178.448
3	24	24.79	23.16	0.14530	0.19370	1.965	0.0247	0.4781	0.9999	217.834
$E(n_{H_0}) = 20.34$			$E(n_{c,H_A}) = 22.83$			$E(n_{e,H_A}) = 19.44$		Sample Size: $n_c = n_e = 31$		

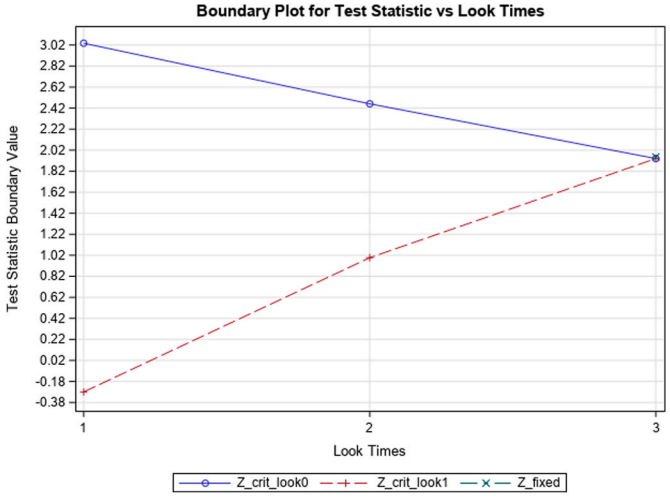


Fig. 2a. Efficacy and futility boundaries for Table 2A on the z-test statistic scale.

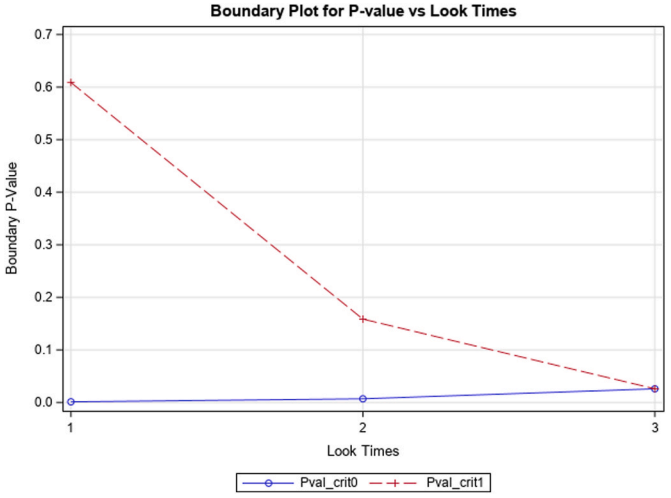


Fig. 2b. Efficacy and futility boundaries for Table 2A on the p-value scale.

Using the general theory mentioned in online Appendix A and the algorithm for efficacy and futility testing mentioned in online Appendix B section 1, the calculations for sample size, expected number of events under H_0 and H_A, and the stop probabilities under H_0 and H_A can be performed. For example, in Table 2A, the efficacy boundary values on the Z scale are 3.041, 2.465 and 1.944 while the futility boundaries on the Z scale are −0.276, 1.001 and 1.944 at the three equidistant look times of 8, 16, and 24 months respectively. These efficacy and futility boundaries are also displayed in Fig. 2a and b. The HSD (OBF) function spends a type I error of 0.025 (one-sided) partitioned into 0.00130, 0.00494 and 0.01875 at the three looks and as both efficacy and futility are tested at each of the three looks, we get stop probabilities under H_0 of 0.3896, 0.4621 and 0.1482. Likewise, the stop probabilities under H_A are 0.1604, 0.4740 and 0.3652 respectively. The expected number of events in both arms under H_0 are 616.1. Under H_A, we expect to observe 703.83 events in the control arm and 647.52 in the treatment arm. The GSD suggests that with an anticipated dropout rate of 20 %, accrual time of 12 months and follow-up time of 12 months, we need to enroll 1392 subjects in each of the two arms maintaining a type

II error of 0.1936 that is split as 0.10009, 0.03828 and 0.14523 at the three looks.

Comparing across Table 2A through 2F, we see that the sample size in each arm expectedly decreases from 1392 to 313 to 130 to 71 to 45 to 31. This is consistent with the results of fixed single-arm (Wu [31]; Phadnis [15]), fixed two-arm trial results (Phadnis and Mayo [17]; Phadnis et al. [18]) and GSD with common Weibull shape (Wu and Xiong [32]) reported in literature specific to the Weibull distribution. For each of these tables, the type II error is split exactly equally using the HSD (OBF) spending function in the first iteration of the algorithm mentioned in online Appendix B section 1, however, owing to the last step of the algorithm (the 'Search ω' step), the final iteration shows small differences in the cumulative type II error used by the GSD.

4.3. Clinical trial with surgical intervention with overall survival (OS) as endpoint

For the surgery versus standard-of-care example of Section 2, we consider designing a GSD with the HSD error spending using $\rho_0 = 1$ for both type I and type II errors to approximate the Pocock approach (equal alpha and beta spending at each look). Again, the Weibull shape parameter for the control arm β_0 is varied from 0.25, 0.5, 0.75, 1.0, 1.25 to 1.50 in Table 3A through 3F with all other design features remaining the same, but with $RT(p_1) = 2$ and $RT(p_2) = 1.5$ implying that the early two-fold improvement in longevity due to surgical intervention diminishes gradually over time towards a 50 % improvement in longevity. As an example, the design features corresponding to Table 3A (with $\beta_0 = 0.25$) are as follows:

Table 3A
GSD – Weibull shape $\beta = 0.25$, equally spaced looks, RT(0.1) = 2, RT(0.9) = 1.5, JT (Pocock) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	559.12	559.12	0.00833	0.00833	2.357	0.0092	0.7265	0.7265	2457.36
2	16	993.21	993.21	0.00833	0.01667	2.238	0.0126	0.2315	0.9579	7926.07
3	24	1080.54	1080.54	0.00833	0.02500	2.056	0.0199	0.0420	0.9999	12616.96
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	558.91	512.50	0.06613	0.06613	0.596	0.2757	0.4781	0.4781	2687.58
2	16	993.30	923.17	0.06613	0.13227	1.604	0.0544	0.3604	0.8385	8802.56
3	24	1080.67	1013.44	0.06613	0.19840	2.057	0.0199	0.1616	1.0001	14181.72
$E(n_{H_0}) = 681.35$			$E(n_{c,H_A}) = 799.87$			$E(n_{e,H_A}) = 741.51$		Sample Size: $n_c = n_e = 1889$		

Table 3B
GSD – Weibull shape $\beta = 0.50$, equally spaced looks, RT(0.1) = 2, RT(0.9) = 1.5, JT (Pocock) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	114.99	114.99	0.00833	0.00833	2.391	0.0084	0.6699	0.6699	575.25
2	16	234.26	234.26	0.00833	0.01667	2.285	0.0112	0.2768	0.9467	1650.74
3	24	267.80	267.80	0.00833	0.02500	2.021	0.0216	0.0532	0.9999	2316.39
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	114.93	95.92	0.06633	0.06633	0.407	0.3421	0.3892	0.3892	665.32
2	16	234.39	205.84	0.06633	0.13267	1.495	0.0674	0.4026	0.7918	2005.78
3	24	267.86	243.89	0.06633	0.19900	2.022	0.0216	0.2082	1.0000	2933.52
$E(n_{H_0}) = 156.12$			$E(n_{c,H_A}) = 194.83$			$E(n_{e,H_A}) = 170.98$		Sample Size: $n_c = n_e = 400$		

Table 3C
GSD – Weibull shape $\beta = 0.75$, equally spaced looks, RT(0.1) = 2, RT(0.9) = 1.5, JT (Pocock) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	46.35	46.35	0.00833	0.00833	2.372	0.0088	0.6332	0.6332	252.499
2	16	104.68	104.68	0.00833	0.01667	2.314	0.0103	0.3097	0.9429	660.640
3	24	121.39	121.39	0.00833	0.02500	2.079	0.0188	0.0573	1.0002	826.640
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	46.34	35.20	0.06623	0.06623	0.329	0.3711	0.3502	0.3502	303.16
2	16	104.70	88.73	0.06623	0.13246	1.503	0.0665	0.4317	0.7819	864.24
3	24	121.40	111.03	0.06623	0.19870	2.079	0.0188	0.2182	1.0001	1161.45
$E(n_{H_0}) = 68.72$			$E(n_{c,H_A}) = 87.91$			$E(n_{e,H_A}) = 74.86$		Sample Size: $n_c = n_e = 164$		

Table 3D
GSD – Weibull shape $\beta = 1.00$, equally spaced looks, RT(0.1) = 2, RT(0.9) = 1.5, JT (Pocock) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	24.08	24.08	0.00833	0.00833	2.422	0.0077	0.5948	0.5948	138.861
2	16	58.46	58.46	0.00833	0.01667	2.281	0.0113	0.3441	0.9389	337.227
3	24	66.94	66.94	0.00833	0.02500	2.105	0.0176	0.0610	0.9999	386.111
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	24.18	16.70	0.06633	0.06633	0.210	0.4169	0.3076	0.3076	171.012
2	16	58.53	48.66	0.06633	0.13267	1.470	0.0707	0.4879	0.7955	468.697
3	24	67.02	62.73	0.06633	0.19900	2.105	0.0177	0.2041	0.9996	586.726
$E(n_{H_0}) = 38.57$			$E(n_{c,H_A}) = 49.59$			$E(n_{e,H_A}) = 41.68$		Sample Size: $n_c = n_e = 86$		

Table 3E
GSD – Weibull shape $\beta = 1.25$, equally spaced looks, RT(0.1) = 2, RT(0.9) = 1.5, JT (Pocock) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	14.78	14.78	0.00833	0.00833	2.321	0.0101	0.5572	0.5572	88.429
2	16	37.60	37.60	0.00833	0.01667	2.294	0.0109	0.3797	0.9369	203.231
3	24	42.09	42.09	0.00833	0.02500	2.064	0.0195	0.0630	0.9999	220.106
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	14.85	9.45	0.06163	0.06163	0.103	0.4589	0.3141	0.3141	111.019
2	16	37.67	31.17	0.06163	0.12327	1.450	0.0735	0.4813	0.7954	295.243
3	24	42.15	40.54	0.06163	0.18490	2.066	0.0194	0.2047	1.0001	349.622
$E(n_{H_0}) = 25.20$			$E(n_{c,H_A}) = 31.36$			$E(n_{e,H_A}) = 26.27$		Sample Size: $n_c = n_e = 53$		

Table 3F
GSD – Weibull shape $\beta = 1.50$, equally spaced looks, RT(0.1) = 2, RT(0.9) = 1.5, JT (Pocock) spending for type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	10.06	10.06	0.00833	0.00833	2.346	0.0095	0.5387	0.5387	61.744
2	16	26.28	26.28	0.00833	0.01667	2.311	0.0104	0.4093	0.9480	136.052
3	24	28.79	28.79	0.00833	0.02500	2.086	0.0185	0.0519	0.9999	142.894
Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	10.05	5.96	0.06333	0.06333	0.090	0.4643	0.2813	0.2813	78.134
2	16	26.28	21.82	0.06333	0.12667	1.503	0.0664	0.5223	0.8036	203.570
3	24	28.76	28.26	0.06333	0.19000	2.086	0.0185	0.1964	1.0000	231.546
$E(n_{H_0}) = 17.67$			$E(n_{c,H_A}) = 22.21$			$E(n_{e,H_A}) = 18.62$		Sample Size: $n_c = n_e = 36$		

- *Number of simulations B = 10 000*
- *Type I error $\alpha = 0.025$ (one-sided test)*
- *Power $1 - \omega = 0.80$*
- *Median survival time in Control arm = 4 (months)*
- *Quantiles at which the effect size is defined: $p_1 = 0.10, p_2 = 0.90$.*
- *Effect size at p_1 defined as : $RT(p_1) = 2.0$.*
- *Effect size at p_2 defined as: $RT(p_2) = 1.5$.*
- *Control arm shape = $\beta_0 = 0.25$.*
- *Allocation ratio r = 1*
- *Proportion loss to follow-up = 0.20*
- *Accrual time a = 12 (months)*
- *Accrual pattern = Uniform*
- *Total time t = 24 (months)*
- *Number of looks m = 3 (equally spaced at 8, 16, and* 24 months*)*
- *Number of skips for futility (binding) = 0*
- *Alpha spending function = Hwang-Shih-DeCani (with $\rho_0 = 1$ i.e. Pocock type)*
- *Beta spending function = Hwang-Shih-DeCani (with $\rho_1 = 1$ i.e. Pocock type)*

From the results displayed in Table 3A through 3F, we see that the sample size in each arm decreases from 1889 to 400 to 164 to 86 to 53 to 36 for different values of β_0. The Pocock spending for type I and type II error results in higher stop probability under H_0 and H_A at the first look in Table 3A compared to Table 2A. Also, the Z statistic for the efficacy boundary in Table 3A is more equally spread (2.357, 2.238, and 2.056) compared to analogous values in Table 2A (3.041, 2.465 and 1.944). The Z statistic for the futility boundary is tighter at the three looks in

Table 3A (0.596, 1.604, 2.057) compared to Table 2A (−0.276, 1.001 and 1.944). Corresponding to the stop probabilities under H_0 and H_A, we see higher expected sample sizes under H_0 and H_A in Table 3Acompared to Table 2A. Similar trends are seen when comparing Table 3B vs 2B, 3C versus 2C and so on. The efficacy and futility boundaries corresponding to these tables can be plotted similar to Fig. 2 (but not shown here for brevity).

4.4. Other variations of design features

Table 4A through 4D display GSDs with variations in design features with $\beta_0 = 0.75$ using the HSD (OBF) error spending function with $RT(p_1) = 1.52$, $RT(p_2) = 1.98$. For example, Table 4A displays a GSD with 1 futility skip for three equally spaced looks at 8, 16, and 24 months. Likewise, Table 4B has 2 futility skips for four equally spaced looks at 6, 12, 18, and 24 months. Comparing these to Table 2C we see that the stop probabilities under H_0 and H_A are considerably reduced at the look times with futility skips since the only reason for stopping the trial at these look times is having overwhelming evidence of early efficacy. Next, Table 4C represents the results of a GSD with user-defined type I and type II error spending whereas Table 4D represents a GSD where the sample size calculations are done at the user defined quantile value of $p_{user} = 0.25$ instead of the default $p_{mid} = \frac{p_1+p_2}{2} = \frac{0.1+0.9}{2} = 0.5$. That is, without extra user input, the sample size calculations assume that p_{mid} is always the midpoint of p_1 and p_2. However, this need not always be the case and any user defined quantile value between p_1 and p_2 can be used to perform the sample size calculations. Since the value of $RT(0.25) = 1.6567$ is smaller than $RT(0.5) = 1.7864$ we require 168

Table 4A
GSD – Weibull shape $\beta = 0.75$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending, 1 futility skip, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	36.73	36.73	0.00130	0.00130	2.906	0.0018	0.0014	0.0014	200.133
2	16	83.05	83.05	0.00494	0.00625	2.530	0.0057	0.8243	0.8257	523.673
3	24	96.26	96.26	0.01875	0.02500	1.969	0.0245	0.1745	1.0002	654.905

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	36.65	28.15	0.00000	0.00000	–	–	0.0949	0.0949	236.318
2	16	82.93	67.48	0.04925	0.04925	0.930	0.1762	0.4696	0.5645	691.786
3	24	96.20	83.62	0.14785	0.19710	1.970	0.0244	0.4355	1.0000	956.703

$E(n_{H_0}) = 85.25$ $E(n_{c,H_A}) = 84.41$ $E(n_{e,H_A}) = 70.78$ Sample Size: $n_c = n_e = 130$

Table 4B
GSD – Weibull shape $\beta = 0.75$, 4 equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) spending, 2 futility skips, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	6	24.55	24.55	0.00080	0.00080	3.191	0.0007	0.0008	0.0008	127.293
2	12	66.60	66.60	0.00218	0.00298	2.903	0.0018	0.0014	0.0022	389.501
3	18	89.94	89.94	0.00592	0.00890	2.438	0.0074	0.8931	0.8953	582.879
4	24	98.42	98.42	0.01610	0.02500	1.989	0.0234	0.1047	1.0000	670.121

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	6	24.60	18.57	0.00000	0.00000	–	–	0.0248	0.0248	145.242
2	12	66.69	52.25	0.00000	0.00000	–	–	0.2422	0.2670	483.542
3	18	90.02	74.40	0.06783	0.06783	1.216	0.1119	0.4071	0.6741	794.129
4	24	98.46	85.44	0.12267	0.19050	1.989	0.0234	0.3259	1.0000	988.607

$E(n_{H_0}) = 90.78$ $E(n_{c,H_A}) = 85.42$ $E(n_{e,H_A}) = 71.26$ Sample Size: $n_c = n_e = 133$

Table 4C
GSD – Weibull shape $\beta = 0.75$, equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, User defined type I and II errors, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Test Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	8	40.11	40.11	0.00500	0.00500	2.593	0.0048	0.5146	0.5146	218.714
2	16	90.67	90.67	0.00750	0.01250	2.313	0.0103	0.3811	0.8957	571.414
3	24	105.13	105.13	0.01250	0.02500	1.989	0.0233	0.1041	0.9998	714.539

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	8	40.19	30.63	0.04518	0.04518	0.025	0.4901	0.2385	0.2385	257.710
2	16	90.68	73.70	0.04518	0.09035	1.202	0.1146	0.4753	0.7138	755.900
3	24	105.12	91.28	0.09035	0.18070	1.989	0.0233	0.2861	0.9999	1055.910

$E(n_{H_0}) = 66.16$ $E(n_{c,H_A}) = 82.75$ $E(n_{e,H_A}) = 68.46$ Sample Size: $n_c = n_e = 142$

Table 4D
GSD – Weibull shape $\beta = 0.75$, unequally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, p_user = 0.25, HSD (OBF) spending, 10000 simulations.

Look #	Look Times	# Events–H_0 Control Arm	# Events–H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Significance Boundary (Efficacy) Z Statistic	Upper Significance **Boundary(Efficacy) p-value**	Stop Probability Under H_0	Cumul. Stop Probability Under H_0	Cumul. Subject time Under H_0
1	12	84.20	84.20	0.00130	0.00130	2.841	0.0023	0.4396	0.4396	491.787
2	20	118.29	118.29	0.00494	0.00625	2.461	0.0069	0.4050	0.8446	782.535
3	24	124.38	124.38	0.01875	0.02500	1.902	0.0286	0.1554	1.0000	846.849

Look #	Look Times	# Events–H_A Control Arm	# Events–H_A Treatment Arm	Beta Spent	Cumul. Beta Spent	Lower Significance Boundary(Futility) Z Test Statistic	Lower Significance Boundary(Futility) p-value	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_A
1	12	84.12	66.11	0.00986	0.00986	−0.141	0.5561	0.2730	0.2730	610.790
2	20	118.23	99.65	0.03741	0.04727	0.990	0.1612	0.3674	0.6404	1096.640
3	24	124.31	108.08	0.14193	0.18920	1.902	0.0286	0.3596	1.0000	1248.230

$E(n_{H_0}) = 104.22$ $E(n_{c,H_A}) = 111.18$ $E(n_{e,H_A}) = 93.52$ Sample Size: $n_c = n_e = 168$

subjects in each of the two arms for the GSD in Table 4D compared to 130 obtained in Table 2C.

4.5. Efficacy only design

In some situations, researchers may only be interested in testing for superior efficacy of treatment over control. That is, owing to practical constraints or ethical considerations, they may not want to terminate a trial at an interim look point for futility. Such an *'efficacy only'* GSD can also be planned with some modifications to the code. See online Appendix B section 2 for the changes that need to be made to the main algorithm of online Appendix B section 1 to design such trials. Table 5A displays the results ($n_c = n_e = 114$) of an *efficacy only* design for the following combination of inputs:

- *Number of simulations B = 10 000*
- *Type I error $\alpha = 0.025$ (one-sided test)*
- *Power $1 - \omega = 0.80$*
- *Mean survival time in Control arm = 4 (months)*
- *Quantiles at which the effect size is defined: $p_1 = 0.10, p_2 = 0.90$.*
- *Effect size at p_1 defined as : $RT(p_1) = 1.52$.*
- *Effect size at p_2 defined as: $RT(p_2) = 1.98$.*
- *Control arm shape = $\beta_0 = 0.75$.*
- *Allocation ratio r = 1*
- *Proportion loss to follow-up = 0*
- *Accrual time a = 12 (months)*
- *Accrual pattern = Uniform*
- *Total time t = 24 (months)*
- *Number of looks m = 3 (equally spaced at 8, 16, and 24 months)*
- *Alpha spending function = Hwang-Shih-DeCani (with $\rho_0 = -4$ i.e. OBF type)*

Table 5B shows the results ($n_c = n_e = 136$) when $RT(p_1) = 2$, $RT(p_2) = 1.5$ and when using the Pocock-type alpha spending function for a GSD with 4 looks. All other inputs are kept the same as mentioned above. Table 5C shows results ($n_c = n_e = 127$) for another scenario when a user-defined alpha spending function is used keeping all other inputs the same. The values for $E(n_{c,H_0})$, $E(n_{c,H_A})$ and $E(n_{e,H_A})$ are also displayed below each table. The *'efficacy only'* GSD calculations were also validated using PASS 2020 for the special case of the exponential distribution ($\beta_0 = 1$) with $RT(p_1) = RT(p_2) = 1.75$ and yielded very closely matching sample sizes (but not shown here).

Table 5A
GSD – Weibull shape $\beta = 0.75$, three equally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, HSD (OBF) for type I, Efficacy Only, 10000 simulations.

Look #	Look Times	#Events –H_0 Control Arm	#Events –H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Efficacy Boundary Z Statistic	Upper Efficacy **Boundary p-value**	#Events –H_A Control Arm	#Events –H_A Treatment Arm	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_0	Cumul. Subject time Under H_A
1	8	34.91	34.86	0.00130	0.00130	2.986	0.0014	34.91	26.02	0.0737	0.0737	194.277	223.277
2	16	82.88	82.82	0.00494	0.00625	2.532	0.0057	82.88	64.91	0.4403	0.5140	538.451	682.668
3	24	99.87	99.79	0.01875	0.02500	1.992	0.0232	99.87	83.11	0.2969	0.8109	708.996	994.389

$E(n_{H_0}) = 99.66$ $E(n_{c,H_A}) = 87.60$ $E(n_{e,H_A}) = 70.89$ Sample Size: $n_c = n_e = 114$

Table 5B
GSD – Weibull shape $\beta = 0.75$, four equally spaced looks, RT(0.1) = 2, RT(0.9) = 1.5, JT (Pocock) for type I, Efficacy Only, 10000 simulations.

Look #	Look Times	#Events –H_0 Control Arm	#Events –H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Efficacy Boundary Z Statistic	Upper Efficacy **Boundary p-value**	#Events –H_A Control Arm	#Events –H_A Treatment Arm	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_0	Cumul. Subject time Under H_A
1	6	26.92	26.88	0.00625	0.00625	2.454	0.0071	26.92	19.42	0.1266	0.1266	141.428	160.59
2	12	75.47	75.49	0.00625	0.01250	2.325	0.0100	75.47	58.74	0.3750	0.5016	452.633	544.99
3	18	105.91	105.90	0.00625	0.01875	2.232	0.0128	105.91	89.13	0.2212	0.7228	709.029	906.37
4	24	119.14	119.12	0.00625	0.02500	2.140	0.0162	119.14	105.70	0.0885	0.8113	845.288	1133.58

$E(n_{H_0}) = 118.20$ $E(n_{c,H_A}) = 88.16$ $E(n_{e,H_A}) = 73.50$ Sample Size: $n_c = n_e = 136$

Table 5C
GSD – Weibull shape $\beta = 0.75$, unequally spaced looks, RT(0.1) = 1.52, RT(0.9) = 1.98, user defined type I, Efficacy Only, 10000 simulations.

Look #	Look Times	#Events –H_0 Control Arm	#Events –H_0 Treatment Arm	Alpha Spent	Cumul. Alpha Spent	Upper Efficacy Boundary Z Statistic	Upper Efficacy **Boundary p-value**	#Events –H_A Control Arm	#Events –H_A Treatment Arm	Stop Probability Under H_A	Cumul. Stop Probability Under H_A	Cumul. Subject time Under H_0	Cumul. Subject time Under H_A
1	5	18.91	18.95	0.00500	0.00500	2.523	0.0058	18.91	13.92	0.0733	0.0733	95.954	105.45
2	10	54.17	54.12	0.00750	0.01250	2.396	0.0083	54.17	40.81	0.2810	0.3543	314.083	369.27
3	24	111.19	111.21	0.01250	0.02500	2.145	0.0160	111.19	92.62	0.4475	0.8018	789.281	1107.67
$E(n_{H_0}) = 110.31$				$E(n_{c,H_A}) = 88.40$				$E(n_{e,H_A}) = 72.29$			Sample Size: $n_c = n_e = 127$		

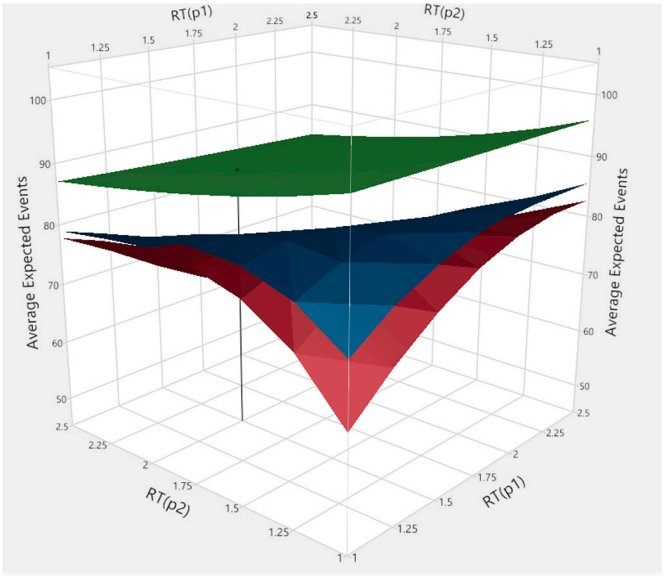


Fig. 3a. Comparing Average Expected Events of Fixed Design vs GSD with no futility skips (Front view).

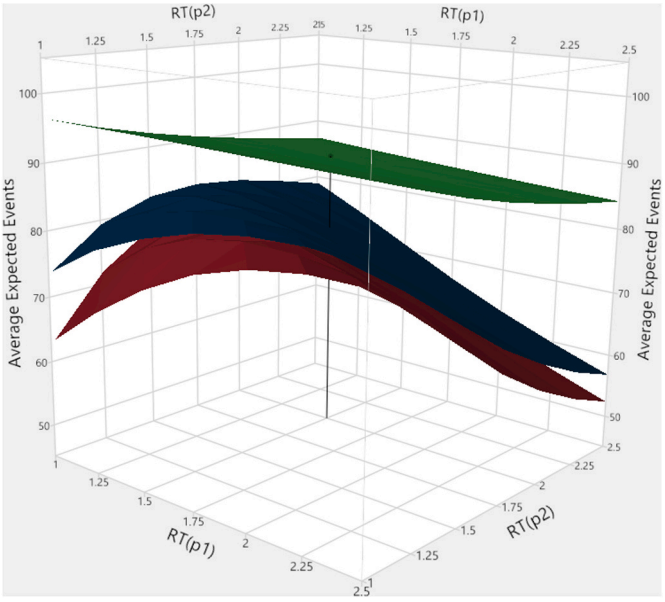


Fig. 3b. Comparing Average Expected Events of Fixed Design vs GSD with no futility skips (Side view).

4.6. Comparison with fixed sample design

One of the vital aspects of implementing a GSD is that it should be able to confer some benefits relative to a usual two-arm fixed design. In the case of a RCT involving a continuous outcome, the average sample size under the two designs can be compared for a given value of an effect size used to plan the trial. That is, for the various types of GSDs (with one or more futility skips) a statistician would be interested in knowing how much reduction (or gain) in sample size can be expected relative to a fixed 'one-look only' design if a treatment effect of a specific magnitude were used to design a study. In the case of a time-to-event endpoint, this comparison can be done in terms of average expected number of events in the control and treatment arms under the two designs. That is, if a GSD trial is stopped early or late (at the final look) for evidence of efficacy or futility, it can be thought to have a desirable operation characteristic if, on an average, it leads to less subjects being enrolled compared to a fixed design.

Fig. 3a (front view) and Fig. 3b (side view) illustrate the above-mentioned comparison for assessing reduction in expected sample size for a 'fixed design' versus a '3 looks efficacy & futility (with no skips) design'. Consider the following design settings for a fixed two-arm design.

- *Type I error $\alpha = 0.025$ (one-sided test),*
- *Power $1 - \omega = 0.80$,*
- *Median survival time in Control arm = 4 (months),*
- *Quantiles at which the effect size is defined $p_1 = 0.10, p_2 = 0.90$,*
- *Effect size at p_1 defined as : $RT(p_1) = 1.50$,*
- *Effect size at p_2 defined as: $RT(p_2) = 2$,*
- *Control arm shape $= \beta_0 = 0.75$,*
- *Allocation ratio $r = 1$,*
- *Proportion loss to follow-up = 0.20,*
- *Accrual time $a = 12$ (months),*
- *Accrual pattern = Uniform,*
- *Total time $t = 24$ (months)*

Under a fixed two-arm trial, N = 139 subjects need to be recruited in each of the two arms for the above-mentioned inputs using the method of Phadnis and Mayo [17] to detect an effect size of $RT(p_{mid}) = 1.788$ with 80 % power resulting in n = 88.606 events (accounting for administrative censoring as well as 20 % loss to follow-up). The n = 88.606 events are indicated by the vertical black line (needle) in Fig. 3a and b. Once this fixed two-arm trial is started, the enrollment of N = 139 subjects cannot be altered. However, the actual effect size observed in the trial can be different from the effect size used for planning the trial and this can affect the actual number of events observed in the trial. The green surface indicates the number of events that would be observed in this fixed two-arm trial as a function of the actual size of the treatment effect. Expectedly, this green surface peaks at the null hypothesis ($RT(p_1) = RT(p_2) = 1$) with n = 96.358 and is lowest at extreme values under the alternate hypothesis ($RT(p_1) = RT(p_2) = 2.5$) with n = 84.244.

Now suppose that we decide to conduct a GSD with the same effect size specifications as earlier. We consider two scenarios of planning this

GSD with the design specifications.

- *Number of looks m = 3 (equally spaced at 8, 16,* 24 months*),*
- *Number of skips for futility = 0*

Scenario #1

- *Alpha spending function = Hwang-Shih-DeCani (with* $\rho_0 = -4$ *i.e. OBF type)*
- *Beta spending function = Hwang-Shih-DeCani (with* $\rho_1 = -4$ *i.e. OBF type)*

Scenario #2

- *Alpha spending function = Hwang-Shih-DeCani (with* $\rho_0 = 1$ *i.e. Pocock type)*
- *Beta spending function = Hwang-Shih-DeCani (with* $\rho_1 = 1$ *i.e. Pocock type)*

Under the OBF design plan (Scenario #1), we need to enroll N = 130 subjects (consistent with the sample size result shown in Section 4.2 for $\beta = 0.75$ and the OBF plan) resulting in n = 77.202 events using the previously mentioned design specifications. Again, once the trial starts after enrollment of N = 130 subjects, the expected number of events will depend on the actual effect sizes observed in the study. These expected number of events are represented by the curved blue surface in Fig. 3a and b. Under the null hypothesis ($RT(p_1) = RT(p_2) = 1$), the trial is likely to stop early for futility (note there are no futility skips) resulting in n = 73.922 events. In this case, the stop probabilities at the three looks are 0.2257, 0.5584 and 0.1858 respectively. Under the most extreme case of the alternative hypothesis ($RT(p_1) = RT(p_2) = 2.5$) also, the trial is likely to stop early for efficacy resulting in n = 56.822 events. In this case, the stop probabilities at the three looks are 0.3652, 0.5745 and 0.0602 respectively. Thus, the trial has a higher chance of getting halted at the first look when $RT(p_1) = RT(p_2) = 2.5$ compared to when $RT(p_1) = RT(p_2) = 1$ thereby reducing the average expected sample size. Under other alternative hypothesis scenarios not as extreme as the above, average expected number of events depend on the corresponding stop probabilities and are well captured by the blue surface. For example, when $RT(p_1) = 1.75$ and $RT(p_2) = 1.25$, the average expected sample size is 84.687 resulting from stop probabilities of 0.0713, 0.3386 and 0.5900 at the three looks respectively. The expected sample size reduction for the blue surface relative to the green surface ranges from 7.94 % to 32.55 %.

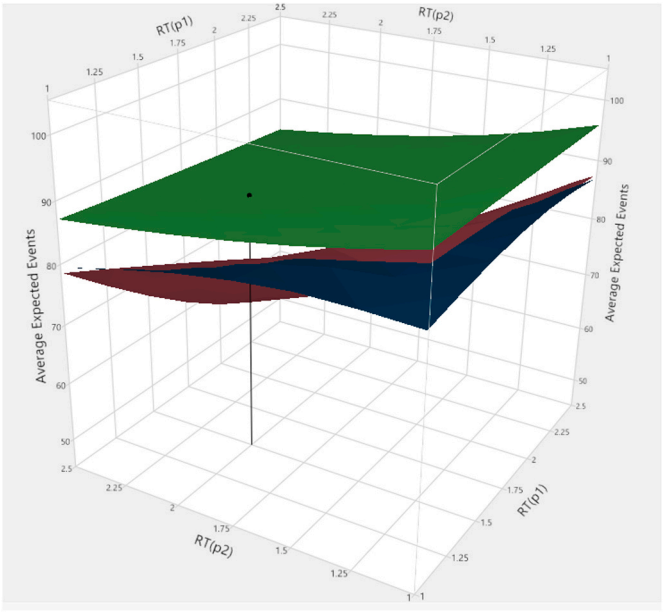


Fig. 3c. Comparing Average Expected Events of Fixed Design vs GSD with 1 futility skip (Front view).

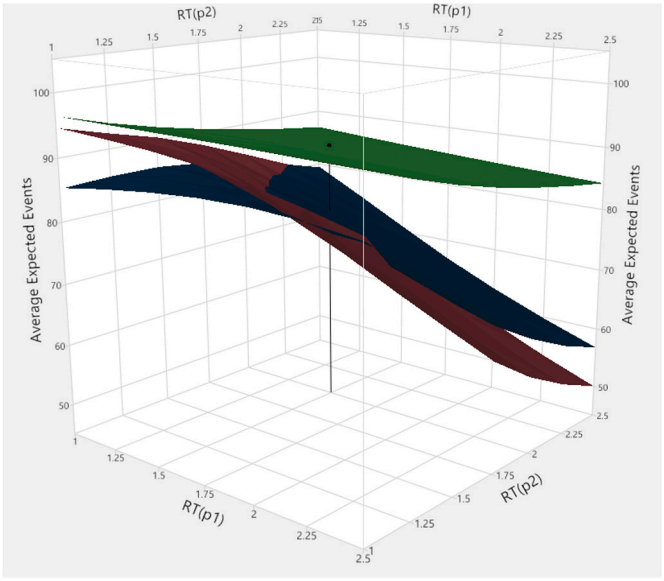


Fig. 3d. Comparing Average Expected Events of Fixed Design vs GSD with 1 futility skip (Side view).

Under the Pocock design plan (Scenario #2), we need to enroll N = 152 subjects resulting in n = 75.201 events using the previously mentioned design specifications. The average expected sample size for this scenario is represented by the red colored surface. Here we see that the red surface is always below the blue surface despite enrolling more subjects (152 vs 130) in the study. This is a consequence of higher stop probabilities at the first look for the Pocock plan compared to the OBF plan. For example, in the Pocock plan, under the null hypothesis the stop probabilities at the three looks are 0.6510, 0.2880, and 0.0611 respectively resulting in an average expected sample size of 63.347. Likewise, under the alternate hypothesis represented by $RT(p_1) = RT(p_2) = 2.5$ the stop probabilities at the three looks are 0.6388, 0.3471, and 0.0140 respectively resulting in an average expected sample size of 52.421. Thus, the expected sample size reduction for the red surface relative to the green surface ranges from 9.79 % to 37.77 %.

Fig. 3c (front view) and Fig. 3d (side view) illustrates comparison of a fixed design versus a GSD for the same design specifications with one notable change – futility is skipped at the first look. The green surface (fixed design) is the same as before but the blue (OBF) and red (Pocock) surfaces are different owing to there being no testing for futility at the first look. The blue and red surfaces are seen to cross each other, and this is a consequence of the stop probabilities under the various effect size combinations of $RT(p_1)$ and $RT(p_2)$. The number of subjects to be enrolled under the fixed design, GSD with OBF, and GSD with Pocock are N = 138, N = 130 and N = 148 respectively. However, under the null hypothesis, the stop probabilities for the OBF plan at the three looks are 0.0013, 0.8244 and 0.1742 respectively resulting in an average expected sample size of 85.358. In contrast, under the null hypothesis, the stop probabilities for the Pocock plan at the three looks are 0.0092, 0.9535 and 0.0374 respectively resulting in an average expected sample size of 94.564. That is, both designs yield a very low stop probability at the first look (due to efficacy only as futility testing has been skipped) but the higher stop probability at the second look under the Pocock plan in combination with the number of events observed in the control and treatment arm results in a larger average expected sample size compared to the OBF plan. Conversely, under the alternate hypothesis represented

by $RT(p_1) = RT(p_2) = 2.5$, the stop probabilities for the Pocock plan at the three looks are 0.6604, 0.3242 and 0.0153 respectively resulting in an average expected sample size of 50.078. Corresponding values of stop probabilities under the OBF plan are 0.3649, 0.5750 and 0.0600 which when combined with the number of events observed in the two arms result in an average expected sample size of 56.831. This explains the crossing of the two surfaces owing to the manner in which type I and type II errors are spent at each of the three looks.

Overall, both error-spending plans indicate a considerable savings in average expected number of events – OBF ranges from 7.94 % to 32.55 % for no futility skips, Pocock ranges from 9.79 % to 37.77 % for no futility skips, OBF ranges from 6.41 % to 32.54 % for one futility skip, Pocock ranges from 1.74 % to 40.56 % for one futility skip. These results demonstrate the advantages of using the proposed GSD.

5. Concluding remarks

In this work, we have shown how a GSD can be implemented for a phase III trial in the case of a time-to-event endpoint under non-proportional hazards when survival times in the two arms come from two different Weibull distributions. The different shape and scale parameters of the two Weibull distributions allow for handling both non-PH and non-PT designs, which is an advantage over the previously proposed GSD method of Phadnis and Mayo [16] that require the PT assumption (AFT model). The proposed method is well suited for a phase III trial as it requires a reasonably accurate estimate of the control arm Weibull shape parameter, and this can be obtained with ease from a previously conducted moderate-sized phase II trial (as discussed in Step 2 of online Appendix B section 2). The proposed method relies on the error-spending approach and the asymptotically normal test statistic makes it easy to calculate the efficacy and futility boundaries both on the z-scale as well as the p-value scale. One of the main advantages of the proposed method is that it allows decision making at the interim looks on the very intuitive and patient-friendly interpretation of treatment benefit measured through the lens of *improvement in longevity*. The algorithm for implementing the GSD is straightforward and poses no computational difficulties since it is based on the Weibull distribution – which allows stable estimates even when there a smaller number of events at the first look. The method also allows crossing of two survival curves (see Step 1 in online Appendix B section 1) with some restrictions. It can be implemented for both scenarios – improving or diminishing treatment benefits – as indicated by the two examples discussed earlier.

Many different methods of *analyzing* specific cases of time-to-event data with non-proportional hazards have been proposed in literature. However, when it comes to *designing* a clinical trial with the non-proportional hazards, only a few methods are available in popular software. Even when such methods are available, they are restricted to *fixed two-arm trials,* and it is not clear how they can be extended to the sequential testing framework. For example, popular commercial software like PASS [14], nQuery [12] and SAS [23] as well as free-to-use software like R allow construction of a GSD only under the PH assumption with an implicit assumption of exponentially distributed times that allow interchangeable inputs of constant hazard rate, median time, or proportion of surviving at a given time. Our proposed method for the non-PH as well as non-PT scenarios is therefore a timely addition to the limited methods available in literature for GSDs. Also, the incorporation of Weibull distribution allows more flexibility in modeling hazards that are increasing over time or decreasing over time with the exponential distribution acting as a special case of constant hazards.

Although, the method in its current form requires the survival times to follow Weibull distributions – a limitation of our approach, it is a step in the right direction facilitating interim testing and therefore decision making when results from a previous trial indicate that the underlying biological process can be well modeled by a Weibull distribution. Our approach in using a simulation-based error-spending algorithm is to provide a platform for motivating future research in the field of complex sequential and adaptive designs with more advanced features. As an example, our method could be extended to the field of multi-arm clinical trials where it is unlikely that the PH assumption will hold between the different treatment arms. In such a case, a GSD based on the concept of Relative Time could prove very useful as it could potentially provide decision rules at the interim looks to implement *arms dropping* (worse performing arms could be dropped at an interim) and *response adaptive randomization* (new patients could be allocated to those arms which are seen to perform better at the interim). The combination of parametric approach along with a simulation-based approach could also potentially allow incorporation of multi-center (random) effects, delayed entry (left-truncation), non-uniform accrual patterns and many such features. Overall, we anticipate that researchers and statisticians will find our proposed method as useful in making decisions when working with phase III trials with time-to-event endpoints in the case of non-proportional hazards.

CRediT authorship contribution statement

Milind A. Phadnis: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Nadeesha Thewarapperuma:** Writing – review & editing, Validation, Software. **Matthew S. Mayo:** Supervision, Project administration, Investigation.

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.

Acknowledgements

This study was supported by a NIH Clinical and Translational Science Award grant (UL1 TR002366) awarded to the University of Kansas Medical Center.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2024.101315.

References

[1] P. Armitage, Sequential Medical Trials, first ed., Thomas, Springfield, 1960.
[2] A. Dmitrienko, G. Molenberghs, C. Chuang-Stein, W. Offen, *Analysis Of Clinical Trials Using SAS: A Practical Guide*. Cary, SAS Press, 2005.
[3] S.S. Ellenberg, T.R. Fleming, D.L. DeMets, Data Monitoring Committees in Clinical Trials: A Practical Perspective, Wiley, New York, 2002.
[4] G.G. Enas, B.E. Dornseif, C.B. Sampson, J. Wuu, *Monitoring versus interim analysis of clinical trials: Perspective from the pharmaceutical industry*, Control, Clin. Trials 10 (1989) 57–70.
[5] C. Jennison, B.W. Turnbull, Statistical approaches to interim monitoring of medical trials: a review and commentary, Stat. Sci. 5 (1990) 299–317.
[6] C. Jennison, B.W. Turnbull, Group Sequential Methods with Applications to Clinical Trials, Chapman & Hall/CRC, Boca Raton, 2000.
[7] Z. Jiang Z, L. Wang, C. Li, J. Xia, H. Jia, A practical simulation method to calculate sample size of group sequential trials for time-to-event under exponential and Weibull distribution, PLoS One 7 (2012) 1–12.
[8] K. Kim, D.L. DeMets, Design and analysis of group sequential tests based on the type I error spending rate function, Biometrika 74 (1987) 149–154.
[9] J.P. Klein, M.L. Moeschberger, Survival Analysis – Techniques for Censored and Truncated Data, second ed., Springer, New York, 2003.
[10] K.K.G. Lan, D.L. DeMets, Discrete sequential boundaries for clinical trials, Biometrika 70 (1983) 659–663.
[11] M. Mazumdar, H. Bang H, Sequential and Group Sequential Designs in Clinical Trials: Guidelines for Practitioners; Handbook of Statistics 27; Series Title: Epidemiology and Medical Statistics, Elsevier, North Holland, 2008.
[12] nQuery Sample Size and Power Calculation, 'Statsols' (Statistical Solutions Ltd), Cork, Ireland, 2017.
[13] P.C. O'Brien, T.R. Fleming, A multiple testing procedure for clinical trials, Biometrics 35 (1979) 549–556.

[14] PASS 20 Power Analysis and Sample Size Software, NCSS, LLC, Kaysville, Utah, USA, 2015 ncss.com/software/pass.
[15] M.A. Phadnis, Sample size calculation for small sample single-arm trials for time-to-event data: logrank test with normal approximation or test statistics based on exact chi-square distribution? Contemp. Clin. Trials Commun. 17 (2019) 1–13, https://doi.org/10.1016/j.conctc.2020.100548.
[16] M.A. Phadnis, M.S. Mayo, Group sequential design for time-to-event data using the concept of Proportional Time, Stat. Methods Med. Res. 29 (2020) 1867–1890.
[17] M.A. Phadnis, M.S. Mayo, Sample size calculation for two-arm trials with time-to-event endpoint for non-proportional hazards using the concept of Relative Time when inference is built on comparing Weibull distributions, Biom. J. 63 (2021) 1406–1433.
[18] M.A. Phadnis, J.B. Wetmore, M.S. Mayo, A clinical trial design using the concept of proportional time using the generalized gamma ratio distribution, Stat. Med. 36 (2017) 4121–4140.
[19] S.J. Pocock, Group sequential methods in the design and analysis of clinical trials, Biometrika 64 (1977) 191–199.
[20] M.A. Proschan, K.K.G. Lan, J.T. Wittes, Statistical Monitoring of Clinical Trials, Springer, New York, 2006.
[21] D.M. Reboussin, D.L. DeMets, K. Kim, K.K.G. Lan, Programs for Computing Group Sequential Bounds Using the Lan-DeMets Method. Technical Report No. 60, UWM, Madison, USA, 1992.
[22] L.A. Renfro, L. Ji, J. Piao, A. Onar-Thomas, J.A. Kairalla, T.A. Alonzo, Trial Design Challenges and Approaches for Precision Oncology in Rare Tumors: Experiences of the Children's Oncology Group, JCO Precision Oncology, 2019, pp. 1–13. Available at: https://ascopubs.org/doi/pdf/10.1200/PO.19.00060.
[23] SAS 9.4, SAS Institute Inc., Cary, NC, USA, 2017.
[24] R. Sposto, D.O. Stram, A strategic view of randomized trial design in low-incidence paediatric cancer, Stat. Med. 18 (1999) 1183–1197.
[25] S. Todd, A 25-year review of sequential methodology in clinical trials, Stat. Med. 26 (2007) 237–252.
[26] A. Wald, Sequential Analysis, Wiley, New York, 1947.
[27] S.K. Wang, A.A. Tsiatis, Approximately optimal one-parameter boundaries for group sequential trials, Biometrics 43 (1987) 193–199.
[28] G. Wassmer, W. Brannath, Group Sequential and Confirmatory Adaptive Designs in Clinical Trials, Springer, Switzerland, 2016.
[29] J. Whitehead, The Design and Analysis of Sequential Clinical Trials, second ed., Wiley, Chichester, 1997.
[30] J. Whitehead, A unified theory for sequential clinical trials, Stat. Med. 18 (1999) 2271–2286.
[31] J. Wu, Sample size calculation for the one-sample log-rank test, Pharmaceut. Stat. 14 (2015) 26–33.
[32] J. Wu, X. Xiong, Group sequential design for randomized trials under the Weibull model, J. Biopharm. Stat. 25 (2015) 1190–1205.