OPTIMAL SAMPLE SIZE DETERMINATION IN ADAPTIVE SEAMLESS PHASE II/III DESIGN

by

Zhongying Xu

B.A. in Radio and Television Journalism, Zhejiang University of Technology, China, 2010

M.A. in Communication and New Media, City University of Hong Kong, Hong Kong, 2011

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Zhongying Xu

It was defended on

April 13, 2016

and approved by

Thesis Advisor and Committee Co-Chair:
Chung-Chou H. Chang, PhD
Professor
Departments of Medicine and Biostatistics
School of Medicine and Graduate School of Public Health
University of Pittsburgh

Committee Co-Chair:
Gary M. Marsh, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
John A. Kellum, MD
Professor
Department of Critical Care Medicine
School of Medicine
University of Pittsburgh

Gary M. Marsh, PhD Chung-Chou H. Chang, PhD

OPTIMAL SAMPLE SIZE DETERMINATION IN ADAPTIVE SEAMLESS PHASE II/III DESIGN

Zhongying Xu, M.S.

University of Pittsburgh, 2016

ABSTRACT

The adaptive seamless phase II/III design combines the conventional separate phases II and III trials into a single trial, and it allows for adaptations (e.g. sample size reassessment and early stopping for futility or success) after the interim analysis. In this study, we propose a simulation-based method to determine the optimal sample size for the adaptive seamless phase II/III design. We assume that a power law relationship exists between the overall sample size and statistical power of the final test. The optimal sample size is defined as the minimum sample size that provides adequate power with overall type I error rate under control. To find the optimal size, we also take correlations between the early and the final outcomes into consideration. The methodology is applied to determining sample sizes in a study for a candidate treatment that can avoid renal damage during cardiac operations while the most effective dose of the treatment will be selected at the interim analysis.

PUBLIC HEALTH SIGNIFICANCE

Adaptive seamless phase II/III design eliminates the time between the traditional separate trials and better utilizes the data collected before the interim analysis, thus will result in faster clinical trials. Treatment effect can be confirmed at the final test if adequate power is achieved and the overall type I error rate is under control. Using these faster clinical trials, effective treatment can

be approved sooner to benefit more patients. In addition, in an adaptive seamless phase II/III design more patients will be allocated to the more effective treatment than they would in conventional clinical trials.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Traditional clinical trials are performed in several independent phases. Phase I is the first trial conducted among humans and the goal is to evaluate the safety of the treatment. In phase II usually less than 100 patients are involved and the goal is to select appropriate dose(s) of the study treatment and further assess the safety. In phase III, hundreds to thousands of patients can be included and the goal is to confirm the safety and effectiveness of the treatments. However, there are several issues violating statistical principles in conventional separated trials. Thall (2008) mentioned that in a conventional phase II trial, the comparison between the treatment and standard therapy was based on the observed data from a single-arm trial and a fixed estimator directly obtained from the historical data. The variance of the test statistics was underestimated since the variability of the estimator from historical data was ignored. He also pointed out that bias could be caused by patient heterogeneity, because patient covariates can make even larger combined effects than treatment in many clinical settings. In addition, phase II trial usually has small sample size with limited reliability and validity.

To improve the scientific reliability and efficiency of conventional phase II and phase III trials, adaptive seamless phase II/III design (ASD) was proposed, and it has become popular in the pharmaceutical industry (Chow and Chang, 2008). An ASD combines the conventional separate phases II and III trials into a single trial, and allows adaptations (e.g. sample size reassessment, and early stopping for futility or success) after the interim analysis at midterm

(Figure 1). Ellenberg and Eisenberger (1985) presented a similar concept as the two-stage phase II/III design but restricted their method to binary outcome for a single treatment versus placebo. Thall et al. (1988a) gave a formal and complete presentation of an ASD, while they considered the same outcomes in both stages. In the setting of evaluating several experimental treatments versus control group, Thall et al. (1988b) proposed a modified two-stage phase II/III design with treatment selection in stage 1 and two-arm comparison in stage 2. When final outcome is not available for all patients, early outcome is used for treatment selection at the interim analysis. Stallard (2010) and Friede et al. (2011) take the correlation of early and late outcomes into consideration in the analysis of stage 2, the confirmatory stage.

An ASD is more efficient since it eliminates the time between the trials conducted separately ("seamless"). Therefore, an ASD better utilizes the data before the interim analysis, and increases the total follow-up time for patients (Maca et al., 2006). The final analysis is conducted by using the data from patients enrolled before and after the midterm adaptation, so the total available sample sizes increases. Thus, it provides more reliable inferences than the traditional design.

We need to consider whether an ASD is the proper design to the study, for efficiency and ethical issues. Firstly, a study with relatively shorter length of time for decision making than recruitment speed is desirable for an ASD. Otherwise, it losses efficiency since we have to pause the enrollment during the study. Secondly, pivotal studies are required to provide adequate information about the treatment. Separate phase II trial will be recommended if too many unknown information on the treatment. Moreover, the surrogate marker for treatment selection and study endpoints must be validated and accepted (Gaydos et al., 2009).

A number of authors proposed statistical approaches to an ASD, while controlling the overall type I error rate at a prespecified level. Group-sequential method, adaptive Dunnett method, and combination test method are the three main approaches (Stallard & Todd, 2011). The null hypothesis for these methods is that the effectiveness of the study treatments is the same as that of the placebo. Stallard and Todd (2003) proposed a group-sequential method based on a cumulative normally distributed test statistics. Under this method, only one treatment could be selected to continue into the subsequent stage (stage 2). Koenig et al. (2008) proposed an adaptive Dunnett test based on a conditional error function. Other adaptations could be made in the subsequent stage after an interim analysis. Bauer and Kieser (1999) proposed a combination test method, where the final decision for rejection was made by a combination of stage 1 and stage 2 *P* values. Comparing with the group-sequential method, both the adaptive Dunnett method and the combination test method allow the selection of more than one treatment at the interim and other adaptations.

Based on the combination test method proposed by Bauer and Kieser, Posch et al. (2005) gave a general formulation of the adaptive testing procedure in the context of treatment selection. Friede et al. (2011) considered another setting when only information from stage 1 (early outcome) was available in the interim analysis for treatment selection, and confirmatory testing was exclusively based on the final outcomes.

A few methods have been developed to estimate the optimal sample size for an ASD while controlling the prespecified overall type I error rate and achieving adequate power of the overall test. Bischoff and Miller (2009) suggested that sample size reassessment for the stage 2 could be derived based on the estimated variance of outcome from the interim analysis. Fisher (1998) proposed a sample size estimation based on conditional power when no early stopping

was planned. On the other hand, Cui et al. (1999) gave an unconditional sample size calculation, and the new sample size was based on the assumed effect size and the estimated effect size from the observed data in the previous stage. However, this method might provide biased conditional power (Gao et al., 2008).

A recent trial of oral propranolol treatments on complicated infantile hemangiomas was conducted using the ASD (Léauté-Labrèze et al., 2015). The interim analysis of the study was performed after the first 188 patients completed 24 weeks. The best propranolol regimen (3mg/kg/day for 6 months) was then selected based on the early outcome (Léauté-Labrèze et al., 2015). In the final analysis, the effect of the selected treatment was confirmed using Posch's method. In this proposal, our aim was to conduct a simulation study to examine the overall Type I error rate and power of the final test based on the information given in the protocol of this study. Our simulation was based on the work of Posch et al. (2005) and Friede et al. (2011). Details of the method will be described in Section 2. After proposal and before graduation, we plan to develop a simulation-based empirical method to estimate the optimal sample size of an ASD and apply it to estimate the sample size for the remote ischemic conditioning to avoid renal damage during cardiac operations (RICARDO) study led by Dr. John Kellum in the Department of Critical Care Medicine. The optimal sample size determinations for an ASD under various scenarios will be given in Section 3. In Section 4, we will present our simulation work for the study of oral propranolol treatments on complicated infantile hemangiomas (Léauté-Labrèze et al., 2015). In Section 5, we will describe the future work of sample size determination for the RICARDO study. A final concluding remark will be given in Section 6.

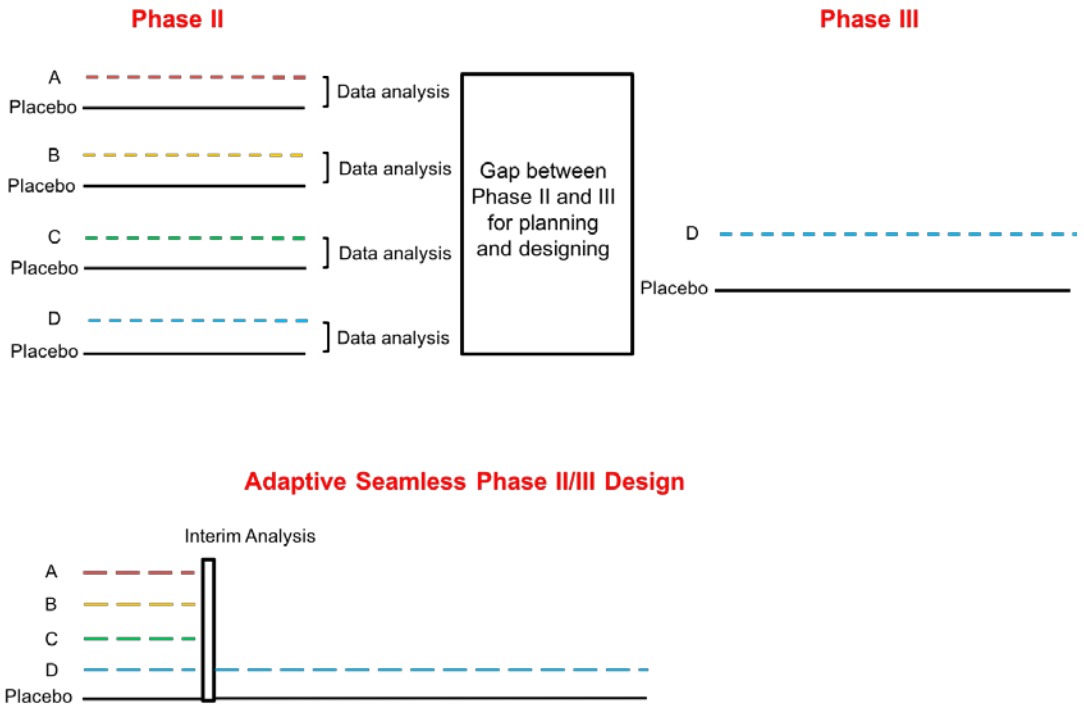

Figure 1. Comparison of traditional clinical trials and ASD.

Adapted from "Adaptive design clinical trials: Methodology, challenges and prospect," by R. Mahajan and K. Gupta, 2010, *Indian journal of pharmacology*, 42(4), 201. Copyright [2010] by Indian Journal of Pharmacology.

2.0 ADAPTIVE TREATMENT SELECTION BASED ON COMBINATION TESTS

An ASD usually contains two stages of traditional clinical trials, a learning stage (phase IIb) and a confirmatory stage (phase III). In phase IIb researchers evaluate efficacy and safety among several candidate treatments. In an ASD, researchers will select one or two treatments based on the interim analysis at the end of learning stage, and then those patients in the selected treatment groups along with those in the control group will continue into the second stage to confirm the estimation of treatment effects in the first stage.

To investigate the effectiveness of study treatment, we define the hypotheses for comparing the treatment success rate θ_i, $i \in \Omega_1 = \{1, \dots, k\}$ to the placebo success rate $\theta_0$be

$$H_i: \theta_i - \theta_0 \leq 0 \quad \text{against} \quad H_i': \theta_i - \theta_0 > 0, i \in \Omega_1.$$

To reject the elementary null hypothesis H_i, $i \in \Omega_1$, at overall type I error rate α, for all subsets $\mathcal{S} \subseteq \Omega_1$ that contain i, the intersection hypotheses H_S have to be rejected at level α, this is known as the closed testing principle (Marcus et al., 1976). For instance, there are three treatments A, B, and C, in order to reject the elementary null hypothesis H_A for treatment A at level α, we need to reject all intersection hypotheses containing treatment A at level α, here the intersection hypotheses are H_A, $H_A \cap H_B$, $H_A \cap H_C$, and $H_A \cap H_B \cap H_C$.

To demonstrate Posch's method of adaptive treatment selection, we first assume the outcome is binary data (e.g. success and failure) sampled from Bernoulli (θ_i) for k experimental treatments $T_1, \dots, T_k$ as well as a control group T_o at stage 1, each arm has n_1 observations. At

stage 2, we also assume binary outcome for the selected treatment T_i, $i \in \Omega_1$, and the control group T_o, with each arm including n_2 observations. Assuming balanced sample size, the overall sample size for each group is $n = n_1 + n_2$.

We denote $r_{j,i}$ the observed success rate of the i^{th} treatment at the j^{th} stage, $r_{j,0}$ the observed success rate for placebo. The standardized test statistic $Z_{j,i}$ under null for treatment $i \in \Omega_1$, at stage $j = 1, 2$ following normal distribution, and $Z_{j,i}$ is given by

$$Z_{j,i} = \frac{\sqrt{n_j}\left(r_{j,i} - r_{j,0}\right)}{\sqrt{2\bar{r}_{j,i}\left(1 - \bar{r}_{j,i}\right)}}, \quad \text{where } \bar{r}_{j,i} = (r_{j,i} + r_{j,0})/2 .$$

Let $p_i = 1 - \Phi(Z_{1,i})$ and $q_i = 1 - \Phi\left(Z_{2,i}\right)$ as p value for stage 1 and stage 2 respectively, where $\Phi(.)$ is the CDF of standard normal distribution. Next, we use Simes's method to get p values for intersection hypotheses $H_{\mathcal{S}}$.

$$p_{\mathcal{S}} = \min_{i \in \mathcal{S}} \frac{s}{l} \vec{p}_i ,$$

where l is the rank of p_i in the vector $\vec{p}$, and $\vec{p}$ includes the p values of all s elementary hypotheses H_i's in the intersection hypothesis $H_{\mathcal{S}}$, $i \in \mathcal{S}$.

We denote $p_{\mathcal{S}}$ as the stage 1 p value for intersection hypotheses $H_{\mathcal{S}}$ and $q_{\mathcal{S}}$ as the stage 2 p value for intersection hypotheses $H_{\mathcal{S}}$ obtained from the Simes test. For simplicity, we assume only one treatment is selected to continue into stage 2. We define $q_{\mathcal{S}} = q_{\mathcal{S} \cap \Omega_2}$ where $\Omega_2 \subseteq \Omega_1$ is the treatment selected for the second stage and $q_{\emptyset} = 1$, $q_{\mathcal{S}}$ will be equal to the p value for the selected treatment in stage 2.

To combine the p values from the two stages, we use the weighted inverse normal method to define a combination p values through the function:

$$C(p_i, q_i) = 1 - \Phi[\upsilon\Phi^{-1}(1 - p_i) + \omega\Phi^{-1}(1 - q_i)] ,$$

where $\upsilon = \sqrt{n_1/n}$, $\omega = \sqrt{n_2/n}$ such that $\upsilon^2 + \omega^2 = 1$.

We define the decision function of a combination test

$$\varphi_C(p_i, q_i) = \begin{cases} 1 & if\ p_i \le a\ or\ both\ p_i \le b\ and\ C(p_i, q_i) \le c \\ 0 & otherwise \end{cases}$$

Therefore, we can make a decision, reject the null for $\varphi_C = 1$, or not. Treatment will be stopped because of early rejection of the null hypothesis ($p_i \le a$), or futility ($p_i > b$). With prespecified *a* and *b*, *c* can be solved by

$$a + \int_a^b \int_0^1 1_{[C(x,y)\le c]}\, dydx = \alpha ,$$

where we define the indicator function

$$1_{[.]} = \begin{cases} 1 & if\ C(x, y) \le c \\ 0 & otherwise \end{cases}$$

In the final analysis, the null hypothesis H_i, $i \in \Omega_1$, will be rejected at familywise type I error rate α, if for each $\mathcal{S} \subseteq \Omega_1$, the intersection hypothesis $H_{\mathcal{S}}$ is rejected at level α, i.e. $\varphi_C(p_{\mathcal{S}}, q_{\mathcal{S}}) = 1$.

Friede et al. (2011) further extended the above method and took the correlation between early and final outcome into consideration. Generally, combination test with closed testing principle becomes more conservative when the correlation between early and final outcome decreases (Friede et al., 2011). Since early and final outcomes are usually different but correlated in many real applications, our simulation was based on the work of Posch et al. (2005) and Friede et al. (2011).

3.0 OPTIMAL SAMPLE SIZE DETERMINATION

Our objective is to determine the optimal sample size for an ASD through simulation using a prespecified type I error rate and power. The total sample size includes estimated sample size n_1 for stage 1 in each arm and reassessed sample size n_2 for stage 2 in each arm based on the interim analysis. We assume a "power law relationship" exists between the overall sample size and statistical power of the final test. A power law is a functional relationship between the two quantities, where one quantity varies as a power of another. That is,

$$P(N) = 1 + \pi N^{-\gamma},$$

where N is the total sample size, $P(N)$ is the statistical power, $\pi < 0$ and $\gamma > 0$ need to be estimated. In this study, we consider three different scenarios as described in detail below.

Paragraph.

3.1 SCENARIO I: ESTIMATE THE OPTIMAL SAMPLE SIZE N_1 FOR STAGE 1 WITH FIXED FUNCTIONAL RELATIONSHIP BETWEEN N_1 AND N_2.

Let N_1 be the total sample size at stage 1, $P(N_1)$ be the power of the final test which is a function of n_1, and $A(N_1)$ be the overall type I error rate which is a function of n_1. We assume that i) the

ratio of sample size in treatment vs. placebo is fixed; ii) the functional relationship between N and N_2, $N_2 = f(N_1)$, is fixed. Therefore, we can estimate the optimal total sample size N by

$$N_{1\ optimal} = min\{N_1 : P(N_1) > P^*, A(N_1) \leq \alpha^*\},$$

$$N_{optimal} = N_{1\ optimal} + f\left(N_{1\ optimal}\right),$$

where α^* is a prespecified overall type I error rate, and we plan to achieve at least P^* power.

For our future work, the RICARDO study fits this scenario. We need to assume a prespecified functional relationship between n_1 and n_2, and conduct simulations to determine the optimal value of n_1 and achieve adequate power with type I error rate under controlled.

3.2 SCENARIO II: ESTIMATE THE OPTIMAL SAMPLE SIZE N_2 FOR STAGE 2 WITH FIXED SAMPLE SIZE N_1 FOR STAGE 1.

With a fixed sample size N_1 for stage 1, we aim to estimate the minimum sample size N_2 for stage 2 in order to achieve the greatest power of the final test under a prespecified overall type I error rate. In other words, we are looking for the best adaptation method, i.e. we will choose the best function $g(.)$ to achieve the greatest power with minimum sample size, $N_2 = g(N_1)$ and $N_{optimal} = N_1 + g(N_1)$.

Sample size estimation in the protocol of Léauté-Labrèze et al. (2015) was performed under this scenario. They fixed stage 1 sample size per arm to be 35, then explored the power change as a function of N_2 with the one-sided overall type I error rate $\alpha = 0.005$.

3.3 SCENARIO III: ESTIMATE THE SET OF OPTIMAL SAMPLE SIZES $\{N_1, N_2\}$ WHERE N_1 AND N_2 ARE INDEPENDENT.

In this scenario, we assume that N_1 and N_2 are independent, thus no sample size adaptation method exists. We will find the set of optimal sample sizes $\{N_1, N_2\}$ when the greatest power is reached and the type I error rate is controlled. The power could be shown in a 3D space, where changing either N_1 or N_2 changes the power.

4.0 SIMULATION-BASED TYPE I ERROR RATE AND POWER

In the protocol of oral propranolol treatments on complicated infantile hemangiomas study (Léauté-Labrèze et al., 2015), four treatment regimens with a control group were considered. Only one treatment was selected and continued along with placebo in stage 2. They denoted n_1 as stage 1 sample size for each arm, n_2 as stage 2 sample size for each arm. Therefore, the minimum total sample size was $5n_1 + 2n_2$. From the interim analysis, the estimated success rates were: 10% for placebo, 30% for regimen 1 (1 mg/kg/day for 3 months), 35% for regimen 2 (1 mg/kg/day for 6 months), 40% for regimen 3 (3 mg/kg/day for 3 months), and 50% for regimen 4 (3 mg/kg/day for 6 months). The overall one-sided type I error rate was $\alpha = 0.005$ and power > 90% was considered adequate.

Using the method proposed by Posch et al. (2005), we simulated the type I error rates and powers under different sets of sample sizes and only the most effective regimen was chosen after the interim analysis. As shown in Tables 1 and 2, we got similar results as those shown in the protocol. We used the R package ASD (Parsons et al., 2012) to produce these results.

Table 1. Estimated one-sided type I error rate when the most effective regimen is chosen at the interim analysis with various (n_1, n_2).

Level (superiority) of Type I Error Rate					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.0047	0.0010	0.0013	0.0012	0.0012	35	40	255
0.0046	0.0010	0.0013	0.0011	0.0012	35	45	265
0.0046	0.0010	0.0013	0.0011	0.0012	35	50	275
0.0047	0.0010	0.0012	0.0014	0.0011	35	55	285
0.0047	0.0009	0.0012	0.0015	0.0011	35	60	295

Note. Results were given based on 10^4 simulations for each scenario.

Table 2. Estimated power when the most effective regimen is chosen at the interim analysis with various (n_1, n_2).

Power (superiority)					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.9431	0.0022	0.0108	0.0449	0.8852	35	40	255
0.9501	0.0022	0.0111	0.0462	0.8906	35	45	265
0.9544	0.0023	0.0115	0.0472	0.8934	35	50	275
0.9576	0.0023	0.0117	0.0480	0.8956	35	55	285
0.9610	0.0023	0.0120	0.0494	0.8973	35	60	295

Note. Results were given based on 10^4 simulations for each scenario.

5.0 RICARDO STUDY

Researches have found that remote ischemic preconditioning (RIPC), having transient external compression of the upper arm, prior to cardiac surgery is associated with reducing the occurrence of acute kidney injury (AKI) and is strongly associated with the release of cell-cycle arrest biomarkers into the urine (Zarbock, et al., 2015a; Kashani, et al., 2013; Bihorac, et al., 2009). John Kellum, MD proposed a multicenter, randomized, double-blind, adaptive seamless phase II/III trial entitled "Remote ischemic conditioning to avoid renal damage during cardiac operations (RICARDO)" to further investigate the effectiveness of this intervention, including an assessment of the effect size and the selection of optimal dose of RIPC.

By using the adaptive seamless phase II/III design, four regimens with different doses of RIPC were used along with the control group (sham-RIPC intervention) at stage 1. In the interim analysis, we will select the regimen that yields the highest proportion of urinary [TIMP-2]•[IGFBP7] $\geq$ 0.5 $(ng/ml)^2/1000$ and without a major adverse event (including any increase in AKI). The selected regimen and the control group will be used in stage 2 to determine the overall effectiveness. The primary endpoint is the major adverse kidney events at 90-days post surgery ($MAKE_{90}$), including death and dialysis or persistent renal dysfunction (2x baseline creatinine). In order to determining the optimal sample size, we take into account the correlation between the early outcome (proportion of biomarker [TIMP-2]•[IGFBP7] $\geq$ 0.5 $(ng/ml)^2/1000$) and the final

outcome ($MAKE_{90}$). In addition, we utilized closed testing procedure together with combination tests to control the family-wise one-sided type I error rate to $\alpha = 0.005$.

In the pilot study conducted by Zarbock, et al.(2015b), patients were randomized into the intervention group and the control group. Patients in the intervention group were treated with 3 cycles of 5-min inflation of a blood-pressure cuff to 200 mm Hg (or at least to a pressure 50 mmHg higher than the systolic arterial pressure) on one upper arm; each inflation is followed by a 5-min reperfusion with the cuff deflated. This intervention is considered as regimen 2 in the RICADO study. Patients in the control group are treated with 3 cycles of upper limb 'pseudo'-ischemia at a lower pressure; in each cycle a 5-min blood-pressure cuff inflation to a pressure of 20 mm Hg higher than the systolic arterial pressure is followed by a 5-min cuff deflation. In the RICARDO study, regimen 1 has a lower dose (2 cycles for 3 minutes) than regimen 2 dose; regimen 3 and regimen 4 have higher doses with 3 cycles of 7 minutes and 4 cycles of 5 minutes, respectively.

We estimated the proportion of $MAKE_{90}$ for sham-RIPC ($\hat{\theta}_0 = 0.25$) and regimen 2 ($\hat{\theta}_2 = 0.14$) from the previous study (Zarbock, et al., 2015b). We further assumed the proportion of $MAKE_{90}$ for regimen 1, 3, and 4 as $\hat{\theta}_1 = 0.2$, $\hat{\theta}_3 = 0.12$, and $\hat{\theta}_4 = 0.10$, respectively. The early outcome is defined as the proportion of urinary [TIMP-2]•[IGFBP7] ≥ 0.5 $(ng/ml)^2/1000$ for each regimen and for the control. From the pilot data, we estimated this proportion as $\hat{p}_0 = 0.25$ for sham-RIPC and $\hat{p}_2 = 0.37$ for regimen 2. For regimen 1, 3, and 4, we assumed the proportion as $\hat{p}_1 = 0.31$, $\hat{p}_3 = 0.39$, and $\hat{p}_4 = 0.41$ respectively. Moreover, we estimated the correlation between early outcome and $MAKE_{90}$ from the pilot data, and it is denoted by phi coefficient, $\hat{\phi} = -0.15$. We plan for having equal sample size per arm, and we expected the overall power to be at least 80%.

Sample size estimation for the RICARDO study is performed under scenario I as mentioned in Section 3. We do not have enough information about the required sample size for stage 1, so we explored different values of n_1 in our simulations. Moreover, we controlled the ratio of stage 2 sample size (n_2) to stage 1 sample size (n_1) in between 5 and 20. For each value of n_1, we estimated the total sample size with four different ratios of n_2 to n_1. The results are shown in Tables 3 to 8.

Table 3. Estimated one-sided type I error rate when the most effective regimen is chosen at the interim analysis with n_1 = 15 and different n_2.

Level (superiority) of Type I Error Rate					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.0038	0.0004	0.0009	0.0016	0.0009	15	150	375
0.0040	0.0004	0.0009	0.0017	0.0010	15	200	475
0.0040	0.0003	0.0010	0.0016	0.0011	15	250	575
0.0042	0.0003	0.0012	0.0016	0.0011	15	300	675

Note. Results were given based on 10^4 simulations for each scenario.

Table 4. Estimated power when the most effective regimen is chosen at the interim analysis with n_1 = 15 and different n_2.

Power (superiority)					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.5565	0.0058	0.0973	0.1775	0.2759	15	150	375
0.6865	0.0071	0.1335	0.2258	0.3201	15	200	475
0.7709	0.0094	0.1618	0.2584	0.3413	15	250	575
0.8217	0.0109	0.1830	0.2781	0.3497	15	300	675

Note. Results were given based on 10^4 simulations for each scenario.

Table 5. Estimated one-sided type I error rate when the most effective regimen is chosen at the interim analysis with n_1 = 20 and different n_2.

Level (superiority) of Type I Error Rate					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.0040	0.0004	0.0009	0.0018	0.0009	20	150	400
0.0038	0.0004	0.0009	0.0016	0.0009	20	200	500
0.0039	0.0004	0.0009	0.0017	0.0009	20	250	600
0.0040	0.0004	0.0010	0.0015	0.0011	20	300	700

Note. Results were given based on 10^4 simulations for each scenario.

Table 6. Estimated power when the most effective regimen is chosen at the interim analysis with n_1 = 20 and different n_2.

Power (superiority)					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.5779	0.0047	0.0977	0.1834	0.2921	20	150	400
0.7054	0.0058	0.1316	0.2303	0.3377	20	200	500
0.7869	0.0079	0.1588	0.2624	0.3578	20	250	600
0.8376	0.0091	0.1802	0.2812	0.3671	20	300	700

Note. Results were given based on 10^4 simulations for each scenario.

Table 7. Estimated one-sided type I error rate when the most effective regimen is chosen at the interim analysis with n_1 = 30 and different n_2.

Level (superiority) of Type I Error Rate					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.0038	0.0003	0.0009	0.0018	0.0008	30	150	450
0.0040	0.0004	0.0009	0.0018	0.0009	30	200	550
0.0039	0.0004	0.0009	0.0017	0.0009	30	250	650
0.0038	0.0004	0.0009	0.0016	0.0009	30	300	750

Note. Results were given based on 10^4 simulations for each scenario.

Table 8. Estimated power when the most effective regimen is chosen at the interim analysis with n_1 = 30 and different n_2.

Power (superiority)					Sample Size		
Overall	Regimen 1	Regimen 2	Regimen 3	Regimen 4	n_1	n_2	min(N)
0.6152	0.0036	0.0969	0.1943	0.3204	30	150	450
0.7364	0.0048	0.1278	0.2394	0.3644	30	200	550
0.8154	0.0068	0.1550	0.2695	0.3841	30	250	650
0.8633	0.0079	0.1749	0.2875	0.3930	30	300	750

Note. Results were given based on 10^4 simulations for each scenario.

From Tables 3-8, the overall type I error rate was controlled under 0.005 for all settings. We found that statistical power increases as total sample size gets larger. The power of the final test achieves 80% when total sample size is closed to 650. Based on these results, we suggest the acceptable sample size to be 675 patients in total, with recruiting 15 patients per arm in stage 1 and 300 patients per arm in stage 2.

Moreover, we compared the required stage 1 sample size using an ASD with the required sample size using other four traditional phase II trial designs: the two-arm trial design, the group-sequential design, and the multi-arm trial designs with and without treatment selection. We used the early outcome (the proportion of urinary [TIMP-2]•[IGFBP7] $\geq$ 0.5 $(ng/ml)^2/1000$) as the endpoint for hypothesis testing in the four traditional designs. The same settings as those in the RICARDO study were used, which include four candidate treatment regimens and a control group. Sample size estimations were done using the R package ASD and the R packages MAMS (Magirr et al., 2012). The results of the estimated sample sizes using different trial designs are summarized in Table 9. If we conduct four separate two-armed trials, the required total sample size is 88; if we conduct four separate group-sequential trials, the maximum required sample size is 112; if we use a multi-arm study, 90 and 75 patients with and without treatment selection, respectively, will be required for the design. As shown in Tables 3-8, we only need 75 patients for stage 1 if using an ASD. These results showed that an ASD can reduce sample size for stage 1 and allocate more patents to more effective treatment in stage 2.

Table 9. Comparison of maximum sample size required for different designs.

Design Option	Maximum Sample Size
Four separate two-armed trials	88
Four separate group-sequential trials	112
Multi-arm study with treatment selection	90
Multi-arm study without treatment selection	75
Adaptive seamless design (stage 1)	75

We checked the power law assumption by plotting estimated total sample size and power under twelve twelve different settings of n_1 and n_2 in Tables 3-8 and comparing their deviation from the theoretical curve. Figure 2 shows that the points scattered around the theoretical curve with relatively small departures. We concluded that a power law relationship between total sample size and statistical power is probable.

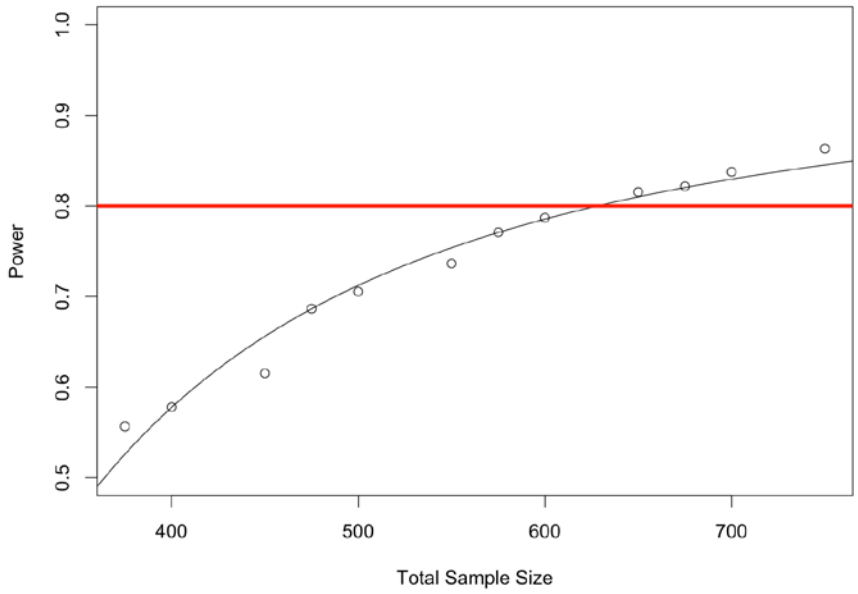

Figure 2. Power law relationship between total sample size and statistical power.

6.0 DISCUSSION

For a trial with complex setting such as an adaptive seamless design, it is more feasible to determine the sample size through simulations than deriving the closed form of the sample size formula. Posch et al. (2005) proposed an empirical method to estimate the effect of the selected treatment. They used the closed testing principle together with the combination tests in the final analysis. Friede et al. (2011) extended the method of Posch et al. and took the correlation of early and final outcomes into consideration. Our simulations were based on the work of both Posch et al. and Friede et al. and a power law relationship between the total sample size and the statistical power was assumed.

Based on the information provided in the protocol of the trial of oral propranolol in infantile hemangioma (Léauté-Labrèze et al., 2015), we estimated the first-stage and the second-stage sample sizes by controlling the overall type I error rate and the power of the final test, using the method of Posch et al. (2005). For the RICARDO study, we applied the same method to estimate the required total sample size via simulations. The power law relationship is a reasonable assumption since the statistical power is expected to converge with infinitely large sample size. With a prespecified functional relationship between the sample sizes in stages 1 and in stage 2 ($5 \leq n_2/n_1 \leq 20$), we estimated the total required sample size by assuming different values of n_1. Based on the simulation results, we suggested the total required sample size to be 675 patients under the condition that the overall type I error rate is less than 0.005 and the overall

power is greater than 80%. Moreover, our simulation results showed that the required sample size in the RICARDO study is smaller if an ASD is used in stage 1, as compared with the sample size required by other traditional phase II designs.

An ASD not only shortens the total length of time, but it also can assign more patients to the more effective treatment. It is worth noting that despite many strengths of an ASD, it may not always be the best choice. When there is a lack of information about the treatment of interest, the final treatment recommendation derived from a study using ASD could be misleading. In addition, an ASD will lose its efficiency if the recruitment rate is relatively fast with respect to the waiting time for the interim analysis. In the RICARDO study, we had obtained sufficient information about the treatment of interest from pilot studies, and the expected waiting time for decision making is short relative to recruitment time. Therefore, it is feasible and reasonable to use an adaptive seamless design in the RICARDO study.

One of the limitations in our sample size estimation for the RICARDO study is that we did not simulate an intensive set of (n_1, n_2), sample sizes for stage 1 and stage 2 within a prespecified range of the total sample size to check the power law assumption. Currently, we fitted a curve with twelve points of (n_1, n_2) with a total sample size ranging from 375 to 750 patients. Simulations with more data points of (n_1, n_2) will be needed to capture the lower part features and the convergence limit of the power law curve.

Besides, a statistical study is required to assess the fit of the power law curve. We will need to investigate the goodness of fit of the power law relationship. An alternative way to assess the fit is to conduct multiple simulations and generate an interval of power estimates for each sample size (e.g. with the first and the third quartiles being the lower and upper ends of the

interval), then we can check whether the power law curve fitted by the median powers will pass through all intervals.

We have tried different packages in R and different macros in SAS for sample size and power estimations. Among them, the R package ASD performs the best in replicability also takes the correlation between the early and the final outcomes into account. Our simulations showed that package ASD gave different estimation results for different seed numbers. To overcome the variability, we suggest running simulations with random choices of different seed numbers and reporting the median of the estimated sample sizes.

APPENDIX: R CODE FOR SAMPLE SIZE SIMULATION IN THE RICARDO STUDY

```
### Simulate sample size in ASD
library(asd)
# Estimated one-sided type I error rate when the most effective regimen is chosen at the interim analysis
result_type1<- treatsel.sim(n=list(stage1=15, stage2=150),
                    effect=list(early=rep(0.75,5), final=rep(0.25,5)),
                    outcome=list(early="B", final="B"), nsim=10000,
                    corr=0.15, seed=4358098, select=1,
                    weight=NULL, level=0.005, ptest=c(1,4),
                    method="invnorm", fu=FALSE, file = "")
# Estimated power when the most effective regimen is chosen at the interim analysis
result_p<- treatsel.sim(n=list(stage1=15, stage2=150),
                  effect=list(early=c(0.75,0.69,0.63,0.61,0.59), final=c(0.25,0.2,0.14,0.12,0.10)),
                  outcome=list(early="B", final="B"), nsim=10000,
                  corr=0.15, seed=4358098, select=1,
                  weight=NULL, level=0.005, ptest=c(1:4),
                  method="invnorm", fu=FALSE, file = "")
```

```
### Simulate sample size in other study designs
library(MAMS)
delta <- 6
sigma <- sqrt(25)
# four separate two-arm
mams.rev(K=1, J=1, alpha=0.025, power=0.8, r=1, r0=1,
        p=pnorm(delta/(sqrt(2)*sigma)) , p0=0.5)
# four separate group-seq
mams.rev(K=1, J=2, alpha=0.025, power=0.8, r=1:2, r0=1:2,
        p=pnorm(delta/(sqrt(2)*sigma)) , p0=0.5,
        u.shape="triangular", l.shape="triangular")
# multi-arm w/o selection
mams(K=4, J=1, alpha=0.025, power=0.8, r=1, r0=1,
      p=pnorm(delta/(sqrt(2)*sigma)) , p0=0.5)
# multi-arm with treatment selection
mams(K=4, J=2, alpha=0.025, power=0.8, r=1:2, r0=1:2,
    p=pnorm(delta/(sqrt(2)*sigma)) , p0=0.5,
    u.shape="triangular", l.shape="triangular")
```

BIBLIOGRAPHY

Bauer, P., & Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in medicine*, 18(14), 1833-1848.

Bihorac, A., Yavas, S., Subbiah, S., Hobson, C. E., Schold, J. D., Gabrielli, A., Layon, A.J., & Segal, M. S. (2009). Long-term risk of mortality and acute kidney injury during hospitalization after major surgery. *Annals of surgery*, 249(5), 851-858.

Bischoff, W., & Miller, F. (2009). A seamless phase II/III design with sample-size re-estimation. *Journal of biopharmaceutical statistics*, 19(4), 595-609.

Chow, S. C., & Chang, M. (2008). Adaptive design methods in clinical trials-a review. *Orphanet Journal of Rare Diseases*, 3(11), 169-90.

Cui, L., Hung, H. J., & Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, 853-857.

Ellenberg, S. S., & Eisenberger, M. A. (1985). An efficient design for phase III studies of combination chemotherapies. *Cancer treatment reports*, 69(10), 1147-1154.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in medicine*, 17(14), 1551-1562.

Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes Marquez, E., Chataway, J., & Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine*, 30(13), 1528-1540.

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

Gaydos, B., Anderson, K. M., Berry, D., Burnham, N., Chuang-Stein, C., Dudinak, J., Fardipour, P., Gallo, P.,Givens, S., Lewis, R., Maca, J., Pinheiro, J., Pritchett, Y., & Krams, M. (2009). Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal*, 43(5), 539-556.

Kashani, K., Al-Khafaji, A., Ardiles, T., Artigas, A., Bagshaw, S. M., Bell, M., Bihorac, A., Birkhahn,R., Cely, C.M., Chawla, L.S., Davison, D.L., Feldkamp, T., Forni, L.G., Gong, M.N., Gunnerson, K.J., Haase, M., Hackett, J., Honore, P.M., Hoste, E.AJ., Joannes-Boyau, O., Joannidis, M., Kim, P., Koyner, J.L., Laskowitz, D.T., Lissauer, M.E., Marx,

G., McCullough, P.A., Mullaney, S., Ostermann, M., Rimmelé, T., Shapiro, N.I., Shaw, A.D., Shi, J., Sprague, A.M., Vincent, J., Vinsonneau, C., Wagner, L., Walker, M.G., Wilkerson, R.G., Zacharowski, K., & Kellum, J.A. (2013). Discovery and validation of cell cycle arrest biomarkers in human acute kidney injury. *Critical Care,* 17(1), R25.

Koenig, F., Brannath, W., Bretz, F., & Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statistics in Medicine,* 27(10), 1612-1625.

Léauté-Labrèze, C., Hoeger, P., Mazereeuw-Hautier, J., Guibaud, L., Baselga, E., Posiunas, G., Phillips, R.J., Caceres, H., Lopez Gutierrez, J.C., Ballona, R., Friedlander, S.F., Powell, J., Perek, D., Metz, B., Barbarot, S., Maruani, A., Szalai, Z.Z., Krol, A., Boccara, O., Foelster-Holst, R., Febrer Bosch, M.I., Su, J., Buckova, H., Torrelo, A., Cambazard, F., Grantzow, R., Wargon, O., Wyrzykowski, D., Roessler, J., Bernabeu-Wittel, J., Valencia, A.M., Przewratil, P., Glick, S., Pope, E., Birchall, N., Benjamin, L., Mancini, A.J., Vabres, P., Souteyrand, P., Frieden, I.J., Berul, C.I., Mehta, C.R., Prey, S., Boralevi, F., Morgan, C.C., Heritier, S., Delarue, A., & Voisard, J. J. (2015). A randomized, controlled trial of oral propranolol in infantile hemangioma. *New England Journal of Medicine*, 372(8), 735-746.

Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., & Krams, M. (2006). Adaptive seamless phase II/III designs-background, operational aspects, and examples. *Therapeutic Innovation & Regulatory Science,* 40(4), 463.

Magirr, D., Jaki, T., & Whitehead, J. (2012). A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*,99(2), 494-501.

Mahajan, R., & Gupta, K. (2010). Adaptive design clinical trials: Methodology, challenges and prospect. *Indian journal of pharmacology*, 42(4), 201.

Marcus, R., Eric, P., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3), 655-660.

Parsons, N., Friede, T., Todd, S., Marquez, E. V., Chataway, J., Nicholas, R., & Stallard, N. (2012). An R package for implementing simulations for seamless phase II/III clinical trials using early outcomes for treatment selection. *Computational Statistics & Data Analysis*, 56(5), 1150-1160.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., & Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine,* 24, 3697–3714.

Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in medicine,* 29(9), 959-971.

Stallard, N., & Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in medicine,* 22(5), 689-703.

Stallard, N., & Todd, S. (2011). Seamless phase II/III designs. *Statistical methods in medical research*, 20(6), 623-634.

Thall, P. F., Simon, R., Ellenberg, S. S., & Shrager, R. (1988a). Optimal two-stage designs for clinical trials with binary response. *Statistics in medicine*, 7(5), 571-579.

Thall, P. F., Simon, R., & Ellenberg, S. S. (1988b). Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, 75(2), 303-310.

Thall, P. F. (2008). A review of phase 2–3 clinical trial designs. *Lifetime data analysis*, 14(1), 37-53.

Zarbock, A., Schmidt, C., Van Aken, H., Wempe, C., Martens, S., Zahn, P. K., Wolf, B., Goebel, U., Schwer, C.I., Rosenberger, P., Haeberle, H., Görlich, D., Kellum, J.A., & Meersch, M. (2015a). Effect of remote ischemic preconditioning on kidney injury among high-risk patients undergoing cardiac surgery: a randomized clinical trial. *JAMA*, 313(21), 2133-2141.

Zarbock, A., Kellum, J., Van Aken, H., Schmidt, C., Martens, S., Görlich, D., & Meersch, M. (2015b). Long-term effects of remote ischaemic preconditioning in high risk patients undergoing cardiac surgery: follow-up of a randomised clinical trial. *Intensive Care Medicine Experimental*, 3(Suppl 1), A411.