METHODS FOR THE SEQUENTIAL PARALLEL COMPARISON DESIGN

Rachel Kloss Silverman

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2017

Approved by:

Anastasia Ivanova

Jason Fine

Gary Koch

Richard Zink

John Baron

ABSTRACT

Rachel Kloss Silverman: Methods for the Sequential Parallel Comparison Design
(Under the direction of Anastasia Ivanova and Jason Fine)

Sequential parallel comparison design (SPCD) has been proposed to increase the likelihood of success of clinical trials, especially trials with a possibly high placebo effect. SPCD is conducted with two stages, and subjects are randomized into three groups: (1) placebo in both periods, (2) placebo in the first period and active therapy in the second period, and (3) active therapy in both periods. Efficacy analysis of the study data includes all data from stage 1 and all placebo non-responding subjects from stage 2. Each stage is analyzed separately then the data are pooled to yield a single p-value.

We first describe methods to use in a trial where we combine SPCD with the group sequential approach. We examine how to increase the sample size and adjust the design parameters during an interim analysis to increase power; these design parameters include allocation proportion to placebo in stage 1 of SPCD and weight of stage 1 data in the overall efficacy test statistic.

Next, we develop new methods for SPCD with binary and time-to-event outcomes. These methods allow us to analyze SPCD stage-wise using the model of interest with adjustment for covariates. We show that under certain conditions the covariance between the estimated treatment effects in the two periods of SPCD is 0 under both null and alternative hypotheses. We also show that the stage-wise test statistics are uncorrelated under the null hypothesis. As a result, we can

omit covariance in the construction of the overall test statistic and the confidence interval for the weighted sum of treatment effects.

We develop framework and implementation of SPCD using permutation tests and bootstrap hypothesis testing. This approach allows the flexibility to use SPCD with any outcome. We examine two variations of permutation tests and three variations of the bootstrap. We show that the overall permutation as well as the stage-wise permutation test preserve type I error. Additionally, the bootstrap that maintains the original stage 1 group sample sizes and the stage-wise bootstrap also preserve type I error. The stage-wise permutation test and bootstrap make it easy to evaluate SPCD data with popular software.

To Justin, I could not have done this without you.
And to Don and Michelle, thank you for all your support along the way.

ACKNOWLEDGEMENTS

I would like to thank Justin Silverman for his extremely helpful R coding knowledge and his willingness to run my code on his cluster; Maurizio Fava for providing ADAPT-A data in use in the examples in this dissertation; the associate editors and anonymous reviewers for their helpful comments for Chapters 2-3; Jason Fine for his guidance; and most importantly, Anastasia Ivanova for her wonderful mentorship, guidance, and support throughout.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

1.1 Overview

The sequential parallel comparison design (SPCD) was developed to combat the issue of large placebo response rates that can occur in traditional randomized clinical trials (Fava, Evins, Dorer, & Schoenfeld, 2003). High rates of placebo response in clinical trials, even of drugs previously approved by the FDA for a particular condition, have often failed to demonstrate a significant difference in the active treatment from placebo (Baer & Ivanova, 2013). This failure arises because large placebo response rates that can occur in traditional randomized clinical trials will decrease the appearance of any true effect size and increase the likelihood of concluding a false-negative at the trial end or an outcome that is no longer clinically meaningful. Such a situation makes it necessary to increase the sample size to achieve the power that is necessary to conclude a true result. As these are extremely costly consequences, there are various trial designs devoted to reducing this placebo effect. In addition to the monetary ramifications of an extensive placebo response rate, there are additional public health implications. Negative results and failed trials can delay the introduction of new therapies on the market which, in turn, raises the cost of development or can cause companies to abandon development of treatments that are effective. In fact, some companies have decided to abandon their drug development efforts in certain fields, such as depression because the risk of an effective treatment failing to show efficacy is so high (Baer & Ivanova, 2013).

The popular placebo lead-in ("placebo run-in" or "placebo wash-out" design) is a common design used when researchers expect moderate to high placebo response rates. In this

design (Figure 1), all patients first receive placebo, and then only those non-responding placebo patients are randomized to the placebo versus active therapy arms for the efficacy analysis treatment phase (Fedorov & Lui, 2007). Ultimately, the placebo lead-in design has been shown to be unsuccessful (Trivedi & Rush, 1994). In fact, on average, the use of the placebo lead-in approach often added time and cost to trials without increasing the effect sizes (Baer & Ivanova, 2013).

In 2003, Fava et al. proposed SPCD where subject assignment is as in a three-sequence, two-stage crossover design with PP, PA, and AA sequences where "P" stands for placebo and "A" stands for active therapy. SPCD takes a different approach to data analysis compared to a crossover design in order to avoid making assumptions about equality of treatment effects in the two stages. SPCD analyzes data from each stage separately and then combines the two analyses are combined to yield a single p-value. To avoid dealing with carry-over effects, SPCD does not utilize stage 2 data from the AA group. Additionally, subjects who responded to placebo in stage 1 (as determined a priori and by some clinically relevant cut point) are identified and their data are not included in the primary efficacy analysis. A more in-depth discussion of the trial design, implementation, and testing is provided later in this chapter. This design is useful in trials with high placebo response. SPCD might yield a higher power than a conventional single-stage design because (1) placebo non-responders contribute two data points to the primary efficacy analysis and (2) the exclusion of placebo responders in stage 2 might lead to an increased treatment effect in stage 2. Ivanova and Tamura (2011) extended SPCD by re-randomizing active treatment responders to P and A in stage 2.

Baer and Ivanova (2013) discuss that SPCD can be advantageous in trials with pediatric and adolescent populations, orphan diseases, post-marketing commitment trials, and dose-

response studies. This advantage exists because, in these cases, it is difficult to recruit for these trials, and there is often a desire to generate adequate power while minimizing the investment or exposure of young people to new compounds. Additionally, SPCD can allow a sponsor to compare a new treatment at several alternative doses to placebo in the population of patients in stage 1, and to compare some of the doses in placebo non-responders in stage 2.

SPCD, when compared to the popular placebo lead-in trials, can be considered a more clinically relevant and generalizable trial (Baer & Ivanova, 2013). This is because SPCD trials give weight to the results of everyone enrolled in the trial. Furthermore, in clinical practice, the more resistant patients (placebo non-responders) seek treatment; therefore, a design that focuses on placebo non-responders can, in fact, address the population that new treatments seek to help. Thus, we can view stage 1 of SPCD as the generalizable stage and stage 2 as the more clinically realistic stage.

We can attribute additional merits of SPCD to the fact that an SPCD trial is typically longer than a placebo lead-in or a single-stage trial, with some patients staying on placebo or active drug therapy for the duration of the SPCD trial. This allows the trial sponsor and the FDA to obtain valuable data on response over time and additional safety measurements (Baer & Ivanova, 2013).

Since its introduction, a variety of tests have been proposed for SPCD. Fava et al. (2003) proposed a test based on linear combination of the estimated treatment effects from the two stages when the response is binary. Also for binary outcomes, Huang and Tamura (2010) and Ivanova, Qaqish, and Schoenfeld (2011) developed a score test with one and two degrees of freedom. We are not aware of the methods allowing adjusting for covariates while analyzing SPCD with binary outcomes. Currently, there is no existing methodology to construct confidence

intervals for the weighted combination of the stage-wise treatment effects when the outcome is binary for SPCD. Data analysis strategies for SPCD with continuous outcomes can be found in Tamura and Huang (2007), Chen et al. (2011), and Doros et al. (2013). These researchers base the test statistic on the linear combination of the estimated treatment effects. It is possible to adjust for covariates, but such methods are not applicable to binary data. Baer and Ivanova (2013) reviewed data analysis methods for SPCD and summarized completed trials that used SPCD.

Specifically, the trial is conducted in two stages (Figure 2): randomization of all subjects to active therapy or placebo in stage 1 and then a re-randomization of stage 1 placebo non-responders in stage 2. Placebo responders are usually re-randomized in stage 2 as well and subjects who received A in stage 1 usually continue on A in stage 2. If randomization is conducted by flipping a biased coin, an equivalent format would be to randomize all subjects once at the onset of the trial into three groups: (1) placebo in stage 1 and placebo in stage 2 (PP); (2) placebo in stage 1 and active therapy in stage 2 (PA); and (3) active therapy in stage 1 and active therapy in stage 2 (AA). In this format, the primary analysis includes all stage 1 and stage 2 data from placebo non-responders from groups PP and PA. Irrespective of randomization format, all subjects are followed for the duration of both stages to maintain blinding.

Let the total sample size in the trial be n with n_1 subjects in the placebo group, n_2 subjects in the active therapy group in stage 1, $n_1 + n_2 = n$, with $n_1 = bn$ where b, $0 < b \leq 1$, is the allocation proportion to placebo in stage 1. Note that when $b = 1$, the SPCD design reduces to a placebo lead-in study. To simplify the presentation, we assume that n_1 is even and that subjects $1,\ldots, n_1/2$ are randomized to PP sequence, and subjects $n_1/2+1,\ldots, n_1$ are randomized to PA sequence. The stage 1 allocation proportion to placebo b is usually higher than 0.5 in order to

have enough subjects to evaluate in stage 2. Allocation proportions above 0.75 are not recommended because the design becomes very similar to the placebo lead-in design which has been shown to be ineffective in identifying placebo non-responders (Trivedi and Rush, 1994). Despite guidelines suggesting *b* between 0.50 and 0.75, *b* could be arbitrary. In stage 2, subjects are re-randomized to placebo and active therapy with 50:50 allocation.

For binary responses, denote p_1 = Pr(active therapy response in stage 1), q_1 = Pr(placebo response in stage 1), p_2 = Pr(active therapy response in stage 2 | placebo non-responder in stage 1), and q_2 = Pr(placebo response in stage 2 | placebo non-responder in stage 1). Let the treatment effects in stage 1 and stage 2 be defined as follows: $\Delta_1 = p_1 - q_1$ and $\Delta_2 = p_2 - q_2$. We are interested in testing the null hypothesis, $H_0 : \Delta_1 = \Delta_2 = 0$ with the alternative hypothesis that at least one of the treatment effects is different from zero. One possible approach to test $H_0 : \Delta_1 = \Delta_2 = 0$ is to combine tests of Δ_1 and Δ_2. This approach requires the knowledge of the correlation between the tests under the null hypothesis. Alternatively, one may consider the test statistic based on the weighted average of the estimated treatment effects as described. In practice, we let (X_i, Y_i) be a pair of binary responses of subject *i* in stage 1 and stage 2 of SPCD, respectively. Then the estimated treatment effects are:

$$\Delta_1 = \sum_{i=1}^{n_1} X_i / n_1 - \sum_{i=n_1+1}^{n_1+n_2} X_i / n_2 \text{ and}$$

$$\Delta_2 = \sum_{i=1}^{n_1/2} (1 - X_i) Y_i / (n_1 / 2) - \sum_{i=n_1/2+1}^{n_1} (1 - X_i) Y_i / (n_1 / 2)$$

With the SPCD test statistic,

$$T_{\mathrm{I}} = \frac{w\Delta_1 + (1-w)\Delta_2}{\sqrt{w^2 Var(\Delta_1) + (1-w)^2 Var(\Delta_2)}},$$

where weight w, $0 \le w \le 1$, is chosen a priori and based on the assumed stage-wise treatment effects, $Var(\Delta_1) = \tilde{\Delta}_1(1-\tilde{\Delta}_1)/(n_1+n_2)$ and $Var(\Delta_2) = \tilde{\Delta}_2(1-\tilde{\Delta}_2)/\sum_{i=1}^{n_1}(1-X_i)$ with $\tilde{\Delta}_1 = \sum_{i=1}^{n_1+n_2} X_i/(n_1+n_2)$ and $\tilde{\Delta}_2 = \sum_{i=1}^{n_1}(1-X_i)Y_i/n_1$.

Fava et al. (2003) proposed a similar test statistic, but derived the denominator was derived from the asymptotic distribution of a multinomial vector of counts in stages 1 and 2 of SPCD.

For continuous outcomes, the treatment effect in stage 1 of SPCD, D_1, is measured in the population of "all comers," and the treatment effect in stage 2, D_2, is measured in the population of placebo non-responders. The null hypothesis is $H_0 : D_1 = D_2 = 0$ with an alternative hypothesis that at least one of the treatment effects is larger than zero. One approach to testing the intersection null hypothesis is by using the weighted average of the estimated treatment effects, $w\hat{D}_1 + (1-w)\hat{D}_2$, where the weight w ($0 < w < 1$) is pre-specified (Chen et al., 2011).

We assume that responses from stages 1 (represented by X) and 2 (represented by Y) from patients assigned to placebo in both stages (PP group) of SPCD follow bivariate normal distribution:

$$(X_P, Y_P) \sim \left(N\begin{pmatrix} \mu_{P1} \\ \mu_{P2} \end{pmatrix}, \begin{pmatrix} \sigma_P^2 & \rho_{PP}\sigma_P^2 \\ \rho_{PP}\sigma_P^2 & \sigma_P^2 \end{pmatrix} \right),$$

and that responses from patients assigned to placebo in stage 1 of SPCD and active treatment in stage 2 (PA group) of SPCD also follow bivariate normal distribution:

$$(X_P, Y_A) \sim \left(N\begin{pmatrix} \mu_{P1} \\ \mu_{A2} \end{pmatrix}, \begin{pmatrix} \sigma_P^2 & \rho_{PA}\sigma_P\sigma_A \\ \rho_{PA}\sigma_P\sigma_A & \sigma_A^2 \end{pmatrix} \right).$$

We further assume that responses of patients receiving active treatment in stage 1 of SPCD are $X_A \sim N(\mu_{A1}, \sigma_A^2)$. Treatment effect for stage 1 is $D_1 = \mu_{A1} - \mu_{P1}$. The stage 1 placebo group response probability is $r = \Pr(X_P > c)$, where *c* represents the known response cutoff. The treatment effect in stage 2 is $D_2 = E\{Y_A \mid X_P < c\} - E\{Y_P \mid X_P < c\}$. The test statistic given by (Chen et al., 2011):

$$T_{SPCD} = \frac{wD_1 + (1-w)D_2}{\sqrt{w^2 Var(D_1) + (1-w)^2 Var(D_2)}} \quad (1)$$

has a standard normal distribution $T_{SPCD} \sim N(0,1)$ under the null hypothesis (Chen et al., 2011), and we reject the null hypothesis when $T_{SPCD} > 1.96$. In some therapeutic areas that use SPCD (e.g., psychiatry), it is often more appropriate to examine the decline of symptoms and, as such, negative responses with large absolute value correspond to good treatment response. In this case, one would reject a one-sided null hypothesis when $T_{SPCD} < 1.96$. One can apply the methods described here to psychiatry trials after multiplying responses by (-1).

While the framework of SPCD with continuous and binary outcomes has been explored, we are not aware of published methods for the analysis of SPCD with time-to-event outcomes. Instead of comparing the number of occurrences between active therapy and placebo groups, one can evaluate the time to the first occurrence and determine if that time to the first occurrence is elongated or shortened on active therapy as compared to the placebo group. Alternatively, if a recurrent event is measured, one could estimate the mean cumulative function and answer questions about average number of events by some meaningful amount of time between placebo and active treatment groups (Lawless and Nadeau, 1995; Diao, Cook and Lee, 2015; Hengelbrock et al., 2016). However, time-to-event analyses might be preferred if there are likely to be drop-outs in the study. When the event is favorable, we hope that the active treatment will

shorten the time to first event compared to placebo. Addyi (flibanserin) was FDA approved for the treatment of hypoactive sexual desire disorder (https://clinicaltrials.gov/show/NCT00996164). The conducted trials compared Addyi to placebo by measuring the number of satisfying sexual events in the placebo and the active therapy arm. The baseline counts of satisfying events for the study participants were low, with 25-50% of participants having 0-1 events; as such, time to event could be a better endpoint to evaluate active treatment effect. The active-placebo comparison can be based on the average number of events as well as the time to the first event.

1.2 Next Chapters

We organize the remainder of this dissertation in the following way. For Chapter 2, our purpose is to evaluate various adaptive strategies in an SPCD trial with one interim analysis and continuous outcomes and to provide recommendations regarding their implementation in an SPCD trial. We consider five adaptive strategies in which we modify the following design parameters in period 2 with the goal of increasing the power of treatment comparison: (1) possibly increase sample size; (2) possibly increase sample size, update the weight and allocation in stage 1 of SPCD with the weight, w^*, and allocation, b^*, that maximize power based on period 1 data; (3) update w and b, with w^* and b^*; (4) update w, with w^*; and (5) update b, with b^*. Additionally, in each of the strategies, we determine if we can stop the trial for futility or efficacy at the interim analysis.

In Chapter 3, we discuss SPCD testing with binary outcomes and with time to positive event outcomes. For both setups, we show that the test statistics and treatment effect estimators from stages 1 and 2 are asymptotically normal and uncorrelated, facilitating simple and easy-to-implement inferences. Additionally, we address how to test the SPCD null hypothesis when

adjusting for covariates and how to construct a confidence interval for the weighted combination of the treatment effects.

Finally, in Chapter 4, we develop framework and implementation of SPCD using a permutation test and bootstrapping. This would allow the flexibility to use SPCD with any outcome. We examine two variations of permutation tests and three variations of the bootstrap and discuss type I error preservation. Furthermore, we address how to implement these testing procedures for SPCD with popular statistical software such as SAS.

CHAPTER 2: SAMPLE SIZE RE-ESTIMATION AND OTHER MIDCOURSE ADJUSTMENTS WITH SEQUENTIAL PARALLEL COMPARISON DESIGN[1]

2.1 Introduction

For a completely known distribution of responses in stages 1 and 2 of SPCD, one can compute the optimal pair of the weight, *w*, and allocation to placebo in stage 1, *b*, which maximizes the power of the SPCD analysis. Since we cannot know the treatment effects before the trial (this is especially true for stage 2 treatment effects), one might consider utilizing a blinded interim analysis to estimate the treatment effects from both stages. Then we can compute the optimal pair of the weight and allocation and use that for the remainder of the trial with the intention to increase power. Alternatively, one can ensure adequate power for the trial by increasing the sample size based on the estimated effect sizes and the proportion of placebo non-responders. Mi and Betensky (2013) considered SPCD with binary outcomes with one interim analysis. They performed adaptations that (1) converted SPCD to a single-stage design if the stage 1 placebo response rate was small, (2) converted SPCD to a single-stage design if the stage 1 treatment effect was large, and (3) used sample size re-estimation. Their simulations showed that the type I error rate was inflated to as much as 0.06 when they performed each of the three adaptations was performed; therefore, they proposed an *ad hoc* adjustment of the critical value to preserve the type I error rate. Wang and Ivanova (2014) described a multi-arm SPCD trial with

[1] The contents of this chapter previously appeared as an article in the *Journal of Biopharmaceutical Statistics*. The original citation is as follows: Silverman, R.K. Ivanova, A. (2017). Sample size re-estimation and other midcourse adjustments with sequential parallel comparison design. *Journal of Biopharmaceutical Statistics*.

an interim analysis to possibly drop some of the arms or change the allocation proportion to the arms, depending on the observed responses.

Interim analyses in an SPCD trial are conducted after a proportion (e.g., half) of the planned sample size has completed both stages of SPCD (Figure 3). It is desirable to conduct an interim look in order to check previous placebo response rates and treatment effect assumptions. Thus, if the previous assumptions were not correct, adaptations can be made to increase the power of the study. Let *m* be the number of subjects enrolled in the study before the interim analysis. We denote the study before the interim look as period 1, and the study after the interim look as period 2. We compute period specific test statistics for SPCD, T_1 and T_2, for periods 1 and 2, respectively, using Equation 1 from Chapter 1. The final test statistic is (Lehmacher and Wassmer, 1999):

$$T = \sqrt{v}T_1 + \sqrt{1-v}T_2 . \quad (2)$$

Here, $v = m/n$, where *n* is the originally planned sample size. This test statistic is the same as proposed by Cui, Hung, and Wang (1999). If we perform an interim analysis after half of the subjects complete their follow-up, we set $v = 0.5$. In psychiatry, a typical SPCD trial has two stages, each 4 weeks long; hence, some subjects will be randomized while subjects 1,…, *m* complete their follow-up. T_1 and T_2 are independent, since T_1 is computed based on the data from subjects 1,…, *m*, and T_2 is computed based on the data from subjects $m + 1$, …, *n*. As a result, $T \sim N(0,1)$ and we reject the null hypothesis when $T > 1.96$. (Ivanova, Li, Silverman, Wiener, & Koch, n.d.)

For SPCD with sample size re-estimation, the interim look is used to calculate the conditional power under observed treatment effects from period 1 (Mi & Betensky, 2013). Another possibility is to evaluate conditional power under the treatment effects for which the

trial was originally powered (Liu and Chi, 2001) as the observed treatment effect can be rather variable. In SPCD, little information is usually available about the second stage treatment effect, which is why estimated effects are used. There have been a number of proposals on how to set a new sample size(Gao, Ware, & Mehta, 2008; Jennison & Turnbull, 2015; Liu & Chi, 2001; Mehta & Pocock, 2010; Timmesfeld, Schäfer, & Müller, 2007; Wan, Ellenberg, & Anderson, 2015) . The method proposed by Jennison and Turnbull (2015) maximizes an objective function:

$$CP(t_1, n^*) - \gamma(n^* - n) . \qquad (3)$$

Here, $T_1 = t_1$ is the test statistic observed in period 1, n^* is the new total sample size, *CP* is a conditional power given the trend observed in period 1, and γ is a penalty parameter. The penalty parameter is set by the trial sponsor to achieve a trade-off between the cost of adding extra subjects and a power increase. The test statistic at the end of the trial, given period 1 data, is $\left(\sqrt{v}T_1 + \sqrt{1-v}T_2\right) | t_1 = \sqrt{v}t_1 + \sqrt{1-v}T_2$. Given the current trend, $T_2 \sim N(t_1\sqrt{(n^* - m)/(n-m)}, 1)$ and hence $CP(t_1, n^*) = P(X > 1.96)$, where $X \sim N(\sqrt{v}t_1 + \sqrt{1-v}t_1\sqrt{(n^* - m)/(n-m)}, 1)$. The new total sample size, n^*, is the size that maximizes $CP(t_1, n^*) - \gamma(n^* - n)$. One can also set n_{max}, the pre-specified maximum allowed total sample size, and choose n^* such that $n \le n^* \le n_{max}$. Mehta and Pocock (2010) proposed to increase the sample size when conditional power is in a certain range (i.e., "promising zone") and set the new sample size n^* such that $CP(t_1, n^*) = 0.8$ or other desired power benchmark. Although this approach increases the sample size substantially, it does so over a relatively narrow range of conditional power values. In contrast, the method of Jennison and Turnbull (2015) increases sample size by a smaller amount, but over a larger range of conditional power values. The Jennison and Turnbull (2015) method yields better results in terms of achieving the same power with a smaller expected sample size; however, it requires the

specification of the penalty parameter. Smaller penalty terms yield more substantial sample size increases and hence higher power. Since effect size is not precisely known at the interim analysis, one cannot guarantee a certain power at the end of the trial with any sample size re-estimation method.

2.2 Simulations

In period 1, the stage 1 weight, *w*, and the allocation proportion to placebo, *b*, are set to be equal to the commonly used stage 1 weight of $w = 0.5$ and allocation proportion of $b = 0.67$. To re-evaluate the weight and/or allocation to placebo in SPCD with an interim look, we utilize the data collected in period 1 (from subjects who have completed both stages of SPCD) to estimate the weight and/or allocation that would have produced the largest power in period 2 given the estimated treatment effects, variability of the outcome, and proportion of placebo non-responders. Thus, we found the optimal parameters w^* and b^* by maximizing the function $\Phi(-1.96 - T_{SPCD}(w,b))$ with respect to *w*, *b*, or both *w* and *b*. For computational simplicity, we used a grid search method and searched over [0,1] for optimal weight and [0.5,1) for optimal allocation. We then used these optimal weights and/or allocations in period 2.

We conducted an SPCD trial with an interim analysis with initial weight of $w = 0.5$ and the allocation to placebo in stage 1 of SPCD of $b = 0.67$. The total planned sample size was $n = 300$. We conducted the interim look after half of the subjects, $n/2 = 150$, were enrolled and had completed both stages of SPCD. We also considered trials with an interim look after 100 patients with total planned sample size of $n = 200$. Placebo non-response probability in stage 1 of SPCD was set to $r = 0.75$ in all scenarios except the final one where it was $r = 0.60$. We set the marginal variances in bivariate normal distributions to be equal to 1 in all groups. For the null scenario, we set $\rho_{PP} = \rho_{PD} = 0.5$. For non-null scenarios, we set $\rho_{PP} = 0.8$ and $\rho_{PD} = 0.3$

yielding the variances of the outcomes in stage 2, conditional variances of 0.7 in placebo group, and 0.96 in active treatment group. A typical SPCD trial has two stages, each 3-5 weeks long; hence, we randomized some subjects belonging to period 2 at a time when an updated allocation was not yet available. Upon the availability of the updated allocation, one can set the new period 2 allocation to take into account several subjects randomized with initial allocation. Since the optimal weight is used in the calculation of the period 2 test statistic, there is no issue with any delay in the interim analysis. Therefore, we considered that all period 2 subjects were randomized using the updated allocation. We simulated data with no dropouts. For the strategy in which we re-evaluated both allocation and weight, we selected the allocations to placebo in stage 1 from [0.5,1]. If we re-evaluated the allocation only, we examined allocations to placebo in stage 1 from [0.5,1), excluding 1. When an allocation to placebo was 1 in stage 1 of SPCD, it was equivalent to the placebo lead-in design. In that case, we based the analysis on stage 2 data only. Therefore, because the notion of weight no longer applied, we excluded 1 when we re-evaluated the allocation only. In all of the adaptive strategies, we considered stopping for futility and efficacy. We stopped for futility at the interim when $T_1 < 0$. To stop for efficacy, we used the O'Brien-Fleming boundary, performed an interim look after the first half of the trial, and stopped after period 1 if $T_1 > 2.7959$. The final efficacy was established when $T > 1.977$ (Fleming, Harrington, & O'Brien, 1984). When we stopped the trial for futility or efficacy, there was no need to re-evaluate the allocation and weight; therefore, we excluded those simulated trials from the summary distribution of the estimated optimal weight and allocation reported in Tables 1-3. All tests performed were one-sided. All simulations were run using R version 3.1.1.

2.3 Simulation Results

Table 1 displays the simulation results of SPCD with sample size re-estimation and sample size re-estimation with additional estimation of weight and allocation proportion to placebo (i.e., adaptive strategies (1) and (2) from Section 1.2). Scenario 1 is a null scenario. Figure 4 shows how the rules for the sample size increase at the interim analysis for several penalty parameters as a function of conditional power, CP. We selected the penalty term of 0.0007 because it attains 80% power in scenarios 2-5 in our simulation study. When designing a study, the penalty parameter can be chosen to yield a required power for potential efficacy scenarios. The penalty parameter can also be viewed as guiding a cost/benefit trade-off per each patient added. When the originally planned sample size is 300 and the penalty parameter is 0.0007, the sample size is increased for conditional powers of 25% to 94% (Figure 4), or, alternatively, for T_1 values between 0.93 and 2.49 with the maximum sample size increase to the total sample size of 535 when conditional power is 45%. When the sample size is 300 and the penalty parameter is 0.00085, the sample size is increased for conditional powers of 32% to 92%, or, alternatively, for T_1 values between 1.03 and 2.38 with the maximum sample size increase is to the total size of 452 when conditional power is 52%. When the sample size is 200 and the penalty parameter is 0.001, the sample size is increased for conditional powers of 24% to 94%, or, alternatively, for T_1 values between 0.91 and 2.53 with the maximum sample size increase is to the total sample size of 373 when conditional power is 45%. Given the placebo response probability and variability of the outcomes, the sample size of 300 yields power of 80% when both stage 1 and stage 2 treatment effects are equal to 0.27. Since treatment effects in Table 1 are smaller and given the originally planned total sample size of 300 (Table 1), the trials in scenarios 2-5 are underpowered with a power of 74% (68% for scenario 6). Re-evaluating sample size after

the first 150 patients leads to a better-powered trial. The median total sample size after sample size re-estimation is 430 and 392 for a sample size of 300 and penalty terms 0.0007 and 0.00085, respectively. When the original sample size is 200, the median total sample size after sample size re-estimation with penalty parameter 0.001 is 301, yielding the power of 66% for scenarios 2-5. In comparison, when we used a fixed sample size of 300 is used, the power in scenarios 2-5 is 74%.

Re-estimation of weight and allocation together with sample size shows a further power increase, but not in all scenarios. For example, for scenarios 4 and 5, in which the original sample size is 300, after sample size re-estimation, the power is 81% and 78% for penalty parameters of 0.0007 and 0.00085, respectively. If we additionally re-estimate w and b, the power goes to 90% for scenario 4 (for both penalty parameters), 90% for scenario 5 (when the penalty is 0.0007), and 92% for scenario 5 (when the penalty parameter is 0.00085). In scenarios 2, 3, and 6, re-estimation of weight and allocation together with sample size decreases the gained power from the sample size re-estimation. This finding is a result of the initial weight and allocation being close to optimal. As such, we see that re-estimating weight and allocation leads to selecting a sub-optimal weight and/or allocation.

To shed light on the advantages and disadvantages of re-estimating weight and allocation to placebo, we simulated the adaptive strategies (3), (4), and (5) (outlined in Section 1.2), fixed the sample size at 300, and re-estimated the weight and/or allocation. Table 2 shows the asymptotic power for SPCD where we employ the theoretical optimal weight and/or allocation in period 2. We can view these power values as the maximum theoretical benefit one derives if we adjust the weight and/or allocation after period 1 based on period 1 data. Table 2 shows that the recommended choice of $w = 0.5$ and $b = 0.67$ is a good choice of parameters for most scenarios

unless the effect size in one of the stages is 0, in which case we can increase the power by using an optimal *w* or an optimal pair *w* and *b*. Interestingly, in the six scenarios we considered, there is no power benefit when using the optimal allocation alone while the weight remains $w = 0.5$; in fact, when we fix weight at 0.5, the optimal allocation for scenarios 1-5 is 0.70 and 0.72 for scenario 6. For this reason, we do not recommend this strategy or the strategy where the allocation to placebo is updated alone with the sample size re-estimation. In scenario 4, when we adjust weight, the power can increase from 74% to 84% and further to 92% when we adjust both weight and allocation. In scenario 5, the power increases from 74% to 91% to 93%.

Table 3 shows the results from the corresponding simulations. For the scenarios where the initial weight and allocation were close to optimal (scenarios 2, 3, and 6), re-estimating the weight and/or allocation leads to selecting a sub-optimal weight and/or allocation and therefore to decreased power by 0-5%. When there is no treatment effect in one of the stages (i.e., scenarios 4 and 5), we can increase the power from 73% to 81% to 88% (as in scenario 4) or from 73% to 89% to 90% (as in scenario 5) when the weight is updated and when both the weight and allocation are updated.

2.4 Discussion

We examined several adaptive strategies implemented within an SPCD trial with one interim look. Regarding weight and allocation re-evaluation adaptation strategies, these approaches can be beneficial when the weight and allocation at the start of the SPCD trial are suboptimal. Such a situation occurs when placebo response is very high in stage 1 of SPCD, resulting in almost no treatment effect in stage 1; in this case, we observed that we could increase the power by 10-14% simply by updating the weight and allocation. In contrast, when treatment effects in both SPCD stages are similar, using optimal weight and/or allocation does not appear

to improve power. As such, we advise using the default parameters in period 2 when treatment effects appear similar in both stages in period 1. We observed that changing the allocation alone did not lead to increased power in any of the scenarios; however, changing the weight can be beneficial. If the weight is changed and there is an interest in reporting the estimate of the weighted treatment effect $wD_1 + (1\text{-}w)D_2$ at the end of the trial, the estimate of this quantity can be obtained for any given weight w using the four estimated treatment effects (from the two stages of SPCD for each of the two periods of the trial). For example, it might be beneficial, in order to preserve blinding, to always assign a certain proportion of patients to active arm in the first stage of SPCD, that is, to limit allowable allocation proportions to placebo to [0.5,.8]. As previously mentioned, SPCD generally has higher power than a single-stage design because some patients contribute two data points. We examined this potential increase in power through computing the expected value of the test statistic after the trial and extracting an equivalent effect size from it. For a single stage parallel arm trial, the expected value of the test statistic after a total of n patients complete the study is $(\delta / \sigma)\sqrt{n} / 2$, where δ / σ is an effect size. With stage 1 weight of 0.5, the weighted sum of effect sizes in the four non-null scenarios of SPCD is 0.25. When n = 300, for example, the expected value of the Z score for any of these scenarios is 2.60 = $0.3\sqrt{300} / 2$, corresponding to an effect size equivalent of 0.3. Hence, one can think of SPCD as detecting an effect size of 0.25 as if it were an effect size of 0.30 (20% increase) in a traditional randomized clinical trial. Alternatively, one can think of SPCD as decreasing the sample size by a factor of 0.69 more than a traditional randomized clinical trial since $0.3\sqrt{n} / 2 = 0.25\sqrt{1.2^2 n} / 2 = 0.25\sqrt{1.44n} / 2$.

We have shown that it is possible to use sample size re-estimation strategies with SPCD in a manner similar to their use in traditional trials. The length of follow-up in SPCD is generally

longer due to the two-stage nature of the design; however, this aspect does not interfere with the successful application of these adaptive methods. In this paper, we examined continuous endpoint measures; nevertheless, we can similarly handle the binary outcome can be handled similarly.

CHAPTER 3: SEQUENTIAL PARALLEL COMPARISON DESIGN WITH BINARY AND TIME-TO-EVENT OUTCOMES

3.1 Introduction

A number of recent publications discussed crossover trials with time-to-event outcomes (Buyze & Goetghebeur, 2013; Makubate & Senn, 2010; Nason & Follmann, 2010). Care is needed when integrating the time-to-event endpoints into the SPCD framework. In each of the two stages, we define the time variable is defined as the time of event (for uncensored outcomes) or the end of follow-up for that stage (for censored outcomes). Stage 1 assesses this time-to-event endpoint for all randomized to active therapy or placebo. We consider the case where events are favorable and therefore regard as non-responders those subjects without events in stage 1. We would evaluate only those censored in the stage 1 placebo group in stage 2. As such, we condition the SPCD test statistic from stage 2 on a negative response (in this case, censored time) from stage 1. In fact, this setup is analogous to the case of SPCD with binary outcomes. Inferential results for continuous data are not applicable with time-to-event outcomes, owing to the specialized methods needed in the presence of right censoring.

Recall that with binary outcomes, the SCPD test statistic is as follows:

$$T_{\mathrm{I}} = \frac{w\Delta_1 + (1-w)\Delta_2}{\sqrt{w^2 Var(\Delta_1) + (1-w)^2 Var(\Delta_2)}},$$

where weight w, $0 \le w \le 1$, is chosen a priori, $Var(\Delta_1) = \tau_1(1-\tau_1)(1/n_1 + 1/n_2)$ and

$Var(\Delta_2) = \tau_2(1-\tau_2)\left(1/\sum_{i=1}^{n_1/2}(1-X_i) + 1/\sum_{i=n_1/2+1}^{n_1}(1-X_i)\right)$ with $\tau_1 = \sum_{i=1}^{n_1+n_2} X_i/(n_1+n_2)$ and

$\tau_2 = \sum_{i=1}^{n_1}(1-X_i)Y_i / \sum_{i=1}^{n_1}(1-X_i)$. We will show in Section 3.2 that this and other test statistics defined here are valid test statistics to test the SPCD null hypothesis that preserve the type I error rate under the null hypothesis.

The most common approach to adjust for covariates is to fit a logistic regression model to stage 1 SPCD data and separately to stage 2 SPCD data to test the null hypothesis $H_0 : \theta_1 = \theta_2 = 0$, where θ_1 and θ_2 are log odds ratios in stages 1 and 2. Let $\hat{\theta}_1$ be the estimated log odds ratio from the stage 1 logistic regression model, and T_1 be the test statistic based on $\hat{\theta}_1$. Similarly, $\hat{\theta}_2$ and T_2 are the log odds ratio and the test statistic from stage 2 of SPCD. Consider the following test statistics:

$$T_{\text{II}} = \frac{w\theta_1 + (1-w)\theta_2}{\sqrt{w^2 Var(\theta_1) + (1-w)^2 Var(\theta_2)}},$$

$$T_{\text{III}} = \sqrt{v}T_1 + \sqrt{1-v}T_2 .$$

In the above, $Var(\theta_1)$ and $Var(\theta_2)$ are the corresponding elements of the asymptotic variance-covariance matrix that is computed as the inverse of the Fisher information. The weight v, $0 \le v \le 1$ should be chosen in advance and plays a similar role to the weight w in T_{I} and T_{II}.

To address which test statistic, T_{II} or T_{III}, yields better power, we examine the optimal weights w^* and v^*. Writing T_{II} as:

$$T_{\text{II}} = \frac{w\sqrt{Var(\theta_1)}}{\sqrt{w^2 Var(\theta_1) + (1-w)^2 Var(\theta_2)}} T_1 + \frac{(1-w)\sqrt{Var(\theta_2)}}{\sqrt{w^2 Var(\theta_1) + (1-w)^2 Var(\theta_2)}} T_2$$

illustrates the connection between the two test statistics. Given observed data, one can choose the optimal weights w^* and v^* so that $\max(T_{II}(w)) = \max(T_{III}(v)) = T_{II}(w^*) = T_{III}(v^*)$. This demonstrates that with the right choice of weight, T_{II} and T_{III} have the same power. Since we are

interested in finding the optimal weight to maximize power, we consider the case when $T_1 > 0$ and $T_2 > 0$. It is easy to see that when $T_1 > 0$ and $T_2 > 0$, the optimal weight for T_{III} is $v^* = T_1^2 / (T_1^2 + T_2^2)$. The optimal weight, w^*, for T_{II} can be obtained from the equation:

$$\frac{w^* \sqrt{Var(\theta_1)}}{\sqrt{w^{*2} Var(\theta_1) + (1 - w^*)^2 Var(\theta_2)}} = \sqrt{\frac{T_1^2}{T_1^2 + T_2^2}},$$

$$w^* = \frac{\sqrt{Var(\theta_2)} / T_2}{\sqrt{Var(\theta_1)} / T_1 + \sqrt{Var(\theta_2)} / T_2}.$$

The weights *w* and *v* should be chosen in advance. However, the formulas for w^* and v^* can be useful if an investigator has prior knowledge about treatment effects and their variability in the two stages of SPCD.

In the next section, we will show that the treatment effect estimates from the two stages are uncorrelated and that any of T_{I}, T_{II}, and T_{III} are asymptotically mean zero under the null hypothesis and can be used to test the SPCD hypothesis.

3.2 Inferences in SPCD with Binary Outcomes

Let $\boldsymbol{X}_P = (X_1, ..., X_{n_1})$ be a vector of responses of subjects assigned to placebo in stage 1, and $\boldsymbol{Y}_P = (Y_1, ..., Y_{n_1})$ be a corresponding vector of stage 2 responses. Let $\boldsymbol{X}_A = (X_{n_1+1}, ..., X_{n_1+n_2})$ be a vector of responses of subjects assigned to active therapy in stage 1, and $\boldsymbol{Y}_A = (Y_{n_1+1}, ..., Y_{n_1+n_2})$ be the corresponding vector of stage 2 responses. Define all stage 1 data as $\boldsymbol{X} = (\boldsymbol{X}_P, \boldsymbol{X}_A)$, all stage 2 data as $\boldsymbol{Y} = (\boldsymbol{Y}_P, \boldsymbol{Y}_A)$, and let $\boldsymbol{Z}$ be the matrix of baseline covariates. Let $\delta_i = I(X_i = 0)$ be an indicator that is equal to 1 if subject's stage 2 data are included in the primary efficacy analysis and 0 otherwise. Since only stage 2 data from stage 1 placebo non-responders are included in the primary analysis, $\delta_i = I(X_i = 0) = 1 - X_i$. We set $\delta_i = 0$ for all

participants n_1+1 through n_1+n_2 as their stage 2 data are not included in the primary analysis. We consider the inferential issues associated with combining test statistics and treatment effect estimators from the two stages with binary data.

Consider a function $f(\boldsymbol{X},\boldsymbol{Z})$ of stage 1 data, which may involve stage 1 responses and baseline covariates. This can be an estimated treatment effect or test statistic, potentially adjusted for covariates. Since our analysis only includes stage 2 data from stage 1 placebo non-responders, the stage 2 test statistics and treatment effect estimates are functions not only of stage 2 responses but also of stage 1 responses, denoted by a function $g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ of data from both stages. It is important to recognize that the stage 2 analysis is based on the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and potentially $\boldsymbol{Z}$. Thus, with binary data, the full likelihood function based on both stage 1 and stage 2 data factors into the stage 1 likelihood based on the marginal distribution of the stage 1 responses and the stage 2 likelihood based on the conditional distribution of the stage 1 responses given the data observed at stage 1. Such likelihood factorization does not occur with continuous outcomes when response is defined by dichotomizing the continuous response. The lack of factorization in the case of continuous outcomes arises because the distribution of the stage 2 data conditions on an event that subject's stage 1 outcome belongs to the set of stage 1 placebo non-responders and not the underlying continuous outcome itself. Factorization of the likelihood, and the fact that the expectation in stage 2 of SPCD is conditioned on the stage 1 outcomes, allows us to write:

$$\begin{aligned}\operatorname{cov}(f(\boldsymbol{X},\boldsymbol{Z}),g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})) &= E\left[f(\boldsymbol{X},\boldsymbol{Z})g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})\right] = E\left[E\left[f(\boldsymbol{X},\boldsymbol{Z})g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})\mid \boldsymbol{X},\boldsymbol{Z}\right]\right] \\ &= E\left[f(\boldsymbol{X},\boldsymbol{Z})E\left[g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})\mid \boldsymbol{X},\boldsymbol{Z}\right]\right] = E[f(\boldsymbol{X},\boldsymbol{Z})\cdot 0] = 0.\end{aligned}$$

whenever $E\left[g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})\mid \boldsymbol{X},\boldsymbol{Z}\right]=0$.

Let $f(\boldsymbol{X},\boldsymbol{Z}) = n^{1/2}(\beta_1 - \beta_1)$ and $g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = n^{*1/2}(\beta_2 - \beta_2)$, with $n^* = \sum_{i=1}^{n_1} \delta_i$, and β_j and β_j are the estimated and true treatment effects at stage j (j = 1 and 2), respectively. Note that when considering the distribution of the appropriately standardized β_j, under the alternative, for each fixed value of β_j, the functions $f(\boldsymbol{X},\boldsymbol{Z})$ and $g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ are not functions of unknown β_j and are mean zero when β_j is equal to the true value of the parameter, for j = 1, 2. Under the usual regularity conditions, the standardized maximum likelihood estimators $f(\boldsymbol{X},\boldsymbol{Z}) = n^{1/2}(\beta_1 - \beta_1)$ and $g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = n^{*1/2}(\beta_2 - \beta_2)$ have an asymptotic multivariate normal distribution for all values of β_1 and β_2. When β_1 and β_2 are true parameter values, asymptotically, the stage 1 quantity is unconditionally mean zero and the stage 2 quantity is mean zero conditioned on the results of stage 1. When $g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = n^{*1/2}(\beta_2 - \beta_2) = 0$, asymptotically, using the above result, $\text{cov}(f(\boldsymbol{X},\boldsymbol{Z}), g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})) = 0$ and the vector $\left(n^{1/2}(\beta_1 - \beta_1), n^{*1/2}(\beta_2 - \beta_2)\right)$ converges in distribution to bivariate normal with mean zero and a diagonal variance-covariance matrix. Hence, the covariance of the asymptotic distribution of β_1 and β_2 is 0. Because the treatment effect estimators from the two stages are asymptotically normal and uncorrelated, they are independent. A similar argument can be made when $f(\boldsymbol{X},\boldsymbol{Z})$ and $g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ are test statistics as long as $E\left[g(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \mid \boldsymbol{X},\boldsymbol{Z}\right] = 0$, which is the case under the null hypothesis. These results may be summarized as follows:

(1) In the absence of covariates, the standardized estimated treatment effects from stages 1 and 2 of SPCD with binary outcomes, Δ_1 and Δ_2, are uncorrelated under the null and alternative hypotheses.

(2) In the presence of covariates, the estimated logistic regression coefficients $\hat{\theta}_1$ and $\hat{\theta}_2$ are uncorrelated under the null and alternative hypotheses. As a result of potentially differing sample sizes in each stage, it is important to note here that all parameter estimates are standardized, that is, $n^{-1/2}(\theta_1 - \theta_1)$ and $n^{*-1/2}(\theta_2 - \theta_2)$.

(3) The test statistics T_1 and T_2 computed at stages 1 and 2, respectively, are asymptotically uncorrelated under the null hypothesis.

The standardized test statistics are only mean zero under the null hypothesis and not the alternative hypothesis. Note that if we center the test statistics around their mean, they will be uncorrelated. The standardized treatment effect estimators are mean zero for fixed values of the β_1 and β_2 under both the null and alternative hypotheses. Having shown that the standardized estimators are asymptotically normally distributed and uncorrelated for all values of the parameters, we can then use the delta method to establish the asymptotic normality of the weighted combination of the estimators (after standardization) and derive its variance. This result enables the construction of confidence intervals which are valid under both the null and alternative hypotheses.

Specifically, we have the following:

(1) The distribution of each standardized T_{I}, T_{II}, and T_{III} is asymptotically $N(0,1)$ under the null hypothesis;

(2) The confidence interval $\left(w\Delta_1 + (1-w)\Delta_2 - Z_{1-\alpha/2}SE, w\Delta_1 + (1-w)\Delta_2 + Z_{1-\alpha/2}SE\right)$ for $w\Delta_1 + (1-w)\Delta_2$ has the coverage of $1-\alpha$. Here $Z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of standard normal distribution, and the standard error, *SE*, is computed as $SE = \sqrt{w^2 Var(\Delta_1) + (1-w)^2 Var(\Delta_2)}$, with

$$Var(\Delta_1) = \hat{p}_1(1-\hat{p}_1)/n_1 + \hat{q}_1(1-\hat{q}_1)/n_2,$$

$$Var(\Delta_2) = \hat{p}_2(1-\hat{p}_2)/\sum_{i=1}^{n_1/2}(1-X_i) + \hat{q}_2(1-\hat{q}_2)/\sum_{i=n_1/2}^{n_1}(1-X_i),$$

where $\hat{p}_1 = \sum_{i=1}^{n_1} X_i / n_1$, $\hat{q}_1 = \sum_{i=n_1+1}^{n_1+n_2} X_i / n_2$,

$\hat{p}_2 = \sum_{i=1}^{n_1/2}(1-X_i)Y_i / \sum_{i=1}^{n_1/2}(1-X_i)$ and $\hat{q}_2 = \sum_{i=n_1/2}^{n_1}(1-X_i)Y_i / \sum_{i=n_1/2}^{n_1}(1-X_i)$.

The confidence interval $(w\theta_1 + (1-w)\theta_2 - Z_{1-\alpha/2}SE, w\theta_1 + (1-w)\theta_2 + Z_{1-\alpha/2}SE)$ for $w\theta_1 + (1-w)\theta_2$ has the coverage of $1-\alpha$. Here θ_1 and θ_2 are the log odds ratios and $SE = \sqrt{w^2 Var(\theta_1) + (1-w)^2 Var(\theta_2)}$.

3.3 Time-to-Event Outcomes

Under the SPCD setting, we define time-to-event outcomes in the classical way for both stages 1 and 2, and censoring for each stage occurs at the end of the stage. Let $T_i^{(1)}$ be subject *i*'s first event time in the first stage of SPCD. If the subject does not have an event until the end of stage one, the subject is censored at the end of stage 1. Let $C_i^{(1)}$ be the subject's censoring time in stage 1, $X_i = T_i^{(1)} \wedge C_i^{(1)}$ be the observed stage 1 time for the subject, and $\delta_i^{(1)} = I(T_i^{(1)} < C_i^{(1)})$ be the indicator of whether the event was observed in stage 1. Let $\boldsymbol{Z}_i$ be the subject-specific vector of baseline covariates. Thus, the observed data for subject *i* in stage 1 are $\{X_i, \delta_i^{(1)}, \boldsymbol{Z}_i\}$. Let $T_i^{(2)}$ be subject *i*'s event time in the second stage of SPCD where the time starts from the beginning of stage 2. Similarly to stage 1, if the subject does not have an event until the end of stage 2, the subject is censored at the end of stage 2. Let $C_i^{(2)}$ be the subject's censoring time in stage 2, $Y_i = T_i^{(2)} \wedge \mathrm{C}_i^{(2)}$ be the observed stage 1 time for the subject, and $\delta_i^{(2)} = I(T_i^{(2)} < C_i^{(2)})$ be the indicator of whether the event was observed in stage 2. All subjects usually participate in both

stages to maintain blinding, but primary analysis in SPCD only includes stage 2 data from subjects who did not respond to placebo in stage 1. Thus, we include only data from subjects assigned to placebo in stage 1 with $X_i \geq C_i^{(1)}$ in SPCD primary analysis. Therefore, $(X_i, \delta_i^{(1)}, Y_i, \delta_i^{(2)}, \boldsymbol{Z}_i)$ are the full data of subject *i* in stage 1 and stage 2 of SPCD. The vectors of responses for each treatment group and stage are defined similar to the binary case.

A popular choice in evaluating the time-to-event data is the Cox proportional hazards model (Cox, 1972). The model is $\lambda(t \mid \boldsymbol{Z}_i) = \lambda_0(t)\exp\{\beta' \boldsymbol{Z}_i\}$ where $\lambda_0(t)$ is the baseline hazard and β is the hazard ratio for active treatment versus placebo. To analyze SPCD data, we can consider stage-wise Cox models without or with covariates:

$$\lambda_j(t \mid \boldsymbol{Z}_i) = \lambda_{j0}(t)\exp\{\beta_j I(\text{Treatment}_j = \text{Active})\}, \text{ or}$$

$$\lambda_j(t \mid \boldsymbol{Z}_i) = \lambda_{j0}(t)\exp\{\beta_j I(\text{Treatment}_j = \text{Active}) + \gamma_j \boldsymbol{Z}_i\}.$$

We are interested in testing the null hypothesis that the treatment effects from both stages are zero, $H_0 : \beta_1 = \beta_2 = 0$. Similarly to the binary data setup, with time-to-event outcomes, test statistics and treatment effect estimators computed at stage 2 are evaluated conditionally on the stage 1 results. That is, β_2 refers to the hazard ratio for active treatment versus placebo conditionally on an event not occurring during the stage 1 follow-up. One may argue along the lines in Section 3.2 to show that the likelihood function for the stage 1 and stage 2 survival outcomes factors under our definition of response and thus the estimated regression parameters based on partial likelihood are the maximum likelihood estimators based on separate estimation using the stage 1 and stage 2 data. As a result, the test statistics derived from the partial likelihood estimators are asymptotically bivariate normal and uncorrelated under the null hypothesis, and the estimated treatment effects, after standardization, are asymptotically

bivariate normal and uncorrelated under both null and alternative hypotheses. The test statistics and parameter estimates may be combined, as with binary outcomes, with the theoretical properties of the weighted combination derived from the delta method. Analogous results in Section 3.2 will also apply to count data (e.g., Poisson type outcome) as long as the definition of placebo non-responder is a subject with no events.

To test the SPCD null hypothesis, we can use the weighted average of the estimated treatment effects, $w\beta_1 + (1-w)\beta_2$, where $0 \le w \le 1$ and is chosen a priori, with the following test statistic:

$$T_{\mathrm{IV}} = \frac{w\beta_1 + (1-w)\beta_2}{\sqrt{w^2 Var(\beta_1) + (1-w)^2 Var(\beta_2)}}.$$

Alternatively, we can use the test statistic $T_{\mathrm{V}} = \sqrt{v}Z_1 + \sqrt{1-v}Z_2$, where T_1 and T_2 are stage-wise test statistics. These test statistics can be for testing the coefficients β_1 and β_2 in the Cox model, or, for example, can be stage-wise log-rank test statistics.

3.4 ADAPT-A Trial Example

ADAPT-A trial was a clinical trial to assess the efficacy of low-dose aripiprazole added to antidepressant therapy in patients with major depressive disorder and inadequate response to prior antidepressant therapy (Fava et al., 2012). SPCD was used with two stages of 30 days each. In the first stage, 167 patients were randomized to placebo and 54 to aripiprazole (3:1 randomization ratio). All patients participated in stage 2 of the trial regardless of their stage 1 responses. The primary analysis included all stage 1 data and stage 2 data from patients who did not respond to placebo in stage 2. The primary endpoint was a dichotomized Montgomery-Asberg Depression Rating Scale (MADRS) score with success defined as a 50% or greater reduction in MADRS scores compared to baseline. In stage 1 of SPCD, 10 out of 54 (18.5%)

patients responded to aripiprazole, and 29 out of 167 (17.4%) responded to placebo. Including stage 2 data from stage 1 placebo non-responders into the primary analysis yields the following counts in stage 2: 14 out of 65 (21.5%) patients responded to aripiprazole, and 5 out of 65 (7.7%) responded to placebo. The primary analysis reported by Fava et al. (2012) used binomial repeated-measures regression, accounting for correlation between subject data in stages 1 and 2. The model was analyzed using SAS Proc Genmod (with identity link, binomial repeated measures) and included study stage, treatments and their interaction, and control for categorical study center variables as randomization was stratified by center. In contrast, our method allows for performing stage-wise analyses and then combining either the estimated treatment effects or the test statistics without a need to estimate correlation between estimates in the two stages. Since there were a total of 21 centers, for the sake of this illustration, we combined the centers creating two larger centers. We performed stage-wise logistic regression analysis unadjusted and adjusted for center (Tables 4 and 5).

Adjusting for covariates correlated with the measured outcome generally reduces the standard error of the estimate. However, this was not the case in our ADAPT-A re-analysis example. However, the treatment effect estimates increased in the adjusted analysis compared to the unadjusted analysis. This, in turn, led to smaller p-values after adjusting for center (Table 5). In our example, the p-value corresponding to the test statistic based on the sum of weighted treatment effects, T_{II}, is smaller than the p-value corresponding to the weighted sum of the test statistics, T_{III}, with equal weighting of the two stages. In fact, the p-value from T_{II} in the adjusted analysis is significant at 0.05 level. In our example, the p-value corresponding to the test statistic based on the sum of weighted treatment effects, T_{II}, is smaller than the p-value corresponding to the weighted sum of the test statistics, T_{III}, with equal weighting of the two stages. To illustrate

the discussion about optimal weights in Section 2.1, we computed optimal weights given observed data and corresponding test statistics. As discussed earlier, if the optimal weight computed based on observed data is used, the two test statistics are exactly the same. Table 5 also shows the results for the score test (Ivanova et al., 2011) with default parameter $r = 1$ and with the optimal parameter computed from observed data. The optimal parameter is equal to the ratio of the treatment effects $r^* = (14/65 - 5/65)/(10/54 - 29/167) = 12.0$.

We see in Table 4 that the effect size, defined as the treatment effect divided by the pooled standard deviation, is close to zero in stage 1, equaling to 0.03 and 0.04 for the unadjusted and adjusted analyses, respectively. Clearly, using a single-stage design would give the impression that there is no benefit to low-dose aripiprazole in reducing the MADRS score. In contrast, the effect size is substantial in stage 2 with 0.38 for unadjusted and 0.41 for adjusted analyses, respectively. Since the treatment effect is much higher in stage 2 of SPCD, one might argue that the placebo lead-in design would have led to an even smaller p-value if used in ADAPT-A trial. This is because it would have had all patients assigned to placebo in stage 1, resulting in more subjects contributing to the primary analysis in stage 2. However, the placebo lead-in design might not have been as effective as SPCD in identifying placebo non-responders and increasing the treatment effect in stage 2 (Trivedi & Rush, 1994) resulting in a smaller treatment effect in stage 2. Also, if stage 1 treatment effect is slightly higher than observed in ADAPT-A trial, the SPCD is more powerful than the placebo lead-in design due to combining data on treatment comparison from both stages.

3.5 Simulation Study, Binary Outcomes

In each simulation, allocation to placebo was $b = 2/3$, and the total sample size was 300 for all scenarios. The additional simulations under the null hypothesis with sample sizes of 40

and 80 show that type I error is preserved even when the sample size is low. Data were generated using the inverse logit function with pre-specified treatment effect and covariate parameters. The correlations from the observations from the same subject were set to $\rho_{PP} = 0.8$ and $\rho_{PA} = 0.3$. The covariate is assumed to be normally distributed and increases the odds of an event by 22% with a unit increase in the covariate (corresponds to a logistic regression parameter of 0.2). Simulation results are based on 150,000 runs. All tests were two-sided with a significance level of 0.05 and were conducted in R version 3.3.1.

Table 6 displays the results of SPCD with binary outcomes with and without adjustment for covariates. In both cases, type I error is maintained at 0.05. Coverage of the 95% confidence interval for the overall treatment effect, defined as the weighted average of stage 1 and stage 2 treatment effects, is maintained for all scenarios. To verify the uncorrelatedness of stage 1 and 2 analyses, we estimated the correlation between the estimated treatment effects in the two stages of SPCD. With M simulation runs, the standard error for the estimated correlation is $se = \sqrt{(1-r^2)/(M-2)}$ and therefore does not exceed 0.0026 with M = 150,000. If correlation is outside the interval (-1.96se,1.96se) = (-0.005, 0.005), we conclude that the correlation is not 0. For the estimated treatment effects, the correlation is always within (-0.005, 0.005) both under the null and alternative hypotheses. The same is true for the correlation between stage 1 and stage 2 test statistics under the null hypothesis. This confirms our result in Section 3.2. The correlation between the test statistics under the alternative hypothesis is always outside the interval (-0.005, 0.005), indicating that the true correlation is not 0. This is as expected. If we subtract the true mean of each stage 1 and stage 2 test statistics under a given alternative, they would be uncorrelated.

3.6 Simulations, Time-to-Event Outcomes

In each simulation, both stages 1 and 2 were 28 days long, and stage 1 allocation to placebo was $b = 2/3$. The sample size was 300 for all scenarios with additional null simulations having sample sizes of 40 and 80 to show that type I error is preserved even when the sample size is low. We used exponential distribution with a scale parameter of 0.01 to simulate time to event on placebo in stage 1, yielding a stage 1 placebo response rate of 25%. We chose the exponential distribution to help uphold the proportional hazards assumption, which is not an unreasonable assumption when the stages are short in duration as in SPCD. A placebo non-responder is a subject who did not have an event in the first 28 days. Since having an event is a favorable outcome, active therapy is expected to reduce time to event compared to placebo. We sampled time to event in stage 1 in both the placebo and active therapy group using the inverse probability method proposed by Bender, Augustin, and Blettner (2005). This method ensures that the time to event follows the proportional hazards model assumptions. We censored all stage 1 times at 28 days. For those placebo non-responders re-randomized to placebo in stage 2, their stage 2 times to event were calculated as 28 days less than their original time to event. Thus, in the PP group, $\boldsymbol{Y}_P = \boldsymbol{X}_P - 28 \cdot \overrightarrow{1_P}$. This makes intuitive sense, as nothing has changed for these subjects. Times to event for placebo non-responders that are re-randomized to active therapy in stage 2 were calculated as, $\boldsymbol{Y}_P = \exp\{-\beta_2 I(\text{Treatment}_2 = \text{Active})\}(\boldsymbol{X}_P - 28 \cdot \overrightarrow{1_P})$. These subjects, had they remained on placebo, would have had an event at time $T_i - 28$. However, because they switched to active therapy in the second stage, their time to event was adjusted by the stage 2 active therapy hazard. This ensures that the proportional hazards assumptions are upheld in stage 2. Again, all stage 2 times were censored at 28 days. We reasonably assumed stage 2 treatment effects to be larger than in stage 1 because of the exclusion of the placebo responders. Tables 7

and 8 report the results of these simulations with and without a covariate. Simulation results are based on 150,000 trial runs. All tests were two-sided with a significance level of 0.05.

As in the binary outcome simulations, the type I error rate is preserved. The 95% confidence interval of the overall treatment effect provides correct coverage with 95% coverage for all scenarios both without and with covariates (Tables 7 and 8). The estimated treatment effects from stages 1 and 2 are uncorrelated under both the null and alternative hypotheses as expected. The results also confirm the uncorrelatedness of the stage-wise log-rank statistics under the null hypothesis but not under the alternative hypothesis, where the estimated mean correlation is outside the interval (-0.005,0.005), indicating that the true correlation is not 0.

We also compared SPCD with a single stage trial. Consider a scenario where we enroll 300 subjects into a time-to-event SPCD trial with 2:1 allocation and 4 weeks follow-up in each stage, 8 weeks total. Instead of dividing 8 weeks into two stages, we can have a trial with 8 weeks follow-up and 300 subjects equally assigned to active treatment and placebo. In this case, the sample size is the same and follow-up time is the same, but the person-months is smaller for SPCD compared to a standard trial because we do not include stage 2 data from stage 1 placebo responders. If the treatment effect β is 0.25 in a standard 8-week duration trial, and in each stage of SPCD, $\beta_1 = \beta_2 = 0.25$, the power for SPCD with weight $w = 0.5$ is 53%, and the power for a single-stage design is 67%. If the treatment effect in placebo non-responders is higher than 0.37, the power for SPCD will be higher than the power for a trial with a standard design having an 8-week follow-up, and potentially much higher as the stage 2 treatment effect increases. That is why SPCD is usually recommended for trials with high placebo response where the treatment effect in stage 1 is rather low and much higher treatment effects are expected in placebo non-responders in stage 2.

3.7 Discussion

SPCD is an efficient design for comparing a novel intervention with placebo. Similar to a crossover design, it utilizes two data points from subjects, though not from all subjects as in a crossover, instead of one data point as with a standard trial design. Unlike crossover, one does not need to worry about the carry-over effect to analyze SPCD data as there is no required assumption about the relevant magnitude of treatment effects in the two stages. Researchers propose SPCD for trials with high placebo response because they believe it eliminates placebo responders from stage 2 and leads to higher treatment effect in stage 2. This, combined with the ability to collect two data points from subjects, in general, leads to a higher power of SPCD compared to a standard trial. One disadvantage of SPCD is that not all data are used in the primary analysis. For example, data of subjects who received active treatment in stage 2 are not utilized. These data, however, allow for important secondary analysis. For example, one can compare placebo and active treatment for the duration of the two stages of SPCD by comparing the responses from PP and AA groups at the end of stage 2.

For normal outcomes, many authors have proposed combining treatment effects with weights for the primary SPCD analysis. Chen et al. (2011) showed that the covariance between the estimated treatment effects is zero under the null hypothesis and, therefore, can be omitted in the denominator of the test statistic based on the weighted combination of the estimated treatment effects. Since covariance might not be zero under the alternative, one needs to estimate the covariance between the estimated treatment effects to construct a proper confidence interval when the outcome is normal. We proved that in SPCD with binary and time-to-event outcomes, covariance is zero between the estimated treatment effects. Therefore, it can be omitted in construction of the test statistic and the confidence interval for the weighted sum of treatment

effects. We also showed that stage-wise tests statistics are uncorrelated under the null hypothesis; therefore, SPCD hypothesis can also be tested based on the weighted combination of the test statistics. Our result applies to binary, count, and time-to-event outcomes where placebo non-responder is defined as a subject with no event in stage 1. It is not clear if this result holds if placebo non-responder is defined differently or for continuous outcomes.

CHAPTER 4: PERMUTATION-BASED INFERENCE FOR SEQUENTIAL PARALLEL COMPARISON DESIGN

4.1 Permutation Test

Let $U = (u_1, ..., u_n)$ and $V = (v_1, ..., v_m)$ be two independent random samples, potentially drawn from two different distributions, *F* and *G*. The two-sample permutation test tests the null hypothesis, $H_0 : F = G$. Let N represent the combined sample of U and V of size $n+m$. Additionally, let $Z = (U, V)$ and Z' be the ordered vector of size N of all the data. Let $\pi = (\pi_1, ..., \pi_N)$ be a vector that indicates the sample (U or V) that each ordered observation belongs to. We have that $\hat{\theta} = f(\pi, Z') = \bar{u} - \bar{v}$ where $\bar{u}$ and $\bar{v}$ are the computed means of the observed samples of U and V. All $\binom{N}{n} = \frac{N!}{m!(n-m)!}$ permutations of π are equally likely under the null hypothesis that $F = G$. We denote π^* as any one of these permutations of π and can compute all $\binom{N}{n}$ permutation replications of $\hat{\theta}$ as $\theta(\pi^*) = f(\pi^*, Z')$. This provides the permutation distribution of $\hat{\theta}$, and the probability that θ^* exceeds $\hat{\theta}$ represents the permutation significance level. Measuring the observed test statistic in relation to the null distribution provides the exact p-value for the test. Thus, we can reject the null hypothesis at the 0.05 significance level when $\Pr(\theta^* \geq \hat{\theta}) < 0.05$. In theory, one can compute the test statistic for every permutation of the data, but in practice this can be computationally intensive as the sample size increases. To reduce the computational burden, one can draw a random subset of permutations of

the data. Dwass (1957) proposed using "modified randomization tests" which randomly samples a subset of reference datasets from the set of all data permutations to provide a valid test. The validity of the test is maintained even when the subsets are small in comparison to the possible number of permutations.

4.2 Permutation test for SPCD

Let n_1, n_2 and n_3 be the stage 1 sample sizes for the PP, PA, and AA groups, respectively. For the PP group, let $\boldsymbol{X}_{PP} = (X_1, ..., X_{n_1})$ be a vector of stage 1 responses, and $\boldsymbol{Y}_{PP} = (Y_1, ..., Y_{n_1})$ a corresponding vector of stage 2 responses. For the PA group, let $\boldsymbol{X}_{PA} = (X_{n_1+1}, ..., X_{n_1+n_2})$ be a vector of stage 1 placebo responses, and $\boldsymbol{Y}_{PA} = (Y_{n_1+1}, ..., Y_{n_1+n_2})$ a corresponding vector of stage 2 active treatment responses. For the AA group, let $\boldsymbol{X}_{AA} = (X_{n_1+n_2+1}, ..., X_{n_1+n_2+n_3})$ be a vector of stage 1 active treatment responses, and $\boldsymbol{Y}_{AA} = (Y_{n_1+n_2+1}, ..., Y_{n_1+n_2+n_3})$ be the corresponding vector of stage 2 active treatment responses. Also, let $\boldsymbol{X}_P = (\boldsymbol{X}_{PP}, \boldsymbol{X}_{PA})$ and $\boldsymbol{Z} = (\boldsymbol{X}_P, \boldsymbol{X}_{AA}, \boldsymbol{Y}_{PA}, \boldsymbol{Y}_{AA})$ be the collection of all the study data. We assume that $\boldsymbol{X}_P$ and $\boldsymbol{Y}_{PP}$ are sampled from some probability distribution *F* and that $\boldsymbol{X}_{AA}, \boldsymbol{Y}_{AA}$ and $\boldsymbol{Y}_{PA}$ are sampled from some probability distribution *G*. We are interested in testing the null hypothesis that $H_0 : F = G$.

Here, $\boldsymbol{Z}'$ is a vector of size *n* of the ordered data, and π is the corresponding group assignment (P or A) vector of size *n* for $\boldsymbol{Z}'$. We choose B independent vectors, $\pi^*(1), ..., \pi^*(B)$, each consisting of n_1 PP's, n_2 PA's, and n_3 AA's, from the set of all $\binom{N}{n_1 + n_2 + n_3}$ potential vectors.

For SPCD, $\theta(\pi^*) = f(\pi^*, Z')$ is the test statistic (see Equation 1). We compute this for each of the permutation replicates. We reject the null hypothesis at the 0.05 significance level when $\Pr(\theta^* \geq \hat{\theta}) < 0.05$. Since the permutation test essentially shuffles the treatment labels for the study participants, data from all participants from each stage need to be collected.

We also examine, for SPCD, a stage-wise permutation test, which means that we perform two permutation tests and combine the resulting test statistics for the calculation of the overall test statistic, $T = \sqrt{0.5}T_1 + \sqrt{0.5}T_2$, where T_1 and T_2 represent the test statistics from stage 1 and stage 2, respectively. The stage 1 permutation test permutes the labels of $\boldsymbol{X}_{PP}, \boldsymbol{X}_{PA}$, and $\boldsymbol{X}_{AA}$, and the stage 2 permutation test permutes the labels of the stage 2 responses of stage 1 placebo non-responders (subset of $\boldsymbol{Y}_{PP}$ and $\boldsymbol{Y}_{PA}$).

4.3 The Bootstrap

The bootstrap hypothesis test for $H_0 : F = G$ is similar to the permutation test except that it samples with replacement and only uses information from stage 1 and stage 2 placebo non-responders. When using a bootstrap to hypothesis test a difference in means, there is no need to assume that the two samples have equal variances, only equal means. In this case, one computes a common mean, centers both samples to this common mean, and resamples each population separately (Efron & Tibshirani, 1993). Slight modification to the algorithm that Efron and Tibshirani provide is needed to extend this testing procedure to SPCD.

4.4 The Bootstrap for SPCD

In order to perform the bootstrap for hypothesis testing, we center the original data for each stage and then compute the SPCD primary efficacy analysis test statistic. For $x_{PP} \in X_{PP}$, we compute $x_{PP}^{\ *} = x_{PP} - \overline{x_P} + \overline{x}$, where $\overline{x_P}$ is the mean of all the stage 1 responses for the PP and

PA groups, and $\bar{x}$ is the mean over all the stage 1 responses. In a similar manner, for $x_{PA} \in X_{PA}, x_{AA} \in X_{AA}$, we compute $x_{PA}^{*} = x_{PA} - \overline{x_P} + \bar{x}$ and $x_{AA}^{*} = x_{AA} - \overline{x_A} + \bar{x}$, where $\overline{x_A}$ is the mean of all the stage 1 responses in the AA group. Stage 2 data is transformed in a similar manner. For $y_{PP} \in Y_{PP}$, we compute $y_{PP}^{*} = y_{PP} - \overline{y_P} + \bar{y}$, where $\overline{y_P}$ is the mean of all the stage 2 responses for PP, and $\bar{y}$ is the mean over all the stage 2 responses from placebo non-responders. Similarly, for $y_{PA} \in Y_{PA}$, we compute $y_{PA}^{*} = y_{PA} - \overline{y_A} + \bar{y}$, where $\overline{y_A}$ is the mean of all the stage 2 responses in the PA group. Using these centered data, we use the test statistic in Equation 1, $t(z^{*b})$, for all $b = 1,...,B$ bootstrapped datasets. We then are able to approximate the achieved significance level, ASL_{boot}, by $ASL_{boot} = \#\{t(z^{*b}) \geq t_{obs}\} / B$, where t_{obs} is the observed test statistic.

We examine three different methods for obtaining the B bootstrapped datasets. For Method 1, we sample n_1 centered observations with replacement from the PP group; n_2 centered observations with replacement from the PA group; and n_3 centered observations with replacement from the AA group. The SPCD test statistic is obtained using the formula in Equation 1 such that all stage 1 and only stage 2 placebo non-responder information is used. Method 2 is similar to Method 1 except that we not only preserve the stage 1 group sizes but also the stage 2 group sizes. Let $n_{1,NR}$ and $n_{2,NR}$ represent the total number of placebo non-responders in the PP and the PA groups, respectively, as determined by the original data. For Method 2, we sample $n_{1,NR}$ centered observations with replacement from the $n_{1,NR}$ placebo non-responding PP group; $n_1 - n_{1,NR}$ centered observations with replacement from the $n_1 - n_{1,NR}$ placebo responding PP group; $n_{2,NR}$ centered observations with replacement from the $n_{2,NR}$ placebo non-responding

PA group; $n_2 - n_{2,NR}$ centered observations with replacement from the $n_2 - n_{2,NR}$ placebo responding PA group; and n_3 centered observations with replacement from the AA group. Again, the SPCD test statistic is obtained using the formula in Equation 1 for each of the bootstrapped samples. Method 3 bootstraps stages 1 and 2 separately, obtaining two stage-wise SPCD test statistics and combining them to yield the overall SPCD test statistic using a weighted combination of the test statistics, $\sqrt{0.5}T_1 + \sqrt{0.5}T_2$, where T_1 and T_2 represent the test statistics from stage 1 and stage 2, respectively. Additionally, one could combine the p-values from stage 1 and stage 2 using Fisher's method, where $-2\sum_{i=1}^{k}\ln(p_i) \sim \chi^2_{2k}$, and solve for the overall SPCD p-value. For the stage 1 bootstrapped samples, we sample n_1 centered stage 1 observations with replacement from the PP group; n_2 centered stage 1 observations with replacement from the PA group; and n_3 centered stage 1 observations with replacement from the AA group. For the stage 2 bootstrapped samples, we sample $n_{1,NR}$ centered stage 2 observations with replacement from the placebo non-responding PP group, and $n_{2,NR}$ centered stage 2 observations with replacement from the placebo non-responding PA group.

The merits of the bootstrap for SPCD lie in the fact that we need only stage 1 data and stage 2 data for placebo. Thus, we do not need stage 2 data from the PP and PA placebo responders and the AA group to perform the bootstrap for SPCD. In contrast, for the permutation test, all data from stages 1 and 2 for each of the three groups, PP, PA, and AA, are essential. One potential result of this is that if one has only stage 1 and stage 2 information from the placebo non-responders, one can still perform the SPCD bootstrap.

4.5 Permutation and Bootstrap SPCD Simulation Study

We conducted a simulation study of the SPCD permutation test and the SPCD bootstrap by examining the type I error and the power of these tests under different sample sizes and outcome distributions. Type I error was evaluated with total sample sizes of 45, 60, and 90 and for normal, gamma, and Poisson distributed outcomes. Power for the SPCD permutation test and SPCD bootstrap was evaluated with total sample sizes such that the SPCD test statistic described in Equation 1 would have 80% power, and for normal, gamma, exponential, and Poisson distributed outcomes. For both type I error and power simulations, unadjusted and adjusted (for baseline and a normally distributed covariate) test statistics were used. Tables 9 and 10 show the results of these simulations. For every scenario, we ran 20,000 simulations, each with 2,500 bootstrapped or permutated datasets.

Simulations of the normally distributed data were simulated in the same manner as in Ivanova et al. (submitted 2017) where the authors randomly generated outcome Y_{ij} of patients from $Y_{ij} = \xi_{ij} + \psi_{ij} + e_{ij}$, with group i = PP, PA, AA, stage $j = 0,1,2$ for baseline, stage 1, and stage 2, respectively. Here, $\xi_i = (\xi_0, \xi_{i1}, \xi_{i2})'$ represents the means for the stages with ξ_0 as the common baseline mean; ξ_{i1} as the stage 1 mean for group i; and ξ_{i2} as the stage 2 mean for placebo non-responders for group i. Additionally, $\psi_i = (0, 0, (1 - Z_{i1})(\xi_{i3} - \xi_{i2}))'$, where Z_{i1} is the indicator for a placebo non-responder for group i, that is, $Z_{i1} = 1$ when $Y_{i1} > c$, where c is some predetermined response cut point, and ξ_{i3} is the stage 2 mean for placebo responders.

$e_i = (e_{i0}, e_{i1}, e_{i2})'$ are normally distributed, $N(\vec{0}_3, \Sigma)$, $\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_{01} & \rho_{02} \\ \rho_{01} & 1 & \rho_{12} \\ \rho_{02} & \rho_{12} & 1 \end{bmatrix}$, with variance σ^2 and correlation between stages j and j', $\rho_{jj'}$.

For the evaluation of type I error with a normally distributed outcome, we set $\xi_0 = 40$, $\xi_{PP,1} = \xi_{PA,1} = \xi_{PP,2} = \xi_{PA,2} = \xi_{AA,2} = 35$ and $\xi_{PP,3} = \xi_{PA,3} = \xi_{AA,3} = 32$, $\sigma^2 = 36$ and assumed an exchangeable correlation structure, $\rho_{01} = \rho_{02} = \rho_{12} = 0.5$. For the evaluation of power with a normally distributed outcome, we set $\xi_0 = 40$, $\xi_{PP,1} = \xi_{PA,1} = \xi_{PP,2} = 35$, $\xi_{AA,1} = \xi_{AA,2} = \xi_{PA,2} = 33$, and $\xi_{PP,3} = 32.5$, $\xi_{PA,3} = 31$, $\xi_{AA,3} = 33$, which equates to a stage 1 and stage 2 treatment effect of -2. Additionally, we set $\sigma^2 = 36$ and assumed an autoregressive correlation structure, $\rho_{01} = \rho_{02}^{0.5} = \rho_{12} = 0.7$. For both type I error and power evaluations with normal outcomes, we chose a placebo response cut point of c = 33, which yields a placebo non-response rate of 63%.

For the gamma distributed outcomes, we used a normal copula to impose an exchangeable correlation structure, $\rho_{01} = \rho_{02} = \rho_{12} = 0.5$, for type I error, and we used an autoregressive correlation structure, $\rho_{01} = \rho_{02}^{0.5} = \rho_{12} = 0.7$, for power simulations. For type I error simulations, all outcomes (baseline, stage 1, and stage 2) regardless of group, were drawn from a gamma with shape parameter 6 and rate parameter 0.17. For power simulations, baseline outcomes, and stages 1 and 2, placebo outcomes were drawn from a gamma with shape parameter 6 and rate parameter 0.17. Stage 1 active outcomes and stage 2 active outcomes were drawn from a gamma with shape parameter 6 and rate parameter 0.13. For both type I error and power evaluations with gamma outcomes, we chose a placebo response cut point of c = 29.5, which yields a placebo non-response rate of 63%.

For the exponentially distributed outcomes, we used a normal copula to impose the same exchangeable correlation structure for type I error and autoregressive correlation structure for the power as with the gamma distributed outcomes. For power simulations, the baseline outcomes, stage 1 placebo outcomes, and stage 2 placebo outcomes were drawn from an exponential with rate parameter 0.25. Stage 1 active outcomes and stage 2 active outcomes were drawn from an exponential with rate parameter 0.53. We chose a placebo response cut point of $c = 1.85$, which yields a placebo non-response rate of 63%.

For the Poisson distributed outcomes, we used a normal copula to impose the same exchangeable correlation structure for type I error and autoregressive correlation structure for the power as with the gamma and exponentially distributed outcomes. For power simulations, baseline outcomes, stage 1 placebo outcomes, and stage 2 placebo outcomes were drawn from a Poisson distribution with rate parameter 2.13. Stage 1 active outcomes and stage 2 active outcomes were drawn from a Poisson distribution with rate parameter 1.35. We used a placebo response cut point of $c = 1$.

When adjusting for a normally distributed covariate and baseline measurements when the outcomes are normal, gamma, and Poisson, we modified means and rate parameters to achieve 80% power.

4.6 Simulation Results

Table 9 reveals the results of the type I error simulation study. It shows that both the permutation test and combined stage-wise permutation test preserve the type I error at the 0.05 level for total sample sizes of 45, 60, and 90 with normal, gamma, and Poisson distributed outcomes. Furthermore, we demonstrate that the type I error is preserved at the 0.05 level when adjusting for baseline and a covariate. For the bootstrap Methods 1 and 3, the type I error is

preserved at the 0.05 level for total sample sizes of 45, 60, and 90 with normal, gamma, and Poisson distributed outcomes with and without adjusting for baseline and a covariate. Bootstrap Method 2, which maintains stage 1 and stage 2 group sizes, inflates the type I error (as much as 0.067) for sample sizes 45, 60, and 90 when not adjusting and adjusting for baseline and a covariate.

Table 10 shows the power from the simulation study and directly compares the power achieved with the typical SPCD test statistic from Equation 1, the SPCD randomization-based test statistic (Ivanova, Li, Silverman, Wiener, & Koch, n.d.), the permutation test, the stage-wise permutation test, and the three different bootstrap tests discussed. Power is also shown for the adjusted (for baseline and a covariate) test in Equation 1, permutation tests, and the bootstrap tests.

For a normal outcome and a total sample size of 210, the typical SPCD test statistic from Equation 1, the permutation tests, and the bootstrap tests perform equally well. The two permutation tests yield similar power with 80-81% for unadjusted and 81-82% for adjusted. The three methods of bootstrap unadjusted have a power of 80-83% and, after adjusting for baseline and a covariate, have a power of 79-81%.

For the gamma outcome and sample size of 60, the typical SPCD test statistic from Equation 1 yields 81% power unadjusted and 82% power adjusted for baseline and a covariate. Again, the permutation tests yield similar power for both unadjusted and adjusted. Bootstrap Methods 1 and 2 achieve similar power with 77% and 80% power for unadjusted and 79% and 79% when adjusting for baseline and a covariate. Bootstrapping Method 3 achieves lower power with 71% power unadjusted and 68% when adjusting.

For an exponential outcome and total sample size of 45, the SPCD test statistic from Equation 1 yields 79% power unadjusted and 78% power adjusted for baseline and a covariate. The permutation tests yield similar power for unadjusted and slightly less power when adjusting for baseline and a covariate (74-75% power). All three methods of the bootstrap perform worse with unadjusted powers of 69%, 73%, and 44%, and adjusted powers of 69%, 71%, and 42% for Methods 1, 2, and 3 respectively.

When the outcome is Poisson and the sample size is 60, the SPCD test statistic from Equation 1 yields 81% power for both unadjusted and adjusted. Again, the permutation tests yield similar power for both unadjusted (80% and 79%) and adjusted (80% and 78%). Bootstrap Methods 1 and 2 have similar power of 81% power for unadjusted and 79% and 80% when adjusting for baseline and a covariate. However, bootstrapping Method 3 has 76% power unadjusted and 71% power adjusted.

4.7 ADAPT-A Example

We re-evaluate the ADAPT-A data (described in Section 3.4) using the actual MADRS scores and treating these scores as continuous outcomes with the discussed permutation tests and the bootstraps. The permutation tests use all data from both stages, whereas the bootstraps use all stage 1 information and stage 2 information from placebo non-responders. For each of the five tests (two permutation tests and three bootstrap tests), we evaluate 150,000 permuted or bootstrapped datasets.

In stage 1 of SPCD, 10 out of 54 (18.5%) patients responded to aripiprazole, and 29 out of 167 (17.4%) responded to placebo. As addressed in the original paper, Fava et al. (Fava et al., 2012) claim that because of the high placebo response rate, ending the trial after stage 1 would

result in a failed trial. However, including the stage 2 information in the primary efficacy analysis increases the effect size and decreases the p-value.

Using the test statistic in Equation 1 and treating the MADRS score as a continuous outcome, the overall SPCD p-value is 0.074. When we use the permutation test, the overall p-value is 0.076. Using the stage-wise permutation test, the p-value is 0.082. Conducting the Method 1 bootstrap where we maintain the group sample sizes in stage 1, we get a p-value of 0.067, and when we maintain the group sample sizes in stages 1 and 2 (Method 2), we get a p-value of 0.059. The stage-wise bootstrap (Method 3) produces a p-value of 0.076. We see that the permutation and the bootstrap tests yield results similar to the SPCD test statistic in Equation 1.

When we adjust for the baseline MADRS score, and use the test statistic in Equation 1, the overall SPCD p-value is 0.15. For the permutation test and the stage-wise permutation test, the p-value becomes 0.15 and 0.17 respectively. Bootstrap Methods 1 – 3 produce p-values of 0.15, 0.14, and 0.17, respectively.

4.8 Stage-wise Testing in SAS

In Section 4.6, we demonstrated that the stage-wise permutation test, while not as powerful when the outcomes are not normally distributed, preserves the type I error. As a result, we can employ the use of popular statistical software such as SAS, to perform the stage-wise permutation testing procedure. We employ the help of the %NParCov4 macro developed for advanced randomization based-methods (Zink, Koch, Chung, & Wiener, 2017). The %NParCov4 macro takes the following inputs: outcomes (OUTCOMES), covariates (COVARS), exposures (EXPOSURES), treatment groups (TRTGRPS), hypothesis (HYPOTH), outcome transformations (TRANSFORM), strata (STRATA), how the analysis will accommodate

covariates within strata (COMBINE), strata weights (C), confidence limit for intervals (ALPHA), exact analysis (EXACT), random seed (SEED), number of random data sets to generate (NREPS), space requirements for SAS/IML (SYMSIZE), print option (DETAILS), input dataset (DSNIN), and prefix for output dataset (DSNOUT). For the SPCD stage-wise permutation test, we first read in the ADAPT-A dataset (ADAPTA). The variables of interest in the ADAPT-A dataset are madrs_visit3 (stage 1 outcome), madrs_visit6 (stage 2 outcome), treatment_stage1 (0/1 for placebo/drug assignment in stage 1), and treatment_stage2 (0/1 for placebo/drug assignment in stage 2). We run the %NParCov4 macro, %NPACOV4(OUTCOMES = madrs_visit3, COVARS =, TRTGRPS = treatment_stage1, HYPOTH = NULL, ALPHA = 0.05, EXACT = YES, SEED = 44, SYMSIZE = 200000, NREPS = 150000, DSNIN = ADAPTA, DSNOUT = stage1). With this statement, we perform the permutation test on the stage 1 information, at the 0.05 significance, with 50,000 replicates. We can find all the permuted treatment differences in the dataset _STAGE1_BETASAMP, which is produced by the macro. The first row of this dataset shows us the treatment difference for our original dataset. We can run a PROC MEANS to obtain the standard error (0.0061) of these permuted treatment differences for stage 1. Next, we fit %NPACOV4(OUTCOMES = madrs_visit6, COVARS =, TRTGRPS = treatment_stage2, HYPOTH = NULL, ALPHA = 0.05, EXACT = YES, SEED = 44, SYMSIZE = 200000, NREPS = 50000, DSNIN = ADAPTA_STAGE2, DSNOUT = stage2) to the stage 2 ADAPT-A dataset with only placebo non-responders (ADAPTA_STAGE2). This results in a dataset _STAGE2_BETASAMP for the treatment differences for stage 2. Again, we run a PROC MEANS to obtain the standard error (0.0067) of the stage 2 permuted treatment differences. We combine the datasets _STAGE1_BETASAMP and _STAGE2_BETASAMP and create the derived variable,

$\sqrt{0.5} * Stage\,1\, treatment\ difference + \sqrt{0.5} * Stage\ 2\ treatment\ difference$, which is the stage-wise SPCD test statistic. This will produce 50,001 SPCD test statistics and the first row of this dataset will have the SPCD test statistic for the original ADAPT-A data. The p-value of this stage-wise permutation test is the rank of the original ADAPT-A SPCD test statistic divided by 50,001. The p-value of the stage-wise permutation test is 0.079. When adjusting for baseline MADRS score, the p-value of the stage-wise permutation test is 0.17. We see that this is similar to the results produced in Section 4.7.

We chose not to employ the %NParCov4 to perform a bootstrap using the HYPOTH = ALT option because the macro does not center the stage 1 and stage 2 data as described in this paper.

4.9 Discussion

We showed that with four differently distributed outcomes, both permutation tests, the bootstrap, and the stage-wise bootstrap hypothesis tests (Methods 1 and 3) preserve type I error under the null hypothesis and are valid for SPCD data. The permutation test and the stage-wise permutation test achieve similar power to the conventional SPCD test statistic but the permutation test requires that all data from stage 1 and stage 2, including placebo responders. The bootstrap maintaining stage 1 group sizes achieves similar power or slightly less power to the conventional SPCD test statistic and the stage-wise bootstrap is more conservative and less powerful. However, the benefit of using the bootstrap for SPCD is that one only needs SPCD primary efficacy analysis data, meaning all stage 1 data and stage 2 data from placebo non-responders.

We find that the SPCD Bootstrap Method 2 which fixes both the stage 1 and stage 2 group sizes does not preserve type I error. Although the reason behind this is not apparent, we

hypothesis that by fixing stage 2 group sizes and not allowing the number of placebo non-responders to be random, we are not drawing from the true null distribution. When we fix stage 1 group size and allow the number of placebo non-responders to be random as in SPCD Bootstrap Method 1, type I error is strictly preserved at 0.05.

The permutation test and the bootstrap, both being non-parametric tests, require fewer distributional assumptions to be valid than the conventional SPCD test statistic and therefore can be safely applied in SPCD trials were the outcome is known to be non-normally distributed. However, in general both the permutation test and the bootstrap require larger sample sizes than the conventional SPCD test statistic because they do not impose parametric assumptions.

We have found that the stage-wise permutation and the stage-wise bootstrap tests are valid for SPCD. As a result, software that can easily compute permutation and bootstrap tests for a standard parallel arm single stage two-treatment trial can be used for SPCD, as shown in Section 6. One needs only to compute the permutation test or bootstrap on stage 1 and stage 2 data separately, extract the test statistics (or p-values) and then combine these test statistics for an overall SPCD test statistic. The stage-wise bootstrap, however, is much less powerful than the bootstrap that maintains stage 1 group sizes when the outcomes are not normally distributed. It is therefore, recommended, when only SPCD data is available (all stage 1 information and only stage 2 placebo non-responders data), to perform the bootstrap that maintains stage 1 group sizes. If all data is available (all responses from both stages 1 and 2) then it is recommended that the permutation test be used, as it is more powerful than the bootstrap. If only SPCD data is available (stage 1 data and stage 2 data from placebo non-responders) then it is recommended that the bootstrap hypothesis test be used as it both preserves type I error and is not as conservative as the stage-wise bootstrap hypothesis test.

APPENDIX 1: FIGURES AND TABLES

Figure 1. Placebo lead-in study design.

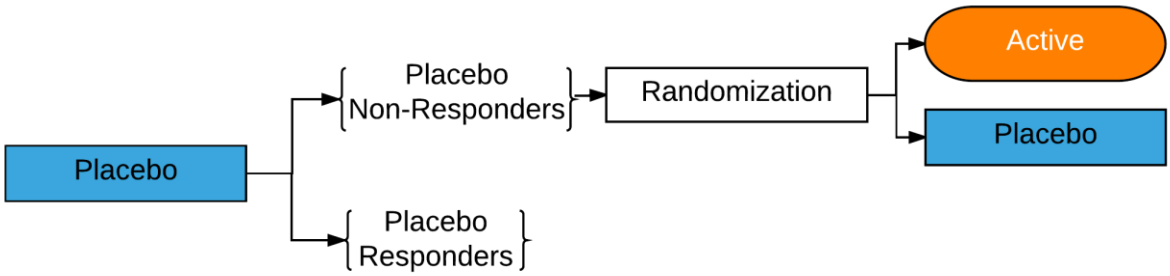

Figure 2. Sequential parallel comparison design. Outcomes highlighted within the grey box are used in the efficacy analysis.

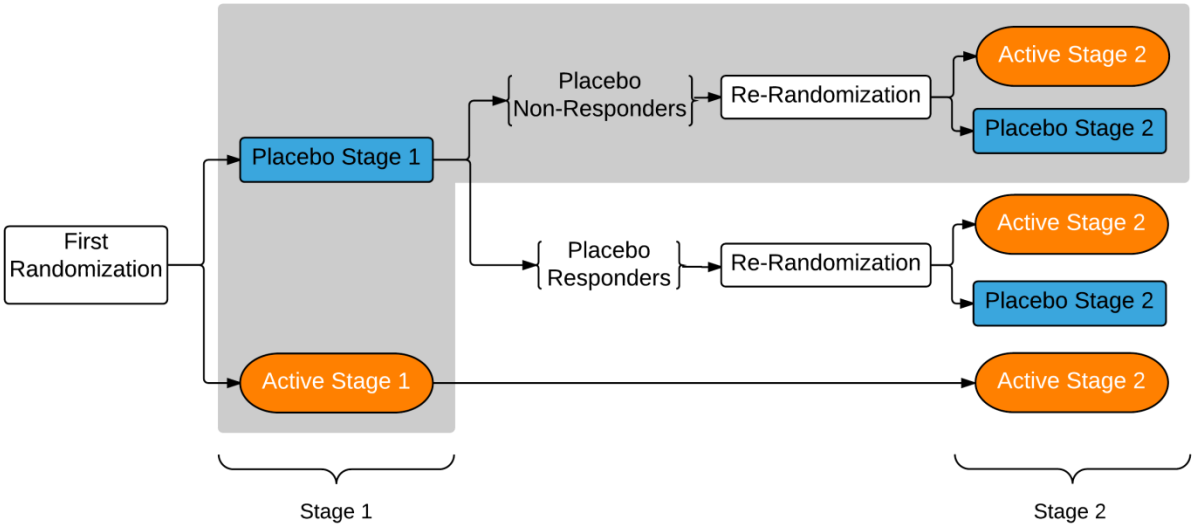

Figure 3: Sequential parallel comparison design with interim analysis.

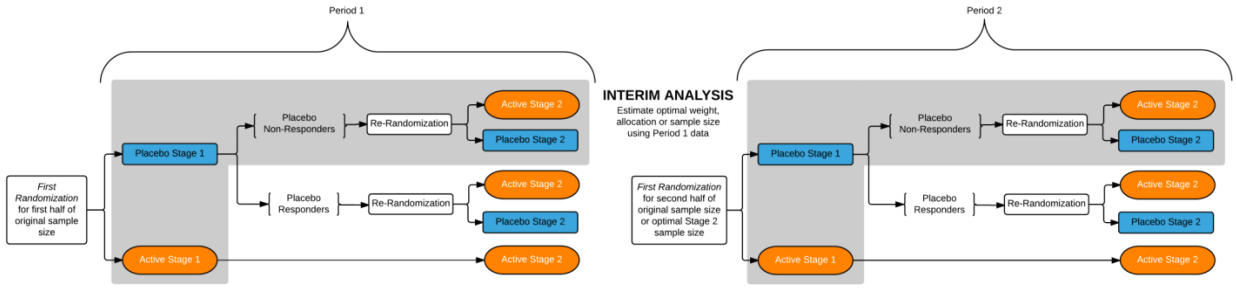

Figure 4. Rules for adding the sample size. Final sample size is plotted by conditional power for different penalty terms and originally planned sample sizes.

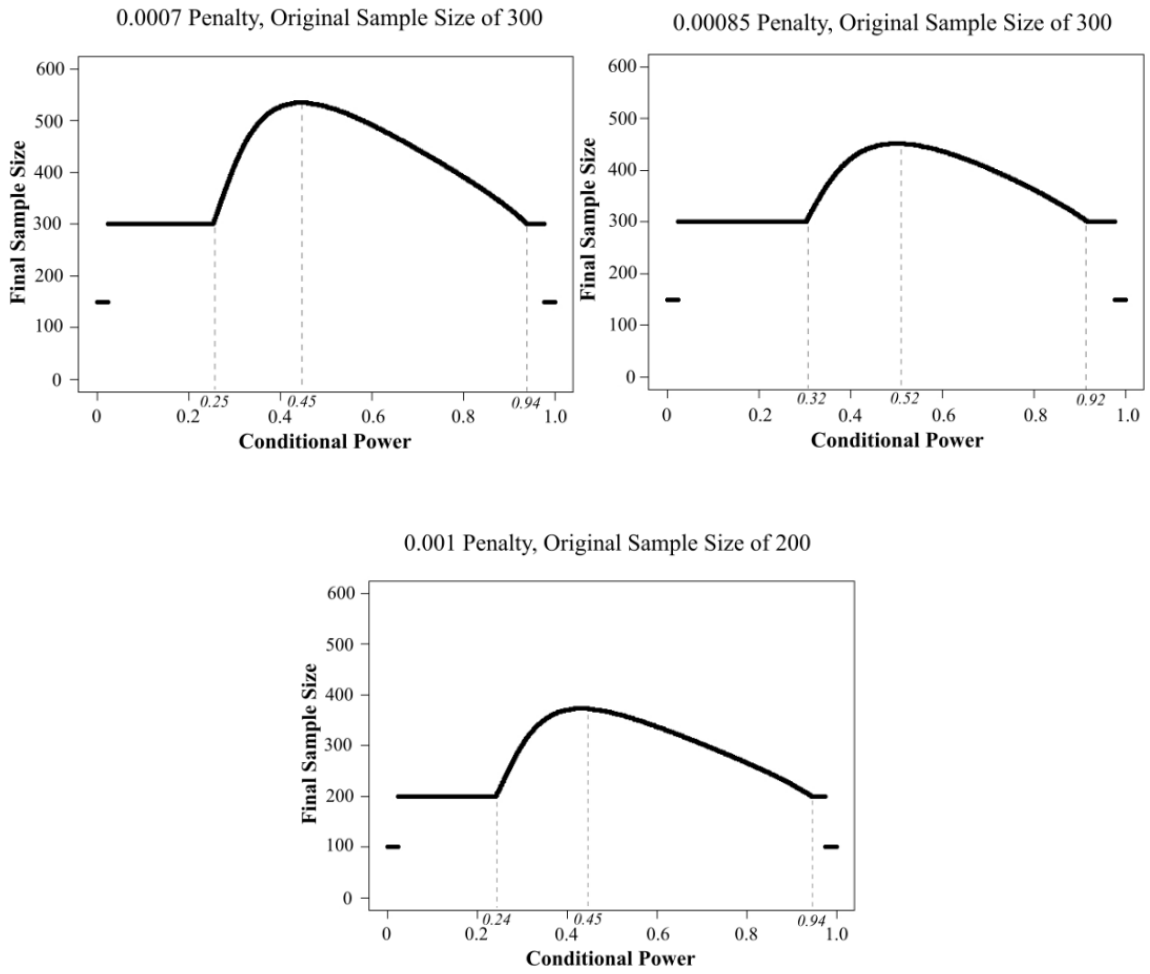

Table 1. Simulated power for SPCD without sample size re-estimation, with sample size re-estimation alone, and with weight and allocation re-adjustment for six scenarios.

	Asymptotic Power, Power under sample size re-estimation, Power under sample size re-estimation with weight and allocation re-estimation		
Effect Size in Stage1, Effect Size in Stage 2 and *r*	$n = 300$ $\gamma = 0.0007$	$n = 300$ $\gamma = 0.00085$	$n = 200$ $\gamma = 0.001$
	$\hat{n}^*_{med}$ =476 (406,519) $\hat{p}_{futile}$ = 0.50 $\hat{p}_{efficacy}$ = 0.003	$\hat{n}^*_{med}$ =415 (368,442) $\hat{p}_{futile}$ = 0.50 $\hat{p}_{efficacy}$ = 0.003	$\hat{n}^*_{med}$ = 327 (273,359) $\hat{p}_{futile}$ = 0.50 $\hat{p}_{efficacy}$ = 0.003
Scenario 1 **0, 0** ***r* = 0.75**	0.025, 0.025, 0.025	0.025, 0.025, 0.025	0.025, 0.025, 0.025
	$\hat{n}^*_{med}$ =430 (361,501) $\hat{p}_{futile}$ = 0.04 $\hat{p}_{efficacy}$ = 0.18	$\hat{n}^*_{med}$ =392 (345,435) $\hat{p}_{futile}$ = 0.04 $\hat{p}_{efficacy}$ = 0.18	$\hat{n}^*_{med}$ =301 (248,348) $\hat{p}_{futile}$ = 0.07 $\hat{p}_{efficacy}$ = 0.11
Scenario 2 **0.25, 0.25** ***r* = 0.75**	0.74, 0.81, 0.77	0.74, 0.78, 0.74	0.56, 0.66, 0.61
Scenario 3 **0.15, 0.35** ***r* = 0.75**	0.74, 0.81, 0.80	0.74, 0.78, 0.78	0.56, 0.66, 0.65
Scenario 4 **0, 0.5** ***r* = 0.75**	0.74, 0.81, 0.90	0.74, 0.78, 0.90	0.56, 0.66, 0.79
Scenario 5 **0.5, 0** ***r* = 0.75**	0.74, 0.81, 0.90	0.74, 0.78, 0.92	0.56, 0.66, 0.81
	$\hat{n}^*_{med}$ =436 (365,503) $\hat{p}_{futile}$ = 0.05 $\hat{p}_{efficacy}$ = 0.16	$\hat{n}^*_{med}$ =396 (348,435) $\hat{p}_{futile}$ = 0.05 $\hat{p}_{efficacy}$ = 0.16	$\hat{n}^*_{med}$ = 307 (254,353) $\hat{p}_{futile}$ = 0.08 $\hat{p}_{efficacy}$ = 0.10
Scenario 6 **0.15, 0.35** ***r* = 0.60**	0.68, 0.77, 0.75	0.68, 0.74, 0.73	0.51, 0.62, 0.60

Note. Original planned sample size is *n* with the interim analysis after *n*/2. Allocation proportion to placebo in stage 1 of SPCD is 0.67 and weight of stage 1 information of 0.50. $\hat{n}^*_{med}$ represents the median total sample estimated size with (25,75) percentiles when SPCD is not stopped for futility or efficacy at the interim. $\hat{p}_{futile}$ and $\hat{p}_{efficacy}$ represent the probability of stopping after period 1 for futility and efficacy, respectively. Proportion of placebo non-responders is given by *r*.

Table 2. Asymptotic power for SPCD in six scenarios.

	SPCD	**SPCD with Optimal Weight in Period 2**		**SPCD with Optimal Allocation in Period 2**		**SPCD with Optimal Weight and Allocation in Period 2**	
Effect Size in Stage1, Effect Size in Stage 2	Power	Power	Optimal weight, w^*, when allocation to placebo is b=0.67	Power	Optimal allocation, b^*, when the weight is w = 0.5	Power	Optimal allocation and weight
Scenario 1 **0,0** ***r* = 0.75**	0.025	0.025	b = 0.67 w^* = [0,1]	0.025	b^* = [0,1] w = 0.50	0.025	b^* = [0,1] w^* = [0,1]
Scenario 2 **0.25,0.25** ***r* = 0.75**	0.74	0.75	b = 0.67 w^* = 0.59	0.74	b^* = 0.70 w = 0.50	0.75	b^* = 0.61 w^* = 0.63
Scenario 3 **0.15,0.35** ***r* = 0.75**	0.74	0.75	b = 0.67 w^* = 0.39	0.74	b^* = 0.70 w = 0.50	0.78	b^* = 1 w^* = 0
Scenario 4 **0,0.5** ***r* = 0.75**	0.74	0.84	b = 0.67 w^* = 0	0.74	b^* = 0.70 w = 0.50	0.92	b^* = 1 w^* = 0
Scenario 5 **0.5,0** ***r* = 0.75**	0.74	0.91	b = 0.67 w^* = 1.0	0.74	b^* = 0.70 w = 0.50	0.93	b^* = 0.50 w^* = 1
Scenario 6 **0.15,0.35** ***r* = 0.60**	0.68	0.68	b = 0.67 w^* = 0.44	0.68	b^* = 0.72 w = 0.50	0.71	b^* = 1 w^* = 0

Note. Recommended parameters (allocation proportion to placebo of b = 0.67 and weight w = 0.50) are used in the first half of the trial (Period 1) and optimal theoretical parameters in the second half of the trial (Period 2). Optimal theoretical parameters are denoted with *. Proportion of placebo non-responders is given by r. Total sample size is 300.

Table 3. Simulated power for SPCD in six scenarios.

	SPCD No Interim Look		**SPCD with Interim Look and Optimal Weight Alone**		**SPCD with Interim Look and Optimal Allocation Alone**		**SPCD with Interim Look and Optimal Weight and Allocation**	
Scenario; Effect Size in Stage1; Effect Size in Stage 2	Power	Weight and allocation	Power	*b* and estimated *w*	Power	*w* and estimated *b*	Power	Estimated *b* and *w*
Scenario 1 0, 0 r = 0.75	0.025	*b* = 0.67 *w* = 0.50	0.025	*b* = 0.67 $\widehat{w^*}$ = 0.50 (0,0.91)	0.025	$\widehat{b^*}$ = 0.70 (0.69,71) *w* = 0.50	0.025	$\widehat{b^*}$ = 0.79 (0.51, 1) $\widehat{w^*}$ = 0.38 (0,0.95)
Scenario 2 0.25, 0.25 r = 0.75	0.73	*b* = 0.67 *w* = 0.50	0.68	*b* = 0.67 $\widehat{w^*}$ = 0.60 (0.41,0.77)	0.73	$\widehat{b^*}$ = 0.70 (0.69,71) *w* = 0.50	0.68	$\widehat{b^*}$ = 0.61 (0.52,1) $\widehat{w^*}$ = 0.65 (0,0.86)
Scenario 3 0.15, 0.35 r = 0.75	0.73	*b* = 0.67 *w* = 0.50	0.68	*b* = 0.67 $\widehat{w^*}$ = 0.39 (0.15,0.57)	0.73	$\widehat{b^*}$ = 0.70 (0.69,71) *w* = 0.50	0.72	$\widehat{b^*}$ = 1 (0.63,1) $\widehat{w^*}$ = 0 (0,0.61)
Scenario 4 0, 0.5 r = 0.75	0.73	*b* = 0.67 *w* = 0.50	0.81	*b* = 0.67 $\widehat{w^*}$ = 0.02 (0,0.26)	0.73	$\widehat{b^*}$ = 0.70 (0.69,71) *w* = 0.50	0.88	$\widehat{b^*}$ = 1 (1,1) $\widehat{w^*}$ = 0 (0,0)
Scenario 5 0.5, 0 r = 0.75	0.73	*b* = 0.67 *w* = 0.50	0.89	*b* = 0.67 $\widehat{w^*}$ = 0.99 (0.83,1)	0.73	$\widehat{b^*}$ = 0.70 (0.69,71) *w* = 0.50	0.90	$\widehat{b^*}$ = 0.52 (0.50,0.55) $\widehat{w^*}$ = 1 (0.92,1)
Scenario 6 0.15, 0.35 r = 0.60	0.68	*b* = 0.67 *w* = 0.50	0.64	*b* = 0.67 $\widehat{w^*}$ = 0.43 (0.17,0.62)	0.68	$\widehat{b^*}$ = 0.71 (0.70,73) *w* = 0.50	0.67	$\widehat{b^*}$ = 1 (0.60,1) $\widehat{w^*}$ = 0 (0,0.68)

Note. Recommended parameters (allocation proportion to placebo of *b* = 0.67 and weight *w* = 0.50) are used in the first half of the trial (Period 1) and estimated optimal parameters in the second half of the trial (Period 2). Estimated optimal parameters are denoted with *. The median of the estimated optimal parameters is denoted with $\hat{w}^*$ and $\hat{b}^*$ (25,75) percentiles are in parentheses. Total sample size is 300 and the interim look is at 150 subjects.

Table 4. Estimated treatment effects, standard errors, and confidence intervals from ADAPT-A trial with analysis using unadjusted logistic regression model and with adjustment for center.

Method	*Statistic*	θ_1	θ_2	$0.5\theta_1 + 0.5\theta_2$
Unadjusted	Estimated treatment effect	0.08	1.19	0.63
	Std. Err.	0.41	0.55	0.34
	95% CI	(-0.76, 0.85)	(0.16, 2.38)	(-0.04, 1.31)
	Estimated effect size	Stage 1: 0.03	Stage 2: 0.38	Overall: 0.20
Adjusted	Estimated treatment effect	0.12	1.32	0.72
	Std. Err.	0.41	0.57	0.35
	95% CI	(-0.72, 0.90)	(0.27, 2.53)	(0.04, 1.41)
	Estimated effect size	Stage 1: 0.04	Stage 2: 0.41	Overall: 0.23

Table 5. Test statistics from ADAPT-A trial with analysis using unadjusted logistic regression model, logistic regression with adjustment for center, and the score test.

Test Statistic	Parameter in the test statistic	Unadjusted		Adjusted	
		Test Statistic	P-Value	Test Statistic	P-Value
$T_{II} = \frac{w\theta_1 + (1-w)\theta_2}{\sqrt{w^2 Var(\theta_1) + (1-w)^2 Var(\theta_2)}}$	$w = 1$	0.193	0.847	0.286	0.775
	$w = 0$	2.149	0.032	2.337	0.019
	$w = 0.5$	1.849	0.064	2.062	0.039
	$w = w^*$	2.158 $w^* = 0.008$	0.031	2.355 $w^* = 0.0148$	0.019
$T_{III} = \sqrt{v}T_1 + \sqrt{1-v}T_2$	$v = 0.5$	1.656	0.098	1.855	0.063
	$v = v^*$	2.158 $v^* = 0.110$	0.031	2.355 $v^* = 0.145$	0.019
Score test with parameter r	r = 1	1.691	0.091	-	-
	$r=r^*=12.0$	2.247	0.025	-	-

Note. The weights or the test parameter in the score test(Ivanova et al., 2011) that maximize the value of the test statistic are denoted by *.

Table 6. Binary outcome – logistic regression with and without covariates.

				Power/Type I Error			Correlation Between Stages	
Analysis	Sample Size	θ_1	θ_2	T_{II}	T_{III}	CI Coverage	Test Statistics	Logistic Regression Coefficient
Unadjusted	40	0	0	0.020	0.021	0.956	0.0015	-0.0033
	80	0	0	0.043	0.041	0.957	0.0017	0.0009
	300	0	0	0.051	0.050	0.949	0.0024	0.0025
	80	0.2	0.3	0.08	0.08	0.954	0.009	-0.015
	80	0.4	0.5	0.17	0.17	0.953	0.010	-0.016
	80	0.6	0.7	0.32	0.32	0.951	0.029	-0.015
	80	0.8	0.9	0.51	0.52	0.951	0.043	-0.011
	300	0.2	0.3	0.19	0.19	0.950	0.0096	-0.0004
	300	0.4	0.5	0.51	0.52	0.951	0.0156	-0.0022
	300	0.6	0.7	0.83	0.84	0.950	0.0270	0.0022
	300	0.8	0.9	0.97	0.98	0.949	0.0354	0.0022
Adjusted	40	0	0	0.035	0.037	0.978	0.0005	0.0027
	80	0	0	0.048	0.047	0.955	0.0009	-0.0044
	300	0	0	0.050	0.050	0.950	0.0006	0.0007
	80	0.2	0.3	0.14	0.14	0.951	0.0204	-0.0020
	80	0.4	0.5	0.27	0.27	0.951	0.0241	-0.0111
	80	0.6	0.7	0.45	0.46	0.950	0.0218	-0.0159
	80	0.8	0.9	0.64	0.65	0.954	0.0294	-0.0159
	300	0.2	0.3	0.29	0.29	0.950	0.0110	0.0001
	300	0.4	0.5	0.65	0.64	0.949	0.0172	-0.0013
	300	0.6	0.7	0.91	0.90	0.950	0.0226	-0.0035
	300	0.8	0.9	0.99	0.99	0.949	0.0304	-0.0040

Note. Type I error rate, power and confidence interval (CI) coverage for the weighted combination of the estimated treatment effects for SPCD with binary outcome, $w = 0.5$, unadjusted and adjusted for a baseline covariate. Parameters θ_1 and θ_2 are true log odds ratios in stage 1 and stage 2 of SPCD. T_{III} is the test statistic based on the weighted combination of stage 1 and stage 2 test statistics from logistic model with $v = 0.5$.

Table 7. Time-to-event analysis – without covariate.

			Power/Type I Error				Correlation Between Stages		
Sample Size	β_1	β_2	T_{IV}	T_V	T_{VI}	CI Coverage	Log Rank Test Statistics	Cox PH Test Statistics	Cox PH Coefficients
40	0	0	0.028	0.049	0.035	0.985	0.0015	-0.0004	-0.00002
80	0	0	0.041	0.049	0.044	0.967	0.00009	-0.0023	-0.0053
300	0	0	0.050	0.050	0.050	0.952	-0.00005	-0.00002	0.00024
80	-0.2	-0.3	0.05	0.08	0.05	0.972	-0.0056	-0.0226	0.0157
80	-0.4	-0.5	0.10	0.11	0.10	0.972	-0.0125	-0.0389	0.0077
80	-0.6	-0.7	0.17	0.19	0.17	0.973	-0.0147	-0.0353	0.0076
80	-0.8	-0.9	0.23	0.27	0.24	0.975	-0.0304	-0.0528	0.0081
300	-0.2	-0.3	0.19	0.19	0.18	0.951	-0.0063	-0.0135	-0.0022
300	-0.4	-0.5	0.46	0.49	0.47	0.951	-0.0116	-0.0163	0.0023
300	-0.6	-0.7	0.74	0.77	0.76	0.952	-0.0161	-0.0205	0.0060
300	-0.8	-0.9	0.90	0.93	0.92	0.949	-0.0254	-0.0301	0.0067

Note. Type I error rate, power and confidence interval (CI) coverage for the weighted combination of the estimated treatment effects for SPCD, $w = 0.5$, with time to event outcome. Parameters β_1 and β_2 are true hazard ratios in stage 1 and stage 2 of SPCD. T_{IV} is the test statistic computed based on the weighted combination of the estimated treatment effects, with estimates from Cox model without covariates; T_V is computed as the weighted combination of stage 1 and stage 2 log-rank tests; and T_{VI} is computed as the weighted combination of test statistics from the Cox model, both with $v = 0.5$.

Table 8. Time-to-event analysis – with covariates.

			Power/Type I Error			Correlation Between Stages	
Sample Size	β_1	β_2	T_{IV}	T_{VI}	CI Coverage	Cox PH Test Statistics	Cox PH Coefficients
40	0	0	0.023	0.026	0.977	-0.0013	0.0032
80	0	0	0.040	0.041	0.959	-0.0014	-0.0032
300	0	0	0.049	0.048	0.951	0.0006	0.0008
80	-0.2	-0.3	0.06	0.06	0.966	-0.0212	0.0103
80	-0.4	-0.5	0.12	0.12	0.968	-0.0387	0.0128
80	-0.6	-0.7	0.20	0.19	0.972	-0.0465	0.0250
80	-0.8	-0.9	0.28	0.27	0.976	-0.0723	0.0026
300	-0.2	-0.3	0.20	0.20	0.951	-0.0165	-0.0013
300	-0.4	-0.5	0.50	0.52	0.952	-0.0204	0.0049
300	-0.6	-0.7	0.77	0.80	0.952	-0.0266	0.0088
300	-0.8	-0.9	0.93	0.95	0.954	-0.0437	0.0071

Note. Type I error rate, power and confidence interval (CI) coverage for the weighted combination of the estimated treatment effects for SPCD, $w = 0.5$, with time to event outcome. Parameters β_1 and β_2 are true hazard ratios in stage 1 and stage 2 of SPCD in the adjusted model. T_{IV} is the test statistic computed based on the weighted combination of the estimated treatment effects, with estimates from the Cox model with covariates; T_{VI} is computed as the weighted combination of test statistics from the Cox model, both with $v = 0.5$.

Table 9. Type I error for normal, gamma, and Poisson distributed outcomes when the total sample size is 45, 60, and 90.

Method		Normal Observations			Non-Normal observations (Gamma)			Non-Normal Observations (Poisson)		
		N=45	*N*=60	*N*=90	*N*=45	*N*=60	*N*=90	*N*=45	*N*=60	*N*=90
Permutation Test Statistic	Unadjusted	0.047	0.049	0.050	0.049	0.048	0.048	0.050	0.048	0.050
Combined Stage-wise Permutation test	Unadjusted	0.050	0.050	0.050	0.050	0.049	0.049	0.048	0.048	0.049
Method 1: Bootstrapped Test Statistic	Unadjusted	0.050	0.050	0.049	0.046	0.049	0.048	0.050	0.050	0.050
Method 2: Bootstrapped Test Statistic stage group sizes maintained	Unadjusted	0.067	0.062	0.060	0.061	0.056	0.058	0.057	0.058	0.054
Method 3: Combined Stage-wise Bootstrap test	Unadjusted	0.027	0.032	0.040	0.020	0.031	0.043	0.029	0.033	0.041
Permutation Test Statistic	Adjusted	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
Combined Stage-wise Permutation test	Adjusted	0.050	0.050	0.050	0.048	0.050	0.048	0.50	0.048	0.049
Method 1: Bootstrapped Test Statistic	Adjusted	0.038	0.045	0.048	0.042	0.046	0.049	0.051	0.050	0.053
Method 2: Bootstrapped Test Statistic using only SPCD data	Adjusted	0.049	0.051	0.056	0.047	0.050	0.048	0.050	0.050	0.050
Method 3: Combined Stage-wise Bootstrap test	Adjusted	0.014	0.025	0.033	0.017	0.025	0.032	0.021	0.027	0.038

Note. Type I errors for both unadjusted and adjusted (for baseline and a covariate) are given for each of two methods of permutation test (overall permutation test and stage-wise) and three methods of the bootstrap (maintaining stage 1 group sizes, maintaining stage 1 and stage 2 group sizes, and stage-wise).

Table 10. Power for normal, gamma, exponential, and Poisson distributed outcomes for both unadjusted and adjusted (for baseline and a covariate) permutation and bootstrapped test statistics.

Method		**Normal Outcome** ***N* = 210**	**Gamma Outcome** ***N* = 60**	**Exponential Outcome** ***N* = 45**	**Poisson Outcome** ***N* = 60**
Chen Test Statistic	Unadjusted	0.81	0.81	0.79	0.81
Randomization Test Statistic*	Unadjusted	0.81	0.79	0.82	-
Permutation Test Statistic	Unadjusted	0.80	0.80	0.78	0.80
Combined Stage-wise Permutation test	Unadjusted	0.82	0.82	0.77	0.79
Method 1: Bootstrapped Test Statistic	Unadjusted	0.80	0.77	0.69	0.81
Method 2: Bootstrapped Test Statistic stage group sizes maintained	Unadjusted	0.83	0.80	0.73	0.81
Method 3: Combined Stage-wise Bootstrap test	Unadjusted	0.81	0. 71	0.44	0.76
Chen Test Statistic	Adjusted	0.81	0.82	0.78	0.81
Permutation Test Statistic	Adjusted	0.82	0.81	0.75	0.80
Combined Stage-wise Permutation test	Adjusted	0.81	0.80	0.74	0.78
Method 1: Bootstrapped Test Statistic	Adjusted	0.81	0.79	0.69	0.79
Method 2: Bootstrapped Test Statistic using only SPCD data	Adjusted	0.81	0.79	0.71	0.80
Method 3: Combined Stage-wise Bootstrap test	Adjusted	0.79	0.68	0.42	0.71

Note. Two methods of permutation test (overall permutation test and stage-wise) and three methods of the bootstrap (maintaining stage 1 group sizes, maintaining stage 1 and stage 2 group sizes, and stage-wise) are shown.

REFERENCES

Baer, L., & Ivanova, A. (2013). When should the sequential parallel comparison design be used in clinical trials? *Clinical Investigation*, *3*(9), 823–833. http://doi.org/10.4155/cli.13.74

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine, 24, 1713–1723.

Buyze, J., & Goetghebeur, E. (2013). Crossover studies with survival outcomes. *Statistical Methods in Medical Research*, *22*(6), 612–29. http://doi.org/10.1177/0962280211402258

Chen, Y.-F., Yang, Y., Hung, H. M. J., & Wang, S.-J. (2011). Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemporary Clinical Trials*, *32*(4), 592–604. http://doi.org/10.1016/j.cct.2011.04.006

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Cui, L., Hung, H. M. J., & Wang, S.-J. (1999). Modification of Sample Size in Group Sequential Clinical Trials. Biometrics, 55(3), 853–857.

Diao, L., Cook, R. J., & Lee, K.-A. (2015). Statistical Analysis of Recurrent Adverse Events. *Statistics in Practice Ser. : Statistical Methods for Evaluating Safety in Medical Product Development*. John Wiley & Sons, Incorporated. http://doi.org/10.1002/9781118763070.ch7

Doros, G., Pencina, M., Rybin, D., Meisner, A., & Fava, M. (2013). A repeated measures model for analysis of continuous outcomes in sequential parallel comparison design studies. *Statistics in Medicine*, *32*, 2767–2789. http://doi.org/10.1002/sim.5728

Dwass, M. (1957). Modified Randomization Tests for Nonparametric Hypotheses. The Annals of Mathematical Statistics, 28(1), 181–187.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. (D. R. Cox, D. V. Hinkley, N. Reid, D. . Rubin, & B. W. Silverman, Eds.)*Springer-Science+Business Media, B.V.* Springer-Science+Business Media, B.V.

Fava, M., Evins, A. E., Dorer, D. J., & Schoenfeld, D. A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: Culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*, *72*, 115–127. http://doi.org/10.1159/000069738

Fava, M., Mischoulon, D., Iosifescu, D., Witte, J., Pencina, M., Flynn, M., … Pollack, M. (2012). A double-blind, placebo-controlled study of aripiprazole adjunctive to antidepressant therapy among depressed outpatients with inadequate response to prior antidepressant therapy (ADAPT-A Study). *Psychotherapy and Psychosomatics*, *81*(2), 87–97. http://doi.org/10.1159/000332050

Fedorov, W., & Lui, T. (2007). Enrichment Design. In *Wiley Encyclopedia of Clinical Trials*. Hoboken: John Wiley & Sons, Inc.

Fleming, T. R., Harrington, D. P., & O'Brien, P. C. (1984). Designs for group sequential tests. *Controlled Clinical Trials*, *5*(4 SUPPL. 1), 348–361. http://doi.org/10.1016/S0197-2456(84)80014-8

Gao, P., Ware, J. H., & Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(6), 1184–1196. http://doi.org/10.1080/10543400802369053

Hengelbrock, J., Gillhaus, J., Kloss, S., & Leverkus, F. (2016). Safety data from randomized controlled trials: applying models for recurrent events. *Pharmaceutical Statistics*, *15*(4), 315–323. http://doi.org/10.1002/pst.1757

Huang, X., & Tamura, R. N. (2010). Comparison of Test Statistics for the Sequential Parallel Design. *Statistics in Biopharmaceutical Research*, *2*(1), 42–50. http://doi.org/10.1198/sbr.2010.08015

Ivanova, A., Li, S., Silverman, R., Wiener, L. E., & Koch, G. G. (n.d.). Randomization-based analusis of covariance for inference in sequential parallel comparison design, Submitted.

Ivanova, A., Qaqish, B., & Schoenfeld, D. A. (2011). Optimality, sample size, and power calculations for the sequential parallel comparison design. *Statistics in Medicine*, *30*(23), 2793–803. http://doi.org/10.1002/sim.4292

Ivanova, A., & Tamura, R. N. (2011). A two-way enriched clinical trial design: combining advantages of placebo lead-in and randomized withdrawal. *Statistical Methods in Medical Research*. http://doi.org/10.1177/0962280211431023

Jennison, C., & Turnbull, B. W. (2015). Adaptive sample size modification in clinical trials: start small then ask for more? *Statistics in Medicine*, *34*(29), 3793–3810. http://doi.org/10.1002/sim.6575

Lawless, J. F., & Nadeau, C. (1995). Some Simple Robust Methods for the Analysis of Recurrent Events. *Technometrics*, *37*(2), 158–168.

Lehmacher, W., & Wassmer, G. (1999). Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics*, *55*(4), 1286–1290.

Liu, Q., & Chi, G. Y. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics*, *57*(March), 172–177. http://doi.org/10.1111/j.0006-341X.2001.00172.x

Makubate, B., & Senn, S. (2010). Planning and analysis of cross-over trials in infertility. *Statistics in Medicine*, *29*(30), 3203–3210. http://doi.org/10.1002/sim.3981

Mehta, C. R., & Pocock, S. J. (2010). Adaptive increase in sample size when inteirm results are promising: A practical guide with examples. *Statistics in Medicine*, *30*(28), 3267–3284.

http://doi.org/10.1002/sim.4102

Mi, M. Y., & Betensky, R. a. (2013). An analysis of adaptive design variations on the sequential parallel comparison design for clinical trials. *Clinical Trials (London, England)*, *10*(2), 207–15. http://doi.org/10.1177/1740774512468806

Nason, M., & Follmann, D. (2010). Design and analysis of crossover trials for absorbing binary endpoints. *Biometrics*, *66*(September), 958–965. http://doi.org/10.1111/j.1541-0420.2009.01358.x

Silverman, R.K. Ivanova, A. (2017). Sample size re-estimation and other midcourse adjustments with sequential parallel comparison design. Journal of Biopharmaceutical Statistics.

Sprout Pharmaceuticals, Inc. Fixed 100 mg Every Evening of Flibanserin vs Placebo in Premenopausal Women With Hypoactive Sexual Desire Disorder. In: ClinicalTrials.gov [Internet]. (n.d.). Retrieved from https://clinicaltrials.gov/show/NCT00996164 NLM Identifier: NCT00996164

Tamura, R. N., & Huang, X. (2007). An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clinical Trials (London, England)*, *4*, 309–317. http://doi.org/10.1177/1740774507081217

Timmesfeld, N., Schäfer, H., & Müller, H. (2007). Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine*, *26*(12), 2449–2464. http://doi.org/10.1002/sim.2725

Trivedi, M. H., & Rush, J. (1994). Does a Placebo Run-In or a Placebo Treatment Cell Affect the Efficacy of Antidepressant Medications ?, *11*(1), 33–43.

Wan, H., Ellenberg, S. S., & Anderson, K. M. (2015). Stepwise two-stage sample size adaptation. *Statistics in Medicine*, *34*(1), 27–38. http://doi.org/10.1002/sim.6311

Zink RC, Koch GG, Chung Y & Wiener LE. (2017). Advanced randomization-based methods in clinical trials. In: Dmitrienko A & Koch GG, eds. Analysis of Clinical Trials Using SAS: A Practical Guide, Second Edition. Cary, NC: SAS Institute Inc.