THE RANDOMIZED PLACEBO-PHASE DESIGN: EVALUATION, INTERIM MONITORING AND ANALYSIS

by

Stephanie Shook

B.S. in Mathematics,

University of Texas at Austin, 2006

Submitted to the Graduate Faculty of

the Department of Biostatistics in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH
DEPARTMENT OF BIOSTATISTICS

This dissertation was presented

by

Stephanie L. Shook

It was defended on

July 29, 2010

and approved by

Dissertation Director:
Howard Rockette, Ph.D., Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Committee Member:
Jong Jeong, Ph.D., Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Committee Member:
Abdus Wahed, Ph.D., Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Committee Member:
Steven Belle, Ph.D., Professor, Department of Epidemiology
Graduate School of Public Health, University of Pittsburgh

THE RANDOMIZED PLACEBO-PHASE DESIGN: EVALUATION, INTERIM MONITORING AND ANALYSIS

Stephanie Shook, PhD

University of Pittsburgh, 2010

The randomized placebo-phase design, also known as the randomized delayed-start design, has been proposed as an approach to circumvent the reluctance of patients and physicians to participate in trials with a placebo control. Although there is some practical appeal to the design and it has been used in an increasing number of active and ongoing trials, there are often overlooked issues relative to statistical power, estimating sample size and determining plans for interim analysis that may limit its usefulness. We developed a general model for describing the pattern of treatment response and based on the specified parameters of this model, derive and compare different strategies for interim monitoring. In addition to statistical power considerations, we also provide results from extensive simulations investigating the robustness of the proposed procedures since the efficiency of the randomized placebo-phase design is highly dependent on the assumptions made about the form of the alternative hypotheses.

Public Health Relevance: The randomized clinical trial is the gold standard for evaluating new medical treatments/public health interventions. Indiscriminate use of the RPPD may result in failure to identify important new treatment/interventions because of low statistical power.

Keywords: Randomized placebo-phase design, Delayed treatment, Statistical power, Clinical trials, Interim monitoring.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Clinical trials for diseases in which there are no standard therapies and rare diseases in which there isn't a large population to sample from encounter problems [12]. Recruitment is one major issue, especially for the randomized controlled trial [22]. The problem of recruitment can be greater for trials with a placebo arm. Patients as well as doctors may not want to participate in a trial in which they might receive a placebo rather than the actual treatment particularly if they believe the treatment will be beneficial. Feldman et al [8] proposed a study design, the randomized placebo-phase design (RPPD), to help increase recruitment in trials with a placebo arm. There are two phases in this design; the first is similar to a randomized clinical trial with one treatment arm and one placebo arm. In the second phase, both arms receive treatment. The time to event for both groups is then analyzed.

The RPPD, as proposed by Feldman, uses standard survival analysis to compare two treatment arms where the same treatment is received in both groups but there's a delay in its administration to one of the arms. The basic idea underlying this study design is that if a treatment is actually effective, then patients who start sooner should also respond sooner [8]. Assuming time to improvement is generated from an exponential, Feldman defines a high, moderate, intermediate and low level of potency to correspond to median times of response of 7, 14, 42, and 150 days, respectively with corresponding hazards of 0.099, 0.050, 0.017 and 0.0046. Feldman assumes that the placebo has a hazard of 0.0023 which corresponds to a median time to spontaneous response of 300 days. The hazard ratio for each treatment hazard compared to the placebo hazard, from the greatest hazard to the smallest, is 43, 22, 7, and 2. Feldman then assumes statistical power for a limited number of scenarios was estimated using simulations [8]. Another design in which both groups recieve treatment

involving a delayed start in the active treatment group is called the randomized delayed-start design (RDSD). This study design was created specifically for clinical trials for Alzheimer disease [20] and has been used in trials for Parkinson's disease [10,24,25]. Unlike the RPPD, which as proposed by Feldman has a "time to event" as the primary outcome, most of the applications of the RDSD appear to have repeated continuous measures as the primary outcome [8]. The motivation behind the RDSD isn't to increase recruitment, rather it's more disease-specific. For diseases like Parkinson's and Alzheimer disease, the treatments tend to treat symptoms rather than the actual progression of the disease. The randomized placebo-controlled clinical trial doesn't attempt to distinguish between these two scenarios. Through the incorporation of a delayed active treatment in the control arm, the RDSD attempts to address this dilemma.

In the RDSD, there are two phases. In the first, the early-start arm receives treatment and the delayed-start arm receives placebo. In the second phase, both arms receive treatment. The purpose of the first phase is to confirm that the treatment has an effect on symptoms. Therefore, the improvement experienced by the early-start group may be solely due to treatment of the symptoms rather than treatment of the actual disease progression. Distinguishing between these two possibilities is the reason for the second phase. At the end of the second phase, both groups should be experiencing improvement; however the early start group should be constantly better than the delayed start group. This is interpreted to mean that the early-start group is experiencing neuroprotection while the benefits of the delayed-start group could be due to the treatment of the symptoms (similar to the early-start group in the first phase) [6]. If the delayed-start arm experiences the same improvement as the early-start arm in the second phase then the improvement is viewed as purely symptomatic. There are three different hypotheses that are tested to show that the treatment is actually slowing disease progression. At the end of Phase I the slope estimates between both arms are compared to show that a difference exists. The second compares the estimate of change in both arms from baseline to the end of Phase II. Finally, the third hypothesis tests noninferiority of the slope estimates between the groups for only Phase II. Significant differences in favor of the early-start arm must be apparent in all three of these tests to

declare the study positive. Repeated measures analysis, specifically a mixed models analysis of covariance, is used to test the above hypotheses [10].

There are several criticisms of the RDSD. D'Agostino describe the decisions that must be made when conducting a RDSD as compared with a randomized clinical trial, such as the length of each of the phases, the number of repeated measures, multiple comparisons, the statistical analysis and the handling of missing data [6]. Clarke's critique focuses specifically on flaws of the studies for Parkinson's disease, such as clinically significant differences, generalizability of study results, and a conflicting hypothesis by Shapira and Obeso [3,31], that any early symptomatic therapy will be beneficial for treating the actual progression of the disease. Therefore, the early-start group is intrinsically better off than the delayed-start group. Contrary to Clarke, Kieburtz thinks that the RDSD is well suited to test for neuroprotection in Parkinson's disease by eliminating the confounder, treatment of symptoms [15].

The RPPD and RDSD are similar in that both entail a delay in the active treatment in the control arm. However, the RDSD attempts to relate the design more closely to a hypothesized disease process and proposes multiple hypothesis tests (conducted at different phases of the study) to assess disease progression. The RPPD analysis is more simplistic. Our focus in the first several specific aims will be on the usefulness and limitations of the general concept of a delayed active treatment serving as a control arm and given that emphasis the less complex and, to some extent, more general RPPD appears to be a more reasonable framework in which to investigate the problem. However, we will integrate aspects of the RDSD approach when they appear appropriate.

2.0 AIMS

To evaluate the RPPD, different scenarios regarding statistical power will be analyzed such as the effects of the lag time (length of the first phase) and the length of the study. Statistical power depends on α (the type I error rate), δ (effect size), length of follow-up, N (sample size), the underlying distribution of response and the statistical test used. The statistical power at the end of the trial also depends on the percentage of the length of Phase 1 as compared with the total length of follow-up. In addition, the statistical power at different time points in the follow-up period will depend on the extent to which the response is time dependent.

Simulations will be used to investigate the statistical power as a function of α, δ, N (the total sample size), length of follow-up, and the percentage of the length of Phase 1 as compared with the total length of follow-up. This will be done using a proportional hazards model with the time to event assuming first an exponential then a Weibull distribution. However, since the nature of this design violates the proportional hazards assumption, an investigation into the statistical power for alternative tests to the proportional hazards model will also be done.

In addition, more appropriate methods of sample size estimation will be developed specifically for this design since standard formulas tend to assume proportional hazards. This will entail creating a general format for describing the pattern of response to therapy. The RPPD assumes treatment effects that may not be reflected with some therapies. This new format will allow flexibility to portray different types of treatments. In fact, various scenarios representing different types of therapies will be used to illustrate the effect this format has on statistical power and sample size when utilized under the RPPD. Finally, more appropriate

methods of interim monitoring will be developed since the RPPD differs so much from a randomized placebo-controlled clinical trial. With this complete evaluation, a set of conditions will be made as to when it is feasible to use this study design and how it should be monitored. In summary relative to the RPPD, we will:

1. Conduct simulations to determine the effect of length of follow-up and length of delay in treatment on statistical power for selected sample size, Type I error and δ. We will also determine the difference in sample size that is required for different lengths of delay to have the same statistical power as the standard placebo-controlled clinical trial. Simulations will be conducted for exponential and for Weibull distributed outcomes using the proportional hazards model to perform the analysis.
2. The proportional hazards model was used for analysis to address Specific Aim 1 because this was suggested in the paper by Feldman et al [8]. However, results of the RPPD clearly violate the proportionality assumption if treatment is effective. Therefore, we consider alternative methods to analyze the RPPD. Simulations will be conducted and asymptotic relative efficiency calculations will be performed to compare results obtained using this method to results obtained using the proportional hazards model.
3. Previous simulations of this design done by others have assumed a model in which the hazards in the control group change instantaneously at the end of the delay period. We developed a more general model which incorporates an additional delay time after an active treatment is initiated as well as a gradual change in hazard once the effect is manifested. In addition, the model incorporates a pattern of attenuation after treatment is terminated. Although simulations were selectively done for Specific Aims 1 and 2 under the new model, our primary rationale for developing this more general simulation framework was to provide a more robust framework to develop guidelines for interim monitoring.
4. Strategies for interim monitoring for the RPPD have not been formulated with consideration to the unique properties of the RPPD (e.g. tendency to lose power with increased follow-up) or with the potential impact that complex patterns of time dependent response may have on selecting an appropriate monitoring plan. The recommended approaches in the standard placebo-controlled trial usually entails

adjusting the Type I error with consideration being given to the number of planned interim analysis, a decision on the rate of spending of the Type I error and a recommendation to plan analyses to have approximately an equal number of events between consecutive analyses. These strategies and formulations may not be appropriate for the RPPD. Furthermore, unlike the RDSD, the RPPD does not consider any underlying biological considerations about the type or pattern of response. We developed strategies and provide interim monitoring rules within the framework of the model developed for Specific Aim 3.

5. Standard sample size formulas were evaluated and compared with simulation results to determine their accuracy. Procedures for estimating sample size will also be developed that take into consideration specific interim monitoring plans.

3.0 EVALUATING THE RANDOMIZED PLACEBO-PHASE DESIGN

3.1 DISTRIBUTIONAL PROPERTIES

3.1.1 Probability Density Function and Expected Time to Event

In this section we provide the formula for the probability density function and the expectation of time to event in the two arms of a RPPD, designated as "treatment" and "control". The treatment arm receives the active treatment immediately while the control arm has a lag of a set amount of time until it receives the treatment. Denote the hazards for the treatment and placebo as by λ_1 and λ_0, respectively. Let t_i be the corresponding time until event or censoring time for the i^{th} subject. Assuming that time to event follows an exponential distribution, the probability density function of the treatment group is

$$f(t) = \lambda_1 e^{-\lambda_1 t} \tag{3.1}$$

with expected value

$$E(t) = \frac{1}{\lambda_1} \tag{3.2}$$

The control group follows a piecewise exponential distribution with PDF

$$f(t) = \begin{cases} \lambda_0 e^{-\lambda_0 t_{P1}} & 0 \leq t < t_{P1} \\ \lambda_1 e^{-\lambda_0 t_{P1}} e^{-\lambda_1 (t - t_{P1})} & t_{P1} \leq t < \infty \end{cases} \tag{3.3}$$

where t_{P1} is the time until active treatment is given, also the length of Phase I. The expected value is

Table 3.1: Comparison of E(t) between Equation and Simulation Results

	Treatment Group E(t)		Control Group E(t)	
Hazard Ratio	Equation 3.4	Simulation	Equation 3.2	Simulation
22	20.00	20.01	73.47	73.48
7	58.82	59.13	107.89	107.68
4	111.11	110.64	152.83	155.44
2	217.39	217.29	245.41	254.69

Assuming a year study and a 60 day lag

$$\begin{aligned} E(t) &= \int t\lambda_0 e^{-\lambda_0 t_{P1}} dt + \int t\lambda_1 e^{-\lambda_0 t_{P1}} e^{-\lambda_1 (t - t_{P1})} \\ &= \frac{1}{\lambda_0} e^{-\lambda_0 t_{P1}} \left(t_{P1} - \frac{1}{\lambda_0} \right) + e^{t_{P1}(\lambda_0 - \lambda_1)} e^{-\lambda_1 t_{P1}} \left(t_{P1} + \frac{1}{\lambda_0} \right) \end{aligned} \tag{3.4}$$

The expected values for the four treatment potency levels considered by Feldman et al for both the treatment and control groups are given in table 3.1 assuming exponentially distributed times to improvement with $t_{P1} = 60$ for a year-long study [8]. Table 3.1 also compares these results against simulation results. Simulations were run with a study length of one year and a lag time of 60 days. There were 500 trials simulated, each with a total sample size of N = 100 (50/group). The agreement between actual and simulated values was good. Therefore, our subsequent simulation studies will be based on 500 replications.

3.1.2 Expected Number of Events

In clinical trials using "time to event" as the primary outcome, the sample size is usually estimated based on the expected number of events. With the RPPD, the expected number of events must be found separately for the treatment and control groups. Let n_1 be the

treatment group sample size and t_s be the total study length. For the treatment group assuming an exponential distribution,

$$E(events) = n_1 F(t_s) = n_1(1 - e^{-\lambda_1 t_s}) \tag{3.5}$$

Also, let n_0 be the control group sample size and t_{P1} be the length of Phase I (before initiation of treatment in the control group). For the control group,

$$\begin{aligned} E(events) &= n_0 F(t_{P1}) + n_0 S(t_{P1}) F(t_s - t_{P1}) \\ &= n_0(1 - e^{-\lambda_0 t_{P1}}) + n_0 e^{-\lambda_0 t_{P1}}(1 - e^{-\lambda_1(t_s - t_{P1})}) \end{aligned} \tag{3.6}$$

The results for each treatment hazard given $n_1 = n_0 = 100$, $\lambda_0 = 0.0023$, $t_{P1} = 60$ days for studies of length $t_S = 60$, 180 and 365 days assuming exponential times to improvement can be found in table 3.2. The table also compares these results to the simulation results with the same parameter values. The results show that for the longer follow-up period, the number of events in the two treatment groups is virtually the same. For a study of 2 months, which is comparable to a randomized placebo-controlled trial since the treatment delay in the control group is also 2 months, the difference between treatment arms in the expected number of events is much larger than the 6 month and 1 year studies, especially for the higher hazard ratios. This suggests that the groups will become more similar as the study continues which will lower statistical power for longer lengths of follow-up.

3.2 COX PROPORTIONAL HAZARDS MODEL

In order to evaluate the RPPD, different scenarios were considered and assessed based on the statistical power. We conducted 500 simulations for each scenario. The hazards being used are those previously described and the times to improvement are exponentially distributed. Treatments were compared using the Cox proportional hazards model [4] with $\alpha = 0.05$. The control group had a hazard rate of 0.0023 for the first time period, Phase I, and then had a hazard rate equivalent to the active treatment arm.

Table 3.2: Expected Number of Events Assuming an Exponential Distribution

		Treatment Group		Control Group	
Length of Study	Hazard Ratio	Equation 3.5	Simulation	Equation 3.6	Simulation
2 Months	22	95	96	13	13
	7	64	64	13	13
	4	42	42	13	13
	2	25	25	13	14
6 Months	22	100	100	100	100
	7	96	96	90	90
	4	81	81	71	72
	2	57	58	51	51
1 Year	22	100	100	100	100
	7	100	100	100	100
	4	96	97	94	95
	2	81	82	79	79

By treatment group, hazard ratio and study length for a 60 day lag period

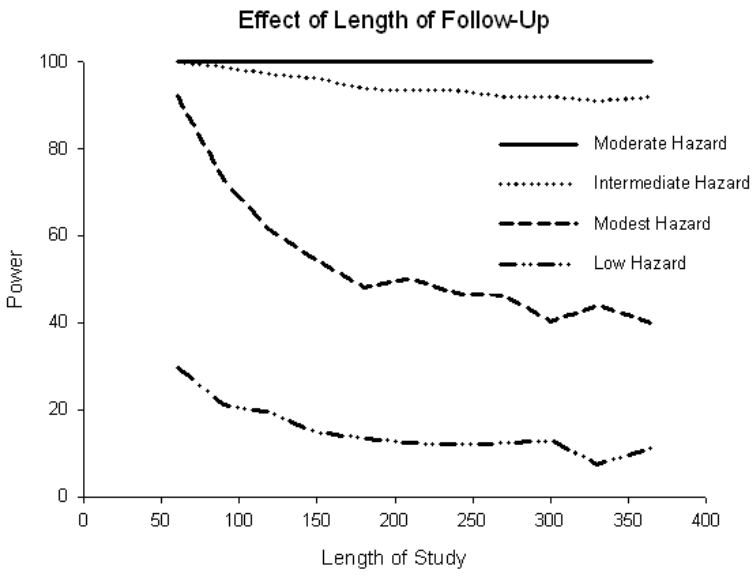

Figure 3.1: Effect of length of follow-up for each hazard with a 60 day lag assuming exponential times to improvement and a total sample size N=100.

Additional simulations were done using a Weibull distribution to investigate the robustness of our conclusions to other underlying distributions. An additional potency level with a hazard of 0.009 was created to be approximately halfway between the intermediate and low levels due to the gap between intermediate and low. We designate this potency level as modest. Also, the high and moderate levels produced similar results. Therefore, we do not present the high potency level. For investigating the effect of length of follow-up and loss of power due to delay in treatment, a total sample size of N=100 (50 per group) was used. Increases in both the lag time and length of study are presented in units of a month (30 days).

3.2.1 Effect of Length of Follow-Up

Usually for a randomized clinical trial with a placebo arm, as the duration of the study increases, the statistical power will also increase. However, this is not necessarily the case in the RPPD. Figure 3.1 exhibits this trend. A lag time of 60 days was used for the treatment

delay in the control group since this is the recommended lag by Feldman et al [8]. The different study lengths that were used started at 60 days and increased up to one year. For a study length of 60 days, the trial is a standard randomized clinical trial with a placebo arm since the length of delay of the treatment in the control arm is equal to the length of the study.

For the moderate potency level and this sample size, the statistical power is $99^{+}\%$ at all time points. In fact, for this potency level at the end of the 60 day delay period, an average of 48 (96%) events has occurred in the treatment group. For the intermediate level of potency, all of the different study lengths had power of at least 80%. The modest level also has reasonable power, at least 80%, for the 60 day trial length but there are large decreases in statistical power as follow-up increases. The statistical power to detect low potency is low and becomes even lower as the length of the study increases.

3.2.2 Loss of Power as a Function of Length of Delay in Treatment

The highest statistical power is achieved when the control arm does not receive any treatment, which is not consistent with the concept of the RPPD. A 0 day lag is uninformative since both arms would be receiving treatment for the same length of time. For simulations, we assume that the length of the study is 365 days. The delay of 365 days is used as a reference point since it represents a standard randomized clinical trial with a placebo arm. The results are presented in Table 3.3. A delay of 0 provides an estimate of the Type I error which is reasonably close to the nominal 0.05 level.

For the moderate potency level, the lag time has no effect on statistical power since the hazard is so great that the majority of patients in the active arm responds in less than 30 days and is sufficient to result in early rejection of the null hypothesis. Even with the intermediate potency, the early responses in the active treatment group dominates the comparison except for short lag periods. For the lower hazards, there is a clear decrease in statistical power as the lag time decreases.

Another way to view the loss of power due to the length of delay of treatment in the control arm is by the minimum hazard needed to achieve approximately 80% power when the

Table 3.3: Loss of Power due to Length of Treatment Delay

	Hazard Ratio			
Lag (days)	22	7	4	2
365	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	80.4%
330	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	76.8%
300	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	70.0%
270	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	65.0%
240	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	60.2%
210	$99^{+}\%$	$99^{+}\%$	99.6%	50.8%
180	$99^{+}\%$	$99^{+}\%$	98.4%	44.6%
150	$99^{+}\%$	$99^{+}\%$	93.8%	33.4%
120	$99^{+}\%$	$99^{+}\%$	85.8%	26.8%
90	$99^{+}\%$	$99^{+}\%$	77.8%	16.4%
60	$99^{+}\%$	92.6%	41.6%	12.2%
30	$99^{+}\%$	49.4%	21.2%	6.2%
0	5.6%	4.6%	5.2%	6.0%

For a year study assuming exponential times to improvement with a total sample size N=100 for each hazard ratio

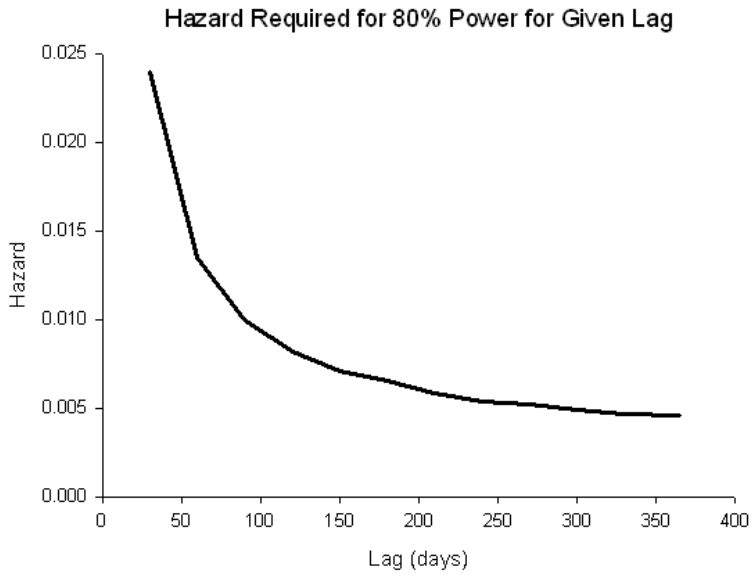

Figure 3.2: Treatment hazard required for approximately 80% statistical power for the given lag time for a year study assuming exponential times to improvement with a total sample size N=100.

Table 3.4: Sample Size Required for 80% Statistical Power

Hazard Ratio	3 Month Study	6 Month Study	1 Year Study
22	10	10	10
7	44	62	64
4	114	198	258
2	600	1180	1580

For a 3 month, 6 month and 1 year study by each hazard ratio assuming exponential times to improvement with a 60 day lag time

sample size is held constant, as displayed in Figure 3.2. As shown in table 3.3 with N=100, for a lag of one year (a pure control arm) the low potency level has approximately 80% statistical power. For a lag of 30 days, a hazard 5 times larger than the low hazard, 10 times larger than the placebo, is required to have 80% power. Consequently, for larger hazard ratios, a shorter lag can be used and still have at least 80% power. It appears that a lag period of 150 days or greater does not strongly affect the hazard required for approximately 80% statistical power.

3.2.3 Effect on Sample Size

With the RPPD, a larger sample size than the standard placebo controlled trial is needed to compensate for the loss in statistical power compared to a randomized placebo-controlled trial. The goal of the RPPD is to make recruiting easier so it is possible a larger sample size may be reached.

We have demonstrated in Section 3.2.1 that with this study design the statistical power may decrease with increasing follow-up. For longer studies, statistical power could be maintained by increasing sample size. With a 60 day delay period in the control arm, the sample size needed to obtain approximately 80% power at 90, 180, and 365 days for the various

Table 3.5: Total Sample Size Required for 80% Power for Given Lag

	Hazard Ratio			
Lag (days)	22	7	4	2
365	6	12	22	96
330	6	12	24	104
300	6	12	26	118
270	6	12	30	136
240	6	14	34	174
210	6	14	40	192
180	6	14	50	228
150	6	20	60	322
120	6	26	84	440
90	6	38	140	796
60	10	62	260	1520
30	30	200	840	5800

For a year study by each hazard ratio assuming exponential times to improvement

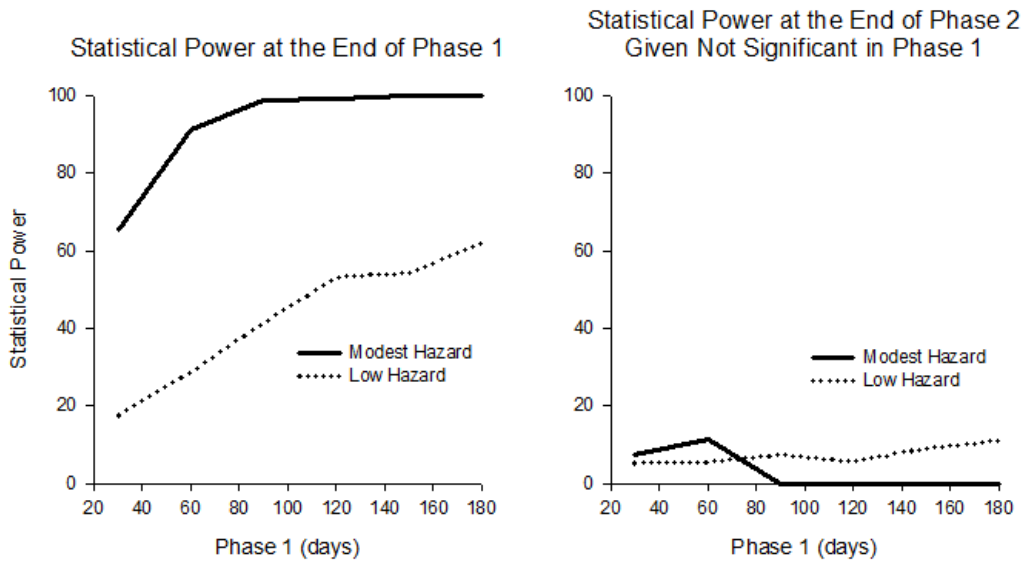

Figure 3.3: The statistical power for each phase of the RPPD for a year study assuming exponential times to improvement with a total sample size N=100. The first phase includes all trials while the second phase only includes those trials not rejected in Phase I.

hazard rates are shown in table 3.4. Table 3.5 depicts the sample size needed for 80% power when changing the lag period. As mentioned in the previous section, as the lag time increases, the sample size decreases. It will reach a point, however, when an increase in lag time will not substantially decrease the sample size required to maintain statistical power. The increase in required sample size is substantial at lower potency levels.

As noted in section 3.2.2, it appears that shorter lag periods tend to have lower statistical power. This raises the question as to how much statistical power is gained in Phase II. Although in Phase II there is no difference in hazards in the two groups, imbalances in events accumulated in Phase I may still result in a rejection rate higher than the $\alpha = 0.05$ type I error rate. Assuming a year long study, N=100 and exponential times to improvement, we vary the length of Phase I in Figure 3.3. We only show the modest and low hazards since the moderate and intermediate hazards have high statistical power as previously shown.

The first graph in Figure 3.3 shows the statistical power at the end of Phase I. The power increases as the length of Phase I increases. This is to be expected since Phase I is a randomized placebo-controlled clinical trial. The second graph in Figure 3.3 depicts the statistical power at the end of the study, or the end of Phase II, for each length of Phase I given that the trial was not significant in Phase I. This graph shows how much statistical power is added in Phase II for various lengths of Phase I. There is some additional power to be gained in Phase II, but it is very minimal. From Figure 3.3, when the length of Phase I is 180 days assuming the low hazard, 62.2% reject in Phase I. Out of the remaining 37.8% of the trials carried over to Phase II, 10.1% reject H_0 which is greater than the type I error of $\alpha = 0.05$. Regardless, most of the statistical power for this design in found in Phase I.

3.2.4 Effect of Nonexponential Improvement Times

3.2.4.1 Distributional Properties Originally, all simulations were done assuming exponential time to event. Yet this assumption may not be satisfied for many types of disease responses. To investigate the robustness of our results, we also assumed improvement times followed a Weibull distribution. Specifically, we repeated the previously presented set of simulations assuming a Weibull distribution with varying shape parameters, $\kappa = 0.5$, 0.75, 2.0, 3.0. Since the hazard function for the Weibull distribution, $H(t) =$ depends on time,a generalized form from the exponential, $\frac{1}{\lambda}$, was used for all calculations and simulations. Therefore, we denote $\theta_1 = \frac{1}{\lambda_1}$ as the treatment hazard and $\theta_0 = \frac{1}{\lambda_0}$ as the baseline hazard. The particular parameterization chosen corresponds with that of the Weibull function in R Statistical Computing Software, the program used to run the simulations. The PDF of the Weibull distribution is

$$f(t) = \frac{\kappa}{\theta_1^{\kappa}} t^{\kappa-1} e^{-\left(\frac{t}{\theta_1}\right)^{\kappa}} \tag{3.7}$$

with expected value

$$E(t) = \theta_1 \Gamma\left(1 + \frac{1}{\kappa}\right) \tag{3.8}$$

The Weibull distribution is robust since it includes densities with decreasing ($\kappa < 1$) and increasing ($\kappa > 1$). For $\kappa = 1$, the density reduces to an exponential (constant hazard) and for $\kappa = 3.6$, the Weibull is approximately normally distributed. Similar to the exponential distribution, the expected number of events are can be estimated in the two groups. For the treatment group,

$$E(events) = n_1 F(t_s) = n_1(1 - e^{-(\lambda_1 t_s)^\kappa}) \tag{3.9}$$

and for the control group,

$$\begin{aligned} E(events) &= n_0 F(t_{P1}) + n_0 S(t_{P1}) F(t_s - t_{P1}) \\ &= n_0(1 - e^{-(\lambda_0 t_{P1})^\kappa}) + n_0 e^{-(\lambda_0 t_{P1})^\kappa}(1 - e^{-(\lambda_1 (t_s - t_{P1}))^\kappa}) \end{aligned} \tag{3.10}$$

Tables 3.6, 3.7 and 3.8 give the expected number of events for the Weibull distributed times to response for studies of length 60 days, 6 months and 1 year, respectively, given $\kappa =$ 0.5, 0.75, 1.0, 2.0, 3.0 with the same hazard ratios that were used for the exponential results. The exponential results $\kappa = 1$ are included as a reference since the form of the Weibull varies for $\kappa < 1$ (event rate decreases over time), $\kappa = 1$ (event rate is constant over time) and $\kappa > 1$ (event rate increases over time). The simulation results are very similar to those calculated from equations 3.9 and 3.10, reinforcing the use of 500 replications for the simulation study.

3.2.4.2 Results In the simulation investigating the effect of length of follow-up (Figure 3.4), the Weibull distribution tends to have lower statistical power for the lower shape parameters but higher power for the higher shape parameters. For higher hazards a large number of events occur in Phase I and there are few patients at risk in Phase II. Therefore, the power is close to 1.0 and no decrease occurs. For most of the remaining scenarios, there is a pattern of power loss as follow-up increases. The exception to this occurs for high shape parameter and low potency.

We also now highlight the pattern for these scenarios. Figure 3.5 plots only the low hazard in regard to the effect of length of follow-up for each shape parameter. An increase in statistical power is apparent for the high shape parameters ($\kappa > 1$) as follow-up is increased,

Table 3.6: Number of Events Assuming a Weibull for 60 Day Study

Shape	Hazard	Treatment Group		Control Group	
Parameter	Ratio	Equation 3.9	Simulation	Equation 3.10	Simulation
0.5	22	83	83	31	31
	7	64	64	31	32
	4	52	52	31	32
	2	41	41	31	31
0.75	22	90	90	21	21
	7	64	64	21	21
	4	47	47	21	20
	2	32	32	21	21
1.0	22	95	96	13	13
	7	64	64	13	13
	4	42	42	13	13
	2	25	25	13	14
2.0	22	100	100	2	2
	7	65	65	2	2
	4	26	26	2	2
	2	8	8	2	2
3.0	22	100	100	1	1
	7	66	65	1	1
	4	15	15	1	1
	2	2	3	1	1

For selected values of the shape parameter ($\kappa = 0.5$, 0.75, 2.0 and 3.0) and hazard ratio assuming a Weibull distribution for a 60 day study with a 60 day lag

Table 3.7: Number of Events Assuming a Weibull for 6 Month Study

Shape	Hazard	Treatment Group		Control Group	
Parameter	Ratio	Equation 3.9	Simulation	Equation 3.10	Simulation
0.5	22	96	95	95	95
	7	83	83	84	84
	4	73	73	76	77
	2	60	60	68	68
0.75	22	100	100	99	99
	7	91	91	86	86
	4	77	77	73	73
	2	59	59	59	59
1.0	22	100	100	100	100
	7	96	96	90	90
	4	81	81	71	72
	2	57	58	51	51
2.0	22	100	100	100	100
	7	100	100	99	99
	4	94	94	71	71
	2	51	50	29	29
3.0	22	100	100	100	100
	7	100	100	100	100
	4	99	99	74	74
	2	45	45	17	17

For selected values of the shape parameter ($\kappa = 0.5$, 0.75, 2.0 and 3.0) and hazard ratio assuming a Weibull distribution for a 6 month study with a 60 day lag

Table 3.8: Number of Events Assuming a Weibull for 1 Year Study

Shape	Hazard	Treatment Group		Control Group	
Parameter	Ratio	Equation 3.9	Simulation	Equation 3.10	Simulation
0.5	22	99	100	99	100
	7	92	100	93	100
	4	84	96	87	94
	2	73	81	79	79
0.75	22	100	100	100	100
	7	99	100	98	100
	4	92	96	91	94
	2	78	81	79	79
1.0	22	100	100	100	100
	7	100	100	100	100
	4	96	97	94	95
	2	81	82	79	79
2.0	22	100	100	100	100
	7	100	100	100	100
	4	100	100	100	100
	2	95	94	85	86
3.0	22	100	100	100	100
	7	100	100	100	100
	4	100	100	100	100
	2	100	99	94	94

For selected values of the shape parameter ($\kappa = 0.5$, 0.75, 2.0 and 3.0) and hazard ratio assuming a Weibull distribution for a 1 year study with a 60 day lag

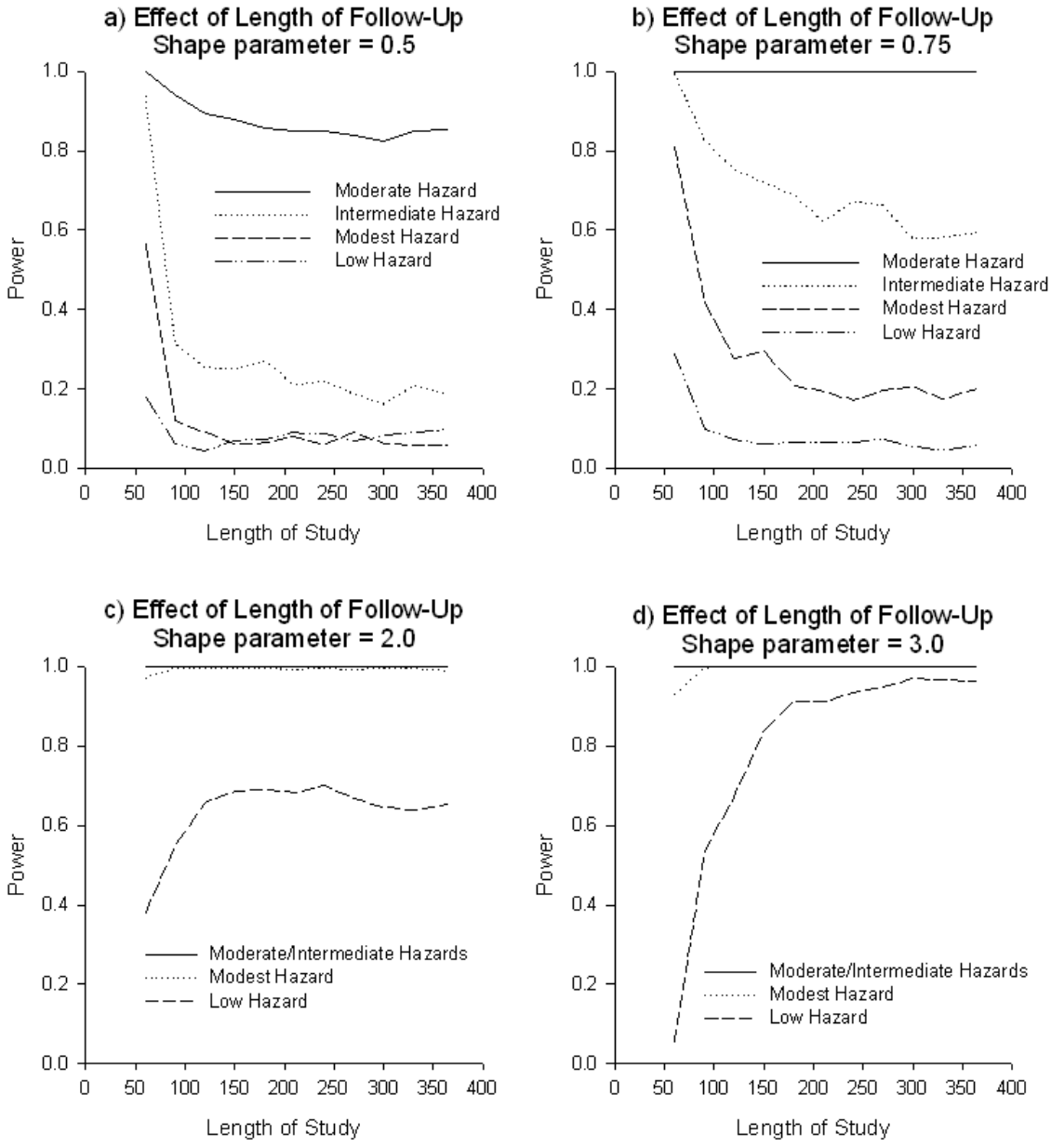

Figure 3.4: The effect of length of follow-up with a 60 day lag and total sample size N=100 assuming times to response are distributed as a Weibull with a shape parameter a) $\kappa = 0.5$, b) $\kappa = 0.75$ c) $\kappa = 2.0$, and d) $\kappa = 3.0$

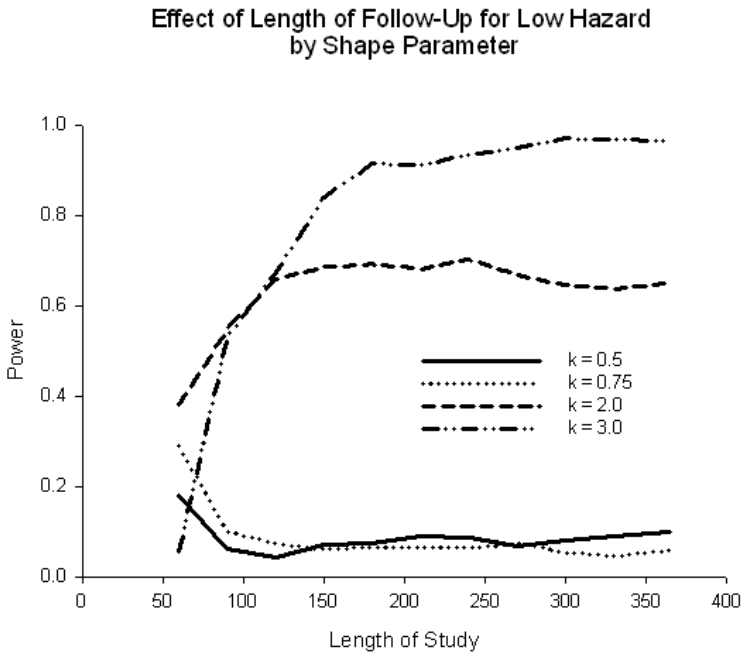

Figure 3.5: The effect of length of follow-up by shape parameter assuming times to response are distributed as a Weibull for the low hazard with a 60 day lag and total sample size N=100.

however a slight decline is seen for the low shape parameters ($\kappa < 1$) over time and which is very low for all study lengths considered.

Examination into why this combination results in increased statistical power with increased follow-up is performed. Figure 3.6 looks at the probability of event when $\kappa = 3$ for all hazards including the baseline hazard that the control group experiences. A shape parameter of 3 was chosen since this is where the increase in statistical power is most extreme for the parameters examined. The higher hazards have an immediate high response rate while the lower hazards experience a delay until events begin to occur which escalates over time. Therefore, the initial low statistical power for the low hazard ($\lambda_1 = 0.0046$) is due to the lack of events in both groups early in the study. The statistical power dramatically increases as more events begin to occur in the treatment group. This relates to the expected number of events found in tables 3.6 and 3.7. For *kappa* = 3 with the low hazard, at 60 days there are 2 events in the treatment group against 1 event in the control group while there are 45 events in the treatment group at 180 days and only 17 in the control group.

Similar to the results observed for the exponential distribution, an increased length of Phase I results in loss of statistical power which can be seen in table 3.9 except for high hazards where most of the patients in the treatment group have an event in Phase I making power close to 1.0. The lower shape parameters have lower power and a steeper decline in power while the high shape parameters have higher statistical power and appear to be less affected by the change in lag time. The lower power occurs because there are less events in Phase I. The increase in required sample size to obtain approximately 80% power continues to remain substantial at lower potency levels when data are generated from a Weibull distribution. The increase in sample size is greater for lower values of the shape parameters. These results are shown in table 3.10.

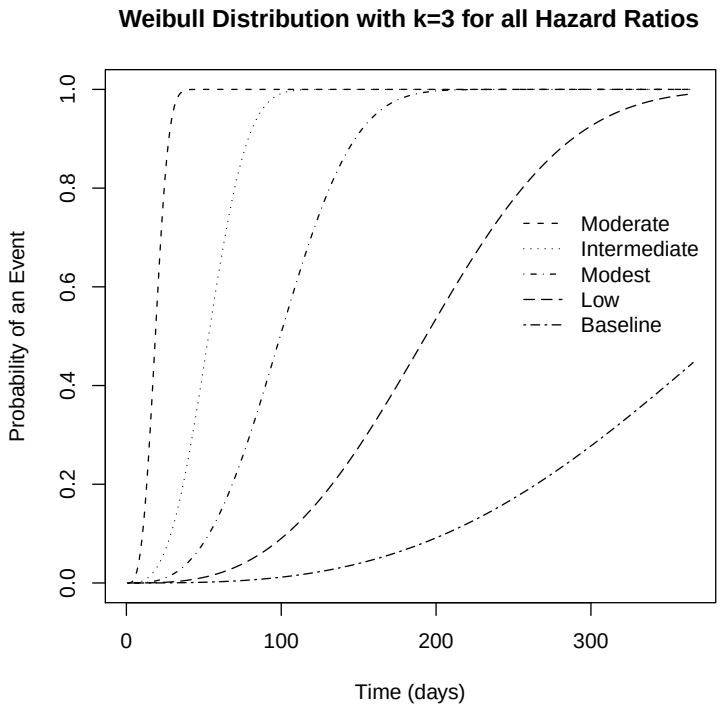

Figure 3.6: Probability of an event for the Weibull distribution with shape parameter $\kappa = 3$ by all treatment hazards and the baseline hazard.

Table 3.9: Loss of Power due to Lag Time Assuming a Weibull Distribution

		Hazard Ratio			
Shape Parameter	Lag (days)	22	7	4	2
0.5	365	99$^+$%	99.4%	82.2%	29.8%
	300	99$^+$%	90.8%	41.2%	9.2%
	240	99$^+$%	76.4%	33.2%	7.2%
	180	99.8%	59.4%	18.6%	5.8%
	120	98.0%	40.6%	9.8%	8.6%
	60	85.0%	19.6%	4.6%	6.8%
0.75	365	99$^+$%	99$^+$%	98.4%	60.4%
	300	99$^+$%	99$^+$%	95.4%	31.0%
	240	99$^+$%	99$^+$%	88.2%	22.4%
	180	99$^+$%	99.2%	71.2%	10.4%
	120	99$^+$%	93.4%	46.0%	5.8%
	60	99$^+$%	62.0%	18.6%	6.0%
2.0	365	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	300	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	240	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	180	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	120	99$^+$%	99$^+$%	99$^+$%	99.6%
	60	99$^+$%	99$^+$%	99.8%	66.6%
3.0	365	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	300	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	240	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	180	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	120	99$^+$%	99$^+$%	99$^+$%	99$^+$%
	60	99$^+$%	99$^+$%	99$^+$%	96.6%

For a year study and N=100 by each hazard ratio

Table 3.10: Sample Size Required for 80% Power Assuming a Weibull Distribution

Shape Parameter	Hazard Ratio	3 Month Study	6 Month Study	1 Year Study
0.5	22	70	84	90
	7	340	590	800
	4	1800	9400	> 10000
	2	> 10000	> 10000	> 10000
0.75	22	22	24	24
	7	92	126	162
	4	260	540	880
	2	2200	> 10000	> 10000
2.0	22	6	6	6
	7	12	10	12
	4	34	36	38
	2	168	140	150
3.0	22	6	6	6
	7	6	6	6
	4	26	16	16
	2	196	80	60

For studies of length $t_S = 60$, 180 and 365 days with a 60 day lag for each shape parameter κ by each hazard ratio

3.3 ASSESSMENT OF THE METHOD OF ANALYSIS

3.3.1 Peto and Peto Generalized Wilcoxon Test

Another problematic area of the RPPD is that it violates the assumptions of the proportional hazards model. Harrington and Fleming [11] devised the G^ρ family of tests in which $\rho = 0$ reduces to the log rank test and $\rho = 1$ is equivalent to the generalized Wilcoxon test by Peto & Peto [27]. In the case of no covariates, the Cox proportional hazards model corresponds to the log-rank test. Lee, Desu and Gehan show that the Peto & Peto and Gehan generalizations of the Wilcoxon test have more power than the log rank test when the hazard ratio is nonconstant [21]. When no censoring is present, both generalizations of the Wilcoxon will have the same results, however when the data are censored, the weight function used in Gehan's test depends on the censoring pattern [2, 27]. Therefore, the Peto & Peto test is recommended for use with the RPPD since it does not depend on the censoring pattern. The Peto & Peto test statistic in the context of the G^ρ family of test statistics has weights on each event of $S(t)$ which is the Kaplan-Meier estimate of survival.

The previous types of simulation models are used to compare the statistical power between the Peto & Peto test and the Cox model or log rank test. Figure 3.7 shows the effect on power of increasing the length of follow-up, which appears to be higher for the Peto & Peto test as compared with the Cox model, illustrated in Figure 3.1. In fact, both the moderate and intermediate potency levels achieve $99^+\%$ statistical power for all study durations. The modest potency level also has higher power than the Cox model, however there's not much of a difference for the low potency level. The higher statistical power when using Peto & Peto's test compared to the Cox proportional hazards model is reflected in the results of a statistical power comparison between various tests presented in Lee et al [21].

The Peto & Peto test also gives higher statistical power when increasing the lag time as seen in table 3.11. The results were the same as the Cox model (3.3) for the moderate hazard. For the intermediate hazard, the Peto & Peto test achieves at least 80% statistical power for all lag times while the Cox model achieved this type of power for lags of 60 days or greater. Whereas the Cox model achieved at least 80% power for lags of 120 days or

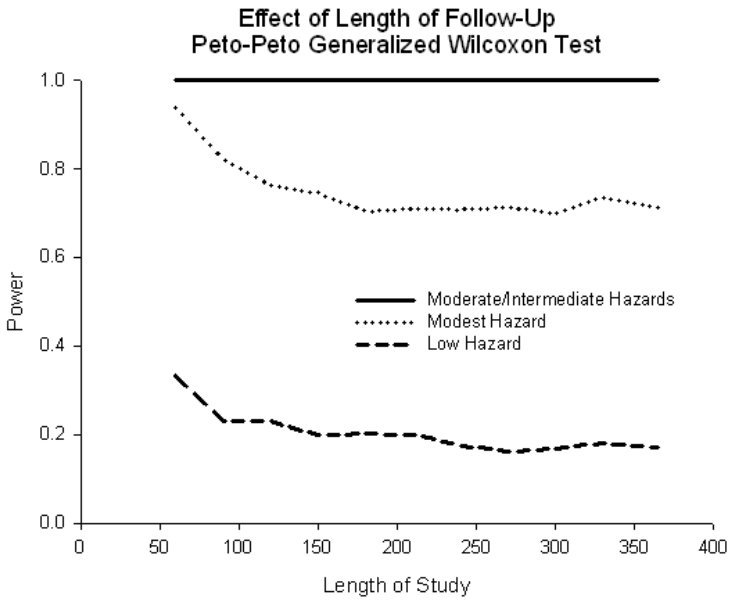

Figure 3.7: Effect of length of follow-up assuming exponential times to improvement when using the Peto & Peto test

greater for the modest potency level, the Peto & Peto test achieves this for lags of 60 days or greater. For both tests, the low potency level only achieves 80% power at a lag of year which is compatible with a randomized placebo-controlled trial. It should be noted that the lag of 0 days is included to assess the type I error.

The most apparent differences between the two models can be seen when comparing the sample size required for 80% power which is given in table 3.12. The sample sizes for the moderate potency level are the same for both models due to the high and immediate response rate of this 22-fold hazard ratio. The remaining potency levels see a large decrease in sample size when using the Peto & Peto model, with the difference increasing as the hazard decreases. Furthermore, for a year-long study with the remaining three potency levels, the sample size using the Peto & Peto test is about half of the Cox model.

Table 3.11: Loss of Power due to Lag Time for Each Hazard

	Hazard Ratio			
Lag (days)	22	7	4	2
365	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	80.0%
330	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	79.4%
300	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	70.8%
270	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	70.2%
240	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	68.0%
210	$99^{+}\%$	$99^{+}\%$	$99^{+}\%$	62.0%
180	$99^{+}\%$	$99^{+}\%$	99.8%	55.8%
150	$99^{+}\%$	$99^{+}\%$	98.8%	45.8%
120	$99^{+}\%$	$99^{+}\%$	98.6%	36.2%
90	$99^{+}\%$	$99^{+}\%$	91.4%	29.0%
60	$99^{+}\%$	99.6%	71.6%	19.0%
30	$99^{+}\%$	83.2%	31.0%	8.6%
0	5.4%	5.4%	6.6%	5.4%

For each hazard ratio assuming exponential times to improvement for a year study and total sample size N=100 using the Peto & Peto test

Table 3.12: Total Sample Size Required for 80% Power for Peto & Peto test

	3 Month Study		6 Month Study		1 Year Study	
Hazard Ratio	P&P	LR	P&P	LR	P&P	LR
22	10	10	10	10	10	10
7	32	44	34	62	36	64
4	96	114	118	198	126	258
2	540	600	670	1180	800	1580

For studies of length $t_S = 60$, 180 and 365 days with a 60 day lag assuming exponential times to improvement when using the Peto & Peto test. P&P represents the sample size when the Peto & Peto test is used while LR stands for the log rank test or Cox model

3.3.2 Asymptotic Relative Efficiency

The previous section showed that the Peto & Peto test appears to achieve higher statistical power than the log rank test, especially for lower hazard ratios and longer lengths of follow-up. The asymptotic relative efficiency (ARE) is an appropriate way to compare the Peto & Peto test and the log rank test for non-proportional hazards. Chen used the Pittman efficiency to show that the Peto & Peto has higher ARE than the log rank test under the scale family of alternatives which assumes non-proportional hazards [2]. The Pittman relative efficiency as $\delta \to 0$ is [5]

$$\lim_{n\to\infty} \frac{n}{n'} \tag{3.11}$$

where n' is the sample size for statistic V that provides the same statistical power as sample size n does for statistic T. The RPPD, however, does not fall into the scale family but rather the location family. Let V be the Peto & Peto test statistic and T be the log rank test statistic. Table 3.13 provides the relative efficiency for T (log rank statistic) and V (Peto &

Table 3.13: Relative Efficiency for Peto & Peto Test Against Log Rank Test

Power	Hazard Ratio	n'	n	ARE
0.70	7	49	29	1.690
	4	198	96	2.063
	2	1300	600	2.167
	1.5	4600	2400	1.917
	1.2	31600	17000	1.859
0.80	7	64	36	1.778
	4	258	126	2.048
	2	1580	800	1.975
	1.5	5400	2950	1.831
	1.2	30000	16000	1.875
0.90	7	90	43	2.093
	4	330	154	2.143
	2	2150	1160	1.853
	1.5	7500	4200	1.786
	1.2	45400	25400	1.787
0.95	7	110	60	1.833
	4	410	206	1.990
	2	2100	1120	1.875
	1.5	8200	4300	1.907
	1.2	50800	24800	2.048

For increasing statistical power and decreasing hazard ratios assuming exponential times to improvement for a year study with a 60 day lag

Peto statistic) for selected values of the hazard ratio assuming a 60 day lag for a year-long study assuming exponential times to improvement and for statistical power ranging from 0.70 to 0.95. The treatment hazards that are examined in this table are those previously discussed plus two additional hazards, λ_1=0.00345 and 0.00276 corresponding to hazard ratios of 1.5 and 1.2, respectively. In all cases except the moderate hazard (the hazard ratio of 22), the Peto & Peto test performs better than the log rank test. For the moderate hazard, a hazard ratio of 22, both tests perform equally well. As the statistical power increases and for the remaining lower hazards, $\delta \rightarrow 0$, the relative efficiency is in the range of 1.8-2.0 suggesting that the Peto & Peto test is approximately twice as efficient as the log rank test under the non-proportional hazards assumption built into the RPPD.

4.0 DESIGN ISSUES

The RPPD creates several challenges when designing a study. Two of these challenges are closely examined in this section. Due to the non-constant hazard ratio and the large increase in sample size required to have the same statistical power compared to a randomized placebo-controlled trial, sample size estimation becomes problematic. In addition to the hazard and the length of the study, the length of Phase I needs to be taken into account when estimating sample size.

Sample size also may depend on the interim monitoring scheme. Due to the decrease of statistical power over time and the non-constant hazard ratio, standard approaches assuming proportional hazards may not be applicable. Since ethical and cost considerations recommend the use of interim monitoring in randomized clinical trials, an appropriate method needs to be developed.

4.1 INTERIM MONITORING IGNORING MULTIPLE COMPARISONS

Having a delayed active treatment serving as a control arm usually leads to decreased statistical power for increased follow-up which provides part of the motivation for a new interim monitoring scheme. First, an investigation into the statistical power at different points during the study needs to be performed before any monitoring schemes are suggested.

As an example, we provide a simplistic monitoring scheme for the RPPD with a 60 day lag period assuming exponential times to response. In relation to the RDSD described in the introduction, an analysis is performed at the end of Phase I and another at the end of the study. We extend this so that three analyses are conducted during the duration of the

Table 4.1: Comparison of statistical power for interim analysis ignoring multiple comparisons to statistical power of the SSD

			Final Analysis		
Hazard	Analysis 1	Analysis 2	Analysis 3	Power	SSD
22	$99^{+}\%$	-	-	$99^{+}\%$	$99^{+}\%$
7	$99^{+}\%$	-	-	$99^{+}\%$	99.8%
4	92.8%	0.6%	0%	93.4%	71.6%
2	32.4%	6.0%	1.8%	40.2%	15.0%

Assuming exponential times to improvement for a RPPD with a 60 day lag and total sample size $N = 100$

study. These analyses are conducted at the end of Phase I (60 days), 6 months and at the end of the study (1 year). The first analysis includes all trials while the second includes all trials not significant (p-value ≤ 0.05) in the first analysis. The final analysis included all trials not significant in first and second analyses. For comparison purposes, we also present the statistical power if analysis is only conducted at the end of the trial (i.e. single-stage design, SSD).

Table 4.1 shows that the first analysis contains the majority of the statistical power. For the higher hazard ratios, almost all of the patients in the treatment arm have an event by the end of Phase I and thus the trial has power close to 1.0. Only for a hazard ratio as low as 2.0 is there a relevant contribution to statistical power at later analyses. In fact, as expected, the analysis at the end of Phase I has higher power than the SSD. Although, we still need to address the elevated type I error due to multiple comparisons. The results in table 4.1 suggest that interim monitoring for the RPPD includes an analysis defined by the time at which Phase I ends as was seen in the RDSD approach.

4.2 INTERIM MONITORING ISSUES SPECFIC TO THE RPPD

Certain aspects of the RPPD may create obstacles for the creation of an interim monitoring plan. For most situations, the statistical power decreases as follow-up is increased. This is due to the the hazard ratio of 1 that occurs during Phase II of the RPPD. Thus, the pattern of the hazard ratio under the alternative hypothesis is unique, with the maximum hazard ratio occuring during Phase I or into Phase II rather than at the end of the study. The interim monitoring plan should account for this challenge by conducting an analysis when this maximum hazard ratio occurs to achieve the greatest statistical power. Knowledge of the mechanism of action of the treatment would be useful in overcoming this obstacle. For instance, if the treatment is not effective immediately or it takes time to reach its full effect, the maximum hazard ratio may occur well into Phase II rather than the end of Phase I. The treatment may also attenuate before the end of the study, causing its hazard to become that of a placebo. For this particular case, the control group, which is still on treatment, may actually have a higher hazard than the treatment group, which has returned to the baseline or placbeo hazard. An appropriate interim monitoring plan and study design needs to account for these challenges in order to maintain the desired statistical power.

The previous section illustrated the statistical power for a very simplistic monitoring scheme. This scheme, however, did not take the multiple comparisons into account, thereby inflating the type I error. The monitoring plan will need to be able to maintain the appropriate type I and type II error rates without drastically increasing the expected sample size in addition to accounting for the pattern of the hazard ratio under the alternative hypothesis.

4.3 GENERAL SIMULATION MODEL

In practice, the statistical power of the RPPD is dependent on the pattern of the response to the active treatment. Prior simulations, were based on the assumption that the effect of the active treatment is immediate and lasts for the duration of the trial regardless of whether or not the active treatment is maintained. We now introduce a more general format

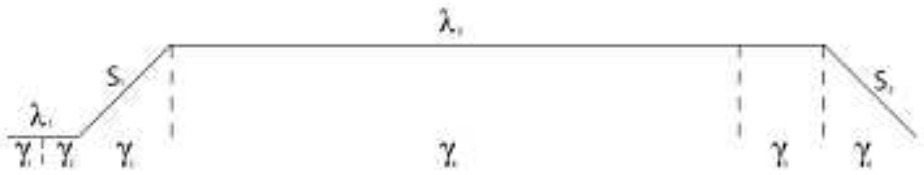

Figure 4.1: Depiction of the general simulation model and its various parameters

to characterize response to therapy that incorporates delay periods prior to the active treatment having a full effect and permits a decline in treatment effectiveness after treatment is discontinued. The parameters characterizing the response to active treatment are as follows:

λ_0 = baseline/untreated hazard

S_1 = slope between baseline and treatment hazards

S_2 = slope of the decline between treatment and baseline hazards

λ_1 = hazard associated with treatment

γ_1 = lag until treatment is given

γ_2 = lag before treatment has any effect

γ_3 = time treatment effect is increasing

γ_4 = time the full effect of treatment is sustained

γ_5 = lag until effect of treatment begins to decline

γ_6 = time the effect of treatment is declining

The baseline and treatment hazards, λ_0 and λ_1, respectively, will be as previously defined. Both slopes, S_1 and S_2, are assumed to be linear and vary depending on the treatment hazard. We recognize that a linear change in the hazard is not what is actually occurring but that for time points in the transition periods the average hazard is more likely to be a summary of a mixture of individuals, some of whom have experienced the effect of active treatment and some who are still in the delay phase. The $\gamma_i's$ represent different time periods and, as a group, characterize the hazard over time for each treatment. The parameter γ_1 is part of

the study design and for the RPPD as previously described, $\gamma_1 = 0$ in the treatment group and γ_1 is the period of treatment delay in the control group. For the situation in which the treatment effect attenuates before the study terminates, λ_0 is also the hazard after the treatment effect has diminshed in both arms. Figure 4.1 illustrates the time periods defined by the parameters of this model. A few examples will be simulated to show how different circumstances can affect the statistical power of the study. The effect of length of follow-up and the sample size required for 80% power will be investigated. Since the length of treatment delay has been incorporated into the general simulation model and its effects on statistical power have already been illustrated, the loss of power due to the length of delay in treatment will not be explored here. Since we have shown in section 3.3.1 that the generalized Wilcoxon test by Peto and Peto [11, 27] has superior power and in section 3.3.2 that it it more efficient than the log rank test, this test will be incorporated into the simulations for the general simulation model.

4.3.1 Scenarios for the General Simulation Model

Three scenarios of the general simulation model are presented here. They are:

1. $\gamma_1 = 60$ days, $\gamma_2 = 0$ days, $\gamma_3 = 60$ days, $\gamma_4 = 90$ days, $\gamma_5 = 0$ days, $\gamma_6 = 60$ days
2. $\gamma_1 = 60$ days, $\gamma_2 = 1$ day, $\gamma_3 = 3$ days, $\gamma_4 = 110$ days, $\gamma_5 = 1$ day, $\gamma_6 = 3$ days
3. $\gamma_1 = 60$ days, $\gamma_2 = 0$ days, $\gamma_3 =30$ days, $\gamma_4 = 30$ days, $\gamma_5 = 0$ days, $\gamma_6 = 30$ days
4. $\gamma_1 = 60$ days, $\gamma_2 = 0$ days, $\gamma_3 =60$ days, $\gamma_4 =$maintained for remainder of study

All of the scenarios use a lag time of $\gamma_1 = 60$ days until active treatment is initiated in the control group since the effect of this parameter has been previously discussed. The first scenario involves a long period of increasing and decreasing treatment effectiveness. The treatment is effective immediately but takes 60 days to reach its full effect which lasts for 90 days. Then its effectiveness declines for 60 days and at this time the hazard returns to the baseline value. For the second scenario, the treatment has a longer time period of full effect. The treatment takes 1 day to become effective, however it takes 3 days to reach its full effect. This effect lasts for 110 days, then after 1 day the effectiveness begins to decline. It takes 3 days to reach baseline. The third scenario involves the full effect of treatment lasting for

a short duration. The treatment is effective immediately; however it takes 30 days for it to reach its full potential which lasts for only 30 days. It experiences 30 days of decline until it reaches baseline. Finally, the last scenario has an immediate effect of treatment with 60 days until it reaches its full effect. The treatment is then maintained for the remainder of the study at its full effect.

4.3.2 Results for the General Simulation Model

First, the effect of length of follow-up was examined. For the moderate hazard, the statistical power remains $99^{+}\%$ regardless of the length of follow-up, the same result obtained for previous simulations in section 3.2.1. The other potency levels experienced similar trends to those using the basic RPPD but with a steeper decline in statistical power as length of follow-up increases for scenarios 2 and 3. Particularly, the third scenario appears to have somewhat less power than the other three which is due to the short period of maximum treatment effectiveness. The fourth scenario doesn't experience much change within each hazard as the length of follow-up continues due to the maintenance of active treatment. The first scenario experiences a different trend that has not yet been seen. The results are shown in Figure 4.2.

The different trend for the first scenario is shown in Figure 4.3 by looking at the effect of length of follow-up for only the modest hazard. Therefore, we repeated scenario 1 using a Weibull time to event. Recall that in this scenario, there is a long period of time when the treatment effect is increasing as well as a long decline in treatment effectiveness when it is not maintained. This produces an increase in statistical power then a decrease as length of follow-up increases. This trend is much more substantial and extreme for $\kappa > 1$. For $\kappa \leq 1$, an initial slight increase occurs, but generally experiences a decrease in statistical power over time. This scenario is of practical importance since it provides a useful scenario where the maximum statistical power occurs well after (90 days) the end of Phase I (60 days).

Table 4.2 displays the sample size needed to achieve approximately 80% statistical power. The sample sizes required for the low potency level for trial lengths of 6 months and 1 year are much higher for the general simulation model scenarios than the basic RPPD. This is

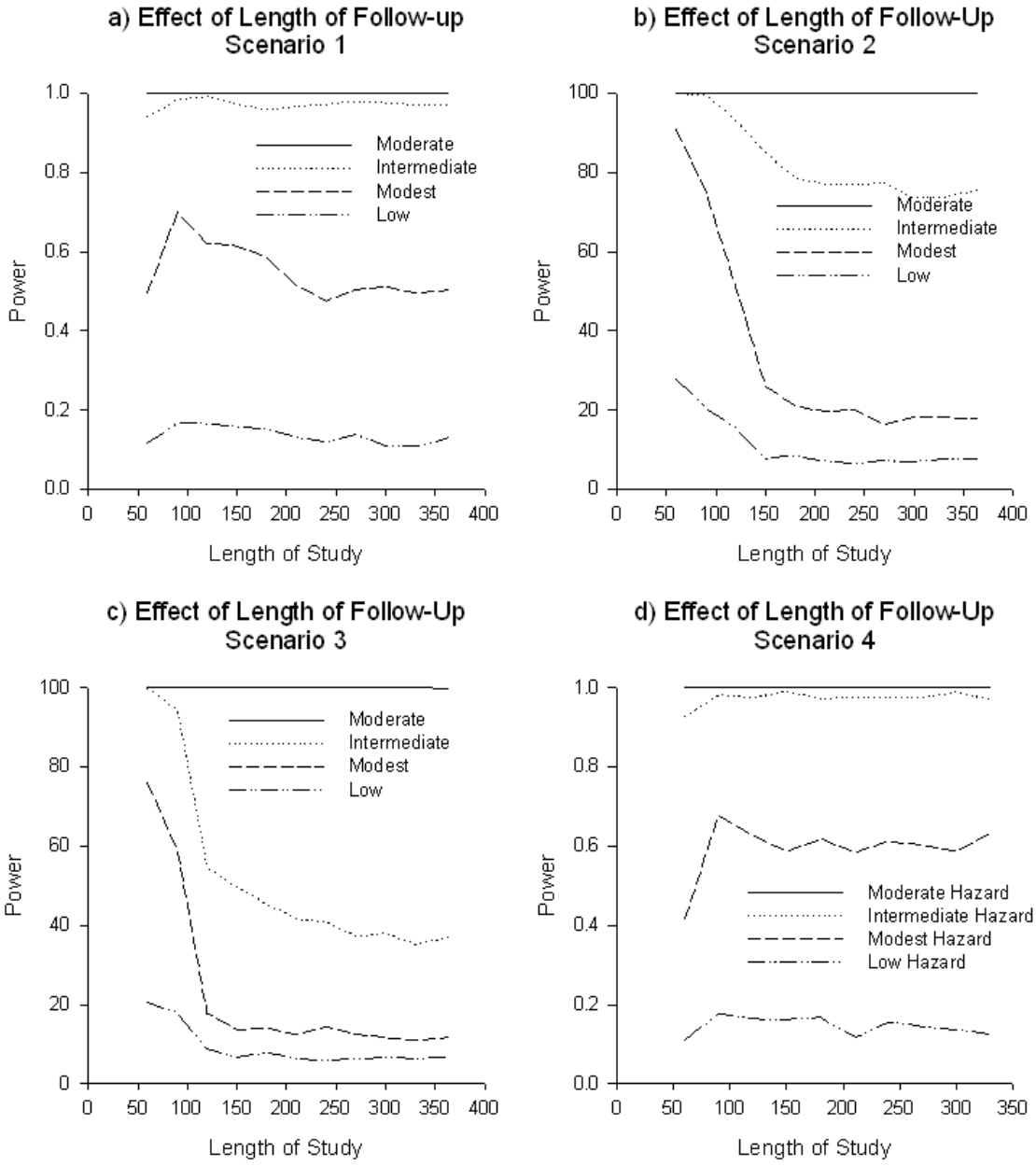

Figure 4.2: Effect of length of follow-up for each general simulation model scenario assuming exponential times to improvement for each hazard ratio.

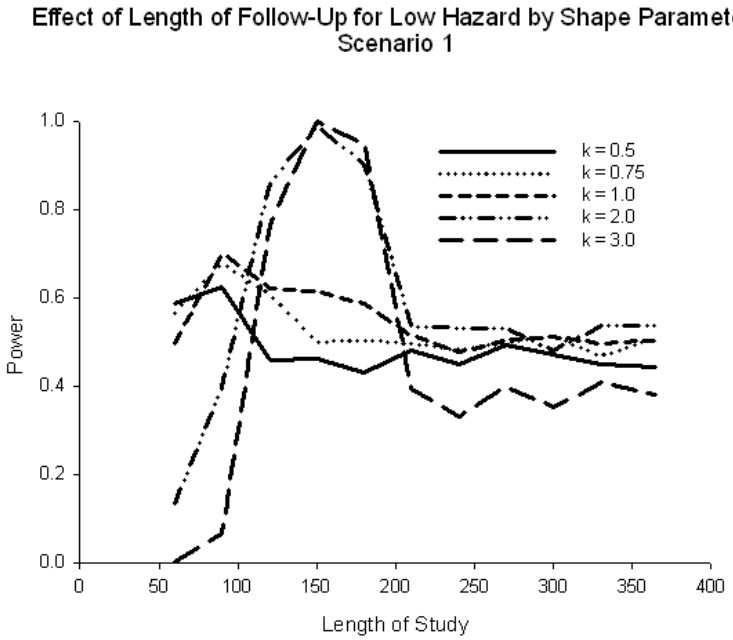

Figure 4.3: Effect of length of follow-up for the first general simulation model scenario (γ_1 = 60, γ_2=0, γ_3=60, γ_4=90, γ_5=0, γ_6=60) assuming Weibull times to improvement for the 4-fold hazard ratio by shape parameter.

Table 4.2: Sample size required for approximately 80% statistical power for selected scenarios from the general simulation model

Scenario	Hazard Ratio	3 Month Study	6 Month Study	1 Year Study
1	22	16	14	16
	7	48	50	56
	4	124	168	218
	2	760	1000	1460
2	22	10	10	10
	7	34	38	40
	4	96	176	198
	2	540	2100	2400
3	22	12	14	14
	7	50	92	100
	4	140	440	530
	2	800	3400	4200
4	22	16	14	16
	7	48	48	52
	4	124	158	170
	2	760	960	1120

By selected hazard ratios and study lengths assuming an exponential distribution

Table 4.3: Sample size required for approximately 80% statistical power for selected hazard ratios, shape parameters and study lengths for general simulation model scenario 1

		Study Length		
Hazard Ratio	κ	3 Months	6 Months	1 Year
22	0.5	34	36	38
	0.75	20	20	20
	1.0	16	14	16
	2.0	6	6	6
	3.0	6	6	6
7	0.5	74	92	98
	0.75	50	66	72
	1.0	48	50	56
	2.0	46	18	21
	3.0	70	10	10
4	0.5	160	220	220
	0.75	140	204	204
	1.0	124	168	218
	2.0	220	80	180
	3.0	630	62	276
2	0.5	370	780	770
	0.75	460	1000	1160
	1.0	760	1000	1460
	2.0	5000	600	1800
	3.0	> 10000	600	2700

Assuming Weibull distributed times to response

due to the loss of statistical power as length of follow-up increases which is the most severe for the low hazard. Similar to the prior simulations utlitizing the basic RPPD to investigate sample size, an increase in sample size to maintain statistical power is required as length of follow-up increases. The first and fourth scenarios have the same sample sizes for the 3 month study and similar sample sizes for the 6 month study due to the first three parameters in the general simulation model being identical since the study ends before the other parameters take effect.

Since scenario 1 has such a different trend than the others, table 4.3 displays the sample size required to acheive approximately 80% statistical power for each hazard ratio by shape parameter for the Weibull distribution. Restricting our focus to the modest and low hazard ratios and the shape parameters $\kappa = 2$ and κ=3, the 6 month study has a much lower sample size than the 3 month and 1 year studies. This coincides with the results in Figure 4.3 where the highest statistical power is located between 150 and 180 days. The extremely large sample size required for the low hazard and κ=3 reflects the very low statistical power seen in Figure 4.3 for a 90 days study when N=100 and is due to the low number of events as seen in Figure 3.6.

4.3.3 Interim Monitoring Scheme Ignoring Multiple Comparisons

We extend the previous simplistic monitoring scheme for several different scenarios within the general simulation model framework as was done with the basic RPPD. Three analyses were conducted during the duration of the study in a similar manner as previously discussed. The times of the first and second analyses vary for each setting. Five settings were examined, all of which assumed exponential times to improvement and incorporated the general simulation model scenario 4 with either a 90 or 60 day lag. Recall that scenario 4 assumes an immediate treatment effect, 60 days to reach full effect which is then maintained for the remainder of study. Only the results for the modest and low hazard ratios are shown due to the high and immediate response rate which results in very high statistical power at the first analysis for the moderate and intermediate hazards.

Table 4.4: Statistical power for a simplistic interim monitoring scheme for various general simulation model scenarios ignoring multiple comparisons

Hazard		$\gamma_1 = 60$ days			$\gamma_1 = 90$ days	
Ratio	Analysis	60, 180	90, 210	120, 240	90, 210	120, 240
4	1	50.4%	68.6%	59.6%	80.6%	87.0%
	2	37.9%	15.9%	11.9%	15.5%	26.2%
	Final	5.2%	3.0%	0.6%	4.3%	0%
	SSD	58.5%	58.5%	58.5%	85.6%	85.6%
2	1	12.8%	17.0%	19.0%	23.0	26.8%
	2	7.2%	4.2%	3.6%	8.6%	6.4%
	Final	2.6%	1.4%	0.8%	1.2%	1.0%
	SSD	11.8%	11.8%	11.8%	21.6	21.6

Assuming exponential times to response and a modest hazard ($\lambda_1 = 0.009$) ignoring multiple comparisons. Each setting assumes a year-long study under general simulation model scenario 4 with the columns split to consider either a 60 or 90 day lag. Recall the remaining parameters are $\gamma_2 = 0, \gamma_3 = 60, \gamma_4 = \textit{maintained}$. The row below the settings lists the times of the first and second analyses in days. The final analysis occurs at the end of the study.

Table 4.4 shows that $\gamma_1 = 90$ gives the highest statistical power for the first analysis as well as for the SSD although the hazard ratio of 2 has very low power for all analysis plans considered. The maximum statistical power occurs approximately one month after the end of the lag period for both the 60 and 90 day lags. Similar to the basic RPPD with a 60 day lag, the first analysis provides the majority of the power out of the three analyses. However, within the general model framework and for the specific parameters selected, we obtain additional power in the second analysis.

We extended this example to the Weibull distribution since it has been shown that Weibull times to response can have different trends than the exponential distribution. Here we assess the same five scenarios assuming 4-fold hazard ratio, a year long study and N=100 but for Weibull distributed times to improvement with shape parameters of $\kappa = 0.5$ and $\kappa = 2.0$. The trends for the Weibull with $\kappa < 1$ are comparable, likewise for $\kappa > 1$, so only one κ was chosen to depict the general trends for each case.

For $\kappa = 0.5$ in table 4.5, it appears that hardly any information is contained in the last two analyses, which varies from the exponential. Stopping early at the first analysis, results in higher statistical power compared with the SSD. This is most likely due to the loss of power as follow-up is increased, which is more severe for the lower shape parameters of the Weibull. The statistical power when $\kappa < 1$ is very similar at the end of the lag and one month later for both $\gamma_1 = 60$ and $\gamma_1 = 90$ days.

The shape parameter can change the results drastically. For $\kappa = 2.0$ in table 4.5, the second analysis holds some, if not more, statitsical power than the first analysis due to maximum hazard ratio occurring well after the end of the lag period. Hardly any power comes from the final analysis, which is similar to the results for $\kappa = 0.5$. In fact, for the scenarios where $\gamma_1 = 90$ days, all 500 simulated trials rejected in either the first or second analysis. Even though all of the trials reject by the final analysis, the SSD has statistical power close to 1.0 due to the high number of events occurring in the treatment group. This is the special case that was highlighted in section 3.2.4.2 involving the high shape parameter and lower hazard ratio. The results from these tables suggest that the distribution of the times to response as well as the pattern of effect expected by the active treatment are crucial

Table 4.5: Statistical power for a simplistic interim monitoring scheme for various general simulation model scenarios ignoring multiple comparisons

		$\gamma_1 = 60$ days			$\gamma_1 = 90$ days	
κ	Analysis	60, 180	90, 210	120, 240	90, 210	120, 240
0.5	1	56.8%	59.2%	50.0%	82.6%	78.4%
	2	2.0%	0%	0.2%	0%	0%
	Final	0.6%	0%	0%	0%	0%
	SSD	43.0%	41.8%	39.8%	59.6%	63.0%
2.0	1	17.8%	41.6%	86.2%	35.6%	92.0%
	2	81.8%	58.2%	13.6%	64.4%	8.0%
	Final	0.4%	0.2%	0.2%	-	-
	SSD	99^{+}%	99.8%	99.8%	99^{+}%	99^{+}%

Assuming Weibull times to response with $\kappa = 0.5, 0.75, 2.0, 3.0$ and a modest hazard ($\lambda_1 = 0.009$). Each scenario assumes a year-long study. The columns are split into scenarios with a 60 or 90 day lag with remaining parameters: $\gamma_2 = 0, \gamma_3 = 60, \gamma_4 =$ *maintained*. The row below the scenarios lists the times of the first and second analyses in days. The final analysis occurs at the end of the study.

to the design of the study and need to be taken into account when planning a study in addition to the parameters of the general simulation model.

4.4 STANDARD STOPPING RULES

Group sequential designs are used to monitor clinical trials as they progress. These designs allow for early rejection of the null hypothesis. One type of design strategy is the α-spending rate functions that allow significance testing throughout the study while keeping the type I error fixed. Two commonly used methods are Pocock's test and the O'Brien & Fleming test. The hypothesis of interest is $H_0 : \frac{\lambda_1}{\lambda_0} = 1$ against a two-sided alternative. When describing the test notation, let α' be the nominal significance level with α being the overall significance level or type I error. The basic RPPD rather than the general simulation model will be used for simplicity.

It has been shown that Peto & Peto's test should be used in place of the log rank test so the question now becomes whether this test can be used with group sequential designs, specifically the Pocock and O'Brien & Fleming stopping rules. Peto & Peto's generalization of the Wilcoxon is approximately multivariate normal [7, 11, 14]. As previously described, this test is a member of Harrington & Fleming's G^{ρ} family of test statistics ($\rho = 1$) thereby making it a member of Fleming & Harrington's $G^{\rho,\gamma}$ family of test statistics ($\rho = 1, \gamma = 0$) [9, 11]. Demirhan & Bacahn (2005) show that this latter family of tests, $G^{\rho,\gamma}$, can be used with group sequential designs [7]. Thus, the Peto & Peto test will be used in conjunction with standard stopping rules within the class of group sequential designs.

4.4.1 O'Brien and Fleming Stopping Rule

Assuming a two treatment study with arms represented as A and B, let S_k be the score statistic and $I_k = \widehat{Var}(S_k)$. This leads to the standardized statistic $Z_k = \frac{S_k}{\sqrt{I_k}}$ [13]. Let δ be

Table 4.6: Statistical power at each analysis using the Peto & Peto test with the O'Brien & Fleming stopping rule for the basic RPPD

	Analysis 1		Analysis 2		Final Analysis			
Hazard Ratio	Power	$\mathcal{I}_1$	Power	$\mathcal{I}_2$	Power	% Accepting H_0	$\mathcal{I}_3$	SSD
22	$99^+\%$	6.71	-	-	-	-	-	$99^+\%$
7	98.4%	6.20	1.2%	8.20	0.4%	0%	8.30	99.8%
4	44.6%	5.11	22.2%	8.17	11.0%	22.2%	8.35	71.6%
2	1.6%	3.85	7.0%	7.52	6.6%	85.2%	8.33	15.0%

Assuming a 60 day lag and exponential distributed times to response with N=100. The study is 1 year long with analyses conducted at 60 days, 180 days and 1 year.

the hazard ratio under H_A. The information for δ required for the fixed sample test is:

$$\mathcal{I}_{f,2} = \frac{[\Phi^{-1}(1-\frac{\alpha}{2}) + \Phi^{-1}(1-\beta)]^2}{\delta^2} \tag{4.1}$$

making the final information level

$$\mathcal{I}_F = \mathcal{I}_{f,2} R_B(K, \alpha, \beta) \tag{4.2}$$

where $R_B(K, \alpha, \beta)$ can be found in table 2.4 from Jennison & Turnball [13]. Each interim analysis should be planned to produce information levels close to $\mathcal{I}_k = \mathcal{I}_F \left(\frac{k}{K}\right), k = 1, \ldots, K$. At each analysis, the null hypothesis will be rejected if

$$|Z_k| \geq C_B(K, \alpha)\sqrt{\frac{K}{k}}, k = 1, \ldots, K \tag{4.3}$$

otherwise the study will continue. Jennison & Turnball have tabulated the critical value, $C_B(K, \alpha)$ in table 2.5 [13]. The type I error probability will be close to α as long as the information levels $I_1, \ldots, I_K$ are approximately equally spaced. The power at the final analysis, $1 - \beta$, depends on the information at the final analysis I_K.

Assuming $K = 3$ analyses and $\alpha = 0.05$, the nominal significance levels are $\alpha'_1 = 0.0006$, $\alpha'_2 = 0.014$ and $\alpha'_3 = 0.045$. Simulations were performed to assess the statistical power using O'Brien & Fleming's stopping rule for the RPPD. Compared to table 4.1, this test has much lower power for the first analysis, the loss being much more severe as the hazard decreases. Futhermore, a larger percentage of trials fail to reject H_0 for this stopping rule in the final analysis. This is due to the test being more conservative for early analyses, saving the type I error for the final analysis which for the RPPD, is a time point at which the number of events in the two treatment arms are more similar.

There appears to be higher statistical power for the interim analyses compared to the single-stage design since the attenuation of the treatment differences does not occur with the earlier analyses. The difference in power increases as the hazard decreases. However, the information is very similar for the last two analyses. This suggests a problem with the locations of the analyses. Also due to the design of the study, differences between the arms are apparent early in the study with the difference attenuating as the study continues. This trend was seen when examining the effect of length of follow-up on statistical power. This suggests that O'Brien & Fleming stopping rule may not be the most appropriate analysis to be used since it is more conservative at the beginning of the study where the differences are expected to exist.

4.4.2 Pocock Stopping Rule

Pocock's test uses the same nominal significance level at each interim analysis, which differs from the O'Brien & Fleming test. This, however, may lead to higher statistical power since Pocock's stopping rule is less conservative at the early analyses were in the RPPD, larger differences in treatment groups are more likely to exist.

Here we use similar notation found in the previous section. The information for δ required for the fixed sample test is the same as the O'Brien & Fleming stopping rule. The final information level, however, differs from that used with O'Brien & Fleming stopping rule.

$$\mathcal{I}_F = \mathcal{I}_{f,2} R_P(K, \alpha, \beta) \tag{4.4}$$

Table 4.7: Statistical power at each analysis using the Peto & Peto test with the Pocock stopping rule for the basic RPPD

	Analysis 1		Analysis 2		Final Analysis			
Hazard Ratio	Power	$\mathcal{I}_1$	Power	$\mathcal{I}_2$	Power	% Accepting H_0	$\mathcal{I}_3$	SSD
22	$99^+\%$	6.71	-	-	-	-	-	$99^+\%$
7	99.8%	6.22	0%	8.36	0%	0.2%	8.36	99.8%
4	84.6%	5.11	2.0%	8.23	0%	13.4%	8.38	71.6%
2	18.4%	3.82	4.0%	7.50	0%	77.6%	8.33	15.0%

Assuming a 60 day lag and exponential distributed times to response with N=100. The study is 1 year long with analyses conducted at 60 days, 180 days and 1 year.

where $R_P(K, \alpha, \beta)$ can be found in table 2.2 from Jennison & Turnball [13]. Similar to O'Brien & Fleming stopping rule, each interim analysis should be planned to produce information levels close to $\mathcal{I}_k = \mathcal{I}_F\left(\frac{k}{K}\right)$.
At each analysis, the null hypothesis will be rejected if

$$|Z_k| \geq C_P(K, \alpha) \tag{4.5}$$

otherwise the study will continue. Jennison & Turnball have tabulated the critical value, $C_P(K, \alpha)$ in table 2.1 [13]. This test works best when there are no more than 5 analyses performed $K \leq 5$ [13]. The type I error probability will be close to α so long as the information levels $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are approximately equally spaced. The power, $1 - \beta$, depends on the information at the final analysis $\mathcal{I}_K$.

Again, assuming $K = 3$ analyses and $\alpha = 0.05$, the nominal significance level for each interim analysis is $\alpha' = 0.022$. Table 4.7 shows the results of simulations performed to assess the statistical power of Pocock's stopping rule. The Pocock stopping rule, similar to O'Brien & Fleming stopping rule, results in higher statistical power compared to the SSD with the same type of trend. The same issues found with the O'Brien & Fleming test involving information levels are also found here. The Pocock stopping rule also experiences a decrease

in statistical power compared to the simulations ignoring multiple comparisons, however the loss of power experienced here is not as severe. The Pocock stopping rules allows for a higher nominal significance level for the first analysis compared to the O'Brien & Fleming stopping rules which explains the higher statistical power.

The information levels appear to still be a concern since they are not equally spaced. In particular, the last two analyses have similar information levels which is to be expected since the difference between the treatment arms attenuates after Phase I where the first analysis takes place. Although simultaneous accrual is not realistic so an accrual pattern needs to be built in. Accrual may also help to even out the information levels. However, if the accrual phase is short compared to the length of follow-up there will still be effect of no accumulating information for the final analyses.

4.4.3 Addition of Accrual

Simulations were conducted considering 6 months of uniform accrual for a year study with a 60 days lag using Pocock's stopping rules with the Peto & Peto test. As before, the nominal significance level will be $\alpha' = 0.022$ for an overall type I error of $\alpha = 0.05$ with N=100.

Three methods of analyzing the patient accrual are considered. In Method I, when 30% of the patients have experienced the lag, an analysis is conducted that includes all patients enrolled at that time (even if they have not experienced all of Phase I) and patients who are beyond Phase I. The second analysis is conducted once 60% of patients have experienced Phase I (in a manner similar to the first analysis) and the final analysis includes all patients. When simulations were run under H_0 for the Peto & Peto test,
the type I error was 3.4% when using Pocock's stopping rule. The results from these simulations are found in table 4.8.

In contrast, Methods II and III attempt to focus on finding the maximum hazard ratio, and in turn, the maximum statistical power. For Method II, when 30% of the patients have experienced Phase I, an analysis is conducted that includes all patients enrolled at that time even if they haven't experienced all of Phase I but only up to the end of Phase I; no information beyond Phase I is analyzed. The second analysis is conducted once 60%

Table 4.8: Interim monitoring using Pocock's boundaries and the Peto & Peto test with uniform patient accrual assuming a basic RPPD

	Analysis 1		Analysis 2		Final Analysis		
Hazard Ratio	Power	$\mathcal{I}_1$	Power	$\mathcal{I}_2$	Power	$\mathcal{I}_3$	Power
22	98.2%	4.70	1.4%	7.52	0%	8.26	
7	76.6%	4.05	9.2%	6.98	0%	8.23	
4	30.8%	3.27	13.4%	6.01	0%	8.28	
2	6.2%	2.42	4.8%	4.67	0%	8.05	

Assuming exponential times to response with a type I error of $\alpha = 0.05$ and N=100 for a study of length 1 year. When 30% of the patients have experienced the lag, an analysis is conducted that includes all patients enrolled at that time and patients who are beyond the lag. The second analysis is conducted once 60% of patients have experienced the lag (in a manner similar to the first analysis) and the final analysis includes all patients enrolled.

Table 4.9: Interim monitoring using Pocock's boundaries and the Peto & Peto test with uniform patient accrual assuming a basic RPPD

	Analysis 1		Analysis 2		Final Analysis		
Hazard Ratio	Power	$\mathcal{I}_1$	Power	$\mathcal{I}_2$	Power	$\mathcal{I}_3$	Power
22	$99^+\%$	4.23	-	-	-	-	$99^+\%$
7	88.8%	3.39	5.4%	7.04	0%	8.27	
4	44.6%	2.62	12.2%	6.04	0%	8.29	
2	6.2%	1.90	5.2%	4.64	0%	8.05	

Assuming exponential times to response with a type I error of $\alpha = 0.05$ and N=100 for a study of length 1 year. When 30% of the patients have experienced the lag, an analysis is conducted that includes all patients enrolled at that time (even if they haven't experienced the lag) but only up to the lag (no information beyond the lag is analyzed so that the maximum hazard ratio can be found). The second analysis is conducted once 60% of patients have experienced the lag (in a manner similar to the first analysis) and the final analysis includes all patients enrolled.

Table 4.10: Interim monitoring using Pocock's boundaries and the Peto & Peto test with uniform patient accrual assuming a basic RPPD

	Analysis 1		Analysis 2		Final Analysis		
Hazard Ratio	Power	$\mathcal{I}_1$	Power	$\mathcal{I}_2$	Power	$\mathcal{I}_3$	Power
22	99.8%	2.07	0.2%	4.88	-	-	$99^{+}\%$
7	72.8%	1.84	18.2%	4.57	5.2%	8.28	96.2%
4	26.4%	1.50	21.4%	4.15%	12.6%	8.34	60.4%
2	4.6%	1.09	6.2%	3.33	3.4%	8.14	14.2%

Assuming exponential times to response with a type I error of $\alpha = 0.05$ and N=100 for a study of length 1 year. When 30% of the patients have experienced the lag, an analysis is conducted that includes only those patients who have experienced the lag but only up to the lag. The second analysis is conducted once 60% of patients have experienced the lag (in a manner similar to the first analysis) and the final analysis occurs once all patients have experienced the lag.

of patients have experienced Phase I in a manner similar to the first analysis and the final analysis occurs once all patients have experienced the lag. Table 4.9 shows the results from these simulations which show higher statistical power for the intermediate and modest hazards. The moderate hazard has statistical power close to 100% for both methods while the low hazard has statistical power around 11.0% for both. These represent the two extremes. The hazard ratio of 22, or the moderate hazard has so many immediate responses that the accrual pattern does not affect the outcome of the test while the hazard ratio of 2, or the low hazard, has so few events resulting in statistical power not much higher than the type I error. When simulations were run under H_0 for the Peto & Peto test, the type I error was 3.8% when using Pocock's stopping rule and 4.0% for the SSD.

Method III only includes those who have experienced the lag in each analysis. The results in table 4.10 show that the statistical power is similar, but slightly higher than Method II. The information levels for this method appear to have a larger gap between the third and final analyses, most likely due to the uneven proportions of sample size between each analysis. Under H_0, the statistical power is 3.8% for Method II and 4.4% for Method III.

In regard to the information levels for all tables, they appear to be more evenly spaced than the previous simulations assuming simultaneous accrual, yet they are not equally spaced as desired. The moderate hazard for all methods ocnsidered rejects H_0 in the first or second analysis so there are no information levels to compare. For the remaining hazards in tables 4.8 and 4.10, the information levels observe a larger gap between the second and third interim analyses. This might be due to the fact that 40% of the patients instead of 30% as in the first two analyses, reach the end of the Phase I for the last analysis. Regardless, a longer accrual period paired with a longer follow-up period may result in non-accumulating information levels for the last two analyses which can affect the type I error. Method II in table 4.9 sees a larger gap between the first two analyses except for the low hazard which has approximately equally spaced levels of information. Method II also appears to be more affected by the attenuation of the treatment difference between both arms.

4.5 AN APPROPRIATE MONITORING SCHEME

An important component of the design of a clinical trial is the strategy for interim monitoring. Investigations in previous chapters demonstrate that many of the characteristics of the RPPD create problems for some of the standard procedures. Specifically, rules such as the O'Brien & Fleming are very conservative at early analyses and less conservative at later analyses. However, for the RPPD under many scenarios, differences are expected early rather than late. Thus, as our early simulations show, we may lose power to reject the large differences at the beginning of the study and later, when the rejection criteria is less strict, the differences in treatment groups has attenuated. The stopping rules introduced by Pocock spread the type I error evenly across all analyses which is better suited for the RPPD since it is less conservative at the beginning of the study.

As previously discussed, many interim procedures, including Pocock's stopping rules, suggest or even require approximately equal steps of "information" between analyses. However, at the later stages of follow-up for the RPPD, there is little if any information being added. The fact there is little or no information being added at the end of the trial raises the question of the benefit of increased follow-up past the end of Phase I. Our previous investigations have identified two possible reasons for continuing past Phase I:

1. A lag in the treatment effect results in the difference in events between the two groups occurring at a later time than the end of Phase I
2. The time to event is nonexponential with an increasing hazard, resulting in expected treatment differences to exist even after Phase I has ended.

The general simulation model took into account the lags in treatment effectiveness while the increasing hazard was illustrated through the incorporation of times to response distributed as a Weibull with shape parameter $\kappa > 1$. The reasoning behind the increased follow-up in Phase II is because of these two possibilities. Therefore, the pertient question appears to be are there sufficient early failures to result in rejection of H_0 prior to the attenuation of differences which will occur at some point. The problem is that we don't know exactly where the point of expected maximum cumulative differences in events will occur.

Therefore, we propose to conduct a study of length $t_S = \gamma_1 + \gamma_2$. This allows the study to end at the point where the expected maximum hazard ratio occurs as well as allowing for a lag in treatment effectiveness portrayed in the general simulation model. Using the fact that accrual is not simultaneous to our advantage, once 25%, 50%, 75% and 100% of the total patients reach the time point t_S an analysis is to be conducted. Since the effect of attenuation in Phase II is reduced, the information levels is believed to be approximately equally spaced. This type of analysis also allows for either the Pocock or O'Brien & Fleming stopping rule to be used due to the accumulation of early events not being affected by the attenuation of increased follow-up. The only issue that remains is choosing an appropriate sample size to achieve $(1 - \beta) * 100\%$ statistical power.

4.6 SAMPLE SIZE ESTIMATION

Since the RPPD has non-proportional hazards that are built into the design of the study, usual sample size formulas are often not applicable. For clinical trials in which time to event is the primary outcome, sample size formulas typically estimate the number of required events as a function of the ratio of the hazards in the two treatment arms. For example, to obtain $(1-\beta)\%$ statistical power, the total number of required events D, can be found using a standard sample size formula [28] known as Method A

$$D = \frac{4(Z_\alpha + Z_\beta)^2}{ln(\Delta)^2} \tag{4.6}$$

where $\Delta = \frac{\lambda_1}{\lambda_0}$ and Z_α and Z_β are the normal quantiles for α and β. Although the equation assumes an exponential distribution it is approximately valid for the Weibull distribution as well [28]. Statistical power can be calculated from this equation by solving

$$Z_\beta = \frac{\sqrt{D}ln(\Delta)}{2} - Z_\alpha^2 \tag{4.7}$$

Table 4.11: Comparison between the number of events required for Method A (parametric equation, 4.6) and Method B (non-parametric equation 4.9) for approximately 80% statistical power

	Equation	
Hazard	Method A	Method B
0.05	1095	1100
0.017	1359	1365
0.009	1852	1858
0.0046	3803	3809

which can be substituted into

$$Power = \Phi(-Z_\beta) \tag{4.8}$$

to find the statistical power given D, α and Δ.

An alternative formula is sometimes used which is less dependent on distributional assumptions and will be called Method B. Using the same parameters as before, let

$$D = \frac{(Z_\alpha + Z_\beta)^2(\Delta + 1)^2}{(\Delta - 1)^2} \tag{4.9}$$

This approach can also be formatted in terms of statistical power rather than sample size. Table 4.11 compares the number of events required for approximately 80% statistical power between the parametric equation (4.6), Method A, and the non-parametric equation (4.9), Method B. The calculated values are extremely close for both equations so the parametric equation will be used since the generation of the simulated data is parametric.

These formulas assume that the two treatment groups have constant (but different) hazards, an assumption violated by the RPPD. We evaluated the formula using the average hazard (weighted by the length of Phase I and Phase 2) in the control arm. All calculations assume a type I error of $\alpha = 0.05$. The number of events required for 80% statistical power is found in table 4.12 for each of the treatment hazards assuming a study of length 365 days

Table 4.12: Comparison of sample size for 80% statistical power between Method A and simulation

			Results	
Hazard	λ_0^*	Δ	Formula	Simulation
0.05	0.0422	1.18	864	10
0.017	0.0146	1.16	1075	64
0.009	0.0079	1.14	1531	258
0.0046	0.0042	1.10	3749	1518

Assuming a year study using the basic RPPD with a 60 day lag by each hazard ratio using the weighted average control group hazard to calculate the hazard ratio

with a lag of 60 days. The baseline hazard λ_0 is 0.0023. Similar results were found for Weibull distributed times to response, but are not shown here. The calculated sample sizes from the equation are much larger than those simulated. This suggests that the average hazard ratio doesn't account for the large difference occurring early in the study due to the delay of active treatment in the control group. Consequently, a new method needs to be found.

4.6.1 Simulation to Estimate Sample Size

For a conventional randomized placebo-controlled trial, once a test statistic is selected, we select δ_M, the minimum clinically important difference, supplemented by α, β, whether the hypothesis is one or two-sided and estimates of the variability of the summary statistic, to obtain estimates of N. In reality, for many of these conventional trials, δ_M is not used because of unrealistically large sample size requirements. Thus, there exists a difference δ_P that is believed will actually occur and the the difference δ' used in the study is selected between

δ_M and δ_P i.e. $\delta_M < \delta' \leq \delta_P$. For example, if $\delta_M = 20\%$ reduction in death but requires a total sample size of N=4000 while an actual reduction of $\delta_P = 50\%$ is expected to occur, a feasible sample size may lead to a study being able to detect a $\delta' = 30\%$ reduction in death.

Our goal in this section is to discuss a reasonable procedure for obtaining sample size with the RPPD. Again, this problem is more complicated than the conventional trials due to

1. nonproportionality of the treatment effect
2. the additional parameter of a lag time in treatment effectiveness, and
3. the strong possibilty that increased follow-up decreases statistical power.

We handle (1) by conducting a variety of simulations for different scenarios. (2) and (3) require additional decisions to be made in the design process such as the estimated length of the delay in treatment effectiveness and the length of the study in proportion to the length of Phase I. We suggest the following steps as a reasonable process to arrive at the sample size.

1. Estimate the baseline hazard λ_0
2. Specify the "minimum clinical difference" and provide a sample size estimate N_P for this in a regular placebo-controlled trial. If this sample size is feasible continue to the next step.
3. Decide what is the difference (δ_P) and associated hazard (λ_1) one expects to obtain based on literature, pilot data or biological considerations. For the RPPD to be feasible, λ_1 should be much greater than λ_0 i.e. $\lambda_1 \gg \lambda_0$.
4. The longest lag period that is ethical for the study needs to be determined. If longer lags are deemed ethical then a randomized placebo-controlled study should be reconsidered. Otherwise compute the sample size required if the study had length of Phase I γ_1. Another situation that could arise indicating the use of the RPPD, is that it is ethical to have a pure placebo arm, but it is believed that this will cause the accrual to be dramatically reduced.
5. Estimate, from literature, pilot data or biological considerations, the delay in the treatment effectiveness γ_2. We recommend that the analysis be conducted at $\gamma_1 + \gamma_2$.

Tables 4.13 - 4.19 provide estimates for the basic RPPD assuming exponentially distributed times to improvement with various values of the hazard ratio, length of follow-up and length of Phase I. The Peto & Peto test is used to determine the sample size for approximately 80% and 90% statistical power for different lengths of follow-up and Phase I. These tables can be used to obtain a sample size estimate that will maintain power at the desired effect.

It should also be noted that use of the RPPD requires a price to be paid in terms of a decreased statistical power or an increased sample size. The percent increase in sample size required is

$$\frac{N_{RPPD} - N_P}{N_P} * 100 \tag{4.10}$$

The percent increase in sample size may really be less than this value if the RPPD increases accrual.

The tables provided do include two parameters, γ_1, γ_2 of the general simulation model but the time of increasing treatment effectiveness γ_3 and any attenuation of treatment effects have not been considered. In the case where more parameters of the general simulation model are estimated or time to event is believed to be a Weibull distribution, then the same approach can be used but tables analogous to 4.13 - 4.19 need to be generated.

4.7 IMPLEMENTATION OF THE RPPD

In this section, we provide a practical example to demonstrate how to design a study using the RPPD. In particular, the sample size estimation procedure and interim monitoring scheme previously discussed is utilized, explained and simulated to find statistical power under various parameters associated with the RPPD such as the length of Phase I γ_1, the lag in treatment effectiveness γ_2, the type I error α, the type II error β, the baseline hazard λ_0, the minimally clinical significant difference δ_M and its associated hazard λ_M, the expected

Table 4.13: Total sample size estimation table with $\lambda_0 = 0.0023$, $\lambda_1 = 0.0046$

		Length of Follow-Up								
Power	Lag	120	150	180	210	240	270	300	330	365
80%	30	2030	2200	2400	2600	2700	2700	2740	2780	2780
	60	600	730	780	800	802	804	806	806	806
	90	300	360	390	400	406	414	426	426	430
	120	-	240	260	276	280	284	286	288	290
	150	-	-	190	196	210	216	220	222	224
	180	-	-	-	160	170	178	188	192	192
	210	-	-	-	-	150	154	156	158	160
	240	-	-	-	-	-	130	134	134	136
	270	-	-	-	-	-	-	120	123	124
	300	-	-	-	-	-	-	-	118	118
	330	-	-	-	-	-	-	-	-	104
90%	30	2200	3000	3200	3340	3420	3510	3580	3620	3670
	60	800	880	940	990	1038	1060	1090	1110	1140
	90	440	476	500	524	550	572	594	600	604
	120	-	300	330	354	380	400	404	406	408
	150	-	-	260	290	294	296	298	300	302
	180	-	-	-	210	240	242	242	244	244
	210	-	-	-	-	182	206	218	228	234
	240	-	-	-	-	-	178	194	196	198
	270	-	-	-	-	-	-	150	160	162
	300	-	-	-	-	-	-	-	144	148
	330	-	-	-	-	-	-	-	-	140

Assuming exponential times to improvement to achieve 80% and 90% statistical power by the lag time and length of follow-up in days using the Peto & Peto test

Table 4.14: Total sample size estimation table with $\lambda_0 = 0.0023$, $\lambda_1 = 0.009$

		Length of Follow-Up								
Power	Lag	120	150	180	210	240	270	300	330	365
80%	30	580	680	760	800	820	840	860	890	900
	60	160	180	200	216	232	240	250	258	270
	90	70	88	98	110	120	126	130	132	134
	120	-	52	60	68	72	78	82	84	86
	150	-	-	40	46	52	58	60	62	64
	180	-	-	-	36	40	46	46	46	46
	210	-	-	-	-	30	32	34	36	38
	240	-	-	-	-	-	28	30	32	34
	270	-	-	-	-	-	-	26	26	28
	300	-	-	-	-	-	-	-	26	26
	330	-	-	-	-	-	-	-	-	24
90%	30	660	900	980	1020	1070	1100	1120	1140	1160
	60	200	230	260	280	300	320	330	340	340
	90	100	110	126	150	158	164	170	174	180
	120	-	72	80	86	90	100	108	116	116
	150	-	-	56	62	68	74	80	80	80
	180	-	-	-	48	56	56	60	60	60
	210	-	-	-	-	40	44	48	50	52
	240	-	-	-	-	-	38	42	42	44
	270	-	-	-	-	-	-	36	38	40
	300	-	-	-	-	-	-	-	34	36
	330	-	-	-	-	-	-	-	-	32

Assuming exponential times to improvement to achieve 80% and 90% statistical power by the lag time and length of follow-up in days using the Peto & Peto test

Table 4.15: Total sample size estimation table with $\lambda_0 = 0.002$, $\lambda_1 = 0.004$

		Length of Follow-Up								
Power	Lag	120	150	180	210	240	270	300	330	365
80%	30	2600	2820	3160	3420	3580	3580	3580	3580	3580
	60	720	740	880	920	952	980	1000	1024	1040
	90	400	436	460	500	510	520	536	550	550
	120	-	280	300	310	320	324	332	340	346
	150	-	-	220	230	238	246	260	264	270
	180	-	-	-	180	190	196	204	212	214
	210	-	-	-	-	160	170	178	180	180
	240	-	-	-	-	-	140	148	152	154
	270	-	-	-	-	-	-	128	136	136
	300	-	-	-	-	-	-	-	122	128
	330	-	-	-	-	-	-	-	-	116
90%	30	3200	3420	3700	3840	4600	4600	4600	4600	4600
	60	1000	1020	1100	1200	1290	1360	1380	1390	1390
	90	500	560	600	620	630	660	700	720	730
	120	-	350	390	410	432	442	450	660	668
	150	-	-	280	290	300	300	316	326	334
	180	-	-	-	248	256	264	270	276	282
	210	-	-	-	-	210	230	240	246	250
	240	-	-	-	-	-	190	198	206	212
	270	-	-	-	-	-	-	176	184	190
	300	-	-	-	-	-	-	-	160	164
	330	-	-	-	-	-	-	-	-	154

Assuming exponential times to improvement to achieve 80% and 90% statistical power by the lag time and length of follow-up in days using the Peto & Peto test

Table 4.16: Total sample size estimation table with $\lambda_0 = 0.002$, $\lambda_1 = 0.006$

		Length of Follow-Up								
Power	Lag	120	150	180	210	240	270	300	330	365
80%	30	770	830	860	888	900	906	912	916	920
	60	230	240	252	276	290	298	298	298	298
	90	124	138	142	144	150	158	160	162	162
	120	-	94	98	104	110	110	110	110	110
	150	-	-	70	76	80	80	82	84	86
	180	-	-	-	64	70	70	72	72	72
	210	-	-	-	-	58	60	62	62	62
	240	-	-	-	-	-	52	52	56	56
	270	-	-	-	-	-	-	46	48	48
	300	-	-	-	-	-	-	-	46	46
	330	-	-	-	-	-	-	-	-	44
90%	30	1080	1140	1180	1210	1232	1232	1234	1234	1236
	60	320	332	344	356	366	384	400	416	416
	90	180	190	192	194	198	202	204	204	
	120	-	106	122	130	140	148	154	154	154
	150	-	-	96	102	106	110	114	114	114
	180	-	-	-	86	88	92	96	96	98
	210	-	-	-	-	70	76	80	82	84
	240	-	-	-	-	-	56	57	57	58
	270	-	-	-	-	-	-	52	56	56
	300	-	-	-	-	-	-	-	51	51
	330	-	-	-	-	-	-	-	-	51

Assuming exponential times to improvement to achieve 80% and 90% statistical power by the lag time and length of follow-up in days using the Peto & Peto test

Table 4.17: Total sample size estimation table with $\lambda_0 = 0.002$, $\lambda_1 = 0.008$

		Length of Follow-Up								
Power	Lag	120	150	180	210	240	270	300	330	365
80%	30	380	420	440	440	450	460	460	460	460
	60	120	134	140	142	142	144	146	148	149
	90	70	76	80	80	80	82	84	84	84
	120	-	49	52	54	56	58	59	60	60
	150	-	-	40	42	44	45	46	46	47
	180	-	-	-	36	37	37	38	39	40
	210	-	-	-	-	30	32	34	35	36
	240	-	-	-	-	-	30	31	32	33
	270	-	-	-	-	-	-	29	30	31
	300	-	-	-	-	-	-	-	29	30
	330	-	-	-	-	-	-	-	-	28
90%	30	460	520	560	590	610	614	620	622	622
	60	180	182	188	188	190	191	192	192	192
	90	100	101	102	104	104	106	106	106	108
	120	-	70	73	74	76	78	78	79	80
	150	-	-	58	60	60	60	62	62	63
	180	-	-	-	44	48	50	51	52	54
	210	-	-	-	-	42	43	44	46	46
	240	-	-	-	-	-	40	42	43	44
	270	-	-	-	-	-	-	38	40	42
	300	-	-	-	-	-	-	-	34	36
	330	-	-	-	-	-	-	-	-	34

Assuming exponential times to improvement to achieve 80% and 90% statistical power by the lag time and length of follow-up in days using the Peto & Peto test

Table 4.18: Total sample size estimation table with $\lambda_0 = 0.002$, $\lambda_1 = 0.0010$

		Length of Follow-Up								
Power	Lag	120	150	180	210	240	270	300	330	365
80%	30	240	254	260	264	270	270	272	272	272
	60	80	82	84	86	88	89	90	90	92
	90	48	50	52	54	54	56	56	56	56
	120	-	33	34	36	38	38	38	38	38
	150	-	-	28	30	30	32	32	32	32
	180	-	-	-	28	28	28	28	28	28
	210	-	-	-	-	24	24	24	24	25
	240	-	-	-	-	-	22	22	24	24
	270	-	-	-	-	-	-	22	22	22
	300	-	-	-	-	-	-	-	22	22
	330	-	-	-	-	-	-	-	-	22
90%	30	330	360	370	372	374	374	374	374	374
	60	120	120	120	120	122	122	122	122	122
	90	60	62	64	66	68	68	70	72	72
	120	-	50	50	50	51	52	52	52	52
	150	-	-	36	38	40	42	42	42	42
	180	-	-	-	34	34	36	36	36	36
	210	-	-	-	-	28	30	32	32	32
	240	-	-	-	-	-	28	28	30	30
	270	-	-	-	-	-	-	26	28	28
	300	-	-	-	-	-	-	-	26	27
	330	-	-	-	-	-	-	-	-	26

Assuming exponential times to improvement to achieve 80% and 90% statistical power by the lag time and length of follow-up in days using the Peto & Peto test

Table 4.19: Total sample size estimation table with $\lambda_0 = 0.002$, $\lambda_1 = 0.012$

		Length of Follow-Up								
Power	Lag	120	150	180	210	240	270	300	330	365
80%	30	180	182	182	182	182	182	182	182	182
	60	58	58	62	62	62	62	62	62	62
	90	34	36	36	38	38	38	38	38	38
	120	-	28	28	28	28	28	28	28	28
	150	-	-	24	24	24	24	24	24	24
	180	-	-	-	21	21	22	22	22	22
	210	-	-	-	-	20	20	20	20	20
	240	-	-	-	-	-	18	18	19	19
	270	-	-	-	-	-	-	18	18	18
	300	-	-	-	-	-	-	-	18	18
	330	-	-	-	-	-	-	-	-	17
90%	30	216	232	244	244	244	246	246	246	246
	60	80	80	80	80	82	82	82	82	82
	90	50	50	50	50	50	50	50	50	50
	120	-	35	36	36	36	36	38	38	38
	150	-	-	32	32	32	32	32	32	32
	180	-	-	-	26	26	26	26	28	28
	210	-	-	-	-	24	26	26	26	26
	240	-	-	-	-	-	24	24	24	24
	270	-	-	-	-	-	-	24	24	24
	300	-	-	-	-	-	-	-	22	22
	330	-	-	-	-	-	-	-	-	22

Assuming exponential times to improvement to achieve 80% and 90% statistical power by the lag time and length of follow-up in days using the Peto & Peto test

treatment difference to actually occur δ_P and its associated hazard λ_1, the total sample size of the randomized placebo-controlled trial N_P and the total sample size of the RPPD N_{RPPD}.

We begin by following a similar procedure found in the previous section on sample size estimation. When meeting with clinicians to design a study, first the baseline hazard λ_0 needs to be specified. The minimum clinically significant difference δ_M and its associated hazard relative to the baseline hazard λ_1 also need to be determined. The sample size for a 1 year study N_Y and a 6 month study N_M should be calculated using equation 4.6 or 4.9 for specific values of α and β. If these sample sizes are feasible, then a randomized clinical trial with a pure placebo arm could be conducted if desired. However, if the sample size estimates are too large for the trial to be implemented then an RPPD will not be practical since the sample size requirements will likely be even higher.

If it is deemed unethical to place patients in a pure placebo arm or if adequate accrual will be difficult because of the presence of a placebo arm, then the RPPD could be considered. The next question is to determine what difference δ_P is expected to actually occur with the treatment. As previously stated, the difference used in the design of the study δ' should be such that $\delta_M < \delta' \leq \delta_P$. If δ_P is much larger than δ_M then we could consider running at a higher difference than δ_M and still obtain a significant difference.

Note this is an important conceptual difference in the RPPD and the standard clinical trial. In designing the standard trial, the investigators utilize the minimum clinically significant difference. The RPPD is often applied in situations where it is expected that the event rate without treatment is low and the expected effect of treatment is high. The loss of statistical power due to the shortened period of time when the two arms are treated differently, is compensated for by a high difference in δ. However this is no longer the minimal clinically significant difference.

A decision on the length of the delay in treatment in the control arm γ_1 is the next step to be taken. Usually ethical considerations will guide the decision-making process. Using the same standard sample size equations as before, the sample size for a study of length γ_1 can be determined. If this sample size is reasonable, theoretically a randomized clinical trial with a pure placebo arm could be used but due to the long process of patient accrual, those

in the placebo arm would need to wait a period of time until the study ends and the results are analyzed before they can be given treatment while the RPPD would provide treatment sooner.

The introduction of the general simulation model provides a framework to characterize the various responses to therapy. Adding another parameter γ_2 to take into account the period of time until the treatment becomes effective will better estimate when the maximum hazard ratio will occur. We assume that once the treatment takes effect there will not be a period of time while it is increasing to its full effect i.e. it is fully effective immediately. With the addition of γ_2 the maximum difference between the two treatment arms will be at time $t_S = \gamma_1 + \gamma_2$. Using tables 4.13-4.19, the sample size N_{RPPD} can be found given t_S and γ_1 for either 80% or 90% statistical power assuming exponential times to improvement.

It is unlikely that knowledge of the exact value of γ_2 exists so it should be noted that a maximum value should be used. If the maximum difference between the groups occurs before t_S, there will still be sufficient power for an analysis conducted at t_S by using the appropriate sample size from the one of the tables. If too small a value for γ_2 is chosen, the analysis will take place before the maximum difference thus lowering statistical power.

Given that the study length t_S, the treatment and baseline hazards, λ_1 and λ_0, respectively, and the sample size N_{RPPD} have been identified, the interim monitoring procedure should be chosen. Assuming uniform accrual patterns, four analyses will be conducted once 25%, 50%, 75% and 100% of patients have completed the study however only the last two procedures will be utilized here. Method I will not be examined due to the low statistical power seen in the previous simulations. Recall that Method II includes the percentage of patients (25%, 50%, 75% and 100%) who have completed the study plus any other enrolled patients while Method III only incorporates the percentage that have completed the study.

Assume that the baseline hazard is $\lambda_0 = 0.0023$ and a minimal clinically significant difference is $\delta_M = 2$. The required total sample size for a year study is $N_Y = 64$ and for a 6 month study is $N_M = 108$. The treatment, however, is expected to have an actual difference of $\delta_P = 4$ with a hazard of $\lambda_1 = 0.009$. Suppose it is deemed unethical to have a pure placebo arm longer than 2 months so the RPPD with $\gamma_1 = 60$ days has been chosen as the study design. The total sample size N_P is found to be $N_P = 126$ for a study of length $\gamma_1 = 60$.

However, it is realized that the treatment usually doesn't take effect until 2 months later ($\gamma_2 = 60$ days) so this sample size is no longer applicable. Using table 4.14 a total sample size of $N_{RPPD} = 160$ should result in 80% statistical power given $t_S = \gamma_1 + \gamma_2 = 120$ days. These N_{RPPD} patients will be accrued assuming uniform accrual over a period of 6 months or 180 days.

Both Pocock and O'Brien & Fleming's stopping rules will be used to account for the multiple comparisons. The desirability of not being conservative at early analyses observed in table 4.7 is not applicable here. In that analysis, all patients entered at the same time point and repeated analyses were done over time while the proposed monitoring procedure eliminates the attenuation of treatment differences by stopping the study at the maximum hazard ratio. The nominal significance level for Pocock's stopping rule is $\alpha' = 0.0182$ while the nominal significance levels for O'Brien & Fleming's stopping rule are: $\alpha'_1 = 0.00006$, $\alpha'_2 = 0.00388$, $\alpha'_3 = 0.01838$ and $\alpha'_4 = 0.04116$ [13]. Due to the inclusion of these group sequential methods, the sample sizes need to be altered the maintain the type II error. Using tables 2.2 and 2.4 from Jennison & Turnball (2000), the adjusted sample sizes for the Pocock and O'Brien & Fleming stopping rules are

$$\begin{aligned} N_{POC} &= N_{RPPD} * R_P(K, \alpha, \beta) = 160 * 1.202 = 194 \\ N_{OBF} &= N_{RPPD} * R_B(K, \alpha, \beta) = 160 * 1.024 = 164 \end{aligned} \tag{4.11}$$

respectively where $K = 4$ is the maximum number of analyses to be conducted. Each interim analysis should occur once $\frac{N}{K}$ patients have been accrued. This will not occur when using Method II due to the fact that the third analysis will take place at approximately 255 days into the study (average accrual time for the 75^{th} percentile of $N_{RPPR} = 160$ patients being accrued for 6 months is 135 days being followed until the end of the study, $t_S = 120$ days) while the accrual has ended at 180 days. This means that the third and final analyses will have the same sample size since all patients have been accrued before the third analysis. Not all patients have completed the study for the third analysis, but are included. All patients will have completed the study for the final analysis though. The results for Method II are

Table 4.20: Interim monitoring using Pocock and O'Brien & Fleming (OBF) boundaries

Stopping	Analysis 1		Analysis 2		Analysis 3		Final Analysis		Overall
Rule	Power	$\mathcal{I}_1$	Power	$\mathcal{I}_2$	Power	$\mathcal{I}_3$	Power	$\mathcal{I}_4$	Power
Pocock	28.8%	7.64	18.0%	9.98	20.4%	11.09	6.0%	11.29	73.2%
OBF	2.6%	6.89	17.6%	8.47	38.6%	9.41	14.4%	9.63	73.2%

Assuming uniform patient accrual for 6 months with exponential times to response utilizing the Method II analysis plan and the Peto & Peto test with $\lambda_0 = 0.0023$ and $\lambda_1 = 0.009$. The type I error is $\alpha = 0.05$, the type II error is $\beta = 0.20$ with $N_{RPPD} = 194$ for the Pocock stopping rule and $N_{RPPD} = 164$ for the O'Brien & Fleming stopping rule assuming a study of length 4 months.

included in table 4.20 to illustrate that the type II error is no longer maintained which can be seen with the low statistical power and similar information levels at the last two analyses. Under H_0, the type I error was 3.4% for Pocock's boundary and 5.2% for O'Brien & Fleming's boundary.

Table 4.21 shows the statistical power and information level at each analysis using Pocock and the O'Brien & Fleming boundaries for Method III. The information levels are approximately equally spaced with approximately 80% statistical power, suggesting accurate type I and type II errors. When simulations were run under H_0 using the Peto & Peto test, the type I error was 5.8% for Pocock's boundary and 5.2% for O'Brien & Fleming's boundary with Method III. Our proposal is that interim analyses be conducted only at a time point in the range of values where differences are likely to be maintained and the repeated analysis is done based on the proportion of patients reaching that point as done in Method III. The O'Brien & Fleming or the Pocock stopping rule are reasonable procedures to consider when the sample size is adjusted to account for the group sequential method being implemented.

Table 4.21: Interim monitoring using Pocock and O'Brien & Fleming (OBF) boundaries

Stopping	Analysis 1		Analysis 2		Analysis 3		Final Analysis		
Rule	Power	$\mathcal{I}_1$	Power	$\mathcal{I}_2$	Power	$\mathcal{I}_3$	Power	$\mathcal{I}_4$	Power
Pocock	20.6%	3.00	27.2%	5.87	17.4%	8.63	15.2%	11.32	80.4%
OBF	1.0%	2.48	15.2%	4.94	36.4%	7.33	24.8%	9.70	77.4%

Assuming uniform patient accrual for 6 months with exponential times to response utilizing the Method III analysis plan and the Peto & Peto test with $\lambda_0 = 0.0023$ and $\lambda_1 = 0.009$. The type I error is $\alpha = 0.05$, the type II error is $\beta = 0.20$ with $N_{RPPD} = 194$ for the Pocock stopping rule and $N_{RPPD} = 164$ for the O'Brien & Fleming stopping rule assuming a study of length 4 months.

5.0 CONCLUSIONS AND DISCUSSION

The RPPD may facilitate recruitment in trials with a placebo control arm. However, there are several issues that should be carefully considered prior to implementing the RPPD. These are as follows:

1. There can be a large loss in statistical power compared to the standard randomized clinical trial with a "true" placebo arm. This effect is particularly large if the delay period for treatment in the control arm is short compared to the study length but may not occur if the response rate is high in the treatment group.
2. The large loss of power may not occur if the response rate is very high in the active treatment group. The hazard rates evaluated by Feldman et al [8] ranged from 2 (low potency) to 43 (high potency). For many diseases, a response to treatment is not realistic. Even if such high responses are reasonably expected to occur, it should be recognized that is does not represent a δ corresponding to the minimal clinically significant difference (the convention adapted in traditional design).
3. The statistical power may be highly dependent on whether there is a time-dependent response to active treatment (e.g. delay in response after receiving treatment or attenuation of effect once treatment is discontinued). As part of the sensitivity analysis in estimating the required sample size, simulations should be conducted considering the possibility of a time-dependent response to active treatment.
4. For many of the possible scenarios that can occur, the RPPD loses statistical power as the follow-up increases. Thus the standard approaches for interim monitoring are not applicable if the analysis includes a large number of patients whose follow-up period is in the range of expected power loss (Phase II).

5. The design clearly violates the assumptions of the Cox proportional hazards model. Alternative statistical tests may increase the statistical power. Our initial simulations show that the generalization of the Wilcoxon test by Peto and Peto appears to have greater statistical power than the Cox proportional hazards model suggested by Feldman [8].

In addition to these concerns, the design issues of interim monitoring and sample size are vital to the proper implementation of the RPPD. As mentioned in item 3 above, the response to therapy may be time-dependent. A general simulation model was created as a tool to identify an appropriate monitoring scheme and maximize the potential statistical power of the study. This model takes into account the treatment delay period in the control arm, the effectiveness of the treatment and its attenuation. This model is crucial when executing the RPPD since the maximum hazard ratio will usually occur shortly after the end of Phase I due to the delayed effect of treatment and the initiation of treatment in the control group. A lag in the treatment effect could extend the expected time at which the difference in treatment arms is likely to occur. These factors should be built into the design.

As suggested by Feldman et al [8], the RPPD should be used for studies of rare diseases or diseases in which there is no standard therapy. Preferably, the treatment should be much more effective than the comparison therapy to achieve at least 80% statistical power. Other aspects of the RPPD such as the length of Phase I and the length of the study need to be appropriate to maintain a feasible sample size since most of the information is contained in the first phase of the design.

The non-proportionality of the hazards and the attenuation of the difference between treatment arms, creates difficulty in choosing an appropriate monitoring scheme. The monitoring scheme should allow for the advantages of using the RPPD, such as ethical concerns and accrual difficulties, while maintaining the desired type I and type II errors with a feasible sample size. The sample size estimation procedure uses standard formulas in conjunction with simulation tables to ease the implementation of this design.

There are limitations to this monitoring procedure. First, the concept of a futility analysis may be relevant. Under simultaneous accrual, a futility analysis is preferred so the trial can be stopped if no early difference is detected since the difference only attenuates as the study

progresses. However, ending the study at time point t_S limits the need for a futility analysis. Regardless, it may still need to be incorporated if no difference or a negative difference is seen in an early analysis. A conditional power approach, particularly stochastic curtailment, can be utilized ad hoc to predict the probability of rejecting H_0 given the current data. It should be noted that conditional power approaches can result in a decrease of statistical power, particularly a trade-off between cost in power and gain with respect to sample size which can be controlled by the choice of the stopping threshold [32]. This stopping threshold, the value of the parameter of interest and "optimal" information fraction used to calculate the conditional power are all arbitrary choices that need to be made with this approach [33].

There are group sequential designs that provide a format to support early stopping under H_0, H_A or both. We used the Pocock [29] and O'Brien & Fleming [23] stopping rules for early rejection of H_0. There are other types of α-spending methods, such as those developed by Lan & DeMets [19], Wang & Tsiatis [34] and Kim & DeMets [16] among others, that could also be incoporated. Pampallona & Tsiatis (1994) extended the methods of Wang & Tsiatis (1987) to include early stopping for futility [26, 34]. These boundaries also include as special cases, Pocock and O'Brien & Fleming stopping rules and thus should occur at equally spaced information levels [26, 34]. Kittelson & Emerson (1999) created a unifying family of group sequential tests encompassing the α-spending methods and Whitehead's treiangular test among others, that includes early stopping for efficacy and futility [17]. Whitehead's triangular test and a similar type of test, the sequential probability ratio test (SPRT), incorporate early stopping in favor of H_0. These tests allow great flexibility in the timing of the analyses without altering the statistical properties [30]. The SPRT has parallel boundaries leading to a open continuation region while the triangular test, on the other hand, has convergent boundaries. Initially, Whitehead's triangular test appeared to be sufficient for this design, however construction of the boundaries requires information about the hazard ratio under the alternative hypothesis H_A. Due to the non-proportionality of the hazards built into the framework of the RPPD, this is difficult to specify. A larger hazard ratio will lead to boundaries closer together while a small hazard ratio will lead to boundaries that are further apart and hence more difficult to cross. Another dilemma when using this design along with the boundaries constructed by Pampallona & Tsiatis is

the assumption that $H_A : \theta > 0$ where $\theta = -ln\left(\frac{\lambda_1}{\lambda_0}\right)$ which corresponds to reducing the number of events as in the case where the event is death or failure. Implementation of these designs using $H_A : \theta < 0$ to correspond to the treatment increasing the number of events, as is the case with the RPPD, causes difficulty with calculating the boundaries. The lower and upper boundaries are reversed which changes the statistical properties of the test. It is assumed that with algebraic manipulation, these boundaries could possibly be incorporated into the RPPD, however the approach taken here successfully outlines a procedure that can successfully be used to monitor the RPPD while maintaining the desired type I and type II errors without large increases in sample size.

In addition to a stochastic curtailment procedure, a test of increasing hazard in the treatment group could be conducted if no differences are apparent. To employ this test, we can obtain maximum likelihood estimates of the shape parameter κ and test $H_0 : \kappa = 1$ vs $H_A : \kappa > 1$. Cumulative differences in events of the two groups also could be investigated. If there is no significant difference, no evidence of increasing hazards and no evidence of increasing cumulative difference of events in the two groups, consideration should be given to stopping the trial, except for possible continuation of toxicity monitoring. Again, this is a process to be determined as the study is progressing.

Another limitation of this monitoring scheme is the additional parameters in the general simulation model. Assuming maintenance of treatment, another parameter of interest is γ_3 that represents the time while the hazard is increasing from baseline to its full effect. We did not include sample size estimation tables incorporating γ_3 but simulations could be performed to do so. In this case, there are two possible stopping points. One is having the study end at time point $t_S = \gamma_1 + \gamma_2 + \gamma_3$, however results in table 4.4 suggest that when $\gamma_1 = 60$ days, $\gamma_2 = 0$ and $\gamma_3 = 60$ days the maximum hazard occurs at $\gamma_1 + \gamma_2 + \frac{1}{2}\gamma_3$. As previously mentioned, the maximum values for γ_2 and γ_3 should be used to ensure that the analysis is conducted where the maximum difference between the treatment groups occurs or shortly after rather than before this point. Simulations can be used to assess both stopping points.

A final consideration is the fact that treatment may affect each patient differently. Thus, the parameters γ_1 and γ_2 could vary by patient. This was not taken into account in any of

the simulations but could affect the statistical power of the study. This is a valid concern in all types of clinical trials, however the RPPD is more sensitive to these parameters than the usual clinical trial.

BIBLIOGRAPHY

[1] L. J. Bain and M. Engelhardt. *Introduction to Probability and Mathematical Statistics*. Duxbury, Pacific Grove, California, 1992.

[2] Y. Chen. Simple-tree weighted logrank tests for right-censored data. *Annals of the Institute of Statistical Mathematics*, 50:311–324, 1998.

[3] C. E. Clark. Are delayed-start design trials to show neuroprotection in parkinson's disease fundamentally flawed? *Movement Disorders*, 23:784–789, 2008.

[4] D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society Series B*, 34:187–220, 1972.

[5] T. Colton D. Armitage, editor. *Encyclopedia of Biostatistics*, volume 4, pages 3380–3384. John Wiley and Sons LTD., New York, 1998.

[6] R. B. D'Agostino Sr. The delayed-start study design. *New England Journal of Medicine*, 361:1304–1306, 2009.

[7] Y. P. Demirhan and S. Bacanh. Group sequential test of non-parametric statistics for survival data. *Hacettepe Journal of Mathematics and Statistics*, 34:67–74, 2005.

[8] W. E. Feldman. The randomized placebo-phase design for clinical trials. *Journal of Clinical Epidemiology*, 54:550–557, 2001.

[9] T. R. Fleming and D. P. Harrington. A class of hypothesis tests for one and two samples of censored survival data. *Comm. Statistics*, 10:763–794, 1981.

[10] Parkinson Study Group. A controlled, randomized, delayed-start study of rasagiline in early parkinson disease. *Archives of Neurology*, 61:561–566, 2004.

[11] D. Harrington and T. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69:553–566, 182.

[12] M. J. H. Huibers. An alternative trial design to overcome validity and recruitment problems in primary care research. *Family Practice*, 21:213–218, 2004.

[13] C. Jennison and B. W. Turnball. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall, New York, 2000.

[14] C. Jennison and B. W. Turnball. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22:850–857, 2003.

[15] K. Kieburtz. Issues in neuroprotection clinical trials in parkinson's disease. *Neurology*, 66:850–857, 2006.

[16] K. Kim and D. L. DeMets. Design and analysis of group sequential tests based on the type i spending rate functions. *Biometrika*, 74:149–154, 1987.

[17] J. M. Kittelson and S. S. Emerson. A unifying family of group sequential test designs. *Biometrics*, 55:874–882, 1999.

[18] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2003.

[19] K. K. G Lan and D. L. DeMets. Discrete sequential boundaries for monitoring clinical trials. *Biometrika*, 70, 1983.

[20] P. Leber. Observations and suggestions on antidementia drug development. *Alzheimer Disease and Associated Disorders*, 10:31–35, 1996.

[21] E. T Lee, M. M. Desu, and E. A. Gehan. A monte carlo study of the power of some two-sample tests. *Biometrika*, 62:425–432, 1975.

[22] A. M. McDonald, R. C. Knight, M. K. Campbell, V. A. Entwistle, A. M. Grant, J. A. Cook, D. R. Elbourne, D. Francis, J. Garcia, I. Roberts, and C. Snowdone. What influences recruitment of randomized controlled trials? a review of trials funded by two uk funding agencies. *Trials*, 7:9–16, 2006.

[23] P. C. O'Brien and T. R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

[24] C. W. Olanow, R. A. Hauser, J. Jankovic, W. Langston, A. Lang, W. Poewe, E. Tolosa, F. Stocchi, E. Melamed, E. Eyal, and O. Rascol. A randomized, double-blind, placebo-controlled, delayed-start study to assess rasagiline as a disease modifying therapy in parkinson's disease (the adagio study): rationale, design and baseline characteristics. *Movement Disorders*, 23:2194–2201, 2008.

[25] C. W. Olanow, O. Rascol, R. A. Hauser, P. D. Feigin, J. Jankovic, A. Lang, W. Langstoni, E. Melamed, W. Poewe, F. Stocch, and E. Tolosa. A double-blind, delayed-start trial of rasagiline in parkinson's disease. *New England Jounral of Medicine*, 361:1268–1278, 2009.

[26] S. Pampallona and A. A. Tsiatis. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, 42:19–35, 1994.

[27] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society*, 135:185–206, 1972.

[28] S. Piantadosi. *Clinical Trials: A Methodological Perspective*. John Wiley and Sons LTD., New Jersey, 2005.

[29] S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199, 1977.

[30] V. Sebille and E. Bellissant. Sequential methods and group sequential designs for comparative clinical trials. *Fundamental and Clinical Pharmacology*, 17:505–516, 2003.

[31] A. H. V. Shapira and J. Obeso. Timing of treatment initiation in parkinson's disease: a need for reqppraisal? *Annals of Neurology*, 59:559–565, 2006.

[32] S. Snapinn, M. Chen, and T. Koutsoukos. Assessment of futility in clinical trials. *Pharmaceutical Statistics*, 5:273–281, 2006.

[33] I. van der Tweel and P. van Noord. Early stopping in clinical trials and epidemiologic studies for futility: Conditional power versus sequential analysis. *Journal of Clinical Epidemiology*, 56:610–617, 2003.

[34] S. K. Wang and A. A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43:193–200, 1987.