Evaluation of Performance of Adaptive Designs Based on Treatment Effect Intervals

Fang Fang[1], Yong Lin[2,3], Weichung Joe Shih[2,3], Shou-En Lu[2,3] & Guangrui Zhu[4]

[1] Kiniksa Pharmaceuticals, 100 Hayden Ave. Lexington, MA 02421, USA

[2] Department of Biostatistics, School of Public Health, Rutgers, The State University of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA

[3] Division of Biometrics, Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08901, USA

[4] Allergan Inc., Irvine, CA 92612, USA

Correspondence: Yong Lin, Department of Biostatistics, School of Public Health, Rutgers, The State University of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA.

Received: July 18, 2018 Accepted: August 28, 2018 Online Published: September 17, 2018

doi:10.5539/ijsp.v7n6p81 URL: https://doi.org/10.5539/ijsp.v7n6p81

Abstract

The accuracy of the treatment effect estimation is crucial to the success of Phase 3 studies. The calculation of sample size relies on the treatment effect estimation and cannot be changed during the trial in a fixed sample size design. Oftentimes, with limited efficacy data available from early phase studies and relevant historical studies, the sample size calculation may not accurately reflect the true treatment effect. Several adaptive designs have been proposed to address this uncertainty in the sample size calculation. These adaptive designs provide flexibility of sample size adjustment during the trial by allowing early trial stopping or sample size adjustment at interim look(s). The use of adaptive designs can optimize the trial performance when the treatment effect is an assumed constant value. However in practice, it may be more reasonable to consider the treatment effect within an interval rather than as a point estimate. Because proper selection of adaptive designs may decrease the failure rate of Phase 3 clinical trials and increase the chance for new drug approval, this paper proposes measures and evaluates the performance of different adaptive designs based on treatment effect intervals, and identifies factors that may affect the performance of adaptive designs.

Keywords: group sequential design, Haybittle-Peto boundary, Pocock boundary, sample size re-estimation design

1. Introduction

It is well-reported that the cost of drug development keeps rising at a high rate while the new drug applications do not keep up with the same rate (Lesko, 2006). It was estimated that the failure rate for Phase 3 trials exceeds 50% (Chuang-Stein, 2004). A poorly designed Phase 3 trial may be a likely reason to account for the high failure rate. It costs both money and patient lives (Thoelke, 2007). The accuracy of the treatment effect assumption is crucial to the success of Phase 3 studies. The calculation of sample size in fixed sample size (FS) designs relies on the assumption of treatment effect and cannot be changed during the trial. With limited efficacy data available from early phase or other relevant historical studies, the sample size calculation may not accurately reflect the true treatment effect. This lack of knowledge leads to the calculated sample size either too small or too large. Thus, the trial may be either oversized or underpowered. The results could be either a waste of finances and patient resources or trial failure.

Many adaptive designs have been proposed to address this issue (e.g., Armitage et al., 1969; Cui et al., 1999; Gould and Shih, 1992 and 1998; Hybittle, 1971; Jennison and Turnbull, 1990 and 2006; Lan and DeMets, 1994; Li et al., 2002 Li et al., 2005; Liu et al., 2008; OBrien and Fleming, 1979; Peto et al., 1976; Pocock, 1977; Proschan et al., 2006; Tsiatis and Mehta, 2003; Wittes and Brittain, 1990). A broad definition of adaptive designs given by Shih (2006) is used in this paper. All classical group sequential (GS) designs and sample size re-estimation (SSR) designs fall into adaptive design scope per this definition. For (classical) GS designs, under a pre-specified total number of looks and maximum sample size, a study is allowed to stop for efficacy or futility at the interim analysis but no extension is allowed beyond the pre-specified maximum sample size. An adaptive GS design is a two-phase GS design with sample size adjustment as illustrated in Figure 1G, which conducts the sample size adjustment at the j^{th} interim analysis (phase 1 portion of the trial), then the remaining duration of the study with the modified sample size is the phase 2 portion. It has been called SSR design in the group sequential setting in the literature (e.g., Gould and Shih, 1998; Cui et al., 1999). We simply call it SSR design

in this paper since it is equivalent to the SSR design with the sample size adjustment at the first look ($j = 1$). Shih et al. (2016) gave a review on popular weighted and unweighted SSR designs. In all adaptive designs, since interim analyses are built into the traditional studies and the results at the interim look(s) are used to adjust the future course of the study, the overall type I error rate and adequate power or conditional power need to be maintained in all adaptive designs.

Currently, the discussion in the literature on adaptive designs mostly focused on a single specified value as a representation of the unknown treatment difference. However, in practice what is often known is a treatment effect interval (Liu et al., 2008), as illustrated in the following.

It is believed that if an experimental drug is added to the current standard therapy (Carboplatin plus Paclitaxel), the remission time for ovarian cancer patients after surgery will be prolonged. A Phase 3 confirmatory trial is planned to compare the treatment effect of the combination therapy versus the standard therapy alone. Progression free survival (PFS) is used as the primary efficacy endpoint. The treatment effect is estimated based on the results of Phase 2 proof of concept studies for the experimental drug and published median PFS for the standard treatment. However, different PFS medians for standard therapy are found in the literature. In the Hellenic Cooperative Oncology Group (HeCOG) study (Aravantinos et al., 2005), the median PFS of 121 patients randomized to the standard (Carboplatin plus Paclitaxel) therapy was 38 months. In the Gynecologic Oncology Group (GOG) study, the median PFS of 392 patients receiving standard therapy was 19.4 months (Ozols et al., 2003). In another Phase 3 study supported by Bristol-Meyers Squibb, the median PFS of patients receiving standard therapy was 16 months (Neijt et al., 2000). Two other randomized trials indicated the median PFS was around 17.5 months (DuBois et al., 2003; Parmar et al., 2002). The standard therapy regimens including dose levels and dose frequencies were similar in these studies. There were also differences in cancer stages and tumor sizes among the patient populations enrolled into these studies. Thus, it became very difficult to find an accurate point estimate of the true median PFS for the standard therapy. After careful comparison of study designs including inclusion/exclusion criteria and treatment schedules, the median PFS for standard therapy was most likely between 15 to 20 months.

Most previous research and designs focused on how to maximize the study efficiency when the treatment effect was estimated as a point value. From the example above, the question arises as to how the current available adaptive designs can be used to maximize the study efficiency on a treatment effect interval? How can we use mathematical framework to evaluate the performance of different designs? What factors will affect the performance of adaptive designs? Under the same constrains, whether certain GS designs (e.g., with different boundaries or different sample size increments) have a similar performance as SSR designs?

In this paper, we address these issues. To be specific, we develop a method to evaluate and compare the performance of adaptive designs including GS, weighted and unweighted SSR designs on a pre-specified treatment effect interval. We first introduce the measures and indicators to be used for evaluating adaptive design performance in Section 2. The performance of adaptive designs when the treatment effect follows a uniform and in general, a location-scaled beta distribution are discussed in Section 3 and Section 4, respectively. Discussion and conclusion are given in Section 5.

2. Methods of Evaluation of Adaptive Design Performance

2.1 Determination of Treatment Effect Interval

Treatment effect interval should be determined based on the combination of multiple considerations. The lower limit of treatment effect interval should be based on (1) clinical meaningful treatment difference, (2) medical policies, such as restriction of medication price, and (3) company financial considerations. The upper limit of treatment effect interval should be based on (1) the minimum number of patients needed for adequate safety evaluation of the test drug and (2) a realistic estimate of the largest treatment difference.

2.2 Measures of Performance

Several authors have evaluated adaptive designs in different contexts. For example, Xi et al. (2017) investigated the optimal timing of interim analyses for making the futility decision. Chen et al. (2016) proposed a biomarker-based subgroup analysis at the end of a Phase III trial to fine-tune the statistical design including hypothesis adjustment. Levin et al. (2013) compared adaptive designs with the test statistics which are based on minimal sufficient statistic, thus included GS and unweighted SSR designs but excluded the weighted SSR design (Cui et al. 1999). All these literature used average sample size to compare the performance of different adaptive designs. In this paper, we use functions of the true treatment effect δ on the treatment effect interval $[\delta_L, \delta_U]$ as the measures and construct performance indicators based on the measures. When comparing the sample size, this function is

$$u(\delta) = \frac{2\left(z_\beta + z_{\alpha/2}\right)^2}{\delta^2},$$

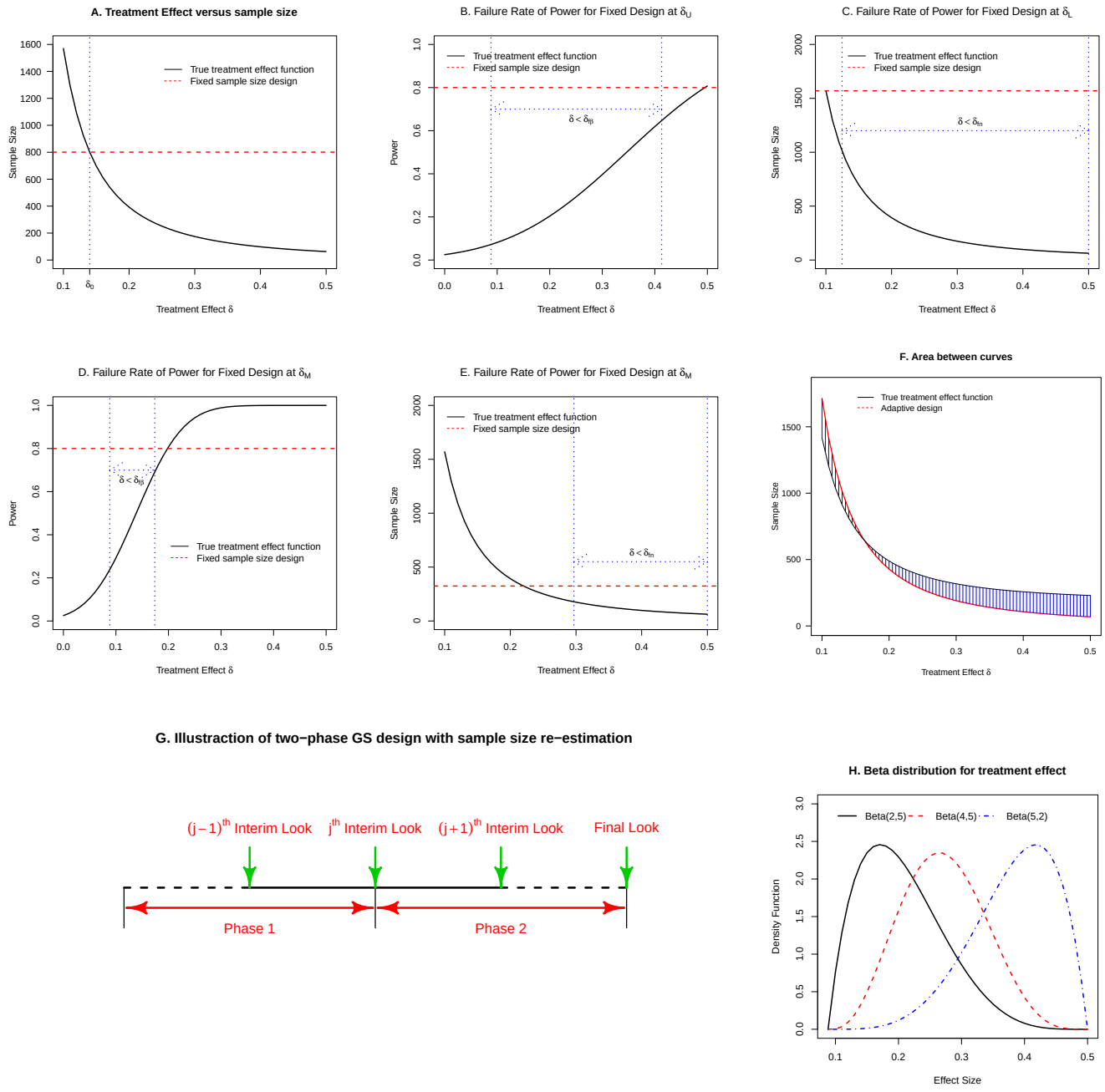

Figure 1. Concept Illustrations

where α and β are the pre-specified type I error and type II error for the study. When comparing the power, this function is

$$p(\delta) = 1 - \beta.$$

These functions can be used separately or be combined as seen in the sequel. Let us first illustrate with the FS design.

2.3 True Treatment Effect Function for the Fixed Size Design

Without knowing the true treatment effect δ, the sample size for FS design has to be calculated based on an estimated effect $\hat{\delta}$ from a previous study or historical data. Sample size cannot be changed during the study. Sample size curves as function of the true treatment effect for the FS design are illustrated in Figure 1A. Only when $\hat{\delta}$ is equal to true treatment effect δ, the FS design will have the ideal performance. When $\hat{\delta}$ is shifted away from the δ, the FS design will be either under powered (if $\hat{\delta} > \delta$) or oversized (if $\hat{\delta} < \delta$). Its performance will be worse as the difference between the true and estimated value gets larger.

2.4 Indicators to Compare Performance

2.4.1 Failure Rate

Failure and Failure Rate:

At each point on the treatment effect interval, it is considered a failure when the sample size for a particular design at that point is more than $\frac{1}{f_s}$ (usually $\frac{1}{f_s} = 2$) times the sample size based on the true treatment effect or the power decreased more than f_p (usually 20%) of the power based on the true treatment effect, where $0 < f_s < 1$ and $0 < f_p < 1$. For example, when $f_p = 20\%$ and the nominal power is 80%, failure occurs if the power of the design is no more than 64% $(= 80\% \times (1 - 0.2))$. The failure rate is defined as the proportion of points that meet the failure criteria on the treatment effect interval. A lower failure rate indicates a better performance of the adaptive design.

Failure Rate for Fixed Sample Size Design:

Denote the sample size at the true treatment effect δ as $u(\delta)$ and n_0 the sample size calculated based on the FS design. Based on the above criterion, a failure occurs when n_0 is larger than $\frac{1}{f_s}u(\delta)$. Because the sample size curve for the true treatment effect is monotone, it is very easy to see that failure occurs for all $\delta > \delta_{f_n}$ on the interval, where $\delta_{f_n} = \sqrt{\frac{2\left(Z_{\alpha/2}+Z_{\beta}\right)^2}{f_s n_0}}$.

The targeted power on the interval for true treatment effect function is always $1 - \beta$. Using the above criterion, it is a failure when power is decreased more than f_p times of the targeted power. Also, because of the monotonicity of power curves given the sample size n_0 calculated based on the FS design, it can be seen that the failure occurs for all $\delta < \delta_{f_\beta}$ on the interval, where $\delta_{f_\beta} = \sqrt{\frac{2\left(Z_{\alpha/2}+Z_{1-f_p(1-\beta)}\right)^2}{n_0}}$.

Thus the total failure rate on a treatment effect interval that combines both sample size and power measures is defined as:

$$R_f = \frac{\left(\delta_U - \delta_{f_n}\right) + \left(\delta_{f_\beta} - \delta_L\right)}{\delta_U - \delta_L} \times 100\%.$$

Below is an example of failure rate for FS designs. When treatment effect interval is $[0.0882, 0.5]$ and assume $f_s = \frac{1}{2}$ and $f_p = 0.2$, the failure rates at different sample sizes for FS designs are

1. Failure rate for a FS design with sample size calculation at δ_U:

   $R_f = (0.4131 - 0.0882)/(0.5 - 0.0882) \times 100\% = 78.9\%$. As indicated in Figure 1B, when treatment effect falls into the region $(0.0882 < \delta < 0.4131)$, power decreases more than 20% of the true treatment effect function for power $(1 - \beta = 80\%)$.

2. Failure rate for a FS design with sample size calculation at δ_L:

   $R_f = (0.5 - 0.1247)/(0.5 - 0.0882) \times 100\% = 91.1\%$. As indicated in Figure 1C, when treatment effect falls into the region $(0.1247 < \delta < 0.5)$, sample size is more than two times of the sample size from the true treatment function.

3. Failure rate for a FS design with sample size calculation at $\delta_M = \sqrt{\delta_L \delta_U}$:

   $R_f = [(0.5 - 0.2968) + (0.1738 - 0.0882)]/(0.5 - 0.0882) * 100\% = 70.1\%$. As indicated in Figure 1D and Figure 1E, when treatment difference falls into the red region $(0.0882 < \delta < 0.1738)$, power decreases more than 20% of

the power from the true treatment effect function; when treatment difference falls into the red region ($0.2968 < \delta < 0.5$), sample size is more than two times of the sample size from the true treatment effect function.

Generalization of Failure Rate:

A generalized formula for failure rate on a treatment effect interval is:

$$R_f = \int_{\delta_L}^{\delta_U} I_{\left(\frac{g(\delta)}{u(\delta)} > \frac{1}{f_s} \text{ or } 1-\frac{f(\delta)}{p(\delta)} > f_p\right)}(\delta)\, W(\delta)\, d\delta,$$

where $g(\delta)$ and $f(\delta)$ denote the sample size and power for adaptive design when treatment effect is δ, $u(\delta)$ and $p(\delta)$ denote the sample size and power from the true treatment effect function when treatment effect is δ, and $I_A(\delta)$ is the indicator function of A. $W(\delta)$ denotes the weight assigned to $\delta \in [\delta_L, \delta_U]$ which will be a probability density function on the treatment effect interval. When treatment effects have a uniform distribution over $[\delta_L, \delta_U]$, $W(\delta)$ is equal to $1/(\delta_U - \delta_L)$. Treatment effect also can be assumed to follow other distributions. In this paper, we will consider the treatment effect follow a uniform and a location-scaled beta distribution in Sections 3 and 4, respectively. The generalized failure rate is used as an indicator of performance in the paper. For simplicity, we just call it failure rate in the sequel.

2.4.2 Area between Log Curves (ABLC)

Loss function:

Loss function (or regret) in decision theory can also be used to evaluate the performance of a design at the true treatment effect δ on the interval $[\delta_L, \delta_U]$. The ratio of the upper limit to lower limit of the interval is called adaptive index (AI), i.e., $AI = \delta_U/\delta_L$. We consider the absolute loss function:

$$l\left(\hat{S}(\delta); S(\delta)\right) = \left|\hat{S}(\delta) - S(\delta)\right|, \delta \in [\delta_L, \delta_U],$$

where $S(\delta)$ is the sample size or power at the true δ and $\hat{S}(\delta)$ is the estimated sample size or the power based on the design. The smaller the loss is, the better the adaptive design performance is. In the following, we consider the loss in terms of sample size only. The study of the loss in terms of power is available from the authors upon request.

Area Between Curves:

Loss only accounts for the performance of adaptive design at a particular point on a treatment effect interval. It is important to identify a criterion to account for the cumulative performance on the treatment effect interval. The risk function is the average of loss over the treatment effect interval. Since the length of the treatment effect interval is the same, comparisons of risks is the same as comparisons of the areas between curves as defined in the following.

Area Between Curves (ABC) of Adaptive Design and True Treatment Effect Function:

To account for the deviation from the true treatment effect function, ABC for each adaptive design can be calculated on the treatment effect interval (see Figure 1F) by

$$ABC = \int_{\delta_L}^{\delta_U} \left|\hat{S}(\delta) - S(\delta)\right| d\delta.$$

Performance can be evaluated by comparing ABC for different designs. The smaller the area between the curves is, the better the performance is.

Area between Log Curves (ABLC) for Adaptive Design and True Treatment Effect Function:

One can interpret $\hat{S}(\delta) - S(\delta)$ as the sample size difference between the adaptive design and the true treatment effect function. More appropriately though, the ratio of $\hat{S}(\delta)$ to $S(\delta)$ indicates the relative difference. For easier interpretation, log ratio can be calculated. We have

$$\log\left(\frac{\hat{S}(\delta)}{S(\delta)}\right) = \log(\hat{S}(\delta)) - \log(S(\delta)).$$

Define ABLC as

$$\text{ABLC} = \int_{\delta_L}^{\delta_U} \left|\log(\hat{S}(\delta)) - \log(S(\delta))\right| d\delta.$$

In addition to the failure rate R_f, ABLC is used as another performance indicator of a design in Sections 3 and 4.

3. Performance of Adaptive Designs when Treatment Effect Follows a Uniform Distribution

3.1 Method

In this section, the performance of (classical) GS designs, weighted SSR designs, and unweighted SSR designs on a treatment effect interval are evaluated, assuming that the treatment effect follows a uniform distribution. CHW design (Cui et al., 1999) is used as the representative of the weighted SSR design. For the unweighted (likelihood) SSR design (e.g., Li et al., 2002; Shih et al., 2016), boundaries for the interim looks after the sample size adjustment are recalculated to maintain the overall type I error rate. For weighted and unweighted SSR designs, an initial sample size is selected based on the sample size at δ_L, δ_U, and δ_M. Sample size can then be increased based on the interim finding on treatment effect size during the study. However, there is a restriction for the maximum allowed sample size. Sample size adjustment can be done based on the conditional power at selected looks before the final analysis. Effects on the patient increment patterns and different adaptive indices are also studied.

For the (classical) GS and weighted SSR designs, boundaries at each interim and final looks are fixed at the design stage through pre-specified alpha-spending. Four kinds of discrete boundaries - O'Brien and Fleming boundaries (OBF), Pocock boundaries (PK), Haybittle-Peto boundaries with critical value α_0 of 0.01 (HP01), and with critical value α_0 of 0.005 (HP005) are considered in the performance comparisons. Boundaries are calculated from exact methods (as opposed to the approximation of alpha-spending function for continuous boundaries). More specifically, let K be the total number of looks, $t_1, t_2, \cdots, t_K$ denote information fraction at each look, δ denotes the treatment effect, and $Z_{t_1}, Z_{t_2}, \cdots, Z_{t_K}$ denote the test statistic at each look, then,

1. Pocock Boundary c satisfies

$$P\left(Z_{t_1} > c, \text{ or } Z_{t_2} > c, \cdots, \text{ or } Z_{t_K} > c | \delta = 0\right) = \alpha.$$

2. O'Brien-Fleming Boundary c satisfies

$$P\left(Z_{t_1}\sqrt{t_1} > c, \text{ or } Z_{t_2}\sqrt{t_2} > c, \cdots, \text{ or } Z_{t_K}\sqrt{t_K} > c | \delta = 0\right) = \alpha.$$

3. Haybittle-Peto Boundaries c_0 and $c_{\alpha-\alpha(K-1)}$ satisfy

$$P\left(Z_{t_1} > c_0, \text{ or } Z_{t_2} > c_0, \cdots, \text{ or } Z_{t_K} > c_{\alpha-\alpha_0(K-1)} | \delta = 0\right) = \alpha,$$

where $c_0 = \Phi^{-1}\left(1 - \alpha_0\right)$, α_0 is predetermined and $\alpha_0\left(K-1\right)$ is the cumulative alpha spending in the first $K-1$ looks.

For unweighted SSR design, since sample size will be updated based on the interim finding, the information time needs to be updated. Thus, the boundaries after the sample size adjustment needs to be recalculated as well.

The effect of different patient increment patterns are also studied. Performance is compared when patient increment is equally-spaced or unequally-spaced with 2 time increments. Information time for each analysis is calculated based on the total number of looks and the patterns of increment.

3.2 Simulation Plan

Simulations for adaptive designs are based on the steps outlined below:

Step 1: Identify an interval of exploration and the maximum and minimum allowed sample size on the basis of early study results and literature.

Step 2: Choose candidate adaptive designs to be considered - GS designs with OBF boundary, Pocock boundary, Haybittle-Peto boundary or SSR designs.

Step 3: Determine the following design parameters:

- Adaptive index
- Maximum sample size for GS designs and initial sample size for SSR designs
- Total number of looks
- Types of information increment
- Time of sample size adjustment for SSR designs
- Adjustment of sample size at the predetermined interim look

Step 4: Obtain average sample size and power at 11 points of treatment effect for 10 evenly divided sections on the selected treatment effect interval for each design via Monte Carlo simulations

Step 5: Get the sample size and power curve on the treatment effect interval through interpolation

Step 6: Evaluate the performance for each adaptive design

All simulation results are based on 10,000 runs for each treatment effect. Simulations are repeated based on different simulation parameters specified in step 3.

3.3 Results

3.3.1 Performance of Adaptive Designs

Performance for each adaptive design on the treatment effect interval $[0.0882, 0.5]$ is evaluated by failure rate R_f and area between log curves (ABLC). For GS design, performance is evaluated at different maximum sample sizes - $2018, 356$, and 63 which are the sample sizes in the FS design when treatment effects are δ_L, δ_M, and δ_U, respectively. For SSR designs, 356 is used as initial sample size (n_{init}) and the maximum allowed sample size after sample size adjustment is 2018. Sample size adjustment is based on a targeted conditional power of 80%.

In Figure 2, the top row shows the performance for GS designs with $n_{max} = 2018, 356$ and 63. Rows 2 to 5 are graphs for the weighted SSR designs with $n_{init} = 2018, 356$ and 63. The last row shows the performance for unweighted SSR designs with $n_{init} = 365$. Both equally-spaced and unequally-spaced (double increment) information time are considered.

For GS designs (top row), the design with PK boundary has the best performance (low failure rate R_f and low ABLC) when n_{max} is not small. When n_{max} is small (= 63), OBF is the best. In terms of the total number of looks, when n_{max} is large (= 2018), the performance of GS designs gets better as the number of looks increases, regardless of the kind of boundaries. When n_{max} is not large, the performance does not alter as much with the increase of the number of looks and is similar among the GS designs with different boundaries. In terms of the maximum allowed sample size, $n_{max} = 356$ gives better performance than $n_{max} = 63$ or 2018. There is an exception though with unequally-spaced information time, where the best performance is observed when n_{max} is 2018.

For the weighted SSR design (rows 2 to 5), the best performance is still with the PK boundary. The focus here is the initial sample size. The best performance is obtained when the initial sample size $n_{init} = 365$ for the following reason: When the initial sample size is small (= 63), the sample size adjustment is done based on very limited information and is not reliable. When the initial sample size is already large (= 2018), no sample size adjustment could occur, thus the SSR designs become GS designs.

Based on the observations for the weighted SSR designs, the unweighted SSR designs (last row) are examined with the initial sample size = 356 only. For unweighted SSR designs, the focus is on the total number of looks, the timing of the sample size adjustment, and the pattern of sample size increment. As shown, the SSR with PK boundary still performs the best overall. Performance improves for all boundaries when the total number of looks increases. However, the failure rate R_f remains similarly low after 3 looks. The timing of the sample size adjustment does not matter for R_f, but for ABLC it is the earlier adjustment, the lower ABLC.

3.3.2 Performance of SSR Designs with $j = 1$

Comparison of SSR designs with $j = 1$ are presented on the first row of Figure 3. In general, ABLC of the unweighted SSR are slightly smaller than the weighted SSR designs, while failure rate for these two designs are almost identical.

3.3.3 Comparison of Maximum of 5 Looks Unequally-Spaced GS Designs Using HP Boundaries Versus SSR Designs Using OBF and Pocock Boundaries

Since sample size can be adjusted based on interim findings, failure rate and ABLC for SSR designs are usually low. However in practice, there is still a lack of understanding on SSR designs compared to the (classical) GS designs. Thus, SSR designs are not well accepted by the regulatory agency (FDA). In the second and third rows of Figure 3 we compare unequally-spaced (double increment) maximum of 5 looks GS designs using HP01 and HP005 boundaries versus SSR designs using OBF and PK boundaries. As shown, the failure rate (second row) and ABLC (third row) for GS designs with HP01 and HP005 boundaries are lower than or similar to that of SSR designs with OBF boundary. However, the opposite is true for SSR designs with PK boundary except for two-stage designs. We discuss more on the special two-stage design in the next sub-section.

3.3.4 Performance of Two-Stage Adaptive Designs

Two-stage adaptive designs have one interim analysis and one final analysis. Because of the simplicity of two-stage designs, they are most often used in clinical trials. Comparison of performance of GS design with maximum sample size of 2018 and weighted and unweighted SSR designs with initial sample size of 356 is presented on the last row of Figure 3. As there is only one interim analysis before the final look, the average sample size for GS design is at least 1009

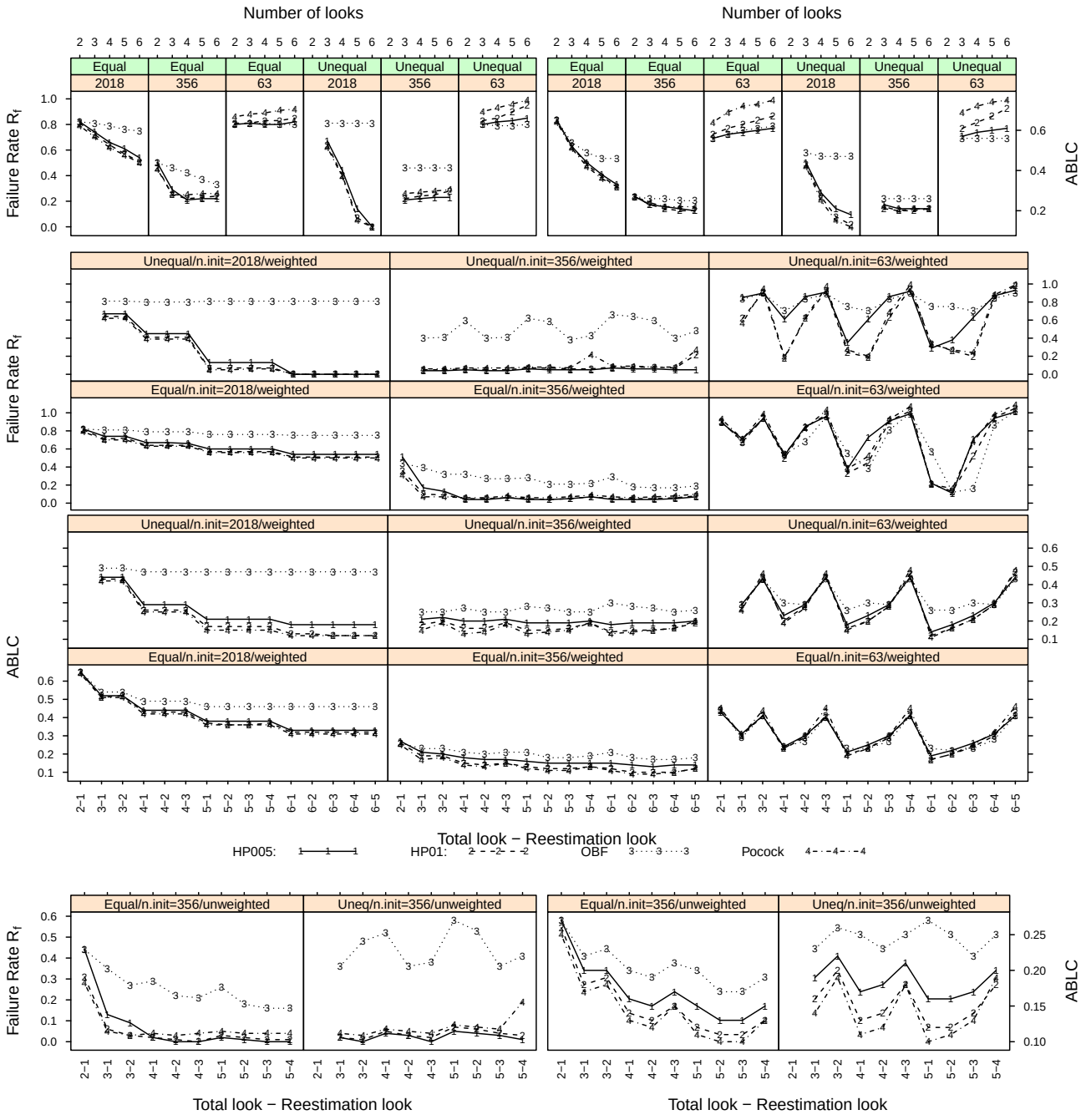

Figure 2. Performance Indicators for GS Design and Sample Size Re-estimation Design (Top row is for GS designs, rows 2 to 5 are for weighted sample size re-estimation designs, last row is for unweighted sample size re-estimation designs with $n_{init} = 356$,)

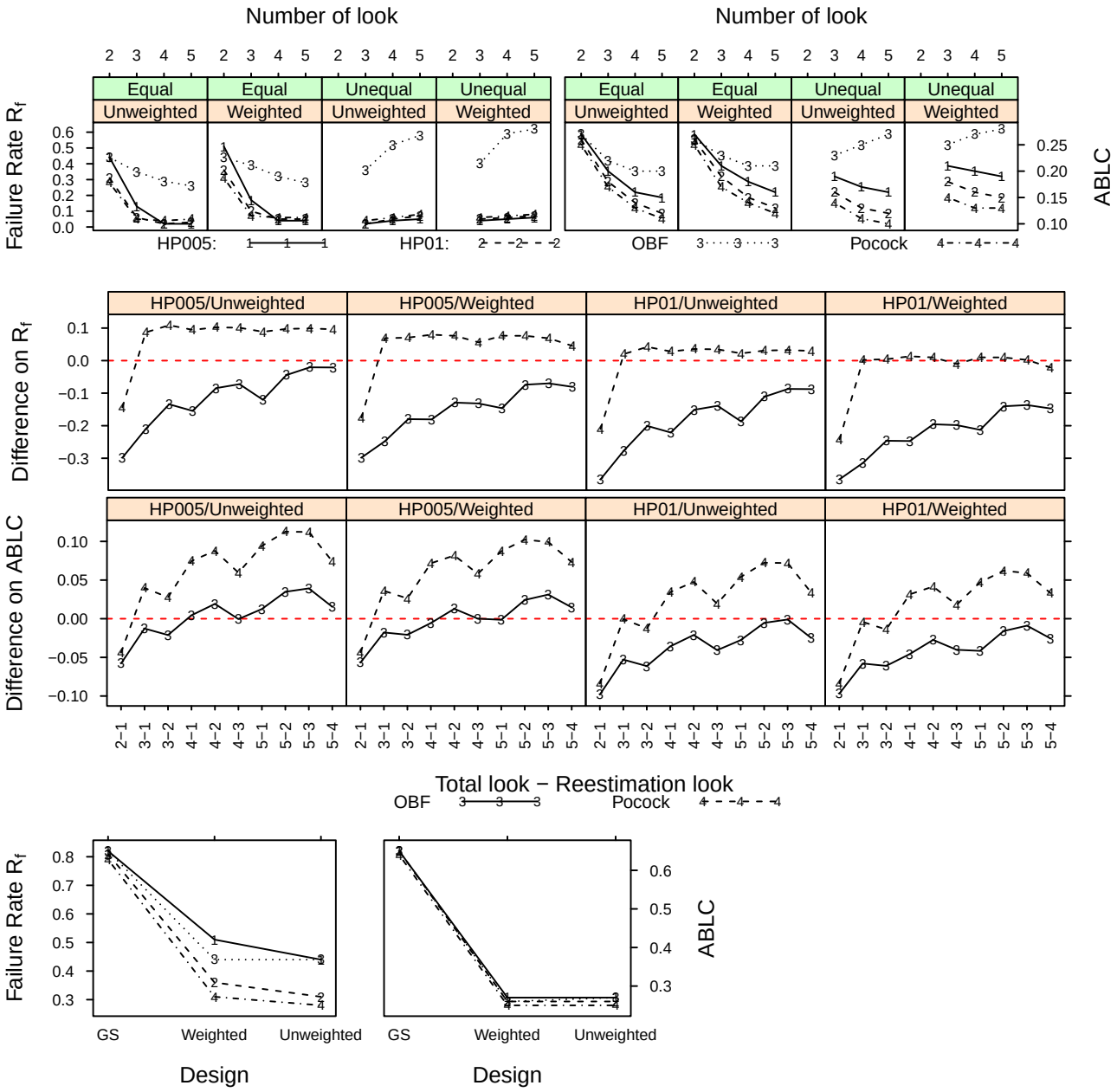

Figure 3. Comparisons of Performance Indicators (Top row is for comparisons of SSR designs with $j = 1$, and 2nd and third rows are for comparisons between maximum of 5 looks unequal increment GS designs and equal increment sample size re-estimation designs, last row is for comparisons of two-stage adaptive designs,)

(= 2018/2) and it is over-sized for a large portion of the treatment effect intervals. Weighted or unweighted SSR designs start with a much smaller initial sample size and sample size will only be increased when the interim finding indicates a small treatment effect. Thus, the performance of two-stage SSR designs is better than the two-stage GS designs. The performance of weighted SSR designs is almost identical to the performance of unweighted SSR designs.

4. Performance of Adaptive Designs when Treatment Effect Follows a Location-Scaled Beta Distribution

In this section, we report results from the same simulation procedures and parameters as in the previous section, except that treatment effects follow the location-scaled beta distributions: $Beta\,(5, 2)$, $Beta\,(2, 5)$, or $Beta\,(4, 5)$ on the treatment effect interval [0.0882, 0.5] (see Figure 1H).

4.1 Results

The top row of Figure 4 shows the performance indicators of GS designs with the maximum sample size n_{max} = 2018. As in the previous uniform distribution case, GS design with PK boundary has the best performance (low failure rate R_f and low ABLC). In terms of the total number of looks, also similar to the uniform distribution case, the performance of GS designs gets better as the number of looks increases, regardless of the kind of boundaries, equal or unequal patient increments. The performance seems robust to the true treatment distribution, especially with the ABLC indicator.

The second and third rows of Figure 4 presents the performance indicators of the weighted and unweighted SSR designs with the initial sample size n_{init} = 365 and the sample size adjustable to maximum = 2018 after sample size adjustment. (See previous discussion on n_{init} = 365 in Section 3.3.1). Only either 2 or 5 looks are considered here. As shown, the best performance is still with the Pocock boundary. The most striking result here is that designs with treatment effect following the location-scaled $beta(5, 2)$ perform much better than designs with treatment effect following a location-scaled $beta(2, 5)$ or $beta(4, 5)$, also the performance indicators with $beta(5, 2)$ are relatively insensitive to the number of looks, the timing of the sample size adjustment, or the pattern of sample size increment.

The last row of Figure 4 presents the comparison of 2-look or 5-look unequally-spaced (double increments) GS designs using HP01 and HP005 boundaries versus SSR designs using OBF and PK equally-spaced boundaries. As shown, when the treatment distribution is $beta(5, 2)$ and looks=5, similar to the uniform distribution case, GS designs with HP01 and HP005 boundaries perform better than or similar to SSR designs with OBF boundary, but worse than SSR designs with PK boundary. When the treatment distribution is $beta(2, 5)$ or $beta(4, 5)$ and looks=5, GS designs with HP01 and HP005 boundaries may perform better or worse than SSR designs with OBF boundary or PK boundary, depending on the timing of the sample size adjustment. In general, early or middle timing is better than late adjustment. For two-stage designs (looks=2), conclusions are the same as the uniform distribution case.

5. Discussion and Conclusion

Among the various adaptive designs, the use of the classical GS designs in clinical trials is well established. However, as commented in a recently published US FDA's Guidance for Industry, the adaptation of sample size based on the interim treatment effect estimates is still regarded as a less understood area (FDA 2010). In this paper, we attempt to contribute to the understanding of SSR designs by comparing performance of GS designs versus SSR designs with different design parameters. One aim is to examine situations where the performance of a SSR design may also be achieved a classical GS design, perhaps with different design parameters.

In this paper, the performance of adaptive designs is based on the measure of sample size and/or power function over the treatment effect interval. The design parameters include the maximum sample size for GS designs, initial sample size for SSR designs, alpha-spending boundaries, total number of looks, types of information increment (equal or double increment), timing of sample size adjustment, etc. Treatment effect is assumed to follow either a uniform distribution or a general location-scaled beta distribution.

There are several interesting findings. First, with the performance indicators defined by failure rate and ABLC in terms of the sample size and/or power measure, PK boundary is the best choice in most cases. GS designs perform better with interim analyses at double increment of information time than at equally-spaced increment. Not surprisingly, the more interim looks the better performance for GS, but not necessarily for SSR designs. Of course, more interim analyses requires more logistic efforts.

For the more common two-stage design, SSR designs perform better than GS designs, regardless of alpha-spending boundary or timing of the interim analysis. The weighted and unweighted SSR designs perform similarly.

Most interestingly, we find that 5-look unequally-spaced GS design with HP01 and HP005 boundaries can achieve similar or better performance than the SSR designs with OBF, but not necessarily better than SSR with PK boundary.

Finally, adaptive designs may not always be the best choice. When the treatment effect interval is narrow, indicating

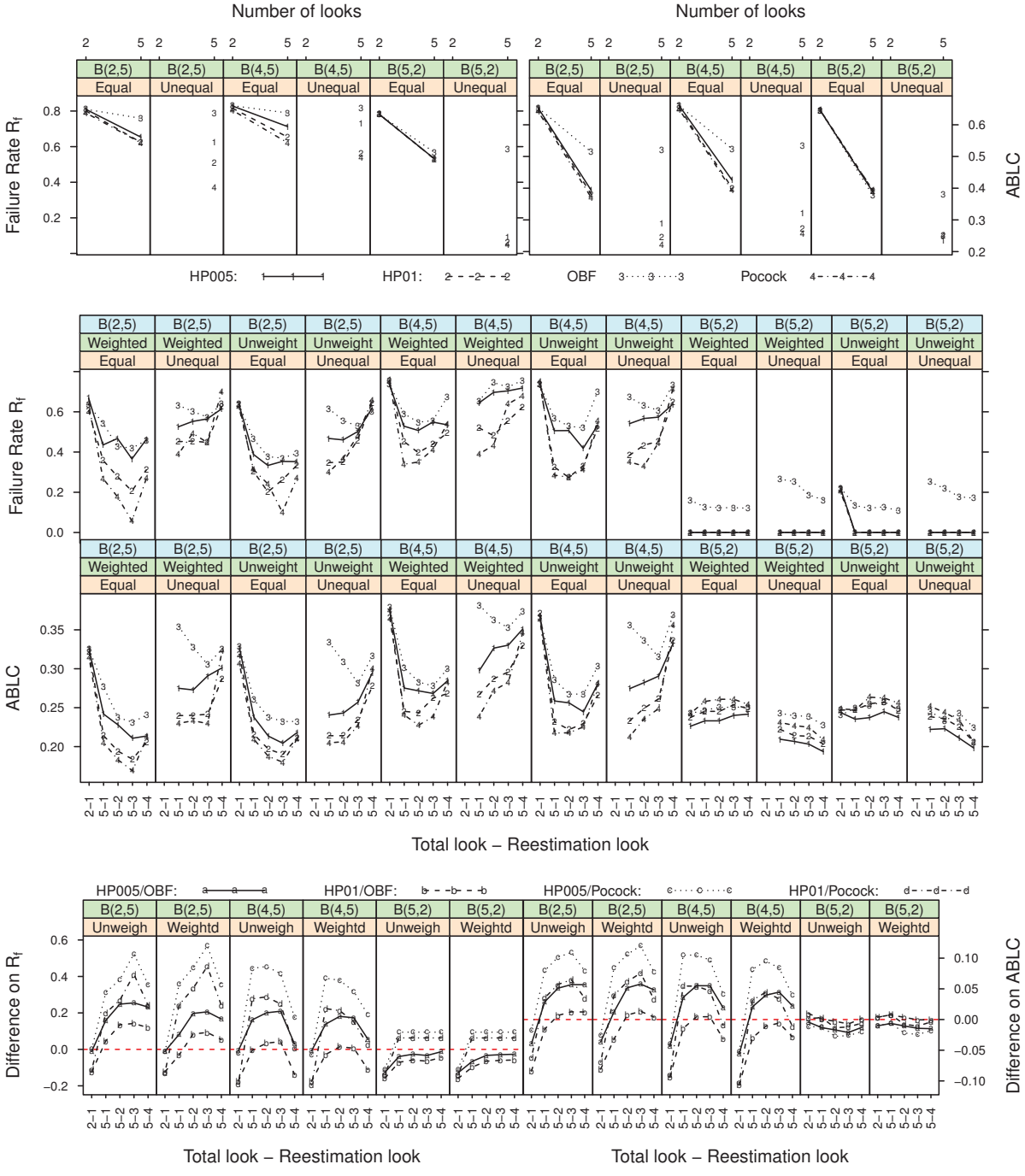

Figure 4. Performance Indicators for designs when treatment effect follows a location-scaled beta distribution (Top row is for GS designs, the second and third rows are for sample size re-estimation designs, and the bottom row is for comparisons of GS design with HP boundary versus SSR with OBF or PK boundaries)

relatively accurate estimation of the treatment effect, performance is robust on the interval regardless of the design chosen. Thus, a fixed sample size design may be good for this circumstance. However, because of the difficulty of obtaining such a narrow treatment effect interval, one should be cautious and may use simulations to confirm the point treatment estimate before the fixed sample size design instead of adaptive design is used.

Acknowledgements

The research of Y. L., S. L. and W. J. S. was partially supported by NIH/NCI CCSG Grant 3P30CA072720.

References

Aravantinos, G., Fountzilas, G., & Kosmidis, P. et al. (2005). Paclitaxel plus carboplatin versus paclitaxel plus alternating carboplatin and cisplatin for initial treatment of advanced ovarian cancer: long-term efficacy results: a Hellenic Cooperative Oncology Group (HeCOG) study. *Annals of Oncology, 16*, 1116-1122.

Armitage, P., McPherson, K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data Journal of the Royal Statistical Society, *132*, 235-244.

Chen, C., Li, N., Shentu, Y., Pang, L., & Beckman, R. A. (2016). Adaptive Informational Design of Confirmatory Phase III Trials With an Uncertain Biomarker Effect to Improve the Probability of Success. *Statistics in Biopharmaceutical Research, 8*(3), 237-247. *Biometrics*, *55*, 853-857.

Chuang-Stein, C. (2004). Seizure the opportunities. *Pharmaceutical Statistics, 3*, 157-159.

Cui, L., Hung, H. M., & Wang, S. J. (1999). Modification of Sample Size in Group Sequential Clinical Trials.

Du Bois, A., Luck, H. J., & Meier, W. et al. (2003). A Randomized Clinical Trial of Cisplatin/Paclitaxel Versus Carboplatin/Paclitaxel as First-Line Treatment of Ovarian Cancer. *Journal of the National Cancer institute*, *95*, 1320-1330.

Ennison, C., & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science*, *5*, 299-317.

FDA. (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics.

Gould, A. L., & Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communication in Statistics, Theory and Methods, 21*, 2833-2853.

Gould, A. L., & Shih, W. J. (1998). Modifying the design of ongoing trials without unblinding. *Statistics in Medicine*, *17*, 89-100.

Hybittle, J. L. (1971). repeated assessment of results in clinical trials of cancer treatment. *Journal of Radiology, 44*, 793-797.

Jennison, C., & Turnbull, B. W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, *25*, 917-932.

Lan, K. K. G., & DeMets, D. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine, 13*, 1341-1352.

Lesko, L. J. (2006). Tools to reduce Phase III trial failures. Proceedings of the AGAH annual meeting.

Levin, G. P., Emerson, S. C., & Emerson, S. S. (2013). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. *Statistics in Medicine, 32*(8), 1259-75.

Li, G., Shih, W. J., Xie, T., & Lu, J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*, *3*, 277–287.

Li, G., Shih, W. J., & Wang, Y. (2005). Two-stage adaptive design for clinical trials with survival data. *Journal of Biopharmaceutical Statistics*, *15*, 707-718.

Liu, G. F., Zhu, R., & Cui, L. (2008). Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Statistics in Medicine, 27*, 584-596.

Neijt, J. P., Engelholm, S. A., Tuxen, M. K. et al. (2000). Exploratory Phase III study of paclitaxel and cisplatin versus paclitaxel and carboplatin in advanced ovarian cancer. *Journal of Oncology*, *18,* 3084-3092.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*, 549-556.

Ozols, R. F., Bundy, B. N., Greer, B. E. et al. (2003). Phase III Trial of Carboplatin and Paclitaxel Compared With Cisplatin and Paclitaxel in Patients With Optimally Resected Stage III Ovarian Cancer: A Gynecologic Oncology Group Study. *Journal of Clinical Oncology*, *21,* 3194-3200.

Parmar, M. K. B., Adams, M., & Balestrino, M. et al. (2002). Paclitaxel plus carboplatin versus standard chemotherapy with either single-agent carboplatin or cyclophosphamide, doxorubicin, and cisplatin in women with ovarian cancer: the ICON3 randomized trial. *The Lancet, 360,* 505-515.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, D. R., Mantel, N., McPherson, K., Peto, J., & Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cance, 34*, 585-612.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika. 64*, 191-199.

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). Statistical monitoring of clinical trials. Springer, New York.

Shih, W. J. (2006). Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: a comparison. *Statistics in Medicine*, *25*, 933-941.

Shih, W. J., Li, G., & Wang, Y. (2016). Methods for flexible sample-size design in clinical trials: Likelihood, weighted, dual test, and promising zone approaches. *Contemporary Clinical Trials, 47*, 40-48.

Thoelke, K. R. (2007). Creating a Framework for Success: A Clinical Strategy for FDA Approval. Proceedings of the Conference: From Pipeline to Product: Navigating the FDA Approval Process.

Tsiatis, A. A., & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, *90*, 367-378.

Wittes, J., & Brittain, E. (1990). The role of pilot studies in increasing the efficiency of clinical trials (with comments). *Statistics in Medicine*, *9*, 65-72.

Xi, D., Gallo, P., & Ohlssen, D. (2017). On the Optimal Timing of Futility Interim Analyses. *Statistics in Biopharmaceutical Research , 9*(3), 293-301.